**ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ**
**Τμημα Μαθηματικων**

**Αθανάσιος Γεωργίου**

---

# Ελλιπή δεδομένα: Έλεγχος της υπόθεσης MCAR

---

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΑΤΡΙΒΗ

Ιωάννινα, 2025

**UNIVERSITY OF IOANNINA**
**Department of Mathematics**

**Athanasios Georgiou**

# Missing data: testing the MCAR hypothesis

Master's Thesis

Ioannina, 2025

*Αφιερώνεται στην οικογένειά μου*

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

Εγκρίθηκε την 01/07/2025 από την εξεταστική επιτροπή:

| **Ονοματεπώνυμο** | **Βαθμίδα** |
|---|---|
| Κωνσταντίνος Ζωγράφος | Καθηγητής |
| Δημήτριος Μπάγκαβος | Αν. Καθηγητής |
| Απόστολος Μπατσίδης | Αν. Καθηγητής (επιβλέπων) |

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."

Αθανάσιος Γεωργίου

# Ευχαριστιες

Η παρούσα διατριβή εκπονήθηκε στο πλαίσιο της επιτυχούς περάτωσης των μεταπτυχιακών μου σπουδών, για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στον τομέα Πιθανοτήτων, Στατιστικής και Επιχειρησιακής Έρευνας του τμήματος Μαθηματικών του Πανεπιστημίου Ιωαννίνων. Η τριμελής εξεταστική επιτροπή αυτής της διατριβής απαρτίζεται από τον επιβλέποντα Αναπληρωτή Καθηγητή κ. Απόστολο Μπατσίδη, τον Αναπληρωτή Καθηγητή κ. Δημήτριο Μπάγκαβο και τον Καθηγητή κ. Κωνσταντίνο Ζωγράφο. Στη συνέχεια θα ήθελα να εκφράσω τις ευχαριστίες μου σε όσους συνέβαλαν στην ολοκλήρωση αυτής της διατριβής.

Αρχικά θα ήθελα να εκφράσω την ευγνωμοσύνη μου για τον επιβλέποντα Αναπληρωτή Καθηγητή κ. Απόστολο Μπατσίδη. Οι εύστοχες παρατηρήσεις για συμπερίληψη αποτελεσμάτων που παρουσιάζουν ενδιαφέρον και η διαρκής πίεση για καλύτερη έκφραση, διαμόρφωσαν την τελική έκβαση αυτής της διατριβής, με τρόπο μεστό και ουσιώδη. Παράλληλα, η διαθεσιμότητα που επέδειξε, η συνεχής παρότρυνση για την επίτευξη του βέλτιστου αποτελέσματος, αλλά και η αντιμετώπιση των λαθών μου, πάντοτε χιουμοριστικά, δηλώνουν πολλά για τον ευγενή και εργατικό του χαρακτήρα. Τα χαρακτηριστικά αυτά προσπάθησε να τα κληροδοτήσει και σε εμένα. Επομένως, μπορώ να πω ειλικρινά, ότι πέρα από τη συνεργασία μου με έναν εξαιρετικό επιβλέποντα, συνεργάστηκα με έναν εξαιρετικό άνθρωπο και την όποια προσπάθεια αυτοβελτίωσης κατέβαλα, την οφείλω σε εκείνον.

Ακόμη, θα ήθελα να εκφράσω τις ευχαριστίες μου προς τα υπόλοιπα μέλη και καθηγητές μου κ. Δημήτριο Μπάγκαβο και κ. Κωνσταντίνο Ζωγράφο. Τους ευχαριστώ για τα σχόλια, τις υποδείξεις και τον χρόνο που αφιέρωσαν σε αυτή τη μεταπτυχιακή διατριβή και γενικότερα σε όλη τη διάρκεια των σπουδών μου.

Επιπρόσθετα, θα ήθελα να ευχαριστήσω την οικογένεια μου για την συνεχή ηθική και οικονομική υποστήριξή της. Δίχως τη συμπαράστασή τους, θα είχα προ πολλού παραιτηθεί από το όνειρό μου για ενασχόληση με την επιστήμη των Μαθηματικών και συγκεκριμένα τη Στατιστική. Αυτός είναι και ο λόγος που

αυτή η διατριβή είναι αφιερωμένη σε αυτούς.

Τέλος, θα ήθελα να ευχαριστήσω όλους μου τους φίλους για την ανοχή, τη διαρκή πίστη και συμπαράστασή τους σε μένα. Όσοι ήταν κοντά και όσοι ήταν μακριά συνέβαλαν με τον τρόπο τους στην επιτυχή περάτωση των σπουδών μου. Για αυτό και τη χαρά της επιτυχίας αυτής θα την μοιραστώ, ευελπιστώ σύντομα μαζί τους!

# Περιληψη

Συχνά σε πρακτικές εφαρμογές δεν είναι διαθέσιμες όλες οι παρατηρήσεις και ο στατιστικός έχει τελικά στη διάθεσή του ένα σύνολο δεδομένων που χαρακτηρίζεται από ελλιπείς τιμές. Οι λόγοι που παρουσιάζονται σύνολα δεδομένων με ελλιπείς τιμές είναι ποικίλοι και οδηγούν σε διάφορες μορφές-μοτίβα ελλιπών δεδομένων. Ο Rubin (1976) θέλοντας να εξηγήσει τον μηχανισμό εμφάνισης των ελλιπών τιμών και να δώσει απάντηση στο ερώτημα αν το γεγονός ότι κάποιες μεταβλητές εμφανίζουν ελλιπείς τιμές σχετίζεται με την τιμή των μεταβλητών αυτών στο σύνολο των δεδομένων περιέγραψε τρεις μηχανισμούς που οδηγούν σε ελλιπή δεδομένα. Πιο συγκεκριμένα, σύμφωνα με τον Rubin (1976), έχουμε τα Πλήρως Τυχαία Ελλιπή (Missing Completely at Random (MCAR)), τα Τυχαία Ελλιπή (Missing at Random (MAR)) και τα Μη Τυχαία Ελλιπή (Non Missing at Random (NMAR)) δεδομένα. Μεταξύ αυτών, για λόγους που θα παρουσιαστούν στο πρώτο κεφάλαιο αυτής της μεταπτυχιακής διατριβής, ξεχωριστή θέση κατέχουν τα MCAR δεδομένα και για αυτόν τον λόγο είναι απαραίτητο όταν υπάρχουν ελλιπή δεδομένα να επιβεβαιώνεται ότι αυτά μπορούν να θεωρηθούν ότι έχουν προέλθει από έναν τέτοιον μηχανισμό. Για τον λόγο αυτό στη βιβλιογραφία, ιδιαίτερα τα τελευταία έτη, παρατηρείται ιδιαίτερο ενδιαφέρον για την εισαγωγή μεθόδων για τον έλεγχο αν ο μηχανισμός έλλειψης των δεδομένων είναι MCAR ή όχι. Στο πλαίσιο αυτό, έχουν παρουσιαστεί μεθοδολογίες ελέγχου τόσο για τη γενική περίπτωση που τα διαθέσιμα δεδομένα μπορούν να θεωρηθούν ένα τυχαίο δείγμα από τον υπό μελέτη πληθυσμό, αλλά και μεθοδολογίες για ειδικές περιπτώσεις δεδομένων (διαχρονικά, επαναλαμβανόμενες μετρήσεις κ.ο.κ.). Επιπλέον, κάποιες από τις μεθοδολογίες είναι ειδικά σχεδιασμένες για διακριτά/κατηγορικά δεδομένα ή μεικτού τύπου δεδομένα. Στο πλαίσιο αυτής της διατριβής, θα περιοριστούμε στην παρουσίαση των μεθοδολογιών εκείνων που είναι σχεδιασμένες για τη γενική περίπτωση που έχουμε διαθέσιμο ένα τυχαίο σύνολο συνεχών δεδομένων ή που μπορούν να εφαρμοστούν σε τέτοιες περιπτώσεις χωρίς να είναι απαραίτητη η διακριτοποίηση των δεδομένων.

Στο παραπάνω πλαίσιο, η διάρθρωση του υπολοίπου της διατριβής έχει ως εξής. Στο Κεφάλαιο 1 (Εισαγωγή) παρουσιάζονται διάφορα πιθανά μοτίβα ελ-

λιπών δεδομένων, ορίζονται οι τρεις μηχανισμοί δημιουργίας ελλιπών δεδομένων και αναδεικνύεται τόσο η σπουδαιότητα των MCAR δεδομένων όσο και η σπουδαιότητα ύπαρξης στατιστικών ελέγχων ότι τα διαθέσιμα ελλιπή δεδομένα είναι τέτοιου τύπου. Στη συνέχεια, στο Κεφάλαιο 2 παρουσιάζονται οι έλεγχοι εκείνοι των οποίων η κεντρική ιδέα είναι η ακόλουθη: τα διαθέσιμα ελλιπή δεδομένα χωρίζονται σε διακεκριμένα μοτίβα και ο έλεγχος της υπόθεσης MCAR ανάγεται στον έλεγχο ότι κάθε ένα από αυτά τα μοτίβα προέρχεται από τον ίδιο πληθυσμό. Αυτό πρακτικά σημαίνει ότι ο έλεγχος της υπόθεσης MCAR ανάγεται στον έλεγχο αν τα δεδομένα ικανοποιούν την υπόθεση της ομοιογένειας. Είναι προφανές ότι απόρριψη της υπόθεσης της ομοιογένειας συνεπάγεται απόρριψη της υπόθεσης MCAR, ενώ αν η υπόθεση της ομοιογένειας δεν μπορεί να απορριφθεί αυτό δεν σημαίνει ότι επιβεβαιώνεται η υπόθεση MCAR. Από την άλλη, όσοι έλεγχοι δεν μπορούν να ταξινομηθούν σε αυτήν την κατηγορία ελέγχων παρουσιάζονται στο Κεφάλαιο 3. Σε κάθε ενότητα των δύο κεφαλαίων οι έλεγχοι υλοποιούνται σε ένα συγκεκριμένο σύνολο δεδομένων, με χρήση συναρτήσεων της R που είναι διαθέσιμες σε κάποια πακέτα της, αλλά και άλλων που υλοποιήθηκαν στο πλαίσιο αυτής τη διατριβής. Στο Κεφάλαιο 4 το ενδιαφέρον επικεντρώνεται στην παρουσίαση των αποτελεσμάτων των συγκριτικών μελετών απόδοσης των στατιστικών ελέγχων που παρουσιάστηκαν στα προηγούμενα δύο κεφάλαια τόσο ως προς τη διατήρηση του επιπέδου σημαντικότητας όσο και ως προς την ισχύ. Στο Κεφάλαιο 5 συνοψίζονται τα συμπεράσματα αυτής της μεταπτυχιακής διατριβής, αναφέρονται σημαντικοί έλεγχοι που έχουν παρουσιαστεί στη βιβλιογραφία για άλλους τύπους δεδομένων και δίνονται προτάσεις για περαιτέρω έρευνα. Τέλος, η μεταπτυχιακή διατριβή ολοκληρώνεται με το Παράρτημα, όπου παρατίθενται οι συναρτήσεις υλοποίησης κάποιων εκ των στατιστικών ελέγχων στη γλώσσα προγραμματισμού R, και τη Βιβλιογραφία.

# Abstract

In practical applications, not all observations are often available, and the statistician ultimately works with a dataset with missing values. The reasons for missingness are diverse and lead to various patterns of incomplete data. Rubin (1976) aimed to explain the mechanisms that lead to missing values and to address the question of whether the fact that some variables have missing values is related to the values of these variables within the data set. In this framework, he described three mechanisms that result in missing data: the Missing Completely at Random - MCAR, the Missing at Random - MAR, and the Non Missing at Random - NMAR. Among these, for reasons that will be discussed in the first chapter of this master's thesis, MCAR data hold particular importance. Therefore, when dealing with missing data, it is essential to verify whether they can be considered to have arisen from such a mechanism. Consequently, there exists a growing interest in the development of methods to test whether the missing data mechanism is MCAR or not. In this context, methodologies for testing have been developed both for the general case where the available data can be considered as a random sample from the population under study, as well as for specific types of data (longitudinal, repeated measurements, etc.). Additionally, some of these methodologies are specifically designed for discrete/categorical data or mixed data. In this thesis, we will focus on presenting those methodologies that are intended for the general case where a random set of continuous data is available or methods that can be applied in such cases without the need for data discretization.

In the above context, the structure of the remainder of this thesis is as follows. Chapter 1 (Introduction) presents various possible missing data patterns, defines the three mechanisms by which missing data can arise, and highlights the importance of MCAR, as well as the necessity of conducting statistical tests to determine whether the available missing data fall under this mechanism. Specifically, Chapter 2 introduces the class of tests based on the following central idea: the available incomplete data are divided into distinct patterns, and testing the MCAR hypothesis is reduced to testing whether each of the-

se patterns originates from the same population. In practice, this means that testing the MCAR hypothesis is equivalent to testing the assumption of homogeneity. Clearly, rejection of the homogeneity assumption implies rejection of the MCAR hypothesis, whereas failure to reject the homogeneity assumption does not confirm the MCAR hypothesis. On the other hand, tests that cannot be classified into this category are presented in Chapter 3. In each section, we apply the MCAR at a specific data set. Chapter 4 focuses on presenting the results of Monte Carlo studies performed in the statistical literature in order to evaluate the performance of the statistical tests presented in the previous two chapters, in terms of both maintaining the significance level and statistical power. Chapter 5 summarizes the conclusions of this master's thesis, mentions important tests that have been presented in the literature for other data types, and provides suggestions for further research. Finally, the thesis concludes with the Appendix, which includes R code implementations of some of the statistical tests, and the References section.

# Περιεχομενα

# ΚΕΦΑΛΑΙΟ 1

# ΕΙΣΑΓΩΓΗ

Στο παρόν κεφάλαιο παρουσιάζονται διάφορα πιθανά μοτίβα ελλιπών δεδομένων, ορίζονται οι τρεις μηχανισμοί δημιουργίας ελλιπών δεδομένων και αναδεικνύεται τόσο η σπουδαιότητα των MCAR δεδομένων όσο και η σπουδαιότητα ύπαρξης στατιστικών ελέγχων ότι τα διαθέσιμα ελλιπή δεδομένα είναι τέτοιου τύπου.

## 1.1 Μοτίβα ελλιπών δεδομένων και μηχανισμοί δημιουργίας ελλιπών δεδομένων

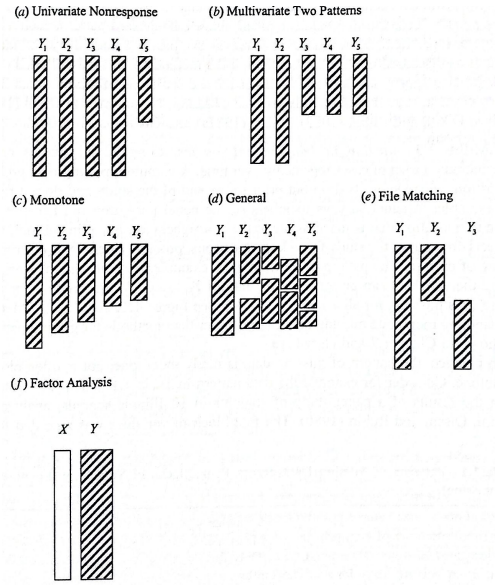
Οι περισσότερες μεθοδολογίες της Πολυμεταβλητής Στατιστικής Ανάλυσης στηρίζονται στην υπόθεση ότι τα διαθέσιμα δεδομένα είναι πλήρη. Ωστόσο, στην πράξη, οι ερευνητές έρχονται, συχνά, αντιμέτωποι με την ανάλυση συνόλων δεδομένων που δεν είναι πλήρη. Στη συνέχεια, ο όρος ελλιπή δεδομένα ή ελλιπής τιμή αναφέρεται σε όλες εκείνες τις τιμές οι οποίες είναι σαφώς καθορισμένες, αλλά για κάποιον λόγο δεν είναι γνωστές στον/στην ερευνητή/τρια. Για παράδειγμα, σε μία έρευνα που πραγματοποιείται μέσω της συλλογής δεδομένων με ερωτηματολόγιο κάποια άτομα παραλείπουν είτε τυχαία ή σκοπίμως να απαντήσουν σε κάποια ερωτήματα που αφορούν π.χ. το εισόδημά τους, το βάρος τους ή τον τόπο καταγωγής τους. Οι τιμές που παραλείπονται είναι σαφώς καθορισμένες, απλά δεν γίνονται γνωστές σε εμάς. Από τα παραπάνω γίνεται σαφές ότι η έννοια των ελλιπών δεδομένων είναι διαφορετική από αυτήν των λογοκριμένων δεδομένων (censored data), που είναι εκείνα τα δεδομένα που δεν μας είναι γνωστά π.χ. γιατί διακόπηκε η διαδικασία συλλογής των χρόνων ζωής των ατόμων πριν την εμφάνιση του γεγονότος ενδιαφέροντος.

Η παρουσία ελλιπών δεδομένων μπορεί να συμβεί για διάφορους λόγους. Ενδεικτικά, μπορεί να οφείλεται σε σφάλματα κατά τη συλλογή των δεδομένων,

σε περιορισμένους πόρους για την καταγραφή ή τη μετάδοση των δεδομένων ή ακόμα και λόγω φυσικών περιορισμών που δεν επιτρέπουν την πλήρη συγκέντρωση πληροφοριών. Η ποικιλία των λόγων δημιουργίας ελλιπών δεδομένων οδηγεί σε ποικιλία μορφών που αυτά εμφανίζονται. Στο Σχήμα 1.1 το οποίο προέρχεται από τη μονογραφία των Little and Rubin (2002), παρουσιάζονται ενδεικτικά παραδείγματα μοτίβων ελλιπών δεδομένων που παρουσιάζουν θεωρητικό ενδιαφέρον και ταυτόχρονα συναντώνται συχνά στην πράξη. Πιο συγκεκριμένα, στα Σχήματα 1.1 (a) και (b) εμφανίζονται ελλιπή δεδομένα σε μία μεταβλητή ή σε ένα διάνυσμα μεταβλητών, αντίστοιχα, και είναι γνωστά ως Univariate Nonresponse και Multivariate Two Patterns, αντίστοιχα. Στο Σχήμα 1.1 (c) αποτυπώνεται μια πολύ γνωστή περίπτωση ελλιπών δεδομένων, τα βηματικά μονότονα ελλιπή δεδομένα. Ειδικότερα, στη συγκεκριμένη περίπτωση $N_1$ από τους ερωτηθέντες δίνουν πληροφορίες για το χαρακτηριστικό $Y_1$, $N_2 < N_1$ δίνουν ταυτόχρονα πληροφορίες για τα χαρακτηριστικά $(Y_1, Y_2)$ και με ανάλογο τρόπο $N_5$ από τους ερωτηθέντες δίνουν πληροφορίες για όλα τα υπό μελέτη χαρακτηριστικά, με $N_5 < N_4 < N_3 < N_2 < N_1$. Στο Σχήμα 1.1 (d) δίνεται μια γενική μορφή ελλιπών δεδομένων, στο Σχήμα 1.1 (e) τα δεδομένα σε δύο από τις υπό μελέτη μεταβλητές (τις $Y_2$ και $Y_3$) δεν παρατηρούνται ποτέ ταυτόχρονα, γνωστό ως File Matching, δηλαδή για μία πειραματική μονάδα καταγράφεται η τιμή είτε στη μεταβλητή $Y_2$ ή στην $Y_3$. Από την άλλη μεριά, στο Σχήμα 1.1 (f) η μεταβλητή $X$ δεν παρατηρείται ποτέ (latent variable, Factor Analysis).

Ο Rubin (1976) θέλοντας να εξηγήσει τον μηχανισμό εμφάνισης των ελλιπών τιμών και να δώσει απάντηση στο ερώτημα αν το γεγονός ότι κάποιες μεταβλητές εμφανίζουν ελλιπείς τιμές σχετίζεται με την τιμή των μεταβλητών αυτών στο σύνολο των δεδομένων περιέγραψε τρεις μηχανισμούς που οδηγούν σε ελλιπή δεδομένα (βλέπε και Little and Rubin (2002)). Ειδικότερα, ο Rubin (1976) στην προσπάθειά του να απαντήσει στο ερώτημα αν η παρουσία ελλιπών τιμών σε κάποιες εκ των μεταβλητών σχετίζεται με την τιμή των μεταβλητών αυτών στο σύνολο των δεδομένων οδηγήθηκε στην περιγραφή τριών μηχανισμών εμφάνισης ελλιπών δεδομένων. Η παρουσίαση αυτών των μηχανισμών θα ακολουθήσει, αφού εισάγουμε τον απαραίτητο συμβολισμό.

Πιο συγκεκριμένα, σε όσα ακολουθούν, $\boldsymbol{Y} = (y_{ij})$ είναι ο $n \times p$ πίνακας των δεδομένων, όπου $y_{ij}$ είναι η τιμή της $i$-οστής παρατήρησης, $i = 1, ..., n$, στην $j$-οστή μεταβλητή $Y_j$, $j = 1, ..., p$. Επιπλέον, $\boldsymbol{R}$ είναι ο $n \times p$ πίνακας που καθορίζει το μοτίβο των ελλιπών δεδομένων υποδεικνύοντας αν το αντίστοιχο στοιχείο του πίνακα $\boldsymbol{Y}$ έχει παρατηρηθεί ή όχι. Αυτό πρακτικά σημαίνει ότι ο πίνακας $\boldsymbol{R}$ ουσιαστικά υποδεικνύει αν μια τιμή είναι ελλιπής ή όχι και για αυτόν τον λόγο ο πίνακας αυτός καλείται indicator matrix. Επομένως, είναι $\boldsymbol{R} = (r_{ij})$, όπου $r_{ij} = 1$ αν η τιμή του $y_{ij}$ λείπει ελλιπής και $r_{ij} = 0$ αν η




Σχήμα 1.1: Παραδείγματα μοτίβων ελλιπών δεδομένων. Πηγή: Little and Rubin (2002).

τιμή του αντίστοιχου $y_{ij}$ παρατηρείται, για $i = 1, ..., n$, και $j = 1, ..., p$. Σε όσα ακολουθούν $f(\boldsymbol{R}|\boldsymbol{Y}, \boldsymbol{\phi})$, συμβολίζει τη δεσμευμένη κατανομή του $\boldsymbol{R}$ δοθέντος του $\boldsymbol{Y}$, με $\boldsymbol{\phi}$ να είναι μία άγνωστη παράμετρος. Τέλος, με $\boldsymbol{Y}_{obs}$ συμβολίζονται οι παρατηρούμενες συνιστώσες ή στοιχεία του $\boldsymbol{Y}$ και με $\boldsymbol{Y}_{mis}$ τα ελλιπή.

Μετά από την εισαγωγή του απαραίτητου συμβολισμού, το ενδιαφέρον επικεντρώνεται στην παράθεση των ορισμών για τους τρεις μηχανισμούς έλλειψης (βλέπε Little and Rubin (2002)).

**Ορισμός 1.1.1** *Τα δεδομένα ονομάζονται Πλήρως Τυχαία Ελλιπή (Missing Completely at Random (MCAR)) όταν ο μηχανισμός που δημιουργεί την έλλειψη είναι εντελώς ανεξάρτητος τόσο από τις τιμές των υπό μελέτη μεταβλητών όσο και από τη φύση αυτών των μεταβλητών. Σε αυτήν την περίπτωση ισχύει ότι:*

$$f(\boldsymbol{R}|\boldsymbol{Y}, \boldsymbol{\phi}) = f(\boldsymbol{R}|\boldsymbol{\phi}), \text{ για κάθε } \boldsymbol{Y}, \boldsymbol{\phi}.$$

Από τον παραπάνω ορισμό προκύπτει ότι σε αυτήν την περίπτωση παρότι το μοτίβο της έλλειψης αυτό καθαυτό δεν είναι τυχαίο, η έλλειψη δεν εξαρτάται από τις τιμές των δεδομένων. Ένα τέτοιο σύνολο για παράδειγμα είναι εκείνο που

προκύπτει όταν κατά τη διαδικασία καταμέτρησης του βάρους κάποιων ατόμων τελειώσει η μπαταρία της ζυγαριάς, γεγονός που θα οδηγήσει σε έλλειψη δεδομένων καθαρά από τύχη (βλέπε van Buuren (2018)).

**Ορισμός 1.1.2** *Τα δεδομένα χαρακτηρίζονται ως Τυχαία Ελλιπή (Missing at Random (MAR)) όταν δεν υπάρχει εξάρτηση του μηχανισμού της έλλειψης και των ελλιπών παρατηρήσεων, δηλαδή όταν*

$$f(\boldsymbol{R}|\boldsymbol{Y}, \boldsymbol{\phi}) = f(\boldsymbol{R}|\boldsymbol{Y}_{obs}, \boldsymbol{\phi}), \text{ για κάθε } \boldsymbol{Y}_{mis}, \boldsymbol{\phi}.$$

Στην πράξη σύνολα δεδομένων MAR μπορεί να προκύψουν όταν τα ελλιπή στοιχεία λείπουν τυχαία και η έλλειψη τους είναι πιθανό να συνδεθεί με την ύπαρξη μιας παραμέτρου. Ένα τέτοιο σύνολο προκύπτει για παράδειγμα σε μία έρευνα που διεξάγεται μέσω ερωτηματολογίου όταν κάποια άτομα δεν αναφέρουν το βάρος τους, αλλά η απουσία αυτών των τιμών δεν εξαρτάται από το βάρος τους, αλλά από άλλες παρατηρήσιμες μεταβλητές, όπως η ηλικία ή το φύλο τους. Δηλαδή, μπορεί να παρατηρήσουμε ότι τα άτομα σε μεγαλύτερες ηλικίες (π.χ. άνω των 60) είναι πιο πιθανό να μην αναφέρουν το βάρος τους. Αυτό δεν σημαίνει ότι τα άτομα μεγαλύτερης ηλικίας έχουν διαφορετικό βάρος από τα νεότερα άτομα, αλλά ότι το βάρος τους δεν καταγράφεται επειδή υπάρχει κάποια συστηματική τάση στους ηλικιωμένους να αποφεύγουν να αναφέρουν το βάρος τους.

