ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ
ΤΟΜΕΑΣ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΕΠΙΧΕΙΡΗΣΙΑΚΗΣ ΕΡΕΥΝΑΣ

Βασίλειος Μουλόπουλος

Μια ανασκοπηση του intraclass μοντελου και του intraclass συντελεστη συσχετισης για οικογενειακα δεδομενα

Μεταπτυχιακή Διατριβή

Ιωάννινα, 2019

Η παρούσα Μεταπτυχιακή Διατριβή εκπονήθηκε στο πλαίσιο των σπουδών για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική και Επιχειρησιακή Έρευνα που απονέμει το Τμήμα Μαθηματικών του Πανεπιστημίου Ιωαννίνων.

Εγκρίθηκε την 26/06/2019 από την εξεταστική επιτροπή:

Ονοματεπώνυμο	**Βαθμίδα**
Κωνσταντίνος Ζωγράφος	Καθηγητής
Σωτήριος Λουκάς	Καθηγητής
Απόστολος Μπατσίδης	Επίκουρος Καθηγητής

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ

"Δηλώνω υπεύθυνα ότι η παρούσα διατριβή εκπονήθηκε κάτω από τους διεθνείς ηθικούς και ακαδημαϊκούς κανόνες δεοντολογίας και προστασίας της πνευματικής ιδιοκτησίας. Σύμφωνα με τους κανόνες αυτούς, δεν έχω προβεί σε ιδιοποίηση ξένου επιστημονικού έργου και έχω πλήρως αναφέρει τις πηγές που χρησιμοποίησα στην εργασία αυτή."

Βασίλειος Μουλόπουλος

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα Μεταπτυχιακή Διατριβή εκπονήθηκε στον Τομέα Στατιστικής και Επιχειρησιακής Έρευνας του Τμήματος Μαθηματικών της Σχολής Θετικών Επιστημών του Πανεπιστημίου Ιωαννίνων, με επιβλέποντα τον Καθηγητή κ. Κωνσταντίνο Ζωγράφο. Αρχικά, θα ήθελα να ευχαριστήσω τον Καθηγητή κ. Κωνσταντίνο Ζωγράφο που πρόθυμα δέχτηκε να αναλάβει την επίβλεψη και εισηγήθηκε το θέμα της διατριβής. Τον ευχαριστώ για την καθοδήγησή του και τον πολύτιμο χρόνο που διέθεσε. Ακόμη, θα ήθελα να ευχαριστήσω τον κ. Σωτήριο Λουκά, Καθηγητή του Τμήματος Μαθηματικών του Πανεπιστημίου Ιωαννίνων και τον κ. Απόστολο Μπατσίδη, Επίκουρο Καθηγητή του Τμήματος Μαθηματικών του Πανεπιστημίου Ιωαννίνων για τη συμμετοχή τους στην Τριμελή Επιτροπή αξιολόγησης της εργασίας. Επιπλέον, ευχαριστώ όλα τα μέλη του Τομέα Στατιστικής και Επιχειρησικής Έρευνας του Τμήματος Μαθηματικών του Πανεπιστημίου Ιωαννίνων για την καθοδήγησή τους και τις γνώσεις που μου παρείχαν όλα αυτά τα χρόνια. Θέλω να ευχαριστήσω θερμά τους φίλους και συναδέλφους μου μεταπτυχιακούς φοιτητές Θεόδωρο Εβρένογλου και Γεώργιο Καρακατσούλη για την υποστήριξη και τις συμβουλές τους. Επίσης, ευχαριστώ τους φίλους μου για τη συμπαράσταση που μου παρείχαν σε όλη τη διάρκεια εκπόνησης της διατριβής. Ακόμη, θα ήθελα να ευχαριστήσω την Βάγια Κυριαζοπούλου για την υποστήριξή της σε περιόδους πίεσης. Τέλος, δε θα μπορούσα να μην ευχαριστήσω την μητέρα μου Ευμορφία και τον αδερφό μου Αντώνη για κάθε είδους συμπαράσταση από μεριάς τους και τον πατέρα μου Γεώργιο, που παρόλο που έχει φύγει από τη ζωή, έχει συμβάλλει τα μέγιστα για να φτάσω ως εδώ.

ΠΕΡΙΕΧΟΜΕΝΑ

Κεφάλαιο 1

ΕΙΣΑΓΩΓΗ

Ο intraclass (ενδοκατηγορικός) συντελεστής συσχέτισης χρησιμοποιείται συνήθως ως μέτρο του βαθμού της σχέσης των μελών της ίδιας οικογένειας, ως προς κάποιο βιολογικό, περιβαλλοντολογικό ή ψυχολογικό χαρακτηριστικό, όπως για παράδειγμα το επίπεδο πίεσης του αίματος, το ύψος, ο δείκτης ευφυίας (βλ. Donner and Koval, 1980a). Η έννοια του intraclass συντελεστή συσχέτισης χρησιμοποιείται στα πλαίσια της εκτίμησης της κληρονομικότητας (heritability). Η κληρονομικότητα (heritability) είναι ένα στατιστικό, που χρησιμοποιείται στα πλαίσια της γενετικής και εκτιμά το βαθμό της μεταβλητότητας ενός χαρακτηριστικού σε ένα πληθυσμό, που οφείλεται στη γενετική μεταβλητότητα, που υπάρχει μεταξύ ατόμων στον πληθυσμό (βλ. Wray and Visscher, 2008). Μελέτες κληρονομικότητας δίνουν τη δυνατότητα να απαντήσουμε σε ερωτήματα όπως για παράδειγμα πόσο μεγάλο ρόλο παίζουν οι γενετικοί παράγοντες σε διαφορές που υπάρχουν στο ύψος μεταξύ ανθρώπων. Το παραπάνω ερώτημα είναι φυσικά διαφορετικό από το πόσο επηρεάζουν οι γενετικοί παράγοντες το ύψος οποιουδήποτε ανθρώπου. Ακόμη, ο intraclass συντελεστής συσχέτισης στη γενετική παίζει κεντρικό ρόλο στην εκτίμηση της κληρονομικότητας επιλεγμένων χαρακτηριστικών σε ζωϊκούς και φυτικούς πληθυσμούς (βλ. Donner, 1986). Ο κύριος σκοπός της εργασίας αυτής είναι να παρουσιαστεί μια ανασκόπηση της σημαντικότερης βιβλιογραφίας που είναι διαθέσιμη γύρω από συμπερασματολογία για τον intraclass συντελεστή συσχέτισης για οικογενειακά δεδομένα (family data).

Ο όρος της intraclass συσχέτισης συναντάται στο βιβλίο του Fisher (1954). Εκεί, παρουσιάζεται ο παλαιότερος τρόπος προσέγγισης του προβλήματος της εκτίμησης του intraclass συντελεστή συσχέτισης, την οποία επιγραμματικά παρουσιάζουμε στη συνέχεια. Πιο συγκεκριμένα και ακολουθώντας τον Fisher (1954, σ. 211), ας υποθέσουμε ότι διαθέτουμε n ζεύγη παρατηρήσεων $(X_1, X_1'), \ldots, (X_n, X_n')$. Τα $(X_i, X_i'), i = 1, \ldots, n$ είναι μετρήσεις που αφορούν αδέρφια n το πλήθος οικογενειών ως προς κάποιο βιολογικό χαρακτηριστικό. Σκοπός είναι η εκτίμηση της συσχέτισης μεταξύ του βιολογικού χαρακτηριστικού των δύο αδελφών με τη βοήθεια των n-παρατηρήσεων. Σύμφωνα με τον Fisher (1954, σ. 211) ένας τρόπος αντιμετώπισης του προβλήματος θα ήταν να χωρίσουμε τις μετρήσεις σε δύο κλάσεις σύμφωνα με κάποιο κριτήριο, όπως η ηλικία και να υπολογίσουμε τον συνήθη εκτιμητή του Pearson για τα ζεύγη των παρατηρήσεων. Ωστόσο, εάν ένας τέτοιος διαχωρισμός δεν εξυπηρετεί τους σκοπούς της έρευνάς μας, μιας και οι μετρήσεις αφορούν αδέρφια, τότε σύμφωνα με τον Fisher (1954, σ. 212) το πρόβλημα της εκτίμησης της συσχέτισης μπορεί να αντιμετωπιστεί με τον υπολογισμό των παρακάτω στατιστικών,

$$\bar{X} = \frac{1}{2n}\sum_{i=1}^{n}(X_i + X_i'),$$
$$S^2 = \frac{1}{2n}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(X_i' - \bar{X})^2\right],$$
$$r = \frac{1}{nS^2}\left[\sum_{i=1}^{n}(X_i - \bar{X})(X_i' - \bar{X})\right].$$

Τότε το στατιστικό r είναι ο εκτιμητής της intraclass συσχέτισης μεταξύ των παρατηρήσεων. Η παραπάνω σχέση μπορεί να επεκταθεί και για την περίπτωση όπου έχουμε μετρήσεις που αφορούν οικογένειες αποτελούμενες από k παιδιά. Έτσι, εάν X_{ij} παριστάνει τη j-οστή παρατήρηση της i-οστής οικογένειας ($i = 1, \ldots, n, j = 1, \ldots, k$), τότε σύμφωνα με τον Harris (1913), ο εκτιμητής της intraclass συσχέτισης μπορεί να υπολογιστεί από την παρακάτω εξίσωση,

$$k \sum_{i=1}^{n} (\bar{X}_i - \bar{X})^2 = nS^2 \left[1 + (k-1)r\right],$$

όπου

$$\bar{X}_i = \frac{1}{k} \sum_{j=1}^{k} X_{ij}, \quad i = 1, 2, \ldots, n,$$

$$\bar{X} = \frac{1}{nk} \sum_{i=1}^{n} \sum_{j=1}^{k} X_{ij},$$

$$S^2 = \frac{1}{nk} \sum_{i=1}^{n} \sum_{j=1}^{k} (X_{ij} - \bar{X})^2.$$

Ο εκτιμητής r είναι γνωστός στη βιβλιογραφία ως pairwise (κατά ζεύγη) εκτιμητής. Μπορεί να χρησιμοποιηθεί και για την περίπτωση, όπου το μέγεθος των οικογενειών ποικίλλει από οικογένεια σε οικογένεια. Στο πλαίσιο αυτό, έστω ότι έχουμε n οικογένειες και k_i παρατηρήσεις $X_{i1}, \ldots, X_{ik_i}, i = 1, \ldots, n$ είναι διαθέσιμες για κάθε οικογένεια. Σχηματίζουμε όλα τα δυνατά ζεύγη $(X_{il}, X_{ij}), l \neq j$, για κάθε οικογένεια. Να σημειωθεί ότι για κάθε (X_{il}, X_{ij}), σχηματίζεται και το ζεύγος που προκύπτει με εναλλαγή των συνιστωσών του, δηλαδή το (X_{ij}, X_{il}) (Fieller and Smith, 1951). Για κάθε οικογένεια θα σχηματιστούν $k_i(k_i - 1)$ ζεύγη παρατηρήσεων. Τότε υπολογίζεται ο συντελεστής συσχέτισης του Pearson (product-moment correlation) για όλα τα ζεύγη παρατηρήσεων. Δηλαδή ο intraclass συντελεστής συσχέτισης θα δίνεται από τη σχέση (βλ. Karlin et al., 1981),

$$r = \frac{\sum\limits_{i=1}^{n} \sum\limits_{j=1}^{k_i} \sum\limits_{\substack{l=1 \\ j \neq l}}^{k_i} (X_{ij} - \bar{X})(X_{il} - \bar{X})}{\sum\limits_{i=1}^{n} (k_i - 1) \sum\limits_{j=1}^{k_i} (X_{ij} - \bar{X})^2},$$

όπου

$$\bar{X} = \frac{1}{\sum\limits_{i=1}^{n} k_i(k_i - 1)} \sum_{i=1}^{n} \sum_{j=1}^{k_i} (k_i - 1) X_{ij}.$$

Ωστόσο, ο παραπάνω εκτιμητής δεν είναι κατάλληλος για μελέτες στις οποίες υπάρχει ουσιώδης μεταβλητότητα μεταξύ των μεγεθών των οικογενειών, k_i. Αυτό συμβαίνει επειδή οι μεγαλύτερες σε μέγεθος οικογένειες θα συνεισφέρουν ένα δυσανάλογο αριθμό ζευγών, που θα οδηγήσει σε μια παραπλανητική τιμή για τη συσχέτιση. Δηλαδή, ο παραπάνω εκτιμητής δίνει πολύ μεγάλο βάρος στις μεγάλες σε μέγεθος οικογένειες (βλ. Fieller and Smith, 1951). Για παράδειγμα, μια οικογένεια μεγέθους 2 συνεισφέρει 2 ζεύγη, ενώ από μια οικογένεια με 5 παιδιά θα προκύψουν 20 ζεύγη, χωρίς αυτό να σημαίνει ότι η δεύτερη οικογένεια δίνει και την ανάλογη ποσότητα πληροφορίας.

Όπως προαναφέρθηκε, η εργασία αυτή στοχεύει σε μια ανασκόπηση της υπάρχουσας βιβλιογραφίας της σχετικής με την ανάπτυξη συμπερασματολογίας (εκτίμηση, διαστήματα εμπιστοσύνης, τεστ) για τον intraclass συντελεστή συσχέτισης για οικογενειακά δεδομένα (familial data). Το πρόβλημα αντιμετωπίζεται από τη σκοπιά και με βάση τις μεθόδους της πολυδιάστατης ανάλυσης καθώς επίσης και αυτής της ανάλυσης διακύμανσης με την υιοθέτηση ενός μοντέλου τυχαίων επιδράσεων (one-way random effect model).

Στο πλαίσιο αυτό, στο 2ο κεφάλαιο εξετάζεται το πρόβλημα από τη σκοπιά της πολυδιάστατης ανάλυσης. Ο intraclass συντελεστής συσχέτισης ρ ορίζεται ως η συνήθης συσχέτιση μεταξύ παιδιών της ίδιας οικογένειας. Οι παρατηρήσεις των παιδιών της κάθε οικογένειας, που αφορούν κάποιο χαρακτηριστικό, θα παριστάνονται από ένα τυχαίο διάνυσμα. Υποθέτουμε ότι οι παρατηρήσεις έχουν την ίδια μέση τιμή, την ίδια διακύμανση και τον ίδιο συντελεστή συσχέτισης ρ. Ακόμη, υποθέτουμε ότι οι παρατηρήσεις κάθε οικογένειας ακολουθούν μια πολυδιάστατη κανονική κατανομή. Η ανάπτυξη συμπερασματολογίας βασίζεται στη μέθοδο μέγιστης πιθανοφάνειας. Στο παραπάνω πλαίσιο, διακρίνονται δύο περιπτώσεις. Στην πρώτη το μέγεθος των οικογενειών παραμένει ίδιο από οικογένεια σε οικογένεια, ενώ στη δεύτερη ποικίλλει. Αρχικά ακολουθώντας τη βιβλιογραφία, υπολογίζονται οι εκτιμητές μέγιστης πιθανοφάνειας των παραμέτρων του μοντέλου (βλ. Srivastava (2002), Wilks et al. (1946), Mentz (2001), Donner and Koval (1980a,b)). Στη συνέχεια κατασκευάζονται ασυμπτωτικά διαστήματα εμπιστοσύνης και τέλος στατιστικά τεστ, που βασίζονται στη μέθοδο μέγιστης πιθανοφάνειας, για τον έλεγχο υποθέσεων γύρω από την παράμετρο της intraclass συσχέτισης. Το κεφάλαιο ολοκληρώνεται με αναφορά στους εκτιμητές που έχει προτείνει ο Srivastava (1984) και μπορούν να χρησιμοποιηθούν ως εναλλακτικοί των εκτιμητών μέγιστης πιθανοφάνειας. Οι παραπάνω εκτιμητές μπορούν να χρησιμοποιηθούν για την κατασκευή διαστημάτων εμπιστοσύνης και στατιστικών τεστ.

Το 3ο κεφάλαιο επικεντρώνεται στην ύπαρξη δύο ή και περισσότερων δειγμάτων. Στο πλαίσιο αυτό κατασκευάζονται στατιστικά τεστ για τον έλεγχο της υπόθεσης κοινού συντελεστή συσχέτισης. Το κεφάλαιο βασίστηκε σε εργασίες των Donner and Bull (1983), Khatri (1989), Konishi (1989), Helu (2007).

Στο 4ο κεφάλαιο υποθέτουμε ότι οι παρατηρήσεις που αφορούν τα παιδιά κάθε οικογένειας περιγράφονται από ένα μοντέλο τυχαίων επιδράσεων (one-way random effect model). Υπό τις υποθέσεις του παραπάνω μοντέλου και της κανονικής κατανομής αναπτύσσεται συμπερασματολογία γύρω από την παράμετρο της intraclass συσχέτισης. Η μελέτη πάλι γίνεται ξεχωριστά για δύο περιπτώσεις ανάλογα με το αν το μέγεθος μεταξύ των οικογενειών είναι ίδιο ή μεταβάλλεται από οικογένεια σε οικογένεια. Το κεφάλαιο βασίστηκε στους Searle (1997), Graybill (1961), Donner (1986), Donner and Koval (1983), McGraw and Wong (1996), Mian (1989), Wald (1940), Thomas and Hultquist (1978).

Κεφάλαιο 2

ΣΤΑΤΙΣΤΙΚΗ ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑ ΓΙΑ ΤΟ **INTRACLASS** ΜΟΝΤΕΛΟ

2.1 Εισαγωγή

Έστω X κωδικοποιεί ένα χαρακτηριστικό γνώρισμα ή ιδιότητα των μελών μιας οικογένειας και στη διατριβή αυτή θα επικεντρώσουμε το ενδιαφέρον μας στα παιδιά της οικογένειας. Η X είναι μια τυχαία μεταβλητή και συμβολίζουμε με $X_i = (X_{i1}, X_{i2}, \ldots, X_{ik_i})'$, $i = 1, \ldots, n$, παρατηρήσεις του χαρακτηριστικού X στα k_i παιδιά της i-οστής οικογένειας, για $i = 1, \ldots, n$. Το X_i είναι ένα τυχαίο διάνυσμα και X_{ij} είναι η παρατήρηση του χαρακτηριστικού X του j-οστού παιδιού της i-οστής οικογένειας, για $j = 1, \ldots, k_i, i = 1, \ldots, n$. Υποθέτουμε ότι το τυχαίο διάνυσμα X_i $(i = 1, 2, \ldots, n)$ ακολουθεί μια πολυδιάστατη κανονική κατανομή με παραμέτρους μ_i και Σ_i. Ακόμη, υποθέτουμε ότι οι συνιστώσες του διανύσματος X_i έχουν την ίδια μέση τιμή, έστω α $(\alpha \in \mathbb{R})$, έχουν την ίδια διακύμανση σ^2 και ότι οι συνδιακυμάνσεις των X_{ij} και X_{il}, για $j \neq l$, είναι ίσες, έστω σ_{12}. Έστω ρ_{jl} $(j \neq l)$ ο συντελεστής συσχέτισης μεταξύ των συνιστωσών j και l του τυχαίου διανύσματος X_i. Τότε λόγω της σχέσης:

$$\rho_{jl} = \frac{Cov(X_{ij}, X_{il})}{[Var(X_{ij})]^{\frac{1}{2}}[Var(X_{il})]^{\frac{1}{2}}} = \frac{\sigma_{12}}{\sigma^2},$$

προκύπτει το συμπέρασμα ότι για $j \neq l$ οι συνιστώσες έχουν κοινό συντελεστή συσχέτισης, έστω ρ. Αυτή είναι η συσχέτιση μεταξύ μελών της ίδιας οικογένειας και ονομάζεται intraclass συσχέτιση. Οι παραπάνω υποθέσεις υποχρεώνουν το μέσο διάνυσμα μ_i και τον πίνακα διακυμάνσεων-συνδιακυμάνσεων Σ_i του τυχαίου διανύσματος X_i να αποκτούν μια συγκεκριμένη μορφή. Έτσι, το μ_i είναι ένα διάνυσμα διάστασης k_i της μορφής $\mu_i = (\alpha, \alpha, \ldots, \alpha)'$ και Σ_i είναι ένας $k_i \times k_i$ πίνακας ο οποίος γράφεται στη μορφή $\Sigma_i = \sigma^2[(1-\rho)I_{k_i} + \rho J_{k_i}]$, όπου I_{k_i} είναι ο ταυτοτικός πίνακας τάξης k_i και J_{k_i} είναι ένας $k_i \times k_i$ πίνακας με όλα τα στοιχεία του ίσα με τη μονάδα, για $i = 1, \ldots, n$. Να τονιστεί ότι προκειμένου ο πίνακας Σ_i να είναι θετικά ορισμένος ($\Sigma_i > 0$), πρέπει να ισχύει ότι: $\frac{-1}{k_i-1} < \rho < 1$ για $i = 1, \ldots, n$. Με βάση τα παραπάνω θεωρούμε ότι:

$$X_i \sim N_{k_i}(\mu_i, \Sigma_i), \tag{2.1}$$

όπου $\mu_i = (\alpha, \alpha, \ldots, \alpha)'$ και $\Sigma_i = \sigma^2[(1-\rho)I_{k_i} + \rho J_{k_i}]$, για $i = 1, \ldots, n$.

Το μοντέλο (2.1) είναι γνωστό στη βιβλιογραφία ως intraclass μοντέλο (ενδοκατηγορικό μοντέλο) και χρησιμοποιείται για την μοντελοποίηση familial δεδομένων (οικογενειακών δεδομένων). Σκοπός του κεφαλαίου αυτού είναι η ανάπτυξη συμπερασματολογίας (εκτίμηση, διαστήματα εμπιστοσύνης, στατιστικά τεστ) για τον intraclass συντελεστή συσχέτισης ρ, ο οποίος εκφράζει το βαθμό της σχέσης μεταξύ των μελών της ίδιας οικογένειας σε σχέση με το υπό μελέτη χαρακτηριστικό. Το παραπάνω πρόβλημα θα εξεταστεί για δύο περιπτώσεις. Αρχικά, για την περίπτωση, όπου το πλήθος των μελών της οικογένειας είναι το ίδιο για κάθε οικογένεια, balanced (ισορροπημένη) περίπτωση, $k_i = k, i = 1, \ldots, n$ και στη συνέχεια για την περίπτωση όπου το μέγεθος των οικογενειών ποικίλλει από οικογένεια σε οικογένεια, unbalanced περίπτωση. Η ανασκόπηση, στα εδάφια που ακολουθούν στηρίζεται στις εργασίες και τα βιβλία των Srivastava (1984, 2002), Srivastava and Katapa (1986), Donner and Koval (1980a,b), Wilks (1946), Mentz (2001) και στις αναφορές που υπάρχουν εκεί.

2.2 Balanced περίπτωση

Στην περίπτωση αυτή υποθέτουμε ότι το πλήθος των παιδιών σε κάθε οικογένεια είναι ίσο με k. Έστω $X_1, X_2, \ldots, X_n$ ένα τυχαίο δείγμα μεγέθους n από την $N_k(\mu, \Sigma)$, όπου $\mu = (\alpha, \alpha, \ldots, \alpha)'$ και $\Sigma = \sigma^2[(1-\rho)I_k + \rho J_k]$. Ο πίνακας των παρατηρήσεων θα έχει την παρακάτω μορφή:

$$X = \begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1k} \\ X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix}_{n \times k}.$$

Έστω $\bar{X}_i = \frac{1}{n}\sum\limits_{j=1}^{n} X_{ji}$, $i = 1, \ldots, k$, $\quad \bar{X} = \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_k \end{pmatrix}, \quad \bar{\bar{X}} = \frac{1}{nk}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{k} X_{ij}$.

Επιπλέον, έστω $V = \sum\limits_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})' = \|v_{ij}\|$, όπου

$$v_{jj} = \sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2 \quad \text{και} \quad v_{jl} = \sum_{i=1}^{n}(X_{ij} - \bar{X}_j)(X_{il} - \bar{X}_l),$$

για $j, l = 1, 2, \ldots, k$.

2.2.1 Εκτίμηση

Στο εδάφιο αυτό, θα υπολογιστούν οι εκτιμητές μέγιστης πιθανοφάνειας (ΕΜΠ) των παραμέτρων α, σ^2 και ρ. Έστω $X_1, X_2, \ldots, X_n$ ένα τυχαίο δείγμα από την κατανομή $N_k(\mu, \Sigma)$, όπου $\mu = (\alpha, \alpha, \ldots, \alpha)'$ και $\Sigma = \sigma^2[(1-\rho)I_k + \rho J_k]$. Να σημειωθεί ότι προκειμένου ο πίνακας Σ να είναι θετικά ορισμένος ($\Sigma > 0$) θα πρέπει $-\frac{1}{k-1} < \rho < 1$. Έστω Ω ο παραμετρικός χώρος του προβλήματος. Σύμφωνα με τα παραπάνω ο Ω έχει την παρακάτω μορφή,

$$\Omega = \{(\mu, \Sigma) : \mu = (\alpha, \alpha, \ldots, \alpha)', \Sigma = \sigma^2[(1-\rho)I_k + \rho J_k], \alpha \in \mathbb{R}, \sigma^2 > 0,$$
$$-\frac{1}{k-1} < \rho < 1\}. \tag{2.2}$$

Η συνάρτηση πιθανοφάνειας του τυχαίου δείγματος $x_1, x_2, \ldots, x_n$ είναι,

$$L(\mu, \Sigma | x_1, x_2, \ldots, x_n) = (2\pi)^{-nk/2} |\Sigma|^{-n/2}$$
$$\times \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)'\Sigma^{-1}(x_i - \mu)\right\}.$$

Σύμφωνα με το Mardia et al. (1979, σ. 97) η παραπάνω εξίσωση απλοποιείται με τη χρήση της παρακάτω σχέσης,

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

Γράφοντας

$$\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})' = V,$$

η συνάρτηση πιθανοφάνειας παίρνει την παρακάτω μορφή,

$$L(\alpha, \sigma^2, \rho | x_1, x_2, \ldots, x_n) = (2\pi)^{-nk/2} |\Sigma|^{-n/2}$$
$$\times \exp\left\{-\frac{1}{2}tr(\Sigma^{-1}V) - \frac{n}{2}(\bar{x} - \mu)'\Sigma^{-1}(\bar{x} - \mu)\right\}.$$

Έστω ℓ η log-likelihood συνάρτηση. Τότε,

$$\begin{aligned}\ell(\alpha,\sigma^2,\rho|x_1,x_2,\ldots,x_n) = &-\frac{nk}{2}\log(2\pi)-\frac{n}{2}\log|\Sigma|-\frac{1}{2}tr(\Sigma^{-1}V)\\ &-\frac{n}{2}(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu).\end{aligned}$$

Για τον πίνακα Σ ισχύουν οι παρακάτω σχέσεις (βλ. Mardia et al., 1979, σ. 461-462),

$$\Sigma^{-1} = \frac{1}{\sigma^2(1-\rho)}\left[I_k-\rho[1+(k-1)\rho]^{-1}J_k\right], \tag{2.3}$$

$$|\Sigma| = \sigma^{2k}(1-\rho)^{k-1}[1+\rho(k-1)]. \tag{2.4}$$

Επομένως,

$$\begin{aligned}tr(\Sigma^{-1}V) &= tr\left(\frac{1}{\sigma^2(1-\rho)}\left[I_k-\rho[1+(k-1)\rho]^{-1}J_k\right]V\right)\\ &= \frac{1}{\sigma^2(1-\rho)}tr(V)-\frac{\rho}{\sigma^2(1-\rho)[1+(k-1)\rho]}tr(J_kV)\\ &= \frac{1}{\sigma^2(1-\rho)}tr(V)-\frac{\rho}{\sigma^2(1-\rho)[1+(k-1)\rho]}1_k'V1_k,\end{aligned}$$

αφού $tr(J_kV) = tr(1_k1_k'V) = tr(1_k'V1_k) = 1_k'V1_k$, όπου $1_k'$ είναι ένα διάνυσμα διάστασης k με όλες τις συνιστώσες του ίσες με τη μονάδα. Ακόμη,

$$\begin{aligned}(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu) &= (\bar{x}-\mu)'\frac{1}{\sigma^2(1-\rho)}\left[I_k-\rho[1+(k-1)\rho]^{-1}J_k\right](\bar{x}-\mu)\\ &= \frac{1}{\sigma^2(1-\rho)}(\bar{x}-\mu)'(\bar{x}-\mu)-\frac{\rho}{\sigma^2(1-\rho)[1+(k-1)\rho]}(\bar{x}-\mu)'J_k(\bar{x}-\mu)\\ &= \frac{1}{\sigma^2(1-\rho)}\sum_{i=1}^{k}(\bar{x}_i-\alpha)^2-\frac{\rho}{\sigma^2(1-\rho)[1+(k-1)\rho]}\left(\sum_{i=1}^{k}(\bar{x}_i-\alpha)\right)^2.\end{aligned}$$

Αντικαθιστώντας στην εξίσωση της log-likelihood συνάρτησης, λαμβάνοντας υπόψιν και τη σχέση (2.4) καταλήγουμε στην παρακάτω εξίσωση,

$$\begin{aligned}\ell(\alpha,\sigma^2,\rho) = &-\frac{nk}{2}\log(2\pi) - \frac{nk}{2}\log\sigma^2 - \frac{n}{2}\log\Big[(1-\rho)^{k-1}[1+(k-1)\rho]\Big] \\ &- \frac{1}{2\sigma^2(1-\rho)}trV + \frac{\rho}{2\sigma^2(1-\rho)[1+(k-1)\rho]}1_k'V1_k \\ &- \frac{n}{2\sigma^2(1-\rho)}\sum_{i=1}^{k}(\bar{x}_i-\alpha)^2 + \frac{n\rho}{2\sigma^2(1-\rho)[1+(k-1)\rho]}\left(\sum_{i=1}^{k}(\bar{x}_i-\alpha)\right)^2.\end{aligned}$$

Παρακάτω, περιγράφεται η διαδικασία που ακολουθήθηκε για την απόκτηση των εκτιμητών. Αρχικά, παραγωγίζεται η log-likelihood συνάρτηση ως προς την παράμετρο α.

$$\frac{\partial\ell(\alpha,\sigma^2,\rho)}{\partial\alpha} = \frac{n}{\sigma^2[1+(k-1)\rho]}\sum_{i=1}^{k}(\bar{x}_i-\alpha).$$

Εξισώνοντας με το 0 και λύνοντας ως προς το α καταλήγουμε στον εκτιμητή μέγιστης πιθανοφάνειας της παραμέτρου α, έστω $\hat{\alpha}$, που δίνεται από την παρακάτω σχέση,

$$\hat{\alpha} = \frac{\sum_{i=1}^{k}\bar{x}_i}{k} = \bar{\bar{x}}. \tag{2.5}$$

Αντικαθιστούμε τον παραπάνω εκτιμητή στην log-likelihood συνάρτηση, παραγωγίζουμε ως προς σ^2 και εξισώνουμε με το 0. Έτσι, προκύπτει η παρακάτω σχέση,

$$nk\sigma^2(1-\rho)[1+(k-1)\rho] = [1+(k-1)\rho]trV - \rho 1_k'V1_k + n[1+(k-1)\rho]\sum_{i=1}^{k}(\bar{x}_i-\bar{\bar{x}})^2.$$

Στη συνέχεια παραγωγίζουμε ως προς ρ και εξισώνουμε με το μηδέν, για να καταλήξουμε στη σχέση,

$$\begin{aligned}nk\rho(k-1)\sigma^2(1-\rho)[1+(k-1)\rho] = &\, trV[1+(k-1)\rho]^2 - 1_k'V1_k[1+(k-1)\rho^2] \\ &+ n[1+(k-1)\rho]^2\sum_{i=1}^{k}(\bar{x}_i-\bar{\bar{x}})^2.\end{aligned}$$

Από τις δύο προηγούμενες σχέσεις προκύπτει μια εξίσωση με μόνο άγνωστο την παράμετρο ρ από τη λύση της οποίας θα προκύψει ο εκτιμητής μέγιστης πιθανοφάνειας του ρ, έστω $\hat{\rho}$. Έτσι, καταλήγουμε στο παρακάτω συμπέρασμα,

$$\hat{\rho} = \frac{1_k' V 1_k - trV - n\sum_{i=1}^{k}(\bar{x}_i - \bar{\bar{x}})^2}{(k-1)\left\{trV + n\sum_{i=1}^{k}(\bar{x}_i - \bar{\bar{x}})^2\right\}}. \tag{2.6}$$

Επιστρέφοντας στη σχέση που προέκυψε ύστερα από την παραγώγιση ως προς σ^2, αντικαθιστώντας το $\hat{\rho}$ στη θέση του ρ και λύνοντας ως προς σ^2 προκύπτει ο εκτιμητής μέγιστης πιθανοφάνειας του σ^2, έστω $\hat{\sigma}^2$, που δίνεται από τη σχέση,

$$\hat{\sigma}^2 = \frac{trV + n\sum_{i=1}^{k}(\bar{x}_i - \bar{\bar{x}})^2}{nk}. \tag{2.7}$$

Η προηγούμενη ανάλυση οδηγεί στο ακόλουθο θεώρημα στο οποίο διατυπώνονται οι ΕΜΠ των παραμέτρων α, σ^2 και ρ του intraclass μοντέλου (2.1) (βλ. Srivastava, 2002, σ. 163).

Θεώρημα 2.2.1. *Εάν $X_1, X_2, \ldots, X_n$ είναι ένα τυχαίο δείγμα από $N_k(\mu, \Sigma)$, για την οποία ισχύει $\mu = (\alpha, \alpha, \ldots, \alpha)'$ και $\Sigma = \sigma^2[(1-\rho)I_k + \rho J_k]$, με παραμετρικό χώρο Ω της σχέσης (2.2), οι εκτιμητές μέγιστης πιθανοφάνειας $\hat{\alpha}, \hat{\sigma}^2, \hat{\rho}$ των παραμέτρων α, σ^2, ρ, αντίστοιχα είναι οι,*

$$\hat{\alpha} = \frac{\sum_{i=1}^{k}\bar{x}_i}{k} = \bar{\bar{x}},$$

$$\hat{\sigma}^2 = \frac{trV + n\sum_{i=1}^{k}(\bar{x}_i - \bar{\bar{x}})^2}{nk},$$

$$\hat{\rho} = \frac{1_k' V 1_k - trV - n\sum_{i=1}^{k}(\bar{x}_i - \bar{\bar{x}})^2}{(k-1)\left\{trV + n\sum_{i=1}^{k}(\bar{x}_i - \bar{\bar{x}})^2\right\}}.$$

Να σημειωθεί στο σημείο αυτό ότι ο παραπάνω εκτιμητής της intraclass συσχέτισης ρ που προσδιορίστηκε, διαφοροποιείται λίγο από τον αντίστοιχο εκτιμητή που δίνεται στη σελίδα 163 του Srivastava (2002), είναι όμως αποτέλεσμα ενδελεχούς διερεύνησης.

Παρατηρήσεις

Στην balanced περίπτωση ο εκτιμητής μέγιστης πιθανοφάνειας της intraclass συσχέτισης ρ, $\hat{\rho}$, που δίνεται από τη σχέση (2.6), ταυτίζεται με τον pairwise εκτιμητή (βλ. Donner and Koval, 1980a). Πιο συγκεκριμένα, εάν X_{ij} παριστάνει τη j-οστή παρατήρηση της i-οστής οικογένειας ($i = 1, 2, \ldots, n, j = 1, 2, \ldots, k$), τότε σύμφωνα με τον Harris (1913) ο εκτιμητής της intraclass συσχέτισης, που συμβολίζεται παρακάτω με r, μπορεί να υπολογιστεί από την παρακάτω εξίσωση,

$$k \sum_{i=1}^{n} (\bar{X}_i - \bar{\bar{X}})^2 = nS^2 \left[1 + (k-1)r\right], \tag{2.8}$$

όπου

$$\bar{X}_i = \frac{1}{k} \sum_{j=1}^{k} X_{ij}, \quad i = 1, 2, \ldots, n,$$

$$\bar{\bar{X}} = \frac{1}{nk} \sum_{i=1}^{n} \sum_{j=1}^{k} X_{ij},$$

$$S^2 = \frac{1}{nk} \sum_{i=1}^{n} \sum_{j=1}^{k} (X_{ij} - \bar{\bar{X}})^2.$$

Παρατηρούμε ότι παραπάνω το $\bar{X}_i$ για $i = 1, \ldots, n$, συμβολίζει το μέσο όρο των παρατηρήσεων της i-οστής οικογένειας. Ο συμβολισμός αυτός χρησιμοποιήθηκε στη διαδικασία εύρεσης των ΕΜΠ των παραμέτρων, όμως συμβόλιζε το μέσο όρο που αντιστοιχεί στις παρατηρήσεις του i-οστού παιδιού. Προκειμένου να αποφύγουμε τη σύγχυση, σε όσα ακολουθούν θα συμβολίζουμε το μέσο όρο που αντιστοιχεί σε κάθε οικογένεια με $\bar{X}_i, i = 1, \ldots, n$ και το μέσο όρο που αντιστοιχεί σε κάθε παιδί με $\bar{X}_j, j = 1, \ldots, k$. Μετά από πράξεις καταλήγουμε στην εξής σχέση,

$$tr(V) + n \sum_{j=1}^{k} (\bar{X}_j - \bar{\bar{X}})^2 = \sum_{i=1}^{n} \sum_{j=1}^{k} (X_{ij} - \bar{\bar{X}})^2 = nkS^2.$$

Ακόμη ισχύει ότι (βλ. Srivastava, 2002, σ. 156),

$$1_k' V 1_k = k^2 \sum_{i=1}^{n} (\bar{X}_i - \bar{\bar{X}})^2.$$

Η σχέση (2.8) γράφεται ισοδύναμα,

$$r = \frac{k \sum_{i=1}^{n} (\bar{X}_i - \bar{\bar{X}})^2 - nS^2}{(k-1)nS^2}.$$

Ο εκτιμητής μέγιστης πιθανοφάνειας της σχέσης (2.6) μετά από αντικατάσταση των σχέσεων που βρέθηκαν παραπάνω γράφεται ισοδύναμα,

$$\hat{\rho} = \frac{k^2 \sum\limits_{i=1}^{n} (\bar{X}_i - \bar{\bar{X}})^2 - nkS^2}{(k-1)nkS^2}$$

ή

$$\hat{\rho} = \frac{k \sum_{i=1}^{n} (\bar{X}_i - \bar{\bar{X}})^2 - nS^2}{(k-1)nS^2},$$

που σημαίνει ότι οι δύο εκτιμητές στην balanced περίπτωση ταυτίζονται. Ο Fisher (1921) υπολόγισε την ακριβή κατανομή του εκτιμητή r, στην περίπτωση όπου έχουμε τυχαίο δείγμα n ζευγών παρατηρήσεων, έχοντας υποθέσει ότι ο πληθυσμός από τον οποίο προέρχεται το δείγμα είναι κανονικός με τις δύο μεταβλητές που περιγράφουν τις παρατατηρήσεις να έχουν την ίδια μέση τιμή, την ίδια τυπική απόκλιση και τον ίδιο συντελεστή συσχέτισης. Στη συνέχεια έδωσε και τη μορφή της κατανομής για την περίπτωση όπου έχουμε n k-άδες παρατηρήσεων. Έτσι, η συνάρτηση πυκνότητας πιθανότητας της δειγματικής κατανομής του r δίνεται από τη σχέση (βλ. Donner, 1986, σ. 74),

$$f(r) = \frac{C_\rho (1-r)^{\frac{1}{2}(n(k-1)-2)} [1+(k-1)r]^{\frac{1}{2}(n-3)}}{[1-\rho+\rho(1-r)(k-1)]^{\frac{1}{2}(nk-1)}}, \quad -\frac{1}{k-1} < r < 1, \qquad (2.9)$$

όπου

$$C_\rho = \frac{\Gamma^{\frac{1}{2}(nk-1)} (k-1)^{\frac{1}{2}n(k-1)} (1-\rho)^{\frac{1}{2}(n-1)} [1+(k-1)\rho]^{\frac{1}{2}n(k-1)}}{\Gamma\left(\frac{1}{2}(n-1)\right) \Gamma\left(\frac{1}{2}n(k-1)\right) k^{\frac{1}{2}(nk-3)}}.$$

Όπως δείξαμε παραπάνω, στην balanced περίπτωση οι εκτιμητές $\hat{\rho}$ και r ταυτίζονται, που σημαίνει ότι από τη σχέση (2.9) περιγράφεται και η δειγματική κατανομή του εκτιμητή $\hat{\rho}$.

2.2.2 Ασυμπτωτική κατανομή εκτιμητών

Στο εδάφιο αυτό θα προσδιορισθεί η ασυμπτωτική κατανομή των εκτιμητών μέγιστης πιθανοφάνειας που υπολογίστηκαν παραπάνω. Οι ΕΜΠ είναι υπό ορισμένες συνθήκες ασυμπτωτικά κανονικοί και αποδοτικοί εκτιμητές. Πιο συγκεκριμένα, έστω $X_1, X_2, \ldots, X_n$ είναι ένα τυχαίο δείγμα από την κατανομή με πυκνότητα $f(x|\theta)$, $\theta \in \Omega$, όπου $\theta = (\theta_1, \theta_2, \ldots, \theta_s)$ και Ω συμβολίζει τον παραμετρικό χώρο. Θεωρούμε ότι οι παρακάτω συνθήκες ικανοποιούνται (βλ. Κουρούκλης, 2015, σ. 222-223),

1) Το κοινό σύνολο τιμών των X_i, $S_1 = \{x : f(x|\theta) > 0\}$, δεν εξαρτάται από το θ.

2) Για κάθε $\theta_1 \neq \theta_2$, σημεία του Ω ισχύει $f(x|\theta_1) \neq f(x|\theta_2)$, $x \in S_2 \subset S_1$ με $P(S_2) > 0$, σε κάθε δηλαδή σημείο του Ω αντιστοιχεί διαφορετική κατανομή πιθανότητας.