**Ορισμός 1.1.3** *Τα δεδομένα χαρακτηρίζονται ως Μη Τυχαία Ελλιπή (Non Missing at Random (NMAR)) όταν η κατανομή του $\boldsymbol{R}$ εξαρτάται από τις ελλιπείς τιμές στον πίνακα δεδομένων $\boldsymbol{Y}$, δηλαδή όταν*

$$f(\boldsymbol{R}|\boldsymbol{Y}, \boldsymbol{\phi}) = f(\boldsymbol{R}|\boldsymbol{Y}_{mis}, \boldsymbol{\phi}), \text{ για κάθε } \boldsymbol{Y}_{obs}, \boldsymbol{\phi}.$$

*Ουσιαστικά στον μηχανισμό αυτόν, η πιθανότητα μία μονάδα να εμφανίσει ελλιπή τιμή εξαρτάται από την τιμή της.*

Στην πράξη σύνολα δεδομένων NMAR μπορεί να προκύψουν όταν το γεγονός ότι η τιμή της μεταβλητής δεν έχει παρατηρηθεί συσχετίζεται με τον λόγο που αυτή δεν έχει παρατηρηθεί. Για παράδειγμα, αν υποθέσουμε ότι διεξάγουμε μια έρευνα για την ικανοποίηση των πελατών από μία υπηρεσία, οι πελάτες που είναι λιγότερο ικανοποιημένοι είναι πιο πιθανό να επιλέξουν να μην απαντήσουν στην ερώτηση.

Ξεχωριστή θέση στη βιβλιογραφία κατέχουν τα ελλιπή δεδομένα που παράγονται με τον μηχανισμό έλλειψης MCAR, γεγονός που οφείλεται σε πολλούς

λόγους. Ειδικότερα, όταν τα δεδομένα είναι MCAR, αυτό σημαίνει ότι η πιθανότητα να λείπουν δεδομένα δεν εξαρτάται από τις παρατηρούμενες ή μη παρατηρούμενες τιμές. Η ιδιότητα αυτή διευκολύνει την ανάλυση για πολλούς λόγους, οι οποίοι εκτενώς παρατίθενται στην εργασία του Little (1988). Ειδικότερα, αν τα δεδομένα είναι MCAR μπορούν να χρησιμοποιηθούν παραδοσιακές και απλές μέθοδοι χειρισμού των ελλιπών δεδομένων, όπως είναι ο περιορισμός της ανάλυσης μόνο στις πλήρεις πειραματικές μονάδες, η αντικατάσταση των ελλιπών τιμών από τη μέση τιμή (mean imputation) ή από την τιμή που προκύπτει με κάποιο μοντέλο παλινδρόμησης (regression imputation). Επιπρόσθετα, ο υπολογισμός εκτιμητών με τη μέθοδο της μέγιστης πιθανοφάνειας καθίσταται απλούστερος, αφού ο όρος που περιλαμβάνει τον μηχανισμό παραγωγής ελλιπών δεδομένων παραλείπεται τόσο υπό την υπόθεση MAR όσο και υπό την υπόθεση MCAR. Επιπρόσθετα, αν ισχύει η υπόθεση MCAR, οι εκτιμώμενοι εκτιμητές είναι συνεπείς, υπό την πρόσθετη υπόθεση ότι οι τέταρτης τάξης ροπές είναι πεπερασμένες. Ακόμα, αν τα δεδομένα είναι MCAR, η ύπαρξη ελλιπών τιμών δεν έχει ως συνέπεια τη μεροληψία, καθώς τα δεδομένα που λείπουν δεν προκαλούν αλλοίωση στα αποτελέσματα της ανάλυσης, διότι δεν υπάρχουν υποκείμενοι παράγοντες που να κάνουν την πιθανότητα ύπαρξης ελλιπούς τιμής να εξαρτάται από τις παρατηρούμενες ή μη παρατηρούμενες τιμές. Επομένως, τα αποτελέσματα της ανάλυσης είναι πιο ακριβή και αξιόπιστα.

Από τα παραπάνω είναι κατανοητό, ότι ο μηχανισμός έλλειψης MCAR μπορεί να θεωρηθεί ιδανικός και για αυτόν τον λόγο είναι απαραίτητο, όταν υπάρχουν ελλιπή δεδομένα, να επιβεβαιώνεται ότι αυτά μπορούν να θεωρηθούν ότι έχουν προέλθει από έναν τέτοιον μηχανισμό. Για τον λόγο αυτό στη βιβλιογραφία, ιδιαίτερα τα τελευταία έτη, παρατηρείται ένα αυξημένο ενδιαφέρον για την εισαγωγή μεθόδων για τον έλεγχο αν ο μηχανισμός έλλειψης των δεδομένων είναι MCAR ή όχι.

## 1.2 Σκοπός και διάρθρωση της μεταπτυχιακής διατριβής

Από όσα προηγήθηκαν έχει γίνει κατανοητή η σπουδαιότητα της υπόθεσης ότι το σύνολο δεδομένων είναι MCAR και επομένως αιτιολογείται το γεγονός ότι για τον έλεγχό της έχουν παρουσιαστεί στη βιβλιογραφία μεθοδολογίες ελέγχου τόσο για τη γενική περίπτωση που τα διαθέσιμα δεδομένα μπορούν να θεωρηθούν ένα τυχαίο δείγμα από τον υπό μελέτη πληθυσμό, αλλά και μεθοδολογίες ειδικά σχεδιασμένες για διαχρονικά δεδομένα (longitudinal data) ή για επαναληπτικές/επαναλαμβανόμενες μετρήσεις (repeated measurements). Επιπρόσθετα

στα παραπάνω κάποιες από τις μεθοδολογίες είναι ειδικά σχεδιασμένες για διακριτά/κατηγορικά δεδομένα. Στο πλαίσιο αυτής της διατριβής, θα περιοριστούμε στην παρουσίαση των μεθοδολογιών εκείνων που είναι σχεδιασμένες για τη γενική περίπτωση που έχουμε διαθέσιμο ένα τυχαίο σύνολο συνεχών δεδομένων, καθώς και εκείνων που έχουν σχεδιαστεί για μικτού τύπου δεδομένα και μπορούν να υλοποιηθούν ακόμα και αν υπάρχουν μόνο συνεχή δεδομένα, χωρίς να απαιτείται διακριτοποίησή τους. Στο πλαίσιο αυτό, εκτός και αν ρητά αναφέρεται κάτι διαφορετικό, θα παρουσιασθούν στατιστικές συναρτήσεις ελέγχους της μηδενικής υπόθεσης ότι τα δεδομένα είναι MCAR έναντι της εναλλακτικής υπόθεσης ότι τα δεδομένα δεν μπορούν να θεωρηθούν ότι είναι MCAR. Πιο συγκεκριμένα, η διάρθρωση των υπόλοιπων κεφαλαίων της διατριβής έχει ως εξής. Στο δεύτερο κεφάλαιο αυτής της μεταπτυχιακής διατριβής θα παρουσιαστούν οι έλεγχοι εκείνοι των οποίων η κεντρική ιδέα είναι η ακόλουθη: τα διαθέσιμα ελλιπή δεδομένα χωρίζονται σε διακεκριμένα μοτίβα και ο έλεγχος της υπόθεσης MCAR ανάγεται στον έλεγχο ότι κάθε ένα από αυτά τα μοτίβα προέρχεται από τον ίδιο πληθυσμό. Αυτό πρακτικά σημαίνει ότι ο έλεγχος της υπόθεσης MCAR ανάγεται στον έλεγχο αν τα δεδομένα ικανοποιούν την υπόθεση της ομοιογένειας. Είναι προφανές ότι απόρριψη της υπόθεσης της ομοιογένειας συνεπάγεται απόρριψη της υπόθεσης MCAR, ενώ αν η υπόθεση της ομοιογένειας δεν μπορεί να απορριφθεί, αυτό δεν σημαίνει ότι επιβεβαιώνεται η υπόθεση MCAR. Όσοι έλεγχοι δεν ανήκουν στην παραπάνω κατηγορία παρουσιάζονται στο τρίτο κεφάλαιο αυτής της διατριβής. Οι έλεγχοι εφαρμόζονται σε ένα σύνολο δεδομένων, με χρήση συναρτήσεων της R που είναι διαθέσιμες σε κάποια πακέτα της, αλλά και άλλων που υλοποιήθηκαν στο πλαίσιο αυτής τη διατριβής. Στο Κεφάλαιο 4 το ενδιαφέρον επικεντρώνεται στην παρουσίαση των αποτελεσμάτων συγκριτικών μελετών που έχουν διεξαχθεί στη βιβλιογραφία και αφορούν την απόδοση των στατιστικών ελέγχων που παρουσιάστηκαν στα προηγούμενα δύο κεφάλαια. Στο Κεφάλαιο 5 συνοψίζονται τα συμπεράσματα αυτής της μεταπτυχιακής διατριβής και δίνονται προτάσεις για περαιτέρω έρευνα. Τέλος, η μεταπτυχιακή διατριβή ολοκληρώνεται με το Παράρτημα, όπου παρατίθενται οι συναρτήσεις υλοποίησης στην R κάποιων εκ των στατιστικών ελέγχων, και τη Βιβλιογραφία.

# ΚΕΦΑΛΑΙΟ 2

# Τεστ ομοιογενειας για τον ελεγχο της υποθεσης MCAR

## 2.1 Εισαγωγή

Παρότι πολλές από τις στατιστικές μεθοδολογίες της Πολυδιάστατης Στατιστικής στηρίζονται στην υπόθεση ότι τα δεδομένα είναι MCAR, ελάχιστες διαδικασίες ελέγχου είχαν εμφανιστεί στη βιβλιογραφία μέχρι τα τέλη της δεκαετίας του 1980. Ειδικότερα, η μέχρι τότε συνήθης διαδικασία, η οποία ήταν διαθέσιμη και στο στατιστικό πρόγραμμα BMDP (Biomedical Data Package), ήταν ο έλεγχος να διεξάγεται μέσω της διενέργειας πολλαπλών $t$ τεστ για την ισότητα δύο πληθυσμιακών μέσων με ανεξάρτητα δείγματα, έχοντας τις παρατηρούμενες τιμές για κάθε μεταβλητή και διαμερίζοντάς τες σε δύο δείγματα ανάλογα με το αν η συγκεκριμένη πειραματική μονάδα εμφανίζει ελλιπή τιμή ή όχι στις υπόλοιπες μεταβλητές. Πιο συγκεκριμένα, όταν οι ελλιπείς τιμές περιορίζονταν σε μία μόνο μεταβλητή, έστω την $Y_i$, διεξάγονται $p-1$ το πλήθος έλεγχοι ισότητας μέσων τιμών με δύο ανεξάρτητα δείγματα για κάθε πλήρως παρατηρούμενη μεταβλητή, με τις πειραματικές μονάδες να χωρίζονται σε δύο ομάδες, ανάλογα με το αν έχει ή δεν έχει καταγραφεί η τιμή της τυχαίας μεταβλητής $Y_i$. Επομένως, σύμφωνα με αυτήν την ιδέα, διεξάγονται $p-1$ το πλήθος $t$ τεστ για τον έλεγχο της ισότητας δύο μέσων τιμών. Στην περίπτωση του ελέγχου MCAR σε ένα $n \times p$ σύνολο δεδομένων με ελλιπείς τιμές σε οποιαδήποτε μεταβλητή, τότε για κάθε μεταβλητή διεξάγονται $(p-1)$ έλεγχοι ισότητας δύο μέσων τιμών με ανεξάρτητα δείγματα, που σε καθέναν τα δύο ανεξάρτητα δείγματα καθορίζονται με βάση την ύπαρξη ή όχι ελλιπούς τιμής σε κάθε μία από τις υπόλοιπες μεταβλητές. Κατά αυτόν τον τρόπο, έχουμε συνολικά $p(p-1)$ το πλήθος $t$ τεστ (βλέπε Dixon and Brown (1983)).

Η παραπάνω διαδικασία, παρότι απλή και εύκολα εφαρμόσιμη, έχει το σοβα-

ρό μειονέκτημα ότι οδηγεί σε αυξημένες πιθανότητες σφαλμάτων τύπου I, που οφείλονται στο γεγονός ότι οι έλεγχοι που διενεργούνται συσχετίζονται (βλέπε Little (1988) και Kim and Bentler (2002)). Το γεγονός αυτό ώθησε, αρχικά τον Little (1988) και στη συνέχεια και άλλους ερευνητές να παρουσιάσουν στη βιβλιογραφία έναν ενιαίο τρόπο ελέγχου της MCAR υπόθεσης που στηρίζεται σε όλα τα διαθέσιμα δεδομένα. Ειδικότερα, σε αυτό το κεφάλαιο της διατριβής θα παρουσιαστούν οι έλεγχοι εκείνοι των οποίων, παρακινούμενοι από την πρωτοπόρα εργασία του Little (1988), η κεντρική ιδέα είναι η ακόλουθη: τα διαθέσιμα ελλιπή δεδομένα χωρίζονται σε διακεκριμένα μοτίβα και ο έλεγχος της υπόθεσης MCAR ανάγεται στον έλεγχο ότι κάθε ένα από αυτά τα μοτίβα προέρχεται από τον ίδιο πληθυσμό. Αυτό πρακτικά σημαίνει ότι ο έλεγχος της υπόθεσης MCAR ανάγεται στον έλεγχο αν τα δεδομένα ικανοποιούν την υπόθεση της ομοιογένειας. Είναι προφανές ότι απόρριψη της υπόθεσης της ομοιογένειας συνεπάγεται απόρριψη της υπόθεσης MCAR, ενώ αν η υπόθεση της ομοιογένειας δεν μπορεί να απορριφθεί αυτό δεν σημαίνει ότι επιβεβαιώνεται η υπόθεση MCAR.

Πριν ολοκληρώσουμε αυτήν την εισαγωγική ενότητα θα δώσουμε κάποιες πληροφορίες για το σύνολο δεδομένων στο οποίο θα υλοποιηθούν οι έλεγχοι που παρουσιάζονται. Πιο συγκεκριμένα, θα χρησιμοποιηθεί το σύνολο δεδομένων `PimaIndiansDiabetes2` που είναι διαθέσιμο στην R, και περιλαμβάνεται στο πακέτο mlbench. Ειδικότερα, το αρχείο αφορά τη διάγνωση διαβήτη σε ένα δείγμα 768 γυναικών από την περιοχή Pima της Ινδίας, πάνω στις ακόλουθες εννιά το πλήθος μεταβλητές:

1. pregnant: ο αριθμός των φορών που μία γυναίκα έχει υπάρξει έγκυος,
2. glucose: η συγκέντρωση πλάσματος γλυκόζης (τεστ ανοχής της γλυκόζης),
3. pressure: η διαστολική πίεση του αίματος (mm Hg),
4. triceps: το πάχος πτυχής δέρματος τρικεφάλου (mm),
5. insulin: η συγκέντρωση ινσουλίνης ορού 2 ωρών (mu U/ml),
6. mass: ο δείκτης μάζας σώματος (βάρος σε kg /(ύψος σε μέτρα)$^2$),
7. pedigree: το σκορ στη συνάρτηση πιθανοφάνειας για προδιάθεση σε διαβήτη με βάση το οικογενειακό ιστορικό,
8. age: ηλικία,
9. diabetes: μία κατηγορική μεταβλητή, που δείχνει την ύπαρξη διαβήτη ή όχι στο άτομο.

Οι έλεγχοι της υπόθεσης MCAR που μελετώνται στην παρούσα διατριβή περιορίζονται σε συνεχείς μεταβλητές. Επομένως, για τη σωστή εφαρμογή τους θα εξαιρέσουμε τις μεταβλητές diabetes και pregnant.

## 2.2 Ο έλεγχος του Little (1988)

Στην ενότητα αυτή το ενδιαφέρον επικεντρώνεται στην παρουσίαση του ελέγχου του Little (1988), που είναι, μέχρι και τις μέρες μας, ο πιο διαδεδομένος τρόπος ελέγχου της μηδενικής υπόθεσης ότι τα δεδομένα είναι MCAR έναντι της εναλλακτικής ότι τα δεδομένα δεν είναι MCAR. Σημειώνεται ότι η εργασία του Little (1988) αποτελεί εργασία αναφοράς σε αυτό το αντικείμενο με περισσότερες από δέκα χιλιάδες αναφορές μέχρι τη συγγραφή αυτής της μεταπτυχιακής διατριβής.

Αρχικά, θα εισαχθεί ο απαραίτητος συμβολισμός για την παρουσίαση του ελέγχου του Little. Σε αυτό το πλαίσιο, σε όσα έπονται:

- $\mathbf{y_i}$ συμβολίζει το $p$-διάστατο διάνυσμα γραμμή στην περίπτωση μη ελλιπών δεδομένων για την $i$-οστή παρατήρηση,
- $\mathbf{r}_i$ είναι το $p$-διάστατο διάνυσμα γραμμή με συνιστώσες 1 και 0 ανάλογα αν η τιμή παρατηρείται ή είναι ελλίπουσα στη συγκεκριμένη θέση της $i$-οστής παρατήρησης,
- $J$ είναι ο αριθμός των διακεκριμένων μοτίβων ελλιπών δεδομένων στο σύνολο των δεδομένων, με τις πλήρεις παρατηρούμενες πειραματικές μονάδες να θεωρούμε ότι σχηματίζουν ένα μοτίβο,
- $S_j$ είναι το σύνολο των πειραματικών μονάδων που ανήκουν στο $j$-οστό μοτίβο, για $j = 1, \ldots, J$,
- $m_j$ είναι το πλήθος των πειραματικών μονάδων που ανήκουν στο σύνολο $S_j$, με $\sum_{j=1}^{j} m_j = n$,
- $p_j$ ο αριθμός των παρατηρούμενων μεταβλητών για τις πειραματικές μονάδες που ανήκουν στο σύνολο $S_j$,
- $\mathbf{D}_j$ είναι ο $p \times p_j$ πίνακας που μας υποδεικνύει ποιες μεταβλητές έχουν παρατηρηθεί στο $j-$οστό μοτίβο. Ο πίνακας αυτός έχει μία στήλη για κάθε παρατηρούμενη μεταβλητή και κάθε στήλη αποτελείται από $p - 1$ μηδενικά και μία μονάδα, με τη μονάδα να δηλώνει τη θέση της μεταβλητής που αναγνωρίστηκε, και τέλος

- $\mathbf{y}_{obs.i}$ συμβολίζει το $p_j$-διάστατο διάνυσμα γραμμή των παρατηρούμενων τιμών, στην περίπτωση που υπάρχουν ελλιπείς τιμές στην $i$-οστή πειραματική μονάδα.

Έπειτα από την εισαγωγή του απαραίτητου συμβολισμού, είμαστε σε θέση να περιγράψουμε όσα παρουσιάστηκαν από τον Little (1988). Ο Little θεωρώντας ότι το διάνυσμα $\mathbf{y_i}$ προέρχεται από έναν πληθυσμό που περιγράφεται ικανοποιητικά από την πολυδιάστατη κανονική κατανομή με παραμέτρους θέσης και κλίμακας $\mu$ και $\mathbf{\Sigma}$, αντίστοιχα, προτείνει αρχικά ένα τεστ πηλίκου πιθανοφανείων για τον έλεγχο της MCAR υπόθεσης, θεωρώντας τη μη ρεαλιστική υπόθεση ότι ο πίνακας $\mathbf{\Sigma}$ είναι γνωστός. Πιο συγκεκριμένα, η προτεινόμενη στατιστική συνάρτηση είναι η

$$d_0^2 = \sum_{j=1}^{J} m_j(\bar{\mathbf{y}}_{obs.j} - \mu^*_{obs.j})\mathbf{\Sigma}^{-1}_{obs.j}(\bar{\mathbf{y}}_{obs.j} - \mu^*_{obs.j})^T. \tag{2.1}$$

Σε αυτήν τη στατιστική συνάρτηση κατέληξε χρησιμοποιώντας το σκεπτικό που θα περιγραφεί στη συνέχεια. Αν τα δεδομένα είναι MCAR, τότε για τη δεσμευμένη κατανομή του $\mathbf{y}_{obs,i}$ δοθέντος του $\mathbf{r}_i$ ισχύει ότι:

$$(\mathbf{y}_{obs.i}|\mathbf{r}_i) \sim \mathcal{N}(\mu_{obs.j},\ \mathbf{\Sigma}_{obs.j}),\ \text{για } i \in S_j, 1 \leq j \leq J, \tag{2.2}$$

όπου $\mu_{obs.j}$ είναι το $1 \times p_j$ διάνυσμα των μέσων των παρατηρούμενων μεταβλητών του $j-$οστού μοτίβου και $\mathbf{\Sigma}_{obs.j}$ είναι ο $p_j \times p_j$ πίνακας διακυμάνσεων συνδιακυμάνσεων των παρατηρούμενων μεταβλητών του $j-$οστού μοτίβου, με

$$\mu_{obs.j} = \mu\mathbf{D}_j,$$

και

$$\mathbf{\Sigma}_{obs.j} = \mathbf{D}_j^T\mathbf{\Sigma}\mathbf{D}_j,$$

αντίστοιχα.

Από την άλλη, αν τα δεδομένα δεν είναι MCAR, τότε για τη δεσμευμένη κατανομή του $\mathbf{y}_{obs.i}$ δοθέντος του $\mathbf{r}_i$ ισχύει ότι τα μέσα διανύσματα μπορούν να διαφοροποιούνται μεταξύ των μοτίβων, δηλαδή τώρα ισχύει ότι:

$$(\mathbf{y}_{obs.i}|\mathbf{r}_i) \sim \mathcal{N}(\mathbf{v}_{obs.j},\ \mathbf{\Sigma}_{obs.j}), i \in S_j, 1 \leq j \leq J, \tag{2.3}$$

με $\mathbf{v}_{obs.j}$ να είναι το $1 \times p_j$ διάνυσμα των μέσων των παρατηρούμενων μεταβλητών του $j$-οστού μοτίβου, με $j = 1, \ldots, J$. Επισημαίνεται ότι η διαφοροποίηση στις κατανομές των σχέσεων (2.2) και (2.3) οφείλεται στην ιδιότητα MCAR και αιτιολογείται λαμβάνοντας υπόψη ότι:

$$\begin{aligned} Pr(\mathbf{y}_{obs.i}|\mathbf{r}_i) &= Pr(\mathbf{y}_{obs.i}, \mathbf{r}_i)Pr(\mathbf{r}_i) = \Pr(\mathbf{r}_i|\mathbf{y}_{obs.i})Pr(\mathbf{y}_{obs.i})/Pr(\mathbf{r}_i) \\ &\overset{MCAR}{=} Pr(\mathbf{y}_{obs.i}). \end{aligned}$$

Στο πλαίσιο αυτό, υιοθετώντας τη δεσμευμένη κατανομή της σχέσης (2.3), υπό την ανεξαρτησία των δεδομένων, προκύπτει ότι ο λογάριθμος της συνάρτησης πιθανοφάνειας δίνεται από τη σχέση:

$$l(\mathbf{v}|\mathbf{y}_{obs}) = C - \frac{1}{2}\sum_{j=1}^{J} m_j(\bar{\mathbf{y}}_{obs.j} - \mathbf{v}_{obs.j})\mathbf{\Sigma}_{obs.j}^{-1}(\bar{\mathbf{y}}_{obs.j} - \mathbf{v}_{obs.j})^T, \qquad (2.4)$$

όπου $C$ σταθερά, $\mathbf{v}$ το σύνολο που περιέχει όλα τα άγνωστα μέσα διανύσματα των $J$ το πλήθος μοτίβων και $\bar{\mathbf{y}}_{obs.j}$ το $1 \times p_j$ διάνυσμα με

$$\bar{\mathbf{y}}_{obs.j} = m_j^{-1}\sum_{i \in S_j} \mathbf{y}_{obs.i}.$$

Η σχέση (2.4) ως τετραγωνική μορφή μεγιστοποιείται ως προς $\mathbf{v}_{obs.j}$ για

$$\hat{\mathbf{v}}_{obs.j} = \bar{\mathbf{y}}_{obs.j}$$

και τότε ισχύει ότι $l(\hat{\mathbf{v}}_{obs.j}|\mathbf{y}_{obs.j}) = 0$.

Από την άλλη, υιοθετώντας ότι ισχύει η σχέση (2.2), αν $\mu^*$ είναι ο εκτιμητής μέγιστης πιθανοφάνειας (ΕΜΠ) της παραμέτρου $\mu$, ο οποίος προσδιορίζεται μέσω του EM (Expectation-Maximization) αλγόριθμου, τότε, θέτοντας $\mu_{obs.j}^* \equiv \mu^* D_j$, έχουμε ότι:

$$l(\mu_{obs.j}^*|\mathbf{y}_{obs.j}) = C - \frac{1}{2}d_0^2.$$

Συνδυάζοντας τα παραπάνω προκύπτει ότι το τεστ πηλίκου πιθανοφανείων δίνεται από τη σχέση:

$$-2[l(\mu_{obs.j}^*|\mathbf{y}_{obs.j}) - l(\hat{\mathbf{v}}_{obs.j}|\mathbf{y}_{obs.j})] = d_0^2,$$

δηλαδή ταυτίζεται με τη στατιστική συνάρτηση (σ.σ.) της σχέσης (2.1).

Για να μπορεί να χρησιμοποιηθεί η σ.σ. $d_0^2$ για τον έλεγχο της MCAR υπόθεσης θα πρέπει να προσδιοριστεί η κατανομή της υπό τη μηδενική υπόθεση. Υπό την υπόθεση της κανονικότητας και υπό τη μηδενική υπόθεση ότι τα δεδομένα είναι MCAR, ο Little (1988) απέδειξε ότι η στατιστική συνάρτηση $d_0^2$ ακολουθεί χι-τετράγωνο κατανομή με $f_1 = \sum_{j=1}^{J} p_j - p$ βαθμούς ελευθερίας. Επιπρόσθετα, απέδειξε ότι αν αρθεί η υπόθεση της κανονικότητας, τότε για μεγάλα σε μέγεθος δείγματα, η στατιστική συνάρτηση $d_0^2$ ακολουθεί ασυμπτωτικά την παραπάνω κατανομή. Επιπλέον, από τον τρόπο κατασκευής της σ.σ. προκύπτει ότι μεγάλες τιμές της συνηγορούν υπέρ της απόρριψης της υπόθεσης ότι

τα δεδομένα είναι MCAR, καθώς τότε αναμένουμε μεγάλες διαφορές μεταξύ των δειγματικών μέσων και των ΕΜΠ που προσδιορίζονται υπό την υπόθεση των MCAR δεδομένων. Καθώς η προς έλεγχο υπόθεση απορρίπτεται για μεγάλες τιμές της σ.σ. ελέγχου, έχουμε ότι, σε (ασυμπτωτικό) επίπεδο σημαντικότητας $\alpha$, η υπόθεση ότι τα δεδομένα είναι MCAR απορρίπτεται αν $d_0^2 \geq \chi^2_{f_1,\alpha}$, με $\chi^2_{f_1,\alpha}$ να είναι το άνω $\alpha$-ποσοστιαίο σημείο της χι τετράγωνο κατανομή με $f_1$ βαθμούς ελευθερίας.

Ο παραπάνω έλεγχος ωστόσο δεν έχει μεγάλη πρακτική αξία, καθώς για να χρησιμοποιηθεί πρέπει να γνωρίζουμε τον πληθυσμιακό πίνακα διακυμάνσεων συνδιακυμάνσεων $\mathbf{\Sigma}$. Ωστόσο, αποτέλεσε τη βάση για τη στατιστική συνάρτηση που προτάθηκε από τον Little (1988) για την πιο γενική και ενδιαφέρουσα περίπτωση που ο πίνακας $\mathbf{\Sigma}$ είναι άγνωστος.

Πιο συγκεκριμένα, ο Little τροποποιεί κατάλληλα αυτόν τον έλεγχο και τον γενικεύει στη συνήθη περίπτωση που ο πίνακας διακυμάνσεων συνδιακυμάνσεων είναι άγνωστος, προτείνοντας την αντικατάσταση των $\mu^*$ και

$$\mathbf{\Sigma}$$

από τους εκτιμητές $\hat{\mu}$ και $\tilde{\mathbf{\Sigma}} = \frac{n}{n-1}\hat{\mathbf{\Sigma}}$, όπου $\hat{\mu}$ και $\hat{\mathbf{\Sigma}}$ οι ΕΜΠ των παραμέτρων $\mu$ και $\mathbf{\Sigma}$, αντίστοιχα, που υπολογίζονται μέσω του ΕΜ αλγορίθμου. Επομένως, η σ.σ. που προτάθηκε από τον Little για την πιο γενική περίπτωση δίνεται από τη σχέση:

$$d^2 = \sum_{j=1}^{J} m_j(\bar{\mathbf{y}}_{obs.j} - \hat{\mu}_{obs.j})\tilde{\mathbf{\Sigma}}_{obs.j}^{-1}(\bar{\mathbf{y}}_{obs.j} - \hat{\mu}_{obs.j})^T, \tag{2.5}$$

όπου $\hat{\mu}_{obs.j} = \hat{\mu}D_j$, ενώ $\tilde{\mathbf{\Sigma}}_{obs.j}^{-1} = (\mathbf{D}_j^T\tilde{\mathbf{\Sigma}}\mathbf{D}_j)^{-1}$.

Επισημαίνεται ότι υποθέτοντας ότι τα δεδομένα είναι MCAR και η κατανομή των $\mathbf{y}_i$ έχει τέταρτης τάξης πεπερασμένες ροπές, προκύπτει ότι ο εκτιμητής $\tilde{\mathbf{\Sigma}}$ είναι συνεπής εκτιμητής της παραμέτρου $\mathbf{\Sigma}$ και επομένως η σ.σ. $d^2$ ακολουθεί ασυμπτωτικά τη χι-τετράγωνο κατανομή με $f_1$ βαθμούς ελευθερίας, ήτοι ακολουθεί την ίδια κατανομή με τη σ.σ. $d_0^2$. Με παρόμοιο σκεπτικό με αυτό που προαναφέρθηκε, η υπόθεση ότι τα δεδομένα είναι MCAR απορρίπτεται, σε ασυμπτωτικό επίπεδο σημαντικότητας $\alpha$, αν $d^2 \geq \chi^2_{f_1,\alpha}$.

**Παρατήρηση 2.2.1** *Ένα μειονέκτημα του ελέγχου του Little αποτελεί το γεγονός ότι είναι περισσότερο κατάλληλος όταν τα διαθέσιμα δεδομένα είναι ποσοτικά. Για τον λόγο αυτό ο Little (1988) προτείνει τη χρήση του ελέγχου του Fuchs (1982) όταν τα δεδομένα είναι κατηγορικά. Επιπλέον, τα αποτελέσματα*

*που παρουσιάσθηκαν για τη γενική περίπτωση ισχύουν για μεγάλα σε μέγεθος δείγματα. Η κατανομή της σ.σ. $d^2$, υπό τη μηδενική υπόθεση, για μικρά σε μέγεθος δείγματα είναι ιδιαίτερα περίπλοκη για τη γενική περίπτωση ελλιπών δεδομένων. Ωστόσο, ο Little (1988) κατάφερε να προσδιορίσει την κατανομή της σ.σ. $d^2$ υπό τη μηδενική υπόθεση ότι τα δεδομένα είναι MCAR για την περίπτωση που τα ελλιπή δεδομένα είναι $k$ βηματικά μονότονα. Ειδικότερα, απέδειξε ότι για μικρά σε μέγεθος δείγματα, η κατανομή της σ.σ. είναι άθροισμα συναρτήσεων $F$ τυχαίων μεταβλητών.*

**Παρατήρηση 2.2.2** *Ο Little (1988) στην εργασία του παρουσίασε τα αποτελέσματα μίας μελέτης προσομοίωσης που αφορά την απόδοση του ελέγχου ως προς τη διατήρηση του επιπέδου σημαντικότητας στη βάση 1000 προσομοιωμένων συνόλων δεδομένων μεγέθους $n = 20, 40, 80$, σε $p = 4$ μεταβλητές. Ειδικότερα, τα σύνολα αυτά απαρτίζονται από συνολικά επτά ομάδες που δημιουργούνται με βάση τα διάφορα μοτίβα ελλιπών τιμών, με την ομάδα $S_1$ που περιέχει τις πλήρως παρατηρούμενες πειραματικές μονάδες να αποτελεί το 40% του συνολικού διαθέσιμου δείγματος σε κάθε προσομοιωμένο σύνολο δεδομένων. Αντίστοιχα οι ομάδες $S_2, ..., S_6$, αποτελούν η κάθε μία το 10% του διαθέσιμου δείγματος και έχουν τα μοτίβα ελλιπών τιμών: $r_1 = (0, 0, 0, 1)$, $r_2 = (0, 0, 1, 1)$, $r_3 = (0, 0, 1, 0)$, $r_4 = (0, 1, 1, 0)$, $r_5 = (0, 1, 0, 0)$ και $r_6 = (0, 1, 0, 1)$. Επιπλέον, τα δεδομένα παράγονται από τρεις διαφορετικές κατανομές: την πολυδιάστατη κανονική κατανομή, τη λογαριθμοκανονική και την $t$ κατανομή με 3 βαθμούς ελευθερίας. Από τα αποτελέσματα της προσομοίωσης συμπεραίνουμε ότι καθώς το μέγεθος του δείγματος αυξάνει το εμπειρικό επίπεδο σημαντικότητας είναι κοντά στο ονομαστικό, χωρίς να παίζει ρόλο η κατανομή από την οποία προέρχονται τα δεδομένα. Το γεγονός αυτό επιβεβαιώνει ότι για μεγάλο μέγεθος δείγματος η υπόθεση της κανονικότητας μπορεί να παραληφθεί. Από την άλλη για μικρό μέγεθος δείγματος ο έλεγχος φαίνεται να είναι συντηρητικός.*

**Εφαρμογή στα δεδομένα PimaIndiansDiabetes2** Ο έλεγχος του Little υλοποιείται στη γλώσσα R με χρήση της συνάρτησης `mcar_test(data)`, που είναι διαθέσιμη στο πακέτο `naniar`. Απαραίτητη προϋπόθεση για τη λειτουργία της συνάρτησης είναι το όρισμα των δεδομένων της συνάρτησης (ορίζεται ως data στη συνάρτηση) να ανήκει στην κλάση data.frame() ή matrix() και οι ελλιπείς τιμές να έχουν οριστεί ως NA. Επισημαίνεται ότι αν κάτι τέτοιο δεν έχει γίνει από τον χρήστη, μπορεί να επιτευχθεί μέσω του ορίσματος as.na της συνάρτησης. Η συνάρτηση αυτή δίνει ως έξοδο την τιμή της στατιστικής συνάρτησης του Little (στήλη statistic), τους βαθμούς ελευθερίας της $\chi^2$ κατανομής (υπό τη στήλη df), ενώ η p-τιμή του ελέγχου βρίσκεται κάτω από τη στήλη p-

value. Τέλος, κάτω από τη στήλη missing.patterns, παρατίθεται το πλήθος των μοτίβων των ελλιπών τιμών. Για το σύνολο δεδομένων PimaIndiansDiabetes2 προκύπτει ότι η υπόθεση MCAR απορρίπτεται καθώς η $p$-τιμή του ελέγχου είναι ίση με $1.50e-11$.

## 2.3 Οι έλεγχοι των Kim and Bentler (2002)

Παρότι ο έλεγχος του Little αποτελεί εδώ και πολλά έτη τη μέθοδο αναφοράς για τον έλεγχο της MCAR υπόθεσης παρουσιάζει ένα σημαντικό μειονέκτημα. Πιο συγκεκριμένα, παρατηρώντας τη σχέση (2.3), είναι φανερό ότι η εναλλακτική υπόθεση επιτρέπει η έλλειψη των δεδομένων να επηρεάζει τα μέσα διανύσματα, αλλά επιβάλλει τον περιορισμό οι διακυμάνσεις και οι συνδιακυμάνσεις να είναι ίδιες για όλα τα μοτίβα. Επομένως, ουσιαστικά πρόκειται για έναν έλεγχο της ομοιογένειας των μέσων τιμών ως προς τα μέσα διανύσματα των μοτίβων και δεν εξετάζει την ομοιογένεια ως προς τον πίνακα διακυμάνσεων συνδιακυμάνσεων. Το μειονέκτημα αυτό επισημάνθηκε και από τον ίδιον τον Little και το αντιμετώπισε προτείνοντας μια διαφορετική μοντελοποίηση για τη δεσμευμένη κατανομή του $\mathbf{y}_{obs.i}$ δοθέντος του $\mathbf{r}_i$ όταν δεν ικανοποιείται η MCAR υπόθεση. Ειδικότερα, σε αυτήν την περίπτωση, η δεσμευμένη κατανομή του $\mathbf{y}_{obs.i}$ δοθέντος του $\mathbf{r}_i$ δεν προσδιορίζεται από τη σχέση (2.3), αλλά από τη σχέση:

$$(\mathbf{y}_{obs.i}|\mathbf{r}_i) \sim \mathcal{N}(\mathbf{v}_{obs.j},\ \mathbf{\Gamma}_{obs.j}),\ \text{για } i \in S_j, 1 \leq j \leq J. \tag{2.6}$$

Στο παραπάνω πλαίσιο και ακολουθώντας τη μέθοδο του πηλίκου πιθανοφανειών, ο Little (1988) κατέληξε να προτείνει για τον έλεγχο της MCAR υπόθεσης τη στατιστική συνάρτηση:

$$d^2_{aug} = d^2 + n\sum_{j=1}^{J} m_j(tr[\mathbf{S}_{obs.j}\hat{\mathbf{\Sigma}}^{-1}_{obs.j}] - log|\mathbf{S}^{-1}_{obs.j}| + log|\hat{\mathbf{\Sigma}}^{-1}_{obs.j}| - p_j), \tag{2.7}$$

όπου με $S_{obs.j}$ συμβολίζουμε τον δειγματικό πίνακα συνδιακυμάνσεων των πειραματικών μονάδων που ανήκουν στο $j$-οστό μοτίβο. Το επόμενο βήμα για τη διενέργεια του ελέγχου είναι ο προσδιορισμός της κατανομής της σ.σ. υπό την υπόθεση ότι τα δεδομένα είναι MCAR. Αποδεικνύεται ότι η σ.σ. $d^2_{aug}$ ακολουθεί ασυμπτωτικά χι-τετράγωνο κατανομή με $f_3 = \sum_{j=1}^{J} p_j(p_j+3)/2 - p(p+3)/2$ βαθμούς ελευθερίας. Καθώς η προς έλεγχο υπόθεση απορρίπτεται για μεγάλες τιμές της σ.σ. ελέγχου, έχουμε ότι σε (ασυμπτωτικό) επίπεδο σημαντικότητας $\alpha$ η υπόθεση ότι τα δεδομένα είναι MCAR απορρίπτεται αν $d^2_{aug} \geq \chi^2_{f_3,\alpha}$.

Από τα παραπάνω είναι φανερό ότι η σ.σ. $d^2$ προέκυψε ως έλεγχος της ομοιογένειας των μέσων τιμών, ενώ η σ.σ. $d^2_{aug}$ προέκυψε ελέγχοντας την ομοιογένεια τόσο των μέσων τιμών όσο και των πινάκων διακυμάνσεων συνδιακυμάνσεων. Η παραπάνω παρατήρηση ώθησε τους Kim and Bentler (2002) να προτείνουν για τον έλεγχο της υπόθεσης MCAR τη στατιστική συνάρτηση που προκύπτει ως διαφορά των σ.σ. $d^2_{aug} - d^2$, η οποία ουσιαστικά ελέγχει την ομοιογένεια των πινάκων διακυμάνσεων συνδιακυμάνσεων. Πιο συγκεκριμένα, η σ.σ. που προκύπτει δίνεται από τη σχέση:

$$d^2_{cov} = d^2_{aug} - d^2 = \sum_{j=1}^{J} m_j (tr[\mathbf{S}_{obs.j}\hat{\mathbf{\Sigma}}^{-1}_{obs.j}] - log|\mathbf{S}^{-1}_{obs.j}| + log|\hat{\mathbf{\Sigma}}^{-1}_{obs.j}| - p_j). \quad (2.8)$$

Το επόμενο βήμα για τη διενέργεια του ελέγχου είναι ο προσδιορισμός της κατανομής της σ.σ. υπό την υπόθεση ότι τα δεδομένα είναι MCAR. Αποδεικνύεται ότι η σ.σ. $d^2_{cov}$ ακολουθεί ασυμπτωτικά χι-τετράγωνο κατανομή με $f_2 = \sum_{j=1}^{J} p_j(p_j+3)/2 - p(p+3)/2$ βαθμούς ελευθερίας. Καθώς η προς έλεγχο υπόθεση απορρίπτεται για μεγάλες τιμές της σ.σ. ελέγχου, έχουμε ότι σε (ασυμπτωτικό) επίπεδο σημαντικότητας $\alpha$ η υπόθεση ότι τα δεδομένα είναι MCAR απορρίπτεται αν $d^2_{cov} \geq \chi^2_{f_2,\alpha}$.

**Παρατήρηση 2.3.1** *Οι Kim and Bentler (2002) επισημαίνουν ότι παρότι η σ.σ. $d^2_{cov}$ δεν έχει προηγούμενα προταθεί στη βιβλιογραφία, μπορεί να προκύψει ως ειδική περίπτωση του ελέγχου που προτάθηκε στην εργασία των Tang and Bentler (1998) για έναν κοινό πίνακα διακυμάνσεων για κάθε μοτίβο ελλιπών τιμών.*

Οι έλεγχοι που παρουσιάστηκαν σε αυτήν την ενότητα δεν προτείνονται να χρησιμοποιούνται στην πράξη, καθώς παρουσιάζουν σημαντικά μειονεκτήματα, όπως αυτά αναδείχθηκαν τόσο από τον Little (1988) όσο και από τους Kim and Bentler (2002). Ειδικότερα, όπως αναφέρει ο Little (1988), κατά την υλοποίηση του ελέγχου με τη σ.σ. $d^2_{aug}$ σε περιπτώσεις όπου ο αριθμός των πειραματικών μονάδων είναι μικρότερος από τον αριθμό των μεταβλητών που παρατηρήθηκαν στην ομάδα που διέπεται από ένα συγκεκριμένο μοτίβο ελλιπών τιμών, ο δειγματικός πίνακας διακυμάνσεων συνδιακυμάνσεων $\mathbf{S}_{obs.j}$ θα είναι μη αντιστρέψιμος. Το γεγονός αυτό έχει ως επακόλουθο ότι ο πίνακας $\mathbf{S}^{-1}_{obs.j}$ δεν μπορεί να υπολογιστεί, γεγονός που θα έχει ως συνέπεια να μην είναι εφικτή η υλοποίηση του στατιστικού ελέγχου για τη συγκεκριμένη ομάδα. Άρα, πρακτικά, για την αντιμετώπιση τέτοιου είδους προβλημάτων θα πρέπει να αποκλειστούν δεδομένα από τα μοτίβα εκείνα όπου το πλήθος των παρατηρήσεων τους $m_j$ είναι μικρότερο

από το πλήθος των μεταβλητών $p_j$. Προφανώς, παρόμοιο σχόλιο ισχύει και για τον έλεγχο που στηρίζεται στη σ.σ. $d^2_{cov}$. Η παραπάνω παρατήρηση μας οδηγεί έμμεσα στο συμπέρασμα ότι για δείγματα μικρού ή μεσαίου μεγέθους, οι παραπάνω έλεγχοι ενδεχομένως να είναι αναξιόπιστοι. Η αναξιοπιστία αυτή οφείλεται στην ενδεχόμενη ύπαρξη πολλών ομάδων με μικρό αριθμό πειραματικών μονάδων. Άρα η χρήση των παραπάνω στατιστικών δύναται να προταθεί μόνο για μεγάλο μέγεθος δείγματος. Επιπλέον, οι Kim and Bentler (2002), μέσω προσομοίωσης, επιβεβαίωσαν την υποψία που διατύπωσε ο Little (1988) ότι ο έλεγχος με τη σ.σ. $d^2_{aug}$ διατηρεί το ονομαστικό επίπεδο σημαντικότητας μόνο αν το μέγεθος του δείγματος είναι αρκετά μεγάλο.

Θέλοντας να ξεπεραστούν τα παραπάνω προβλήματα, οι Kim and Bentler (2002) παρουσίασαν στη βιβλιογραφία τρεις ελέγχους της ομοιογένειας (μέσων τιμών, πινάκων διακυμάνσεων συνδιακυμάνσεων και συνδυασμό αυτών) βασιζόμενοι στη θεωρία των γενικευμένων ελαχίστων τετραγώνων. Πιο συγκεκριμένα, οι προτεινόμενες σ.σ. προκύπτουν ως παραλλαγές ή γενικεύσεις των ελέγχων που έχουν μελετηθεί στη βιβλιογραφία, μεταξύ άλλων, από τους Jennrich (1970), Nagao (1973) και Lee and Tsui (1982).

Σε αυτό το πλαίσιο, για τον έλεγχο της ομοιογένειας των μέσων διανυσμάτων, οι Kim and Bentler (2002) καταλήγουν, μέσω της θεωρίας των γενικευμένων ελαχίστων τετραγώνων, να προτείνουν τη σ.σ.

$$G_1 = \sum_{j=1}^{J} \frac{m_j}{n} (\bar{\mathbf{y}}_{obs.j} - \hat{\mu}_{obs.j}) \hat{\mathbf{\Sigma}}^{-1}_{obs.j} (\bar{\mathbf{y}}_{obs.j} - \hat{\mu}_{obs.j})^T, \tag{2.9}$$

η οποία διαφοροποιείται ελάχιστα από τη σ.σ. $d^2$ που προτάθηκε από τον Little (1988). Οι Kim and Bentler (2002) αποδεικνύουν ότι η σ.σ. $nG_1$, υπό την υπόθεση ότι τα δεδομένα είναι MCAR, ακολουθεί ασυμπτωτικά χι-τετράγωνο κατανομή με $f_1 = \sum_{j=1}^{J} p_j - p$ βαθμούς ελευθερίας. Επομένως, η υπόθεση ότι τα δεδομένα είναι MCAR απορρίπτεται, σε (ασυμπτωτικό) επίπεδο σημαντικότητας $\alpha$, αν $nG_1 \geq \chi^2_{f_1,\alpha}$.

Από την άλλη, για τον έλεγχο της ομοιογένειας των πινάκων διακυμάνσεων συνδιακυμάνσεων, οι Kim and Bentler (2002) προτείνουν τη σ.σ.

$$G_2 = \sum_{j=1}^{J} \frac{k_j}{2} tr([\mathbf{S}_{obs.j} - \hat{\mathbf{\Sigma}}_{obs.j}] \hat{\mathbf{\Sigma}}^{-1}_{obs.j})^2, \tag{2.10}$$

όπου $k_j = \frac{m_j - 1}{n}$. Καθώς αποδεικνύουν ότι η σ.σ. $nG_2$, υπό την υπόθεση ότι τα δεδομένα είναι MCAR, ακολουθεί ασυμπτωτικά χι-τετράγωνο κατανομή με

$f_2 = \sum_{j=1}^{J} p_j(p_j+1)/2 - p(p+1)/2$ βαθμούς ελευθερίας, έχουμε ότι σε (ασυμπτωτικό) επίπεδο σημαντικότητας $\alpha$ η υπόθεση ότι τα δεδομένα είναι MCAR απορρίπτεται αν $nG_2 \geq \chi^2_{f_2,\alpha}$. Σημειώνεται ότι η πιθανή μη αντιστρεψιμότητα του πίνακα $\mathbf{S}_{obs.j}$ δεν επηρεάζει αυτήν τη σ.σ., καθώς στον ορισμό της δεν περιλαμβάνεται ο αντίστροφος αυτού του πίνακα. Παρατηρήστε ότι απαιτείται μόνο να είναι αντιστρέψιμος ο πίνακας $\hat{\mathbf{\Sigma}}_{obs.j}$, κάτι που συμβαίνει αν ο πίνακας $\hat{\mathbf{\Sigma}}$ είναι αντιστρέψιμος.

Τέλος, ο συνδυασμός των δύο παραπάνω ελέγχων οδήγησε τους Kim and Bentler (2002) στη σ.σ.

$$G_3 = G_1 + G_2, \tag{2.11}$$

η οποία ακολουθεί ασυμπτωτικά, υπό την υπόθεση ότι τα δεδομένα είναι MCAR, χι-τετράγωνο κατανομή με $f_3 = \sum_{j=1}^{J} p_j(p_j+3)/2 - p(p+3)/2$ βαθμούς ελευθερίας. Επομένως, η υπόθεση ότι τα δεδομένα είναι MCAR απορρίπτεται, σε (ασυμπτωτικό) επίπεδο σημαντικότητας $\alpha$, αν $G_3 \geq \chi^2_{f_3,\alpha}$.

Συνοψίζοντας υπάρχουν δύο έλεγχοι ομοιογένειας μέσων τιμών ($d^2$,$G_1$), δύο έλεγχοι ομοιογένειας πινάκων διακυμάνσεων συνδιακυμάνσεων ($d^2_{cov}, G_2$) και δύο έλεγχοι ομοιογένειας μέσων και πινάκων διακυμάνσεων συνδιακυμάνσεων ($d^2_{aug}$,$G_3$), που έχουν προσδιοριστεί με τη μέθοδο της πιθανοφάνειας και των γενικευμένων ελαχίστων τετραγώνων. Για μεγάλα σε μέγεθος δείγματα, υπό τη μηδενική υπόθεση και την υπόθεση της κανονικότητας, οι έλεγχοι αυτοί, σύμφωνα με τους Kim and Bentler (2002), θα πρέπει να έχουν παρόμοια, ισοδύναμη απόδοση. Ωστόσο, για μικρά σε μέγεθος δείγματα μπορεί η απόδοσή τους να διαφοροποιείται.

**Παρατήρηση 2.3.2** *Οι Kim and Bentler (2002) συμπέραναν, μέσω μίας μελέτης προσομοίωσης που υλοποίησαν, ότι οι έλεγχοι που πρότειναν και ειδικότερα εκείνοι για τον έλεγχο της ομοσκεδαστικότητας υπερτερούν έναντι αυτών που προτείνει ο Little (1988). Ωστόσο, σε μελέτες προσομοίωσης που υλοποιήθηκαν τόσο από τους Bentler et al. (2004), όσο και από τους Jamshidian and Jalal (2010) προέκυψε το συμπέρασμα ότι οι έλεγχοι των Kim and Bentler (2002) δεν διατηρούν το ονομαστικό επίπεδο σημαντικότητας. Πιο συγκεκριμένα, μία εκ των αιτιών της μη διατήρησης παρατηρήθηκε να είναι το μικρό μέγεθος δείγματος, ενώ μία επιπλέον αιτία της ασυνέπειας του ονομαστικού επιπέδου, οφείλεται στην απόκλιση των δεδομένων από την πολυδιάστατη κανονική κατανομή. Συγκεκριμένα, όταν τα δεδομένα προέρχονται από κατανομές με βαριές ουρές (heavy tailed), το εμπειρικό επίπεδο σημαντικότητας είναι πολύ μεγαλύτερο σε σχέση με το εκ των προτέρων προσδιορισμένο. Αντίθετα, σε περιπτώσεις όπου ο κύριος όγκος των δεδομένων της κατανομής δεν κείτεται στις ουρές (short tai-*

*led), το αντίστοιχο εκτιμώμενο ονομαστικό επίπεδο είναι κατώτερο του εκ των προτέρων προσδιορισμένου. Από όσα αναφέρθηκαν προηγούμενα είναι σαφές ότι ο έλεγχος αυτός καλό είναι να αποφεύγεται σε πρακτικές εφαρμογές και να προτιμώνται άλλοι. Για τον λόγο αυτόν και δεν παρατίθεται κώδικας υλοποίησής του και δεν εφαρμόζεται στα δεδομένα PimaIndiansDiabetes2.*

## 2.4 Ο έλεγχος των Jamshidian and Jalal (2010)

Οι Jamshidian and Jalal (2010) έχοντας ως στόχο να παρουσιάσουν έναν έλεγχο της MCAR υπόθεσης, ο οποίος θα έχει καλή απόδοση τόσο όταν το πλήθος των πειραματικών μονάδων που ανήκουν σε κάθε μοτίβο ελλιπών δεδομένων είναι μικρό όσο και όταν τα δεδομένα αποκλίνουν από την υπόθεση της κανονικότητας, πρότειναν δύο διαφορετικές μεθοδολογίες, οι οποίες ανήκουν στην κατηγορία των ελέγχων ομοιογένειας των πινάκων διακυμάνσεων συνδιακυμάνσεων. Οι έλεγχοι αυτοί θα παρουσιαστούν συνοπτικά στην παρούσα ενότητα.