Επιπλέον υποθέτουμε ότι (βλ. Lehmann and Casella, 2006, σελ. 462-463),

i) Υπάρχει ένα ανοιχτό υποσύνολο ω του Ω, που περιέχει την πραγματική τιμή της παραμέτρου θ, έστω θ^0, τέτοιο ώστε για (σχεδόν) όλα τα x, για την $f(x|\theta)$ ορίζονται όλες οι παράγωγοι $\left(\frac{\partial^3}{\partial\theta_j \partial\theta_k \partial\theta_l} f(x|\theta)\right), \forall\ \theta \in \omega$.

ii) Οι πρώτες και δεύτερες παράγωγοι του λογαρίθμου της f ικανοποιούν τις εξισώσεις,

$$E_\theta\left[\frac{\partial}{\partial\theta_j}\log f(X|\theta)\right]=0,\ \text{για}\ j=1,2,\ldots,s$$

και

$$\begin{aligned} I_{jk}(\theta) &= E_\theta\left[\frac{\partial}{\partial\theta_j}\log f(X|\theta)\cdot\frac{\partial}{\partial\theta_k}\log f(X|\theta)\right] \\ &= -E_\theta\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log f(X|\theta)\right], \end{aligned}$$

όπου $I_{jk}(\theta)$ είναι το $(j,k), j,k=1,\ldots,s$, στοιχείο του πίνακα πληροφορίας του Fisher που αντιστοιχεί σε μία παρατήρηση X.

iii) Υποθέτουμε ότι $I_{jk}(\theta)$ είναι πεπερασμένα και ότι ο πίνακας $I(\theta)=[I_{jk}(\theta)]$ είναι θετικά ορισμένος $\forall\ \theta\in\omega$, και ως εκ τούτου τα s στατιστικά,

$$\frac{\partial}{\partial\theta_1}\log f(x|\theta),\ldots,\frac{\partial}{\partial\theta_s}\log f(x|\theta),$$

είναι ανεξάρτητα με πιθανότητα 1.

iv) Τέλος, υποθέτουμε ότι υπάρχουν συναρτήσεις M_{jkl} τέτοιες ώστε,

$$\left|\frac{\partial^3}{\partial\theta_j\partial\theta_k\partial\theta_l}\log f(x|\theta)\right|\leq M_{jkl}(x),\forall\ \theta\in\omega,$$

όπου $m_{jkl}=E_{\theta^0}\left[M_{jkl}(X)\right]<\infty,\ \ \forall\ j,k,l.$

Γνωρίζουμε ότι ισχύει το παρακάτω θεώρημα (βλ. Lehmann and Casella, 2006, σ. 463).

Θεώρημα 2.2.2. *Έστω $X_1,X_2,\ldots,X_n$ τυχαίο δείγμα από μία κατανομή με πυκνότητα $f(x|\theta)$, που ικανοποιεί τις υποθέσεις $1), 2)$ και $i)-iv)$ παραπάνω. Τότε με πιθανότητα που τείνει στο 1, καθώς $n\to\infty$, υπάρχουν λύσεις $\hat\theta_n=\hat\theta_n(X_1,X_2,\ldots,X_n)$ των εξισώσεων πιθανοφάνειας έτσι ώστε,*

1. *$\hat\theta_{jn}$ είναι συνεπής εκτιμητής της παραμέτρου $\theta_j, j=1,\ldots,s$.*

2. *$\sqrt{n}(\hat{\theta}_n - \theta)$ ακολουθεί ασυμπτωτικά κανονική κατανομή με μηδενικό μέσο διάνυσμα και πίνακα διακυμάνσεων-συνδιακυμάνσεων $[I(\theta)]^{-1}$.*

3. *$\hat{\theta}_{jn}$ είναι ασυμπτωτικά αποδοτικός με την έννοια ότι,*

$$\sqrt{n}(\hat{\theta}_{jn} - \theta_j) \xrightarrow{d} N(0, [I(\theta)]_{jj}^{-1}),$$

όπου με $[I(\theta)]_{jj}^{-1}$ συμβολίζεται το $(j, j), j = 1, \ldots, s$, στοιχείο του αντίστροφου του πίνακα πληροφορίας $I(\theta)$ και με $\xrightarrow{d}$ συμβολίζεται η σύγκλιση κατά κατανομή.

Υποθέτουμε ότι στα πλαίσια του δικού μας προβλήματος οι παραπάνω υποθέσεις ικανοποιούνται και τα συμπεράσματα του Θεωρήματος 2.2.2 ισχύουν. Επομένως, καταλήγουμε στο συμπέρασμα ότι,

$$\sqrt{n}\left[(\hat{\alpha}, \hat{\sigma}^2, \hat{\rho})' - (\alpha, \sigma^2, \rho)'\right] \xrightarrow{d} N_3\left(0, [I(\alpha, \sigma^2, \rho)]^{-1}\right). \tag{2.10}$$

Η σχέση (2.10) γράφεται ισοδύναμα ως,

$$(\hat{\alpha}, \hat{\sigma}^2, \hat{\rho})' \xrightarrow{d} N_3\left((\alpha, \sigma^2, \rho)', [nI(\alpha, \sigma^2, \rho)]^{-1}\right), \tag{2.11}$$

αφού με την έννοια της σχέσης (2.10) είναι αποδεκτό. Έστω $I_n(\alpha, \sigma^2, \rho)$ ο πίνακας πληροφορίας του Fisher που αντιστοιχεί σε ολόκληρο το δείγμα μεγέθους n. Τότε γνωρίζουμε ότι $I_n(\alpha, \sigma^2, \rho) = nI(\alpha, \sigma^2, \rho)$. Επομένως η σχέση (2.11) γράφεται ισοδύναμα ως,

$$(\hat{\alpha}, \hat{\sigma}^2, \hat{\rho})' \xrightarrow{d} N_3\left((\alpha, \sigma^2, \rho)', [I_n(\alpha, \sigma^2, \rho)]^{-1}\right). \tag{2.12}$$

Οι Donner and Koval (1980b) υπολόγισαν την ασυμπτωτική διακύμανση του εκτιμητή μέγιστης πιθανοφάνειας $\hat{\rho}$, της παραμέτρου ρ, στην πιο γενική περίπτωση, δηλαδή αυτή στην οποία το πλήθος των παιδιών διαφέρει από οικογένεια σε οικογένεια (unbalanced), αντιστρέφοντας τον πίνακα πληροφορίας του Fisher, που αντιστοιχεί σε ολόκληρο το δείγμα. Υιοθετώντας το συμβολισμό που χρησιμοποίησαν έγιναν οι αντίστοιχοι υπολογισμοί και για την balanced περίπτωση όπου το μέγεθος των παιδιών δε διαφέρει από οικογένεια σε οικογένεια. Πιο συγκεκριμένα, αρχικά υπολογίστηκε ο πίνακας πληροφορίας του Fisher που αντιστοιχεί σε ολόκληρο το δείγμα μεγέθους n και στη συνέχεια βρέθηκε ο αντίστροφός του. Προκύπτει λοιπόν ότι,

$$I_n(\alpha, \sigma^2, \rho) = \begin{pmatrix} \frac{nk}{\sigma^2 W} & 0 & 0 \\ 0 & \frac{nk}{2(\sigma^2)^2} & \frac{nk(k-1)\rho}{2\sigma^2(1-\rho)W} \\ 0 & \frac{nk(k-1)\rho}{2\sigma^2(1-\rho)W} & \frac{nk(k-1)V}{2W^2(1-\rho)^2} \end{pmatrix}_{3\times 3} \tag{2.13}$$

και

$$[I_n(\alpha, \sigma^2, \rho)]^{-1} = \begin{pmatrix} \frac{W\sigma^2}{nk} & 0 & 0 \\ 0 & \frac{2V(\sigma^2)^2}{nk} & -\frac{2\rho\sigma^2(1-\rho)W}{nk} \\ 0 & -\frac{2\rho\sigma^2(1-\rho)W}{nk} & \frac{2(1-\rho)^2W^2}{nk(k-1)} \end{pmatrix}_{3\times 3}, \tag{2.14}$$

όπου $W = 1 + (k-1)\rho$ και $V = 1 + (k-1)\rho^2$.

Έχοντας υπολογίσει τον πίνακα της σχέσης (2.14), η ασυμπτωτική κατανομή των εκτιμητών μέγιστης πιθανοφάνειας $\hat{\alpha}, \hat{\sigma}^2, \hat{\rho}$ προσδιορίζεται από τη σχέση (2.12). Ωστόσο, το ενδιαφέρον μας θα επικεντρωθεί στον εκτιμητή μέγιστης πιθανοφάνειας $\hat{\rho}$ της παραμέτρου ρ. Από το Θεώρημα 2.2.2 για τον εκτιμητή $\hat{\rho}$ ισχύει ότι,

$$\sqrt{n}(\hat{\rho} - \rho) \xrightarrow{d} N\left(0, [I(\alpha, \sigma^2, \rho)]^{-1}_{33}\right),$$

ή

$$\hat{\rho} \xrightarrow{d} N\left(\rho, [I_n(\alpha, \sigma^2, \rho)]^{-1}_{33}\right),$$

ή

$$\hat{\rho} \xrightarrow{d} N\left(\rho, \frac{2(1-\rho)^2W^2}{nk(k-1)}\right). \tag{2.15}$$

Έστω με $Var(\hat{\rho})$ συμβολίζεται η ασυμπτωτική διακύμανση του εκτιμητή $\hat{\rho}$. Τότε,

$$Var(\hat{\rho}) = \frac{2(1-\rho)^2 W^2}{nk(k-1)} = \frac{2(1-\rho)^2[1+(k-1)\rho]^2}{nk(k-1)}. \tag{2.16}$$

Επομένως ένας plug-in εκτιμητής της $Var(\hat{\rho})$ θα δίνεται από τη σχέση,

$$\widehat{Var}(\hat{\rho}) = \frac{2(1-\hat{\rho})^2[1+(k-1)\hat{\rho}]^2}{nk(k-1)}. \tag{2.17}$$

2.2.3 Ασυμπτωτικά διαστήματα εμπιστοσύνης

Στο σημείο αυτό θα κατασκευαστούν ασυμπτωτικά διαστήματα εμπιστοσύνης για την παράμετρο ρ. Γνωρίζουμε ότι ισχύει,

$$\sqrt{n}(\hat{\rho}-\rho) \xrightarrow{d} N\left(0, \frac{2(1-\rho)^2 W^2}{k(k-1)}\right),$$

όπου $W^2 = [1+(k-1)\rho]^2$. Τότε,

$$\frac{(\hat{\rho}-\rho)}{\sqrt{Var(\hat{\rho})}} \xrightarrow{d} N(0,1), \tag{2.18}$$

όπου $Var(\hat{\rho})$ ορίζεται από τη σχέση (2.16). Ο $\hat{\rho}$ είναι συνεπής εκτιμητής της παραμέτρου ρ (βλ. Θεώρημα 2.2.2 (1)). Αυτό σημαίνει ότι,

$$\hat{\rho} \xrightarrow{P} \rho, \ \text{καθώς} \ n \to \infty,$$

όπου με $\xrightarrow{P}$ συμβολίζεται η σύγκλιση κατά πιθανότητα. Επομένως, από το Θεώρημα συνεχούς απεικόνισης (βλ. Van der Vaart, 2000, σ. 7) ένας συνεπής εκτιμητής της $\sqrt{Var(\hat{\rho})}$ είναι ο $\sqrt{\widehat{Var}(\hat{\rho})}$, όπου ο $\widehat{Var}(\hat{\rho})$ ορίζεται από τη σχέση (2.17). Εφαρμόζοντας το Λήμμα του Slutsky (βλ. Van der Vaart, 2000, σ. 11) ισχύει ότι,

$$\frac{(\hat{\rho}-\rho)}{\sqrt{\widehat{Var}(\hat{\rho})}} = \frac{\sqrt{Var(\hat{\rho})}}{\sqrt{\widehat{Var}(\hat{\rho})}} \frac{(\hat{\rho}-\rho)}{\sqrt{Var(\hat{\rho})}} \xrightarrow{d} N(0,1), \tag{2.19}$$

λόγω της σχέσης (2.18) και επειδή $\frac{\sqrt{Var(\hat{\rho})}}{\sqrt{\widehat{Var}(\hat{\rho})}} \xrightarrow{P} 1$. Επομένως, για την κατασκευή ενός ασυμπτωτικού $100\,(1-\alpha)\%$ διαστήματος εμπιστοσύνης έχουμε ότι,

$$P\left(-z_{a/2} < \frac{(\hat{\rho}-\rho)}{\sqrt{\widehat{Var}(\hat{\rho})}} < z_{a/2}\right) = 1-\alpha,$$

ή

$$P\left(\hat{\rho} - z_{a/2}\sqrt{\widehat{Var}(\hat{\rho})} < \rho < \hat{\rho} + z_{a/2}\sqrt{\widehat{Var}(\hat{\rho})}\right) = 1-\alpha,$$

όπου $100\,(1-\alpha)\%, 0 < \alpha < 1$, είναι ο επιθυμητός βαθμός εμπιστοσύνης και $z_{a/2}$ είναι τα εκατοστιαία σημεία της τυπικής κανονικής κατανομής τέτοια ώστε $P(Z \geqslant z_a) = a$. Δηλαδή, ένα ασυμπτωτικό $100\,(1-\alpha)\%$ διάστημα εμπιστοσύνης για την παράμετρο ρ είναι το,

$$\left(\hat{\rho} - z_{a/2}\sqrt{\widehat{Var}(\hat{\rho})}, \hat{\rho} + z_{a/2}\sqrt{\widehat{Var}(\hat{\rho})}\right). \tag{2.20}$$

Η παραπάνω μέθοδος που χρησιμοποιήθηκε είναι γνωστή ως plug-in μέθοδος. Η plug-in μέθοδος είναι η πιο συχνή για την κατασκευή διαστημάτων εμπιστοσύνης, όμως υπάρχουν περιπτώσεις όπου μπορεί να αποφευχθεί. Από τη σχέση (2.18) έχουμε ότι,

$$\frac{(\hat{\rho}-\rho)}{\sqrt{Var(\hat{\rho})}} \xrightarrow{d} N(0,1).$$

Η ποσότητα στο αριστερό μέλος της παραπάνω σχέσης περιέχει ως μόνη άγνωστη την παράμετρο ρ την οποία θέλουμε και να εκτιμήσουμε. Έτσι, εάν υποθέσουμε ότι ισχύει,

$$P\left(-z_{a/2} < \frac{(\hat{\rho}-\rho)}{\sqrt{Var(\hat{\rho})}} < z_{a/2}\right) = 1-\alpha,$$

τότε το σύνολο,

$$\left\{\rho : -z_{a/2} < \frac{(\hat{\rho}-\rho)}{\sqrt{Var(\hat{\rho})}} < z_{a/2}\right\}, \tag{2.21}$$

είναι ένα ασυμπτωτικό $100\,(1-\alpha)\%$ διάστημα εμπιστσύνης για την παράμετρο ρ.

2.2.4 Στατιστικά Τεστ

Στο εδάφιο αυτό θα κατασκευαστούν στατιστικά τεστ για τον έλεγχο υποθέσεων που διατυπώνονται για την παράμετρο ρ του intraclass μοντέλου (2.1). Πιο συγκεκριμένα θα παρουσιαστούν τα εξής στατιστικά τεστ:

i) Τεστ πηλίκου μέγιστων πιθανοφανειών.

ii) Wald τεστ.

iii) Score τεστ.

Πριν διατυπωθούν τα κύρια αποτελέσματα θα δοθεί μια μικρή ανασκόπηση των ανωτέρω μεθόδων ελέγχου υποθέσεων, χάριν πληρότητας της διατριβής.

2.2.4.1 Ανασκόπηση μεθόδων κατασκευής στατιστικών τεστ

Έστω $X_1, X_2, \ldots, X_n$ ένα τυχαίο δείγμα από πληθυσμό με κατανομή $f(x|\theta)$, $\theta = (\theta_1, \ldots, \theta_s) \in \Omega \subseteq \mathbb{R}^s$, όπου Ω είναι ο παραμετρικός χώρος. Τότε γνωρίζουμε ότι η συνάρτηση πιθανοφάνειας ορίζεται ως η από κοινού κατανομή του τυχαίου δείγματος $X_1, X_2, \ldots, X_n$,

$$L(\theta) = L(\theta|x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta).$$

Έστω ότι θέλουμε να εξετάσουμε την μηδενική υπόθεση,

$$H_0 : \theta \in \Omega_0,$$

όπου $\Omega_0 = \{(\theta_1, \ldots, \theta_s)' \in \mathbb{R}^s : \theta_i = \theta_{i0}\ (\theta_{i0}\ \text{καθορισμένο}),\ i = 1, \ldots, r, r \leq s\}$. Ισοδύναμα η μηδενική υπόθεση H_0 γράφεται ως,

$$H_0 : \theta_1 = \theta_{10}, \ldots, \theta_r = \theta_{r0},$$

όπου $r \leqslant s$ και $\theta_{10}, \ldots, \theta_{r0}$ γνωστά.

$i)$ **Το πηλίκο μέγιστων πιθανοφανειών** λ ορίζεται από το λόγο,

$$\lambda = \frac{\sup_{\Omega_0} L}{\sup_{\Omega} L} = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})}, \tag{2.22}$$

όπου με $\sup_{\Omega_0} L$ συμβολίζεται το supremum της L όταν η παράμετρος θ του μοντέλου περιλαμβάνεται στον παραμετρικό χώρο Ω_0, που ορίζεται από την μηδενική υπόθεση, ενώ με $\sup_{\Omega} L$ συμβολίζεται το supremum της L όταν η παράμετρος θ περιλαμβάνεται στον παραμετρικό χώρο Ω. Επίσης, $\hat{\theta}_0$ είναι ο εκτιμητής μέγιστης πιθανοφάνειας της παραμέτρου θ, για $\theta \in \Omega_0$ και $\hat{\theta}$ είναι ο εκτιμητής μέγιστης πιθανοφάνειας της παραμέτρου θ, για $\theta \in \Omega$. Επειδή $\Omega_0 \subseteq \Omega$ ο λόγος πιθανοφανειών της σχέσης (2.22) ικανοποιεί την ανισότητα $0 \leq \lambda \leq 1$. Μικρές τιμές του λόγου λ συνηγορούν υπέρ της απόρριψης της μηδενικής υπόθεσης H_0 και η περιοχή απόρριψης της μηδενικής υπόθεσης H_0 καθορίζεται πλέον από μικρές τιμές του λόγου λ. Δηλαδή η κρίσιμη περιοχή διαμορφώνεται έτσι ώστε $\lambda \leq \lambda_0$ με λ_0 κατάλληλη σταθερά. Το λ_0 προσδιορίζεται από την πιθανότητα σφάλματος Τύπου I, δηλαδή από τη σχέση,

$$P(\text{Απορρ. } H_0 | H_0 \ \text{αληθής}) = P(\lambda \leq \lambda_0 | H_0 \ \text{αληθής}) = \alpha, \tag{2.23}$$

με $\alpha, 0 < \alpha < 1$, το επίπεδο σημαντικότητας του τεστ. Η ακριβής κατανομή του στατιστικού λ σε πολλές περιπτώσεις είναι δύσχρηστη ή είναι αδύνατο να προσδιοριστεί. Στις περιπτώσεις αυτές εξυπηρετεί η ασυμπτωτική κατανομή του λ που δίνεται από το παρακάτω θεώρημα (βλ. Knight, 2000).

Θεώρημα 2.2.3. *Έστω $X_1, X_2, \ldots, X_n$ τυχαίο δείγμα από μία κατανομή $f(x|\theta)$ με $\theta = (\theta_1, \theta_2, \ldots, \theta_s)$. Έστω ότι η $f(x|\theta)$ ικανοποιεί τις υποθέσεις $1), 2)$ και $i) - iv)$ που αναφέρθηκαν στην ενότητα 2.2.2. Εάν για τον εκτιμητή μέγιστης πιθανοφάνειας $\hat{\theta}_n$ ισχύει $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, [I(\theta)]^{-1})$, τότε το πηλίκο μέγιστων* *πιθανοφανειών* λ *για τον έλεγχο της* $H_0 : \theta_1 = \theta_{10}, \ldots, \theta_r = \theta_{r0}$ *ικανοποιεί την παρακάτω σχέση,*

$$-2\log\lambda \xrightarrow[n\to\infty]{d} \chi^2_r, \ \textit{όταν } H_0 \textit{ είναι αληθής}. \tag{2.24}$$

Η κατανομή της σχέσης (2.24) και η πιθανότητα σφάλματος Τύπου I της σχέσης (2.23) οδηγούν στον προσδιορισμό του κρίσιμου σημείου λ_0, ως εξής,

$$P(\text{Απορρ. } H_0 | H_0 \text{ αληθής}) = P(\lambda \leq \lambda_0 | H_0 \text{ αληθής}) = \alpha,$$

ή

$$P(-2\log\lambda \geq -2\log\lambda_0 | H_0 \text{ αληθής}) = \alpha,$$

ή

$$P(\chi^2_r \geq -2\log\lambda_0) = \alpha,$$

από την οποία προκύπτει ότι $-2\log\lambda_0 = \chi^2_{r,\alpha}$.
Υπάρχουν και άλλα τεστ που βασίζονται στη μέθοδο μέγιστης πιθανοφάνειας. Όσα ακολουθούν βασίζονται στο βιβλίο του Knight, 2000.

$ii)$ **Wald τεστ**
Έστω $\theta = (\phi', \tau')$, όπου $\phi = (\theta_1, \theta_2, \ldots, \theta_r)'$ και $\tau = (\theta_{r+1}, \ldots, \theta_s)'$. Η παράμετρος τ δεν παίζει ρόλο στη διαμόρφωση της μηδενικής υπόθεσης, είναι δηλαδή μια nuisance παράμετρος (ενοχλητική παράμετρος). Επομένως, η H_0 γίνεται,

$$H_0 : \phi = \phi_0,$$

όπου $\phi_0 = (\theta_{10}, \ldots, \theta_{r0})'$. Το Wald τεστ συγκρίνει τον δίχως περιορισμούς (unrestricted) εκτιμητή μέγιστης πιθανοφάνειας του ϕ με την τιμή του υπό τη μηδενική υπόθεση. Εάν η απόσταση μεταξύ των δύο είναι μεγάλη, σημαίνει ότι η H_0 είναι εσφαλμένη και αυτό θα πρέπει να αντικατοπτρίζεται από το τεστ. Εάν $\hat{\phi}_n$ είναι ο εκτιμητής μέγιστης πιθανοφάνειας (unrestricted) της παραμέτρου ϕ τότε,

$$\sqrt{n}(\hat{\phi}_n - \phi_0) \xrightarrow{d} N_r\left(0, C(\phi_0, \tau)\right), \text{ εάν η } H_0 \text{ είναι αληθής,} \tag{2.25}$$

όπου ο πίνακας διακυμάνσεων-συνδιακυμάνσεων $C(\phi_0, \tau)$ μπορεί να υπολογιστεί από τον πίνακα πληροφορίας του Fisher,

$$I(\phi_0, \tau) = \begin{pmatrix} I_{11}(\phi_0, \tau) & I_{12}(\phi_0, \tau) \\ I_{21}(\phi_0, \tau) & I_{22}(\phi_0, \tau) \end{pmatrix},$$

από τη σχέση,

$$C(\phi_0, \tau) = [I_{11}(\phi_0, \tau) - I_{21}(\phi_0, \tau) I_{22}^{-1}(\phi_0, \tau) I_{12}(\phi_0, \tau)]^{-1}. \tag{2.26}$$

Το Wald στατιστικό ορίζεται ως,

$$W_n = n(\hat{\phi}_n - \phi_0)' \widehat{C}_n^{-1} (\hat{\phi}_n - \phi_0), \tag{2.27}$$

όπου $\widehat{C}_n^{-1}$ είναι κάποιος εκτιμητής του πίνακα $C^{-1}(\phi_0, \tau)$ που είναι συνεπής υπό την H_0. Μπορούμε να θεωρήσουμε $\widehat{C}_n^{-1} = [C(\hat{\phi}_n, \hat{\tau}_n)]^{-1}$ ή $\widehat{C}_n^{-1} = [C(\phi_0, \hat{\tau}_n)]^{-1}$. Για το στατιστικό της σχέσης (2.27) ισχύει ότι,

$$W_n \xrightarrow[n \to \infty]{d} \chi_r^2, \text{ εάν η } H_0 \text{ είναι αληθής.} \tag{2.28}$$

Απορρίπτουμε την H_0 για μεγάλες τιμές του στατιστικού W_n. Πιο συγκεκριμένα, απορρίπτουμε την H_0 εάν $W_n \geq W_0$, όπου το W_0 ορίζεται από την παρακάτω σχέση,

$$P(\text{Απορρ. } H_0 | H_0 \text{ αληθής}) = P(W_n \geq W_0 | H_0 \text{ αληθής}) = \alpha,$$

ή

$$P(\chi_r^2 \geq W_0) = \alpha,$$

από την οποία προκύπτει ότι $W_0 = \chi_{r,\alpha}^2$.

iii) **Score τεστ**

Το score τεστ βασίζεται στο γεγονός ότι εάν η μηδενική υπόθεση είναι εσφαλμένη, τότε η κλίση (gradient) της log-likelihood συνάρτησης δε θα είναι κοντά στο μηδενικό διάνυσμα. Πιο συγκεκριμένα, εάν $S_i(\phi, \tau), i = 1, 2, \ldots, n$, είναι η κλίση (gradient) της $\log f(X_i|\phi, \tau)$ ως προς ϕ, τότε,

$$S_i(\phi, \tau) = \left(\frac{\partial}{\partial \theta_1} \log f(X_i|\phi, \tau), \ldots, \frac{\partial}{\partial \theta_r} \log f(X_i|\phi, \tau) \right)'. \tag{2.29}$$

Υπό την H_0 (δεδομένου των υποθέσεων 1), 2) και *i*)-*iv*)) ισχύει ότι,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_i(\phi_0, \hat{\tau}_{n0}) \xrightarrow{d} N_r(0, I_{11}(\phi_0, \tau)), \tag{2.30}$$

όπου $\hat{\tau}_{n0}$ είναι ο εκτιμητής μέγιστης πιθανοφάνειας του τ υπό την H_0. Το score στατιστικό ορίζεται από τη σχέση,

$$S_n = \frac{1}{n} \left(\sum_{i=1}^{n} S_i(\phi_0, \hat{\tau}_{n0}) \right)' I_{11}^{-1}(\phi_0, \hat{\tau}_{n0}) \left(\sum_{i=1}^{n} S_i(\phi_0, \hat{\tau}_{n0}) \right). \tag{2.31}$$

Για το στατιστικό της σχέσης (2.31) ισχύει ότι,

$$S_n \xrightarrow[n \to \infty]{d} \chi_r^2, \ \text{ εάν η } \ H_0 \ \text{ είναι αληθής}. \tag{2.32}$$

Απορρίπτουμε την H_0 για μεγάλες τιμές του στατιστικού S_n. Πιο συγκεκριμένα, απορρίπτουμε την H_0 εάν $S_n \geq S_0$, όπου το S_0 ορίζεται από την παρακάτω σχέση,

$$P(\text{Απορρ. } H_0 | H_0 \ \text{ αληθής}) = P(S_n \geq S_0 | H_0 \ \text{ αληθής}) = \alpha,$$

ή

$$P(\chi_r^2 \geq S_0) = \alpha,$$

από την οποία προκύπτει ότι $S_0 = \chi_{r,\alpha}^2$.

2.2.4.2 Τεστ για τον έλεγχο $H_0 : \rho = \rho_0,\ \rho_0$ γνωστό

Μετά την ανασκόπηση των τεστ πηλίκου μέγιστων πιθανοφανειών, Wald και Rao Score τεστ, θα κατασκευαστούν, στη συνέχεια, τα παραπάνω στατιστικά τεστ για τον έλεγχο υποθέσεων που αφορούν την intraclass συσχέτιση ρ. Έστω $X_1, X_2, \ldots, X_n$ τυχαίο δείγμα από την $N_k(\mu, \Sigma)$ με $\mu = (\alpha, \alpha, \ldots, \alpha)'$ και $\Sigma = \sigma^2[(1-\rho)I_k + \rho J_k]$. Έστω ο έλεγχος της υπόθεσης,

$$H_0 : \rho = \rho_0 \quad \text{έναντι} \quad H_1 : \rho \neq \rho_0. \tag{2.33}$$

Η συνάρτηση πιθανοφάνειας για το παραπάνω τυχαίο δείγμα είναι,

$$\begin{aligned}
&L(\alpha, \sigma^2, \rho | x_1, x_2, \ldots, x_n) = (2\pi)^{-nk/2} \left[\sigma^{2k}(1-\rho)^{k-1}[1+(k-1)\rho]\right]^{-n/2} \\
&\times \exp\left\{-\frac{1}{2\sigma^2(1-\rho)} trV + \frac{\rho}{2\sigma^2(1-\rho)[1+(k-1)\rho]} 1_k' V 1_k\right\} \\
&\times \exp\left\{-\frac{n}{2\sigma^2(1-\rho)} \sum_{i=1}^{k} (\bar{x}_i - \alpha)^2 + \frac{n\rho}{2\sigma^2(1-\rho)[1+(k-1)\rho]} \left(\sum_{i=1}^{k} (\bar{x}_i - \alpha)\right)^2\right\}.
\end{aligned}$$

Το πηλίκο μέγιστων πιθανοφανειών ορίζεται ως (βλ. 2.22),

$$\lambda = \frac{\sup_{\Omega_0} L}{\sup_{\Omega} L} = \frac{L(\hat{\alpha}_0, \hat{\sigma}_0^2, \rho_0)}{L(\hat{\alpha}, \hat{\sigma}^2, \hat{\rho})},$$

όπου $\hat{\alpha}_0, \hat{\sigma}_0^2$ είναι οι εκτιμητές μέγιστης πιθανοφάνειας των παραμέτρων α, σ^2 αντίστοιχα, υπό την H_0 και $\hat{\alpha}, \hat{\sigma}^2, \hat{\rho}$ είναι οι εκτιμητές μέγιστης πιθανοφάνειας των παραμέτρων α, σ^2, ρ για $(\alpha, \sigma^2, \rho) \in \Omega$, όπου Ω δίνεται από τη σχέση (2.2). Λαμβάνοντας υπόψη το Θεώρημα 2.2.1, μετά από πράξεις έχουμε ότι,

$$L(\hat{\alpha}, \hat{\sigma}^2, \hat{\rho}) = (2\pi)^{-nk/2} (\hat{\sigma}^2)^{-nk/2} \left[(1-\hat{\rho})^{k-1}[1+(k-1)\hat{\rho}]\right]^{-n/2} e^{-nk/2}.$$

Για να βρεθούν οι $\hat{\alpha}_0, \hat{\sigma}_0^2$ πρέπει να μεγιστοποιηθεί η συνάρτηση πιθανοφάνειας ή ισοδύναμα η log-likelihood συνάρτηση ως προς α, σ^2 αφού αντικατασταθεί στη θέση του ρ το ρ_0. Έτσι έχουμε,

$$\hat{\alpha}_0 = \frac{\sum_{i=1}^{k} \bar{x}_i}{k} = \bar{\bar{x}}, \tag{2.34}$$

Επιπλέον για τον εκτιμητή $\hat{\sigma}_0^2$ ισχύει η σχέση (βλ. ενότητα 2.2.1 αυτής της μετ. διατριβής),

$$nk\hat{\sigma}_0^2(1-\rho_0)[1+(k-1)\rho_0] = [1+(k-1)\rho_0]trV - \rho_0 1_k' V 1_k + n[1+(k-1)\rho_0]\sum_{i=1}^{k}(\bar{x}_i - \bar{\bar{x}})^2.$$

Έτσι με αντικατάσταση των παραπάνω σχέσεων στη συνάρτηση πιθανοφάνειας ισχύει ότι,

$$L(\hat{\alpha}_0, \hat{\sigma}_0^2, \rho_0) = (2\pi)^{-nk/2}(\hat{\sigma}_0^2)^{-nk/2}\left[(1-\rho_0)^{k-1}[1+(k-1)\rho_0]\right]^{-n/2} e^{-nk/2}.$$

Επιπλέον οι εκτιμητές μέγιστης πιθανοφάνειας στον πλήρη παραμετρικό χώρο Ω έχουν προσδιορισθεί στο Θεώρημα 2.2.1. Έτσι, είμαστε πλέον σε θέση να υπολογίσουμε το πηλίκο μέγιστων πιθανοφανειών λ,

$$\begin{aligned}\lambda = \frac{\sup_{\Omega_0} L}{\sup_{\Omega} L} &= \frac{(2\pi)^{-nk/2}(\hat{\sigma}_0^2)^{-nk/2}\left[(1-\rho_0)^{k-1}[1+(k-1)\rho_0]\right]^{-n/2} e^{-nk/2}}{(2\pi)^{-nk/2}(\hat{\sigma}^2)^{-nk/2}\left[(1-\hat{\rho})^{k-1}[1+(k-1)\hat{\rho}]\right]^{-n/2} e^{-nk/2}} \\ &= \left[\left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^k \left(\frac{(1-\hat{\rho})^{k-1}[1+(k-1)\hat{\rho}]}{(1-\rho_0)^{k-1}[1+(k-1)\rho_0]}\right)\right]^{n/2}.\end{aligned}$$

Από το Θεώρημα 2.2.3. γνωρίζουμε ότι για την ασυμπτωτική κατανομή του λ ισχύει ότι,

$$-2\log\lambda \xrightarrow[n\to\infty]{d} \chi_1^2, \text{αν η } H_0 \text{ αληθεύει.}$$

Όπως είδαμε και παραπάνω, απορρίπτουμε την H_0 εάν $-2\log\lambda \geq \chi_{1,\alpha}^2$, όπου α είναι το επιθυμητό επίπεδο σημαντικότητας.

Στη συνέχεια θα προσδιοριστεί το Wald τεστ για τον έλεγχο της $H_0 : \rho = \rho_0$. Ισχύει ότι (βλ. σχέση 2.15),

$$\sqrt{n}(\hat{\rho} - \rho_0) \xrightarrow{d} N\left(0, \frac{2(1-\rho_0)^2[1+(k-1)\rho_0]^2}{k(k-1)}\right) \text{ υπό την } H_0.$$

Τότε το Wald στατιστικό δίνεται από τη σχέση,

$$W_n = n\left\{\frac{k(k-1)}{2(1-\rho_0)^2[1+(k-1)\rho_0]^2}\right\}(\hat{\rho}-\rho_0)^2.$$

Για την ασυμπτωτική κατανομή του W_n ισχύει ότι,

$$W_n \xrightarrow[n\to\infty]{d} \chi_1^2, \text{ εάν η } H_0 \text{ είναι αληθής.}$$

Απορρίπτουμε την H_0 εάν $W_n \geq \chi^2_{1,\alpha}$, όπου α είναι το επιθυμητό επίπεδο σημαντικότητας. Τέλος, θα κατασκευαστεί το Score τεστ. Έχουμε ότι,

$$\log f(X_i|\alpha,\sigma^2,\rho) = -\frac{k}{2}\log(2\pi) - \frac{k}{2}\log\sigma^2 - \frac{1}{2}\log\left[(1-\rho)^{k-1}[1+(k-1)\rho]\right]$$
$$-\frac{1}{2\sigma^2(1-\rho)}\left[\sum_{j=1}^{k}(X_{ij}-\alpha)^2 - \frac{\rho}{1+(k-1)\rho}\left(\sum_{j=1}^{k}(X_{ij}-\alpha)\right)^2\right],$$

για $i = 1, 2, \ldots, n$. Θέτουμε

$$\begin{aligned} S_i(\alpha,\sigma^2,\rho) &= \frac{\partial}{\partial\rho}\log f(X_i|\alpha,\sigma^2,\rho) \\ &= \frac{k\rho(k-1)}{2(1-\rho)[1+(k-1)\rho]} - \frac{1}{2\sigma^2(1-\rho)^2}\sum_{j=1}^{k}(X_{ij}-\alpha)^2 \\ &+ \frac{1+\rho^2(k-1)}{2\sigma^2(1-\rho)^2[1+(k-1)\rho]^2}\left(\sum_{j=1}^{k}(X_{ij}-\alpha)\right)^2, \end{aligned}$$

για $i = 1, \ldots, n$. Τότε υπό την H_0 ισχύει ότι,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_i(\hat{\alpha}_0,\hat{\sigma}_0^2,\rho_0) \xrightarrow{d} N\left(0, \frac{k(k-1)V_0}{2W_0^2(1-\rho_0)^2}\right),$$

όπου $W_0 = 1+(k-1)\rho_0$ και $V_0 = 1+(k-1)\rho_0^2$ (βλ. (2.13)). Το Score στατιστικό ορίζεται από τη σχέση,

$$S_n = \frac{1}{n}\left(\frac{k(k-1)V_0}{2W_0^2(1-\rho_0)^2}\right)^{-1}\left(\sum_{i=1}^{n} S_i(\hat{\alpha}_0,\hat{\sigma}_0^2,\rho_0)\right)^2.$$

Για την ασυμπτωτική κατανομή του S_n ισχύει ότι,

$$S_n \xrightarrow[n\to\infty]{d} \chi_1^2, \text{ εάν η } H_0 \text{ είναι αληθής.}$$

Απορρίπτουμε την H_0 εάν $S_n \geq \chi^2_{1,\alpha}$, όπου α είναι το επιθυμητό επίπεδο σημαντικότητας. Συγκεντρωτικά, η προηγούμενη ανάλυση συνοψίζεται στο παρακάτω θεώρημα.

Θεώρημα 2.2.4. *Έστω $X_1, X_2, \ldots, X_n$ τυχαίο δείγμα από την κατανομή $N_k(\mu, \Sigma)$, όπου $\mu = (\alpha, \alpha, \ldots, \alpha)'$, $\Sigma = \sigma^2[(1-\rho)I_k + \rho J_k]$ και Ω είναι ο παραμετρικός χώρος που ορίζεται από την (2.2). Υποθέτουμε ότι ισχύουν οι συνθήκες που εξασφαλίζουν την ασυμπτωτική κανονικότητα των εκτιμητών μέγιστης πιθανοφάνειας. Έστω ότι θέλουμε να ελέγξουμε την μηδενική υπόθεση,*

$$H_0 : \rho = \rho_0 \quad \text{έναντι της εναλλακτικής} \quad H_1 : \rho \neq \rho_0.$$

i) *Το στατιστικό του λόγου μέγιστων πιθανοφανειών λ ορίζεται από,*

$$\lambda = \left[\left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} \right)^k \left(\frac{(1-\hat{\rho})^{k-1}[1+(k-1)\hat{\rho}]}{(1-\rho_0)^{k-1}[1+(k-1)\rho_0]} \right) \right]^{n/2}.$$

Η ασυμπτωτική κατανομή του $-2\log\lambda$ είναι χ_1^2 και απορρίπτουμε την H_0 εάν $-2\log\lambda \geq \chi^2_{1,\alpha}$, όπου α είναι το επιθυμητό επίπεδο σημαντικότητας.

ii) *Το Wald στατιστικό ορίζεται από,*

$$W_n = n \left\{ \frac{k(k-1)}{2(1-\rho_0)^2[1+(k-1)\rho_0]^2} \right\} (\hat{\rho} - \rho_0)^2.$$

Η ασυμπτωτική κατανομή του W_n είναι χ_1^2 και απορρίπτουμε την H_0 εάν $W_n \geq \chi^2_{1,\alpha}$, όπου α είναι το επιθυμητό επίπεδο σημαντικότητας.

iii) *Το Score στατιστικό ορίζεται από,*

$$S_n = \frac{1}{n} \left(\frac{k(k-1)V_0}{2W_0^2(1-\rho_0)^2} \right)^{-1} \left(\sum_{i=1}^n S_i(\hat{\alpha}_0, \hat{\sigma}_0^2, \rho_0) \right)^2.$$

Η ασυμπτωτική κατανομή του S_n είναι χ_1^2 και απορρίπτουμε την H_0 εάν $S_n \geq \chi^2_{1,\alpha}$, όπου α είναι το επιθυμητό επίπεδο σημαντικότητας.

Σε όλα τα παραπάνω οι εκτιμητές $\hat{\rho}, \hat{\sigma}^2$ και $\hat{\alpha}_0$ ορίζονται από τις σχέσεις (2.6), (2.7) και (2.34) αντίστοιχα. Επιπλέον, για τον εκτιμητή $\hat{\sigma}_0^2$ ισχύει ότι,

$$nk\hat{\sigma}_0^2(1-\rho_0)[1+(k-1)\rho_0] = [1+(k-1)\rho_0]trV - \rho_0 1_k' V 1_k + n[1+(k-1)\rho_0]\sum_{i=1}^{k}(\bar{x}_i - \bar{\bar{x}})^2.$$

2.2.4.3 Έλεγχος ότι τα δεδομένα υποστηρίζουν την intraclass δομή

Τέλος, θα κατασκευαστεί το τεστ πηλίκου πιθανοφανειών για τον έλεγχο της υπόθεσης ότι τα δεδομένα υποστηρίζουν την intraclass δομή. Ο Wilks (1946) κατασκεύασε το παραπάνω στατιστικό τεστ το οποίο ελέγχει την ισότητα των μέσων τιμών, διακυμάνσεων και συνδιακυμάνσεων ενός k-διάστατου κανονικού πληθυσμού. Έστω, λοιπόν ένα τυχαίο δείγμα $X_1, X_2, \ldots, X_n$ από έναν κανονικό πληθυσμό έτσι ώστε να ισχύει,

$$X_i \sim N_k(\mu, \Sigma), \quad i = 1, 2, \ldots, n,$$

όπου $\mu = (\alpha_1, \alpha_2, \ldots, \alpha_k)'$ και $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1k}\sigma_1\sigma_k \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \rho_{2k}\sigma_2\sigma_k \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1}\sigma_k\sigma_1 & \rho_{k2}\sigma_k\sigma_2 & \cdots & \sigma_k^2 \end{pmatrix}_{k\times k}$,

όπου σ_j^2 είναι η διακύμανση της j-συνιστώσας του X_i τυχαίου διανύσματος και ρ_{jl} είναι η συσχέτιση μεταξύ j-συνιστώσας και l-συνιστώσας, $j \neq l$, με $j, l = 1, \ldots, k$ και $i = 1, \ldots, n$. Στόχος είναι να κατασκευαστεί το τεστ του πηλίκου μέγιστων πιθανοφανειών για τον έλεγχο της υπόθεσης της ισότητας των μέσων τιμών, διακυμάνσεων και συνδιακυμάνσεων, ή αλλιώς για τον έλεγχο ότι τα δεδομένα προέρχονται από μία συμμετρική πολυδιάστατη κανονική κατανομή.

(A) Ίσες μέσες τιμές

Σκοπός λοιπόν, είναι να ελεγχθεί ότι,

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = \alpha$$

και ταυτόχρονα

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1k}\sigma_1\sigma_k \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \rho_{2k}\sigma_2\sigma_k \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1}\sigma_k\sigma_1 & \rho_{k2}\sigma_k\sigma_2 & \cdots & \sigma_k^2 \end{pmatrix}_{k\times k} = \begin{pmatrix} \sigma^2 & \rho\sigma^2 & \cdots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \cdots & \rho\sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho\sigma^2 & \rho\sigma^2 & \cdots & \sigma^2 \end{pmatrix}_{k\times k}.$$

Το πηλίκο μέγιστων πιθανοφανειών ορίζεται ως,

$$\lambda = \frac{\sup_{\Omega} L}{\sup_{\Omega'} L} = \frac{L(\hat{\alpha}, \hat{\sigma}^2, \hat{\rho})}{L(\hat{\mu}, \hat{\Sigma})},$$

όπου $\Omega' = \{(\mu, \Sigma) : \mu = (\alpha_1, \alpha_2, \ldots, \alpha_k)', \alpha_i \in \mathbb{R} \text{ για } i = 1, 2, \ldots, k, \ \Sigma > 0\}$, Ω ορίζεται από την (2.2), $\hat{\mu}, \hat{\Sigma}$ οι εκτιμητές μέγιστης πιθανοφάνειας των παραμέτρων για $(\mu, \Sigma) \in \Omega'$ και $\hat{\alpha}, \hat{\sigma}^2, \hat{\rho}$ είναι οι εκτιμητές του Θεωρήματος 2.2.1. Για τους εκτιμητές μέγιστης πιθανοφάνειας των παραμέτρων μ και Σ για $(\mu, \Sigma) \in \Omega'$ ισχύει το παρακάτω θεώρημα (βλ. Watson, 1964, Mardia et al., 1979, σ. 103, Anderson and Olkin, 1985).