Στο πλαίσιο αυτό, η κεντρική ιδέα των Jamshidian and Jalal (2010) ήταν να συμπληρώσουν, αρχικά, τα ελλιπή δεδομένα ακολουθώντας μία μέθοδο συμπλήρωσης αυτών (imputation) και στη συνέχεια να εφαρμόσουν στα πλήρη δεδομένα που προκύπτουν μία υπάρχουσα μεθοδολογία ελέγχου της ομοσκεδαστικότητας, έστω $g$, το πλήθος πολυδιάστατων κανονικών κατανομών. Όλα αυτά με την απαραίτητη επισήμανση ότι η μεθοδολογία που θα επιλεγεί θα πρέπει να έχει καλή απόδοση ακόμη και για μικρά σε μέγεθος δείγματα από αυτούς τους πληθυσμούς.

Από τα παραπάνω είναι προφανές ότι δύο είναι τα ερωτήματα που προκύπτουν: πώς θα συμπληρωθούν τα ελλιπή δεδομένα και ποια μέθοδος θα χρησιμοποιηθεί για τον έλεγχο της ομοσκεδαστικότητας στο πλήρες πλέον σύνολο δεδομένων που θα προκύψει. Όσον αφορά το πρώτο ερώτημα η απάντηση που έδωσαν οι Jamshidian and Jalal (2010) εξαρτάται από το αν η κατανομή των δεδομένων μπορεί να θεωρηθεί ότι είναι η πολυδιάστατη κανονική ή όχι και έτσι προτείνουν έναν παραμετρικό και έναν μη παραμετρικό έλεγχο. Όσον αφορά το δεύτερο ερώτημα σε κάθε περίπτωση υιοθετούν τον έλεγχο του Hawkins (1981) διαφοροποιώντας ωστόσο τον τρόπο λήψης απόφασης. Στη συνέχεια θα παραθέσουμε αναλυτικότερα τις απαντήσεις στα πάνω ερωτήματα για κάθε περίπτωση.

Ως προς το πρώτο ερώτημα που αφορά τον τρόπο συμπλήρωσης των ελλιπών τιμών, αν δεν επαληθεύεται η υπόθεση της πολυδιάστατης κανονικότητας χρησιμοποιείται μία μέθοδος που απαιτεί μόνο την ανεξαρτησία των πειραματικών μονάδων. Πιο συγκεκριμένα, οι Jamshidian and Jalal (2010) προτείνουν να χρησιμοποιείται μία μέθοδος συμπλήρωσης παρόμοια με αυτήν των Srivastava

and Dolatabadi (2009). Για λεπτομέρειες για αυτήν τη μέθοδο παραπέμπουμε στην ενότητα 3.2 της εργασίας τους. Στο πλαίσιο της παρούσας μεταπτυχιακής διατριβής, απλά θα αναφέρουμε ότι οι τιμές που συμπληρώνουν τις ελλιπείς τιμές προκύπτουν προσθέτοντας ένα κατάλληλο τυχαίο σφάλμα στις καλύτερες γραμμικές προβλέψεις των ελλιπών τιμών. Για να εφαρμοστεί αυτή η μέθοδος, απαιτείται μια εκτίμηση του μέσου διανύσματος και του πίνακα διακυμάνσεων συνδιακυμάνσεων, που προσδιορίζεται από τις πλήρως παρατηρούμενες πειραματικές μονάδες. Η διαδικασία αυτή υλοποιείται στο πακέτο `MissMech` της γλώσσας προγραμματισμού R.

Στην περίπτωση τώρα που η υπόθεση της πολυδιάστατης κανονικότητας δεν μπορεί να απορριφθεί, έχουμε ότι αν $\mathbf{Y}_{ij} = (\mathbf{Y}_{obs.ij}, \mathbf{Y}_{mis.ij})^T$ είναι η $j$-οστή παρατήρηση από το $i$-οστό μοτίβο των ελλιπών δεδομένων, με τις συνιστώσες να αντιστοιχούν στο παρατηρούμενο και στο ελλιπές μέρος, αντίστοιχα, τότε

$$\mathbf{Y}_{ij} = \begin{pmatrix} \mathbf{Y}_{obs.ij} \\ \mathbf{Y}_{mis.ij} \end{pmatrix} \sim \mathcal{N}\left(\mu_i, \mathbf{\Sigma}_i\right), \text{ για } i = 1, ..., g, \text{ και } j = 1, ..., n_i,$$

όπου

$$\mu_i = \begin{pmatrix} \mu_{o.i} \\ \mu_{m.i} \end{pmatrix} \text{ και } \mathbf{\Sigma}_i = \begin{pmatrix} \mathbf{\Sigma}_{oo.i} & \mathbf{\Sigma}_{om.i} \\ \mathbf{\Sigma}_{mo.i} & \mathbf{\Sigma}_{mm.i} \end{pmatrix}.$$

Επομένως, από τις ιδιότητες της πολυδιάστατης κανονικής κατανομής έχουμε ότι η δεσμευμένη κατανομή $\mathbf{Y}_{mis.ij}|\mathbf{Y}_{obs.ij}, \mu_i, \mathbf{\Sigma}_i$ είναι $p - p_i$ διάστατη κανονική με παραμέτρους θέσης και κλίμακας

$$\mu_{m.i} + \mathbf{\Sigma}_{mo.i}\mathbf{\Sigma}_{oo.i}^{-1}(\mathbf{Y}_{obs.ij} - \mu_{o.i})$$

και

$$\mathbf{\Sigma}_{mm.i} - \mathbf{\Sigma}_{mo.i}\mathbf{\Sigma}_{oo.i}^{-1}\mathbf{\Sigma}_{om.i},$$

αντίστοιχα. Για τον λόγο αυτό, οι Jamshidian and Jalal (2010) προτείνουν να συμπληρώνονται οι ελλιπείς τιμές από τιμές που επιλέγονται τυχαία από αυτήν τη δεσμευμένη κατανομή. Για να γίνει αυτό είναι απαραίτητο να είναι γνωστές οι τιμές των $\mu_i$ και $\mathbf{\Sigma}_i$. Οι Jamshidian and Jalal (2010) υποθέτουν ότι $\mu_1 = ... = \mu_g = \mu$ και $\mathbf{\Sigma}_1 = ... = \mathbf{\Sigma}_g = \mathbf{\Sigma}$ και προτείνουν να εκτιμηθεί η κοινή μέση τιμή και ο κοινός πίνακας διακυμάνσεων συνδιακυμάνσεων από τον Ε.Μ.Π. Η μέθοδος αυτή υλοποιείται στο πακέτο `MissMech` της R. Από την άλλη, αν τα μέσα διανύσματα δεν μπορούν να θεωρηθούν ίσα, προτείνουν να υπολογίζονται οι Ε.Μ.Π. για το μέσο διάνυσμα κάθε ομάδας. Η επιλογή `ImputationMethod = "Normal"` στη συνάρτηση `TestMCARNormality` του πακέτου `MissMech` χρησιμοποιεί αυτήν την επιλογή. Αντικαθιστώντας λοιπόν τους Ε.Μ.Π. στην παραπάνω

δεσμευμένη κατανομή, αποκτούμε πλήρη γνώση της κατανομής των ελλιπών δεδομένων και επομένως παράγοντας τιμές από την παραπάνω κατανομή, συμπληρώνουμε τις ελλιπείς τιμές. Αφού συμπληρώσουμε τα δεδομένα τότε πρέπει να απαντηθεί το δεύτερο ερώτημα σχετικά με το ποιος έλεγχος ομοσκεδαστικότητας θα χρησιμοποιηθεί.

Ως προς το δεύτερο ερώτημα, οι Jamshidian and Jalal (2010) επέλεξαν να χρησιμοποιήσουν τον έλεγχο που είχε προταθεί, υπό την υπόθεση ότι οι πληθυσμοί περιγράφονται ικανοποιητικά από πολυδιάστατη κανονική κατανόμη, από τον Hawkins (1981). Η επιλογή αυτού του ελέγχου δικαιολογείται από τους συγγραφείς καθώς η απόδοση του είχε αποδειχθεί, μέσω προσομοιώσεων, ότι είναι ικανοποιητική ακόμη και για μικρά σε μέγεθος δείγματα. Σύμφωνα με αυτόν τον έλεγχο αν $X$ είναι ο $n \times p$ πίνακας των πλήρων δεδομένων, με $X_{ij}$ να είναι η $j$-οστή πειραματική μονάδα ($j = 1, \ldots, n_i$) από την $i$-οστή ομάδα ($i = 1, \ldots, g$) τότε η σ.σ. προσδιορίζεται από τη σχέση:

$$F_{ij} = \frac{(n-g-p)n_i V_{ij}}{p\left[(n_i-1)(n-g) - n_i V_{ij}\right]},$$

όπου

$$V_{ij} = (\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)^T \mathbf{S}^{-1} (\mathbf{X}_{ij} - \bar{\mathbf{X}}_i),$$

με

$$\bar{\mathbf{X}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{X}_{ij},$$

$$\mathbf{S}_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)^T$$

και

$$\mathbf{S} = \sum_{i=1}^{g} \left( \frac{n_i - 1}{n - g} \right) \mathbf{S}_i.$$

Σύμφωνα με τον Hawkins (1981), υπό τη μηδενική υπόθεση της ομοσκεδαστικότητας και θεωρώντας ότι ικανοποιείται η υπόθεση της πολυδιάστατης κανονικότητας, η σ.σ. $F_{ij}$ ακολουθεί κατανομή $\mathcal{F}$ με $p$ και $n-p-g$ βαθμούς ελευθερίας. Επομένως, αν $A_{ij}$ είναι οι τυχαίες μεταβλητές που ορίζονται ως η πιθανότητα η τυχαία μεταβλητή $F \sim \mathcal{F}_{p,n-p-g}$ να λαμβάνει τιμή μεγαλύτερη από $F_{ij}$, δηλαδή αν $A_{ij} = Pr[F > F_{ij}]$,τότε αυτές ακολουθούν, υπό τη μηδενική υπόθεση, ομοιόμορφη κατανομή στο διάστημα (0,1). Με αυτό το σκεπτικό, ο Hawkins (1981) ανάγει τον αρχικό έλεγχο ομοιογένειας σε έλεγχο ότι οι τιμές $A_{ij}$ κάθε ομάδας προέρχονται από την ομοιόμορφη κατανομή. Ο έλεγχος αυτός διεξάγεται κάνοντας χρήση του ελέγχου καλής προσαρμογής που έχει

προταθεί από τους Anderson and Darling (1954) και η αρχική υπόθεση της ομοσκεδαστικότητας προτάθηκε από τον Hawkins (1981) να απορρίπτεται όταν μία τουλάχιστον $p$-τιμή, όπως αυτή προκύπτει εφαρμόζοντας κάποια διόρθωση λόγω πολλαπλών ελέγχων, είναι μικρότερη από το καθορισμένο επίπεδο σημαντικότητας. Επισημαίνεται ότι οι Jamshidian and Jalal (2010) προτείνουν στο σημείο αυτό δύο διαφοροποιήσεις σε σχέση με τον έλεγχο του Hawkins (1981). Η πρώτη διαφοροποίηση έγκειται στη χρήση του ελέγχου του Neyman (1937) για τον έλεγχο ότι τα δεδομένα $A_{ij}$ προέρχονται από ομοιόμορφη κατανομή αντί του ελέγχου των Anderson and Darling (1954) που χρησιμοποιήθηκε από τον Hawkins (1981). Η διαφοροποίηση αυτή βασίστηκε σε μία μελέτη προσομοίωσης που διεξήγαγαν και από την οποία προέκυψε ότι η απόδοση αυτού του ελέγχου είναι καλύτερη σε σχέση με την απόδοση άλλων ελέγχων καλής προσαρμογής της ομοιόμορφης κατανομής, μεταξύ των οποίων συμπεριλαμβανόταν ο έλεγχος των Anderson and Darling (1954). Ειδικότερα, στο πλαίσιο των προτάσεων των Jamshidian and Jalal (2010), η σ.σ. του Neyman (1937) που χρησιμοποιείται για τον έλεγχο ότι τα $A_{ij}$ προέρχονται από την ομοιόμορφη κατανομή δίνεται από τη σχέση:

$$N_i = \sum_{l=1}^{4} \left\{ n_i^{-\frac{1}{2}} \sum_{j=1}^{n_i} \pi_l(A_{ij}) \right\}, \text{ για } i = 1, ..., g,$$

με $\pi_l$ να είναι τα κανονικοποιημένα πολυώνυμα Legendre στο (0,1), τα οποία έχουν προσδιοριστεί στην εργασία του David (1939).

Όσον αφορά τη δεύτερη διαφοροποίηση που έχει προταθεί από τους Jamshidian and Jalal (2010), αυτή έγκειται στον συνδυασμό των $g$ το πλήθος $p$-τιμών των ελέγχων καλής προσαρμογής της ομοιόμορφης κατανομής για να προκύψει μία συνολική $p$ τιμή. Ειδικότερα, προτείνουν να χρησιμοποιείται ο συνδυασμός που έχει προταθεί από τον Fisher (1932). Σε αυτό το πλαίσιο, αν $p_1, \ldots, p_g$ είναι οι $g$ το πλήθος $p$ τιμές, τότε, υπό τη μηδενική υπόθεση, ισχύει ότι:

$$P_T = \sum_{i=1}^{g} p_i \sim \chi^2_{2g}.$$

Τα παραπάνω ισχύουν θεωρώντας ότι ικανοποιείται η υπόθεση της πολυδιάστατης κανονικότητας. Αν αυτή η υπόθεση δεν μπορεί να υιοθετηθεί τότε οι Jamshidian and Jalal (2010) αποδεικνύουν ότι η κατανομή των $F_{ij}$ παρότι άγνωστη είναι ίδια για όλα τα $i, j$, όταν $n_1 = \ldots = n_g$ ή ασυμπτωτικά ίδια όταν τα $n_i$ είναι αρκετά μεγάλα. Επομένως, η μηδενική υπόθεση της ομοσκεδαστικότητας απορρίπτεται αν η κατανομή των $F_{ij}$ δεν είναι ίδια στα διαφορετικά γκρουπ. Αυτό ουσιαστικά σημαίνει ότι ο έλεγχος ανάγεται σε έναν μη παραμετρικό $k$-sample έλεγχο. Στο

πλαίσιο αυτό, προτείνουν να χρησιμοποιηθεί ο έλεγχος που προτάθηκε από τους Scholz and Stephens (1987), που είναι γνωστός στην βιβλιογραφία ως "k-sample Anderson-Darling" τεστ. Ο έλεγχος αυτός χρησιμοποιεί τη σ.σ. $T = \frac{1}{n}\sum_{i=1}^{g} T_i$, με

$$T_i = \frac{1}{n_i}\sum_{j=1}^{N-1}\frac{(NM_{ij} - jn_i)^2}{j(N-j)},$$

όπου $N = \sum_{i=1}^{g} n_i$ το μέγεθος του συνολικού δείγματος των $F_{ij}$ και $M_{ij}$ είναι ο αριθμός των παρατηρήσεων του $i$-οστού δείγματος που δεν είναι μεγαλύτερες της $j$-οστής διατεταγμένης στατιστικής συνάρτησης του συνολικού δείγματος των $F_{ij}$. Απόρριψη της ισότητας των κατανομών των $F_{ij}$ στις $g$ ομάδες, συνεπάγεται απόρριψη της ομοιογένειας των πινάκων διακυμάνσεων συνδιακυμάνσεων, άρα απόρριψη της υπόθεσης MCAR.

Οι Jamshidian and Jalal (2010) διεξήγαγαν μία μελέτη προσομοίωσης που αφορά τόσο τον παραμετρικό έλεγχο όσο και τον μη παραμετρικό έλεγχο για να αξιολογήσουν την απόδοσή τους ως προς τη διατήρηση του επιπέδου σημαντικότητας και την ισχύ στη βάση 1000 προσομοιωμένων $p$-διάστατων ($p = 4, 7, 10$) συνόλων δεδομένων μεγέθους $n = 200, 500, 1000$ από πέντε διαφορετικές πολυδιάστατες κατανομές. Όσον αφορά το ποσοστό των ελλιπών τιμών θεώρησαν τρεις περιπτώσεις ($q = 10\%, 20\%, 30\%$). Υπό τη μηδενική υπόθεση, η θέση των ελλιπών τιμών καθορίζεται με βάση τις τιμές ενός $n \times p$ πίνακα με στοιχεία $u_{ij},\ i = 1, ..., n,\ j = 1, ..., p$ που παράγονται από την ομοιόμορφη κατανομή. Ειδικότερα, αν $u_{ij} < q$, τότε σε αυτήν τη θέση του πίνακα των δεδομένων εκχωρείται ελλιπής τιμή. Από την άλλη, για την αξιολόγηση της απόδοσης του ελέγχου ως προς την ισχύ τα δεδομένα παράγονται υπό την υπόθεση MAR. Ειδικότερα, τα ελλιπή δεδομένα δημιουργούνται διαγράφοντας την $i-$οστή πειραματική μονάδα της $j-$οστής υποομάδας, αν η τιμή της $i$-οστης πειραματικής μονάδας της $j-1$ υποομάδας είναι μικρότερη από κάποιο κατώφλι ($i = 1, .., n,\ j = 2, .., p$), το οποίο παράγεται με στόχο να αποκτήσουμε το επιθυμητό ποσοστό ελλιπών τιμών $q$. Ο περιορισμός σε δεδομένα από την κανονική κατανομή εξηγείται καθώς έλαβαν υπόψη τα μη ικανοποιητικά αποτελέσματα ως προς το επίπεδο σημαντικότητας για τις περιπτώσεις των άλλων κατανομών. Από τα αποτελέσματα της μελέτης τους προκύπτουν τα ακόλουθα συμπεράσματα:

- ο παραμετρικός έλεγχος των Jamshidian and Jalal (2010) φαίνεται να διατηρεί το επίπεδο σημαντικότητας αρκετά κοντά στο 5%, για δεδομένα υπό την κανονική κατανομή. Παρατηρείται μόνο μία μεγάλη παρέκκλιση για την περίπτωση όπου $n = 1000,\ p = 7$ και $q = 0.3$, η οποία οφείλεται στην ύπαρξη πολλών ομάδων με μικρό αριθμό πειραματικών μονάδων. Ωστόσο,

ο παραμετρικός έλεγχος φαίνεται να μην διατηρεί το ονομαστικό επίπεδο σημαντικότητας για δεδομένα που αποκλίνουν από την πολυδιάστατη κανονική κατανομή. Για τον λόγο αυτόν η ισχύς του δεν αξιολογείται στις περιπτώσεις των άλλων κατανομών.

- Ο μη παραμετρικός έλεγχος που προτείνουν οι Jamshidian and Jalal (2010) παρουσιάζει αρκετά καλά αποτελέσματα ως προς τη διατήρηση του ονομαστικού επιπέδου σημαντικότητας για όλες τις κατανομές, καθώς σε γενικές γραμμές φαίνεται να διατηρεί το επίπεδο σημαντικότητας.

- Για μέγεθος δείγματος $n = 200$, παρατηρούνται μικρές τιμές για την ισχύ του παραμετρικού ελέγχου. Όπως είναι αναμενόμενο, όταν το μέγεθος του δείγματος αυξάνεται, αυξάνεται και η ισχύς του ελέγχου. Ο παραμετρικός έλεγχος επιδεικνύει μέτρια-κακή ισχύ για $n = 500$, ιδιαίτερα σε περιπτώσεις όπου ο αριθμός των μεταβλητών είναι μεγάλος, ενώ για μεγάλο μέγεθος δείγματος $n = 1000$ και για μικρό αριθμό μεταβλητών, φαίνεται να είναι αξιόπιστος. Επιπλέον παρατηρείται πτώση της ισχύος, όταν αυξάνεται το ποσοστό των ελλιπών δεδομένων.

- Η ισχύς του μη παραμετρικού έλεγχου είναι σε γενικές γραμμές καλή για μέγεθος δείγματος $n = 500$ και αρκετά καλή για μέγεθος δείγματος $n = 1000$. Ειδικότερα, ο έλεγχος επιδεικνύει μεγάλη ισχύ για κατανομές με "ελαφριές ουρές" (light-tailed), ενώ το επίπεδο της ισχύος είναι αρκετά χαμηλό για κατανομές όπως η πολυδιάστατη Weibull. Επίσης, παρατηρείται, ξανά, αύξηση της ισχύος καθώς αυξάνεται το μέγεθος του δείγματος.

Όλα τα παραπάνω μας οδηγούν στο συμπέρασμα ότι ο μη παραμετρικός έλεγχος μπορεί να θεωρηθεί περισσότερο αξιόπιστος.

**Εφαρμογή στα δεδομένα PimaIndiansDiabetes2** Η υλοποίηση των παραπάνω ελέγχων γίνεται με χρήση της συνάρτησης `TestMCARNormality()`, που βρίσκεται στο πακέτο `MissMech` της γλώσσας R και προϋποθέτει ότι το σύνολο των δεδομένων, το οποίο δηλώνεται στο όρισμα data, περιέχει τουλάχιστον δύο στήλες. Επιπλέον, τα δεδομένα που τοποθετούνται στο όρισμα data ανήκουν στην κλάση matrix() ή data.frame() και οι ελλιπείς τιμές έχουν οριστεί στο σύνολο αυτό ως NA. Στο όρισμα method αν ο χρήστης είναι βέβαιος ότι τα δεδομένα έχουν πολυμεταβλητή κανονική κατανομή, θα πρέπει να επιλεγεί η μέθοδος "Hawkins". Από την άλλη, αν τα δεδομένα δεν είναι κανονικά κατανεμημένα, τότε θα πρέπει να χρησιμοποιηθεί η μέθοδος "Nonparametric". Αν ο χρήστης δεν είναι σίγουρος, η προεπιλεγμένη τιμή είναι method="Auto", οπότε

και θα εκτελεστούν και οι δύο δοκιμές, τόσο η Hawkins όσο και η μη παραμετρική. Στο όρισμα imputation.method γράφουμε, όταν τα δεδομένα δεν είναι πολυδιάστατα κανονικά, "Dist.Free" για να αντικατασταθούν τα ελλιπή δεδομένα με τιμές που προκύπτουμε μέσω μιας μη παραμετρικής μεθόδου, ενώ διαφορετικά γράφουμε "Normal". Τέλος, στο όρισμα alpha θέτουμε το επίπεδο σημαντικότητας. Για το σύνολο δεδομένων PimaIndiansDiabetes2 προκύπτει ότι η υπόθεση MCAR απορρίπτεται με τον παραμετρικό έλεγχο, ενώ δεν απορρίπτεται με τον μη παραμετρικό, καθώς οι $p$ τιμές είναι ίσες με $1.810503e-11$ και 0.2108583, αντίστοιχα.

## 2.5 Ο έλεγχος των Li and Yu (2015)

Όπως αναλυτικά είδαμε στις προηγούμενες ενότητες οι έλεγχοι ομοιογένειας που προτάθηκαν από τον Little (1988) και τους Kim and Bentler (2002) βασίζονται στην υπόθεση της πολυδιάστατης κανονικότητας, η οποία συχνά στην πράξη δεν μπορεί να υιοθετηθεί. Η μη υιοθέτηση αυτής της υπόθεσης έχει ως συνέπεια πολλές φορές να προτιμώνται μεθοδολογίες που δεν στηρίζονται σε κάποια υπόθεση για την κατανομή των δεδομένων, δηλαδή να προτιμώνται μη παραμετρικές μεθοδολογίες. Στην ενότητα που προηγήθηκε, παρουσιάστηκε μία τέτοια μεθοδολογία που προτάθηκε από τους Jamshidian and Jalal (2010). Ωστόσο, ο έλεγχος αυτός επικεντρώνεται στον έλεγχο της ομοιογένειας των πινάκων διακυμάνσεων συνδιακυμάνσεων, ενώ, στην πράξη, αν τα δεδομένα δεν είναι πλήρως τυχαία ελλιπή (MCAR), τότε μπορούν να διαφοροποιούνται στατιστικά σημαντικά ως προς οποιοδήποτε άλλο χαρακτηριστικό. Για παράδειγμα, θα μπορούσαν να διαφοροποιούνται ως προς τη λοξότητα ή την κύρτωση κ.ο.κ. Παρακινούμενοι από αυτήν την παρατήρηση οι Li and Yu (2015) ελέγχουν την MCAR υπόθεση μέσω του ελέγχου της ομοιογένειας των κατανομών των διαφορετικών μοτίβων ελλιπών δεδομένων. Σε αυτό το πλαίσιο, προτείνουν έναν μη παραμετρικό έλεγχο, υπό την έννοια ότι δεν απαιτείται κάποια υπόθεση για τη μορφή της κατανομής. Ωστόσο, ο έλεγχός τους υποθέτει ότι υπάρχουν κάποιες πειραματικές μονάδες για τις οποίες έχουν παρατηρηθεί τιμές σε όλες τις $p$ το πλήθος μεταβλητές. Χωρίς βλάβη της γενικότητας αυτές οι πειραματικές μονάδες αποτελούν το πρώτο μοτίβο των ελλιπών δεδομένων και ο πίνακας των δεδομένων τους είναι ο $\mathbb{Y}_1$.