Θεώρημα 2.2.5. *Έστω $X_1, X_2, \ldots, X_n$ τυχαίο δείγμα από $N_k(\mu, \Sigma)$ κατανομή με $\Sigma > 0$ και μ, Σ άγνωστα. Τότε,*

i) *Ο εκτιμητής μέγιστης πιθανοφάνειας της παραμέτρου μ είναι,*

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{2.35}$$

ii) *Αν $n > k$ ο εκτιμητής μέγιστης πιθανοφάνειας του πίνακα Σ είναι,*

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})' = \frac{1}{n}V. \tag{2.36}$$

Αν εφαρμοστεί το θεώρημα αυτό ο παρονομαστής του πηλίκου μέγιστων πιθανοφανειών γίνεται,

$$L(\hat{\mu}, \hat{\Sigma}) = L(\bar{x}, \hat{\Sigma}) = (2\pi)^{-nk/2}|\hat{\Sigma}|^{-n/2}e^{-nk/2}.$$

Ακόμη, $L(\hat{\alpha}, \hat{\sigma}^2, \hat{\rho}) = (2\pi)^{-nk/2}(\hat{\sigma}^2)^{-nk/2}\left[(1-\hat{\rho})^{k-1}[1+(k-1)\hat{\rho}]\right]^{-n/2}e^{-nk/2}$. Επομένως,

$$\lambda = \frac{L(\hat{\alpha}, \hat{\sigma}^2, \hat{\rho})}{L(\hat{\mu}, \hat{\Sigma})} = \left(\frac{|\hat{\Sigma}|}{(\hat{\sigma}^2)^k\left[(1-\hat{\rho})^{k-1}[1+(k-1)\hat{\rho}]\right]}\right)^{n/2}.$$

Όπως και παραπάνω θα χρησιμοποιήσουμε την ασυμπωτική κατανομή του λ. Στην προκειμένη περίπτωση ισχύει ότι,

$$-2\log\lambda \xrightarrow[n\to\infty]{d} \chi^2_{s-r}, \text{αν η } H_0 \text{ αληθεύει},$$

όπου s συμβολίζει το πλήθος των αγνώστων παραμέτρων του μοντέλου και r ($r < s$) συμβολίζει το πλήθος των αγνώστων παραμέτρων του μοντέλου λαμβάνοντας υπόψη τη μηδενική υπόθεση H_0. Για το δικό μας πρόβλημα ισχύει ότι $s = k + \frac{k(k+1)}{2}$ και $r = 3$. Άρα, καταλήγουμε στο συμπέρασμα ότι,

$$-2\log\lambda \xrightarrow[n\to\infty]{d} \chi^2_{\frac{1}{2}k(k+3)-3}, \text{αν η } H_0 \text{ αληθεύει}.$$

Επιπλέον,

$$\begin{aligned}\alpha &= P(\text{Απορρ. } H_0|H_0 \text{ αληθεύει}) = P(\lambda \leq \lambda_0|H_0 \text{ αληθεύει})\\ &= P(-2\log\lambda \geq -2\log\lambda_0|H_0 \text{ αληθεύει}),\end{aligned}$$

ή

$$\alpha = P(\chi^2_{s-r} \geq -2\log\lambda_0) = P(\chi^2_{\frac{1}{2}k(k+3)-3} \geq -2\log\lambda_0),$$

από την οποία προκύπτει ότι $-2\log\lambda_0 = \chi^2_{\frac{1}{2}k(k+3)-3,\alpha}$.
Έτσι καταλήγουμε στο παρακάτω θεώρημα (βλ. Wilks, 1946).

Θεώρημα 2.2.6. *Έστω $X_1, X_2, \ldots, X_n$ ένα τυχαίο δείγμα από έναν κανονικό πληθυσμό έτσι ώστε να ισχύει, $X_i \sim N_k(\mu, \Sigma)$, $i = 1, 2, \ldots, n$, όπου $\mu = (\alpha_1, \alpha_2, \ldots, \alpha_k)'$ και*

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1k}\sigma_1\sigma_k \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \rho_{2k}\sigma_2\sigma_k \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1}\sigma_k\sigma_1 & \rho_{k2}\sigma_k\sigma_2 & \cdots & \sigma_k^2 \end{pmatrix}_{k\times k} .$$

Έστω ότι θέλουμε να εξετάσουμε την μηδενική υπόθεση ότι οι συνιστώσες α_i του μέσου διανύσματος, οι διακυμάνσεις και οι συνδιακυμάνσεις των συνιστωσών των X_i είναι ίσες. Το στατιστικό του λόγου μέγιστων πιθανοφανειών λ ορίζεται από,

$$\lambda = \frac{L(\hat{\alpha}, \hat{\sigma}^2, \hat{\rho})}{L(\hat{\mu}, \hat{\Sigma})} = \left(\frac{|\hat{\Sigma}|}{(\hat{\sigma}^2)^k \left[(1-\hat{\rho})^{k-1}[1+(k-1)\hat{\rho}]\right]} \right)^{n/2},$$

με $\hat{\rho}, \hat{\sigma}^2$ και $\hat{\Sigma}$ να δίνονται αντίστοιχα από τις σχέσεις (2.6), (2.7) και (2.36). Η ασυμπτωτική κατανομή του $-2\log\lambda$ είναι $\chi^2_{\frac{1}{2}k(k+3)-3}$ και απορρίπτουμε την H_0 εάν $-2\log\lambda \geq \chi^2_{\frac{1}{2}k(k+3)-3,\alpha}$, όπου α είναι το επιθυμητό επίπεδο σημαντικότητας.

Παρατηρήσεις

Οι ροπές της ακριβούς κατανομής του στατιστικού $\lambda^{2/n}$, όταν η H_0 είναι αληθής, έχουν προσδιορισθεί για όλες τις τιμές k (μέγεθος οικογένειας) και n (μέγεθος δείγματος) για τις οποίες τέτοιες κατανομές υπάρχουν, δηλαδή για $k \geq 2$ και

$n > k$. Τότε, η g-οστή ροπή της κατανομής του στατιστικού $\lambda^{2/n}$ δίνεται από τη σχέση (βλ. Wilks, 1946),

$$M_g(\lambda^{2/n}) = (k-1)^{g(k-1)} \prod_{i=2}^{k} \frac{\Gamma\left(\frac{1}{2}(n-i)+g\right)}{\Gamma\left(\frac{1}{2}(n-i)\right)} \cdot \frac{\Gamma\left(\frac{1}{2}(k-1)n\right)}{\Gamma\left(\frac{1}{2}(k-1)(n-1)+g(k-1)\right)}.$$

(Β) Άνισες μέσες τιμές

Εάν η υπόθεση του Θεωρήματος 2.2.6 απορριφθεί, είναι εύλογο κάποιος να αναρωτηθεί εάν το δείγμα θα υποστήριζε την παραπάνω υπόθεση, χωρίς τον περιορισμό ότι οι συνιστώσες $\alpha_1, \alpha_2, \ldots, \alpha_k$ του μέσου διανύσματος μ είναι ίσες με α. Πιο συγκεκριμένα, θα ήθελε να εξετάσει εάν το δείγμα υποστηρίζει τη δομή στην οποία ο πίνακας διακυμάνσεων-συνδιακυμάνσεων Σ είναι της μορφής $\Sigma = \sigma^2[(1-\rho)I_k + \rho J_k]$ και το μέσο διάνυσμα μ είναι αυθαίρετο, δηλαδή $\mu = (\alpha_1, \ldots, \alpha_k)'$ με τα $\alpha_i, i = 1, \ldots, k$, όχι κατ΄ ανάγκη ίσα μεταξύ τους. Ο Wilks (1946) και ο Mentz (2001) κατασκεύασαν το τεστ πηλίκου πιθανοφανειών για τον παραπάνω έλεγχο.

Έστω $X_1, X_2, \ldots, X_n$ τυχαίο δείγμα από έναν κανονικό πληθυσμό έτσι ώστε να ισχύει, $X_i \sim N_k(\mu, \Sigma)$, $i = 1, \ldots, n$, όπου $\mu = (\alpha_1, \alpha_2, \ldots, \alpha_k)'$ και

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1k}\sigma_1\sigma_k \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \rho_{2k}\sigma_2\sigma_k \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1}\sigma_k\sigma_1 & \rho_{k2}\sigma_k\sigma_2 & \cdots & \sigma_k^2 \end{pmatrix}_{k\times k},$$

όπως και παραπάνω.

Θα κατασκευαστεί λοιπόν το αντίστοιχο τεστ πηλίκου πιθανοφανειών για τον έλεγχο της υπόθεσης,

$$H_0 : \Sigma = \sigma^2[(1-\rho)I_k + \rho J_k] \quad \text{έναντι} \quad H_1 : \Sigma \neq \sigma^2[(1-\rho)I_k + \rho J_k].$$

Η συνάρτηση πιθανοφάνειας είναι,

$$\begin{aligned} L(\mu,\sigma^2,\rho|x_1,x_2,\ldots,x_n) &= (2\pi)^{-nk/2}\left[\sigma^{2k}(1-\rho)^{k-1}[1+(k-1)\rho]\right]^{-n/2} \\ &\times \exp\left\{-\frac{1}{2\sigma^2(1-\rho)}trV + \frac{\rho}{2\sigma^2(1-\rho)[1+(k-1)\rho]}1_k'V1_k\right\} \\ &\times \exp\left\{-\frac{n}{2}(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu)\right\}. \end{aligned}$$

Το πηλίκο μέγιστων πιθανοφανειών ορίζεται ως,

$$\lambda = \frac{\sup_{\Omega_0'} L}{\sup_{\Omega'}} = \frac{L(\hat{\mu}_0', \hat{\sigma}_0'^2, \hat{\rho}_0')}{L(\hat{\mu}, \hat{\Sigma})},$$

όπου

$$\begin{aligned} \Omega_0' = \{(\mu,\Sigma) : \mu = (\alpha_1,\alpha_2,\ldots,\alpha_k)', \alpha_i \in \mathbb{R} \text{ για } i = 1,2,\ldots,k \\ \text{και } \Sigma = \sigma^2[(1-\rho)I_k + \rho J_k], \sigma^2 > 0, -\frac{1}{k-1} < \rho < 1\}, \end{aligned}$$

$\hat{\mu}_0', \hat{\sigma}_0'^2, \hat{\rho}_0'$ είναι οι εκτιμητές μέγιστης πιθανοφάνειας των παραμέτρων μ, σ^2, ρ αντίστοιχα όταν $(\mu, \sigma^2, \rho) \in \Omega_0'$. Ακόμη, υπενθυμίζουμε ότι για τον Ω' ισχύει, $\Omega' = \{(\mu,\Sigma) : \mu = (\alpha_1,\alpha_2,\ldots,\alpha_k)', a_i \in \mathbb{R} \text{ για } i = 1,2,\ldots,k,\ \ \Sigma > 0\}$. Για να βρεθούν οι εκτιμητές $\hat{\mu}_0', \hat{\sigma}_0'^2, \hat{\rho}_0'$ πρέπει να μεγιστοποιηθεί η συνάρτηση πιθανοφάνειας ή ισοδύναμα η log-likelihood συνάρτηση ως προς μ, σ^2, ρ, υπό την H_0. Παρακάτω θα περιγράψουμε την διαδικασία από την οποία θα προκύψουν οι εκτιμητές.

Η log-likelihood συνάρτηση δίνεται από την εξίσωση,

$$\begin{aligned} \ell(\mu,\sigma^2,\rho|x_1,x_2,\ldots,x_n) &= -\frac{nk}{2}\log(2\pi) - \frac{nk}{2}\log\sigma^2 \\ &- \frac{n}{2}\log\left[(1-\rho)^{k-1}[1+(k-1)\rho]\right] - \frac{1}{2\sigma^2(1-\rho)}trV \\ &+ \frac{\rho}{2\sigma^2(1-\rho)[1+(k-1)\rho]}1_k'V1_k - \frac{n}{2}(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu). \end{aligned}$$

Είναι φανερό ότι η μεγιστοποίηση της ℓ ως προς την παράμετρο μ θα εξαρτηθεί μόνο από τον όρο $(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu)$, καθώς μόνο αυτός εμπεριέχει την παράμετρο

μ. Επειδή, $\Sigma^{-1} > 0$, για την τετραγωνική μορφή $(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu)$ ισχύει ότι,

$$(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu) \geq 0, \quad \forall \mu.$$

Τώρα η μεγιστοποίηση της ℓ ως προς μ είναι ισοδύναμη με την ελαχιστοποίση της παραπάνω τετραγωνικής μορφής. Η ελάχιστη τιμή που παίρνει η τετραγωνική μορφή είναι ίση με 0 και επιτυγχάνεται για $\bar{x} = \mu$. Επομένως,

$$\hat{\mu}_0' = \bar{x},$$

ή

$$(\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_k) = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k),$$

ή

$$\hat{\alpha}_j = \bar{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}, \text{ για } j = 1, 2, \ldots, k. \tag{2.37}$$

Αντικαθιστούμε $\mu = \bar{x}$ στην log-likelihood συνάρτηση, παραγωγίζουμε ως προς σ^2 και ρ, εξισώνουμε με το 0 και οι εξισώσεις πιθανοφάνειας είναι,

$$nk\sigma^2(1-\rho)[1+(k-1)\rho] = tr(V)[1+(k-1)\rho] - \rho 1_k' V 1_k$$

και

$$\begin{aligned} nk\rho(k-1)\sigma^2(1-\rho)[1+(k-1)\rho] = {} & tr(V)[1+(k-1)\rho]^2 \\ & - 1_k' V 1_k [1+(k-1)\rho^2]. \end{aligned}$$

Αντικαθιστώντας την πρώτη στη δεύτερη προκύπτει μια εξίσωση με μόνη άγνωστη παράμετρο το συντελεστή ρ, από τη λύση της οποίας θα προκύψει ο εκτιμητής μέγιστης πιθανοφάνειας $\hat{\rho}_0'$ του ρ. Καταλήγουμε στο παρακάτω συμπέρασμα,

$$\hat{\rho}_0' = \frac{1_k' V 1_k - trV}{(k-1)trV}. \tag{2.38}$$

Επιστρέφοντας στην πρώτη σχέση, αντικαθιστώντας το $\hat{\rho}'_0$ στη θέση του ρ και λύνοντας ως προς σ^2 προκύπτει ο εκτιμητής μέγιστης πιθανοφάνειας $\hat{\sigma}'^2_0$ του σ^2, που δίνεται από τη σχέση,

$$\hat{\sigma}'^2_0 = \frac{trV}{nk}. \tag{2.39}$$

Υπενθυμίζουμε ότι ο πίνακας V που εμφανίζεται στις σχέσεις (2.38) και (2.39) ορίζεται ως $V = \sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})'$. Έτσι έχουμε ότι,

$$L(\hat{\mu}'_0, \hat{\sigma}'^2_0, \hat{\rho}'_0) = (2\pi)^{-nk/2}(\hat{\sigma}'^2_0)^{-nk/2}\left[(1-\hat{\rho}'_0)^{k-1}[1+(k-1)\hat{\rho}'_0]\right]^{-n/2} e^{-nk/2}.$$

Ακόμη,

$$L(\hat{\mu}, \hat{\Sigma}) = (2\pi)^{-nk/2}|\hat{\Sigma}|^{-n/2}e^{-nk/2}.$$

Είμαστε πλέον σε θέση να υπολογίσουμε το πηλίκο μέγιστων πιθανοφανειών,

$$\begin{aligned}\lambda = \frac{L(\hat{\mu}'_0, \hat{\sigma}'^2_0, \hat{\rho}'_0)}{L(\hat{\mu}, \hat{\Sigma})} &= \frac{(2\pi)^{-nk/2}(\hat{\sigma}'^2_0)^{-nk/2}\left[(1-\hat{\rho}'_0)^{k-1}[1+(k-1)\hat{\rho}'_0]\right]^{-n/2} e^{-nk/2}}{(2\pi)^{-nk/2}|\hat{\Sigma}|^{-n/2}e^{-nk/2}} \\ &= \left(\frac{|\hat{\Sigma}|}{(\hat{\sigma}'^2_0)^k\left[(1-\hat{\rho}'_0)^{k-1}[1+(k-1)\hat{\rho}'_0]\right]}\right)^{n/2}.\end{aligned}$$

Θα χρησιμοποιήσουμε κι εδώ την ασυμπωτική κατανομή του λ. Έτσι, έχουμε,

$$-2\log\lambda \xrightarrow[n\to\infty]{d} \chi^2_{s-r}, \text{αν η } H_0 \text{ αληθεύει},$$

όπου τώρα $s = k + \frac{k(k+1)}{2}$ και $r = k+2$. Άρα, καταλήγουμε στο συμπέρασμα ότι,

$$-2\log\lambda \xrightarrow[n\to\infty]{d} \chi^2_{\frac{1}{2}k(k+1)-2}, \text{αν η } H_0 \text{ αληθεύει}.$$

Επιπλέον,

$$\begin{aligned}\alpha &= P(\text{Απορρ. } H_0|H_0 \text{ αληθεύει}) = P(\lambda \le \lambda_0|H_0 \text{ αληθεύει}) \\ &= P(-2\log\lambda \ge -2\log\lambda_0|H_0 \text{ αληθεύει})\end{aligned}$$

ή

$$\alpha = P(\chi^2_{s-r} \geq -2\log\lambda_0) = P(\chi^2_{\frac{1}{2}k(k+1)-2} \geq -2\log\lambda_0),$$

από την οποία προκύπτει ότι $-2\log\lambda_0 = \chi^2_{\frac{1}{2}k(k+1)-2,\alpha}$. Η προηγούμενη ανάλυση οδηγεί στο παρακάτω θεώρημα (βλ. Wilks, 1946 και Mentz, 2001).

Θεώρημα 2.2.7. *Έστω $X_1, X_2, \ldots, X_n$ τυχαίο δείγμα από έναν πληθυσμό, έτσι ώστε να ισχύει, $X_i \sim N_k(\mu, \Sigma)$, $i = 1, 2, \ldots, n$, όπου $\mu = (\alpha_1, \alpha_2, \ldots, \alpha_k)'$ και*

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1k}\sigma_1\sigma_k \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \rho_{2k}\sigma_2\sigma_k \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1}\sigma_k\sigma_1 & \rho_{k2}\sigma_k\sigma_2 & \cdots & \sigma_k^2 \end{pmatrix}_{k\times k}.$$

Έστω ότι θελουμε να ελέγξουμε την υπόθεση,

$$H_0 : \Sigma = \sigma^2[(1-\rho)I_k + \rho J_k] \quad \text{έναντι} \quad H_1 : \Sigma \neq \sigma^2[(1-\rho)I_k + \rho J_k].$$

Το στατιστικό του λόγου μέγιστων πιθανοφανειών λ ορίζεται από,

$$\lambda = \left(\frac{|\hat{\Sigma}|}{(\hat{\sigma}_0'^2)^k\,[(1-\hat{\rho}_0')^{k-1}[1+(k-1)\hat{\rho}_0']]} \right)^{n/2},$$

με $\hat{\rho}_0', \hat{\sigma}_0'^2$ και $\hat{\Sigma}$ να δίνονται από τις σχέσεις (2.38), (2.39) και (2.36) αντίστοιχα. Η ασυμπτωτική κατανομή του $-2log\lambda$ είναι $\chi^2_{\frac{1}{2}k(k+1)-2}$ και απορρίπτουμε την H_0 εάν $-2\log\lambda \geq \chi^2_{\frac{1}{2}k(k+1)-2,\alpha}$, όπου α είναι το επιθυμητό επίπεδο σημαντικότητας.

Παρατηρήσεις

Οι ροπές της ακριβούς κατανομής του στατιστικού $\lambda^{2/n}$, όταν η H_0 είναι αληθής, έχουν προσδιορισθεί για όλες τις τιμές k (μέγεθος οικογένειας) και n (μέγεθος δείγματος) για τις οποίες τέτοιες κατανομές υπάρχουν, δηλαδή για $k \geq 2$ και

$n > k$. Τότε, η g-οστή ροπή της κατανομής του στατιστικού $\lambda^{2/n}$ δίνεται από τη σχέση (βλ. Wilks, 1946),

$$M_g(\lambda^{2/n}) = (k-1)^{g(k-1)} \prod_{i=2}^{k} \frac{\Gamma\left(\frac{1}{2}(n-i)+g\right)}{\Gamma\left(\frac{1}{2}(n-i)\right)} \cdot \frac{\Gamma\left(\frac{1}{2}(k-1)(n-1)\right)}{\Gamma\left(\frac{1}{2}(k-1)(n-1)+g(k-1)\right)}.$$

Στη συνέχεια θα αναπτυχθεί στατιστική συμπερασματολογία, όπως και πριν, με την υπόθεση ότι για ένα τυχαίο δείγμα $X_1, \ldots, X_n$ μεγέθους n, $X_i \sim N_k(\mu, \Sigma)$ για $i = 1, 2, \ldots, n$, όπου $\mu = (\alpha_1, \alpha_2, \ldots, \alpha_k)'$ και $\Sigma = \sigma^2[(1-\rho)I_k + \rho J_k]$. Ο παραμετρικός χώρος του προβλήματος συμβολίζεται με Ω_0' και ορίζεται ως,

$$\Omega_0' = \Big\{(\mu, \Sigma) : \mu = (\alpha_1, \ldots, \alpha_k)', \alpha_i \in \mathbb{R}, i = 1, \ldots, k, \Sigma = \sigma^2[(1-\rho)I_k + \rho J_k], \sigma^2 > 0, -\frac{1}{k-1} < \rho < 1\Big\}. \tag{2.40}$$

2.2.5 Η παράμετρος μ αυθαίρετη

Η εύρεση των εκτιμητών μέγιστης πιθανοφάνειας έχει πραγματοποιηθεί παραπάνω, αφού οι εκτιμητές αυτοί ήταν απαραίτητοι για την κατασκευή του πηλίκου μέγιστων πιθανοφανειών. Συνοπτικά, και χάριν πληρότητας, οι εκτιμητές μέγιστης πιθανοφάνειας των παραμέτρων μ, σ^2 και ρ δίνονται στο ακόλουθο θεώρημα (βλ. Wilks, 1946 και Mentz, 2001).

Θεώρημα 2.2.8. *Έστω $X_1, X_2, \ldots, X_n$ τυχαίο δείγμα μεγέθους n από πληθυσμό που ικανοποιεί την (2.40). Εάν συμβολίσουμε τους εκτιμητές μέγιστης πιθανοφάνειας των παραμέτρων μ, σ^2, ρ με $\hat{\mu}, \hat{\sigma}^2, \hat{\rho}$, αντίστοιχα, τότε,*

$$\hat{\mu} = \bar{x}, \ \ \hat{\sigma}^2 = \frac{trV}{nk}, \ \ \hat{\rho} = \frac{1_k' V 1_k - trV}{(k-1)trV},$$

με $V = \sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})'$.

Στη συνέχεια θα προσδιορισθεί η ασυμπτωτική κατανομή των εκτιμητών στην περίπτωση, όπου το μ είναι αυθαίρετο. Ο πίνακας πληροφορίας του Fisher που αντιστοιχεί σε ολόκληρο το δείγμα μεγέθους n, έστω $I_n(\mu, \sigma^2, \rho)$ και ο αντίστροφός του, $[I_n(\mu, \sigma^2 \rho)]^{-1}$, δίνονται από τις σχέσεις,

$$I_n(\mu, \sigma^2, \rho) = \begin{pmatrix} I_n(\mu)_{k\times k} & 0_{k\times 1} & 0_{k\times 1} \\ 0_{1\times k} & \frac{nk}{2(\sigma^2)^2} & \frac{nk(k-1)\rho}{2\sigma^2(1-\rho)W} \\ 0_{1\times k} & \frac{nk(k-1)\rho}{2\sigma^2(1-\rho)W} & \frac{nk(k-1)V}{2W^2(1-\rho)^2} \end{pmatrix}_{(k+2)\times(k+2)}, \tag{2.41}$$

όπου

$$I_n(\mu) = \begin{pmatrix} \frac{n[1+(k-2)\rho]}{\sigma^2(1-\rho)W} & -\frac{n\rho}{\sigma^2(1-\rho)W} & \cdots & -\frac{n\rho}{\sigma^2(1-\rho)W} \\ -\frac{n\rho}{\sigma^2(1-\rho)W} & \frac{n[1+(k-2)\rho]}{\sigma^2(1-\rho)W} & \cdots & -\frac{n\rho}{\sigma^2(1-\rho)W} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{n\rho}{\sigma^2(1-\rho)W} & -\frac{n\rho}{\sigma^2(1-\rho)W} & \cdots & \frac{n[1+(k-2)\rho]}{\sigma^2(1-\rho)W} \end{pmatrix}_{k\times k}.$$

Ο αντίστροφος του παραπάνω πίνακα δίνεται από τη σχέση,

$$[I_n(\mu, \sigma^2 \rho)]^{-1} = \begin{pmatrix} [I_n(\mu)]^{-1}_{k\times k} & 0_{k\times 1} & 0_{k\times 1} \\ 0_{1\times k} & \frac{2V(\sigma^2)^2}{nk} & -\frac{2\rho\sigma^2(1-\rho)W}{nk} \\ 0_{1\times k} & -\frac{2\rho\sigma^2(1-\rho)W}{nk} & \frac{2(1-\rho)^2W^2}{nk(k-1)} \end{pmatrix}_{(k+2)\times(k+2)}, \tag{2.42}$$

όπου

$$[I_n(\mu)]^{-1} = \begin{pmatrix} \frac{\sigma^2}{n} & \rho\frac{\sigma^2}{n} & \cdots & \rho\frac{\sigma^2}{n} \\ \rho\frac{\sigma^2}{n} & \frac{\sigma^2}{n} & \cdots & \rho\frac{\sigma^2}{n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho\frac{\sigma^2}{n} & \rho\frac{\sigma^2}{n} & \cdots & \frac{\sigma^2}{n} \end{pmatrix}_{k\times k},$$

$W = 1 + (k-1)\rho$ και $V = 1 + (k-1)\rho^2$. Όπως και παραπάνω, επικαλούμαστε το Θεώρημα 2.2.2 που διατυπώνει την ασυμπτωτική κανονικότητα των εκτιμητών μέγιστης πιθανοφάνειας. Έτσι έχουμε,

$$\sqrt{n}\left[(\hat{\mu}, \hat{\sigma}^2, \hat{\rho})' - (\mu, \sigma^2, \rho)'\right] \xrightarrow{d} N_{k+2}\left(0, [I(\mu, \sigma^2, \rho)]^{-1}\right)$$

ή

$$(\hat{\mu}, \hat{\sigma}^2, \hat{\rho})' \xrightarrow{d} N_{k+2}\left((\mu, \sigma^2, \rho)', [I_n(\mu, \sigma^2, \rho)]^{-1}\right).$$

Από το Θεώρημα 2.2.2 προκύπτει ότι για τον εκτιμητή $\hat{\rho}$ ισχύει ότι,

$$\sqrt{n}(\hat{\rho} - \rho) \xrightarrow{d} N\left(0, \frac{2(1-\rho)^2 W^2}{k(k-1)}\right)$$

ή

$$\hat{\rho} \xrightarrow{d} N\left(\rho, \frac{2(1-\rho)^2 W^2}{nk(k-1)}\right),$$

όπου $W^2 = [1 + (k-1)\rho]^2$. Παρατηρούμε λοιπόν ότι οι εκτιμητές μέγιστης πιθανοφάνειας που προέκυψαν για την intraclass συσχέτιση ρ είναι διαφορετικοί για τις περιπτώσεις $\mu = (\alpha, \alpha, \ldots, \alpha)'$ και μ αυθαίρετο, δηλαδή $\mu = (\alpha_1, \ldots, \alpha_k)'$ με τα $\alpha_i, i = 1, \ldots, k$, όχι κατ' ανάγκη ίσα μεταξύ τους. Ωστόσο, η ασυμπτωτική κατανομή τους είναι η ίδια. Αυτό έχει ως αποτέλεσμα τα ασυμπτωτικά διαστήματα εμπιστοσύνης που θα προκύψουν για την περίπτωση αυτή να είναι τα ίδια με αυτά που προέκυψαν για την περίπτωση όπου μ όχι αυθαίρετο, αλλά της μορφής $(\alpha, \ldots, \alpha)'$ (βλ. σχέσεις (2.20) και (2.21)).

Το εδάφιο αυτό θα ολοκληρωθεί με την κατασκευή στατιστικών τεστ για την intraclass συσχέτιση ρ στην περίπτωση αυθαίρετου μ. Στο πλαίσιο αυτό, αρχικά θα κατασκευαστεί το στατιστικό τεστ για τον έλεγχο της υπόθεσης $\rho = \rho_0$ με τη μέθοδο του πηλίκου μέγιστων πιθανοφανειών. Έστω $X_1, X_2, \ldots, X_n$ τυχαίο δείγμα από πληθυσμό με κατανομή $N_k(\mu, \Sigma)$ με $\mu = (\alpha_1, \alpha_2, \ldots, \alpha_k)'$ και $\Sigma = \sigma^2[(1-\rho)I_k + \rho J_k]$. Έστω ο έλεγχος της υπόθεσης,

$$H_0 : \rho = \rho_0 \quad \text{έναντι} \quad H_1 : \rho \neq \rho_0.$$

Η συνάρτηση πιθανοφάνειας είναι,

$$\begin{aligned}L(\mu,\sigma^2,\rho|x_1,x_2,\ldots,x_n) &= (2\pi)^{-nk/2}\left[\sigma^{2k}(1-\rho)^{k-1}[1+(k-1)\rho]\right]^{-n/2}\\ &\times\exp\left\{-\frac{1}{2\sigma^2(1-\rho)}trV+\frac{\rho}{2\sigma^2(1-\rho)[1+(k-1)\rho]}1_k'V1_k\right\}\\ &\times\exp\left\{-\frac{n}{2}(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu)\right\}.\end{aligned}$$

Το πηλίκο μέγιστων πιθανοφανειών ορίζεται ως,

$$\lambda=\frac{L(\hat{\mu}_0,\hat{\sigma}_0^2,\rho_0)}{L(\hat{\mu},\hat{\sigma}^2,\hat{\rho})},$$

όπου $\hat{\mu}_0,\hat{\sigma}_0^2$ είναι οι εκτιμητές μέγιστης πιθανοφάνειας των παραμετρων μ,σ^2 αντίστοιχα υπό την H_0, ενώ οι εκτιμητές $\hat{\mu},\hat{\sigma}^2,\hat{\rho}$ είναι οι εκτιμητές που υπολογίστηκαν παραπάνω (βλ. Θεώρημα 2.2.8). Για να προσδιορίσουμε τους εκτιμητές $\hat{\mu}_0,\hat{\sigma}_0^2$ αντικαθιστούμε στη συνάρτηση πιθανοφάνειας για $\rho=\rho_0$ και μεγιστοποιούμε ως προς τις παραμέτρους μ και σ^2. Τα αποτελέσματα που προέκυψαν για τον $\hat{\mu}_0$ είναι ότι υπό την H_0,

$$\hat{\mu}_0=\hat{\mu}=\bar{x}. \tag{2.43}$$

Αντικαθιστώντας τον εκτιμητή μέγιστης πιθανοφάνειας της παραμέτρου μ, παραγωγίζοντας ως προς σ^2 και εξισώνοντας με το 0 προκύπτει ότι για τον εκτιμητή μέγιστης πιθανοφάνειας $\hat{\sigma}_0^2$ της παραμέτρου σ^2, ισχύει,

$$nk\hat{\sigma}_0^2(1-\rho_0)[1+(k-1)\rho_0]=[1+(k-1)\rho_0]trV-\rho_0 1_k'V1_k. \tag{2.44}$$

Με χρήση της παραπάνω σχέσης είναι,

$$L(\hat{\mu}_0,\hat{\sigma}_0^2,\rho_0)=(2\pi)^{-nk/2}(\hat{\sigma}_0^2)^{-nk/2}\left[(1-\rho_0)^{k-1}[1+(k-1)\rho_0]\right]^{-n/2}e^{-nk/2}.$$

Ακόμη,

$$L(\hat{\mu},\hat{\sigma}^2,\hat{\rho})=(2\pi)^{-nk/2}(\hat{\sigma}^2)^{-nk/2}\left[(1-\hat{\rho})^{k-1}[1+(k-1)\hat{\rho}]\right]^{-n/2}e^{-nk/2}.$$

Είμαστε πλέον σε θέση να υπολογίσουμε το πηλίκο μέγιστων πιθανοφανειών το οποίο δίνεται από την παρακάτω σχέση,

$$\begin{aligned}
\lambda = \frac{L(\hat{\mu}_0, \hat{\sigma}_0^2, \rho_0)}{L(\hat{\mu}, \hat{\sigma}^2, \hat{\rho})} &= \frac{(2\pi)^{-nk/2}(\hat{\sigma}_0^2)^{-nk/2}\left[(1-\rho_0)^{k-1}[1+(k-1)\rho_0]\right]^{-n/2} e^{-nk/2}}{(2\pi)^{-nk/2}(\hat{\sigma}^2)^{-nk/2}\left[(1-\hat{\rho})^{k-1}[1+(k-1)\hat{\rho}]\right]^{-n/2} e^{-nk/2}} \\
&= \left[\left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^k \frac{(1-\hat{\rho})^{k-1}[1+(k-1)\hat{\rho}]}{(1-\rho_0)^{k-1}[1+(k-1)\rho_0]}\right]^{n/2}.
\end{aligned}$$

Για την ασυμπτωτική κατανομή του λ ισχύει ότι (βλ. Θεώρημα 2.2.3),

$$-2\log\lambda \xrightarrow[n\to\infty]{d} \chi_1^2, \text{ αν η } H_0 \text{ αληθεύει.}$$

Απορρίπτουμε την H_0 εάν $-2\log\lambda \geq \chi_{1,\alpha}^2$, όπου α είναι το επιθυμητό επίπεδο σημαντικότητας.

Η κατασκευή του Wald στατιστικού δεν περιλαμβάνει τις παραμέτρους που δεν αφορούν την μηδενική υπόθεση. Ισχύει ότι,

$$\sqrt{n}(\hat{\rho}-\rho_0) \xrightarrow{d} N\left(0, \frac{2(1-\rho_0)^2[1+(k-1)\rho_0]^2}{k(k-1)}\right), \text{ υπό την } H_0.$$

Τότε το Wald στατιστικό δίνεται από τη σχέση,

$$W_n = n\left\{\frac{k(k-1)}{2(1-\rho_0)^2[1+(k-1)\rho_0]^2}\right\}(\hat{\rho}-\rho_0)^2,$$

με $\hat{\rho}$ να δίνεται από τη σχέση (2.38). Για την ασυμπτωτική κατανομή του W_n ισχύει ότι,

$$W_n \xrightarrow[n\to\infty]{d} \chi_1^2, \text{ εάν η } H_0 \text{ είναι αληθής.}$$

Απορρίπτουμε την H_0 εάν $W_n \geq \chi_{1,\alpha}^2$, όπου α είναι το επιθυμητό επίπεδο σημαντικότητας.

Όσον αφορά το Score τεστ έχουμε ότι,

$$\begin{aligned}
\log f(X_i|\mu, \sigma^2, \rho) = &-\frac{k}{2}\log(2\pi) - \frac{k}{2}\log\sigma^2 - \frac{1}{2}\log\left[(1-\rho)^{k-1}[1+(k-1)\rho]\right] \\
&-\frac{1}{2\sigma^2(1-\rho)}\left[(X_i-\mu)'(X_i-\mu) - \frac{\rho}{1+(k-1)\rho}(X_i-\mu)'J_k(X_i-\mu)\right],
\end{aligned}$$

για $i = 1, 2, \ldots, n$. Θέτουμε,

$$S_i(\mu,\sigma^2,\rho) = \frac{\partial}{\partial\rho}\log f(X_i|\mu,\sigma^2,\rho) = \frac{k\rho(k-1)}{2(1-\rho)[1+(k-1)\rho]}$$
$$- \frac{1}{2\sigma^2(1-\rho)^2}\left[(X_i-\mu)'(X_i-\mu) - \frac{1+\rho^2(k-1)}{[1+(k-1)\rho]^2}(X_i-\mu)'J_k(X_i-\mu)\right],$$

για $i = 1, \ldots, n$. Τότε υπό την H_0 ισχύει ότι,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_i(\hat{\mu}_0, \hat{\sigma}_0^2, \rho_0) \xrightarrow[n\to\infty]{d} N\left(0, \frac{k(k-1)V_0}{2W_0^2(1-\rho_0)^2}\right),$$

όπου $W_0 = 1+(k-1)\rho_0$ και $V_0 = 1+(k-1)\rho_0^2$. Το Score στατιστικό ορίζεται από τη σχέση,

$$S_n = \frac{1}{n}\left(\frac{k(k-1)V_0}{2W_0^2(1-\rho_0)^2}\right)^{-1}\left(\sum_{i=1}^{n} S_i(\hat{\mu}_0, \hat{\sigma}_0^2, \rho_0)\right)^2,$$

με $\hat{\mu}_0$ και $\hat{\sigma}_0^2$ να δίνονται από τις σχέσεις (2.43) και (2.44) αντίστοιχα. Για την ασυμπτωτική κατανομή του S_n ισχύει ότι,

$$S_n \xrightarrow[n\to\infty]{d} \chi_1^2, \text{ εάν η } H_0 \text{ είναι αληθής.}$$

Απορρίπτουμε την H_0 εάν $S_n \geq \chi_{1,\alpha}^2$, όπου α είναι το επιθυμητό επίπεδο σημαντικότητας. Όλα τα παραπάνω συνοψίζονται στο παρακάτω Θεώρημα:

Θεώρημα 2.2.9. *Έστω $X_1, X_2, \ldots, X_n$ τυχαίο δείγμα από πληθυσμό που ικανοποιεί την (2.40). Υποθέτουμε ότι ισχύουν οι συνθήκες που εξασφαλίζουν την ασυμπτωτική κανονικότητα των εκτιμητών μέγιστης πιθανοφάνειας. Έστω ότι θέλουμε να ελέγξουμε την παρακάτω υπόθεση,*

$$H_0 : \rho = \rho_0 \quad \text{έναντι} \quad H_1 : \rho \neq \rho_0.$$

i) *Το στατιστικό του λόγου μέγιστων πιθανοφανειών λ ορίζεται από,*

$$\lambda = \left[\left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^k \frac{(1-\hat{\rho})^{k-1}[1+(k-1)\hat{\rho}]}{(1-\rho_0)^{k-1}[1+(k-1)\rho_0]}\right]^{n/2},$$

όπου $\hat{\rho}, \hat{\sigma}^2$ *και* $\hat{\sigma}_0^2$ *δίνονται, αντίστοιχα, από τις σχέσεις (2.38), (2.39) και (2.44). Η ασυμπτωτική κατανομή του* $-2\log\lambda$ *είναι* χ_1^2 *και απορρίπτουμε την* H_0 *εάν* $-2\log\lambda \geq \chi_{1,\alpha}^2$*, όπου* α *είναι το επιθυμητό επίπεδο σημαντικότητας.*

ii) *Το Wald στατιστικό ορίζεται από,*

$$W_n = n\left\{\frac{k(k-1)}{2(1-\rho_0)^2[1+(k-1)\rho_0]^2}\right\}(\hat{\rho}-\rho_0)^2,$$

όπου $\hat{\rho}$ *δίνεται από τη σχέση (2.38). Η ασυμπτωτική κατανομή του* W_n *είναι* χ_1^2 *και απορρίπτουμε την* H_0 *εάν* $W_n \geq \chi_{1,\alpha}^2$*, όπου* α *είναι το επιθυμητό επίπεδο σημαντικότητας.*

iii) *Το Score στατιστικό ορίζεται από,*

$$S_n = \frac{1}{n}\left(\frac{k(k-1)V_0}{2W_0^2(1-\rho_0)^2}\right)^{-1}\left(\sum_{i=1}^{n} S_i(\hat{\mu}_0, \hat{\sigma}_0^2, \rho_0)\right)^2,$$

όπου $\hat{\mu}_0$ *και* $\hat{\sigma}_0^2$ *δίνονται από τις σχέσεις (2.43) και (2.44) αντίστοιχα και* $W_0 = 1+(k-1)\rho_0$, $V_0 = 1+(k-1)\rho_0^2$*. Η ασυμπτωτική κατανομή του* S_n *είναι* χ_1^2 *και απορρίπτουμε την* H_0 *εάν* $S_n \geq \chi_{1,\alpha}^2$*, όπου* α *είναι το επιθυμητό επίπεδο σημαντικότητας.*

2.3 Unbalanced περίπτωση

Εξετάζεται τώρα η περίπτωση στην οποία το μέγεθος των οικογενειών ποικίλλει από οικογένεια σε οικογένεια, περίπτωση η οποία συναντάται πιο συχνά στην πράξη. Στην περίπτωση αυτή οι εκτιμητές μέγιστης πιθανοφάνειας δεν υπάρχουν σε κλειστή μορφή. Ωστόσο, ο εκτιμητής μέγιστης πιθανοφάνειας της intraclass συσχέτισης ρ μπορεί να υπολογιστεί με τη χρήση αριθμητικών μεθόδων. Έστω ότι για το παρακάτω μοντέλο ισχύει, $X_i \sim N_{k_i}(\mu_i, \Sigma_i)$, $i = 1, \ldots, n$, όπου

$\mu_i = (\alpha, \alpha, \ldots, \alpha)'$ και $\Sigma_i = \sigma^2[(1-\rho)I_{k_i} + \rho J_{k_i}]$. Με άλλα λόγια, υποθέτουμε ότι το μοντέλο (2.1) ισχύει.

2.3.1 Εκτίμηση

Στο εδάφιο αυτό περιγράφεται η διαδικασία που ακολουθούμε για την απόκτηση του εκτιμητή μέγιστης πιθανοφάνειας της intraclass συσχέτισης ρ σύμφωνα με τους Donner and Koval (1980a). Έστω $X_1, X_2, \ldots, X_n$ ένα τυχαίο δείγμα από την παραπάνω κατανομή. Τότε η συνάρτηση πιθανοφάνειας του δείγματος δίνεται από τη σχέση,

$$\begin{aligned} L(\alpha, \sigma^2, \rho | x_1, x_2, \ldots, x_n) = &(2\pi)^{-N/2} \prod_{i=1}^{n} |\Sigma_i|^{-1/2} \\ &\times \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (X_i - \mu_i)' \Sigma_i^{-1} (X_i - \mu_i) \right\}, \end{aligned}$$

όπου $N = \sum\limits_{i=1}^{n} k_i$.