Πρακτικά, ο έλεγχος ισότητας των κατανομών δεν μπορεί να υλοποιηθεί άμεσα. Η αδυναμία αυτή, οφείλεται στο ότι η από κοινού κατανομή της κάθε ομάδας απαρτίζεται και από ένα ελλιπές μέρος, για το οποίο δεν μπορούμε να πραγματοποιήσουμε συμπερασματολογία. Επομένως, η υπόθεση των ίσων κατανομών

δεν μπορεί να ελεγχθεί άμεσα. Οι Li and Yu (2015) ξεπερνούν αυτό το εμπόδιο, εξετάζοντας την ισότητα των κατανομών της κάθε ομάδας, μόνο για τις από κοινού παρατηρούμενες μεταβλητές. Εξετάζουν δηλαδή, τις περιθώριες κατανομές των ομάδων ανά δύο, οι οποίες περιθώριες κατανομές αποτελούνται από τις από κοινού παρατηρούμενες μεταβλητές των ομάδων. Σε όσα ακολουθούν με $o_i$, $i = 1, \ldots, s$, συμβολίζεται το υποσύνολο του συνόλου $\{1, 2, \ldots, p\}$ που υποδεικνύει ποιες μεταβλητές παρατηρούνται στην $i$-οστή ομάδα, ενώ με $o_{ij}$ συμβολίζεται το σύνολο των μεταβλητών που παρατηρούνται στην $i$-οστή και $j$-οστή ομάδα, δηλαδή $o_{ij} = o_i \cap o_j$. Στο πλαίσιο αυτό, ελέγχουν τη μηδενική υπόθεση

$$H_0 : F_{i,o_{ij}} = F_{j,o_{ij}} \ , \forall i \neq j \in \{1, ...s\} \ \kappa\alpha\iota \ o_{ij} \neq \emptyset, \tag{2.12}$$

έναντι της

$$H_a : \exists \ i \neq j \in \{1, ...s\} \ : F_{i,o_{ij}} \neq F_{j,o_{ij}} \ \kappa\alpha\iota \ o_{ij} \neq \emptyset,$$

όπου με $F_{i,o_{ij}}$ συμβολίζουμε την από κοινού περιθώρια κατανομή για την $i$-οστή ομάδα που απαρτίζεται από τις από κοινού παρατηρούμενες μεταβλητές για τις ομάδες $i$ και $j$.

Από όσα προηγήθηκαν είναι κατανοητό ότι θα πρέπει να ελεγχθεί η ισότητα κάθε ζεύγους $(F_{i,o_{ij}}, F_{j,o_{ij}})$. Για να μπορεί να γίνει αυτό θα πρέπει να χρησιμοποιηθεί ένα κατάλληλο μέτρο ανομοιότητας μεταξύ του πίνακα των δεδομένων της $i$-οστής ομάδας που περιέχει τις μεταβλητές που παρατηρήθηκαν ταυτόχρονα στις δύο συγκρινόμενες ομάδες, έστω $\mathbb{Y}_{i,o_{ij}}$, και του πίνακα των δεδομένων της $j$-οστής ομάδας που περιέχει τις μεταβλητές που παρατηρήθηκαν ταυτόχρονα στις δύο συγκρινόμενες ομάδες, έστω $\mathbb{Y}_{j,o_{ij}}$. Για τον σκοπό αυτό, οι Li and Yu (2015) πρότειναν να χρησιμοποιηθεί το μέτρο που έχει προταθεί στη βιβλιογραφία από τους Rizzo and Székely (2010), ο ορισμός του οποίου ακολουθεί.

**Ορισμός 2.5.1** *Ας είναι* $\mathbb{X} = (\mathbf{x}_1^T, ..., \mathbf{x}_{n_1}^T)^T$ *και* $\mathbb{Z} = (\mathbf{z}_1^T, ..., \mathbf{z}_{n_2}^T)^T$, *με* $\mathbf{x}_i, \mathbf{z}_j \in \mathbb{R}^p$ *και* $\{\mathbf{x}_1, ..., \mathbf{x}_{n_1}\}$, $\{\mathbf{z}_1, ..., \mathbf{z}_{n_2}\}$ *να είναι δύο τυχαία δείγματα. Το δειγματικό μέτρο ανομοιότητας ορίζεται από τη σχέση:*

$$d(\mathbb{X}, \mathbb{Z}) = 2g(\mathbb{X}, \mathbb{Z}) - g(\mathbb{X}, \mathbb{X}) - g(\mathbb{Z}, \mathbb{Z}),$$

*όπου*

$$g(\mathbb{X}, \mathbb{Z}) = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \|\mathbf{x}_i - \mathbf{z}_j\|,$$

$$g(\mathbb{X}, \mathbb{X}) = \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} \|\mathbf{x}_i - \mathbf{x}_j\|,$$

*και*

$$g(\mathbb{Z},\mathbb{Z}) = \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}\|\mathbf{z}_i - \mathbf{z}_j\|,$$

*με* $\|\cdot\|$ *να είναι η Ευκλείδεια νόρμα.*

Χρησιμοποιώντας το παραπάνω μέτρο ανομοιότητας οι Li and Yu (2015) ορίζουν την ακόλουθη στατιστική συνάρτηση:

$$Q = \frac{B/(s-1)}{W/(n-s)}, \tag{2.13}$$

όπου

$$B = \sum_{1\le i<j\le s\ o_{ij}\neq\emptyset}\left(\frac{n_i n_j}{2n}\right) d(\mathbb{Y}_{i,o_{ij}}, \mathbb{Y}_{j,o_{ij}})$$

και

$$W = \sum_{i=1}^{s}\left(\frac{n_i}{2}\right) g(\mathbb{Y}_{i,o_i}, \mathbb{Y}_{i,o_i}).$$

Οι σ.σ. $B$ και $W$ μετρούν την ανομοιότητα μεταξύ των δειγμάτων και την ανομοιότητα εντός του δείγματος, υιοθετώντας, κατά κάποιο τρόπο, τη λογική των μοντέλων ανάλυσης διακύμανσης. Περιμένουμε λοιπόν μεγάλες τιμές της συνάρτησης $Q$, να συνεπάγονται μεγαλύτερη ανομοιότητα μεταξύ των δειγμάτων σε σχέση με αυτήν εντός των δειγμάτων, αποτελώντας ένδειξη ανομοιότητας των κατανομών των δειγμάτων άρα και διαφορετικότητα των κατανομών. Επομένως, απορρίπτεται, σε επίπεδο σημαντικότητας $\alpha$, η μηδενική υπόθεση που διατυπώθηκε στη σχέση (2.13) και επομένως και η υπόθεση ότι τα δεδομένα είναι MCAR όταν $Q > c_\alpha$, με $c_\alpha$ να συμβολίζει το άνω $\alpha$ ποσοστιαίο σημείο της κατανομής της $Q$. Καθώς η κατανομή της σ.σ. $Q$ δεν έχει προσδιοριστεί, το κρίσιμο σημείο $c_\alpha$ προσδιορίζεται μέσω της μεθόδου bootstrap (Efron and Tibshirani (1994), Davison and Hinkley (1997)). Στη συνέχεια περιγράφεται συνοπτικά αυτή η διαδικασία. Έστω $\mathbb{Y}_1$ ο $n_1 \times p$ πίνακας δεδομένων που αποτελείται από τα διανύσματα που ανήκουν στην ομάδα 1 (ανακαλέστε την υπόθεση στην αρχή αυτής της ενότητας). Επιλέγουμε, με επανάθεση, $n$ το πλήθος παρατηρήσεις από αυτόν τον πίνακα και τις τοποθετούμε σε έναν νέο, $n\times p$, πίνακα, τον οποίο συμβολίζουμε με $\mathbb{Y}^*_{complete}$. Καθώς, η κατανομή του $Q$ εξαρτάται από τα μοτίβα των ελλιπών τιμών, τα bootstrap δείγματα του $\mathbb{Y}$ πρέπει να έχουν το ίδιο μοτίβο ελλιπών τιμών. Άρα, αν πάρουμε τις πρώτες $n_1$ γραμμές του πίνακα $\mathbb{Y}^*_{complete}$, έχουμε ένα νέο δείγμα από την $\mathbb{Y}_1$. Αντίστοιχα παίρνοντας τις επόμενες $n_2$ γραμμές από τον πίνακα $\mathbb{Y}^*_{complete}$ και διαγράφοντας τις παρατηρήσεις που αντιστοιχούν σε μεταβλητές που δεν παρατηρήθηκαν στην ομάδα 2, έχουμε ένα bootstrap δείγμα για

τον $\mathbb{Y}_2$ (αυτό οφείλεται στην υπόθεση ισότητας των κατανομών). Συνεχίζεται η διαδικασία μέχρι να εξαντλήσουμε όλες τις ομάδες και να πάρουμε τον πίνακα $\mathbb{Y}^*$. Η διαδικασία επαναλαμβάνεται $B$ φορές και υπολογίζουμε κάθε φορά την τιμή της σ.σ., έστω $Q^{(1)}, \ldots, Q^{(B)}$, για αρκετά μεγάλη τιμή του $B$. Είναι τότε $\hat{c}_\alpha$ το άνω $\alpha$ ποσοστιαίο σημείο των $Q^{(1)}, \ldots, Q^{(B)}$.

**Παρατήρηση 2.5.1** *Αποδεικνύεται ότι ο έλεγχος των Li and Yu (2015) είναι συνεπής για κάθε εναλλακτική υπόθεση με πεπερασμένες δεύτερης τάξης ροπές. Επιπλέον, μέσω προσομοιώσεων, οι Li and Yu (2015) συμπέραναν ότι ο έλεγχος αυτός μπορεί να συμπεριλάβει περιπτώσεις ομάδων με μοτίβα ελλιπών τιμών που περιέχουν μικρό πλήθος παρατηρήσεων, όμως χρειάζεται ικανοποιητικό μέγεθος δείγματος για την υλοποίηση της διαδικασίας bootstrap.*

**Εφαρμογή στα δεδομένα PimaIndiansDiabetes2** Ο έλεγχος των Li and Yu (2015) δεν υλοποιείται σε κάποιο πακέτο της γλώσσας R, σύμφωνα με όσα γνωρίζουμε. Ο παραπάνω έλεγχος υλοποιείται μέσω της συνάρτησης `Li_Yu_test()` που κατασκευάστηκε και παρατίθεται στο παράρτημα αυτής της μεταπτυχιακής διατριβής. Στο όρισμα αυτής "Y" δίνεται το σύνολο δεδομένων που ανήκει στην κλάση data.frame(), στο όρισμα "n_boot" δηλώνεται το πλήθος B των bootstrap δειγμάτων που θα χρησιμοποιηθούν, ενώ στο όρισμα "alpha" το επίπεδο σημαντικότητας. Η συνάρτηση τυπώνει το κρίσιμο σημείο "Critical Value" που προκύπτει έπειτα από εφαρμογή της διαδικασίας bootstrap, την τιμή που λαμβάνει η στατιστική συνάρτηση "Q_statistic" και το αν απορρίπτεται ή όχι η υπόθεση MCAR. Επισημαίνεται ότι η ορθή λειτουργία του ελέγχου προϋποθέτει την εγκατάσταση των βιβλιοθηκών "dplyr" και "proxy". Για το σύνολο δεδομένων PimaIndiansDiabetes2 προκύπτει ότι η τιμή της στατιστικής συνάρτησης είναι ίση με 0.9293818, με το κρίσιμο σημείο να υπολογίζεται ότι είναι ίσο 0.8262347. Επομένως, συμπεραίνουμε ότι απορρίπτεται η MCAR υπόθεση.

## 2.6 Ο έλεγχος των Chen et al. (2023)

Είναι πολύ σύνηθες τα διαθέσιμα προς ανάλυση δεδομένα να είναι ελλιπή και να περιλαμβάνουν τόσο κατηγορικά όσο και ποσοτικά δεδομένα, ήτοι να είναι μικτού τύπου. Για τον λόγο αυτό είναι αναγκαία η ύπαρξη μεθόδων ελέγχου της υπόθεσης MCAR σε μικτού τύπου δεδομένα. Για την αντιμετώπιση αυτού του ερευνητικού ερωτήματος, πρόσφατα, οι Chen et al. (2023) πρότειναν έναν τρόπο ελέγχου μίας υπό περίπτωσης της υπόθεσης MCAR. Πιο συγκεκριμένα, η προτεινόμενη μέθοδός τους διαπραγματεύεται την υπόθεση των Πραγματοποιημένων

Πλήρως Τυχαία Ελλιπών Δεδομένων (Realized Missing Completetly at Random, RMCAR), έννοια η οποία πρωτοπαρουσιάστηκε στην εργασία των Mealli and Rubin (2015). Πιο συγκεκριμένα, οι Mealli and Rubin (2015) χωρίζουν την υπόθεση MCAR σε δύο κατηγορίες: την RMCAR και την MACAR (Πάντα Πλήρως Τυχαία Ελλιπή Δεδομένα, Missing Always Completely at Random), με τον ορισμό των MACAR να ταυτίζεται με τον ορισμό που έχει δοθεί ως τώρα ως MCAR. Πρακτικά η κατηγορία RMCAR, είναι υπό-κατηγορία της υπόθεσης MCAR, αφού ορίζεται εντελώς ανάλογα με την υπόθεση MCAR, αλλά ισχύει μόνο για τα παρατηρηθέντα μοτίβα ελλιπών τιμών, όπως αυτά ορίστηκαν στην Ενότητα 2.1. Πιο συγκεκριμένα, έχουμε ότι υπό την MCAR ισχύει ότι:

$$Pr(\mathbf{y}_{obs.i}|\mathbf{r}_i) = Pr(\mathbf{y}_{obs.i}), \text{ για κάθε παρατηρηθέν ή μη } \mathbf{r}_i,$$

ενώ για την RMCAR ισχύει ότι:

$$Pr(\mathbf{y}_{obs.i}|\mathbf{r}_i) = Pr(\mathbf{y}_{obs.i}), \text{ για κάθε παρατηρηθέν } \mathbf{r}_i.$$

Από τα παραπάνω γίνεται αντιληπτό το γεγονός ότι η υπόθεση αυτή είναι λιγότερο περιοριστική σε σχέση με την υπόθεση MCAR, αφού αρκεί να ισχύει για τις δείκτριες συναρτήσεις που παρατηρήθηκαν. Ακόμη, η εφαρμογή του ελέγχου σε μικτού τύπου δεδομένα, καθιστά τον έλεγχο αρκετά ευέλικτο και χρήσιμο.

Πρακτικά, η υλοποίηση του προτεινόμενου ελέγχου δεν είναι τίποτε άλλο παρά ένας συνδυασμός ελέγχων για συνεχή και διακριτά δεδομένα. Πιο συγκεκριμένα, παρόμοια με τον έλεγχο του Little (1988), χωρίζουμε τα δεδομένα σε ομάδες ανάλογα με το μοτίβο των ελλιπών δεδομένων. Έπειτα, τις παρατηρούμενες συνιστώσες των διανυσμάτων που ανήκουν σε κάθε μοτίβο, τις χωρίζουμε σε 3 κατηγορίες: στις συνεχείς, στις κατηγορικές και στις διακριτές. Παίρνουμε τις συνιστώσες που προέρχονται από πειραματικές μονάδες, που ανήκουν στην ίδια ομάδα και στην ίδια κατηγορία. Προς στιγμήν, για διευκόλυνση στην επεξήγηση της μεθόδου ας πούμε, ότι αναφερόμαστε στην περίπτωση, όπου μας ενδιαφέρει μία συνεχής μεταβλητή, έστω η $X_1$. Φτιάχνουμε μία ομάδα, η οποία αποτελείται από το σύνολο των πειραματικών μονάδων για τις οποίες έχει παρατηρηθεί πλήρως η μεταβλητή $X_1$. Στη συνέχεια, η ομάδα αυτή προφανώς θα αποτελείται από πειραματικές μονάδες με διαφορετικά μοτίβα ελλιπών τιμών, αλλά σε όλα θα έχει παρατηρηθεί η μεταβλητή $X_1$. Χωρίζουμε αυτήν την ομάδα σε υποομάδες. Κάθε υποομάδα θα έχει ως στοιχεία της τις πειραματικές μονάδες που έχουν το ίδιο μοτίβο ελλιπών τιμών. Για κάθε μία από αυτές κρατάμε μόνο τις παρατηρήσεις για την μεταβλητή $X_1$, αγνοώντας τις παρατηρήσεις για τις υπόλοιπες μεταβλητές. Αν ισχύει η υπόθεση MCAR, τότε κάθε υποομάδα θα αποτελεί ένα υπο-δείγμα για τη μεταβλητή $X_1$.

Προχωρώντας λοιπόν σε έλεγχο της ομοιογένειας μεταξύ αυτών των ομάδων είμαστε σε θέση να ελέγξουμε την υπόθεση MCAR για τη μεταβλητή $X_1$. Επεκτείνοντας την ίδια επιχειρηματολογία για όλες τις μεταβλητές, που ανήκουν σε κάθε κατηγορία, χρησιμοποιώντας μία κατάλληλη στατιστική συνάρτηση ελέγχου, ελέγχουμε τη συνολική ομοιογένεια των μικτού τύπου δεδομένων.

Στο σημείο αυτό επισημαίνεται ότι για τις συνεχείς και διακριτές μεταβλητές ο έλεγχος ομοιογένειας γίνεται μέσω του k-sample Anderson-Darling στατιστικού των Scholz and Stephens (1987), ενώ για τις κατηγορικές μεταβλητές προτείνεται ο έλεγχος μέσω του $\mathcal{X}^2$ στατιστικού του Pearson.

Επισημαίνεται ότι καθώς χρησιμοποιούνται πολλαπλοί έλεγχοι στα ίδια δεδομένα υπάρχει κίνδυνος να απορριφθούν εσφαλμένα αληθείς υποθέσεις. Για την αντιμετώπιση αυτού του προβλήματος οι Chen et al. (2023) προτείνουν την εφαρμογή μεθόδων pFDR (posterior False Discovery Rate) και FDR (False Discovery Rate, q-value). Για παράδειγμα, μπορεί να χρησιμοποιηθεί η μέθοδος των Benjamini and Hochberg (1995).

**Εφαρμογή στα δεδομένα PimaIndiansDiabetes2** Όσο είμαστε σε θέση να γνωρίζουμε δεν υπάρχει διαθέσιμος ο έλεγχος των Chen et al. (2023) σε κάποιο πακέτο της R. Για τον λόγο αυτό στο παράρτημα αυτής της μεταπτυχιακής διατριβής παρατίθεται ο κώδικας της συνάρτησης `testChengChungBasu()` για την υλοποίησή του. Επισημαίνει ότι καθώς η συνάρτηση αξιοποιεί ήδη έτοιμες ρουτίνες από τις βιβλιοθήκες "kSamples" και "stats" απαραίτητη προϋπόθεση είναι η προεγκατάσταση των προαναφερθέντων βιβλιοθηκών. Η συνάρτηση απαιτεί ως όρισμα ένα σύνολο δεδομένων που ανήκει στην κλάση data.frame(), διαχωρίζει τις μεταβλητές σε κατηγορίες και έπειτα τις πειραματικές μονάδες ανάλογα με το μοτίβο ελλιπών τιμών. Για τις περιπτώσεις των συνεχών και διακριτών μεταβλητών εφαρμόζει τον έλεγχο του $k$-sample Anderson-Darling στατιστικού των Scholz and Stephens (1987), ενώ για την περίπτωση κατηγορικών μεταβλητών κατασκευάζει έναν πίνακα συνάφειας με διαστάσεις το πλήθος των διακεκριμένων μοτίβων των ελλιπών τιμών και το πλήθος των επιπέδων των κατηγοριών των μεταβλητών. Στη συνέχεια, εφαρμόζει το $\mathcal{X}^2$ στατιστικό τεστ του Pearson.

Επισημαίνεται, ότι στην περίπτωση που είτε δεν υπάρχουν πάνω από 2 ομάδες ελλιπών δεδομένων ή περισσότερες από δύο πειραματικές μονάδες εντός των ομάδων, ο $k$-sample έλεγχος αγνοεί αυτή τη μεταβλητή και, εκ κατασκευής, συνεχίζει με τις υπόλοιπες (βλέπε ad.test()).

Τελικά, η συνάρτηση testChengChungBasu() τυπώνει τις $p$-τιμές για τις με-

ταβλητές που δεν αγνοήθηκαν και στις οποίες έχει εφαρμοστεί η μέθοδος των Benjamini and Hochberg (1995). Είναι φανερό ότι απόρριψη της MCAR υπόθεσης για μία εξ αυτών των μεταβλητών, οδηγεί στην απόρριψη της MCAR υπόθεσης για όλο το σύνολο των δεδομένων.

Για το σύνολο δεδομένων PimaIndiansDiabetes2 το διάνυσμα των $p$-τιμών που τυπώνεται είναι το

$$(0.0199785, 0.00178182, 0.55565, 0.1133016, 1.34097e-05, 3.73842e-11),$$

γεγονός που υποδηλώνει ότι τα δεδομένα δεν είναι MCAR.

## 2.7 Ο έλεγχος των Jamshidian and Yuan (2014)

Στην ενότητα αυτή το ενδιαφέρον επικεντρώνεται στην παρουσίαση του ελέγχου των Jamshidian and Yuan (2014), ο οποίος βασίζεται στον διαχωρισμό ενός συνόλου δεδομένων σε ομάδες και στον μετέπειτα έλεγχο της ομοιογένειας των περιθωρίων κατανομών, ακολουθώντας κατά μία έννοια τον συλλογισμό που αναπτύχθηκε από τους Chen et al. (2023). Συγκεκριμένα, ο έλεγχος της υπόθεσης MCAR έναντι της MAR ή της NMAR υπόθεσης σε ένα $n \times p$ σύνολο δεδομένων με ελλιπείς τιμές σε οποιαδήποτε μεταβλητή, ανάγεται στον έλεγχο ισότητας των περιθωρίων κατανομών μεταξύ δύο ομάδων, για κάθε μεταβλητή στην οποία παρατηρούνται ελλιπείς τιμές.

Προς το παρόν, για διευκόλυνση στην επεξήγηση της μεθόδου ας περιοριστούμε στην περίπτωση, όπου μας ενδιαφέρει μία μεταβλητή, έστω η $Y_1$. Φτιάχνουμε μία ομάδα, η οποία αποτελείται από το σύνολο των πειραματικών μονάδων για τις οποίες έχει παρατηρηθεί πλήρως η μεταβλητή $Y_1$ και αντίστοιχα μία στις οποίες δεν έχει παρατηρηθεί. Οι ομάδες αυτές προφανώς θα αποτελούνται από πειραματικές μονάδες με διαφορετικά μοτίβα ελλιπών τιμών, αλλά σε όλα θα έχει παρατηρηθεί η μεταβλητή $Y_1$, για την πρώτη ομάδα και αντίστοιχα δεν θα έχει παρατηρηθεί για τη δεύτερη ομάδα. Αγνοούμε τη μεταβλητή με βάση την οποία πραγματοποιήθηκε ο διαχωρισμός και εξετάζουμε τις υπόλοιπες. Για κάθε μία από τις $p-1$ το πλήθος πλέον μεταβλητές στις οποίες παρατηρούνται ελλιπείς τιμές, εφαρμόζουμε τον k-sample Anderson-Darling έλεγχο ομοιογένειας των Scholz and Stephens (1987). Ο έλεγχος εφαρμόζεται μόνο για τις διαθέσιμες παρατηρήσεις για την κάθε μεταβλητή αγνοώντας τις ελλιπείς τιμές. Άρα συνολικά διενεργούνται $r \times (p-1)$ το πλήθος έλεγχοι, όπου $r$ το πλήθος των μεταβλητών που παρουσιάζουν ελλιπείς τιμές.

Ο παραπάνω έλεγχος, παρότι απλός και εύκολα εφαρμόσιμος, έχει το σοβαρό

μειονέκτημα ότι οδηγεί σε αυξημένες πιθανότητες σφαλμάτων τύπου I, που οφείλονται στο γεγονός ότι οι έλεγχοι που διενεργούνται συσχετίζονται. Για την αντιμετώπιση αυτού του προβλήματος οι Jamshidian and Yuan (2014), προτείνουν την εφαρμογή της μεθόδου των Benjamini and Hochberg (1995).

**Παρατήρηση 2.7.1** *Όταν η εναλλακτική υπόθεση είναι η MAR, ο έλεγχος μπορεί να ανιχνεύσει την ή τις σχέσεις ανάμεσα στις μεταβλητές που δημιουργούν τις ελλιπείς τιμές. Υποθέτουμε ξανά ότι μας ενδιαφέρει η μεταβλητή $Y_1$ και ακολουθούμε την ίδια διαδικασία που περιγράφηκε παραπάνω για τον έλεγχο της ομοιογένειας. Αν παρατηρήσουμε διαφοροποίηση ως προς τις περιθώριες κατανομές των δύο ομάδων για μία από τις υπόλοιπες μεταβλητές που περιέχουν ελλιπείς τιμές, τότε η διαφορά αυτή αποτελεί ένδειξη για τη σχέση ανάμεσα στις ελλιπείς τιμές της μεταβλητής $Y_1$ και στις παρατηρούμενες της υπό μελέτη μεταβλητής, στην οποία εφαρμόζεται ο έλεγχος των k-sample Anderson-Darling.*

**Εφαρμογή στα δεδομένα PimaIndiansDiabetes2** Ο έλεγχος που περιγράφηκε σε αυτήν την ενότητα δεν υλοποιείται σε κάποιο διαθέσιμο πακέτο της γλώσσας R, ωστόσο οι Jamshidian and Yuan (2014) παραθέτουν συνάρτηση υλοποίησης του ελέγχου στην εργασία τους με το όνομα `NPV test`. Ειδικότερα, ο κώδικάς τους, ο οποίος για λόγους πληρότητας παρατίθεται στο παράρτημα αυτής της μεταπτυχιακής διατριβής, απαρτίζεται από 2 συναρτήσεις. Η πρώτη συνάρτηση `ADtest()`, απαιτεί ως όρισμα ένα σύνολο δεδομένων το οποίο περιέχει ελλιπείς τιμές και επιπρόσθετα ανήκει στην κλάση `matrix()` της γλώσσας R. Η συνάρτηση αυτή παράγει ως έξοδο έναν πίνακα τριών στηλών. Οι πρώτες δύο στήλες αποτελούνται από τους δείκτες των μεταβλητών που συγκρίνονται, ενώ η τρίτη στήλη περιέχει τις $p$-τιμές που παράγονται από το "K-samples" Anderson-Darling έλεγχο. Η δεύτερη συνάρτηση `BenHoch()` εφαρμόζει τη μέθοδο Benjamini-Hochberg, κατά πλήρη αναλογία με όσα προαναφέρθηκαν στην ενότητα και απαιτεί ως όρισμα το διάνυσμα-στήλη που περιέχει τις $p$-τιμές που είναι διαθέσιμες από την εφαρμογή της συνάρτησης `ADtest()`. Επισημαίνεται ότι το διάνυσμα αυτό πρέπει να ανήκει στην κλάση `vector()`. Για το σύνολο δεδομένων PimaIndiansDiabetes2 τυπώνονται 26 το πλήθος $p$-τιμές. Σύμφωνα με όσα προηγήθηκαν θα έπρεπε να τυπωθούν $r(p-1) = 30$ το πλήθος $p$-τιμές. Ωστόσο, τέσσερις από αυτές δεν τυπώνονται καθώς είχαμε ομάδες με λιγότερες από 2 πειραματικές μονάδες. Από τα αποτελέσματα που είναι διαθέσιμα, αλλά δεν παρατίθενται για λόγους χώρου, συμπεραίνουμε ότι απορρίπτεται η MCAR υπόθεση.