Για τον πίνακα $\Sigma_i, \;\; i = 1, 2, \ldots, n$ ισχύουν (βλ. Mardia et al., 1979, σ. 461-462) οι παρακάτω σχέσεις,

$$\Sigma_i^{-1} = \frac{1}{\sigma^2(1-\rho)}[I_{k_i} + \rho[1 + (k_i - 1)\rho]^{-1} J_{k_i}], \tag{2.45}$$

$$|\Sigma_i| = \sigma^{2k_i}(1-\rho)^{k_i - 1}[1 + (k_i - 1)\rho]. \tag{2.46}$$

Έστω $W_i = 1 + (k_i - 1)\rho, \;\; i = 1, 2, \ldots, n$. Τότε με αντικατάσταση των παραπάνω σχέσεων είναι εύκολο να δειχθεί ότι,

$$\begin{aligned} -2\log L(\alpha, \sigma^2, \rho) = N(\log\sigma^2 + \log 2\pi) + \log(1-\rho)\sum_{i=1}^{n}(k_i - 1) + \sum_{i=1}^{n}\log W_i \\ + \frac{1}{\sigma^2(1-\rho)}\left\{ \sum_{i=1}^{n}\left(\frac{W_i - \rho}{W_i}\right)\sum_{j=1}^{k_i}(X_{ij} - \alpha)^2 - \rho\sum_{i=1}^{n}\sum_{j=1}^{k_i}\sum_{l\neq j}^{k_i}\frac{(X_{ij} - \alpha)(X_{il} - \alpha)}{W_i} \right\}. \end{aligned}$$

Παραγωγίζοντας την παραπάνω σχέση ως προς α και σ^2 και επιλύοντας τις εξισώσεις πιθανοφάνειας προκύπτει ότι οι εκτιμητές μέγιστης πιθανοφάνειας $\hat{\alpha}$ και $\hat{\sigma}^2$ των παραμέτρων α και σ^2 αντίστοιχα, ικανοποιούν τις εξισώσεις,

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} \frac{k_i \bar{X}_i}{W_i}}{\sum_{i=1}^{n} \frac{k_i}{W_i}}, \text{ όπου } \bar{X}_i = \sum_{j=1}^{k_i} \frac{X_{ij}}{k_i}$$

και

$$\hat{\sigma}^2 = \frac{1}{N(1-\rho)} \left\{ \sum_{i=1}^{n} \left(\frac{W_i - \rho}{W_i} \right) \sum_{j=1}^{k_i} (X_{ij} - \hat{\alpha})^2 - \rho \sum_{i=1}^{n} \sum_{j=1}^{k_i} \sum_{l \neq j}^{k_i} \frac{(X_{ij} - \hat{\alpha})(X_{il} - \hat{\alpha})}{W_i} \right\}.$$

Αντικαθιστώντας τις δύο παραπάνω σχέσεις στην εξίσωση της $-2 \log L$ προκύπτει,

$$-2 \log L = N(1 + \log \hat{\sigma}^2 + \log 2\pi) + (N - n) \log (1 - \rho) + \sum_{i=1}^{n} \log W_i.$$

Για να αποκτηθεί ο εκτιμητής μέγιστης πιθανοφάνειας της παραμέτρου ρ, έστω $\hat{\rho}$, θα πρέπει να ελαχιστοποιηθεί η παραπάνω σχέση ως προς ρ. Αυτή η ελαχιστοποίηση πραγματοποιείται με χρήση αριθμητικών μεθόδων.

2.3.2 Ασυμπτωτική κατανομή

Οι Donner and Koval (1980b) προσδιόρισαν την ασυμπτωτική διακύμανση του εκτιμητή $\hat{\rho}$. Αυτό έγινε αντιστρέφοντας τον πίνακα πληροφορίας του Fisher των παραμέτρων α, σ^2, ρ, που αντιστοιχεί στο τυχαίο δείγμα μεγέθους n, $I_n(\alpha, \sigma^2, \rho)$. Έστω, για $i = 1, \ldots, n$,

$$W_i = 1 + (k_i - 1)\rho,$$
$$V_i = 1 + (k_i - 1)\rho^2$$

και
$$d = N\sum_{i=1}^{n} k_i(k_i-1)V_iW_i^{-2} - \rho^2\left(\sum_{i=1}^{n} k_i(k_i-1)W_i^{-1}\right)^2. \qquad (2.47)$$
Τότε ισχύουν τα παρακάτω,

$$I_n(\alpha,\sigma^2,\rho)$$
$$= \begin{pmatrix} \frac{1}{\sigma^2}\sum\limits_{i=1}^{n} k_iW_i^{-1} & 0 & 0 \\ 0 & \frac{1}{2(\sigma^2)^2}N & \frac{\rho}{2\sigma^2(1-\rho)}\sum\limits_{i=1}^{n} k_i(k_i-1)W_i^{-1} \\ 0 & \frac{\rho}{2\sigma^2(1-\rho)}\sum\limits_{i=1}^{n} k_i(k_i-1)W_i^{-1} & \frac{1}{2(1-\rho)^2}\sum\limits_{i=1}^{n} k_i(k_i-1)V_iW_i^{-2} \end{pmatrix}_{3\times 3}$$
και

$$[I_n(\alpha,\sigma^2,\rho)]^{-1}$$
$$= \begin{pmatrix} \frac{\sigma^2}{\sum_{i=1}^{n} k_iW_i^{-1}} & 0 & 0 \\ 0 & \frac{2(\sigma^2)^2\sum_{i=1}^{n} k_i(k_i-1)V_iW_i^{2}}{d} & -\frac{2\sigma^2(1-\rho)\rho\sum_{i=1}^{n} k_i(k_i-1)W_i^{-1}}{d} \\ 0 & -\frac{2\sigma^2(1-\rho)\rho\sum_{i=1}^{n} k_i(k_i-1)W_i^{-1}}{d} & \frac{2N(1-\rho)^2}{d} \end{pmatrix}_{3\times 3}.$$
Οι Donner and Koval (1980b) αναφέρουν πως ένα ασυμπτωτικό σφάλμα για τον εκτιμητή ρ μπορεί να εκτιμηθεί αντικαθιστώντας $\hat{\rho}$ στη θέση του ρ στη σχέση $\frac{2N(1-\rho)^2}{d}$. Έτσι η εκτιμώμενη διακύμανση (ασυμπτωτική) θα δίνεται από τη σχέση:
$$\widehat{Var}(\hat{\rho}) = \frac{2N(1-\hat{\rho})^2}{N\sum_{i=1}^{n} k_i(k_i-1)\hat{V}_i\hat{W}_i^{-2} - \hat{\rho}^2\left(\sum_{i=1}^{n} k_i(k_i-1)\hat{W}_i^{-1}\right)^2}, \qquad (2.48)$$
όπου $\hat{W}_i = 1+(k_i-1)\hat{\rho}, \hat{V}_i = 1+(k_i-1)\hat{\rho}^2$, για $i = 1,2,\ldots,n$.

2.3.3 Ασυμπτωτικά διαστήματα εμπιστοσύνης

Ένα $100(1-\alpha)\%$ ασυμπτωτικό διάστημα εμπιστοσύνης για την παράμετρο ρ είναι το (βλ. Donner and Wells, 1986),
$$\left(\hat{\rho} - z_{a/2}\sqrt{\widehat{Var}(\hat{\rho})}, \hat{\rho} + z_{a/2}\sqrt{\widehat{Var}(\hat{\rho})}\right), \qquad (2.49)$$

όπου $100(1-\alpha)\%$ είναι ο επιθυμητός βαθμός εμπιστοσύνης και $z_{a/2}$ είναι το αντίστοιχο εκατοστιαίο σημείο της τυπικής κανονικής κατανομής.

2.3.4 Στατιστικά τεστ

Όπως αναφέρθηκε και παραπάνω, οι εκτιμητές μέγιστης πιθανοφάνειας δεν υπάρχουν σε κλειστή μορφή. Ωστόσο και στην περίπτωση αυτή μπορούν να κατασκευαστούν στατιστικά τεστ τα οποία βασίζονται στη μέθοδο του πηλίκου μέγιστων πιθανοφανειών. Παρακάτω περιγράφεται η διαδικασία που θα ακολουθηθεί για την κατασκευή αυτών των τεστ για τον έλεγχο της υπόθεσης,

$$H_0 : \rho = \rho_0 \quad \text{έναντι της} \quad H_1 : \rho \neq \rho_0.$$

Έστω λοιπόν $X_1, X_2, \ldots, X_n$ τυχαίο δείγμα από την $N_{k_i}(\mu_i, \Sigma_i), i = 1, 2, \ldots, n,$ (μοντέλο 2.1). Τότε το πηλίκο μέγιστων πιθανοφανειών ορίζεται ως (βλ. Helu, 2007),

$$\lambda = \frac{L(\hat{\alpha}_0, \hat{\sigma}_0^2, \rho_0)}{L(\hat{\alpha}, \hat{\sigma}^2, \hat{\rho})},$$

όπου με L συμβολίζεται η συνάρτηση πιθανοφάνειας που αντιστοιχεί στο παραπάνω δείγμα, $\hat{\alpha}_0, \hat{\sigma}_0^2$ είναι οι εκτιμητές μέγιστης πιθανοφάνειας υπό την H_0 και $\hat{\alpha}, \hat{\sigma}^2, \hat{\rho}$ είναι οι εκτιμητές μέγιστης πιθανοφάνειας που προκύπτουν ακολουθώντας τα βήματα που περιγράφονται στην παράγραφο της εκτίμησης. Τότε, γνωρίζουμε ότι για $n \to \infty$,

$$-2 \log \lambda \xrightarrow{d} \chi_1^2, \ \text{υπό την } H_0.$$

Απορρίπτουμε την H_0 όταν $-2\log\lambda \geq \chi^2_{1,\alpha}$, όπου α είναι το επιθυμητό επίπεδο σημαντικότητας.

Το Wald τεστ κατασκευάζεται με το συνήθη τρόπο, ως το τετράγωνο του λόγου του εκτιμητή μέγιστης πιθανοφάνειας της παραμέτρου ενδιαφέροντος ως προς την ασυμπτωτική τυπική του απόκλιση. Σύμφωνα με τα παραπάνω το Wald στατιστικό δίνεται από (βλ. Helu, 2007),

$$W_n = \frac{d_0}{2N(1-\rho_0)^2}(\hat{\rho}-\rho_0)^2,$$

όπου d_0 υπολογίζεται από τη σχέση για το d, που δίνεται από την (2.47) για $\rho = \rho_0$ και $\hat{\rho}$ είναι ο εκτιμητής μέγιστης πιθανοφάνειας της intraclass συσχέτισης ρ. Όταν $n \to \infty$,

$$W_n \xrightarrow{d} \chi_1^2, \ \text{υπό την } H_0.$$

Απορρίπτουμε την H_0 για μεγάλες τιμές του στατιστικού, δηλαδή όταν $W_n \geq \chi^2_{1,\alpha}$, όπου α είναι το επιθυμητό επίπεδο σημαντικότητας.

Τέλος θα κατασκευαστεί και το Score τεστ. Έστω $S(\alpha,\sigma^2,\rho) = \frac{\partial}{\partial\rho}\log L(\alpha,\sigma^2,\rho)$. Πιο συγκεκριμένα,

$$\begin{aligned} S(\alpha,\sigma^2,\rho) = \frac{\partial}{\partial\rho}\log L(\alpha,\sigma^2,\rho) &= \frac{\rho}{2(1-\rho)}\sum_{i=1}^{n}\frac{k_i(k_i-1)}{1+(k_i-1)\rho} \\ &- \frac{1}{2\sigma^2(1-\rho)^2}\sum_{i=1}^{n}(X_i-\alpha 1_{k_i})'\left[I_{k_i} - \frac{1+(k_i-\rho)^2}{[1+(k_i-\rho)]^2}J_{k_i}\right](X_i-\alpha 1_{k_i}). \end{aligned}$$

Ακόμη, θεωρούμε $S(\hat{\alpha}_0,\hat{\sigma}_0^2,\rho_0)$ και $I_n(\hat{\alpha}_0,\hat{\sigma}_0^2,\rho_0)$, όπου $\hat{\alpha}_0$ και $\hat{\sigma}_0^2$ είναι οι εκτιμητές μέγιστης πιθανοφάνειας των παραμέτρων α και σ^2 αντίστοιχα υπό την H_0 και $I_n(\alpha,\sigma^2,\rho)$ είναι ο πίνακας πηροφορίας του Fisher όπως ορίστηκε παραπάνω. Τότε το Score στατιστικό ορίζεται ως (βλ. Helu, 2007)

$$S_n = \left(\frac{1}{2(1-\rho_0)^2}\sum_{i=1}^{n}k_i(k_i-1)V_{i0}W_{i0}^{-2}\right)^{-1}\left(S(\hat{\alpha}_0,\hat{\sigma}_0^2,\rho_0)\right)^2,$$

όπου $W_{i0} = 1+(k_i-1)\rho_0$ και $V_{i0} = 1+(k_i-1)\rho_0^2$ για $i = 1,\ldots,n$. Όταν $n \to \infty$,

$$S_n \xrightarrow{d} \chi_1^2, \ \text{υπό την } H_0.$$

Απορρίπτουμε την H_0 για μεγάλες τιμές του στατιστικού, δηλαδή όταν $S_n \geq \chi^2_{1,\alpha}$, όπου α είναι το επιθυμητό επίπεδο σημαντικότητας.

2.4 Εναλλακτικοί εκτιμητές

Ο Srivastava (1984) πρότεινε εναλλακτικούς εκτιμητές, αντί των εκτιμητών μέγιστης πιθανοφάνειας. Έστω ότι το μοντέλο (2.1) ισχύει. Θεωρούμε έναν $k_i \times k_i$ πίνακα Γ_i για τον οποίο ισχύει ότι $\Gamma_i' = (\frac{1}{k_i}1_{k_i}, C_i')$, $i = 1, 2, \ldots, n$, όπου ο πίνακας C_i είναι τάξης $(k_i - 1) \times k_i$ και είναι τέτοιος ώστε να ισχύει $C_i 1_{k_i} = 0$ και $C_i C_i' = I_{k_i - 1}$. Τότε, θεωρούμε τον εξής μετασχηματισμό,

$$Y_i = \Gamma_i X_i, \;\; i = 1, 2, \ldots, n. \tag{2.50}$$

Επειδή, ο Γ_i είναι ένας γνωστός αντιστρέψιμος πίνακας το να παρατηρήσουμε το X_i είναι ισοδύναμο με το να παρατηρήσουμε το Y_i και οποιοσδήποτε εκτιμητής που βασίζεται στα X_i είναι ισοδύναμος με αυτόν που βασίζεται στα Y_i. Τα X_i είναι ανεξάρτητα, επομένως και τα Y_i θα είναι ανεξάρτητα. Ακόμη, αφού $X_i \sim N_{k_i}(\mu_i, \Sigma_i)$, (βλ. μοντέλο (2.1)), προκύπτει ότι,

$$Y_i \sim N_{k_i}(\Gamma_i \mu_i, \Gamma_i \Sigma_i \Gamma_i'), \quad i = 1, 2, \ldots, n, \tag{2.51}$$

όπου

$$\Gamma_i \mu_i = \begin{pmatrix} \alpha \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{k_i \times 1} \quad \text{και } \Gamma_i \Sigma_i \Gamma_i' = \begin{pmatrix} \sigma^2[1 + (k_i - 1)\rho]\frac{1}{k_i} & 0 & \cdots & 0 \\ 0 & \sigma^2(1-\rho) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2(1-\rho) \end{pmatrix}_{k_i \times k_i}$$

Θέτουμε

$$\eta_i^2 = \sigma^2 \frac{1}{k_i}[1 + (k_i - 1)\rho], \quad i = 1, 2, \ldots, n, \tag{2.52}$$

και

$$\gamma^2 = \sigma^2(1 - \rho). \tag{2.53}$$

Δηλαδή Y_i είναι ανεξάρτητες κανονικές τυχαίες μεταβλητές. Πιο συγκεκριμένα,

$$Y_{i1} \sim N(\alpha, \eta_i^2), \quad i = 1, 2, \dots, n, \tag{2.54}$$

$$Y_{ij} \sim N(0, \gamma^2), \quad i = 1, 2, \dots, n, \quad j = 2, \dots, k_i. \tag{2.55}$$

Μπορούμε να εκφράσουμε τα Y_i σε όρους των αρχικών μεταβλητών X_i. Ισχύει η παρακάτω σχέση,

$$Y_{i1} = \frac{1}{k_i} \sum_{j=1}^{k_i} X_{ij}, \quad i = 1, 2, \dots, n.$$

Παρόμοια οι μεταβλητές $Y_{i2}, \dots, Y_{ik_i}$ μπορούν να γραφούν σε όρους των μεταβλητών $X_{i1}, \dots, X_{ik_i}$. Ωστόσο αυτό εξαρτάται από την επιλογή του πίνακα C_i, αλλά επειδή ισχύει,

$$C_i' C_i = I_{k_i} - \frac{1}{k_i} 1_{k_i} 1_{k_i}',$$

έχουμε ότι,

$$\begin{aligned}
\sum_{j=2}^{k_i} Y_{ij}^2 &= (X_{i1}, \dots, X_{ik_i}) C_i' C_i (X_{i1}, \dots, X_{ik_i})' \\
&= \sum_{j=1}^{k_i} X_{ij}^2 - \frac{1}{k_i} \left(\sum_{j=1}^{k_i} X_{ij} \right)^2,
\end{aligned}$$

με την παραπάνω σχέση να μην εξαρτάται από την επιλογή του πίνακα C_i. Παραπάνω θέσαμε $\gamma^2 = \sigma^2(1 - \rho)$. Επομένως,

$$\rho = 1 - \frac{\gamma^2}{\sigma^2}. \tag{2.56}$$

O Srivastava (1984) χρησιμοποιώντας την παραπάνω σχέση προτείνει εναλλακτικούς εκτιμητές που είναι ευκολότεροι στον υπολογισμό, σε σχέση με τους εκτιμητές μέγιστης πιθανοφάνειας. Ωστόσο, ο ένας από αυτούς είναι ανεξάρτητος από την επιλογή του πίνακα C_i, ενώ ο άλλος έχει το πλεονέκτημα ότι έχει καλές ιδιότητες, αλλά εξαρτάται από την επιλογή του πίνακα C_i. O Srivastava (1984) προτείνει και για τις δύο παραπάνω περιπτώσεις ως εκτιμητή της παραμέτρου γ^2 τον,

$$\hat{\gamma}^2 = \sum_{i=1}^{n}\sum_{j=2}^{k_i} Y_{ij}^2 \Bigg/ \sum_{i=1}^{n}(k_i - 1), \tag{2.57}$$

ο οποίος είναι αμερόληπτος και συνεπής εκτιμητής της παραμέτρου γ^2. Η επιλογή του εκτιμητή αυτού μπορεί να αιτιολογηθεί διότι,

$$Y_{ij} \sim N(0, \gamma^2), i = 1, 2 \ldots, n, j = 2, \ldots, k_i.$$

Λόγω ανεξαρτησίας ισχύει ότι,

$$\sum_{j=2}^{k_i} \frac{Y_{ij}^2}{\gamma^2} \sim \chi^2_{(k_i - 1)}.$$

Αν επικαλεστούμε ακόμη μια φορά την ανεξαρτησία των μεταβλητών προκύπτει το συμπέρασμα,

$$\sum_{i=1}^{n}\sum_{j=2}^{k_i} \frac{Y_{ij}^2}{\gamma^2} \sim \chi^2_{\sum_{i=1}^{n}(k_i - 1)}.$$

Επομένως,

$$E\left(\frac{1}{\gamma^2}\sum_{i=1}^{n}\sum_{j=2}^{k_i} Y_{ij}^2\right) = \sum_{i=1}^{n}(k_i - 1)$$

και έτσι,

$$\gamma^2 = E\left(\frac{\sum_{i=1}^{n}\sum_{j=2}^{k_i} Y_{ij}^2}{\sum_{i=1}^{n}(k_i - 1)}\right).$$

Ακόμη, παρατηρούμε ότι $\hat{\gamma}^2$ ορίζεται ως η δειγματική διακύμανση των παρατηρήσεων $Y_{ij}, i = 1, \ldots, n, j = 2, \ldots, k_i$, των οποίων η διακύμανση ισούται με γ^2. Επομένως, $\hat{\gamma}^2$ είναι συνεπής εκτιμητής της παραμέτρου γ^2.

2.4.1 Εκτιμητής του σ^2 ανεξάρτητος του C_i

Ο Srivastava (1984) προτείνει ως εκτιμητή του σ^2 τον,

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_{i1}-\bar{Y}_1)^2 + \frac{1}{n}\hat{\gamma}^2\sum_{i=1}^{n}(1-k_i^{-1}), \tag{2.58}$$

όπου $\bar{Y}_1 = \frac{1}{n}\sum_{i=1}^{n} Y_{i1}$ και $\hat{\gamma}^2$ δίνεται από την (2.57). Γνωρίζουμε ότι ισχύει $\sigma^2 = \eta_i^2 + \alpha_i\gamma^2$, όπου $\alpha_i = 1 - k_i^{-1}$. Επομένως,

$$n\sigma^2 = \sum_{i=1}^{n}\eta_i^2 + \gamma^2\sum_{i=1}^{n}\alpha_i$$

ή

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\eta_i^2 + \frac{1}{n}\gamma^2\sum_{i=1}^{n}\alpha_i.$$

Εφόσον γ^2 έχει εκτιμηθεί χρησιμοποιώντας τις τυχαίες μεταβλητές $Y_{i2},\ldots,Y_{ik_i}, i = 1,2\ldots,n$, είναι φυσιολογικό να σκεφτούμε τις τυχαίες μεταβλητές $Y_{i1}, i = 1,2,\ldots,n$. Έχουμε ότι,

$$E\left(\sum_{i=1}^{n}Y_{i1}^2 - n\bar{Y}_1^2\right) = (n-1)\sigma^2 - \gamma^2\left(\sum_{i=1}^{n}\alpha_i\right)\left(1-\frac{1}{n}\right)$$

ή

$$\sigma^2 = E\left((n-1)^{-1}\left(\sum_{i=1}^{n}Y_{i1}^2 - n\bar{Y}_1^2\right) + \frac{1}{n}\gamma^2\sum_{i=1}^{n}\alpha_i\right).$$

Επομένως, θεωρούμε ως εκτιμητή του σ^2 τον $\hat{\sigma}^2$ που ορίζεται από τη σχέση,

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_{i1}-\bar{Y}_1)^2 + \frac{1}{n}\hat{\gamma}^2\sum_{i=1}^{n}(1-k_i^{-1}).$$

Επομένως από τη σχέση (2.56) ο εναλλακτικός εκτιμητής της παραμέτρου ρ που είναι ανεξάρτητος από την επιλογή του πίνακα C_i είναι ο,

$$\hat{\rho} = 1 - \frac{\hat{\gamma}^2}{\hat{\sigma}^2}, \tag{2.59}$$

όπου $\hat{\gamma}^2$ και $\hat{\sigma}^2$ ορίζονται από (2.57) και (2.58) αντίστοιχα.

2.4.2 Εκτιμητής του σ^2 που εξαρτάται από την επιλογή του C_i

Ακολουθώντας τον Srivastava (1984), έστω,

$$\widetilde{Y_{i1}} = Y_{i1} - k_i^{-1/2} \sum_{j=2}^{k_i} Y_{ij}, i = 1, 2, \ldots, n.$$

Ο εκτιμητής του σ^2 που προτείνεται έστω $\hat{\sigma}_\alpha^2$, δίνεται από τη σχέση,

$$\hat{\sigma}_\alpha^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\widetilde{Y_{i1}} - \widetilde{Y_{.1}})^2, \tag{2.60}$$

όπου $\widetilde{Y_{.1}} = \frac{1}{n} \sum_{i=1}^{n} \widetilde{Y_{i1}}$. Ο εκτιμητής του γ^2 μπορεί να παραμείνει ο ίδιος με αυτόν της σχέσης (2.57), ωστόσο μπορεί να εκτιμηθεί και από,

$$\hat{\gamma}_\alpha^2 = \sum_{i=1}^{n} \sum_{j=2}^{k_i} (Y_{ij} - \bar{Y}_{i.})^2 \Big/ \sum_{i=1}^{n} (k_i - 2), \tag{2.61}$$

όπου $\bar{Y}_{i.} = \frac{1}{k_i - 1} \sum_{j=2}^{k_i} Y_{ij}$. Έστω ένας ορθογώνιος μετασχηματισμός τύπου Helmert (βλ. Lancaster, 1965) εφαρμόζεται στις μεταβλητές $Y_{i2}, \ldots, Y_{ik_i}$. Τότε θα προκύψουν ανεξάρτητες $N(0, \gamma^2)$ τυχαίες μεταβλητές $Z_{i2}, \ldots, Z_{ik_i}$, όπου,

$$Z_{i2} = (k_i - 1)^{-1/2} \sum_{j=2}^{k_i} Y_{ij},$$

$$\sum_{j=3}^{k_i} Z_{ij}^2 = \sum_{j=2}^{k_i} Y_{ij}^2 - Z_{i2}^2 = \sum_{j=2}^{k_i} (Y_{ij} - \bar{Y}_{i.})^2.$$

Επομένως,

$$\gamma^{-2} \sum_{j=2}^{k_i} (Y_{ij} - \bar{Y}_{i.})^2 \sim \chi^2_{(k_i - 2)}.$$

Λόγω ανεξαρτησίας των μεταβλητών ισχύει ότι,

$$\gamma^{-2} \sum_{i=1}^{n} \sum_{j=2}^{k_i} (Y_{ij} - \bar{Y}_{i.})^2 \sim \chi^2_{\sum_{i=1}^{n} (k_i - 2)}.$$

Επομένως,

$$E\left(\gamma^{-2}\sum_{i=1}^{n}\sum_{j=2}^{k_i}(Y_{ij}-\bar{Y}_{i.})^2\right)=\sum_{i=1}^{n}(k_i-2)$$

ή

$$\gamma^2=E\left(\frac{\sum_{i=1}^{n}\sum_{j=2}^{k_i}(Y_{ij}-\bar{Y}_{i.})^2}{\sum_{i=1}^{n}(k_i-2)}\right).$$

Επομένως, ένας αμερόληπτος εκτιμητής της παραμέτρου γ^2 είναι ο $\hat{\gamma}_\alpha^2$ που δίνεται από τη σχέση,

$$\hat{\gamma}_\alpha^2=\frac{\sum_{i=1}^{n}\sum_{j=2}^{k_i}(Y_{ij}-\bar{Y}_{i.})^2}{\sum_{i=1}^{n}(k_i-2)}.$$

Ακόμη θεωρώντας το μετασχηματισμό,

$$\begin{pmatrix}\widetilde{Y_{i1}}\\ Z_{i2}\end{pmatrix}=\begin{pmatrix}1 & -\alpha_i^{1/2}\\ 0 & 1\end{pmatrix}\begin{pmatrix}Y_{i1}\\ Z_{i2}\end{pmatrix},$$

θα ισχύει ότι,

$$\begin{pmatrix}\widetilde{Y_{i1}}\\ Z_{i2}\end{pmatrix}\sim N_2\left(\begin{pmatrix}\alpha\\ 0\end{pmatrix},\begin{pmatrix}\sigma^2 & -\alpha_i\gamma^2\\ -\alpha_i\gamma^2 & \gamma^2\end{pmatrix}\right).$$

Παρατηρούμε ότι $E(\widetilde{Y_{i1}})=\alpha$ και $Var(\widetilde{Y_{i1}})=\sigma^2$, επομένως παίρνουμε ως εκτιμητές των παραμέτρων α και σ^2, έστω $\hat{\alpha}$ και $\hat{\sigma}_\alpha^2$ αντίστοιχα, τους εκτιμητές που δίνονται από τις σχέσεις,

$$\hat{\alpha}=\frac{1}{n}\sum_{i=1}^{n}\widetilde{Y_{i1}}, \tag{2.62}$$

$$\hat{\sigma}_\alpha^2=\frac{1}{n-1}\sum_{i=1}^{n}(\widetilde{Y_{i1}}-\widetilde{Y_{.1}})^2. \tag{2.63}$$

Επομένως, ένας εναλλακτικός εκτιμητής που προκύπτει από τη σχέση (2.56) είναι,

$$\hat{\rho}_\alpha=1-\frac{\hat{\gamma}_\alpha^2}{\hat{\sigma}_\alpha^2}, \tag{2.64}$$

όπου $\hat{\gamma}_\alpha^2$ και $\hat{\sigma}_\alpha^2$ ορίζονται από (2.61) και (2.63) αντίστοιχα.

2.4.3 Ασυμπτωτική κατανομή

Οι Srivastava and Katapa (1986) υπολόγισαν τις ασυμπτωτικές κατανομές των εναλλακτικών εκτιμητών που πρότεινε ο Srivastava (1984),

$$\hat{\rho} = 1 - \frac{\hat{\gamma}^2}{\hat{\sigma}^2} \quad \text{και} \quad \hat{\rho}_\alpha = 1 - \frac{\hat{\gamma}_\alpha^2}{\hat{\sigma}_\alpha^2}.$$

Σύμφωνα με την εργασία αυτή, χρησιμοποιώντας το Κεντρικό Οριακό Θεώρημα ισχύει ότι, καθώς $n \to \infty$,

$$\sqrt{n}(\hat{\gamma}^2 - \gamma^2, \hat{\sigma}^2 - \sigma^2)' \xrightarrow{d} N(0, \Sigma_2),$$

όπου

$$\Sigma_2 = 2 \begin{pmatrix} \gamma^4(\bar{k}-1)^{-1} & \gamma^4(\bar{k}-1)^{-1}n^{-1}\sum_{i=1}^n \alpha_i \\ \gamma^4(\bar{k}-1)^{-1}n^{-1}\sum_{i=1}^n \alpha_i & c^2\sigma^4 \end{pmatrix},$$

$$c^2 = 1 - 2(1-\rho)n^{-1}\sum_{i=1}^n \alpha_i + (1-\rho)^2\left[\frac{1}{n}\sum_{i=1}^n \alpha_i^2 + (\bar{k}-1)^{-1}(n^{-1}\sum_{i=1}^n \alpha_i)^2\right],$$

$\bar{k} = \frac{1}{n}\sum_{i=1}^n k_i$ και $\alpha_i = (1 - k_i^{-1})$.
Εφαρμόζοντας τη δέλτα μέθοδο έχουμε ότι, καθώς $n \to \infty$,

$$\sqrt{n}(\hat{\rho} - \rho) \xrightarrow{d} N(0, V_2),$$

όπου

$$V_2 = 2(1-\rho)^2\left[(\bar{k}-1)^{-1} + c^2 - 2(1-\rho)(\bar{k}-1)^{-1}n^{-1}\sum_{i=1}^n \alpha_i\right]. \tag{2.65}$$

Ακόμη, κάνοντας για ακόμη μια φορά χρήση του Κεντρικού Οριακού Θεωρήματος έχουμε, για $n \to \infty$,

$$\sqrt{n}(\hat{\gamma}_\alpha^2 - \gamma^2, \hat{\sigma}_\alpha^2 - \sigma^2) \xrightarrow{d} N(0, \Sigma_3),$$

όπου

$$\Sigma_3 = 2\begin{pmatrix} \gamma^4(\bar{k}-2)^{-1} & 0 \\ 0 & \sigma^4 \end{pmatrix}.$$

Εφαρμόζοντας τη δέλτα μέθοδο προκύπτει το συμπέρασμα ότι, για $n \to \infty$:

$$\sqrt{n}(\hat{\rho}_\alpha - \rho) \xrightarrow{d} N(0, V_3),$$

όπου

$$V_3 = 2(1-\rho)^2[(\bar{k}-2)^{-1} + 1]. \tag{2.66}$$

2.4.4 Διαστήματα εμπιστοσύνης και Στατιστικά Τεστ

Ισχύει ότι $\widetilde{Y_{i1}} \sim N(\alpha, \sigma^2)$, $i = 1, 2, \ldots, n$. Λόγω της κανονικότητας των τυχαίων μεταβλητών έχουμε ότι,

$$\frac{\sum_{i=1}^{n}(\widetilde{Y_{i1}} - \widetilde{Y_{.1}})^2}{\sigma^2} \sim \chi^2_{n-1}$$

και

$$\gamma^{-2}\sum_{i=1}^{n}\sum_{j=2}^{k_i}(Y_{ij} - \bar{Y}_{i.})^2 \sim \chi^2_{\sum_{i=1}^{n}(k_i-2)}.$$

Τα παραπάνω στατιστικά είναι ανεξάρτητα επομένως (βλ. Srivastava, 1984),

$$F = \frac{(1-\rho)^{-1}(n-1)\sum_{i=1}^{n}\sum_{j=2}^{k_i}(Y_{ij} - \bar{Y}_{i.})^2}{\sum_{i=1}^{n}(k_i-2)\sum_{i=1}^{n}(\widetilde{Y_{i1}} - \widetilde{Y_{.1}})^2} \sim F_{\sum_{i=1}^{n}(k_i-2), n-1}. \tag{2.67}$$

Το στατιστικό F μπορεί να χρησιμοποιηθεί για την κατασκευή διαστημάτων εμπιστοσύνης, καθώς και την κατασκευή τεστ για τον έλεγχο στατιστικών υποθέσεων.

Κεφάλαιο 3

ΠΕΡΙΣΣΟΤΕΡΑ ΤΟΥ ΕΝΟΣ ΔΕΙΓΜΑΤΑ

3.1 Εισαγωγή

Στο προηγούμενο κεφάλαιο παρουσιάστηκαν αποτελέσματα σχετικά με την εκτίμηση και τους ελέγχους υποθέσεων στις παραμέτρους του intraclass μοντέλου (2.1) και ειδικότερα σε ότι αφορά την παράμετρο της intraclass συσχέτισης ρ. Τα αποτελέσματα που παρουσιάστηκαν βασίστηκαν σε ένα τυχαίο δείγμα από έναν πολυδιάστατο κανονικό πληθυσμό. Στο κεφάλαιο αυτό θα γίνει μια ανασκόπηση της βιβλιογραφίας που υπάρχει γύρω από την παράμετρο της intracass συσχέτισης στην περίπτωση που έχουμε περισσότερα του ενός δείγματα. Παρακάτω, περιγράφεται το γενικό μοντέλο του προβλήματος.

Έστω ότι υπάρχουν g ανεξάρτητοι πληθυσμοί και δεδομένα, που αφορούν τα παιδιά οικογενειών που επιλέχθηκαν τυχαία από τους διαθέσιμους πληθυσμούς. Υποθέτουμε ότι το τυχαίο δείγμα του i-οστού πληθυσμού αποτελείται από n_i οικογένειες και ο αριθμός των παιδιών στις οικογένειες μπορεί να ποικίλλει. Συμβολίζουμε με k_{ij} τον αριθμό των παιδιών της j-οστής οικογένειας από τον i-οστό πληθυσμό. Υποθέτουμε ότι X_{ijm}, $m = 1, 2, \ldots, k_{ij}$, $j = 1, 2, \ldots, n_i$, $i = 1, 2, \ldots, g$ παριστάνουν τις παρατηρήσεις του m-οστού παιδιού της j-οστής οικογένειας που ανήκει στον i-οστό πληθυσμό. Υποθέτουμε για τον i-οστό πληθυσμό ότι, $E(X_{ijm}) = \alpha_i$, $Var(X_{ijm}) = \sigma_i^2$ και η intraclass συσχέτιση $Corr(X_{ijm}, X_{ijm'})$ $= \rho_i$ για $m \neq m'$. Για κάθε i, ισχύει ότι, $\alpha_i \in \mathbb{R}$, $\sigma_i^2 > 0$ και $-\frac{1}{\max\limits_{1 \leq j \leq n_i}(k_{ij}-1)} < \rho_i < 1$. Έστω ότι το διάνυσμα των παρατηρήσεων της j-οστής οικογένειας του i-οστού πληθυσμού είναι το $X_{ij} = (X_{ij1}, X_{ij2}, \ldots, X_{ijk_{ij}})'$. Τότε $E(X_{ij}) = \mu_{ij} = \alpha_i 1_{k_{ij}}$, όπου γενικά 1_k είναι ένα $k \times 1$ διάνυσμα με όλα του τα στοιχεία ίσα με τη μονάδα και ο πίνακας διακυμάνσεων-συνδιακυμάνσεων του X_{ij} είναι,

$$Var(X_{ij}) = \Sigma_{ij} = \sigma_i^2[(1-\rho_i)I_{k_{ij}} + \rho_i J_{k_{ij}}] = \sigma_i^2 V_{ij}(\rho_i),$$

όπου γενικά I_k είναι ο ταυτοτικός πίνακας τάξης k, J_k είναι ένας $k \times k$ πίνακας με όλα του τα στοιχεία ίσα με τη μονάδα και $V_{ij}(\rho_i)$ ορίζεται να είναι ο πίνακας $[(1-\rho_i)I_{k_{ij}} + \rho_i J_{k_{ij}}]$. Ακόμη, υποθέτουμε ότι κάθε οικογένεια ακολουθεί μια πολυδιάστατη κανονική κατανομή. Πιο συγκεκριμένα, υποθέτουμε ότι,

$$X_{ij} \sim N_{k_{ij}}(\mu_{ij}, \Sigma_{ij}), j = 1, \ldots, n_i, i = 1, \ldots, g. \tag{3.1}$$

Με την υπόθεση του μοντέλου (3.1) θα παρουσιαστούν στατιστικά τεστ, που αφορούν την intraclass παράμετρο ρ_i, υιοθετώντας ειδικές περιπτώσεις του παραπάνω μοντέλου ή υιοθετώντας το στην πιο γενική μορφή του.

3.2 Κατασκευή στατιστικών τεστ

Οι Donner and Bull (1983) υπολόγισαν τον εκτιμητή μέγιστης πιθανοφάνειας του κοινού intraclass συντελεστή συσχέτισης ρ, που βασίζεται σε δύο ανεξάρτητα τυχαία δείγματα που προέρχονται από δύο πολυδιάστατους ανεξάρτητους κανονικούς πληθυσμούς και το τεστ πηλίκου πιθανοφανειών για τον έλεγχο της υπόθεσης κοινού συντελεστή συσχέτισης ρ. Το μοντέλο που υποθέτουμε ότι ι-σχύει είναι το (3.1) για δύο ανεξάρτητους πληθυσμούς, $g = 2$. Ακόμη, το πλήθος των παιδιών σε κάθε οικογένεια του i-οστού πληθυσμού δεν ποικίλλει, δηλαδή θεωρούμε ότι $k_{ij} = k_i$, $\mu_{ij} = \mu_i$ και $\Sigma_{ij} = \Sigma_i, i = 1, 2, j = 1, \ldots, n_i$. Έστω ότι έχουμε δύο ανεξάρτητα τυχαία δείγματα από κάθε πληθυσμό. Στόχος είναι να εκτιμηθεί η κοινή intraclass συσχέτιση ρ ως προς το συνδυασμένο δείγμα, $W = (X_{11}, X_{12}, \ldots, X_{1n_1}, X_{21}, X_{22}, \ldots, X_{2n_2})$. Για τη συνάρτηση πιθανοφάνειας L που αντιστοιχεί στο παραπάνω δείγμα ισχύει η παρακάτω σχέση,

$$\begin{aligned} -2\log L &= (n_1k_1 + n_2k_2)\log(2\pi) + n_1 \log|\Sigma_1| + n_2 \log|\Sigma_2| \\ &\quad + \sum_{i=1}^{n_1}(X_{1i} - \mu_1)'\Sigma_1^{-1}(X_{1i} - \mu_1) + \sum_{i=1}^{n_2}(X_{2i} - \mu_2)'\Sigma_2^{-1}(X_{2i} - \mu_2), \end{aligned}$$

όπου (βλ. Mardia et al., 1979, σ. 97),

$$\log|\Sigma_1| = k_1 \log\sigma_1^2 + (k_1 - 1)\log(1 - \rho) + \log[1 + (k_1 - 1)\rho]$$

και

$$\Sigma_1^{-1} = \frac{1}{\sigma_1^2(1-\rho)}[I_{k_1} - \rho[1 + (k_1 - 1)\rho]^{-1}J_{k_1}].$$

Οι ανάλογες σχέσεις ικανοποιούνται και για $\log|\Sigma_2|$ και Σ_2^{-1}. Να τονιστεί ότι στις παραπάνω σχέσεις έχουμε υποθέσει ότι $\rho_1 = \rho_2 = \rho$.

Αντικαθιστώντας στη σχέση της $-2\log L$ και παραγωγίζοντας ως προς α_1 και α_2 προκύπτουν οι εκτιμητές $\hat{\alpha}_1$ και $\hat{\alpha}_2$ αντίστοιχα, που δίνονται από τις σχέσεις,

$$\hat{\alpha}_1 = \frac{\sum_{i=1}^{n_1}\sum_{j=1}^{k_1} X_{1ij}}{k_1n_1}, \tag{3.2}$$

$$\hat{\alpha}_2 = \frac{\sum_{i=1}^{n_2}\sum_{j=1}^{k_2} X_{2ij}}{k_2n_2}. \tag{3.3}$$

Αντικαθιστώντας τις παραπάνω σχέσεις στη σχέση της $-2\log L$ και παραγωγίζοντας ως προς σ_1^2, σ_2^2 και ρ καταλήγουμε στις σχέσεις,

$$\hat{\sigma}_1^2 = S_1^2 \left[1 - \frac{(k_1-1)\hat{\rho}_M(r_1-\hat{\rho}_M)}{\{1+(k_1-1)\hat{\rho}_M\}(1-\hat{\rho}_M)}\right], \tag{3.4}$$

$$\hat{\sigma}_2^2 = S_2^2 \left[1 - \frac{(k_2-1)\hat{\rho}_M(r_2-\hat{\rho}_M)}{\{1+(k_2-1)\hat{\rho}_M\}(1-\hat{\rho}_M)}\right] \tag{3.5}$$

και

$$\hat{\rho}_M = \frac{n_1k_1(k_1-1)W_1r_1 + n_2k_2(k_2-1)W_2r_2}{n_1k_1(k_1-1)W_1 + n_2k_2(k_2-1)W_2}, \tag{3.6}$$

όπου,

$$S_1^2 = \frac{\sum_{i=1}^{n_1}\sum_{j=1}^{k_1}(X_{1ij}-\hat{\alpha}_1)^2}{n_1k_1}, \quad S_2^2 = \frac{\sum_{i=1}^{n_2}\sum_{j=1}^{k_2}(X_{2ij}-\hat{\alpha}_2)^2}{n_2k_2},$$

$$r_1 = \sum_{i=1}^{n_1}\sum_{\substack{j=1\\ j\neq l}}^{k_1}\sum_{l=1}^{k_1}\left[(X_{1ij}-\hat{\alpha}_1)(X_{1il}-\hat{\alpha}_1)/\{n_1k_1(k_1-1)S_1^2\}\right],$$

$$r_2 = \sum_{i=1}^{n_2}\sum_{\substack{j=1\\ j\neq l}}^{k_2}\sum_{l=1}^{k_2}\left[(X_{2ij}-\hat{\alpha}_2)(X_{2il}-\hat{\alpha}_2)/\{n_2k_2(k_2-1)S_2^2\}\right],$$

$$W_1 = \frac{\hat{\sigma}_1^2 S_1^2}{\left\{n_1k_1(S_1^2-\hat{\sigma}_1^2)\right\}^2} \text{ και } W_2 = \frac{\hat{\sigma}_1^2 S_2^2}{\left\{n_2k_2(S_2^2-\hat{\sigma}_2^2)\right\}^2}.$$

Οι Donner and Bull (1983) προτείνουν ότι ο εκτιμητής $\hat{\rho}_M$ μπορεί να αποκτηθεί αριθμητικά, πιο εύκολα, εάν αντικατασταθούν οι εκφράσεις για $\hat{\alpha}_i$ και $\hat{\sigma}_i^2, i = 1, 2$, στη σχέση όπου δίνεται η $-2\log L$ και ελαχιστοποιηθεί αριθμητικά ως προς ρ το αποτέλεσμα.

Στη συνέχεια οι Donner and Bull (1983) κατασκεύασαν το τεστ πηλίκου πιθανοφανειών για τον έλεγχο της υπόθεσης κοινού συντελεστή συσχέτισης ρ, βασισμένοι στο συνδυασμένο δείγμα W. Το τεστ κατασκευάζεται για τον έλεγχο της παρακάτω υπόθεσης,

$$H_0 : \rho_1 = \rho_2 = \rho \quad \text{έναντι} \quad H_1 : \rho_1 \neq \rho_2.$$

Ακολουθώντας τους Donner and Bull (1983), έστω L_0 είναι η συνάρτηση πιθανοφάνειας για το συνδυασμένο δείγμα $W = (X_{11}, X_{12}, \ldots, X_{1n_1}, X_{21}, X_{22}, \ldots, X_{2n_2})$ υπό την H_0 και L_1 είναι η συνάρτηση πιθανοφάνειας του συνδυασμένου δείγματος υπό την εναλλακτική υπόθεση H_1. Τα αποτελέσματα διατυπώνονται στο επόμενο θεώρημα (βλ. Donner and Bull, 1983).

Θεώρημα 3.2.1. *Έστω $X_{11}, \ldots, X_{1n_1}$ και $X_{21}, \ldots, X_{2n_2}$ ανεξάρτητα τυχαία δείγματα από δύο ανεξάρτητους κανονικούς πληθυσμούς για τους οποίους ισχύουν οι υποθέσεις του μοντέλου (3.1). Για τον έλεγχο της υπόθεσης,*

$$H_0 : \rho_1 = \rho_2 = \rho \quad \textit{έναντι} \quad H_1 : \rho_1 \neq \rho_2,$$

το στατιστικό, που προκύπτει με τη μέθοδο του πηλίκου μέγιστων πιθανοφανειών ορίζεται ως,

$$L_D = -2\log\left(\frac{L_0}{L_1}\right) = \sum_{i=1}^{2} n_i \left[k_i \log\left(\frac{\hat{\sigma}_i^2}{S_i^2}\right) + (k_i - 1)\log\left(\frac{1-\hat{\rho}_M}{1-r_i}\right) + \log\left\{\frac{1+(k_i-1)\hat{\rho}_M}{1+(k_i-1)r_i}\right\}\right].$$

Για μεγάλα n_1 και n_2 $(n_1, n_2 \to \infty)$ έχουμε ότι,

$$L_D \xrightarrow[n_1, n_2 \to \infty]{d} \chi_1^2 \quad \textit{υπό την} \quad H_0.$$

Απορρίπτουμε την H_0 για μεγάλες τιμές του στατιστικού και πιο συγκεκριμένα απορρίπτουμε όταν $L_D \geq \chi^2_{1,\alpha}$, όπου α είναι το επιθυμητό επίπεδο σημαντικότητας.

Ο Khatri (1989) με τη σειρά του ασχολήθηκε με το πρόβλημα της κατασκευής στατιστικών τεστ για την ισότητα των intraclass συσχετίσεων για δύο πολυδιάστατους κανονικούς πληθυσμούς. Στο πλαίσιο αυτό, έστω ότι έχουμε δύο τυχαία δείγματα μεγέθους $n_i, i = 1, 2$, από δύο ανεξάρτητους κανονικούς πληθυσμούς για τους οποίους ισχύουν οι υποθέσεις του μοντέλου (3.1). Σε αντίθεση με την προηγούμενη περίπτωση, υποθέτουμε ότι κάθε οικογένεια, ανεξαρτήτως του πληθυσμού στον οποίο ανήκει, έχει τον ίδιο αριθμό παιδιών, δηλαδή $k_{ij} = k_i = k, i = 1, 2$. Αρχικά, θα κατασκευαστούν στατιστικά τεστ για τον έλεγχο των παρακάτω υποθέσεων,

$$H_{01} : \rho_1 = \rho_2 \quad \text{έναντι} \quad H_1 : \rho_1 > \rho_2,$$
$$H_{02} : \rho_1 = \rho_2 \quad \text{έναντι} \quad H_1 : \rho_1 < \rho_2.$$

Για την κατασκευή στατιστικών τεστ για τον έλεγχο των H_{01} και H_{02} θα αξιοποιηθούν τα παρακάτω χρήσιμα αποτελέσματα. Έστω τα ανεξάρτητα $X_1, X_2, \ldots, X_n$ από μια k-διάστατη κανονική κατανομή, έτσι ώστε $X_i \sim N_k(\alpha 1_k, \sigma^2[(1-\rho)I_k + \rho J_k])$ για $i = 1, \ldots, n$. Τότε έχουμε ότι (βλ. Rao, 1973, σ. 196),

$$\bar{X} = \frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k} X_{ij} \sim N\left(\alpha, \frac{\sigma^2}{nk}[1+(k-1)\rho]\right),$$
$$B = \frac{1}{k}\sum_{i=1}^{n}\left(\sum_{j=1}^{k} X_{ij}\right)^2 - \frac{1}{nk}\left(\sum_{i=1}^{n}\sum_{j=1}^{k} X_{ij}\right)^2 \sim \sigma^2(1+(k-1)\rho)\chi^2_{n-1},$$
$$W = \sum_{i=1}^{n}\sum_{j=1}^{k} X_{ij}^2 - \frac{1}{k}\sum_{i=1}^{n}\left(\sum_{j=1}^{k} X_{ij}\right)^2 \sim \sigma^2(1-\rho)\chi^2_{n(k-1)}.$$

Τα $\bar{X}, B, W$ είναι ανεξάρτητα και επαρκή στατιστικά για τις παραμέτρους α, σ^2 και ρ. Ακόμη, $\frac{B}{W} \sim \theta f$, όπου f είναι η κατανομή του λόγου δύο ανεξάρτητων χ^2 κατανομών και $\theta = (1+(k-1)\rho)/1-\rho$. Μια τυχαία μεταβλητή θα έχει την f με βαθμούς ελευθερίας $n-1$ και $n(k-1)$, κατανομή, εάν η συνάρτηση πυκνότητας πιθανότητας της δίνεται από,

$$\frac{\Gamma\left((nk-1)/2\right)}{\Gamma\left(n(k-1)/2\right)\Gamma\left((n-1)/2\right)} \cdot \frac{f^{(n-3)/2}}{(1+f)^{(nk-1)/2}},$$

και συμβολίζεται ως $f_{n-1,n(k-1)}$.

Έστω λοιπόν, ότι έχουμε στη διάθεσή μας δύο ανεξάρτητα τυχαία δείγματα μεγέθους $n_i, i = 1, 2$, από κάθε πληθυσμό και έχουμε υπολογίσει τα παραπάνω στατιστικά. Για κάθε έναν από τους πληθυσμούς ισχύει ότι,

$$f_i = \frac{B_i}{W_i} \sim \theta_i f_{n_i - 1, n_i(k-1)}, \;\; i = 1, 2,$$

όπου $\theta_i = \frac{1+(k-1)\rho_i}{1-\rho_i}, i = 1, 2$. Υπό την μηδενική υπόθεση $H_0 : \rho_1 = \rho_2$ η στατιστική συνάρτηση,

$$R = \frac{\frac{B_1}{W_1}}{\frac{B_2}{W_2}},$$

έχει κατανομή την ίδια με την κατανομή του λόγου δύο f κατανομών. Η κατανομή του R είναι ανεξάρτητη από την κοινή τιμή του ρ, υπό τη μηδενική υπόθεση και μπορεί να χρησιμοποιηθεί για την κατασκευή τεστ για τον έλεγχο των H_{01} και H_{02}. Οι άνω και κάτω στατιστικά σημαντικές τιμές του στατιστικού R για επίπεδα σημαντικότητας $0.01, 0.05, 0.10$, δίνονται από κατάλληλους πίνακες (βλ. Khatri et al., 1989). Πιο συγκεκριμένα, για τον έλεγχο της υπόθεσης H_{01}, απορρίπτουμε την H_{01} εάν $R \geq R_u(n_1, n_2)$ και για τον έλεγχο της H_{02}, απορρίπτουμε την H_{02} εάν $R \leq R_l(n_1, n_2)$, όπου $R_l(n_1, n_2) = [R_u(n_2, n_1)]^{-1}$. Οι τιμές $R_u(n_1, n_2)$ και $R_l(n_1, n_2)$ υπάρχουν σε πίνακες που έχουν κατασκευαστεί για διάφορα επίπεδα σημαντικότητας, όπως αναφέρθηκε και παραπάνω.

Στη συνέχεια θα κατασκευαστεί το τεστ πηλίκου πιθανοφανειών για τον έλεγχο της υπόθεσης,

$$H_{03} : \rho_1 = \rho_2 \quad \text{έναντι} \quad H_1 : \rho_1 \neq \rho_2.$$

Στα παρακάτω ακολουθούμε την εργασία των Khatri et al. (1989). Στο πλαίσιο αυτό έχουμε τα επαρκή στατιστικά $\bar{X}_i, B_i$ και W_i για τις παραμέτρους $\alpha_i, \sigma_i, \rho_i$ για $i = 1, 2$ τα οποία είναι ανεξάρτητα μεταξύ τους και για τις κατανομές τους ισχύουν τα παρακάτω,

$$\begin{aligned}\bar{X}_i &\sim N(\alpha_i, \theta_i\phi_i/(n_ik)),\\ B_i &\sim \theta_i\phi_i\chi^2_{n_i-1},\\ W_i &\sim \phi_i\chi^2_{n_i(k-1)},\end{aligned}$$

όπου $\phi_i = \sigma_i^2(1-\rho_i)$ και $\theta_i = [1+(k-1)\rho_i]/(1-\rho_i)$. Γράφοντας την από κοινού συνάρτηση πιθανοφάνειας και μεγιστοποιώντας ως προς α_i, ϕ_i για δοθέν θ_i προκύπτουν οι βέλτιστες τιμές,

$$\hat{\alpha}_i = \bar{X}_i \quad \text{και} \quad \hat{\phi}_i = \frac{1}{n_ik}\left(\frac{B_i}{\theta_i} + W_i\right), i = 1, 2.$$

Η τιμή της log-likelihood συνάρτησης, εάν αντικαταστήσουμε τις παραπάνω βέλτιστες τιμές, χωρίς να λάβουμε υπόψιν τις σταθερές δίνεται από τη σχέση,

$$-\frac{1}{2}\sum_{i=1}^{2}\left[n_i\log\theta_i + n_ik\log\left(\frac{B_i}{\theta_i} + W_i\right)\right].$$

Παραγωγίζοντας την παραπάνω σχέση ως προς θ_i και εξισώνοντας με το 0, για να αποκτηθούν οι εκτιμητές μέγιστης πιθανοφάνειας, προκύπτει ότι,

$$\hat{\theta}_i = (k-1)f_i, \text{ με } f_i = \frac{B_i}{W_i}, \ \ i = 1, 2.$$

Οδηγούμαστε έτσι στο δίχως περιορισμούς (δηλ. υπό την H_1) supremum της loglikelihood συνάρτησης το οποίο είναι,

$$-\frac{1}{2}\sum_{i=1}^{2}\left[n_i(k-1)\log\hat{\theta}_i - n_ik\log W_i - n_ik\log(f_i+\hat{\theta}_i)\right]. \qquad (3.7)$$

Υπό την H_0, $\rho_1 = \rho_2$, οπότε έχουμε ότι, $\theta_1 = \theta_2 = \theta$. Έτσι ο εκτιμητής του θ θα αποκτηθεί παραγωγίζοντας τη παραπάνω σχέση ως προς θ και εξισώνοντας την παράγωγο με το 0. Προκύπτει η παρακάτω εξίσωση,

$$\frac{n_1f_1}{f_1+\theta} + \frac{n_2f_2}{f_2+\theta} = \frac{n_1+n_2}{k}.$$

Η παραπάνω εξίσωση είναι τετραγωνική ως προς θ και έχει μια θετική και μια αρνητική ρίζα. Ο εκτιμητής μέγιστης πιθανοφάνειας αντιστοιχεί στη θετική ρίζα, έστω $\hat{\theta}$. Το supremum της loglikelihood συνάρτησης δίνεται από την παρακάτω σχέση,

$$-\frac{1}{2}\sum_{i=1}^{2}\left[n_i(k-1)\log\hat{\theta} - n_i k\log W_i - n_i k\log(f_i+\hat{\theta})\right]. \tag{3.8}$$

Με βάση τις (3.7) και (3.8), το στατιστικό του λόγου μέγιστων πιθανοφανειών λ ή το loglikelihood στατιστικό $L_D = -2\log\lambda$ είναι,

$$L_D = (k-1)\sum_{i=1}^{2} n_i \log\left(\frac{\hat{\theta}_i}{\hat{\theta}}\right) - k\sum_{i=1}^{2} n_i\log\left((f_i+\hat{\theta}_i)/(f_i+\hat{\theta})\right).$$

Γνωρίζουμε ότι για μεγάλα μεγέθη δείγματος το στατιστικό L_D ακολουθεί προσεγγιστικά μια χ^2 κατανομή με έναν βαθμό ελευθερίας. Τα αποτελέσματα συνοψίζονται στο παρακάτω θεώρημα (βλ. Khatri et al., 1989).

Θεώρημα 3.2.2. *Έστω $X_{11},\ldots,X_{1n_1}$ και $X_{21},\ldots,X_{2n_2}$ ανεξάρτητα τυχαία δείγματα από δύο ανεξάρτητους k-διάστατους κανονικούς πληθυσμούς για τους οποίους ισχύουν οι υποθέσεις του μοντέλου (3.1). Για τον έλεγχο της υπόθεσης,*

$$H_{03}: \rho_1 = \rho_2 = \rho \quad \text{έναντι} \quad H_1: \rho_1 \neq \rho_2,$$

το στατιστικό του λόγου μέγιστων πιθανοφανειών λ ή το loglikelihood στατιστικό $L_D = -2\log\lambda$ ορίζεται ως,

$$L_D = (k-1)\sum_{i=1}^{2} n_i \log\left(\frac{\hat{\theta}_i}{\hat{\theta}}\right) - k\sum_{i=1}^{2} n_i\log\left((f_i+\hat{\theta}_i)/(f_i+\hat{\theta})\right).$$

Για μεγάλα n_1 και n_2 ($n_1, n_2 \to \infty$) έχουμε ότι,

$$L_D \xrightarrow[n_1,n_2\to\infty]{d} \chi^2_1 \quad \text{υπό την} \quad H_{03}.$$

Απορρίπτουμε την H_{03} για μεγάλες τιμές του στατιστικού και πιο συγκεκριμένα απορρίπτουμε όταν $L_D \geq \chi^2_{1,\alpha}$, όπου α είναι το επιθυμητό επίπεδο σημαντικότητας.

Οι Konishi and Gupta (1989) κατασκεύασαν ένα προσεγγιστικό τεστ πηλίκου πιθανοφανειών για τον έλεγχο της ισότητας των intraclass συντελεστών συσχέτισης, στη βάση ανεξάρτητων τυχαίων δειγμάτων μεγέθους $n_i, i = 1, \ldots, g$, από g ανεξάρτητους πολυδιάστατους κανονικούς πληθυσμούς. Έστω λοιπόν ότι οι υποθέσεις του μοντέλου (3.1) ισχύουν, όπου το μέγεθος των οικογενειών στον i-οστό πληθυσμό δε μεταβάλλεται. Δηλαδή έχουμε ότι $k_{ij} = k_i$, $\mu_{ij} = \mu_i$ και $\Sigma_{ij} = \Sigma_i$, για $i = 1, \ldots, g$ και $j = 1, \ldots, n_i$. Σκοπός είναι να εξετάσουμε την παρακάτω υπόθεση, $H_0 : \rho_1 = \rho_2 = \cdots = \rho_g$ έναντι της εναλλακτικής H_1 : υπάρχει τουλάχιστον ένα ζεύγος intraclass συντελεστών συσχέτισης που είναι διαφορετικοί μεταξύ τους. Μια πιθανή προσέγγιση στο παραπάνω πρόβλημα είναι η κατασκευή του τεστ πηλίκου πιθανοφανειών. Ωστόσο, γίνεται πολύ δύσκολη η μεγιστοποίηση της συνάρτησης πιθανοφάνειας για περισσότερα από δύο δείγματα, θεωρητικά και υπολογιστικά, υπό την υπόθεση του κοινού συντελεστή συσχέτισης. Έτσι, οι Konishi and Gupta (1989) προτείνουν ένα προσεγγιστικό τεστ πηλίκου πιθανοφανειών για τον έλεγχο της παραπάνω υπόθεσης και προσδιορίζουν την ασυμπτωτική κατανομή του υπό τη μηδενική και εναλλακτική υπόθεση όταν τα δείγματα προέρχονται από μη-κανονικούς πληθυσμούς. Επομένως, τα αποτελέσματα για τους κανονικούς πληθυσμούς προκύπτουν ως μια ειδική περίπτωση του παραπάνω γενικού αποτελέσματος. Για τη συνάρτηση πιθανοφάνειας που αντιστοιχεί στα g ανεξάρτητα τυχαία δείγματα ισχύει η παρακάτω σχέση,

$$
\begin{aligned}
-2\log L = & \left(\sum_{i=1}^{g} n_i k_i\right)\log(2\pi) \\
& + \sum_{i=1}^{g} n_i[k_i \log\sigma_i^2 + (k_i-1)\log(1-\rho_i) + \log[1+(k_i-1)\rho_i]] \\
& + \sum_{i=1}^{g} \frac{[1+(k_i-1)\rho_i]^{-1}}{\sigma_i^2(1-\rho_i)} \sum_{j=1}^{n_i} \Bigg\{ [1+(k_i-2)\rho_i] \sum_{m=1}^{k_i} (X_{ijm}-\alpha_i)^2
\end{aligned}
$$

$$-\rho_i \sum_{m\neq l}^{k_i} (X_{ijm} - \alpha_i)(X_{ijl} - \alpha_i) \Bigg\},$$

όπου έχουν ληφθεί υπόψιν οι σχέσεις (βλ. Mardia et al., 1979, σ. 97),

$$\log|\Sigma_i| = k_i \log\sigma_i^2 + (k_i - 1)\log(1 - \rho_i) + \log[1 + (k_i - 1)\rho_i]$$

και

$$\Sigma_i^{-1} = \frac{1}{\sigma_i^2(1 - \rho_i)}[I_{k_i} - \rho_i[1 + (k_i - 1)\rho_i]^{-1} J_{k_i}].$$

Υπό την H_1 οι εκτιμητές μέγιστης πιθανοφάνειας των παραμέτρων α_i, σ_i^2 και $\rho_i, i = 1, \ldots, g$, δίνονται αντίστοιχα από,

$$\begin{aligned}
\bar{X}_i &= \frac{1}{n_i k_i} \sum_{j=1}^{n_i} \sum_{m=1}^{k_i} X_{ijm}, \\
S_i^2 &= \frac{1}{n_i k_i} \sum_{j=1}^{n_i} \sum_{m=1}^{k_i} (X_{ijm} - \bar{X}_i)^2, \\
r_i &= \sum_{j=1}^{n_i} \sum_{\substack{m=1 \\ m\neq l}}^{k_i} (X_{ijm} - \bar{X}_i)(X_{ijl} - \bar{X}_i) / \{n_i k_i (k_i - 1) S_i^2\},
\end{aligned}$$

για $i = 1, \ldots, g$. Αντικαθιστώντας τους παραπάνω εκτιμητές στην προηγούμενη σχέση προκύπτει η τιμή της $-2\log L$ υπό την εναλλακτική υπόθεση H_1, που δίνεται από τη σχέση,

$$\begin{aligned}
\left(\sum_{i=1}^{g} n_i k_i\right) \log(2\pi) &+ \sum_{i=1}^{g} n_i [k_i \log S_i^2 + (k_i - 1)\log(1 - r_i) + \log[1 + (k_i - 1) r_i]] \\
&+ \sum_{i=1}^{g} n_i k_i.
\end{aligned}$$

Υπό την H_0, η log-likelihood συνάρτηση, έπειτα από μεγιστοποίηση ως προς

$\alpha_i, i = 1, \ldots, g$, μπορεί να γραφεί ως,

$$\begin{aligned}
-2\log L = &\left(\sum_{i=1}^{g} n_i k_i\right)\log(2\pi) \\
&+ \sum_{i=1}^{g} n_i[k_i \log\sigma_i^2 + (k_i-1)\log(1-\rho) + \log[1+(k_i-1)\rho]] \\
&+ (1-\rho)^{-1}\sum_{i=1}^{g} n_i k_i b_i^2[1+(k_i-1)\rho]^{-1}[1+(k_i-2)\rho-(k_i-1)\rho r_i],
\end{aligned}$$

όπου $b_i^2 = S_i^2/\sigma_i^2$. Παραγωγίζοντας την παραπάνω σχέση ως προς σ_i^2 και ρ και εξισώνοντας με το 0 έχουμε ότι,

$$\sigma_i^2 = S_i^2[1-(k_i-1)\rho(r_i-\rho)(1-\rho)^{-1}[1+(k_i-1)\rho]^{-1}],$$

για $i = 1, \ldots, g$ και

$$\sum_{i=1}^{g} n_i k_i (k_i-1) b_i^2 (r_i-\rho)[1+(k_i-1)\rho]^{-2} = 0.$$

Οι εκιμητές μέγιστης πιθανοφάνειας των παραμέτρων $\sigma_i^2, i = 1, \ldots, g$ και ρ δεν μπορούν να εκφραστούν σε κλειστή μορφή. Οι εκτιμητές αυτοί είναι πολύ δύσκολο να αποκτηθούν ακόμη και αριθμητικά. Οι Konishi and Gupta (1989) προτείνουν έναν προσεγγιστικό εκτιμητή για τον κοινό intraclass συντελεστή συσχέτισης ρ που δίνεται από τη σχέση,

$$r = \sum_{i=1}^{g}(n_i k_i (k_i-1) r_i) \Bigg/ \sum_{i=1}^{g}(n_i k_i (k_i-1)). \tag{3.9}$$

Τότε το προσεγγιστικό στατιστικό του likelihood λόγου δίνεται από,

$$-2\log\Lambda = \sum_{i=1}^{g} n_i\left[k_i\log\left(\frac{\hat{\sigma}_i^2}{S_i^2}\right) + (k_i-1)\log\left(\frac{1-r}{1-r_i}\right) + \log\frac{[1+(k_i-1)r]}{1+(k_i-1)r_i}\right],$$

όπου $\hat{\sigma}_i^2 = S_i^2[1-(k_i-1)r(r_i-r)(1-r)^{-1}[1+(k_i-1)r]^{-1}]$ για $i = 1, \ldots, g$.

Το πρόβλημα του υπολογισμού της κατανομής του παραπάνω προτεινόμενου στατιστικού παραμένει, αφού το εν λόγω στατιστικό δεν έχει γενικά για μεγάλο μέγεθος δείγματος μια χ^2 κατανομή υπό την H_0. Οι Konishi and Gupta (1989), αρχικά προσδιορίζουν την ασυμπτωτική κατανομή της στατιστικής συνάρτησης υπό την μηδενική και την εναλλακτική υπόθεση, έχοντας υποθέσει δειγματοληψία από πολυδιάστατους μη κανονικούς πληθυσμούς. Καταλήγουν στο παρακάτω θεώρημα για την κατανομή του στατιστικού υπό την H_0 (βλ. Konishi and Gupta, 1989).

Θεώρημα 3.2.3. *Υπό την H_0, η ασυμπτωτική κατανομή του προσεγγιστικού λόγου πιθανοφανειών $-2\log\Lambda$, που βασίζεται σε τυχαίο δείγμα μεγέθους n_i από k_i-διάστατη κατανομή, $i = 1, \ldots, g$, με πεπερασμένες cumulants 4ης τάξης, είναι ένας γραμμικός συνδυασμός ανεξάρτητων χ^2 κατανομών, δηλαδή, καθώς $n = \sum\limits_{i=1}^{g} n_i \to \infty$,*

$$-2\log\Lambda \xrightarrow[n\to\infty]{d} \sum_{i=1}^{g} \omega_i \chi_1^2(i),$$

όπου $\chi_1^2(i)$ είναι μια χ^2 κατανομή με 1 βαθμό ελευθερίας και ω_i είναι οι ιδιοτιμές του πίνακα ΨG, όπου οι πίνακες Ψ και G ορίζονται πιο κάτω.

Στο παραπάνω θεώρημα ο πίνακας Ψ είναι ένας διαγώνιος πίνακας, δηλαδή $\Psi = diag(\phi_1^2, \ldots, \phi_g^2)$ όπου,

$$\begin{aligned}
\phi_p^2 =& \rho^2(\sqrt{f_p}k_p)^{-2}\sum_{i=1}^{g}\kappa_4^i(p) \\
&+ \{\sqrt{f_p}k_p(k_p-1)\}^{-2}\left[\sum_{i\neq j}^{k_p}\{(2+(k_p-1)^2\rho^2)\kappa_{22}^{ij}(p) - 4(k_p-1)\rho\kappa_{31}^{ij}(p)\}\right. \\
&\left. +2\sum_{i\neq j\neq k}^{k_p}\{2\kappa_{211}^{ijk}(p) - (k_p-1)\rho\kappa_{112}^{ijk}(p)\} + \sum_{i\neq j\neq k\neq r}^{k_p}\kappa_{1111}^{ijkr}(p)\right]
\end{aligned}$$

$$+2\{f_p k_p(k_p-1)\}^{-1}(1-\rho)^2[1+(k_p-1)\rho]^2,$$

για $p = 1, 2, \ldots, g$. Στην παραπάνω σχέση έχουμε $f_p = \frac{n_p}{n}$ και $\kappa^{1,2,\ldots,r}_{i_1,i_2,\ldots,i_r}(p)$ είναι οι cumulants της από κοινού κατανομής των $X_{i_1}, X_{i_2}, \ldots, X_{i_r}$. Επιπλέον, ο πίνακας G είναι ένας $g \times g$ συμμετρικός πίνακας του οποίου τα στοιχεία δίνονται από,

$$b_{pp} = \{\sqrt{2}(1-\rho)\}^{-2}\left\{(1-p_p)^2 h_p + p_p^2 \sum_{w(\neq p)}^{g} h_w\right\},$$

$$b_{pw} = \{\sqrt{2}(1-\rho)^2\}^{-2}\left\{p_p(p_w-1)h_w + p_w(p_p-1)h_p + p_p p_w \sum_{r(\neq p,w)}^{g} h_r\right\}, \; p \neq w,$$

όπου

$$p_p = f_p k_p(k_p-1)\Bigg/\left\{\sum_{p=1}^{g} f_p k_p(k_p-1)\right\},$$

$$h_p = f_p k_p(k_p-1)/\left\{1+(k_p-1)\rho\right\}^2,$$

για $p = 1, 2, \ldots, g$. Οι άγνωστες παράμετροι που υπάρχουν στους συντελεστές $\omega_i, i = 1, \ldots, g$ πρέπει να εκτιμηθούν από τα δεδομένα.

Εάν υποθέσουμε ότι η δειγματοληψία γίνεται από πολυδιάστατους κανονικούς πληθυσμούς, τότε το πλήθος των cumulants για τις οποίες $\sum i_r > 2$ είναι μηδέν. Επομένως, το παρακάτω θεώρημα ισχύει ως ειδική περίπτωση του προηγούμενου θεωρήματος (βλ. Konishi and Gupta, 1989).

Θεώρημα 3.2.4. *Η ασυμπτωτική κατανομή του στατιστικού $-2\log\Lambda$ υπό την H_0, που βασίζεται σε τυχαία δείγματα που προέρχονται από πολυδιάστατες κανονικές κατανομές είναι, καθώς $n = \sum_{i=1}^{g} n_i \to \infty$, ένας γραμμικός συνδυασμός ανεξάρτητων χ^2 κατανομών με ένα βαθμό ελευθερίας των οποίων οι συντελεστές δίνονται από τις ιδιοτιμές του πίνακα ΨG, όπου ο G υπολογίζεται όπως πάνω και*

$\Psi = diag(\phi_1^2, \ldots, \phi_g^2)$ *όπου,*

$$\phi_p^2 = 2\{f_p k_p(k_p - 1)\}^{-1}(1-\rho)^2[1 + (k_p - 1)\rho]^2,$$

για $p = 1, \ldots, g$.

Οι Paul and Barnwal (1990), κατασκεύσαν το τεστ πηλίκου πιθανοφανειών για τον έλεγχο της ισότητας g intraclass συντελεστών συσχέτισης, που βασίζεται σε g τυχαία δείγματα μεγέθους $n_i, i = 1, \ldots, g$, που προέρχονται από g ανεξάρτητους κανονικούς πληθυσμούς. Υποθέτουμε ότι ισχύουν οι υποθέσεις του μοντέλου (3.1) και ότι κάθε οικογένεια του i-οστού πληθυσμού έχει τον ίδιο αριθμό παιδιών, δηλαδή $k_{ij} = k_i$. Επίσης υποθέτουμε ότι $\mu_{ij} = \mu_i$ και $\Sigma_{ij} = \Sigma_i, i = 1, \ldots, g, j = 1, \ldots, n_i$. Πιο συγκεκριμένα, θέλουμε να εξετάσουμε τη μηδενική υπόθεση $H_0 : \rho_1 = \cdots = \rho_g = \rho$ έναντι της εναλλακτικής H_1 : υπάρχει τουλάχιστον ένα ζεύγος intraclass συντελεστών συσχέτισης που είναι διαφορετικοί μεταξύ τους. Σύμφωνα με τους Paul and Barnwal (1990), το στατιστικό του λόγου πιθανοφανειών αποτελεί μια φυσική επέκταση του στατιστικού που έδωσαν οι Donner and Bull (1983), για την περίπτωση $g = 2$. Έτσι, προτείνεται το στατιστικό της επόμενης σχέσης,

$$L_D = \sum_{i=1}^{g} n_i \left[k_i \log\left(\frac{\hat{\sigma}_i^2}{S_i^2}\right) + (k_i - 1)\log\left(\frac{1-\hat{\rho}_M}{1-r_i}\right) + \log\left\{\frac{1+(k_i-1)\hat{\rho}_M}{1+(k_i-1)r_i}\right\}\right],$$

όπου,

$$\hat{\alpha}_i = \sum_{j=1}^{n_i}\sum_{m=1}^{k_i} X_{ijm}/n_i k_i,$$

$$\hat{\sigma}_i^2 = \frac{S_i^2}{1-\hat{\rho}_M} - \frac{k_i\hat{\rho}_M ST_i}{(1-\hat{\rho}_M)[1+(k_i-1)\hat{\rho}_M]},$$

$$S_i^2 = \sum_{j=1}^{n_i}\sum_{m=1}^{k_i}(X_{ijm} - \hat{\alpha}_i)^2/n_i k_i,$$

$$ST_i = \sum_{j=1}^{n_i}(\bar{X}_{ij0} - \hat{\alpha}_i)^2/n_i, \ \ \bar{X}_{ij0} = \sum_{m=1}^{k_i} X_{ijm}/k_i.$$

Ακόμη, $\hat{\rho}_M$ προκύπτει από τη λύση της εξίσωσης,

$$\sum_{i=1}^{g} \frac{n_i k_i (k_i - 1)(\rho - r_i)}{(1 - \rho R_i)[1 + (k_i - 1)\rho]} = 0,$$

όπου,

$$r_i = \frac{1}{k_i - 1}\left(\frac{k_i ST_i}{S_i^2} - 1\right), \ \ R_i = 1 - (k_i - 1)(1 - r_i), i = 1, \ldots, g$$

και $\hat{\rho}_M \in \bigcap\limits_{i=1}^{g}\left(-\frac{1}{k_i - 1}, 1\right)$. Συνοπτικά τα αποτελέσματα διατυπώνονται στο παρακάτω θεώρημα (βλ. Paul and Barnwal, 1990).

Θεώρημα 3.2.5. *Έστω g ανεξάρτητα τυχαία δείγματα μεγέθους n_i από g ανεξάρτητους κανονικούς πληθυσμούς για τους οποίους ισχύουν οι υποθέσεις του μοντέλου (3.1). Για τον έλεγχο της υπόθεσης,*

$H_0 : \rho_1 = \cdots = \rho_g = \rho$ *έναντι* H_1 :*υπάρχει τουλάχιστον ένα ζεύγος intraclass συντελεστών συσχέτισης που είναι διαφορετικοί μεταξύ τους,*

το στατιστικό, που προκύπτει με τη μέθοδο του πηλίκου μέγιστων πιθανοφανειών ορίζεται ως,

$$L_D = \sum_{i=1}^{g} n_i \left[k_i \log\left(\frac{\hat{\sigma}_i^2}{S_i^2}\right) + (k_i - 1)\log\left(\frac{1 - \hat{\rho}_M}{1 - r_i}\right) + \log\left\{\frac{1 + (k_i - 1)\hat{\rho}_M}{1 + (k_i - 1)r_i}\right\}\right].$$

Για μεγάλα μεγέθη δείγματος $n_i, i = 1, \ldots, g$ έχουμε ότι,

$$L_D \xrightarrow[n_1,\ldots,n_g \to \infty]{d} \chi^2_{g-1} \quad \text{υπό την } H_0.$$

Απορρίπτουμε την H_0 για μεγάλες τιμές του στατιστικού και πιο συγκεκριμένα απορρίπτουμε όταν $L_D \geq \chi^2_{g-1,\alpha}$, όπου α είναι το επιθυμητό επίπεδο σημαντικότητας.

Στη συνέχεια οι Paul and Barnwal (1990), κατασκεύσαν ένα $C(\alpha)$ τεστ. Για τον υπολογισμό του παραπάνω τεστ η εναλλακτική υπόθεση μπορεί να γραφεί ως, $\rho_i = \rho + \phi_i, i = 1, \ldots, g$, με $\phi_g = 0$. Έτσι, η μηδενική υπόθεση για την ισότητα των ρ_i είναι ισοδύναμη με τον έλεγχο της μηδενικής υπόθεσης $H_0 : \phi_i = 0, i = 1, \ldots, g$. Οι παράμετροι α_i, σ_i^2 και $\rho, i = 1, \ldots, g$ αντιμετωπίζονται ως ενοχλητικές (nuisance) παράμετροι. Οι Paul and Barnwal (1990), υποθέτουν χωρίς βλάβη της γενικότητας για την κατασκευή του $C(\alpha)$ τεστ ότι $\alpha_1 = \cdots = \alpha_g = 0$. Επομένως, οι ενοχλητικές παράμετροι είναι οι $\sigma_1^2, \ldots, \sigma_g^2$ και ρ. Για τη log-likelihood συνάρτηση, αντικαθιστώντας για $\alpha_1 = \cdots = \alpha_g = 0$ και ρ με $\rho + \phi_i, i = 1, \ldots, g$ ισχύει ότι,

$$\begin{aligned} -2\ell = & \sum_{i=1}^{g} \big(k_i n_i \log \sigma_i^2 + n_i(k_i - 1)\log(1 - (\rho + \phi_i))\big) \\ & + \sum_{i=1}^{g} \Bigg(n_i \log[1 + (k_i - 1)(\rho + \phi_i)] + \frac{\sum_{j=1}^{n_i} \sum_{m=1}^{k_i} X_{ijm}^2}{[1 - (\rho + \phi_i)]\sigma_i^2} \\ & - \frac{(\rho + \phi_i) k_i^2 \sum_{j=1}^{n_i} \bar{X}_{ij0}^2}{[1 - (\rho + \phi_i)][1 + (k_i - 1)(\rho + \phi_i)]\sigma_i^2} \Bigg). \end{aligned}$$

Ορίζουμε $\phi = (\phi_1, \ldots, \phi_{g-1})', \theta = (\theta_1, \ldots, \theta_g, \theta_{g+1})' = (\sigma_1^2, \ldots, \sigma_g^2, \rho)'$ και υπολογίζουμε τα παρακάτω,

$$\psi_i = \left.\frac{\partial \ell}{\partial \phi_i}\right|_{\phi=0}, i = 1, \ldots, g - 1$$

και

$$\gamma_j = \left.\frac{\partial \ell}{\partial \theta_j}\right|_{\phi=0}, j = 1, \ldots, g + 1.$$

Έστω $\hat{\theta}$ ένας $\sqrt{n}$-συνεπής εκτιμητής της παραμέτρου θ υπό την H_0. Τότε, το $C(\alpha)$ τεστ βασίζεται στα,

$$T_i(\hat{\theta}) = \psi_i(\hat{\theta}) - \sum_{j=1}^{g+1} \beta_{ij}\gamma_j(\hat{\theta}), i = 1, \dots, g-1,$$

όπου $\beta_{ij}, j = 1, \dots, g+1$, είναι ο μερικός συντελεστής παλινδρόμησης του ψ_i στο γ_j. Οι Paul and Barnwal (1990) χρησιμοποιούν ως $\sqrt{n}$-συνεπή εκτιμητή τον εκτιμητή μέγιστης πιθανοφάνειας της παραμέτρου θ, άρα $\gamma_j = 0, \forall j$, στην παραπάνω σχέση και $T_i(\hat{\theta}) = \psi_i(\hat{\theta})$. Ο πίνακας διακυμάνσεων-συνδιακυμάνσεων του $T = (T_1(\theta), \dots, T_{g-1}(\theta))$, δίνεται από τον πίνακα $A - BC^{-1}B'$, όπου το (i,j) στοιχείο των πινάκων $A_{(g-1)\times(g-1)}, B_{(g-1)\times(g+1)}$ και $C_{(g+1)\times(g+1)}$ είναι αντίστοιχα,

$$\begin{aligned} A_{ij} &= -E\left(\left.\frac{\partial^2 \ell}{\partial\phi_i \partial\phi_j}\right|_{\phi=0} \right), \ B_{ij} = -E\left(\left.\frac{\partial^2 \ell}{\partial\phi_i \partial\theta_j}\right|_{\phi=0} \right), \\ C_{ij} &= -E\left(\left.\frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j}\right|_{\phi=0} \right). \end{aligned}$$

Αφού αντικατασταθούν τα T, A, B και C με $\hat{\theta}$, (στη θέση του θ), το $C(\alpha)$ στατιστικό είναι το $\chi^2_C = T'(A - BC^{-1}B')^{-1}T$, το οποίο έχει προσεγγιστικά μια χ^2 κατανομή με $g-1$ βαθμούς ελευθερίας. Πιο συγκεκριμένα, το στατιστικό δίνεται από τη σχέση,

$$\chi^2_C = \sum_{i=1}^{g} \frac{n_i k_i (k_i - 1)(\hat{\rho}_M - r_i)^2}{2(1 - \hat{\rho}_M R_i)^2},$$

όπου $\hat{\rho}_M, r_i$ και $R_i, i = 1, \dots, g$, έχουν οριστεί παραπάνω.

Ένας εναλλακτικός τρόπος κατασκευής στατιστικών τεστ για έλεγχο ισότητας intraclass συντελεστών συσχέτισης προτείνεται στην εργασία των Naik and Helu (2007). Στην εργασία αυτή περιγράφεται η διαδικασία με την οποία κατασκευάζονται τρία ασυμπτωτικά τεστ, που βασίζονται στη μέθοδο της μέγιστης πιθανοφάνειας, για τον έλεγχο της υπόθεσης της ισότητας g intraclass συντελεστών συσχέτισης. Υποθέτουμε ότι ισχύει το μοντέλο (3.1) στην πιο γενική μορφή του. Η συνάρτηση πιθανοφάνειας δίνεται από τη σχέση,

$$L(\theta)=\prod_{i=1}^{g}\prod_{j=1}^{n_i}\frac{1}{(2\pi)^{k_{ij}/2}|\Sigma_{ij}|^{1/2}}\exp\left\{-\frac{1}{2}(X_{ij}-\mu_{ij})'\Sigma_{ij}^{-1}(X_{ij}-\mu_{ij})\right\},$$

όπου $\theta=(\alpha_1,\sigma_1^2,\rho_1,\ldots,\alpha_g,\sigma_g^2,\rho_g)$. Λαμβάνοντας υπόψιν τις σχέσεις που ισχύουν για τον πίνακα $\Sigma_{ij}, i=1,\ldots,g, j=1,\ldots,n_i$, (βλ. Mardia et. al., 1979, σ. 97),

$$|\Sigma_{ij}|=(\sigma_i^2)^{k_{ij}}\left[(1-\rho_i)^{k_{ij}-1}[1+(k_{ij}-1)\rho_i]\right]$$

και

$$\Sigma_{ij}^{-1}=\frac{1}{\sigma_i^2(1-\rho_i)}\left[I_{k_{ij}}-\frac{\rho_i}{1+(k_{ij}-1)\rho_i}J_{k_{ij}}\right],$$

η log-likelihood συνάρτηση παίρνει την παρακάτω μορφή,

$$\begin{aligned}&\log L(\theta)=\\&\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left\{-\frac{k_{ij}}{2}\log(2\pi\sigma_i^2)-\frac{1}{2}\left[(k_{ij}-1)\log(1-\rho_i)+\log(1+(k_{ij}-1)\rho_i)\right]\right.\\&\left.-\frac{1}{2\sigma_i^2(1-\rho_i)}(X_{ij}-\alpha_i1_{k_{ij}})'\left[I_{k_{ij}}-\frac{\rho_i}{1+(k_{ij}-1)\rho_i}J_{k_{ij}}\right](X_{ij}-\alpha_i1_{k_{ij}})\right\}.\end{aligned}$$

Έστω $\hat{\theta}$ είναι ο εκτιμητής μέγιστης πιθανοφάνειας της παραμέτρου θ που προκύπτει μεγιστοποιώντας την $L(\theta)$ και $\hat{\theta}_0$ είναι ο εκτιμητής μέγιστης πιθανοφάνειας που προκύπτει μεγιστοποιώντας την $L(\theta)$ υπό την $H_0:\rho_1=\rho_2=\cdots=\rho_g$. Να σημειωθεί ότι, $\theta_0=(\alpha_1,\sigma_1^2,\rho,\mu_2,\sigma_2^2,\rho,\ldots,\alpha_g,\sigma_g^2,\rho)'$, που προκύπτει από την παράμετρο θ υπό την H_0. Όπως αναφέρθηκε παραπάνω θα περιγραφεί η διαδικασία για την κατασκευή τριών τεστ που βασίζονται στη μέθοδο μέγιστης πιθανοφάνειας.