## 2.8 Ο έλεγχος των Jamshidian and Yuan (2013)

Η ανάλυση ευαισθησίας είναι μία ακόμη προσέγγιση που έχει προταθεί για την αντιμετώπιση του προβλήματος του ελέγχου της υπόθεσης MCAR (Diggle (1989)). Σύμφωνα με αυτήν, ο μηχανισμός των ελλιπών δεδομένων και τα δεδομένα μοντελοποιούνται από κοινού και πραγματοποιείται ανάλυση ευαισθησίας ως προς τις παραμέτρους του μοντέλου. Για παράδειγμα, η κοινή κατανομή του μηχανισμού των ελλιπών δεδομένων και των παρατηρούμενων δεδομένων παραγοντοποιείται στην περιθώρια κατανομή των παρατηρούμενων δεδομένων και την υπό συνθήκη κατανομή του μηχανισμού των ελλιπών δεδομένων, όταν δίνονται τα παρατηρούμενα δεδομένα. Στη συνέχεια, οι παράμετροι αυτής της υπό συνθήκη κατανομής, η οποία στοχεύει στη μοντελοποίηση του μηχανισμού των ελλιπών δεδομένων, μεταβάλλονται και/ή ελέγχονται ώστε να εκτιμηθεί η ευαισθησία των εκτιμήσεων των παραμέτρων στην περιθώρια κατανομή των παρατηρούμενων δεδομένων (βλ., π.χ., Diggle (1989), Jamshidian and Yuan (2013)).

Ωστόσο, ένα πρόβλημα αυτής της προσέγγισης είναι το επιπλέον, ας μας επιτραπεί η έκφραση, κόστος που συνεπάγεται η μοντελοποίηση του μηχανισμού των ελλιπών δεδομένων. Στις περισσότερες περιπτώσεις, η επιλογή κατάλληλου μοντέλου για τον μηχανισμό των ελλιπών δεδομένων δεν είναι σαφής. Για την αποφυγή αυτού του προβλήματος οι Jamshidian and Mata (2008) πρότειναν μία μέθοδο ανάλυσης ευαισθησίας που δεν χρειάζεται τον προσδιορισμό ενός μοντέλου για τον μηχανισμό των ελλιπών δεδομένων. Σύμφωνα με τη μέθοδό τους ένα μοναδικό μοντέλο προσαρμόζεται σε όλα τα δεδομένα και ταυτόχρονα σε ένα υποσύνολο των δεδομένων. Έπειτα ένα μέτρο απόκλισης μεταξύ των εκτιμητών των παραμέτρων χρησιμοποιείται για τον έλεγχο της MCAR υπόθεσης (Jamshidian and Yuan (2013)). Ειδικότερα, η μεθοδολογία τους αναπτύσσεται με βάση μια παρατήρηση των συγγραφέων ότι τα θηκογράμματα για τις εκτιμώμενες μεταβλητές υπό την υπόθεση MCAR, θα βρίσκονται πολύ κοντά μεταξύ τους, ενώ θα απομακρύνονται υπό την NMAR υπόθεση. Με βάση αυτήν την παρατήρηση προκύπτει μια στατιστική συνάρτηση, η οποία ποσοτικοποιεί το πόσο "κοντά" βρίσκονται τα θηκογράμματα. Η μέθοδος τους βασίζεται στην εφαρμογή μιας μεθόδου τύπου "bootstrap", για την οποία η υλοποίηση σε πειραματικές μονάδες με ελλιπή δεδομένα δεν είναι ξεκάθαρη και παρουσιάζει θέματα σύγκλισης, ενώ ταυτόχρονα, απαιτείται η ευρετική εκχώρηση μιας τιμής από τον χρήστη η οποία δεν αποτελεί προϊόν κάποιου θεωρητικού αποτελέσματος. Παράλληλα, ο προτεινόμενος έλεγχος υλοποιείται μόνο για την εναλλακτική υπόθεση όπου τα δεδομένα είναι NMAR. Για την αποφυγή των παραπάνω προβλημάτων, οι Jamshidian and Yuan (2013) αντί της bootstrap μεθοδολογίας πρότειναν τη χρήση ασυμπτωτικών μεθόδων για τον προσδιορισμό διαστημάτων εμπιστοσύνης για

τις εκτιμώμενες παραμέτρους στη βάση των δύο διαφορετικών δειγμάτων. Για αυτούς τους λόγους, αντικείμενο μελέτης στο υπόλοιπο αυτής της ενότητας είναι η παρουσίαση του ελέγχου των Jamshidian and Yuan (2013).

Στο παραπάνω πλαίσιο, έστω ότι μας ενδιαφέρει η εφαρμογή ενός μοντέλου $\mathcal{M}$, με $k$ παραμέτρους, πάνω στα διαθέσιμα δεδομένα. Συμβολίζουμε το διαθέσιμο $n \times p$ σύνολο δεδομένων με $\mathbf{Y}$ και το υποσύνολο που αναφέραμε προηγουμένως, ως $\mathbf{S}_j$. Επισημαίνεται ότι το σύνολο αυτό θα αποτελεί μια από τις $J$ διακριτές υποομάδες, που δημιουργείται με βάση το μοτίβο ελλιπών τιμών, για όσα θα περιγραφούν παρακάτω. Επιπλέον, ορίζουμε με $\boldsymbol{\theta}$ το $k$-διάστατο διάνυσμα των παραμέτρων, που προκύπτουν από την προσαρμογή του μοντέλου $\mathcal{M}$ στο σύνολο $\mathbf{Y}$. Αντίστοιχα, συμβολίζουμε με $\boldsymbol{\xi}$ το διάνυσμα των παραμέτρων για το σύνολο $\mathbf{S}_j$. Προφανώς τα δύο διανύσματα αναφέρονται στις ίδιες παραμέτρους, αλλά οι εκτιμητές τους που συμβολίζουμε ως $\hat{\boldsymbol{\theta}}$ και $\hat{\boldsymbol{\xi}}$, γενικά θα διαφέρουν.

Η λογική του προτεινόμενου ελέγχου βασίζεται στο σκεπτικό που θα περιγραφεί στη συνέχεια. Αν ισχύει η υπόθεση MCAR για το διαθέσιμο σύνολο δεδομένων, τότε η υποομάδα $\mathbf{S}_j$ αποτελεί ένα τυχαίο υπο-δείγμα του διαθέσιμου συνόλου και επομένως αναμένουμε οι παράμετροι των δύο μοντέλων να μην διαφέρουν στατιστικά σημαντικά. Αν λοιπόν εξετάσουμε μεμονωμένα κάθε μία από τις $k$ παραμέτρους για τα δύο μοντέλα και έστω και μία διαφέρει σημαντικά τότε απορρίπτεται η υπόθεση MCAR. Επομένως, ο προτεινόμενος έλεγχος βασίζεται σε $k$ διαστήματα εμπιστοσύνης για τον έλεγχο ισότητας των παραμέτρων.

Η κατασκευή των διαστημάτων εμπιστοσύνης θα βασιστεί στην ασυμπτωτική κατανομή των ποσοτήτων $(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\xi}})$ και επομένως τα όποια συμπεράσματα θα ισχύουν ασυμπτωτικά. Στη συνέχεια εκτιμούμε τις παραμέτρους $\boldsymbol{\theta}$ και $\boldsymbol{\xi}$, μέσω της μεθόδου της Μέγιστης Πιθανοφάνειας και οι εκτιμητές $\hat{\boldsymbol{\theta}}$ και $\hat{\boldsymbol{\xi}}$ θα συμβολίζουν τους Ε.Μ.Π., που προκύπτουν μεγιστοποιώντας τον λογάριθμο της συνάρτησης πιθανοφάνειας και οι οποίοι ικανοποιούν τη σχέση:

$$\mathbf{g}(\hat{\boldsymbol{\theta}}) = \mathbf{0}. \tag{2.14}$$

Εφαρμόζοντας στη σχέση (2.14) ανάπτυγμα Taylor γύρω από την αληθινή τιμή $\boldsymbol{\theta}^*$ προκύπτει η σχέση:

$$\mathbf{0} = \mathbf{g}(\hat{\boldsymbol{\theta}}) = \mathbf{g}(\boldsymbol{\theta}^*) + \mathbf{H}(\bar{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*), \tag{2.15}$$

όπου με $\mathbf{H}(\bar{\boldsymbol{\theta}})$ συμβολίζεται ο Εσσιανός πίνακας του λογαρίθμου της συνάρτησης πιθανοφάνειας υπολογιζόμενος στην ενδιάμεση τιμή $\bar{\boldsymbol{\theta}}$. Η παραπάνω διαδικασία εφαρμόζεται εντελώς ανάλογα για το διάνυσμα $\boldsymbol{\xi}$. Έπειτα από πράξεις και εισάγοντας την υπόθεση ότι $n/m_j$ ($m_j$ ο αριθμός των πειραματικών μονάδων που

ανήκουν στην ομάδα $\mathbf{S}_j$) συγκλίνει σε μία σταθερά $r$, καθώς το $n$ τείνει στο άπειρο, αποδεικνύεται ότι (βλέπε Jamshidian and Yuan (2013)):

$$\sqrt{n}[(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\xi}}) - (\boldsymbol{\theta}^* - \boldsymbol{\xi}^*)] \xrightarrow{L} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}), \tag{2.16}$$

όπου

$$\boldsymbol{\Omega} = \left[\frac{-\boldsymbol{A}_{\boldsymbol{\xi}}^{-\mathbf{1}}}{r}, \boldsymbol{A}_{\boldsymbol{\theta}}^{-\mathbf{1}}\right] \mathbf{B} \left[\frac{-\boldsymbol{A}_{\boldsymbol{\xi}}^{-\mathbf{1}}}{r}, \boldsymbol{A}_{\boldsymbol{\theta}}^{-\mathbf{1}}\right]^t,$$

$$\mathbf{B} = lim \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} \mathbf{g_{xi}}(\boldsymbol{\xi}^*) I_{\{i \leq m\}} \\ \mathbf{g_i}(\boldsymbol{\theta}^*) \end{pmatrix} (\mathbf{g_{xi}}(\boldsymbol{\xi}^*)^t I_{\{i \leq m\}}, \mathbf{g_i}(\boldsymbol{\theta}^*)^t)$$

και

$$\boldsymbol{A}_{\boldsymbol{\theta}} = -lim \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{H}_i(\boldsymbol{\theta}^*).$$

Εντελώς ανάλογα με την ποσότητα $\boldsymbol{A}_{\boldsymbol{\theta}}$, ορίζεται και η ποσότητα $\boldsymbol{A}_{\boldsymbol{\xi}}$. Από τα διαγώνια στοιχεία του πίνακα $\boldsymbol{\Omega}$, έχουμε στη διάθεσή μας τα ασυμπτωτικά τυπικά σφάλματα των ποσοτήτων $\sqrt{n}(\hat{\theta}_i - \hat{\xi}_i)$ $, i = 1, ..., k$ και αντικαθιστώντας στις ποσότητες $\boldsymbol{A}_{\boldsymbol{\theta}}$, $\boldsymbol{A}_{\boldsymbol{\xi}}$ και $\boldsymbol{B}$ τις άγνωστες ποσότητες $\boldsymbol{\theta}^*$ και $\boldsymbol{\xi}^*$ με τους Ε.Μ.Π. προκύπτουν τα $100(1-\alpha)\%$ διαστήματα εμπιστοσύνης για τις $k$ παραμέτρους:

$$(\hat{\theta}_i - \hat{\xi}_i) \pm z_{\alpha/2} \sqrt{\frac{\hat{\Omega}_{ii}}{n}}. \tag{2.17}$$

Προφανώς για την αξιοποίηση των διαστημάτων εμπιστοσύνης, λόγω της συσχέτισης μεταξύ τους εφαρμόζεται διόρθωση Bonferroni και από τη σχέση (2.17), τελικά προκύπτουν τα ακόλουθα διαστήματα εμπιστοσύνης:

$$(\hat{\theta}_i - \hat{\xi}_i) \pm z_{\alpha/(2k)} \sqrt{\frac{\hat{\Omega}_{ii}}{n}}, \ i = 1, ..., k. \tag{2.18}$$

Συνοψίζοντας, η μέθοδος των Jamshidian and Yuan (2013) προϋποθέτει την επιλογή ενός κατάλληλου υποσυνόλου από το σύνολο των δεδομένων. Στη συνέχεια, το μοντέλο $M$, το οποίο είναι αποδεκτό αν ο μηχανισμός έλλειψης μπορεί να αγνοηθεί, προσαρμόζεται τόσο στο υποσύνολο του όσο και στο αρχικό σύνολο. Έπειτα, η κεντρική ιδέα είναι να αξιολογηθεί η διαφορά στην προσαρμογή του μοντέλου στα δύο σύνολα μέσω της κατασκευής διαστημάτων εμπιστοσύνης για κάθε συνιστώσα της παραμέτρου. Από τα παραπάνω είναι σαφές ότι η επιλογή του υποσυνόλου είναι καθοριστική για την υλοποίηση αυτού του ελέγχου. Αν

είναι εκ των προτέρων γνωστό ότι ένα από τα δύο υποσύνολα είναι είτε MCAR ή MAR ή ότι αντιπροσωπεύει τον πληθυσμό, τότε η απόρριψη της ισότητας των παραμέτρων θα σημαίνει ότι τα δεδομένα δεν είναι MAR. Ωστόσο, στις περισσότερες πρακτικές εφαρμογές τέτοια πληροφορία δεν είναι διαθέσιμη. Τέλος, η εφαρμογή της μεθόδου απαιτεί τον προσδιορισμό των Εσσιανών πινάκων $\boldsymbol{A_\theta}$ και $\boldsymbol{A_\xi}$. Ο προσδιορισμός αυτός εξαρτάται από το μοντέλο που θα υιοθετηθεί. Τα παραπάνω, ίσως, αιτιολογούν το γεγονός ότι ο έλεγχος αυτός δεν είναι ιδιαίτερα δημοφιλής, δεν έχει υλοποιηθεί στη γλώσσα R, ενώ, όπως επισημαίνεται από τους ίδιους τους συγγραφείς, μια άμεση σύγκρισή του με τους γνωστούς ελέγχους των Little (1988), Kim and Bentler (2002) και Jamshidian and Jalal (2010) δεν είναι εφικτή. Αυτός είναι και ο λόγος που δεν υλοποιείται στο σύνολο δεδομένων PimaIndiansDiabetes2.

# ΚΕΦΑΛΑΙΟ 3

# Ελεγχοι ανεξαρτητοι της ομοιογενειας

## 3.1 Εισαγωγή

Όπως εκτενώς αναφέρθηκε στο προηγούμενο κεφάλαιο, οι έλεγχοι της υπόθεσης MCAR που εκεί παρουσιάστηκαν ουσιαστικά μετατρέπουν τον έλεγχο αυτόν σε έλεγχο ομοιογένειας. Καθώς οι έλεγχοι αυτοί έχουν αρκετές αδυναμίες, όπως για παράδειγμα όταν έχουμε ομάδες δεδομένων με μικρό πλήθος πειραματικών μονάδων και μεγάλο αριθμό μεταβλητών, έχουν παρουσιαστεί διάφοροι άλλοι έλεγχοι, οι οποίοι δεν υλοποιούνται υπό το ίδιο πρίσμα. Στο κεφάλαιο αυτό θα παρουσιαστούν οι κυριότεροι έλεγχοι που δεν μπορούν να ταξινομηθούν στην κατηγορία των ελέγχων ομοιογένειας.

## 3.2 Ο έλεγχος του Aleksić (2024)

Όπως είδαμε σε προηγούμενες ενότητες παρουσιάστηκαν εκτενώς έλεγχοι για την υπόθεση MCAR, που στηρίζονται, κυρίως, σε ελέγχους ομοιογένειας, είτε των παραμέτρων ή των περιθωρίων κατανομών, στις διάφορες ομάδες που δημιουργούνται με βάση τα μοτίβα ελλιπών τιμών. Οι έλεγχοι αυτοί αρκετά συχνά παρουσιάζουν μειονεκτήματα λόγω του τρόπου κατασκευής τους. Εύλογα λοιπόν τίθεται το ερώτημα για το αν είναι εφικτή η δημιουργία ενός ελέγχου που δεν βασίζεται σε αυτήν τη λογική. Απάντηση σε αυτό το ερώτημα δίνει ο έλεγχος του Aleksić (2024) που θα αποτελέσει αντικείμενο παρουσίασης σε αυτήν την ενότητα.

Ο έλεγχος που προτάθηκε από τον Aleksić (2024) είναι μη παραμετρικός και υλοποιείται υπολογίζοντας τις συνδιακυμάνσεις μεταξύ δείκτριων συναρτήσεων

και πλήρως παρατηρούμενων πειραματικών μονάδων για κάποιες υπό μελέτη μεταβλητές. Από την παραπάνω διατύπωση είναι προφανές ότι ο έλεγχος προϋποθέτει την ύπαρξη πλήρως παρατηρηθέντων πειραματικών μονάδων για τουλάχιστον μία μεταβλητή ενδιαφέροντος. Επισημαίνεται ότι μια τέτοια υπόθεση δεν είναι περιοριστική ούτε και μη ρεαλιστική, αφού πρακτικά σε σύνολα μεγάλου πλήθους μεταβλητών συναντάται τουλάχιστον μία πλήρως παρατηρούμενη μεταβλητή. Τέλος, όπως θα δούμε, ο έλεγχος αυτός βασίζεται κυρίως στη θεωρία των U-Statistics του Hoeffding (1948) και στις ασυμπτωτικές ιδιότητες αυτών.

Για διευκόλυνση στην παρουσίαση όσων ακολουθούν υποθέτουμε ότι υπάρχουν $p$ πλήρως παρατηρηθέντα διανύσματα $\mathbf{X}_u,\ u = 1, ..., p$ που κωδικοποιούν τις μεταβλητές αυτές και αντίστοιχα $q$ το πλήθος διανύσματα με ελλιπή δεδομένα $\mathbf{Y},\ v = 1, ..., q$, τα οποία κωδικοποιούν τις εναπομείνασες μεταβλητές, καθώς επίσης και τα αντίστοιχα διανύσματα των δείκτριων συναρτήσεων $\mathbf{r}_v,\ v = 1, .., q$. Ταυτόχρονα θεωρούμε ότι έχουμε στη διάθεσή μας ένα δείγμα $n$ το πλήθος πειραματικών μονάδων. Υπό τον ορισμό του MCAR μηχανισμού, ισχύει, γενικά, ότι:

$$Pr(\mathbf{r}|\mathbf{X}, \mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})) \overset{MCAR}{=} Pr(\mathbf{r}). \tag{3.1}$$

Από τη σχέση (3.1) προκύπτει για τις πλήρως παρατηρούμενες μεταβλητές η σχέση $Pr(\mathbf{X}|\mathbf{r}) = Pr(\mathbf{X})$, αφού τα πλήρη δεδομένα μπορούν να θεωρηθούν ως μία ομάδα ελλιπών δεδομένων. Η προηγούμενη σχέση πρακτικά ισοδυναμεί με ανεξαρτησία των δεδομένων και των δείκτριων συναρτήσεων, δηλαδή ισχύει ότι:

$$Pr(\mathbf{X}, \mathbf{r}) = Pr(\mathbf{X})Pr(\mathbf{r}).$$

Ως αποτέλεσμα του παραπάνω συλλογισμού προκύπτει η σχέση:

$$E(\mathbf{X})E(\mathbf{r}) \overset{MCAR}{=} E(\mathbf{Xr}) \ \Leftrightarrow \ E(\mathbf{X})E(\mathbf{r}) - E(\mathbf{Xr}) = 0. \tag{3.2}$$

Επομένως, ορίζοντας $\theta = E(\mathbf{X})E(\mathbf{r}) - E(\mathbf{Xr})$, ένα κριτήριο για την ισχύ της υπόθεσης MCAR μπορεί να αποτελέσει το κατά πόσο κοντά είναι η παράμετρος $\theta$ στο 0. Αν η παράμετρος $\theta$ διαφέρει στατιστικά σημαντικά από το 0, τότε απορρίπτεται η υπόθεση MCAR. Απαραίτητη για τον έλεγχο της σημαντικότητας είναι η εισαγωγή μιας στατιστικής συνάρτησης. Για αυτόν τον σκοπό αρχικά θα πρέπει να προσδιοριστεί μία αμερόληπτη εκτιμήτρια της προηγούμενης ποσότητας. Ειδικότερα, ο Aleksić (2024) πρότεινε τη στατιστική συνάρτηση:

$$T_n^{(u,v)} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} X_{iu} R_{jv} - \frac{1}{n} \sum_{i=1}^{n} X_{iu} R_{iv} = U_1^{(u,v)} - U_2^{(u,v)} \tag{3.3}$$

για $u = 1, ..., p,\ v = 1, ..., q$, όπου

$$U_1^{(u,v)} = \frac{1}{n(n-1)} \sum_{1 \le i < j \le n} \sum \Phi\left(X_{iu}, Y_{iv}, R_{iv}\right), \tag{3.4}$$

και

$$U_2^{(u,v)} = \frac{1}{n} \sum_{i=1}^{n} \Psi\left(X_{iu}, Y_{iv}, R_{iv}\right), \tag{3.5}$$

με

$$\Phi\left(X_{iu}, Y_{iv}, R_{iv}\right) = \frac{1}{2}(X_{iu}R_{jv} + X_{ju}R_{iv})$$

και

$$\Psi\left(X_{iu}, Y_{iv}, R_{iv}\right) = (X_{iu}R_{iv}).$$

Κατά αυτόν τον τρόπο, ο Aleksić (2024) κατορθώνει να συνδέσει την υπόθεση MCAR με την ασυμπτωτική θεωρία των U–Statistics, αφού οι ποσότητες των σχέσεων (3.4) και (3.5) αποτελούν U–Statistics με πυρήνες τις ποσότητες $\Phi\left(X_{iu}, Y_{iv}, R_{iv}\right)$ και $\Psi\left(X_{iu}, Y_{iv}, R_{iv}\right)$ (βλέπε Wand and Jones (1994)). Ανακαλώντας στη συνέχεια τα αποτελέσματα για τα U-statistics που δόθηκαν από τον Hoeffding (1948), έχουμε ότι για μεγάλο μέγεθος δείγματος:

$$\left(\sqrt{n}T_n^{(1,1)}, ..., \sqrt{n}T_n^{(1,q)}, ..., \sqrt{n}T_n^{(p,q)}\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \tag{3.6}$$

όπου

$$\mathbf{\Sigma} = \Big[Cov\Big((X^{(\lfloor i/p \rfloor)}), X^{(\lfloor j/p \rfloor)}\Big) Cov\Big(r^{(i \ mod \ p)}, r^{(j \ mod \ p)}\Big)\Big]_{i,j \in \{1,...,pq\}},$$

με $i \ mod \ p$ να είναι το υπόλοιπο της διαίρεσης του $i = 1, \ldots, pq$ με το $p$ και $\lfloor \cdot \rfloor$ η συνάρτηση του ακέραιου μέρους.

Από τη σχέση (3.6) προκύπτει ότι:

$$\left(\hat{\mathbf{\Sigma}}^{-1/2} \left(\sqrt{n}T_n^{(1,1)}, ..., \sqrt{n}T_n^{(1,q)}, ..., \sqrt{n}T_n^{(p,q)}\right)^T\right)^T \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

όπου $\hat{\mathbf{\Sigma}}$ ο εκτιμητής του πίνακα $\mathbf{\Sigma}$, που δόθηκε παραπάνω, ο οποίος προκύπτει μέσω των δειγματικών συνδιακυμάνσεων που τον απαρτίζουν.

**Παρατήρηση 3.2.1** *Για την εκτίμηση του πίνακα έχει γίνει χρήση των αμερόληπτων εκτιμητών. Επίσης, απαραίτητη είναι και η αντιστροφή του πίνακα* $\hat{\mathbf{\Sigma}}$*, η οποία εξασφαλίζεται υπό την προϋπόθεση, ότι οι πλήρως παρατηρούμενες μεταβλητές έχουν πεπερασμένες ροπές τέταρτης τάξης.*

Επομένως, ορίζοντας την εκάστοτε συνιστώσα του παραπάνω στατιστικού ως $A_n^{(u,v)}$, προκύπτει έπειτα από τον γραμμικό μετασχηματισμό των συνιστωσών του παραπάνω διανύσματος ότι:

$$A_n = \sum_{u=1}^{p} \sum_{v=1}^{q} \left( A_n^{(u,v)} \right)^2 \sim \chi^2_{pq}. \tag{3.7}$$

Έχοντας προσδιορίσει την ελεγχοσυνάρτηση η οποία υπό την υπόθεση MCAR ακολουθεί κατανομή $\chi^2_{pq}$, μπορούμε να προσδιορίσουμε την κρίσιμη περιοχή του ελέγχου. Προφανώς η μηδενική υπόθεση θα απορρίπτεται για μεγάλες τιμές του στατιστικού, άρα η κρίσιμη περιοχή είναι:

$$A_n > \mathcal{X}^2_{pq,\alpha},$$

με $\chi^2_{pq,\alpha}$ να είναι το άνω $\alpha$ ποσοστιαίο σημείο της χι τετράγωνο κατανομής με $pq$ βαθμούς ελευθερίας.