Τεστ πηλίκου πιθανοφανειών

Το στατιστικό είναι το,

$$-2\log\lambda = 2\log L(\hat{\theta}) - 2\log L(\hat{\theta}_0).$$

Απορρίπτουμε την H_0 για μεγάλες τιμές του στατιστικού και η ασυμπτωτική κατανομή του είναι χ^2 με $g-1$ βαθμούς ελευθερίας.

Score τεστ του Rao

Έστω $S(\theta) = \partial\log L(\theta)/\partial\theta$ είναι ένα $3g \times 1$ διάνυσμα της score συνάρτησης και $I(\theta) = E\left[(\partial\log L(\theta)/\partial\theta)(\partial\log L(\theta)/\partial\theta)'\right]$ είναι ο $3g \times 3g$ πίνακας πληροφορίας του Fisher. Τότε το score στατιστικό τεστ δίνεται από την παρακάτω σχέση,

$$S(\hat{\theta}_0)' I(\hat{\theta}_0)^{-1} S(\hat{\theta}_0).$$

Η ασυμπτωτική κατανομή του στατιστικού είναι και στην περίπτωση αυτή η χ^2 κατανομή με $g-1$ βαθμούς ελευθερίας. Να σημειωθεί ότι $S(\theta) = (S_1(\theta), \ldots, S_g(\theta))'$, όπου,

$$S_i(\theta) = (\partial\log L(\theta)/\partial\alpha_i, \partial\log L(\theta)/\partial\sigma_i^2, \partial\log L(\theta)/\partial\rho_i)'$$

και

$$\begin{aligned}
\frac{\partial\log L(\theta)}{\partial\alpha_i} &= \frac{1}{\sigma_i^2(1-\rho_i)}\sum_{j=1}^{n_i} 1'_{k_{ij}}\left[I_{k_{ij}} - \frac{\rho_i}{1+(k_{ij}-1)\rho_i}J_{k_{ij}}\right]e_{ij}, \\
\frac{\partial\log L(\theta)}{\partial\sigma_i^2} &= -\frac{1}{2\sigma_i^2}\sum_{j=1}^{n_i} k_{ij} + \sum_{j=1}^{n_i}\frac{1}{2\sigma_i^4(1-\rho_i)}e'_{ij}\left[I_{k_{ij}} - \frac{\rho_i}{1+(k_{ij}-1)\rho_i}J_{k_{ij}}\right]e_{ij}, \\
\frac{\partial\log L(\theta)}{\partial\rho_i} &= \frac{\rho_i}{2(1-\rho_i)}\sum_{j=1}^{n_i}\frac{k_{ij}(k_{ij}-1)}{1+(k_{ij}-1)\rho_i} \\
&\quad - \frac{1}{2\sigma_i^2(1-\rho_i)^2}\sum_{j=1}^{n_i} e'_{ij}\left[I_{k_{ij}} - \frac{1+(k_{ij}-1)\rho_i^2}{(1+(k_{ij}-1)\rho_i)^2}J_{k_{ij}}\right]e_{ij},
\end{aligned}$$

όπου $e_{ij} = X_{ij} - \alpha_i 1_{k_{ij}}$.

Παρατηρούμε ότι ο πίνακας πληροφορίας του Fisher $I(\theta)$ είναι ένας block διαγώνιος πίνακας που αποτελείται από g blocks 3×3 πινάκων. Τότε, γνωρίζουμε ότι ο αντίστροφος πίνακας πληροφορίας του Fisher, $I(\theta)^{-1}$, θα είναι επίσης block διαγώνιος. Παρακάτω δίνεται η μορφή του i-οστού block I_i του πίνακα πληροφορίας του Fisher, ο αντίστροφος του οποίου μπορεί να βρεθεί.

$$I_i =$$
$$\begin{pmatrix} \sum_{j=1}^{n_i} \frac{k_{ij}}{\sigma_i^2(1+(k_{ij}-1)\rho_i)} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sum_{j=1}^{n_i} \frac{k_{ij}}{2\sigma_i^4} & \sum_{j=1}^{n_i} \frac{\rho_i k_{ij}(k_{ij}-1)}{2\sigma_i^2(1-\rho_i)(1+(k_{ij}-1)\rho_i)} \\ \mathbf{0} & \sum_{j=1}^{n_i} \frac{\rho_i k_{ij}(k_{ij}-1)}{2\sigma_i^2(1-\rho_i)(1+(k_{ij}-1)\rho_i)} & \sum_{j=1}^{n_i} \frac{k_{ij}(k_{ij}-1)(1+(k_{ij}-1)\rho_i^2)}{2(1-\rho_i)^2(1+(k_{ij}-1)\rho_i)^2} \end{pmatrix}.$$

Τεστ του Wald

Έστω $\rho = (\rho_1, \rho_2, \dots, \rho_g)'$ και $\hat{\rho} = (\hat{\rho}_1, \hat{\rho}_2, \dots, \hat{\rho}_g)'$ είναι ο εκτιμητής μέγιστης πιθανοφάνειας του ρ. Τότε η μηδενική υπόθεση $H_0 : \rho_1 = \rho_2 = \cdots = \rho_g$ μπορεί να εκφραστεί ως $C\rho = 0$, όπου C είναι ένας $(g-1) \times g$ πίνακας της μορφής:

$$C = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}.$$

Έστω V_ρ είναι ο $g \times g$ διαγώνιος πίνακας που κατασκευάζεται από g blocks του $I(\hat{\theta})^{-1}$ χρησιμοποιώντας το 3ο διαγώνιο στοιχείο του κάθε block I_i (ξεκινώντας από το πρώτο block). Τότε το στατιστικό του Wald είναι

$$(C\hat{\rho})' \left[CV_{\hat{\rho}}C'\right]^{-1} (C\hat{\rho}).$$

Η ασυμπτωτική κατανομή του στατιστικού αυτού είναι επίσης μια χ^2 με $g-1$ βαθμούς ελευθερίας.

Κεφάλαιο 4

ΜΕΛΕΤΗ ΤΟΥ **INTRACLASS** ΜΟΝΤΕ-ΛΟΥ ΜΕ ΜΕΘΟΔΟΥΣ ΑΝΑΛΥΣΗΣ ΔΙΑ-ΚΥΜΑΝΣΗΣ

4.1 Εισαγωγή

Μία άλλη προσέγγιση του προβλήματος της εκτίμησης του intraclass συντελεστή συσχέτισης ρ πραγματοποιείται με την υιοθέτηση ενός one-way μοντέλου τυχαίων επιδράσεων (random effect model). Τότε, σύμφωνα με τον Donner (1986) η παρατήρηση ενός χαρακτηριστικού X του j-οστού μέλους της i-οστής οικογένειας, $j = 1, 2, \ldots, k_i, i = 1, 2, \ldots, n$, μπορεί μαθηματικά να περιγραφεί από την παρακάτω σχέση,

$$X_{ij} = \mu + \alpha_i + e_{ij}, \tag{4.1}$$

όπου μ είναι η μέση τιμή όλων των παρατηρήσεων του πληθυσμού, a_i είναι η τυχαία επίδραση της i-οστής οικογένειας και e_{ij} είναι ένας όρος σφάλματος του j-οστού μέλους της i-οστής οικογένειας. Πιο συγκεκριμένα, όπως αναφέρει ο

Searle (1997) οι τυχαίες επιδράσεις α_i είναι ανεξάρτητες τυχαίες μεταβλητές με κοινή μέση τιμή 0 και διακύμανση σ_A^2. Επιπλέον, οι όροι σφάλματος e_{ij} αποτελούν και αυτοί ένα τυχαίο δείγμα με κοινή μέση τιμή 0 και διακύμανση σ_e^2. Στο παραπάνω μοντέλο υποθέτουμε ότι η δειγματοληψία των τυχαίων επιδράσεων α είναι ανεξάρτητη από εκείνη των σφαλμάτων e, γεγονός που σημαίνει ότι οι συνδιακυμάνσεις μεταξύ των α και e είναι μηδέν. Συνοπτικά ισχύουν οι παρακάτω σχέσεις για $i = 1, 2, \ldots, n$ και $j = 1, 2, \ldots, k_i$,

$$\begin{aligned}
E(X_{ij}) &= \mu, \\
E(\alpha_i) &= 0, \\
Var(\alpha_i) &= \sigma_A^2, \\
Cov(\alpha_i, \alpha_j) &= 0, i \neq j, \\
E(e_{ij}) &= 0, \\
Var(e_{ij}) &= \sigma_e^2, \\
Cov(e_{ij}, e_{i'l}) &= \begin{cases} 0, & i \neq i' \\ 0, & i = i', j \neq l \end{cases}, \\
Cov(\alpha_i, e_{ij}) &= 0.
\end{aligned}$$

Επομένως, σύμφωνα με τα παραπάνω και ακολουθώντας τον Donner (1986), η διακύμανση μιας παρατήρησης X_{ij} για οποιαδήποτε i και j δίνεται από τη σχέση,

$$\begin{aligned}
\sigma_X^2 = E(X_{ij} - \mu)^2 = E(\alpha_i + e_{ij})^2 &= E(\alpha_i^2) + 2E(\alpha_i e_{ij}) + E(e_{ij}^2) \\
&= \sigma_A^2 + 2Cov(\alpha_i, e_{ij}) + \sigma_e^2
\end{aligned}$$

ή

$$\sigma_X^2 = \sigma_A^2 + \sigma_e^2. \tag{4.2}$$

O intraclass συντελεστής συσχέτισης ρ ορίζεται ως ο συνήθης συντελεστής συσχέτισης μεταξύ δύο οποιονδήποτε παρατηρήσεων X_{ij} και X_{il} της ίδιας οικογένειας (i σταθερό , $j \neq l$). Έτσι,

$$\begin{aligned}
\rho = \frac{Cov(X_{ij}, X_{il})}{\sigma_X \sigma_X} &= \frac{E[(X_{ij} - \mu)(X_{il} - \mu)]}{\sigma_X^2} = \frac{E(\alpha_i + e_{ij})(\alpha_i + e_{il})}{\sigma_A^2 + \sigma_e^2} \\
&= \frac{E(\alpha_i^2) + E(\alpha_i e_{il}) + E(e_{ij}\alpha_i) + E(e_{ij}e_{il})}{\sigma_A^2 + \sigma_e^2} \\
&= \frac{E(\alpha_i^2) + Cov(\alpha_i, e_{il}) + Cov(e_{ij}, \alpha_i) + Cov(e_{ij}, e_{il})}{\sigma_A^2 + \sigma_e^2} \\
&= \frac{E(a_i^2)}{\sigma_A^2 + \sigma_e^2} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_e^2}
\end{aligned}$$

ή

$$\rho = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_e^2}. \tag{4.3}$$

Οι διακυμάνσεις σ_A^2 και σ_e^2 ονομάζονται συνιστώσες διακύμανσης (variance components), διότι κάθε μία είναι από μόνη της διακύμανση και ταυτόχρονα συνιστώσα της σ_X^2. Το μοντέλο αυτό αναφέρεται κάποιες φορές ως variance components μοντέλο και όταν αυτό υιοθετείται το ενδιαφέρον επικεντρώνεται κυρίως στην εκτίμηση των συνιστωσών διακύμανσης σ_A^2 και σ_e^2. Ο εκτιμητής της intraclass συσχέτισης ρ, έστω $\hat{\rho}_A$, δίνεται από τη σχέση,

$$\hat{\rho}_A = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_A^2 + \hat{\sigma}_e^2}, \tag{4.4}$$

όπου $\hat{\sigma}_A^2$ και $\hat{\sigma}_e^2$ είναι εκτιμητές των συνιστωσών διακύμανσης. Είναι προφανές λοιπόν, ότι ο υπολογισμός των $\hat{\sigma}_A^2$ και $\hat{\sigma}_e^2$ θα εξασφαλίσει και τον υπολογισμό του $\hat{\rho}_A$, που είναι και το ζητούμενο. Θα εξεταστεί η εκτίμηση των σ_A^2 και σ_e^2 για δύο περιπτώσεις. Η πρώτη είναι αυτή των balanced δεδομένων, δηλαδή οι υπό θεώρηση n οικογένειες έχουν τον ίδιο αριθμό παιδιών, όπου $k_i = k$ ($i = 1, 2, \ldots, n$) και η δεύτερη είναι των unbalanced δεδομένων, όπου τα μεγέθη των οικογενειών διαφέρουν μεταξύ τους.

4.2 Balanced περίπτωση

4.2.1 Συνιστώσες διακύμανσης

Σύμφωνα με τον Searle (1997, σ. 384) η εκτίμηση των συνιστωσών διακύμανσης από balanced δεδομένα ($k_i = k$) βασίζεται σχεδόν αποκλειστικά σε μία μέθοδο. Η μέθοδος στηρίζεται στα μέσα τετράγωνα (mean squares) της ανάλυσης διακύμανσης για το αντίστοιχο μοντέλο σταθερών επιδράσεων (fixed effects model). Η γενική διαδικασία έχει ως εξής,

i) Πραγματοποιείται η ανάλυση διακύμανσης αντιμετωπίζοντας το μοντέλο ως ένα μοντέλο σταθερών επιδράσεων.

ii) Υπολογίζονται οι αναμενόμενες τιμές των μέσων τετραγώνων αντιμετωπίζοντας πλέον το μοντέλο ως ένα μοντέλο τυχαίων επιδράσεων (random effect model). Οι αναμενόμενες τιμές θα είναι γραμμικές συναρτήσεις των συνιστωσών διακύμανσης.

iii) Τέλος, εξισώνοντας τις αναμενόμενες τιμές των μέσων τετραγώνων με τις τιμές των μέσων τετραγώνων που προέκυψαν από το βήμα i), προκύπτουν γραμμικές εξισώσεις των συνιστωσών διακύμανσης, οι λύσεις των οποίων είναι οι εκτιμητές αυτών.

Η παραπάνω μέθοδος είναι γνωστή ως μέθοδος της ανάλυσης διακύμανσης. Στο σημείο αυτό θα δοθεί η μορφή του πίνακα διακυμάνσεων-συνδιακυμάνσεων των παρατηρήσεων των μελών της ίδιας οικογένειας. Έστω $X_i = (X_{i1}, \ldots, X_{ik})'$ το διάνυσμα των παρατηρήσεων της i-οστής οικογένειας, $i = 1, 2, \ldots, n$. Από τις παραπάνω σχέσεις έχουμε ότι,

$$Cov(X_{ij}, X_{il}) = \begin{cases} \sigma_A^2 + \sigma_e^2, & j = l \\ \sigma_A^2, & j \neq l \end{cases}.$$

Επομένως, ο πίνακας θα έχει την παρακάτω μορφή,

$$Var(X_i) = \begin{pmatrix} \sigma_A^2 + \sigma_e^2 & \sigma_A^2 & \cdots & \sigma_A^2 \\ \sigma_A^2 & \sigma_A^2 + \sigma_e^2 & \cdots & \sigma_A^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_A^2 & \cdots & \cdots & \sigma_A^2 + \sigma_e^2 \end{pmatrix}_{k \times k} = \sigma_e^2 I_k + \sigma_A^2 J_k, \quad (4.5)$$

όπου I_k είναι ο ταυτοτικός πίνακας τάξης k και J_k είναι ένας πίνακας τάξης k, όπου κάθε στοιχείο του είναι ίσο με τη μονάδα.

Έτσι, εάν θέλουμε να γράψουμε τη μορφή του πίνακα διακυμάνσεων-συνδιακυμάνσεων για το διάνυσμα $X = (X_1', X_2', \ldots, X_n')'$, δηλαδή το διάνυσμα όλων των παρατηρήσεων, αρκεί να λάβουμε υπόψιν μας ότι $Cov(X_i, X_{i'}) = 0_{k \times k}$, για $i \neq i'$ και τη σχέση (4.5). Τότε προκύπτει ότι,

$$Var(X) = \begin{pmatrix} \sigma_e^2 I_k + \sigma_A^2 J_k & 0_{k \times k} & \cdots & 0_{k \times k} \\ 0_{k \times k} & \sigma_e^2 I_k + \sigma_A^2 J_k & \cdots & 0_{k \times k} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{k \times k} & 0_{k \times k} & \cdots & \sigma_e^2 I_k + \sigma_A^2 J_k \end{pmatrix}_{(nk) \times (nk)}, \quad (4.6)$$

όπου I_k και J_k ορίζονται όπως και παραπάνω και $0_{k \times k}$ ο μηδενικός πίνακας τάξης k.

Γνωρίζουμε ότι εάν $A_{p \times q}$ και $B_{r \times s}$ είναι δύο πίνακες, τότε το ευθύ τους γινόμενο ή αλλιώς Kronecker product ορίζεται ως (βλ. Schäcke, 2013),

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1q}B \\ a_{21}B & a_{22}B & \cdots & a_{2q}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1}B & a_{p2}B & \cdots & a_{pq}B \end{pmatrix}.$$

Επομένως, μπορούμε να γράψουμε τον πίνακας της σχέσης (4.6) ως,

$$Var(X) = I_n \otimes (\sigma_e^2 I_k + \sigma_A^2 J_k).$$

Η ανάλυση διακύμανσης του μοντέλου (4.1) και οι αναμενόμενες τιμές των μέσων τετραγώνων φαίνονται στον παρακάτω πίνακα (Πίνακας 4.1), που έχει κατασκευαστεί για την balanced περίπτωση (βλ. Searle, 1997, σ. 405),

Source of Variation	*d.f*	*Sum of Squares*	*Mean Square*	*Expected value of mean square*
Mean	1	$SSM = T_\mu$	$MSM = SSM$	$N\mu^2 + k\sigma_A^2 + \sigma_e^2$
Classes	$n-1$	$SSA = T_A - T_\mu$	$MSA = SSA/(n-1)$	$k\sigma_A^2 + \sigma_e^2$
Residual error	$n(k-1)$	$SSE = T_0 - T_A$	$MSE = SSE/(n(k-1))$	σ_e^2
Total	kn	$SST = T_0$	-	-

Πίνακας 4.1: Ανάλυση διακύμανσης για το μοντέλο (4.1) με k παρατηρήσεις σε κάθε μια από τις n κλάσεις

όπου $N = kn$, $T_0 = \sum_{i=1}^{n}\sum_{j=1}^{k} X_{ij}^2$, $T_A = k\sum_{i=1}^{n} \bar{X}_{i.}^2$ και $T_\mu = N\bar{X}_{..}^2$ με $\bar{X}_{i.} = \frac{1}{k}\sum_{j=1}^{k} X_{ij}$ και $\bar{X}_{..} = \frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k} X_{ij}$.

Παρατηρούμε ότι η $E(MSM)$ περιέχει την παράμετρο μ, που είναι μια σταθερή επίδραση (fixed effect). Ωστόσο, όπως αναφέρει και ο Searle (1997, σ. 388) στη μέθοδο ανάλυσης διακύμανσης θα υπάρχουν τόσα μέσα τετράγωνα των οποίων οι αναμενόμενες τιμές δεν περιέχουν fixed effects, όσες και οι συνιστώσες διακύμανσης που θέλουμε να εκτιμήσουμε. Εξισώνοντας αυτά τα μέσα τετράγωνα με τις αναμενόμενες τιμές τους θα προκύψει ένα σύστημα γραμμικών εξισώσεων με αγνώστους τις συνιστώσες διακύμανσης. Έτσι, εφαρμόζοντας τη μέθοδο ανάλυσης διακύμανσης (βλ. Searle, 1997, σ. 405) προκύπτει το παρακάτω σύστημα εξισώσεων,

$$\begin{cases} E(MSA) = MSA \\ E(MSE) = MSE \end{cases} \quad \text{ή} \quad \begin{cases} k\sigma_A^2 + \sigma_e^2 = MSA \\ \sigma_e^2 = MSE \end{cases} \quad \text{ή} \quad \begin{cases} k\sigma_A^2 = MSA - MSE \\ \sigma_e^2 = MSE \end{cases} .$$

Τελικά έχουμε,

$$\sigma_A^2 = \frac{MSA - MSE}{k},$$
$$\sigma_e^2 = MSE.$$

Επομένως, με τη μέθοδο ανάλυσης διακύμανσης καταλήγουμε στους εκτιμητές $\hat{\sigma}_A^2$, $\hat{\sigma}_e^2$, των παραμέτρων σ_A^2 και σ_e^2 αντίστοιχα, που δίνονται από τις πρακάτω σχέσεις,

$$\hat{\sigma}_A^2 = \frac{MSA - MSE}{k}, \tag{4.7}$$
$$\hat{\sigma}_e^2 = MSE. \tag{4.8}$$

Ιδιότητες των εκτιμητών

Ο Searle (1997, σ. 405-408) σχολιάζει κάποιες από τις ιδιότητες των εκτιμητών των σχέσεων (4.7) και (4.8), που προέκυψαν με τη μέθοδο της ανάλυσης διακύμανσης.

i) **Αμεροληψία**

Έστω ότι $m = \{M_i\}$, για $i = 1, 2, \ldots, p$, είναι το διάνυσμα των μέσων τετραγώνων της ανάλυσης διακύμανσης, έτσι ώστε $E(m)$ δεν περιέχει fixed effects και σ^2 είναι το διάνυσμα των συνιστωσών διακύμανσης που πρέπει να εκτιμηθούν, με $E(m) = P\sigma^2$, για P αντιστρέψιμο πίνακα. Τότε οι εξισώσεις που θα δώσουν τους εκτιμητές είναι,

$$E(m) = m \text{ ή } P\sigma^2 = m.$$

Εάν συμβολίσουμε με $\hat{\sigma}^2$ τις λύσεις των παραπάνω εξισώσεων προκύπτει ότι,

$$\hat{\sigma}^2 = P^{-1}m. \tag{4.9}$$

Τότε,

$$E(\hat{\sigma}^2) = E(P^{-1}m) = P^{-1}E(m) = P^{-1}P\sigma^2 = \sigma^2,$$

γεγονός που επιβεβαιώνει την αμεροληψία των εκτιμητών. Ωστόσο, να τονιστεί ότι αυτή είναι μια ιδιότητα που ισχύει στην balanced περίπτωση. Ακόμη, η παραπάνω απόδειξη έγινε για πιο γενικά μοντέλα, επομένως θα ισχύει και για το μοντέλο (4.1) που είναι μια ειδική περίπτωση των τελευταίων.

ii) **Ελάχιστη διακύμανση**

Ο Searle (1997, σ. 406) αναφέρει πως οι εκτιμητές $\hat{\sigma}^2$ της σχέσης (4.9) έχουν τη μικρότερη διακύμανση στην κλάση των αμερόληπτων εκτιμητών, οι οποίοι είναι τετραγωνικές μορφές των παρατηρήσεων (βλ. Graybill and Hultquist, 1961). Η παραπάνω ιδιότητα ισχύει χωρίς να γίνει κάποια υπόθεση για την κατανομή των τυχαίων επιδράσεων και των όρων σφάλματος. Εάν υποθέσουμε ότι τα παραπάνω ακολουθούν κανονική κατανομή, τότε οι εκτιμητές της σχέσης (4.9) έχουν τη μικρότερη διακύμανση στην κλάση όλων των αμερόληπτων εκτιμητών. Για την τελευταία ιδιότητα γίνεται λόγος στις εργασίες των Graybill (1954) και Graybill and Wortham (1956). Ωστόσο, οι ιδιότητες αυτές ισχύουν για την balanced περίπτωση.

iii) **Αρνητικοί εκτιμητές**

Μια συνιστώσα διακύμανσης είναι εξ΄ ορισμού θετική. Ωστόσο, οι εκτιμητές που προκύπτουν από τη μέθοδο ανάλυσης διακύμανσης από τη σχέση (4.9) μπορεί να είναι αρνητικοί και δεν υπάρχει κάποιος τρόπος να αποφευχθεί αυτή η κατάσταση. Αυτό δε συμβαίνει μόνο σε μια απλή περίπτωση μοντέλου όπως αυτή του (4.1), αλλά και σε μοντέλα με περισσότερους παράγοντες, είτε με balanced είτε με unbalanced δεδομένα. Προφανώς, είναι πρόβλημα η εκτίμηση μιας θετικά ορισμένης παραμέτρου να είναι αρνητική. Υπάρχουν όμως τρόποι διευθέτησης μιας τέτοιας περίπτωσης, όπως αναφέρει ο Searle (1997, σ. 407-408) όμως λίγοι από αυτούς είναι ικανοποιητικοί. Χαρακτηριστικά αναφέρουμε έναν από αυτούς τους τρόπους, σύμφωνα με τον οποίο, εάν ένας εκτιμητής συνιστώσας διακύμανσης προκύψει αρνητικός, τότε δεχόμαστε την αρνητική τιμή του εκτιμητή ως ένδειξη ότι η πραγματική τιμή της αντίστοιχης συνιστώσας είναι μηδέν και χρησιμοποιούμε την τιμή μηδέν για τον εκτιμητή. Η παραπάνω αντικατάσταση φαίνεται λογική, ωστόσο μπαίνει εμπόδιο στις ιδιότητες των εκτιμητών, που θα προέκυπταν με τη μέθοδο της ανάλυσης διακύμανσης, για παράδειγμα παύουν να είναι αμερόληπτοι.

Μέχρι στιγμής δεν έχει γίνει η υπόθεση ότι οι όροι σφάλματος και οι τυχαίες επιδράσεις του μοντέλου (4.1) ακολουθούν κάποια συγκεκριμένη κατανομή. Όλα τα παραπάνω ισχύουν για οποιαδήποτε κατανομή. Στο σημείο αυτό θα κάνουμε την επιπλέον υπόθεση ότι η κατανομή των α_i και $e_{ij}, i = 1, \ldots, n, j = 1, \ldots, k$, είναι η κανονική (βλ. Searle, 1997, σ. 408). Με αυτή την επιπλέον υπόθεση θα εξεταστεί η κατανομή των μέσων τετραγώνων.

Στο πλαίσιο αυτό, έστω f, SS και MS οι βαθμοί ελευθερίας, τα αθροίσματα τετραγώνων και τα μέσα τετράγωνα αντίστοιχα σε μία γραμμή ενός πίνακα ανάλυσης διακύμανσης balanced δεδομένων, όπως για παράδειγμα ο πίνακας της ανάλυσης διακύμανσης 4.1 του μοντέλου (4.1). Καταλήγουμε στο εξής γενικό αποτέλεσμα (βλ. Searle, 1997, σ. 409),

$$SS/E(MS) \sim \chi_f^2 \text{ και οι } SS - \text{όροι είναι ανά δύο ανεξάρτητοι.}$$

Αφού $MS = SS/f$ προκύπτει ότι,

$$fMS/E(MS) \sim \chi_f^2 \text{ και οι } MS - \text{όροι είναι ανά δύο ανεξάρτητοι.}$$

Σύμφωνα με τον Searle (1997, σ. 409) το παραπάνω αποτέλεσμα μπορεί να προκύψει γράφοντας το πηλίκο $SS/E(MS)$ ως τετραγωνική μορφή $X'AX$, για συγκεκριμένο πίνακα A, των παρατηρήσεων X_{ij} και εφαρμόζοντας τα εξής θεωρήματα (βλ. Searle, 1997, σ. 57 και 59).

Θεώρημα 4.2.1. *Όταν $X \sim N(\mu, V)$, τότε $X'AX \sim \chi^2[r(A), \frac{1}{2}\mu' A\mu]$ εάν και μόνο εάν AV είναι ταυτοδύναμος, όπου με $r(A)$ συμβολίζεται η βαθμίδα (rank) του πίνακα A.*

Θεώρημα 4.2.2. *Όταν $X \sim N(\mu, V)$, οι τετραγωνικές μορφές $X'AX$ και $X'BX$ κατανέμονται ανεξάρτητα εάν και μόνο εάν $AVB = 0$ (ή ισοδύναμα $BVA = 0$).*

Κατανομές μέσων τετραγώνων

Σύμφωνα με τα παραπάνω θα υπολογίσουμε τις κατανομές των MSA και MSE για το μοντέλο (4.1). Έστω $X = (X_1', X_2', \ldots, X_n')'$ το διάνυσμα όλων των N παρατηρήσεων. Κάτω από την υπόθεση της κανονικότητας έχουμε ότι,

$$X \sim N_N(\mu 1_N, V), \tag{4.10}$$

όπου $N = nk$ και 1_N είναι το N-διάστατο διάνυσμα με όλες τις συνιστώσες του ίσες με τη μονάδα. Ο πίνακας διακυμάνσεων-συνδιακυμάνσεων V της σχέσης

(4.6) μπορεί να πάρει την εξής μορφή,

$$V = \sigma_e^2 I_N + \sigma_A^2 (I_n \otimes J_k), \tag{4.11}$$

όπου I_N και I_n είναι ταυτοτικοί πίνακες τάξης $N = nk$ και n αντίστοιχα και J_k ο πίνακας τάξης k με στοιχεία του τη μονάδα. Για την κατανομή του MSA έχουμε τα παρακάτω (βλ. Searle, 1997, σ. 409-410),

$$SSA = X' \left(k^{-1}(I_n \otimes J_k) - N^{-1} J_N \right) X \tag{4.12}$$

και

$$E(MSA) = k\sigma_A^2 + \sigma_e^2.$$

Επομένως,

$$SSA/E(MSA) = X'AX, \ \text{ όπου } \ A = \frac{k^{-1}(I_n \otimes J_k) - N^{-1} J_N}{k\sigma_A^2 + \sigma_e^2}.$$

Έτσι λαμβάνοντας υπόψιν ιδιότητες για το Kronecker's product (βλ. Schäke, 2013) και ιδιότητες του πίνακα J (βλ. Searle 1966, σ. 197) έχουμε ότι,

$$\begin{aligned} AV &= \left(\frac{k^{-1}(I_n \otimes J_k) - N^{-1} J_N}{k\sigma_A^2 + \sigma_e^2} \right) \left(\sigma_e^2 I_N + \sigma_A^2 (I_n \otimes J_k) \right) \\ &= \frac{\sigma_e^2 \left(k^{-1}(I_n \otimes J_k) - N^{-1} J_N \right) + \sigma_A^2 (I_n \otimes J_k) \left(k^{-1}(I_n \otimes J_k) - N^{-1} J_N \right)}{k\sigma_A^2 + \sigma_e^2} \\ &= \frac{(I_n \otimes J_k)(k^{-1}\sigma_e^2 + \sigma_A^2) - N^{-1} J_N (\sigma_e^2 + k\sigma_A^2)}{k\sigma_A^2 + \sigma_e^2} \end{aligned}$$

ή

$$AV = k^{-1}(I_n \otimes J_k) - N^{-1} J_N. \tag{4.13}$$

Ακόμη αποδεικνύεται ότι $(AV)^2 = AV$, δηλαδή AV είναι ταυτοδύναμος, όπως ακολουθεί,

$$
\begin{aligned}
(AV)^2 &= \left(k^{-1}(I_n \otimes J_k) - N^{-1}J_N\right)\left(k^{-1}(I_n \otimes J_k) - N^{-1}J_N\right) \\
&= k^{-2}(I_n \otimes kJ_k) - k^{-1}N^{-1}(I_N \otimes J_k)J_N - N^{-1}k^{-1}J_N(I_n \otimes J_k) + N^{-2}(J_N)^2 \\
&= k^{-1}(I_n \otimes J_k) - N^{-1}J_N \\
&= AV.
\end{aligned}
$$

Ακόμη, επειδή ο V είναι αντιστρέψιμος και ο AV ταυτοδύναμος ισχύει ότι,

$$r(A) = r(AV) = tr(AV) = nk\left(\frac{1}{k} - \frac{1}{N}\right) = n - 1. \tag{4.14}$$

Τέλος,

$$\frac{1}{2}(\mu 1'_N A 1_N \mu) = 0, \tag{4.15}$$

διότι $1'_N A = 0$. Επομένως, από το Θεώρημα 4.2.1, λόγω των σχέσεων (4.14) και (4.15) προκύπτει ότι,

$$SSA/E(MSA) \sim \chi^2_{n-1}. \tag{4.16}$$

ή

$$\frac{(n-1)MSA}{k\sigma_A^2 + \sigma_e^2} \sim \chi^2_{n-1}. \tag{4.17}$$

Σειρά έχει τώρα η εύρεση της κατανομής του μέσου τετραγώνου MSE. Αρχικά, όπως και προηγουμένως, πρέπει να εκφράσουμε το άθροισμα τετραγώνων των υπολοίπων SSE ως μια τετραγωνική μορφή. Έχουμε ότι (βλ. Saw, 1992),

$$SSE = X'\left(I_N - k^{-1}(I_n \otimes J_k)\right)X, \tag{4.18}$$

και

$$E(MSE) = \sigma_e^2.$$

Επομένως,

$$SSE/E(MSE) = X'BX, \text{ όπου } B = \frac{I_N - k^{-1}(I_n \otimes J_k)}{\sigma_e^2}.$$

Τότε έχουμε,

$$\begin{aligned}BV &= \frac{1}{\sigma_e^2}\left(I_N - k^{-1}(I_n \otimes J_k)\right)\left(\sigma_e^2 I_N + \sigma_A^2(I_n \otimes J_k)\right)\\ &= \frac{1}{\sigma_e^2}\left(\sigma_e^2 I_N + \sigma_A^2(I_n \otimes J_k) - k^{-1}\sigma_e^2(I_n \otimes J_k) - \sigma_A^2 k^{-1}k(I_n \otimes J_k)\right)\end{aligned}$$

ή

$$BV = I_N - k^{-1}(I_n \otimes J_k). \tag{4.19}$$

Ο BV είναι ταυτοδύναμος διότι,

$$\begin{aligned}(BV)^2 &= (I_N - k^{-1}(I_n \otimes J_k))(I_N - k^{-1}(I_n \otimes J_k))\\ &= I_N - k^{-1}(I_n \otimes J_k) - k^{-1}(I_n \otimes J_k) + k^{-2}k(I_n \otimes J_k)\\ &= I_N - k^{-1}(I_n \otimes J_k)\\ &= BV.\end{aligned}$$

Ακόμη, επειδή ο V είναι αντιστρέψιμος και ο BV ταυτοδύναμος ισχύει ότι,

$$r(B) = r(BV) = tr(BV) = nk\left(1 - \frac{1}{k}\right) = nk - n = n(k-1). \tag{4.20}$$

Τέλος,

$$\frac{1}{2}(\mu 1_N' B 1_N \mu) = 0, \tag{4.21}$$

διότι $1_N' B = 0$. Επομένως, από το Θεώρημα 4.2.1, λόγω των σχέσεων (4.20) και (4.21) προκύπτει ότι,

$$SSE/E(MSE) \sim \chi^2_{n(k-1)}, \tag{4.22}$$

ή

$$\frac{n(k-1)MSE}{\sigma_e^2} \sim \chi^2_{n(k-1)}. \tag{4.23}$$

Επιπλέον, τα μέσα τετράγωνα MSA και MSE είναι ανεξάρτητα. Αυτό συμβαίνει διότι,

$$
\begin{aligned}
AVB &\overset{(4.13)}{=} \left(k^{-1}(I_n \otimes J_k) - N^{-1}J_N\right) \frac{I_N - k^{-1}(I_n \otimes J_k)}{\sigma_e^2} \\
&= \frac{1}{\sigma_e^2}\left(k^{-1}(I_n \otimes J_k) - k^{-2}k(I_n \otimes J_k) - N^{-1}J_N + N^{-1}k^{-1}J_N(I_n \otimes J_k)\right) \\
&= 0.
\end{aligned}
$$

Από το Θεώρημα 4.2.2 προκύπτει ότι,

$$
\frac{(n-1)MSA}{k\sigma_A^2 + \sigma_e^2} \quad \text{και} \quad \frac{n(k-1)MSE}{\sigma_e^2}
$$

είναι ανεξάρτητα, άρα καταλήγουμε στο συμπέρασμα ότι και MSA, MSE είναι ανεξάρτητα μεταξύ τους.

Κατανομές των εκτιμητών

Από τη μέθοδο της ανάλυσης διακύμανσης που ακολουθήσαμε, προέκυψαν εκτιμητές των συνιστωσών διακύμανσης που είναι γραμμικές συναρτήσεις των μέσων τετραγώνων, για τις κατανομές των οποίων δόθηκε παραπάνω ένα γενικό συμπέρασμα. Επομένως, οι εκτιμητές είναι γραμμικές συναρτήσεις χ^2-μεταβλητών, κάποιες από τις οποίες έχουν αρνητικούς συντελεστές. Δεν υπάρχει κλειστή μορφή για την κατανομή γραμμικών συνδυασμών χ^2-μεταβλητών παρότι έχουν προταθεί στη βιβλιογραφία διάφορες τεχνικές προσέγγισής τους. Ακολουθώντας στο θέμα αυτό τον Searle (1997, σ. 410), για το μοντέλο (4.1) για balanced δεδομένα έχουμε για παράδειγμα ότι,

$$
\hat{\sigma}_A^2 = \frac{MSA - MSE}{k} \sim \frac{k\sigma_A^2 + \sigma_e^2}{k(n-1)}\chi^2_{(n-1)} - \frac{\sigma_e^2}{nk(k-1)}\chi^2_{n(k-1)}. \tag{4.24}
$$

Η παραπάνω σχέση προκύπτει από τις σχέσεις (4.17) και (4.23) και την ανεξαρτησία των μέσων τετραγώνων MSA και MSE. Η ακριβής κατανομή της σχέσης

(4.24) δεν μπορεί να υπολογιστεί, λόγω του ότι ο δεύτερος όρος είναι αρνητικός και επειδή σ_A^2 και σ_e^2 που συναντώνται στους συντελεστές είναι άγνωστες παράμετροι.

Ωστόσο, η κατανομή του $\hat{\sigma}_e^2$ της σχέσης (4.8) είναι πάντα γνωστή υπό την κανονικότητα. Λόγω της σχέσης (4.23) έχουμε ότι,

$$\hat{\sigma}_e^2 = MSE \sim \frac{\sigma_e^2}{n(k-1)}\chi^2_{n(k-1)}. \tag{4.25}$$

Ακολουθώντας τον Searle (1997, σ. 411), το αντίστοιχο αποτέλεσμα με αυτό της σχέσης (4.24) για μια πιο γενική περίπτωση μοντέλου, από αυτή του (4.1), δίνεται. Από τη σχέση (4.9) έχουμε ότι $\hat{\sigma}^2 = P^{-1}m$. Για την κατανομή των στοιχείων του m ισχύει το γενικό συμπέρασμα που διατυπώθηκε παραπάνω. Έτσι για παράδειγμα $M_i \sim E(M_i)f_i^{-1}\chi^2_{f_i}, \;\; i = 1, 2, \ldots, p$, όπου M_i είναι το μέσο διάνυσμα που αντιστοιχεί στην i-οστή γραμμή του πίνακα ανάλυσης διακύμανσης και f_i είναι οι βαθμοί ελευθερίας του M_i. Τώρα γράφοντας $C = diag\{f_i^{-1}\chi^2_{f_i}\}$, για $i = 1, 2, \ldots, p$, έχουμε ότι,

$$\hat{\sigma}^2 \sim P^{-1}CE(m) \sim P^{-1}CP\sigma^2, \tag{4.26}$$

με $\sim$ να συμβολίζει ότι η τυχαία ποσότητα του αριστερού μέλους έχει την ίδια κατανομή με εκείνη του δεξιού μέλους. Από τη σχέση (4.26) προκύπτει ότι οι συνιστώσες του διανύσματος των εκτιμητών είναι συναρτήσεις χ^2-μεταβλητών.

4.2.2 Εκτίμηση της intraclass συσχέτισης ρ

Επιστρέφουμε στο αρχικό πρόβλημα της εκτίμησης της intraclass συσχέτισης. Ο εκτιμητής της intraclass συσχέτισης ρ με την υιοθέτηση του μοντέλου (4.1) ονομάζεται $ANOVA$ εκτιμητής, συμβολίζεται με $\hat{\rho}_A$ και υπολογίζεται στην περίπτωση των balanced δεδομένων, σύμφωνα με τη σχέση (4.4) και λαμβάνοντας υπόψιν τις (4.7) και (4.8), από,

$$\hat{\rho}_A = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_A^2 + \hat{\sigma}_e^2} = \frac{\frac{MSA-MSE}{k}}{\frac{MSA-MSE}{k} + MSE} = \frac{MSA - MSE}{MSA + (k-1)MSE}. \qquad (4.27)$$

Στο σημείο αυτό να τονιστεί ότι ο παραπάνω εκτιμητής προέκυψε χωρίς να γίνει κάποια υπόθεση για την κατανομή των τυχαίων επιδράσεων και των όρων σφάλματος. Αναφέραμε παραπάνω ότι οι εκτιμητές των συνιστωσών διακύμανσης, που προέκυψαν από τη μέθοδο ανάλυσης διακύμανσης, μπορεί να είναι αρνητικοί. Έτσι, σε περίπτωση που $\hat{\sigma}_A^2$ προκύψει αρνητικός, τότε και ο εκτιμητής $\hat{\rho}_A$ θα είναι αρνητικός, γεγονός που είναι μη αποδεκτό, καθώς ο $\hat{\rho}_A$ εκτιμά μια εξ' ορισμού θετική παράμετρο. Στην περίπτωση αυτή, όπως προαναφέρθηκε, θεωρούμε $\hat{\rho}_A = 0$ (ή ισοδύναμα $\hat{\sigma}_A^2 = 0$), ως μία truncated μορφή του εκτιμητή (βλ. Karlin, 1981).

4.2.3 Κατανομή του εκτιμητή $\hat{\rho}_A$

Μπορούμε να υπολογίσουμε την ακριβή κατανομή του $\hat{\rho}_A$, υποθέτοντας βέβαια την κανονική κατανομή των τυχαίων επιδράσεων και των όρων σφάλματος (βλ. Donner and Koval, 1983). Γνωρίζουμε από τις σχέσεις (4.17) και (4.23) την κατανομή των μέσων τετραγώνων MSA και MSE, καθώς και το γεγονός ότι είναι ανεξάρτητα. Επομένως, η από κοινού κατανομή τους θα είναι το γινόμενο των περιθωρίων τους. Θεωρώντας τις τυχαίες μεταβλητές,

$$\hat{\rho}_A = \frac{MSA - MSE}{MSA + (k-1)MSE} \quad \text{και} \quad t = MSA + (k-1)MSE,$$

μπορούμε να βρούμε την από κοινού κατανομή των $\hat{\rho}_A$ και t. Ολοκληρώνοντας ως προς t προκύπτει η περιθώρια κατανομή του $ANOVA$ εκτιμητή η οποία δίνεται από τη σχέση,

$$f(\hat{\rho}_A) = \frac{C_\rho (1-\hat{\rho}_A)^{\frac{n(k-1)-2}{2}} [1 + (k-1)\hat{\rho}_A]^{\frac{n-3}{2}}}{\left[\frac{n-1}{1+(k-1)\rho}[1 + (k-1)\hat{\rho}_A] + \frac{n(k-1)}{1-\rho}(1-\hat{\rho}_A)\right]^{\frac{nk-1}{2}}}, \quad -\frac{1}{k-1} < \hat{\rho}_A < 1,$$

όπου
$$C_\rho = \frac{k\left(\frac{n-1}{1+(k-1)\rho}\right)^{\frac{n-1}{2}}\left(\frac{n(k-1)}{1-\rho}\right)^{\frac{n(k-1)}{2}}\Gamma\left(\frac{nk-1}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{n(k-1)}{2}\right)}.$$

Ακόμη η ασυμπτωτική διακύμανση του $\hat{\rho}_A$ δίνεται από (βλ. Fisher, 1954),

$$Var(\hat{\rho}_A) = \frac{2(1-\rho)^2[1+(k-1)\rho]^2}{k(k-1)(n-1)}. \tag{4.28}$$

Οι Donner and Koval (1983) θέλοντας να εξετάσουν την ακρίβεια της παραπάνω προσέγγισης πραγματοποίησαν προσομοιώσεις, οι οποίες τους οδήγησαν στο συμπέρασμα ότι η σχέση (4.28) αποτελεί μια πολύ καλή προσέγγιση της πραγματικής διακύμανσης του εκτιμητή για δείγματα που αποτελούνται από μέτριο πλήθος οικογενειών ($n \geq 30$).

4.2.4 Διαστήματα εμπιστοσύνης για το ρ

Στηριζόμενοι στην υπόθεση της κανονικότητας μπορούμε να βρούμε διαστήματα εμπιστοσύνης για την παράμετρο ρ. Σύμφωνα με τις σχέσεις (4.17) και (4.23) γνωρίζουμε ότι,
$$\frac{(n-1)MSA}{k\sigma_A^2+\sigma_e^2} \sim \chi^2_{n-1}$$

και
$$\frac{n(k-1)MSE}{\sigma_e^2} \sim \chi^2_{n(k-1)}.$$

Ακόμη τα παραπάνω είναι ανεξάρτητα, επομένως,

$$\frac{\frac{MSA}{(k\sigma_A^2+\sigma_e^2)}}{\frac{MSE}{\sigma_e^2}} \sim F_{n-1,n(k-1)},$$

ή

$$\frac{\sigma_e^2 F}{k\sigma_A^2+\sigma_e^2} \sim F_{n-1,n(k-1)},$$

ή

$$\frac{F}{1+k\frac{\sigma_A^2}{\sigma_e^2}} \sim F_{n-1,n(k-1)}, \tag{4.29}$$

όπου $F = \frac{MSA}{MSE}$. Αρχικά, θα αποκτήσουμε το διάστημα εμπιστοσύνης για την παράμετρο σ_A^2/σ_e^2. Σύμφωνα με τον Scheffé (1999, σ. 229) έστω ότι $1-\alpha$ είναι ο επιθυμητός βαθμός εμπιστοσύνης του διαστήματος. Διαλέγουμε $\alpha_1 \geq 0$ και $\alpha_2 \geq 0$, τέτοια ώστε $\alpha_1 + \alpha_2 = \alpha$. Συνήθως, μπορούμε να επιλέξουμε $\alpha_1 = \alpha_2 = \frac{1}{2}\alpha$ ή $\alpha_1 = \alpha, \alpha_2 = 0$. Έστω F_L και F_U τέτοια ώστε,

$$P(F_{n-1,n(k-1)} < F_L) = \alpha_1 \;\text{ και }\; P(F_{n-1,n(k-1)} > F_U) = \alpha_2.$$

Τότε,

$$P(F_L \leq F_{n-1,n(k-1)} \leq F_U) = 1-\alpha,$$

ή

$$P\left(F_L \leq \frac{F}{1+k\frac{\sigma_A^2}{\sigma_e^2}} \leq F_U\right) = 1-\alpha,$$

ή

$$P\left(\frac{1}{k}\left(\frac{F}{F_U}-1\right) \leq \frac{\sigma_A^2}{\sigma_e^2} \leq \frac{1}{k}\left(\frac{F}{F_L}-1\right)\right) = 1-\alpha.$$

Πλέον εύκολα μπορεί να αποκτηθεί ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για τον intraclass συντελεστή συσχέτισης ρ, παρατηρώντας ότι (βλ. Scheffé, 1999, σ. 231),

$$\rho = \left(1+\left(\frac{\sigma_A^2}{\sigma_e^2}\right)^{-1}\right)^{-1}. \tag{4.30}$$

Επομένως, καταλήγουμε στο συμπέρασμα ότι,

$$P\left((1+L^{-1})^{-1} \leq \rho \leq (1+R^{-1})^{-1}\right) = 1-\alpha, \tag{4.31}$$

όπου

$$L = \frac{1}{k}\left(\frac{F}{F_U}-1\right) \quad \text{και} \quad R = \frac{1}{k}\left(\frac{F}{F_L}-1\right).$$

Δηλαδή, ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για την παράμετρο ρ είναι το (L', U'), όπου $L' = (1+L^{-1})^{-1}$ και $U = (1+R^{-1})^{-1}$, όπου L και R ορίζονται όπως παραπάνω.

4.2.5 Στατιστικά τεστ

Θα κατασκευαστούν, στο εδάφιο αυτό, στατιστικά τεστ για τον έλεγχο υποθέσεων σχετικά με την πληθυσμιακή τιμή της intraclass συσχέτισης ρ στην περίπτωση των ισο-μεγεθών οικογενειών, δηλαδή της balanced περίπτωσης. Τα παρακάτω βασίζονται στους McGraw and Wong (1996). Αρχικά, θα εξετάσουμε την παρακάτω υπόθεση,

$$H_0 : \rho = 0 \quad \text{έναντι} \quad H_1 : \rho > 0.$$

Από τη σχέση (4.29) ισχύει ότι,

$$\frac{F}{1 + k\frac{\sigma_A^2}{\sigma_e^2}} \sim F_{n-1,n(k-1)}.$$

Υπό την H_0 ισχύει ότι,

$$\rho = 0 \quad \text{ή} \quad \frac{\sigma_A^2}{\sigma_e^2 + \sigma_A^2} = 0 \quad \text{ή} \quad \sigma_A^2 = 0.$$

Επομένως,

$$F = \frac{MSA}{MSE} \sim F_{n-1,n(k-1)}, \ \text{υπό την} \ H_0.$$

Απορρίπτουμε την H_0 για μεγάλες τιμές του στατιστικού F. Πιο συγκεκριμένα, απορρίπτουμε την H_0, εάν $F \geq F_{n-1,n(k-1),\alpha}$, όπου α, $0 < \alpha < 1$, είναι το επιθυμητό επίπεδο σημαντικότητας.

Ωστόσο, τα τεστ για τον έλεγχο ότι $\rho = 0$, δεν εξυπηρετούν στις περισσότερες περιπτώσεις. Για παράδειγμα, σε μία μελέτη όπου τα μέλη της οικογένειας είναι δίδυμα, υποθέτουμε εξ' αρχής μη μηδενική intraclass συσχέτιση. Επομένως, είναι πιο χρήσιμο να κατασκευαστεί ένα στατιστικό τεστ για τον έλεγχο της πιο γενικής υπόθεσης,

$$H_0 : \rho = \rho_0 \quad \text{έναντι} \quad H_1 : \rho > \rho_0,$$

όπου ρ_0 είναι μια θετική σταθερά ($\rho_0 < 1$). Γνωρίζουμε ότι,

$$\frac{\sigma_e^2 F}{k\sigma_A^2 + \sigma_e^2} \sim F_{n-1,n(k-1)}. \tag{4.32}$$

Ακόμη ισχύουν οι παρακάτω σχέσεις,

$$\sigma_e^2 = \sigma_X^2(1-\rho) \tag{4.33}$$

και

$$k\sigma_A^2 + \sigma_e^2 = \sigma_X^2[1+(k-1)\rho], \tag{4.34}$$

όπου $\sigma_X^2 = \sigma_A^2 + \sigma_e^2$ ορίζεται από την (4.2). Με χρήση των σχέσεων (4.33) και (4.34) η (4.32) γίνεται,

$$\frac{1-\rho}{1+(k-1)\rho} F \sim F_{n-1,n(k-1)}.$$

Υπό την H_0 ισχύει ότι,

$$F' = \frac{1-\rho_0}{1+(k-1)\rho_0} F \sim F_{n-1,n(k-1)}.$$

Απορρίπτουμε την H_0 για μεγάλες τιμές του στατιστικού F'. Πιο συγκεκριμένα, απορρίπτουμε την H_0 εάν $F' \geq F_{n-1,n(k-1),\alpha}$, όπου α είναι το επιθυμητό επίπεδο σημαντικότητας.

4.3 Unbalanced περίπτωση

Στην περίπτωση αυτή το μέγεθος των οικογενειών ποικίλλει από οικογένεια σε οικογένεια. Υποθέτουμε ότι υπάρχουν n οικογένειες και η i-οστή οικογένεια αποτελείται από k_i παιδιά. Δηλαδή έχουμε ότι (βλ. μοντέλο (4.1) και τις σχέσεις που ισχύουν για αυτό),

$$X_{ij} = \mu + \alpha_i + e_{ij}, \quad \begin{cases} j = 1, 2, \ldots, k_i \\ i = 1, 2, \ldots, n \end{cases} .$$

Η μέθοδος που θα εφαρμοστεί για την απόκτηση των εκτιμητών συνιστωσών διακύμανσης είναι η ίδια με αυτήν που εφαρμόστηκε για την balanced περίπτωση, η οποία είναι γνωστή ως Henderson's Method 1 (βλ. Searle, 1997, σ. 424). Για να γίνει αυτό θα πρέπει να κατασκευαστεί ο πίνακας ανάλυσης διακύμανσης για το παραπάνω μοντέλο και να υπολογιστούν οι αναμενόμενες τιμές των μέσων τετραγώνων. Ο πίνακας αυτός φαίνεται παρακάτω (βλ. Graybill, 1961, σ. 353),

Source of Variation	*d.f*	*Sum of Squares*	*Mean Square*	*Expected value of mean square*
Mean	1	$SSM = T^*_\mu$	$MSM = SSM$	-
Classes	$n-1$	$SSA = T^*_A - T^*_\mu$	$MSA = SSA/n-1$	$k_0\sigma^2_A + \sigma^2_e$
Residual error	$N-n$	$SSE = T^*_0 - T^*_A$	$MSE = SSE/N-n$	σ^2_e
Total	N	$SST = T^*_0$	-	-

Πίνακας 4.2: Ανάλυση διακύμανσης για το μοντέλο 4.1 για την unbalanced περίπτωση.

όπου $N = \sum\limits_{i=1}^{n} k_i$, $T^*_0 = \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{k_i} X^2_{ij}$, $T^*_A = \sum\limits_{i=1}^{n} \frac{X^2_{i.}}{k_i}, T^*_\mu = \frac{X^2_{..}}{N}$ με $X_{i.} = \sum\limits_{j=1}^{k_i} X_{ij}$ και $X_{..} = \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{k_i} X_{ij}, i = 1, \ldots, n, j = 1, \ldots, k_i$.

Ακόμη έχουμε ότι,

$$k_0 = \frac{N^2 - \sum_{i=1}^{n} k_i^2}{N(n-1)}.$$

Για τον υπολογισμό των $E(MSA)$ και $E(MSE)$ ακολουθούμε τον Graybill (1961, σ. 353-354). Σύμφωνα με αυτόν ισχύει ότι,

$$\begin{aligned}
E(MSE) &= \frac{1}{N-n} E\left(\sum_{i=1}^{n}\sum_{j=1}^{k_i} X_{ij}^2 - \sum_{i=1}^{n} \frac{X_{i.}^2}{k_i}\right) \\
&= \frac{1}{N-n} E\left[\sum_{i=1}^{n}\sum_{j=1}^{k_i} (X_{ij} - \bar{X}_{i.})^2\right] \\
&\overset{(4.1)}{=} \frac{1}{N-n} E\left[\sum_{i=1}^{n}\sum_{j=1}^{k_i} (\mu + \alpha_i + e_{ij} - \mu - \alpha_i - \bar{e}_{i.})^2\right] \\
&= \frac{1}{N-n} E\left[\sum_{i=1}^{n}\sum_{j=1}^{k_i} (e_{ij} - \bar{e}_{i.})^2\right],
\end{aligned}$$

όπου $\bar{X}_{i.} = \frac{1}{k_i}\sum_{j=1}^{k_i} X_{ij}$ και $\bar{e}_{i.} = \frac{1}{k_i}\sum_{j=1}^{k_i} e_{ij}$.

Στο σημείο αυτό θα χρησιμοποιήσουμε, επίσης, το παρακάτω λήμμα (βλ. Graybill, 1961, σ. 343).

Λήμμα 4.3.1. *Έστω $Z_1, Z_2, \ldots, Z_n$ ασυσχέτιστες τυχαίες μεταβλητές με μέση τιμή μ και διακύμανση σ^2. Τότε,*

$$E\left[\sum_{i=1}^{n} (Z_i - \bar{Z})^2\right] = (n-1)Var(Z_i) = (n-1)\sigma^2.$$

Χρησιμοποιώντας το Λήμμα 4.3.1 έχουμε ότι,

$$E(MSE) = \frac{1}{N-n} \sum_{i=1}^{n} \left\{ E\left[\sum_{j=1}^{k_i} (e_{ij} - \bar{e}_{i.})^2\right] \right\} = \frac{1}{N-n} \sum_{i=1}^{n} (k_i - 1)\sigma_e^2 = \sigma_e^2.$$

Για την $E(MSA)$ έχουμε ότι,

$$E(MSA) = \frac{1}{n-1} E\left(\sum_{i=1}^{n} \frac{X_{i.}^2}{k_i} - \frac{X_{..}^2}{N}\right) = \frac{1}{n-1} E\left[\sum_{i=1}^{n}\sum_{j=1}^{k_i} (\bar{X}_{i.} - \bar{X}_{..})^2\right],$$

όπου $\bar{X}_{..} = \frac{1}{N}\sum_{i=1}^{n} k_i \bar{X}_{i.} = \frac{1}{N}\sum_{i=1}^{n} k_i(\mu + \alpha_i + \bar{e}_{i.})$. Άρα έχουμε,

$$
\begin{aligned}
&E(MSA)\\
&= \frac{1}{n-1} E\left[\sum_{i=1}^{n}\sum_{j=1}^{k_i}\left(\mu + \alpha_i + \bar{e}_{i.} - \mu - \frac{1}{N}\sum_{m=1}^{n} k_m\alpha_m - \frac{1}{N}\sum_{m=1}^{n}\sum_{l=1}^{k_m} e_{ml}\right)^2\right]\\
&= \frac{1}{n-1}\left(E\left[\sum_{i=1}^{n}\sum_{j=1}^{k_i}\left(\alpha_i - \frac{1}{N}\sum_{m=1}^{n} k_m\alpha_m\right)^2\right]\right.\\
&\qquad\left. + E\left[\sum_{i=1}^{n}\sum_{j=1}^{k_i}\left(\bar{e}_{i.} - \frac{1}{N}\sum_{m=1}^{n}\sum_{l=1}^{k_m} e_{ml}\right)^2\right]\right)\\
&= \frac{1}{n-1}\left(E\left[\sum_{i=1}^{n}\sum_{j=1}^{k_i}\left(\alpha_i^2 - \frac{2}{N}\alpha_i\sum_{m=1}^{n} k_m\alpha_m + \frac{1}{N^2}\left(\sum_{m=1}^{n} k_m\alpha_m\right)^2\right)\right]\right.\\
&\qquad\left. + E\left[\sum_{i=1}^{n}\sum_{j=1}^{k_i}\left(\bar{e}_{i.}^2 - \frac{2}{N}\bar{e}_{i.}\sum_{m=1}^{n}\sum_{l=1}^{k_m} e_{ml} + \frac{1}{N^2}\left(\sum_{m=1}^{n}\sum_{l=1}^{k_m} e_{ml}\right)^2\right)\right]\right)\\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}\sum_{j=1}^{k_i}\left(\sigma_A^2 - \frac{2}{N}k_i\sigma_A^2 + \frac{\sum_{m=1}^{n} k_m^2}{N^2}\sigma_A^2\right)\right.\\
&\qquad\left. + \sum_{i=1}^{n}\sum_{j=1}^{k_i}\left(\frac{\sigma_e^2}{k_i} - \frac{2k_i\sigma_e^2}{Nk_i} + \frac{\sigma_e^2}{N}\right)\right)\\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}\left(k_i\sigma_A^2 - \frac{2k_i^2}{N}\sigma_A^2 + k_i\frac{\sum_{m=1}^{n} k_m^2}{N^2}\sigma_A^2\right) + \sum_{i=1}^{n}\left(\sigma_e^2 - \frac{k_i}{N}\sigma_e^2\right)\right)\\
&= \frac{1}{n-1}\left(\sigma_A^2\left(N - \frac{\sum_{i=1}^{n} k_i^2}{N}\right) + \sigma_e^2(n-1)\right)\\
&= \sigma_A^2\left(\frac{N^2 - \sum_{i=1}^{n} k_i^2}{N(n-1)}\right) + \sigma_e^2.
\end{aligned}
$$

Επομένως καταλήγουμε στο συμπέρασμα ότι,

$$E(MSA) = \sigma_e^2 + k_0\sigma_A^2,$$

όπου $k_0 = \frac{N^2 - \sum_{i=1}^{n} k_i^2}{N(n-1)}$.

Έστω $\hat{\sigma}_A^2$ και $\hat{\sigma}_e^2$ οι εκτιμητές των συνιστωσών διακύμανσης των παραμέτρων σ_A^2 και σ_e^2 αντίστοιχα. Οι παραπάνω εκτιμητές θα προκύψουν με την ίδια μέθοδο που εφαρμόστηκε και στη balanced περίπτωση. Δηλαδή, θα εξισώσουμε τα μέσα τετράγωνα MSA, MSE με τις αντίστοιχες αναμενόμενες τιμές τους. Έτσι, έχουμε τα παρακάτω αποτελέσματα (βλ. Searle 1997, σ. 437-438),

$$\begin{cases} E(MSA) = MSA \\ E(MSE) = MSE \end{cases} \quad \text{ή} \quad \begin{cases} k_0\sigma_A^2 + \sigma_e^2 = MSA \\ \sigma_e^2 = MSE \end{cases} \quad \text{ή} \quad \begin{cases} k_0\sigma_A^2 = MSA - MSE \\ \sigma_e = MSE \end{cases}$$

Τελικά, οι εκτιμητές των διακυμάνσεων σ_A^2 και σ_e^2 δίνονται από τις σχέσεις,

$$\hat{\sigma}_A^2 = \frac{MSA - MSE}{k_0}, \tag{4.35}$$

$$\hat{\sigma}_e^2 = MSE. \tag{4.36}$$

4.3.1 Ιδιότητες των εκτιμητών $\hat{\sigma}_A^2$ και $\hat{\sigma}_e^2$

i) **Αμεροληψία**

Οι εκτιμητές των σχέσεων (4.35) και (4.36) είναι αμερόληπτοι. Αυτό είναι προφανές αφού,

$$E(\hat{\sigma}_A^2) = \frac{1}{k_0}(E(MSA) - E(MSE)) = \frac{1}{k_0}(k_0\sigma_A^2 + \sigma_e^2 - \sigma_e^2) = \sigma_A^2$$

και

$$E(\hat{\sigma}_e^2) = E(MSE) = \sigma_e^2.$$

Ωστόσο στο σημείο αυτό να τονιστεί ότι, όπως αναφέρει και ο Searle (1997, σ. 406), η ιδιότητα της αμεροληψίας των εκτιμητών συνιστωσών διακύμανσης, που προέκυψαν από την παραπάνω μέθοδο (Henderson's Method 1) γενικά δεν ισχύει. Δηλαδή, ακόμη και η πιο απλή ιδιότητα ενός εκτιμητή, αυτή της αμεροληψίας, δεν ισχύει γενικά για τους εκτιμητές συνιστωσών διακύμανσης, που προκύπτουν από την παραπάνω μέθοδο για άλλες περιπτώσεις μοντέλων. Παρ΄ όλα αυτά η ιδιότητα της αμεροληψίας των εκτιμητών $\hat{\sigma}_A^2$ και $\hat{\sigma}_e^2$ ικανοποιείται με την υιοθέτηση του μοντέλου (4.1).

ii) **Ελάχιστη διακύμανση**

Είδαμε ότι για τη balanced περίπτωση, οι εκτιμητές με τη μικρότερη διακύμανση στην κλάση των αμερόληπτων εκτιμητών, που είναι τετραγωνικές μορφές των παρατηρήσεων, ταυτίζονται με τους εκτιμητές που προκύπτουν από τη μέθοδο της ανάλυσης διακύμανσης. Ωστόσο, αυτή η ιδιότητα δεν ισχύει για την περίπτωση των unbalanced δεδομένων. Ο Searle (1997, σ. 406) αναφέρει πως συζητήσεις για παρόμοιες ιδιότητες όπως αυτών της αμεροληψίας και της ελάχιστης διακύμανσης για εκτιμητές από unbalanced δεδομένα (όχι απαραίτητα με χρήση της μεθόδου ανάλυσης διακύμανσης) περιορίζονται στο μοντέλο (4.1) και παραπέμπει στους Townsend (1968) και Harville (1969). Ακόμη, οι Townsend and Searle (1971) έχουν υπολογίσει τοπικά αμερόληπτους εκτιμητές που έχουν την μικρότερη διακύμανση και είναι ταυτόχρονα τετραγωνικές μορφές των παρατηρήσεων, για την περίπτωση του μοντέλου (4.1) με $\mu = 0$ και πρότειναν και αντίστοιχους προσεγγιστικούς εκτιμητές και για την περίπτωση όπου $\mu \neq 0$ (βλ. Searle, 1997, σ. 470).

4.3.2 Εκτίμηση της intraclass συσχέτισης ρ

Λαμβάνοντας υπόψη τις σχέσεις (4.35) και (4.36) ο εκτιμητής ANOVA $\hat{\rho}_A$, για την unbalanced περίπτωση ορίζεται σύμφωνα με τη σχέση (4.4) ως,

$$\hat{\rho}_A = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_A^2 + \hat{\sigma}_e^2} = \frac{\frac{MSA-MSE}{k_0}}{\frac{MSA-MSE}{k_0} + MSE} = \frac{MSA - MSE}{MSA + (k_0 - 1)MSE}, \qquad (4.37)$$

όπου $k_0 = \frac{N^2 - \sum_{i=1}^{n} k_i^2}{N(n-1)}$ με $N = \sum\limits_{i=1}^{n} k_i$.
Η ασυμπτωτική διακύμανση του $\hat{\rho}_A$ υπολογίστηκε από τον Smith (1957) υπό την υπόθεση της κανονικότητας και δίνεται από την παρακάτω σχέση (βλ. Donner, 1986),

$$\begin{aligned} Var(\hat{\rho}_A) =& \frac{2(1-\rho)^2}{k_0^2} \left\{ \frac{[1+\rho(k_0-1)]^2}{N-n} + \frac{(n-1)(1-\rho)[1+\rho(2k_0-1)]}{(n-1)^2} \right. \\ & \left. + \frac{\rho^2 \left[\sum_{i=1}^{n} k_i^2 - 2N^{-1} \sum_{i=1}^{n} k_i^3 + N^{-2} \left(\sum_{i=1}^{n} k_i^2 \right)^2 \right]}{(n-1)^2} \right\}. \end{aligned}$$

4.3.3 Διαστήματα εμπιστοσύνης

Είδαμε ότι στην περίπτωση των balanced δεδομένων βρέθηκε ένα ακριβές διάστημα εμπιστοσύνης για την παράμετρο ρ (βλ. σχέση (4.31)). Ο Wald (1940) ανέπτυξε μια διαδικασία για την κατασκευή ενός ακριβούς διαστήματος εμπιστοσύνης για την παράμετρο $\frac{\sigma_A^2}{\sigma_e^2}$. Λαμβάνοντας υπόψη ότι η παράμετρος της intraclass συσχέτισης ρ είναι μονότονη συνάρτηση του $\frac{\sigma_A^2}{\sigma_e^2}$ (βλ. σχέση 4.30)), ισοδύναμα μπορεί να προκύψει ένα ακριβές διάστημα εμπιστοσύνης για την παράμετρο ρ. Στη συνέχεια ακολουθούμε την εργασία του Wald (1940).

Έστω ότι το παρακάτω μοντέλο ισχύει,

$$X_{ij} = \alpha_i + e_{ij}, \tag{4.38}$$

όπου $e_{ij},\ \ i = 1, 2, \ldots, n, j = 1, 2, \ldots, k_i$, έχουν κανονική κατανομή με μέση τιμή μ_e και διακύμανση σ_e^2 και $\alpha_i,\ i = 1, 2, \ldots, n$, έχουν κανονική κατανομή με μέση τιμή μ_A και διακύμανση σ_A^2. Ακόμη, υποθέτουμε ότι όλες οι μεταβλητές e_{ij} και α_i είναι ανεξάρτητες. Να τονιστεί ότι τα μοντέλα (4.1) και (4.38), προφανώς, είναι ισοδύναμα. Συμβολίζουμε με $\bar{X}_i$ τον αριθμητικό μέσο της i-οστής οικογένειας.

Έτσι έχουμε,

$$\bar{X}_i = \frac{1}{k_i}\sum_{j=1}^{k_i} X_{ij} \overset{(4.38)}{=} \frac{1}{k_i}\sum_{j=1}^{k_i}(\alpha_i + e_{ij}) = \alpha_i + \frac{1}{k_i}\sum_{j=1}^{k_i} e_{ij}. \tag{4.39}$$

Έτσι, η διακύμανση του $\bar{X}_i$ δίνεται από,

$$Var(\bar{X}_i) = \sigma^2_{\bar{X}_i} = Var(\alpha_i) + \frac{1}{k_i^2}\sum_{j=1}^{k_i} Var(e_{ij}) = \sigma^2_A + \frac{1}{k_i}\sigma^2_e. \tag{4.40}$$

Θέτουμε $\theta^2 = \frac{\sigma^2_A}{\sigma^2_e}$. Τότε έχουμε,

$$\sigma^2_{\bar{X}_i} = \sigma^2_e\left(\frac{1}{k_i} + \frac{\sigma^2_A}{\sigma^2_e}\right) = \sigma^2_e\left(\frac{1}{k_i} + \theta^2\right) = \frac{\sigma^2_e}{w_i}, \tag{4.41}$$

όπου

$$w_i = \frac{k_i}{1 + k_i\theta^2}.$$

Θα αποδείξουμε, ακολουθώντας τον Wald (1940), ότι,

$$\frac{1}{\sigma^2_e}\sum_{i=1}^{n}\left[w_i\left(\bar{X}_i - \frac{\sum_{i=1}^n w_i\bar{X}_i}{\sum_{i=1}^n w_i}\right)^2\right], \tag{4.42}$$

ακολουθεί μια χ^2-κατανομή με $n-1$ βαθμούς ελευθερίας. Για το σκοπό αυτό έστω,

$$Y_i = \sqrt{w_i}\bar{X}_i, \quad i = 1, 2, \ldots, n.$$

Θεωρούμε τον παρακάτω ορθογώνιο μετασχηματισμό,

$$\begin{aligned}
Y'_1 &= L_1(Y_1, Y_2, \ldots, Y_n),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots,\\
Y'_{n-1} &= L_{n-1}(Y_1, Y_2, \ldots, Y_n),\\
Y'_n &= L_n(Y_1, Y_2, \ldots, Y_n) = \frac{\sqrt{w_1}Y_1 + \cdots + \sqrt{w_n}Y_n}{\sqrt{w_1 + \cdots + w_n}},
\end{aligned}$$

όπου $L_1(Y_1,\ldots,Y_n),\ldots,L_{n-1}(Y_1,\ldots,Y_n)$ είναι αυθαίρετες ομογενείς γραμμικές συναρτήσεις, με το μόνο περιορισμό ότι ο μετασχηματισμός θα πρέπει να είναι ορθογώνιος. Ισχύουν οι παρακάτω σχέσεις,

$$E(Y_i) = \sqrt{w_i}E(\bar{X}_i) = \sqrt{w_i}(\mu_A + \mu_e) \tag{4.43}$$

και

$$Var(Y_i) = w_i Var(\bar{X}_i) = w_i \frac{\sigma_e^2}{w_i} = \sigma_e^2. \tag{4.44}$$

Τότε, λόγω του ότι ο μετασχηματισμός είναι ορθογώνιος ισχύει ότι,

$$E(Y_i') = 0,\ \ i = 1,2,\ldots,n-1, \tag{4.45}$$

$$Var(Y_i') = \sigma_e^2,\ \ i = 1,2,\ldots,n. \tag{4.46}$$

Προκειμένου να αποδείξουμε τον ισχυρισμό μας, αρκεί να δείξουμε ότι η ποσότητα της σχέσης (4.42) είναι ίση με $\frac{1}{\sigma_e^2}(Y_1'^2+\cdots+Y_{n-1}'^2)$. Έχουμε ότι: $\bar{X}_i = \frac{Y_i}{\sqrt{w_i}},\ \ i = 1,2,\ldots,n$. Αντικαθιστούμε στην σχέση (4.42), οπότε,

$$\begin{aligned}
\frac{1}{\sigma_e^2}\sum_{i=1}^{n}\left[w_i\left(\bar{X}_i - \frac{\sum_{i=1}^{n} w_i\bar{X}_i}{\sum_{i=1}^{n} w_i}\right)^2\right] &= \frac{1}{\sigma_e^2}\sum_{i=1}^{n}\left[w_i\left(\frac{Y_i}{\sqrt{w_i}} - \frac{\sum_{i=1}^{n} w_i\frac{Y_i}{\sqrt{w_i}}}{\sum_{i=1}^{n} w_i}\right)^2\right] \\
&= \frac{1}{\sigma_e^2}\sum_{i=1}^{n}\left[w_i\left(\frac{Y_i^2}{w_i} - 2\frac{Y_i}{\sqrt{w_i}}\frac{\sum_{i=1}^{n}\sqrt{w_i}Y_i}{\sum_{i=1}^{n} w_i} + \left(\frac{\sum_{i=1}^{n}\sqrt{w_i}Y_i}{\sum_{i=1}^{n} w_i}\right)^2\right)\right] \\
&= \frac{1}{\sigma_e^2}\left[\sum_{i=1}^{n} Y_i^2 - 2\frac{\left(\sum_{i=1}^{n}\sqrt{w_i}Y_i\right)^2}{\sum_{i=1}^{n} w_i} + \frac{\left(\sum_{i=1}^{n}\sqrt{w_i}Y_i\right)^2}{\sum_{i=1}^{n} w_i}\right] \\
&= \frac{1}{\sigma_e^2}\left[\sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n}\sqrt{w_i}Y_i\right)^2}{\sum_{i=1}^{n} w_i}\right] = \frac{1}{\sigma_e^2}\left[\sum_{i=1}^{n} Y_i^2 - Y_n'^2\right] \\
&= \frac{1}{\sigma_e^2}\left[\sum_{i=1}^{n} Y_i'^2 - Y_n'^2\right] = \frac{1}{\sigma_e^2}(Y_1'^2 + \cdots + Y_{n-1}'^2).
\end{aligned}$$

Επομένως αποδείχθηκε ο ισχυρισμός ότι η ποσότητα της σχέσης (4.42) ακολουθεί μια χ^2 κατανομή με $n-1$ βαθμούς ελευθερίας, δηλαδή έχουμε ότι,

$$\frac{1}{\sigma_e^2}\sum_{i=1}^{n}\left[w_i\left(\bar{X}_i - \frac{\sum_{i=1}^{n} w_i\bar{X}_i}{\sum_{i=1}^{n} w_i}\right)^2\right] \sim \chi_{n-1}^2. \tag{4.47}$$

Επιπλέον ισχύει ότι,

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{k_i}(X_{ij}-\bar{X}_i)^2}{\sigma_e^2} \sim \chi^2_{N-n}, \tag{4.48}$$

όπου $N=\sum_{i=1}^{n} k_i$. Λόγω ανεξαρτησίας των ποσοτήτων των σχέσεων (4.47) και (4.48) καταλήγουμε στο συμπέρασμα ότι,

$$F=\frac{(N-n)\sum_{i=1}^{n}\left[w_i\left(\bar{X}_i-\frac{\sum_{i=1}^{n}w_i\bar{X}_i}{\sum_{i=1}^{n}w_i}\right)^2\right]}{(n-1)\sum_{i=1}^{n}\sum_{j=1}^{k_i}(X_{ij}-\bar{X}_i)^2} \sim F_{n-1,N-n}. \tag{4.49}$$

Έστω F_1 και F_2 το κατώτερο και ανώτερο όριο εμπιστοσύνης για το στατιστικό F της σχέσης (4.49). Αρχικά, πρέπει να δείξουμε ότι το σύνολο των τιμών της παραμέτρου $\theta^2=\frac{\sigma_A^2}{\sigma_e^2}$, για το οποίο το στατιστικό F βρίσκεται μεταξύ των ορίων εμπιστοσύνης F_1 και F_2 είναι διάστημα. Για το σκοπό αυτό αρκεί να δείξουμε ότι,

$$f(\theta^2)=\sum_{i=1}^{n}\left[w_i\left(\bar{X}_i-\frac{\sum_{i=1}^{n}w_i\bar{X}_i}{\sum_{i=1}^{n}w_i}\right)^2\right], \tag{4.50}$$

είναι φθίνουσα συνάρτηση ως προς θ^2. Πράγματι,

$$\begin{aligned}\frac{\partial f(\theta^2)}{\partial\theta^2}=&\sum_{i=1}^{n}\frac{\partial w_i}{\partial\theta^2}\left(\bar{X}_i-\frac{\sum_{i=1}^{n}w_i\bar{X}_i}{\sum_{i=1}^{n}w_i}\right)^2-2\frac{\partial}{\partial\theta^2}\left(\frac{\sum_{i=1}^{n}w_i\bar{X}_i}{\sum_{i=1}^{n}w_i}\right)\\&\times\left[\sum_{i=1}^{n}w_i\left(\bar{X}_i-\frac{\sum_{i=1}^{n}w_i\bar{X}_i}{\sum_{i=1}^{n}w_i}\right)\right].\end{aligned}$$

Επειδή,

$$\sum_{i=1}^{n}w_i\left(\bar{X}_i-\frac{\sum_{i=1}^{n}w_i\bar{X}_i}{\sum_{i=1}^{n}w_i}\right)=0,$$

έχουμε ότι,

$$\begin{aligned}\frac{\partial f(\theta^2)}{\partial\theta^2}&=\sum_{i=1}^{n}\frac{\partial w_i}{\partial\theta^2}\left(\bar{X}_i-\frac{\sum_{i=1}^{n}w_i\bar{X}_i}{\sum_{i=1}^{n}w_i}\right)^2\\&=\sum_{i=1}^{n}-w_i^2\left(\bar{X}_i-\frac{\sum_{i=1}^{n}w_i\bar{X}_i}{\sum_{i=1}^{n}w_i}\right)^2<0.\end{aligned}$$

Επομένως, το κατώτερο όριο εμπιστοσύνης, θ_1^2 της παραμέτρου θ^2 δίνεται από τη ρίζα, ως προς θ^2, της εξίσωσης,

$$F = F_2, \tag{4.51}$$

ενώ το ανώτερο όριο εμπιστοσύνης θ_2^2 του θ^2 δίνεται από τη λύση της εξίσωσης,

$$F = F_1. \tag{4.52}$$

Επειδή $f(\theta^2)$ είναι φθίνουσα, οι παραπάνω εξισώσεις έχουν το πολύ μια ρίζα ως προς θ^2. Εάν η (4.51) ή η (4.52) δεν έχει ρίζα, το αντίστοιχο όριο εμπιστοσύνης ισούται με 0. Εάν ούτε η εξίσωση (4.51) ούτε η εξίσωση (4.52) έχουν λύση τότε πρέπει να απορρίψουμε τουλάχιστον μια από τις παρακάτω υποθέσεις,

1) $X_{ij} = \alpha_i + e_{ij}$.

2) Οι μεταβλητές a_i και e_{ij} $(i = 1, 2, \ldots, n, j = 1, 2, \ldots, k_i)$ είναι κανονικές και ανεξάρτητες.

3) Κάθε μία από τις μεταβλητές e_{ij} έχει την ίδια κατανομή.

4) Κάθε μία από τις μεταβλητές α_i έχει την ίδια κατανομή.

Όπως αναφέρει ο Wald (1940) οι εξισώσεις (4.51) και (4.52) είναι πολύπλοκες αλγεβρικές εξισώσεις ως προς θ^2. Για τον υπολογισμό των ριζών αυτών των εξισώσεων, γνωστές μέθοδοι προσέγγισης μπορούν να εφαρμοστούν, αξιοποιώντας και το γεγονός ότι το στατιστικό F είναι μονότονη συνάρτηση του θ^2. Για την εφαρμογή οποιασδήποτε μεθόδου προσέγγισης είναι χρήσιμο να ξεκινάμε με δύο όρια της ρίζας που δεν απέχουν μεταξύ τους. Ο Wald (1940) περιγράφει μια μέθοδο για την εύρεση τέτοιων ορίων. Ακολουθώντας τον Wald (1940) συμβολίζουμε με $\bar{F}$ τη συνάρτηση που προκύπτει από την F (βλ. (4.49)) ανικαθιστώντας

$$\bar{w}_i = \frac{l_i}{1 + l_i\theta^2}, \;\; \text{στη θέση του} \;\; w_i, \;\; i = 1, 2, \ldots, n.$$

Ακόμη, έστω $\bar{f}$ (βλ. (4.50)) είναι η συνάρτηση που αποκτούμε από την f με την ίδια διαδικασία. Συμβολίζουμε με $\phi(\kappa, \theta^2)$ τη συνάρτηση που αποκτούμε από την $\bar{F}$, αντικαθιστώντας το κ στη θέση των $l_1, \ldots, l_n$. Πρώτα θα δείξουμε ότι η $\bar{F}$ είναι μη-φθίνουσα, καθώς τα l_p, $p = 1, 2, \ldots, n$, αυξάνονται, δηλαδή ότι $\frac{\partial \bar{F}}{\partial l_p} \geq 0$. Αρκεί να δείξουμε ότι $\frac{\partial \bar{f}}{\partial l_p} \geq 0$. Έχουμε ότι,

$$\begin{aligned}
\frac{\partial \bar{f}}{\partial l_p} &= \sum_{i=1}^{n} \frac{\partial \bar{w}_i}{\partial l_p} \left(\bar{X}_i - \frac{\sum_{i=1}^{n} \bar{w}_i \bar{X}_i}{\sum_{i=1}^{n} \bar{w}_i} \right)^2 \\
&\qquad - 2 \frac{\partial}{\partial l_p} \left(\frac{\sum_{i=1}^{n} \bar{w}_i \bar{X}_i}{\sum_{i=1}^{n} \bar{w}_i} \right) \left[\sum_{i=1}^{n} \bar{w}_i \left(\bar{X}_i - \frac{\sum_{i=1}^{n} \bar{w}_i \bar{X}_i}{\sum_{i=1}^{n} \bar{w}_i} \right) \right] \\
&= \sum_{i=1}^{n} \frac{\partial \bar{w}_i}{\partial l_p} \left(\bar{X}_i - \frac{\sum_{i=1}^{n} \bar{w}_i \bar{X}_i}{\sum_{i=1}^{n} \bar{w}_i} \right)^2 \\
&= \frac{1}{(1 + l_p \theta^2)^2} \left(\bar{X}_p - \frac{\sum_{i=1}^{n} \bar{w}_i \bar{X}_i}{\sum_{i=1}^{n} \bar{w}_i} \right)^2 \geq 0.
\end{aligned}$$

Συμβολίζουμε με k' τη μικρότερη και με k'' τη μεγαλύτερη αντίστοιχα από τις τιμές $k_1, k_2, \ldots, k_n$. Τότε έχουμε ότι,

$$\phi(k', \theta^2) \leq F \leq \phi(k'', \theta^2). \tag{4.53}$$

Ακολουθώντας τον Wald (1940), συμβολίζουμε με $\theta_1'^2, \theta_1''^2, \theta_2'^2, \theta_2''^2$ τις λύσεις ως προς θ^2 των παρακάτω εξισώσεων αντίστοιχα,

$$\begin{aligned}
\phi(k', \theta^2) &= F_2, \\
\phi(k'', \theta^2) &= F_2, \\
\phi(k', \theta^2) &= F_1, \\
\phi(k'', \theta^2) &= F_1.
\end{aligned}$$

Επειδή, η F είναι φθίνουσα ως προς θ^2, λαμβάνοντας υπόψιν τις σχέσεις (4.51), (4.52) και (4.53) καταλήγουμε στις παρακάτω σχέσεις,

$$\theta_1'^2 \leq \theta_1^2 \leq \theta_1''^2$$

και

$$\theta_2'^2 \leq \theta_2^2 \leq \theta_2''^2.$$

Οι παραπάνω σχέσεις δίνουν τα επιθυμητά διαστήματα των λύσεων.

Ωστόσο, όπως αναφέρει και ο Donner (1986), λόγω της υπολογιστικής απλότητάς τους, προσεγγιστικές μέθοδοι εφαρμόζονται συχνά στη πράξη για την κατασκευή διαστημάτων εμπιστοσύνης για την παράμετρο της intraclass συσχέτισης ρ στην περίπτωση των unbalanced δεδομένων. Αρκετοί συγγραφείς (βλ. Thomas and Hultquist, 1978, Donner, 1979) έχουν προτείνει ότι επαρκή διαστήματα εμπιστοσύνης για κάποιους unbalanced σχεδιασμούς μπορούν να αποκτηθούν χρησιμοποιώντας σχέσεις της balanced περίπτωσης, αντικαθιστώντας τον όρο k με k_0 (βλ. Donner and Wells, 1986). Δηλαδή με την παραπάνω αντικατάσταση ένα προσεγγιστικό $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για την intraclass συσχέτιση ρ, είναι το (βλ. σχέση (4.31)),

$$\left(\left(1+L_0^{-1}\right)^{-1}, \left(1+R_0^{-1}\right)^{-1}\right), \tag{4.54}$$

όπου

$$L_0 = \frac{1}{k_0}\left(\frac{F}{F_U}-1\right), \quad R_0 = \frac{1}{k_0}\left(\frac{F}{F_L}-1\right) \quad \text{με} \quad k_0 = \frac{N^2-\sum_{i=1}^{n} k_i^2}{N(n-1)}.$$

Το παραπάνω διάστημα εμπιστοσύνης είναι προσεγγιστικό επειδή το $\frac{SSA}{k_0\sigma_A^2+\sigma_e^2}$ ή ισοδύναμα $\frac{(n-1)MSA}{k_0\sigma_A^2+\sigma_e^2}$ δεν ακολουθεί χ^2 κατανομή με $n-1$ βαθμούς ελευθερίας. Αυτό συμβαίνει εάν και μόνο εάν $\sigma_A^2=0$ ή $\rho=0$. Αυτό έχει ως αποτέλεσμα το στατιστικό $\frac{F}{1+k_0\frac{\sigma_A^2}{\sigma_e^2}}$ να μην έχει μια κεντρική F κατανομή, όπως συμβαίνει στην balanced περίπτωση.