**Παρατήρηση 3.2.2** *Για την ειδική περίπτωση όπου παρατηρούνται ελλιπείς τιμές μόνο σε μία μεταβλητή, ο Aleksić (2024) αποδεικνύει ότι η στατιστική συνάρτηση $A_n$, ταυτίζεται με τη στατιστική συνάρτηση που προτείνει ο Little για τον έλεγχο της υπόθεσης MCAR.*

Ο έλεγχος του Aleksić παρουσιάζει το μειονέκτημα ότι η υλοποίησή του προϋποθέτει την ύπαρξη μίας πλήρως παρατηρούμενης μεταβλητής (στήλης) και στηρίζεται σε εκείνες που περιέχουν ελλιπείς τιμές, δηλαδή τις $Y_{iv}$ που ορίστηκαν προηγουμένως. Αυτό έχει ως συνέπεια να μην μπορεί να ανιχνεύσει αποκλίσεις από την MCAR υπόθεση σε κάποιες περιπτώσεις, όπως για παράδειγμα είναι εκείνη όπου οι δείκτριες συναρτήσεις βασίζονται στις στήλες $Y_{iv}$. Ο Aleksić παρακινούμενος από αυτές τις αδυναμίες γενικεύει τον έλεγχο του, στοχεύοντας στη συμπερίληψη των ποσοτήτων που περιγράφονται από τη σχέση (3.3) αντικαθιστώντας τις ποσότητες $X_{iu}$, με τις $Y_{iv}$. Ειδικότερα, ακολουθώντας παρόμοια διαδικασία με αυτήν που περιγράφτηκε σε αυτήν την ενότητα, στο υπό κρίση άρθρο του που είναι διαθέσιμο στο αποθετήριο Arxiv (βλέπε Aleksić (2025)) ορίζει μία νέα γενικευμένη στατιστική συνάρτηση $A_n^{'}$, η οποία, υπό τη μηδενική υπόθεση, ακολουθεί, ασυμπτωτικά, χι τετράγωνο κατανομή με $pq + q * (q-1)$ βαθμούς ελευθερίας. Η διαφοροποίηση των βαθμών ελευθερίας οφείλεται στο γεγονός ότι σε αυτόν τον έλεγχο συμπεριλαμβάνονται οι επιπλέον $q$ το πλήθος μη παρατηρούμενες μεταβλητές. Επομένως, η υπόθεση MCAR απορρίπτεται, σε ασυμπτωτικό $\alpha$ επίπεδο σημαντικότητας, αν $A_n^{'} > \mathcal{X}^2_{pq+q*(q-1),\alpha}$.

**Εφαρμογή στα δεδομένα PimaIndiansDiabetes2** Η υλοποίηση του ελέγχου του Aleksić (2024), δηλαδή του ελέγχου με τη σ.σ. $A_n$ περιέχεται στο παράρτημα της μεταπτυχιακής διατριβής. Η υλοποίησή του επιτυγχάνεται στην R μέσω της συνάρτησης Aleksic_Test(). Η συνάρτηση απαιτεί την εκχώρηση στο όρισμα "data" ενός συνόλου δεδομένων της κλάσης data.frame() και στο όρισμα "alpha" την εκχώρηση του επιπέδου σημαντικότητας για τον έλεγχο. Επισημαίνεται ότι αναγκαία για τη λειτουργία της συνάρτησης είναι τα πακέτα "expm" και "MVN", ενώ απαραίτητη προϋπόθεση για τη σωστή λειτουργία του ελέγχου είναι το σύνολο των συνεχών δεδομένων να περιέχει τουλάχιστον μία πλήρως παρατηρούμενη μεταβλητή και δεν λειτουργεί για την περίπτωση που έχουμε μόνο μία πλήρως παρατηρούμενη και μία που περιέχει ελλιπείς τιμές μεταβλητή. Η συνάρτηση Aleksic_Test(), τυπώνει την τιμή του στατιστικού, το κρίσιμο σημείο, την $p-$τιμή του ελέγχου και ελέγχει αν η τιμή της στατιστικής συνάρτησης ξεπερνά αυτή του κρίσιμου σημείου ώστε να απορριφθεί η υπόθεση MCAR. Επιπλέον, κώδικες στην R για την υλοποίηση των ελέγχων μέσω των στατιστικών συναρτήσεων $A_n^{'}$ και $A_n$ παρέχονται στο προφίλ του συγγραφέα στη πλατφόρμα Github μέσω του συνδέσμου https://github.com/danijel-g-aleksic. Για το σύνολο δεδομένων PimaIndiansDiabetes2 η $p$-τιμή της στατιστικής συνάρτησης $A_n$ είναι ίση με $1.550767e-13$ και υποδηλώνει ότι απορρίπτεται η MCAR υπόθεση. Στο ίδιο συμπέρασμα καταλήγουμε και με τη στατιστική συνάρτηση $A_n^{'}$, καθώς η $p$-τιμή του ελέγχου είναι ίση με $1.560749e-91$. Για την υλοποίηση του τελευταίου ελέγχου είναι απαραίτητη η λήψη της βιβλιοθήκης `corpor` και ο καθορισμός των στηλών που περιέχουν πλήρη και ελλιπή δεδομένα.

## 3.3 Ο έλεγχος των Rouzinov and Berchtold (2022)

Στη βιβλιογραφία των ελέγχων για την υπόθεση MCAR κρίνεται συχνά απαραίτητη η χρήση μοντέλων παλινδρόμησης και η συμπλήρωση ελλιπών τιμών, με στόχο τη διερεύνηση του μηχανισμού παραγωγής ελλιπών τιμών. Σε αυτό το πλαίσιο περιλαμβάνεται και ο έλεγχος που προτείνουν οι Rouzinov and Berchtold (2022). Ο έλεγχος των Rouzinov and Berchtold (2022) είναι αρκετά γενικός, καθώς υλοποιείται σε περιπτώσεις συνεχών, διακριτών και κατηγορικών ελλιπών δεδομένων, ωστόσο θα πρέπει να επισημανθεί το μειονέκτημά του ότι δεν ανιχνεύει την περίπτωση, όπου τα ελλιπή δεδομένα παράγονται από τον μηχανισμό NMAR. Ουσιαστικά δηλαδή πρόκειται για έλεγχο της εξής υπόθεσης:

$$H_0 : MCAR \text{ έναντι της } H_1 : MAR.$$

Η υλοποίηση του ελέγχου γίνεται εύκολα αντιληπτή, αν αρχικά περιοριστούμε

στον έλεγχο της υπόθεσης MCAR σε μία μόνο μεταβλητή. Ειδικότερα, σε αυτό το πλαίσιο, ας υποθέσουμε ότι θέλουμε να ελέγξουμε την υπόθεση MCAR στη μεταβλητή $Y_1$, θεωρώντας ότι οι υπόλοιπες έχουν παρατηρηθεί πλήρως. Όπως και στις προηγούμενες ενότητες αναφερόμαστε σε ένα σύνολο $n$ το πλήθος πειραματικών μονάδων στις οποίες μετρώνται $p$ το πλήθος μεταβλητές. Χωρίζουμε κατά τα γνωστά τις πειραματικές μονάδες σε δύο ομάδες ανάλογα με το μοτίβο ελλιπών τιμών. Άρα η ομάδα $S_1$ θα περιέχει $m_1$ πλήρεις πειραματικές μονάδες και η $S_2$ τις υπόλοιπες. Χρησιμοποιώντας μια κατάλληλη συνάρτηση σύνδεσης (link function) $g(\cdot)$, σε πρώτη φάση για την ομάδα $S_1$ δημιουργούμε ένα μοντέλο παλινδρόμησης για την εκτίμηση της μεταβλητής απόκρισης την $Y_1$ με όλες τις υπόλοιπες μεταβλητές ως επεξηγηματικές, δηλαδή έχουμε το μοντέλο πρόβλεψης:

$$\widehat{Y}_{1,S_1} = g(Y_2, ..., Y_p) \tag{3.8}$$

Από τη σχέση (3.8) προκύπτει ένα διάνυσμα εκτιμητών $\hat{\beta}$. Με βάση αυτούς τους εκτιμητές και τις πειραματικές μονάδες της ομάδας $S_2$, εκτιμούμε την μεταβλητή $Y_1$ όπου είναι ελλιπής και προκύπτει:

$$\widehat{Y}_{1,S_2} = g(Y_2, ..., Y_p) \tag{3.9}$$

Υπό την υπόθεση MCAR η παρουσία των ελλιπών δεδομένων στη μεταβλητή $Y_1$, δεν μπορεί να επεξηγηθεί από το σύνολο των επεξηγηματικών μεταβλητών που ανήκουν στο σύνολο, αφού η δημιουργία των ελλιπών τιμών είναι ανεξάρτητη των δεδομένων. Σε διαφορετική περίπτωση, αν για τα δεδομένα ισχύει η υπόθεση MAR, ένα μέρος των ελλιπών τιμών μπορεί να επεξηγηθεί από τα δεδομένα, εξαιτίας της συσχέτισης μηχανισμού και δεδομένων. Άρα η απόρριψη ή μη της υπόθεσης MCAR έγκειται στη διαφορετικότητα των εκτιμώμενων ποσοτήτων, η οποία οφείλεται στη διαφορετικότητα των δεδομένων των συνόλων $S_1$ και $S_2$.

Απομένει η κατασκευή μιας στατιστικής συνάρτησης για τον έλεγχο της ισότητας των εκτιμώμενων ποσοτήτων. Σύμφωνα με τους Rouzinov and Berchtold (2022), εφόσον ο έλεγχος υλοποιείται για διαφορετικού τύπου δεδομένα, έχουμε διαφορετικού τύπου ελεγχοσυναρτήσεις, όπως και συναρτήσεις σύνδεσης. Πιο συγκεκριμένα, για τη συνεχή περίπτωση η συνάρτηση σύνδεσης είναι η κλασική που εφαρμόζεται στα γραμμικά μοντέλα παλινδρόμησης. Στη συνέχεια, η σύγκριση ανάγεται στον έλεγχο της ισότητας ότι τα δύο δείγματα προέρχονται από τον ίδιο πληθυσμό. Η σύγκριση αυτή επιτυγχάνεται στην περίπτωση που τα δεδομένα προέρχονται από κανονικό πληθυσμό μέσω του ελέγχου των Noé (1972), ενώ σε διαφορετική περίπτωση προτείνεται να χρησιμοποιείται η πολυδιάστατη εκδοχή του μη παραμετρικού έλεγχου που προτάθηκε από τον Marozzi (2009), η οποία

είναι διαθέσιμη στη γλώσσα προγραμματισμού R στο παράρτημα της εργασίας του Marozzi (2014).

Στην περίπτωση που τα δεδομένα είναι κατηγορικά με δύο κατηγορίες $a$ και $b$, ουσιαστικά αναφερόμαστε σε μοντέλα λογιστικής παλινδρόμησης με συνάρτηση σύνδεσης $g(\cdot) = \log\left(\frac{p}{1-p}\right),\ p = Pr(Y_{1,S_1} = a),\ 1-p = Pr(Y_{1,S_1} = b)$. Σε αυτήν την περίπτωση πρέπει να υλοποιηθούν δύο έλεγχοι καθώς έχουμε τέσσερα συνολικά ενδεχόμενα. Ένας έλεγχος για τη σύγκριση των εκτιμώμενων τιμών των πιθανοτήτων $Pr(Y_{1,S_1} = a)$ και $Pr(Y_{1,S_2} = a)$ και ένας έλεγχος για τη σύγκριση των εκτιμώμενων τιμών των πιθανοτήτων $Pr(Y_{1,S_1} = b)$ και $Pr(Y_{1,S_2} = b)$. Για τον έλεγχο της ισότητας αυτών των εκτιμώμενων πιθανοτήτων γίνεται ξανά χρήση του ελέγχου του Noé (1972), όμως καθώς χρησιμοποιούνται πολλαπλοί έλεγχοι στα ίδια δεδομένα υπάρχει κίνδυνος να απορριφθούν εσφαλμένα αληθείς υποθέσεις, άρα είναι αναγκαία η εφαρμογή μιας μεθόδου FDR (False Discovery Rate, q-value). Συγκεκριμένα προτείνεται η μέθοδος των Benjamini and Hochberg (1995).

Τέλος, σε περίπτωση ύπαρξης περισσοτέρων των δύο κατηγοριών προτείνεται να εφαρμοστεί το μοντέλο πολυωνυμικής παλινδρόμησης. Αν υποθέσουμε ότι υπάρχουν $j$ διαφορετικές κατηγορίες τότε, όπως στην προηγούμενη περίπτωση εφαρμόζουμε $j$ το πλήθος φορές τον έλεγχο του Noé (1972) και τη μέθοδο των Benjamini and Hochberg (1995) ως μέθοδο FDR.

Η παραπάνω διαδικασία περιγράφηκε για την περίπτωση μίας μεταβλητής με ελλιπή δεδομένα. Οι Rouzinov and Berchtold (2022) παρέχουν έναν τρόπο υλοποίησης της διαδικασίας για περισσότερες από μία μεταβλητές, ο οποίος τρόπος υλοποιείται σε δύο στάδια. Στο πρώτο στάδιο (το οποίο ονομάζουν στάδιο εκτίμησης) οι μεταβλητές ελέγχονται διαδοχικά για ελλιπείς τιμές. Ξεκινώντας από την πρώτη μεταβλητή που θα έχει ελλιπείς τιμές, υλοποιούν τη διαδικασία που περιγράφηκε χρησιμοποιώντας όμως ως επεξηγηματικές μεταβλητές μόνο εκείνες τις μεταβλητές που έχουν παρατηρηθεί πλήρως. Έπειτα ανάλογα με το αποτέλεσμα του ελέγχου που προηγήθηκε αποφασίζουν αν για τη συγκεκριμένη μεταβλητή ισχύει η υπόθεση MCAR ή η υπόθεση MAR. Ακολούθως, αν δεν έχει απορριφθεί η υπόθεση MCAR, χρησιμοποιείται ένα τυχαίο δείγμα από τα πλήρως παρατηρούμενα δεδομένα για να συμπληρωθούν οι ελλιπείς τιμές, ενώ αν έχει απορριφθεί για τη συμπλήρωση των δεδομένων εφαρμόζεται η μέθοδος συμπλήρωσης MICE (βλέπε Van Buuren and Oudshoorn (1999)), που έχει προταθεί για τέτοιες περιπτώσεις. Κατ' αυτόν τον τρόπο επαναλαμβάνοντας τη διαδικασία για τις υπόλοιπες μεταβλητές, αποκτούμε ένα πλήρες σύνολο δεδομένων. Στο δεύτερο στάδιο επαναλαμβάνεται η διαδικασία για κάθε μεταβλητή με ελλιπείς τιμές, όπως είχε περιγραφεί στο πρώτο στάδιο. Η σημαντική διαφοροποίηση

σε αυτό το στάδιο είναι ότι ως επεξηγηματικές μεταβλητές δεν χρησιμοποιούνται μόνο εκείνες που έχουν παρατηρηθεί πλήρως, άλλα και οι υπόλοιπες που έχουν γίνει πλήρεις μέσω της συμπλήρωσης των δεδομένων στο προηγούμενο στάδιο, με εξαίρεση τη μεταβλητή που σε κάθε επανάληψη έχει τον ρόλο της μεταβλητής απόκρισης. Ακολούθως ελέγχεται η υπόθεση της MCAR έναντι της MAR για κάθε μεταβλητή. Αν για την υπό μελέτη μεταβλητή προκύψει διαφορετικό συμπέρασμα από εκείνο που έχουμε καταλήξει στο προηγούμενο στάδιο, συμπεραίνουμε ότι πρέπει να κάνουμε εκ νέου συμπλήρωση των ελλιπών δεδομένων. Η διαδικασία επαναλαμβάνεται μέχρις ότου δεν υπάρχει διαφορά στα αποτελέσματα των δύο σταδίων μίας επανάληψης.

**Παρατήρηση 3.3.1** *Ο έλεγχος των Rouzinov and Berchtold (2022) παρά την ευελιξία που παρουσιάζει στη χρήση του περιέχει αρκετά μειονεκτήματα. Υποθέτει για την υλοποίησή του, ότι οι ελλιπείς τιμές παράγονται μόνο από τους μηχανισμούς MCAR και MAR. Αυτή η υπόθεση συνεπάγεται, ότι έχουμε λάβει υπόψη όλες τις σχετικές με τη μελέτη μεταβλητές. Αν όμως έχει αγνοηθεί έστω και μία μεταβλητή αυτό σημαίνει ότι οι ελλιπείς τιμές ενδεχομένως να οφείλονται στη μη ύπαρξη αυτής και επομένως ο μηχανισμός παραγωγής να είναι NMAR. Πρακτικά σημαίνει ότι για κάθε σύνολο δεδομένων πρέπει εξ' αρχής να γνωρίζουμε όλες τις μεταβλητές που πρέπει να συμπεριληφθούν, σενάριο που δεν είναι πάντα ρεαλιστικό. Επίσης, η εφαρμογή της μεθόδου στη γενική περίπτωση προϋποθέτει την ύπαρξη μίας πλήρως παρατηρούμενης μεταβλητής στο σύνολο. Για τους παραπάνω λόγους ο έλεγχος αυτός προτείνεται να χρησιμοποιείται συνδυαστικά μαζί με άλλους ελέγχους, έχοντας μία επιβεβαιωτική μορφή τουλάχιστον για τους μηχανισμούς που μελετά.*

**Εφαρμογή στα δεδομένα PimaIndiansDiabetes2** Η υλοποίηση της γενικής μορφής του ελέγχου πραγματοποιείται στη γλώσσα R με τη συνάρτηση `RBtest()` του πακέτου `RBtest`. Η συνάρτηση απαιτεί την εκχώρηση ενός συνόλου δεδομένων στο όρισμα data, το οποίο απαραίτητα θα πρέπει να περιέχει μία πλήρως παρατηρούμενη μεταβλητή. Υλοποιεί τη διαδικασία που περιγράφτηκε και μέσω του `$final` μπορεί κανείς να δει αν για την κάθε μεταβλητή ισχύει ή όχι η υπόθεση MCAR. Η τιμή 0 σημαίνει ότι έχουμε μεταβλητή με πλήρη δεδομένα, η τιμή -1 υποδηλώνει ότι για τη συγκεκριμένη μεταβλητή δεν μπορεί να απορριφθεί η MCAR έναντι της υπόθεσης MAR, ενώ η τιμή 1 υποδεικνύει ότι ισχύει η υπόθεση MAR. Για το σύνολο δεδομένων PimaIndiansDiabetes2 στην έξοδο των αποτελεσμάτων λαμβάνουμε 0,0,1,1,0,-1,-1, γεγονός που υποδεικνύει ότι η πρώτη (glucose), η δεύτερη (pressure) και η πέμπτη (mass) μεταβλητή δεν έχουν ελλιπείς τιμές, η τρίτη (triceps) και η τέταρτη (insulin) μεταβλητή έχουν

MAR δεδομένα, ενώ για τις δύο τελευταίες μεταβλητές (pedigree και age) δεν μπορεί να απορριφθεί η MCAR υπόθεση έναντι της MAR.

# ΚΕΦΑΛΑΙΟ 4

# ΣΥΓΚΡΙΤΙΚΕΣ ΜΕΛΕΤΕΣ

## 4.1 Εισαγωγή

Το ενδιαφέρον στα προηγούμενα δύο κεφάλαια αυτής της μεταπτυχιακής διατριβής επικεντρώθηκε στην παρουσίαση των κυριότερων μεθοδολογιών ελέγχου της υπόθεσης MCAR, αλλά και στη δημιουργία συναρτήσεων στην R για την υλοποίηση αυτών σε πρακτικές εφαρμογές. Ειδικότερα, οι μεθοδολογίες αυτές ταξινομήθηκαν σε δύο μεγάλες κατηγορίες ανάλογα με την κεντρική ιδέα τους. Ένα εύλογο ερώτημα που προκύπτει είναι αν υπάρχει κάποια αναγκαιότητα για την ύπαρξη διαφορετικών μεθοδολογιών ή μπορούμε να ισχυριστούμε ότι κάποια ή κάποιες εξ αυτών έχουν καλύτερη απόδοση συγκριτικά με τις υπόλοιπες. Η απάντηση σε αυτό το ερώτημα μπορεί να δοθεί μέσω της διεξαγωγής προσομοιώσεων για να διαπιστωθεί αν οι προτεινόμενες διαδικασίες ελέγχου διατηρούν το ονομαστικό επίπεδο ελέγχου και αν έχουν ικανοποιητική ισχύ. Η παραπάνω διαπίστωση οδήγησε πολλούς συγγραφείς να διεξάγουν συγκριτικές μελέτες στο πλαίσιο της δημοσίευσης μίας νέας μεθόδου. Στη συνέχεια αυτού του κεφαλαίου θα παραθέσουμε τα αποτελέσματα των συγκριτικών μελετών που έχουν διενεργηθεί στη βιβλιογραφία τόσο για τη διατήρηση του επιπέδου σημαντικότητας όσο και της ισχύος. Σε κάθε περίπτωση θα αναφέρονται οι συγκρινόμενες μεθοδολογίες και τα συμπεράσματα που έχουν εξαχθεί.

## 4.2 Αποτελέσματα συγκριτικών μελετών

Στην ενότητα αυτή θα παραθέσουμε τα αποτελέσματα των συγκριτικών μελετών που έχουν εμφανισθεί στη βιβλιογραφία.

**Συγκριτική μελέτη των Jamshidian and Yuan (2014)** Οι Jamshidian and Yuan (2014) διεξήγαγαν μία διεξοδική συγκριτική μελέτη στην οποία εξετάζεται η απόδοση του ελέγχου του Little (1988), του παραμετρικού και μη παραμετρικού ελέγχου των Jamshidian and Jalal (2010) και η απόδοση του ελέγχου που οι ίδιοι πρότειναν, ήτοι του ελέγχου των Jamshidian and Yuan (2014). Σε πρώτο στάδιο, οι συγγραφείς παράγουν 5000 προσομοιωμένα σύνολα δεδομένων μεγέθους $n$ από τη δισδιάστατη κανονική κατανομή με μηδενικό μέσο διάνυσμα και πίνακα διακυμάνσεων συνδιακυμάνσεων με διαγώνια στοιχεία ίσα με τη μονάδα και μη διαγώνια στοιχεία ίσα με $\rho$. Επιπλέον, παράγουν MAR και NMAR ελλιπή δεδομένα με τον τρόπο που θα περιγραφεί στη συνέχεια. Έστω ότι $(y_{i1}, y_{i2})^t$ είναι η $i$-οστή πειραματική μονάδα που παράγεται από τη δισδιάστατη κανονική κατανομή, τότε δημιουργούνται MAR δεδομένα αν εκχωρείται ελλιπής τιμή στη δεύτερη συνιστώσα κάθε φορά που: α) $y_{i1} \in [-\infty, -1.214425] \bigcup [-0.8, 0.8] \bigcup [1.214425, \infty]$ ή β) $y_{i1} \in [-\infty, -1] \bigcup [1, \infty]$. Από την άλλη δημιουργούνται NMAR δεδομένα αν εκχωρείται ελλιπής τιμή στη δεύτερη συνιστώσα κάθε φορά που γ) $y_{i2} \in [-\infty, -1.214425] \bigcup [-0.8, 0.8] \bigcup [1.214425, \infty]$ ή δ) $y_{i2} \in [-\infty, -1] \bigcup [1, \infty]$. Ουσιαστικά, σε όλες τις περιπτώσεις προκύπτει το μοτίβο των ελλιπών δεδομένων που είναι γνωστό ως 2 βηματικό μονότονα ελλιπές. Τα συμπεράσματα που προκύπτουν σε αυτό το πλαίσιο είναι ότι:

- ο έλεγχος του Little (1988), παρουσιάζει εξαιρετικά μικρή ισχύ, ενώ οι έλεγχοι των Jamshidian and Jalal (2010) παρουσιάζουν αρκετά υψηλό επίπεδο ισχύος ($\sim 99\%$), με την απόδοση του παραμετρικού ελέγχου να είναι καλύτερη.

- Ο έλεγχος των Jamshidian and Yuan (2014) έχει πολύ καλή απόδοση όταν τα δεδομένα είναι MAR, ωστόσο δεν έχει το ίδιο καλή απόδοση όταν τα δεδομένα είναι NMAR. Η απόδοσή του σε αυτήν την περίπτωση εξαρτάται από την παράμετρο $\rho$ και δεν αυξάνεται με την αύξηση του μεγέθους δείγματος. Συγκεκριμένα, για $n = 300, \rho = 0.5$ για τα δεδομένα που προέρχονται από τη δισδιάστατη κανονική κατανομή με τις ελλιπείς τιμές να παράγονται υπό τον πρώτο NMAR μηχανισμό, απέρριψε μόλις 4.8% των περιπτώσεων για επίπεδο σημανικότητας $\alpha = 5\%$. Παρατηρήθηκε ότι η αύξηση του μεγέθους του δείγματος δεν επηρέασε την ισχύ. Όμως, για $\rho = 0.8$ η ισχύς του ελέγχου έφτασε το επίπεδο του 12% για $n = 300$, 28% για $n = 1000$ και πολύ κοντά στο 100% για $n = 5000$.

Στη συνέχεια, οι Jamshidian and Yuan (2014) εξετάζουν την απόδοση των παραπάνω ελέγχων όταν τα 5000 σε πλήθος MAR και NMAR σύνολα

δεδομένων με ποσοστό ελλιπών δεδομένων 20% δημιουργούνται χρησιμοποιώντας μοντέλο λογιστικής παλινδρόμησης για την εκχώρηση ελλιπούς τιμής. Ειδικότερα, αφού έχουν παραχθεί δεδομένα από τη δισδιάστατη κανονική κατανομή, δημιουργούνται MAR σύνολα δεδομένων αν εκχωρείται ελλιπής τιμή στην παρατήρηση $y_{i2}$ με πιθανότητα $\exp(-2-1.65y_{i1})/(1+\exp(-2-1.65y_{i1})$ και NMAR αν εκχωρείται ελλιπής τιμή στην παρατήρηση $y_{i2}$ με πιθανότητα $\exp(-2-1.65y_{i2})/(1+\exp(-2-1.65y_{i2})$. Τα συμπεράσματα που προκύπτουν σε αυτό το πλαίσιο είναι ότι:

- ο έλεγχος του Little (1988) είναι εξαιρετικά ισχυρός για κάθε περίπτωση σε σχέση με τους ελέγχους των Jamshidian and Jalal (2010). Ειδικότερα, ο μη παραμετρικός έλεγχος των Jamshidian and Jalal (2010) δεν μπορεί να ανιχνεύσει ότι τα δεδομένα είναι MAR ή NMAR, ενώ η απόδοση του παραμετρικού ελέγχου των Jamshidian and Jalal (2010) εξαρτάται από την τιμή της παραμέτρου $\rho$, με την ισχύ να αυξάνει όταν μεγαλώνει η τιμή της παραμέτρου.
- Ο έλεγχος των Jamshidian and Yuan (2014) είναι αρκετά ισχυρός (απόδοση κοντά στο 100%).