Ωστόσο, οι Thomas and Hultquist (1978) πρότειναν μια καλύτερη προσέγγιση από αυτή της σχέσης (4.54) και η οποία έχει ένα ευρύτερο πεδίο εφαρμογών σε one-way σχεδιασμούς. Έστω ότι το μοντέλο (4.1) ισχύει με την επιπλέον υπόθεση της κανονικότητας των τυχαίων επιδράσεων α_i και των όρων σφάλματος e_{ij}. Το μοντέλο γράφεται σε μορφή πίνακα ως εξής,

$$X = \mu j_N + ZA + E, \tag{4.55}$$

όπου j_N είναι ένα διάνυσμα μήκους N με στοιχεία του τη μονάδα, $A \sim N_n(0_n, \sigma_A^2 I_n)$ είναι ένα διάνυσμα μήκους n με στοιχεία του τις τυχαίες επιδράσεις α_i, $E \sim N_N(0_N, \sigma_e^2 I_N)$ είναι διάνυσμα μήκους N με στοιχεία του τους όρους σφάλματος e_{ij} και τέλος Z είναι ένας $N \times n$ πίνακας (γνωστός). Τότε έχουμε ότι $E(X) = \mu j_N$ και $Var(X) = \sigma_A^2 ZZ' + \sigma_e^2 I_N$.

Το πρώτο βήμα είναι να εφαρμόσουμε το μετασχηματισμό $R = \begin{pmatrix} R_1 \\ R_2 \end{pmatrix}$, όπου $R_1 = diag\left((1/k_1)j'_{k_1}, \ldots, (1/k_n)j'_{k_n}\right) = (Z'Z)^{-1}Z'$ και R_2 αποτελείται από $N-n$ ορθοκανονικές γραμμές στο ορθογώνιο συμπλήρωμα του γραμμικού χώρου του $(Z'Z)^{-1}Z'$. Επομένως, $E(R_1X) = R_1\mu j_N = \mu j_n$, $E(R_2X) = 0_{N-n}$ και

$$Var\begin{pmatrix} R_1X \\ R_2X \end{pmatrix} = \sigma_A^2 RZZ'R' + \sigma_e^2 RR' = \sigma_A^2 \begin{pmatrix} I_{n\times n} & 0_{n\times(N-n)} \\ 0_{(N-n)\times n} & 0_{(N-n)\times(N-n)} \end{pmatrix}$$
$$+ \sigma_e^2 \begin{pmatrix} (Z'Z)^{-1} & 0_{n\times(N-n)} \\ 0_{(N-n)\times n} & I_{(N-n)\times(N-n)} \end{pmatrix},$$

και έτσι $X'R_2'R_2X \sim \chi^2_{(N-n)}$ και R_1X, R_2X είναι ανεξάρτητα.

Στη συνέχεια σχηματίζουμε τον $n\times n$ ορθογώνιο πίνακα H, όπου $H = \begin{pmatrix} h_1 \\ H_2 \end{pmatrix}$, με $h_1 = (1/\sqrt{n})j'_n$ και H_2 έχει $n-1$ ορθοκανονικές γραμμές στο ορθογώνιο συμπλήρωμα του h_1. Τότε έχουμε ότι,

$$E(HR_1X) = E(Hj_n) = (\mu\sqrt{n}, 0'_{n-1})'$$

και

$$Var(HR_1X) = \sigma_A^2 I_n + \sigma_e^2 H(Z'Z)^{-1}H'.$$

Για το H_2R_1X έχουμε ότι $E(H_2R_1X) = 0_{n-1}$ και $Var(H_2R_1X) = \sigma_A^2 I_{n-1} + \sigma_e^2 H_2(Z'Z)^{-1}H_2'$. Έστω P ένας $(n-1)\times(n-1)$ ορθογώνιος πίνακας που διαγωνοποιεί τον $H_2(Z'Z)^{-1}H_2'$. Έστω δηλαδή ότι έχουμε $PH_2(Z'Z)^{-1}H_2'P' = \Lambda = diag(\lambda_1, \ldots, \lambda_{n-1})$, όπου $\lambda_1, \ldots, \lambda_{n-1}$ είναι οι ιδιοτιμές του $H_2(Z'Z)^{-1}H_2'$.

Τώρα έχουμε, $PH_2R_1X \sim N_{n-1}(0_{n-1}, \sigma_A^2 I_{n-1} + \sigma_e^2\Lambda)$ με PH_2R_1X και R_2X να είναι ανεξάρτητα. Συμβολίζοντας με P_i την i-οστή γραμμή του πίνακα P, τότε,

$$\upsilon_i = P_iH_2R_1X \sim N(0, \sigma_A^2 + \lambda_i\sigma_e^2), \quad i = 1, 2, \ldots, n-1. \tag{4.56}$$

Επομένως έχουμε ότι,

$$\frac{\upsilon_i^2}{\sigma_A^2 + \lambda_i\sigma_e^2} \sim \chi_1^2 \quad \text{και} \quad W^* = \sum_{i=1}^{n-1} \frac{\upsilon_i^2}{\sigma_A^2 + \lambda_i\sigma_e^2} \sim \chi_{n-1}^2. \tag{4.57}$$

Ακόμη, $tr(H_2(Z'Z)^{-1}H_2') = tr((Z'Z)^{-1}H_2'H_2) = tr((Z'Z)^{-1}[I_n - J_n/n]) = \sum_{i=1}^{n}(1/k_i) - (1/n)j'(Z'Z)^{-1}j = ((n-1)/n)\sum_{i=1}^{n}\frac{1}{k_i}$. Έτσι, $\bar{\lambda} = \sum_{i=1}^{n}\lambda_i/(n-1) = (1/n)\sum_{i=1}^{n}(1/k_i)$ και γενικά $\bar{\lambda} = 1/\widetilde{n}$, όπου $\widetilde{n} = n/\sum_{i=1}^{n}(1/k_i)$ είναι ο αρμονικός μέσος των k_i.

Παρατηρώντας ότι $\bar{\lambda} = \frac{1}{\widetilde{n}}$ συμπεραίνουμε ότι τα λ_i είναι συνήθως μικρά και φαίνεται ότι το W^*, συνήθως, θα διαφέρει λίγο από το,

$$W = \sum_{i=1}^{n-1} \frac{\upsilon_i^2}{\sigma_A^2 + \bar{\lambda}\sigma_e^2}. \tag{4.58}$$

Οι Thomas και Hultquist (1978), υποστηρίζουν ότι η τυχαία μεταβλητή W έχει προσεγγιστικά μια χ^2 κατανομή με $n-1$ βαθμούς ελευθερίας. Αφού P είναι ορθογώνιος πίνακας έχουμε ότι,

$$\sum_{i=1}^{n-1}\upsilon_i^2 = \sum_{i=1}^{n-1} X'R_1'H_2'P_i'P_iH_2R_1X = X'R_1'H_2'H_2R_1X. \tag{4.59}$$

Παρατηρούμε από τη σχέση (4.59) ότι δεν είναι απαραίτητο να υπολογίσουμε τον πίνακα P, για να βρούμε το W. Επιπλέον,

$$X'R_1'H_2'H_2R_1X = \sum_{i=1}^{n} \bar{X}_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} \bar{X}_i\right)^2 = (n-1)S_{\bar{X}}^2,$$

όπου $S_{\bar{X}}^2$ είναι η δειγματική διακύμανση των $\bar{X}_i = \frac{1}{k_i}\sum_{j=1}^{k_i} X_{ij}$. Δηλαδή αντικαθιστώντας στη σχέση (4.58) έχουμε ότι,

$$W = \frac{(n-1)S_{\bar{X}}^2}{\sigma_A^2 + \bar{\lambda}\sigma_e^2}. \tag{4.60}$$

Μπορεί να βρεθεί η ακριβής κατανομή του W, δοθέντων τιμών για τις παραμέτρους σ_A^2 και σ_e^2, χρησιμοποιώντας ένα θεώρημα των Robbins and Pitman (1949). Έστω $\lambda_m = \min_{1\leq i\leq n-1} \lambda_i$ και $b = \frac{\sigma_A^2+\lambda_m\sigma_e^2}{\sigma_A^2+\bar{\lambda}\sigma_e^2}$. Η αθροιστική συνάρτηση κατανομής του W δίνεται από,

$$P(W < w) = \sum_{j=0}^{\infty} c_j F_{n-1+2j}(w/b),$$

όπου $F_\upsilon(x)$ είναι η αθροιστική συνάρτηση της χ^2 κατανομής,

$$c_j = \sum_{i_1+\cdots+i_r=j} c_{1,i_1}\cdot c_{2,i_2}\cdots c_{r,i_r},$$

με

$$c_{i,j} = b_i^{-1/2}\frac{\frac{1}{2}\left(\frac{1}{2}+1\right)\cdots\left(\frac{1}{2}+j-1\right)}{j!}\left[1-\frac{1}{b_i}\right]^j,$$

$c_{i,0} = b_i^{-1/2}$, $i = 1, 2, \ldots, r$ και $b_j = (\sigma_A^2 + \lambda_i\sigma_e^2)/(\sigma_A^2 + \lambda_m\sigma_e^2)$.

Οι Thomas and Hultquist (1978) με προσομοιώσεις, αλλά όχι εκτεταμένες, κατέληξαν στο συμπέρασμα ότι η κατανομή του W προσεγγίζεται ικανοποιητικά από την χ^2 κατανομή. Ακόμη ισχύει ότι,

$$R_2X \sim N_{N-n}(0_{N-n}, \sigma_e^2 I_{N-n}), \quad X'R_2'R_2X/\sigma_e^2 = SSE/\sigma_e^2 \sim \chi_{N-n}^2.$$

Τα W και SSE είναι συναρτήσεις των R_1X και R_2X αντίστοιχα. Αφού, R_1X και R_2X είναι ανεξάρτητα προκύπτει ότι και W και SSE είναι ανεξάρτητα. Οπότε θα περίμενε κανείς ότι,

$$G = \frac{\frac{W}{n-1}}{\frac{SSE}{(N-n)\sigma_e^2}} \overset{\text{προσεγγ.}}{\sim} F_{n-1,N-n}. \tag{4.61}$$

Χρησιμοποιώντας πάλι ένα θεώρημα των Robbins and Pitman (1949) μπορεί να αναπαρασταθεί η ακριβής κατανομή του στατιστικού G. Η προσέγγιση της κατανομής αυτής από την F κατανομή μελετήθηκε από τους Thomas and Hultquist (1978) και το συμπέρασμα που προέκυψε είναι ότι η προσέγγιση είναι καλή και η σχέση (4.61) επιβεβαιώνεται.

Το στατιστικό G είναι αυτό που θα χρησιμοποιηθεί για την κατασκευή του προσεγγιστικού διαστήματος εμπιστοσύνης για την intraclass συσχέτιση ρ. Η σχέση (4.61) γράφεται ισοδύναμα (βλ. σχέση 4.60),

$$G = F^* \frac{\sigma_e^2}{\widetilde{n}\sigma_A^2 + \sigma_e^2} \overset{\text{προσεγγ.}}{\sim} F_{n-1,N-n}, \tag{4.62}$$

όπου $F^* = \frac{\widetilde{n}S_{\bar{X}}^2}{MSE}$ και $\widetilde{n} = n/\sum_{i=1}^{n}(1/k_i)$. Επομένως, η διαδικασία που ακολουθούμε είναι παρόμοια με αυτή που ακολουθήσαμε για την κατασκευή του διαστήματος εμπιστοσύνης στην balanced περίπτωση (βλ. σχέσεις (4.29)-(4.31)). Δηλαδή θα κατασκευάσουμε το $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για την παράμετρο $\frac{\sigma_A^2}{\sigma_e^2}$ και στη συνέχεια θα εκμεταλλευτούμε το γεγονός ότι η παράμετρος ρ είναι μονότονη συνάρτηση του $\frac{\sigma_A^2}{\sigma_e^2}$. Ακολουθώντας την παραπάνω διαδικασία το επιθυμητό διάστημα εμπιστοσύνης θα προκύψει από τη σχέση (4.54) αντικαθιστώντας στη θέση του k_0 το $\widetilde{n}$ και στη θέση του F το F^*. Επομένως, ένα προσεγγιστικό $100(1-\alpha)\%$ διάστημα εμπιστοσύνης είναι το,

$$\left(\left(1+L_*^{-1}\right)^{-1}, \left(1+R_*^{-1}\right)^{-1}\right), \tag{4.63}$$

όπου
$$L_* = \frac{1}{\tilde{n}}\left(\frac{F^*}{F_U} - 1\right) \quad \text{και} \quad R_* = \frac{1}{\tilde{n}}\left(\frac{F^*}{F_L} - 1\right).$$

Μια άλλη προσέγγιση ακολουθήθηκε από τους Donner and Wells (1986) οι οποίοι κατασκευάζουν ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για την παράμετρο ρ, που βασίζεται στο μετασχηματισμό του Fisher. Πιο συγκεκριμένα, ο Fisher (1954) έδειξε ότι,
$$Z_F = \frac{1}{2}\log\left[\frac{1+(k_0-1)\hat{\rho}_A}{1-\hat{\rho}_A}\right], \tag{4.64}$$
ακολουθεί προσεγγιστικά μια κανονική κατανομή με μέση τιμή και διακύμανση, που δίνονται αντίστοιχα από τις παρακάτω σχέσεις,
$$E(Z_F) = \frac{1}{2}\log\left[\frac{1+(k_0-1)\rho}{1-\rho}\right] \tag{4.65}$$
και
$$Var(Z_F) = \frac{1}{2}\left[(n-1)^{-1} + (N-n)^{-1}\right]. \tag{4.66}$$
Έστω ότι,
$$I(Z_F) = \frac{\exp(2Z_F - 1) - 1}{\exp(2Z_F) + k_0 - 1}, \tag{4.67}$$
συμβολίζει τον αντίστροφο του παραπάνω μετασχηματισμού. Τότε ένα προσεγγιστικό $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για την παράμετρο ρ δίνεται από,
$$\left(I\left(Z_F - z_{a/2}\sqrt{Var(Z_F)}\right), I\left(Z_F + z_{a/2}\sqrt{Var(Z_F)}\right)\right). \tag{4.68}$$
Ακόμη, όπως αναφέρθηκε και πιο πάνω η ασυμπτωτική διακύμανση του $\hat{\rho}_A$ υπολογίστηκε από τον Smith (1957) υπό την υπόθεση της κανονικότητας και δίνεται από τη σχέση,
$$\begin{aligned} Var(\hat{\rho}_A) = & \frac{2(1-\rho)^2}{k_0^2}\left\{\frac{[1+\rho(k_0-1)]^2}{N-n} + \frac{(n-1)(1-\rho)[1+\rho(2k_0-1)]}{(n-1)^2}\right. \\ & \left. + \frac{\rho^2\left[\sum_{i=1}^n k_i^2 - 2N^{-1}\sum_{i=1}^n k_i^3 + N^{-2}\left(\sum_{i=1}^n k_i^2\right)^2\right]}{(n-1)^2}\right\}. \end{aligned}$$

Είναι γνωστό ότι για μεγάλο μέγεθος δείγματος ο εκτιμητής $\hat{\rho}_A$ ακολουθεί προσεγγιστικά μια κανονική κατανομή, άρα ένα προσεγγιστικό $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για την παράμετρο ρ δίνεται από (βλ. Donner and Wells, 1986),

$$\left(\hat{\rho}_A - z_{a/2}\sqrt{\hat{Var}(\hat{\rho}_A)}, \hat{\rho}_A + z_{a/2}\sqrt{\hat{Var}(\hat{\rho}_A)}\right), \tag{4.69}$$

όπου η άγνωστη τιμή της παραμέτρου ρ αντικαθίσταται από την αντίστοιχη τιμή του εκτιμητή $\hat{\rho}_A$ στον υπολογισμό των παραπάνω ορίων.

4.3.4 Στατιστικά τεστ

Ολοκληρώνοντας το κεφάλαιο αυτό θα κατασκευαστούν, στη συνέχεια, στατιστικά τεστ για τον έλεγχο,

$$H_0 : \rho = \rho_0 \quad \text{έναντι} \quad H_1 : \rho \neq \rho_0 \quad (\rho_0 \geq 0),$$

με την υπόθεση ότι το μοντέλο (4.1) ισχύει και με την επιπλέον υπόθεση της κανονικής κατανομής των τυχαίων επιδράσεων α_i και των όρων σφάλματος e_{ij}. Όσα ακολουθούν βασίζονται στην εργασία των Mian et al. (1989).

Όπως αναφέρθηκε και παραπάνω ο Smith (1957) υπολόγισε την ασυμπτωτική διακύμανση $Var(\hat{\rho}_A)$ του εκτιμητή $\hat{\rho}_A$. Έστω ότι με $Var_0(\hat{\rho}_A)$ συμβολίζεται η τιμή της $Var(\hat{\rho}_A)$ αντικαθιστώντας στη θέση του ρ το ρ_0. Ακόμη ο $\hat{\rho}_A$ είναι συνεπής εκτιμητής του ρ και ισχύει ότι,

$$Z_A = \frac{\hat{\rho}_A - \rho_0}{\sqrt{Var_0(\hat{\rho}_A)}} \xrightarrow[n\to\infty]{d} N(0,1), \quad \text{υπό την} \quad H_0.$$

Επομένως, ένα ασυμπτωτικό τεστ μεγέθους α είναι εκείνο που απορρίπτει την H_0 εάν $|Z_A| > z_{\frac{a}{2}}$.

Ένα άλλο στατιστικό για τον έλεγχο της H_0 μπορεί να σχηματιστεί με τη χρήση του μετασχηματισμού του Fisher (βλ. σχέση (4.64)). Επομένως, ένα ασυμπτωτικό τεστ μεγέθους α για τον έλεγχο της H_0 είναι αυτό που απορρίπτει την H_0 εάν,

$$|Z| = \frac{|\hat{Z}_F - E(Z_F^0)|}{\sqrt{Var(Z_F)}} > z_{\frac{\alpha}{2}},$$

όπου $\hat{Z}_F$ ορίζεται από την (4.64), $E(Z_F^0)$ υπολογίζεται από τη σχέση (4.65) με αντικατάσταση του ρ με ρ_0 και $Var(Z_F)$ ορίστηκε στη σχέση (4.66). Ακόμη, είδαμε ότι ο Wald (1940) απέδειξε (βλ. σχέση (4.49)), ότι,

$$F = \frac{(N-n)\sum_{i=1}^{n}\left[w_i\left(\bar{X}_i - \frac{\sum_{i=1}^{n} w_i\bar{X}_i}{\sum_{i=1}^{n} w_i}\right)^2\right]}{(n-1)\sum_{i=1}^{n}\sum_{j=1}^{k_i}(X_{ij}-\bar{X}_i)^2} \sim F_{n-1,N-n}.$$

Το παραπάνω στατιστικό είναι εκφρασμένο ως συνάρτηση του λόγου $\frac{\sigma_A^2}{\sigma_e^2}$. Ισχύουν όμως οι παρακάτω σχέσεις,

$$\sigma_A^2 = \rho\sigma_X^2 \tag{4.70}$$

και

$$\sigma_e^2 = (1-\rho)\sigma_X^2. \tag{4.71}$$

Έτσι, το παραπάνω στατιστικό F γράφεται ισοδύναμα με τη χρήση αυτών των σχέσεων ως,

$$F(\rho) = \frac{(1-\rho)\sum_{i=1}^{n}\alpha_i(\rho)\left(\bar{X}_i - \hat{\mu}(\rho)\right)^2}{(n-1)MSE} \sim F_{n-1,N-n}, \tag{4.72}$$

όπου $\alpha_i(\rho) = \frac{k_i}{1+(k_i-1)\rho}$ και $\hat{\mu}(\rho)$ προκύπτει από το $\frac{\sum_{i=1}^{n} w_i\bar{X}_i}{\sum_{i=1}^{n} w_i}$, αντικαθιστώντας στη θέση των w_i χρησιμοποιώντας τις σχέσεις (4.70) και (4.71) και μετατρέποντας με τον τρόπο αυτό τα w_i σε συναρτήσεις του ρ, δηλαδή $w_i = (1-\rho)\frac{k_i}{1+(k_i-1)\rho} = (1-\rho)\alpha_i(\rho), i = 1,\ldots,n$. Δηλαδή τελικά θα έχουμε ότι $\hat{\mu}(\rho) = \frac{\sum_{i=1}^{n}\alpha_i(\rho)\bar{X}_i}{\sum_{i=1}^{n}\alpha_i(\rho)}$. Έτσι, αντικαθιστούμε την τιμή ρ_0 στη θέση του ρ στο στατιστικό της σχέσης (4.71) και μικρές ή μεγάλες τιμές του στατιστικού υποδεικνύουν απόρριψη της μηδενικής υπόθεσης.

Τέλος τα παρακάτω δύο στατιστικά θα έχουν προσεγγιστικά F-κατανομή.

1) Γνωρίζουμε ότι, στην unbalanced περίπτωση, το $\frac{(n-1)MSA}{k_0\sigma_A^2+\sigma_e^2}$ δεν ακολουθεί μια χ^2 κατανομή με $n-1$ βαθμούς ελευθερίας, εκτός εάν $\sigma_A^2 = 0$. Ωστόσο, όπως και αναφέρθηκε, στην κατασκευή διαστημάτων εμπιστοσύνης μπορούμε να χρησιμοποιήσουμε σχέσεις που προέκυψαν από τη balanced περίπτωση, αντικαθιστώντας τον όρο k με k_0. Τότε έχουμε ότι,

$$\frac{F}{1+k_0\frac{\sigma_A^2}{\sigma_e^2}} \overset{\text{προσεγγ.}}{\sim} F_{n-1,N-n}.$$

Με αντικατάσταση των σχέσεων (4.70) και (4.71) προκύπτει ότι,

$$F\left(\frac{1-\rho}{1+(k_0-1)\rho}\right) \overset{\text{προσεγγ.}}{\sim} F_{n-1,N-n}, \tag{4.73}$$

όπου $F = MSA/MSE$. Επομένως, ένα στατιστικό που μπορεί να χρησιμοποιηθεί για τον έλεγχο της H_0 είναι το,

$$F(\rho_0) = F\left(\frac{1-\rho_0}{1+(k_0-1)\rho_0}\right) \overset{\text{προσεγγ.}}{\sim} F_{n-1,N-n}, \text{ υπό την } H_0. \tag{4.74}$$

Ωστόσο, όπως αναφέραμε και παραπάνω η κατανομή του $F(\rho_0)$ στην unbalanced περίπτωση θα είναι η $F_{n-1,N-n}$ εάν $\rho_0 = 0$. Επομένως, η παραπάνω προσέγγιση αποκλίνει καθώς ρ_0 αυξάνεται.

2) Οι Thomas and Hultquist (1978) κατέληξαν στο συμπέρασμα ότι,

$$G = F^* \frac{\sigma_e^2}{\widetilde{n}\sigma_A^2+\sigma_e^2} \overset{\text{προσεγγ.}}{\sim} F_{n-1,N-n}.$$

Με αντικατάσταση των σχέσεων (4.70) και (4.71) έχουμε ισοδύναμα ότι,

$$G = \frac{\widetilde{n}S_{\bar{X}}^2(1-\rho)}{MSE[1+\rho(\widetilde{n}-1)]} \overset{\text{προσεγγ.}}{\sim} F_{n-1,N-n}. \tag{4.75}$$

Επομένως, ένα στατιστικό που μπορεί να χρησιμοποιηθεί για τον έλεγχο της H_0 είναι το,

$$G(\rho_0) = \frac{\widetilde{n}S_{\bar{X}}^2(1-\rho_0)}{MSE[1+\rho_0(\widetilde{n}-1)]} \overset{\text{προσεγγ.}}{\sim} F_{n-1,N-n}, \text{ υπό την } H_0. \tag{4.76}$$

Μικρές ή μεγάλες τιμές των παραπάνω προσεγγιστικών στατιστικών υποδεικνύουν την απόρριψη της H_0.

ΕΠΙΛΟΓΟΣ

Σκοπός της εργασίας αυτής ήταν μια ανασκόπηση της κύριας βιβλιογραφίας, που υπάρχει σχετικά με συμπερασματολογία (σημειοεκτιμητική, διαστήματα εμπιστοσύνης, στατιστικά τεστ) για τον intraclass συντελεστή συσχέτισης για οικογενειακά δεδομένα (familial data). Το πρόβλημα μελετήθηκε από τη σκοπιά της πολυδιάστατης ανάλυσης. Η εξαγωγή συμπερασματολογίας βασίστηκε στη μέθοδο της μέγιστης πιθανοφάνειας με ένα ή και περισσότερα τυχαία δείγματα. Αναφέρονται και εναλλακτικοί εκτιμητές, που μπορεί να χρησιμοποιηθούν αντί των εκτιμητών μέγιστης πιθανοφάνειας. Ακόμη, μελετήθηκε το πρόβλημα υπό την υπόθεση ότι οι παρατηρήσεις περιγράφονται από ένα μοντέλο τυχαίων επιδράσεων (one-way random effect model). Τα συμπεράσματα διατυπώθηκαν ξεχωριστά για δύο περιπτώσεις. Στη μία το μέγεθος των οικογενειών είναι σταθερό μεταξύ των οικογενειών, ενώ στην άλλη μεταβάλλεται από οικογένεια σε οικογένεια. Στην παρούσα εργασία θεωρήσαμε ότι τα δεδομένα που είχαμε στη διάθεσή μας είναι συνεχή. Ωστόσο, η έννοια της intraclass συσχέτισης εμφανίζεται και σε οικογενειακά μοντέλα (familial models) για διακριτά και διχότομα δεδομένα (βλ. Sutradhar, 2011). Στην περίπτωση αυτή κάθε παρατήρηση (διακριτή ή διχότομη), που αφορά το παιδί μιας οικογένειας συνοδεύεται από ένα σύνολο συμμεταβλητών (covariates), όπως το φύλο, το επίπεδο εκπαίδευσης κτλ. Ωστόσο, οι παρατηρήσεις παιδιών που ανήκουν στην ίδια οικογένεια επηρεάζονται από την παρουσία μίας κοινής τυχαίας 'οικογενειακής' επίδρασης. Δηλαδή οι παρατηρήσεις οποιονδήποτε μελών της ίδιας οικογένειας είναι συσχετισμένες. Σκοπός, στο πλαίσιο αυτό, είναι η εκτίμηση των επιδράσεων που έχουν οι συμμεταβλητές στις απαντήσεις των παιδιών, λαμβάνοντας υπόψιν και τη συσχέτιση που υπάρχει μεταξύ τους. Διάφορες μέθοδοι προσέγγισης υπάρχουν στη βιβλιογραφία σχετικά με το πρόβλημα της εκτίμησης των παραπάνω παραμέτρων (βλ. Breslow and Clayton (1993), Lee and Nelder (1996), Jiang (1998), Jiang and Zhang (2001), Sutradhar (2004), Sutradhar and Mukerjee (2005), Sutradhar and Rao (2003)).

ΠΕΡΙΛΗΨΗ

Ο intraclass (ενδοκατηγορικός) συντελεστής συσχέτισης χρησιμοποιείται συνήθως ως μέτρο του βαθμού της σχέσης των μελών της ίδιας οικογένειας, ως προς κάποιο βιολογικό, περιβαλλοντολογικό ή ψυχολογικό χαρακτηριστικό, όπως για παράδειγμα το επίπεδο πίεσης του αίματος, το ύψος, ο δείκτης ευφυίας. Η έννοια του intraclass συντελεστή συσχέτισης χρησιμοποιείται στα πλαίσια της εκτίμησης της κληρονομικότητας (heritability). Η κληρονομικότητα (heritability) είναι ένα στατιστικό, που χρησιμοποιείται στα πλαίσια της γενετικής και εκτιμά το βαθμό της μεταβλητότητας ενός χαρακτηριστικού σε ένα πληθυσμό, που οφείλεται στη γενετική μεταβλητότητα, που υπάρχει μεταξύ ατόμων στον πληθυσμό. Ακόμη, ο intraclass συντελεστής συσχέτισης στη γενετική παίζει κεντρικό ρόλο στην εκτίμηση της κληρονομικότητας επιλεγμένων χαρακτηριστικών σε ζωϊκούς και φυτικούς πληθυσμούς. Ο κύριος σκοπός της εργασίας αυτής είναι να παρουσιαστεί μια ανασκόπηση της σημαντικότερης βιβλιογραφίας που είναι διαθέσιμη γύρω από συμπερασματολογία για τον intraclass συντελεστή συσχέτισης για οικογενειακά δεδομένα (family data).

Έστω X η τυχαία μεταβλητή που κωδικοποιεί ένα χαρακτηριστικό των μελών

μιας οικογένειας και πιο συγκεκριμένα των παιδιών μιας οικογένειας. Έστω $X_i = (X_{i1}, X_{i2}, \ldots, X_{ik_i})', i = 1, \ldots, n$, είναι το τυχαίο διάνυσμα που παριστάνει τις παρατηρήσεις της i-οστής οικογένειας, που αποτελείται από k_i παιδιά. Υποθέτουμε ότι το τυχαίο διάνυσμα $X_i, i = 1, \ldots, n$, ακολουθεί κανονική κατανομή με μέσο διάνυσμα μ_i και πίνακα διακυμάνσεων-συνδιακυμάνσεων Σ_i. Επιπλέον, υποθέτουμε ότι κάθε συνιστώσα του διανύσματος X_i έχει την ίδια μέση τιμή $\alpha, \alpha \in \mathbb{R}$, την ίδια διακύμανση σ^2 και ότι οι συνδιακυμάνσεις μεταξύ των X_{ij} και X_{il}, για $j \neq l$, είναι ίσες. Ως αποτέλεσμα τα παιδιά της ίδιας οικογένειας έχουν κοινή συσχέτιση ρ. Αυτή η συσχέτιση ορίζεται ως η intraclass συσχέτιση. Συνοψίζοντας τα παραπάνω υποθέτουμε ότι $X_i \sim N_{k_i}(\mu_i, \Sigma_i)$, όπου $\mu_i = (\alpha, \ldots, \alpha)'$ και $\Sigma_i = \sigma^2[(1-\rho)I_{k_i} + \rho J_{k_i}]$ για $i = 1, \ldots, n$. Το παραπάνω μοντέλο είναι γνωστό ως intraclass μοντέλο και χρησιμοποιείται για τη μοντελοποίηση οικογενειακών δεδομένων.

Στο πλαίσιο αυτό, στο πρώτο κεφάλαιο με την υπόθεση του intraclass μοντέλου, αναπτύσσεται συμπερασματολογία που βασίζεται στη μέθοδο μέγιστης πιθανοφάνειας. Δύο περιπτώσεις διακρίνονται. Στην πρώτη περίπτωση, το μέγεθος των οικογενειών είναι το ίδιο μεταξύ των οικογενειών, σε αντίθεση με τη δεύτερη στην οποία ποικίλλει. Το κεφάλαιο κλείνει με προτεινόμενους εκτιμητές που μπορούν να χρησιμοποιηθούν αντί των εκτιμητών μέγιστης πιθανοφάνειας.

Στο δεύτερο κεφάλαιο στατιστικά τεστ κατασκευάζονται για τον έλεγχο της υπόθεσης κοινού intraclass συντελεστή συσχέτισης ρ όταν έχουμε στη διάθεσή μας 2 ή περισσότερα τυχαία δείγματα.

Στο τρίτο κεφάλαιο συμπερασματολογία αναπτύσσεται με τη χρήση μεθόδων ανάλυσης διακύμανσης. Πιο συγκεκριμένα, υποθέτουμε ότι το j-οστό μέλος της i-οστής οικογένειας ($j = 1, \ldots, k_i, i = 1, \ldots, n$) μπορεί μαθηματικά να περιγραφεί από ένα μοντέλο τυχαίων επιδράσεων. Δύο περιπτώσεις διακρίνονται κι εδώ.

ABSTRACT

The intraclass correlation coefficient is used to measure the degree of resemblance, the degree of correlation between the members of a family, in respect to some biological, environmental or phychological attribute, such as blood pressure level, height or I.Q. The main aim of this Thesis is to present a review of the existing methods of inference about the intraclass correlation coefficient which is always present in analyzing familial data.

In this context, let X be a random variable that represents a trait or feature of the members of a family and more specifically of the children of a family. Let $X_i = (X_{i1}, X_{i2}, \ldots, X_{ik_i})', i = 1, \ldots, n$, be the random vector that represents the observations, which are collected on the feature X of the children the trait X of the i-th family, which consists on k_i children. It is assumed that the random vector $X_i, i = 1, \ldots, n$, is normally distributed with mean vector μ_i and variance-covariance matrix Σ_i. In addition, we assume that each component of the vector X_i has the same mean $\alpha, \alpha \in \mathbb{R}$, the same variance σ^2 and that the covariances between X_{ij} and X_{il}, for $j \neq l$, are equal. As a result, the children of the same family have a common correlation ρ. This correlation is

defined to be the intraclass correlation. Summarizing the above, we assume that $X_i \sim N_{k_i}(\mu_i, \Sigma_i)$, where $\mu = (\alpha, \ldots, \alpha)'$ and $\Sigma_i = \sigma^2[(1-\rho)I_{k_i} + \rho J_{k_i}]$ for $i = 1, \ldots, n$. The above model is known as intraclass model and it is used for modeling familial data.

In this context, in the second chapter and subject to the assumption of the intraclass model, inference is being developed for the intraclass correlation which is based on the method of maximum likelihood. Two cases are distinguished. In the first, family sizes are considered to be the same among families, in contrast with the second one in which the respective family sizes are considered to be different. This chapter is integrated with an alternative to the maximum likelihood method for the estimation of the intraclass correlation coefficient.

Chapter 3 focuses on the development of tests, for testing the null hypothesis of a common intraclass correlation coefficient ρ, on the basis of two or more random samples.

In the final chapter 4, inference is being developed for the intraclass correlation by using methods of analysis of variance. More specefically, it is assumed that the j-th member of the i-th family, X_{ij}, $j = 1, \ldots, k_i, i = 1, \ldots, n$ can be statistically described by a one-way random effect model. Two cases, the same as those in chapter 2 depending on the family sizes, are also considered.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Anderson, T. W. and Olkin, I. (1985). Maximum-likelihood estimation of the parameters of a multivariate normal distribution. *Linear algebra and its applications*, **70**, 147-171.

2. Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, **88**(421), 9-25.

3. Donner, A. (1979). The use of correlation and regression in the analysis of family resemblance. *American journal of epidemiology*, **110**(3), 335-342.

4. Donner, A. (1986). A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review*, **54**, 67-82.

5. Donner, A. and Bull, S. (1983). Inferences concerning a common intraclass correlation coefficient. *Biometrics*, **39**, 771-775.

6. Donner, A. and Koval, J. J. (1980a). The estimation of intraclass correlation in the analysis of family data. *Biometrics*, **36**, 19-25.

7. Donner, A. and Koval, J. J. (1980b). The large sample variance of an intraclass correlation. *Biometrika*, **67**(3), 719-722.

8. Donner, A. and Koval, J. J. (1983). A note on the accuracy of fisher's approximation to the large sample variance of an intraclass correlation. *Communications in Statistics-Simulation and Computations*, **12**(4), 443-449.

9. Donner, A. and Wells, G. (1986). A comparison of confidence interval methods for the intraclass correlation coefficient. *Biometrics*, **42**, 401-412.

10. Fieller, E. and Smith, C. A. (1951). Note on the analysis of variance and intraclass correlation. *Annals of Eugenics*, **16**(1), 97-104.

11. Fisher, R. (1954). *Statistical methods for research workers.* Oliver and Boyd, Edinburgh.

12. Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, **1**, 3-32.

13. Graybill, F. A. (1961). *An introduction to linear statistical models.* McGraw-Hill Book Company, New York.

14. Graybill, F. A. et al. (1954). On quadratic estimates of variance components. *The Annals of Mathematical Statistics*, **25**(2), 367-372.

15. Graybill, F. A. and Hultquist, R. A. (1961). Theorems concerning eisenhart's model ii. *The Annals of Mathematical Statistics*, **32**(1), 261-269.

16. Graybill, F. A. and Wortham, A. (1956). A note on uniformly best unbiased estimators for variance components. *Journal of the American Statistical Association*, **51**(274), 226-268.

17. Harris, J. A. (1913). On the calculation of intra-class and inter-class coefficients of correlation from class moments when the number of possible combinations is large. *Biometrika*, **9**(3/4), 446-472.

18. Harville, D. A. (1969). Quadratic unbiased estimation of variance components for the one-way classification. *Biometrika*, **56**(2), 313-326.

19. Helu, A. and Naik, D. N. (2007). Estimation of sib-sib correlation via a kotz-type density function. *Communications in Statistics-Theory and Methods*, **36**(5), 1021-1029.

20. Jiang, J. (1998). Consistent estimators in generalized linear mixed models. *Journal of the American Statistical Association*, **93**(442), 720-729.

21. Jiang, J. and Zhang, W. (2001). Robust estimation in generalised linear mixed models. *Biometrika*, **88**(3), 753-765.

22. Karlin, S., Cameron, E. C. and Williams, P. T. (1981). Sibling and parent-offspring correlation estimation with variable family size. *Proceedings of the National Academy of Sciences*, **78**(5), 2664-2668.

23. Khatri, C., Pukkila, T. and Rao, C. R. (1989). Testing intraclass correlation coefficients. *Communications in Statistics-Simulation and Computation*, **18**(1), 15-30.

24. Knight, K. (2000). *Mathematical Statistics.* Chapman & Hall/CRC, Boca Raton.

25. Konishi, S. and Gupta, A. K. (1989). Testing the equality of several intraclass correlation coefficients. *Journal of Statistical Planning and Inference*, **21**(1), 93-105.

26. Kourouklis, S., Petropoulos, K. and Piperigkou, V. (2015). *Θέματα παραμετρικής στατιστικής συμπερασματολογίας*. [ηλεκτρ. βιβλ.]. Εκδόσεις Κάλλιπος, Αθήνα.

27. Lancaster, H. (1965). The helmert matrices. *The American Mathematical Monthly*, **72**(1), 4-12.

28. Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(4), 619-656.

29. Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation.* Springer Science & Business Media.

30. Mardia, K., Kent, J. and Bibby, J. (1979). *Multivariate Analysis.* Academic Press, INC, San Diego.

31. McGraw, K. O. and Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods*, **1**(1), 30.

32. Mentz, R. P. (2001). Likelihood ratio test of equality of several variances in the intraclass correlation model. *Communications in Statistics-Theory and Methods*, **30**(1), 75-91.

33. Mian, I., Shoukri, M. and Tracy, D. (1989). A comparison of significance testing procedures for the intraclass correlation from family data. *Communications in Statistics-Simulation and Computations*, **18**(2), 613-631.

34. Naik, D. N. and Helu, A. (2007). On testing equality of intraclass correlations under unequal family sizes. *Computational Statistics & Data Analysis*, **51**(12), 6498-6510.

35. Paul, S. and Barnwal, R. (1990). Maximum likelihood estimation and a c(a) test for a common intraclass correlation. *Journal of the Royal Statistical Society, Series D (The Statistician)*, **39**(1), 19-24.

36. Rao, C. R. (1973). *Linear statistical inference and its applications.* John Wiley and Sons, New York.

37. Robbins, H. and Pitman, E. (1949). Applications of the method of mixtures to quadratic forms in normal variates. *The Annals of mathematical statistics*, **20**, 552-560.

38. Saw, S. L. (1992). Anova sums of squares as quadratic forms. *The American Statistician*, **46**(4), 288-290.

39. Schäke, K. (2013). On the Kronecker Product.

40. Scheffe, H. (1999). *The analysis of variance.* John Wiley & Sons, New York.

41. Searle, S. R. et al. (1966). *Matrix algebra for the biological sciences.* Wiley, New York.

42. Searle, S. R. (1997). *Linear Models.* John Wiley and Sons, New York.

43. Smith, C. (1957). On the estimation of intraclass correlation. *Annals of human genetics*, **21**(4), 363-373.

44. Srivastava, M. (1984). Estimation of interclass correlations in familial data. *Biometrika*, **71**(1), 177-185.

45. Srivastava, M. and Katapa, R. (1986). Comparison of estimators of interclass and intraclass correlations from familial data. *Canadian Journal of Statistics*, **14**(1), 29-42.

46. Srivastava, M. S. (2002). *Methods of multivariate statistics.* John Wiley and Sons, New York.

47. Sutradhar, B. C. (2004). On exact quasilikelihood inference in generalized linear mixed models. *Shankyā: The Indian Journal of Statistics*, **66**, 263-291.

48. Sutradhar, B. C. (2011). *Dynamic mixed models for familial longitudinal data.* Springer Science & Business Media, New York.

49. Sutradhar, B. C. and Mukerjee, R. (2005). On likelihood inference in binary mixed model with an application to copd data. *Computational & data analysis*, **48**(2), 345-361.

50. Sutradhar, B. C. and Rao, R. P. (2003). On the quasi-likelihood inference in generalized linear mixed models with two components of dispersion. *The Canadian Journal of Statistics*, **31**, 415-435.

51. Thomas, J. D. and Hultquist, R. A. (1978). Interval estimation for the unbalanced case of the one-way random effects model. *The Annals of Statistics*, **6**, 582-587.

52. Townsend, E. C. (1968). Unbiased estimators of variance components in simple unbalanced designs. Ph. D. Thesis, Cornell University, Ithaca, N.Y.

53. Townsend, E. C. and Searle, S. R. (1971). Best quadratic unbiased

estimation of variance components from unbalanced data in the 1-way classification. *Biometrics*, **27**, 643-657.

54. Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press, United Kingdom.

55. Wald, A. (1940). A note on the analysis of variance with unequal class frequencies. *The Annals of Mathematical Statistics*, **11**(1), 96-100.

56. Watson, G. (1964). A note on maximum likelihood, *Shankyā*, **26**, 303-304.

57. Wilks, S. S. et al. (1946). Sample criteria for testing equality of means, equality of variances and equality of covariances in a normal multivariate distribution. *The Annals of Mathematical Statistics*, **17**(3), 257-281.

58. Wray, N. and Visscher, P. (2008). Estimating trait heritability. *Nature education*, **1**(1), 29.