Συνοψίζοντας, από τη συγκριτική μελέτη των Jamshidian and Yuan (2014) συμπεραίνουμε ότι ο έλεγχος τους είχε καλή απόδοση σε κάθε περίπτωση που τα δεδομένα ήταν MAR, ενώ μπορεί να προταθεί για τον έλεγχο της υπόθεσης MCAR έναντι της εναλλακτικής NMAR μόνο όταν γνωρίζουμε ότι οι μεταβλητές σχετίζονται σε μεγάλο βαθμό. Επιπρόσθετα, η απόδοση των ελέγχων των Jamshidian and Jalal (2010) επηρεάζεται από τον τρόπο δημιουργίας των ελλιπών δεδομένων. Τέλος, επισημαίνεται ότι τα όποια αποτελέσματα έχουν προέλθει προσομοιώνοντας τα αρχικά δεδομένα από δισδιάστατη κανονική κατανομή, γεγονός που περιορίζει σε αυτήν την περίπτωση τα όποια συμπεράσματα έχουν εξαχθεί.

**Συγκριτική μελέτη των Li and Yu (2015)** Οι Li and Yu (2015) διεξήγαγαν μία διεξοδική συγκριτική μελέτη του ελέγχου που προτάθηκε στην εργασία τους με τον έλεγχο των Jamshidian and Jalal (2010). Ακολουθώντας τον σχεδιασμό τον Jamshidian and Jalal (2010), που περιγράφηκε στην Ενότητα 2.4, θεωρούν 1000 το πλήθος προσομοιωμένα $p$-διάστατα σύνολα δεδομένων, $p = 4, 10$, μεγέθους $n$, με $n = 200, 500, 1000$ και ποσοστό πλήρων περιπτώσεων $r = 0.65, 0.35$. Τα δεδομένα που δημιουργούν προέρχονται από οκτώ διαφορετικές κατανομές, στις οποίες περιλαμβάνονται συμμετρικές κατανομές, όπως

η πολυδιάστατη κανονική, κατανομές με βαριές ουρές, όπως η πολυδιάστατη $t$, κατανομές με ελαφριές ουρές, όπως η πολυδιάστατη ομοιόμορφη κατανομή, ενώ η πολυδιάστατη Weibull μπορεί να θεωρηθεί ως ένα παράδειγμα λοξευμένης κατανομής.

Από τα αποτελέσματα που δίνονται σχετικά με την απόδοση αυτών των ελέγχων ως προς τη διατήρηση του επιπέδου σημαντικότητας συμπεραίνουμε ότι:

- το εμπειρικό επίπεδο σημαντικότητας του ελέγχου των Li and Yu (2015) για τους διάφορους συνδυασμούς μεγέθους δείγματος, μεταβλητών και ποσοστού ελλιπών τιμών διατηρείται κοντά στο ονομαστικό επίπεδο σημαντικότητας, ενώ
- ο έλεγχος των Jamshidian and Jalal (2010) παρουσιάζει διογκωμένο επίπεδο σημαντικότητας, γεγονός που έρχεται σε πλήρη συμφωνία με τα ευρήματα που παρατέθηκαν στην Ενότητα 2.4.

Για τη διερεύνηση της απόδοσης των ελέγχων ως προς την ισχύ οι Li and Yu (2015) παράγουν ελλιπή δεδομένα υπό τους μηχανισμούς MAR και NMAR. Από τα αποτελέσματα των προσομοιώσεών τους συμπεραίνουμε ότι:

- σε γενικές γραμμές ο έλεγχος που προτείνουν οι Li and Yu (2015), παρουσιάζει υψηλή ισχύ για όλες τις εναλλακτικές MAR και NMAR που παρουσιάζονται στην εργασία τους.
- Η ισχύς του ελέγχου των Li and Yu (2015) εξαρτάται σε μεγάλο βαθμό από τις διαφορές των κατανομών μεταξύ των υποομάδων, που δημιουργούνται με βάση τα μοτίβα ελλιπών τιμών. Επιπλέον, εξαρτάται από το μέγεθος του δείγματος και τα μεγέθη δείγματος των μοτίβων ελλιπών τιμών.
- Σε κάποιες περιπτώσεις που τα δεδομένα είναι NMAR ο έλεγχος των Jamshidian and Jalal (2010) παρουσιάζει υψηλότερη ισχύ συγκριτικά με τον έλεγχο των Li and Yu (2015). Κάτι τέτοιο συμβαίνει όταν τα μοτίβα ελλιπών δεδομένων έχουν διαφοροποιήσεις ως προς τους πίνακες διακυμάνσεων συνδιακυμάνσεων.
- Οι περισσότερες τιμές της ισχύος του ελέγχου των Jamshidian and Jalal (2010) δεν είναι μεγαλύτερες από τις τιμές του εμπειρικού επιπέδου σημαντικότητας, γεγονός που αποτελεί ισχυρό μειονέκτημα.
- Η ισχύς των δύο ελέγχων άλλοτε αυξάνει και άλλοτε όχι, όταν αυξάνει η διάσταση $p$.

**Συγκριτική μελέτη του Aleksić (2024)** Στη συνέχεια αντικείμενο παρουσίασης θα αποτελέσει η συγκριτική μελέτη που διεξήγαγε ο Aleksić (2024), με στόχο να συγκρίνει τον έλεγχό του με αυτόν που προτείνει ο Little (1988). Ανακαλώντας όσα έχουν αναφερθεί στην Ενότητα 3.2, γίνεται αντιληπτό ότι σε αυτήν τη μελέτη προσομοίωσης θα πρέπει για μία μεταβλητή τουλάχιστον να είναι διαθέσιμες όλες οι τιμές. Σε αυτό το πλαίσιο, 5000 σε πλήθος προσομοιωμένα σύνολα δεδομένων προέρχονται από $p$-διάστατους πληθυσμούς με $p = 3, 5$. Στην περίπτωση των τρισδιάστατων πληθυσμών ο Aleksić (2024) θεωρεί ότι μία μεταβλητή είναι πλήρως παρατηρούμενη και οι άλλες δύο περιέχουν ελλιπή δεδομένα, ενώ στην περίπτωση που $p = 5$ εξετάζει τόσο την περίπτωση που δύο μεταβλητές είναι πλήρως παρατηρούμενες όσο και την περίπτωση που τρεις εξ αυτών είναι πλήρως παρατηρούμενες. Επιπρόσθετα, τα αρχικά πλήρη δεδομένα παράγονται από την τυπική πολυδιάστατη κανονική κατανομή και από την Clayton copula με παράμετρο 1 και περιθώριες από την εκθετική κατανομή με παράμετρο 1 και την $\chi^2_4$. Επισημαίνεται ότι η Clayton copula με παράμετρο 1 χρησιμοποιείται με στόχο να παραχθεί συσχέτιση 1/3 με βάση τον συντελεστή του Kendall. Τέλος, τα ελλιπή δεδομένα υπό την υπόθεση MCAR δημιουργούνται μέσω της συνάρτησης `delete_MCAR()` του πακέτου "missMethods". Στο όρισμα "ds" αυτής της συνάρτησης δίνεται το σύνολο δεδομένων που θα ανήκει στην κλάση "matrix()" ή "data.frame()", στο όρισμα `cols_mis()` δίνονται ως τιμές εισόδου διανύσματα φυσικών αριθμών που ανήκουν στην κλάση "vector()", ενώ το όρισμα "p" δέχεται ως τιμές εισόδου διανύσματα πραγματικών αριθμών του συνόλου (0,1). Με βάση τις τιμές εισόδου στο όρισμα `cols_mis()` εξάγονται ως τιμές εξόδου ελλιπείς τιμές υπό τη μορφή "NA", στις αντίστοιχες στήλες με τους αριθμούς που εισάγαμε, ενώ το ποσοστό των ελλιπών τιμών στις στήλες που επιλέχθηκαν στο όρισμα `cols_mis()` καθορίζεται από το διάνυσμα που δηλώθηκε στο όρισμα `cols_mis()`. Πιο συγκεκριμένα, ο σχεδιασμός της προσομοίωσης του Aleksić (2024) θεωρεί τιμές για την πιθανότητα εμφάνισης ελλιπούς τιμής που κυμαίνονται από 0.03 έως 0.24 με βήμα 0.03.

Από τα αποτελέσματα που παρατίθενται στην εργασία του Aleksić (2024) συμπεραίνουμε ότι:

- ο έλεγχος του Aleksić (2024) διατηρεί το επίπεδο σημαντικότητας κοντά στο ονομαστικό επίπεδο σημαντικότητας ακόμη και για μικρό μέγεθος δείγματος, για κάθε περίπτωση.
- Από την άλλη, ο έλεγχος του Little (1988) διατηρεί το επίπεδο σημαντικότητας όταν τα δεδομένα έχουν παραχθεί από την πολυδιάστατη κανονική κατανομή, ενώ για δεδομένα που προέρχονται από τις υπόλοιπες κατανομές, το εμπειρικό επίπεδο σημαντικότητας φαίνεται να είναι σε γενικές γραμμές

διογκωμένο.

Όσον αφορά τη σύγκριση των δύο ελέγχων ως προς την ισχύ, ο Aleksić (2024) προσομοιώνει δεδομένα που είναι ελλιπή άλλα όχι MCAR χρησιμοποιώντας ειδικές για αυτόν τον σκοπό συναρτήσεις του πακέτου `missMethods` της γλώσσας R. Πιο συγκεκριμένα, χρησιμοποιείται η συνάρτηση "delete_MAR_1_to_x()" όπου στο όρισμα "x" εκχωρείται η τιμή 9, η συνάρτηση "delete_MAR_rank()" και η συνάρτηση "delete_MAR_mean()". Επισημαίνεται ότι η μελέτη προσομοίωσης της ισχύς γίνεται μόνο για εκείνα τα μεγέθη δείγματος όπου και οι δύο έλεγχοι διατηρούν το επίπεδο σημαντικότητας υπό τις κατανομές που παράγονται τα δεδομένα. Από τα αποτελέσματα που παρατίθενται στην εργασία του συμπεραίνουμε ότι:

- στην περίπτωση που έχουμε $n = 100$ το πλήθος πειραματικές μονάδες από την τυπική πολυδιάστατη κανονική κατανομή, ο έλεγχος του Aleksić (2024) έχει μεγαλύτερη ισχύ για κάθε διάσταση και όλα τα μοτίβα ελλιπών τιμών υπό την "MAR 1 to 9" εναλλακτική. Όπως αναμενόταν, για μεγαλύτερα σε μέγεθος δείγματα η ισχύς αυξάνει, αλλά η σχέση παραμένει αναλλοίωτη.

- Στην περίπτωση δειγματοληψίας από την Clayton copula υπό την "MAR 1 to 9" εναλλακτική σε γενικές γραμμές ο έλεγχος του Little (1988) παρουσιάζει υψηλότερη ισχύ συγκριτικά με αυτόν του Aleksić (2024). Όμως, γνωρίζοντας από τη μελέτη του εμπειρικού επιπέδου σημαντικότητας, ότι ο έλεγχος του Little παρουσιάζει αυξημένα σφάλματα τύπου I, η πραγματική διαφορά μεταξύ των δύο ελέγχων αναμένεται να είναι αρκετά μικρότερη.

- Στην περίπτωση δειγματοληψίας MAR rank για την περίπτωση της τυπικής πολυδιάστατης κανονικής κατανομής η ισχύς του ελέγχου του Aleksić (2024) παραμένει υψηλότερη, ενώ τα συμπεράσματα που αναφέρθηκαν προηγούμενα για την περίπτωση της δειγματοληψίας από την Clayton copula υπό την "MAR 1 to 9" εναλλακτική συνεχίζουν να ισχύουν.

- Στην περίπτωση που τα δεδομένα προέρχονται από τον τρίτο μηχανισμό έλλειψης τα αποτελέσματα του ελέγχου του Aleksić (2024) είναι καλύτερα ως προς την ισχύ.

- Τέλος, για μεγάλο μέγεθος δείγματος ($n = 500$) και οι δύο έλεγχοι έχουν πολύ μεγάλη ισχύ.

Από τα προηγούμενα γίνεται ακόμη μία φορά σαφές ότι ο έλεγχος του Little (1988) θα πρέπει να χρησιμοποιείται με προσοχή όταν έχουμε ενδείξεις ότι τα

δεδομένα δεν προέρχονται από πολυδιάστατη κανονική κατανομή, ενώ ο έλεγχος του Aleksić (2024) αποτελεί μία σημαντική προσθήκη στη βιβλιογραφία, υπό τις υποθέσεις που πρέπει να ικανοποιούνται.

**Συγκριτική μελέτη των Chen et al. (2023)** Στη συνέχεια, αντικείμενο μελέτης αποτελεί η συγκριτική μελέτη των Chen et al. (2023), όπου ο έλεγχός τους συγκρίνεται με τα αποτελέσματα του ελέγχου του Little (1988). Στη μελέτη αυτή, τα προσομοιωμένα σύνολα δεδομένων κατασκευάζονται έτσι ώστε τα δεδομένα να εξαρτώνται μεταξύ τους με βάση τη δομή "Block AR(1)". Αυτό σημαίνει ότι για την εκάστοτε πειραματική μονάδα, η πρώτη με τη δεύτερη συνιστώσα θα έχει αυτοσυσχέτιση $\rho$, η πρώτη με την τρίτη $\rho^2$ κ.ο.κ.. Τα δεδομένα παρήχθησαν υπό την πολυδιάστατη κανονική κατανομή για τιμές της παραμέτρου $\rho = 0, 0.35, 0.7$. Παράλληλα ο αριθμός των μεταβλητών επιλέχθηκε να είναι $p = 5, 10, 15$. Στο πλαίσιο αυτό για την αξιολόγηση της απόδοσης του ελέγχου ως προς τη διατήρηση του επιπέδου σημαντικότητας παράγουν MAR δεδομένα με βάση μοντέλο λογιστικής παλινδρόμησης κατά τέτοιον τρόπο ώστε περίπου το 40% των υπό μελέτη μεταβλητών-στηλών να είναι πλήρως παρατηρούμενες. Ο αριθμός των πειραματικών μονάδων για τα διάφορα σύνολα δεδομένων στα οποία εφαρμόστηκε ο έλεγχος, είναι τέτοιος ώστε $n/p = \{20, 26, 32\}$ και η εκτίμηση της πιθανότητας σφάλματος τύπου I προκύπτει από 1000 επαναλήψεις της διαδικασίας. Από τα αποτελέσματα που παρατίθενται στην εργασία τους συμπεραίνουμε ότι ο έλεγχος των Chen et al. (2023) διατηρεί το ονομαστικό επίπεδο σημαντικότητας για τις διαφορετικές περιπτώσεις του συντελεστή συσχέτισης $\rho$. Για $n = 100,\ p = 5$ και $\rho = 0.35$ το προσομοιωμένο επίπεδο σημαντικότητας είναι 6%, ενώ για $n = 320,\ p = 15$ και ξανά $\rho = 0.35$ είναι 5%, όπου και στις 2 περιπτώσεις έχει εφαρμοστεί η μέθοδος "FDR", όπως περιγράφηκε στο Κεφάλαιο 2.

Στη συνέχεια, θέλοντας να εξετάσουν την απόδοση του ελέγχου ως προς την ισχύ και να την συγκρίνουν με τα αποτελέσματα του ελέγχου του Little (1988) παράγουν δεδομένα από την πολυδιάστατη κανονική κατανομή που περιγράφηκε παραπάνω για $\rho = 0$ και από την πολυδιάστατη $t$ με 2.2 βαθμούς ελευθερίας. Οι έλεγχοι εφαρμόστηκαν και στις δύο κατανομές για μεγέθη δείγματος $n = 100$ και $n = 200$, δημιουργώντας δεδομένα που ήταν MAR και NMAR, στη βάση ενός μοντέλου λογιστικής παλινδρόμησης. Σημειώνουμε ότι τα ελλιπή δεδομένα τόσο υπό τη μηδενική υπόθεση όσο και υπό την εναλλακτική δημιουργούνται με το μοντέλο λογιστικής παλινδρόμησης:

$$logit(P_r(r_j = 0)) = \beta_{0j} + \sum \beta_{kj} Y_k. \tag{4.1}$$

Ειδικότερα, ο μηχανισμός MCAR προκύπτει για $\beta_{kj},\ k \geq 1,\ \forall j$, ο MAR για

$\beta_{kj} \neq 0$ με τους δείκτες $j$ και $k$ να αναφέρονται στις μεταβλητές που δεν είναι πλήρως παρατηρούμενες στο σύνολο δεδομένων. Τέλος, ο μηχανισμός NMAR δημιουργείται για $\beta_{jj} \neq 0$ με τον δείκτη $j$ να αναφέρεται μόνο στις μεταβλητές που εμφανίζονται ελλιπή δεδομένα και $\beta_{lj} \neq 0$, για τουλάχιστον έναν δείκτη $l$ που αναφέρεται στις πλήρως παρατηρούμενες μεταβλητές. Στη μελέτη προσομοίωσης θεώρησαν ότι όλες οι παράμετροι $\beta_{kj}$ είναι ίσες μεταξύ τους με τιμές να κυμαίνονται έως 1.25.

Από τα αποτελέσματα που δίνονται στην εργασία των Chen et al. (2023) συμπεραίνουμε ότι ο έλεγχος που προτείνουν οι Chen et al. (2023), είναι πιο ισχυρός από αυτόν που προτάθηκε από τον Little (1988), για τους μηχανισμούς MAR. Σε κάθε περίπτωση ως προς την ισχύ ο έλεγχος των Chen et al. (2023) είναι καλύτερος από του Little (1988) τόσο όταν τα δεδομένα είναι εξαρτημένα όσο και όταν είναι ανεξάρτητα. Επιπλέον, η ισχύς του ελέγχου των Chen et al. (2023) αυξάνεται όταν αυξάνεται το μέγεθος της αυτοσυσχέτισης και το μέγεθος του δείγματος.

**Συγκριτική μελέτη των Rouzinov and Berchtold (2022)** Η συγκριτική μελέτη των Rouzinov and Berchtold (2022) περιλαμβάνει τον έλεγχο που προτάθηκε σε αυτήν την εργασία καθώς και τους ελέγχους των Little (1988) και Jamshidian and Jalal (2010). Ειδικότερα, στη μελέτη τους οι Rouzinov and Berchtold (2022) θεωρούν προσομοιωμένα δεδομένα, μεγέθους $n = 1000$, που προέρχονται από την πολυδιάστατη κανονική κατανομή και την πολυδιάστατη ομοιόμορφη. Θέλοντας να αξιολογήσουν την απόδοση των ελέγχων ως προς το επίπεδο σημαντικότητας προσομοιώνουν MCAR δεδομένα, ενώ για την αξιολόγηση της ισχύς προσομοιώνουν MAR δεδομένα τεσσάρων μορφών (αναφέρονται ως MAR1-MAR4), με τον κώδικα υλοποίησης να δίνεται στο άρθρο τους. Το ποσοστό των ελλιπών τιμών στο σύνολο δεδομένων κυμαίνεται από 1%-50%.

Στην περίπτωση που τα δεδομένα είναι ανεξάρτητα από την πολυδιάστατη ομοιόμορφη κατανομή προκύπτει ότι:

- όλοι οι έλεγχοι διατηρούν το επίπεδο σημαντικότητας κοντά στο 5%, όταν τα ελλιπή δεδομένα έχουν παραχθεί υπό τον μηχανισμό MCAR, για σχεδόν όλα τα ποσοστά ελλιπών τιμών.
- Ο έλεγχος του Little (1988) έχει πολύ μεγάλη ισχύ για όλους τους μηχανισμούς και όλα σχεδόν τα ποσοστά ελλιπών δεδομένων.
- Ο έλεγχος των Jamshidian and Jalal (2010) δεν μπορεί να απορρίψει τη μηδενική υπόθεση όταν υπάρχει είτε μεγάλο ποσοστό ελλιπών δεδομένων

ή αρκετά μικρό. Ειδικά για τα μεγάλα ποσοστά, η συμπεριφορά του ελέγχου, δεν είναι αναμενόμενη και ενδέχεται να οφείλεται στην εφαρμογή ταυτόχρονα του παραμετρικού και μη παραμετρικού ελέγχου, οδηγώντας σε αύξηση των σφαλμάτων τύπου I λόγω συσχέτισης και άρα σε χαμηλή ισχύ. Ακόμη, αιτία του φαινομένου, ενδέχεται να είναι η απόδοση του ελέγχο του Hawkins (1981), πάνω στον οποίο βασίζεται ο έλεγχος και το γεγονός ότι αυτός αδυνατεί να απορρίψει τη μηδενική υπόθεση για μικρό μέγεθος δείγματος.

- Η ισχύς του ελέγχου των Rouzinov and Berchtold (2022) αυξάνεται όσο το ποσοστό των ελλιπών τιμών αυξάνεται, με το επίπεδο ισχύος να είναι σε γενικές γραμμές ικανοποιητικό. Ωστόσο, η απόδοσή του δεν είναι καλύτερη αυτής του τεστ του Little (1988).

Στην περίπτωση που τα δεδομένα είναι περιγράφονται ικανοποιητικά από την πολυδιάστατη κανονική κατανομή προκύπτει ότι:

- όλοι οι έλεγχοι φαίνεται να διατηρούν το ονομαστικό επίπεδο σημαντικότητας για όλα τα ποσοστά ελλιπών τιμών.

- Ο έλεγχος του Little (1988) επιδεικνύει πολύ υψηλή ισχύ για τις περισσότερες εναλλακτικές υποθέσεις και ποσοστά ελλιπών τιμών.

- Τα αποτελέσματα που αφορούν τον έλεγχο των Rouzinov and Berchtold (2022) είναι παρόμοια με αυτά που πρωτύτερα σχολιάσθηκαν.

- Η ισχύς του ελέγχου των Jamshidian and Jalal (2010) είναι εξαιρετικά χαμηλή σε δύο περιπτώσεις μηχανισμού ελλιπών δεδομένων για ποσοστά ελλιπών δεδομένων πολύ μικρά ή πολύ μεγάλα.

- Όλοι οι έλεγχοι παρουσιάζουν χαμηλά επίπεδα ισχύος για έναν συγκεκριμένο μηχανισμό έλλειψης (MAR4). Για την εναλλακτική αυτή ο έλεγχος των Jamshidian and Jalal (2010) έχει εξαιρετικά υψηλή ισχύ, μόνο όμως για ποσοστά ελλιπών δεδομένων που κυμαίνονται στο 2%-30%. Σημειώνεται ότι ο μηχανισμός αυτός παράγει τα ελλιπή δεδομένα με βάση την αλληλεπίδραση μεταξύ δύο επεξηγηματικών μεταβλητών, χωρίς όμως η αλληλεπίδραση να περιλαμβάνεται ως επιπλέον μεταβλητή, με αποτέλεσμα το ποσοστό των ελλιπών δεδομένων να μην μπορεί να αιτιολογηθεί με βάση κάποια/ες μεταβλητές, οδηγώντας έτσι στη μη απόρριψη της υπόθεσης MCAR.

**Συνοψίζοντας**, συνδυάζοντας τα αποτελέσματα των προηγούμενων συγκριτικών μελετών, καθώς και τα αποτελέσματα που παρατέθηκαν στις Ενότητες 2.2 και 2.4, συμπεραίνουμε όσα ακολουθούν.

- Δεν υπάρχει μοναδικός έλεγχος που να μπορεί να προταθεί.
- Ο έλεγχος των Jamshidian and Jalal (2010) δεν προτείνεται, ειδικά, σε περιπτώσεις με λίγες πειραματικές μονάδες και πολλές μεταβλητές, καθώς δεν διατηρεί το επίπεδο σημαντικότητας, ενώ επηρεάζεται από τον τρόπο δημιουργίας των ελλιπών δεδομένων.
- Ο μη παραμετρικός έλεγχος των Jamshidian and Jalal (2010) είναι πιο αξιόπιστος από τον αντίστοιχο παραμετρικό.
- Ο έλεγχος του Little (1988) για μικρό μέγεθος δείγματος είναι συντηρητικός, διατηρεί τις περισσότερες φορές το επίπεδο σημαντικότητας, και προτείνεται να χρησιμοποιείται κυρίως όταν τα δεδομένα προέρχονται από πολυδιάστατο κανονικό πληθυσμό.
- Ο έλεγχος των Jamshidian and Yuan (2014) έχει πολύ καλή απόδοση όταν τα δεδομένα είναι MAR, ωστόσο δεν έχει το ίδιο καλή απόδοση όταν τα δεδομένα είναι NMAR.
- Ο έλεγχος που προτείνουν οι Li and Yu (2015) παρουσιάζει υψηλή ισχύ για κάποιες εναλλακτικές MAR και NMAR που παρουσιάζονται στην εργασία τους για $n > 500$.
- Ο έλεγχος των Chen et al. (2023) έχει το πλεονέκτημα ότι μπορεί να εφαρμοσθεί και σε μεικτού τύπου δεδομένα, ενώ ταυτόχρονα φαίνεται να έχει καλύτερη απόδοση συγκριτικά με αυτήν του ελέγχου του Little (1988) για MAR και NMAR δεδομένα.
- Τέλος, ο έλεγχος του Aleksić (2024) αποτελεί σημαντική προσθήκη, ωστόσο για να χρησιμοποιηθεί απαιτεί την ύπαρξη μίας τουλάχιστον μεταβλητής με πλήρη δεδομένα.

Θα ήταν παράλειψη να μην αναφερθεί ότι σε όσα προηγήθηκαν παρουσιάστηκαν τα αποτελέσματα συγκριτικών μελετών στις οποίες συμμετείχαν οι έλεγχοι που παρουσιάστηκαν στα προηγούμενα δύο κεφάλαια αυτής της διατριβής, όπως εμφανίζονται στον Πίνακα 4.1. Από τον πίνακα αυτό γίνεται σαφές ότι παραμένει ανοικτό θέμα προς διερεύνηση η συγκριτική μελέτη όλων των ελέγχων

που παρατέθηκαν. Επιπλέον, επισημαίνεται ότι σε όλες τις εργασίες δεν χρησιμοποιήθηκαν ούτε οι ίδιες κατανομές για τη δημιουργία των δεδομένων ούτε οι ίδιοι μηχανισμοί δημιουργίας MCAR, MAR και NMAR δεδομένων. Θα ήταν επιθυμητό να συγκριθούν οι μέθοδοι σε μία μελέτη προσομοίωσης για όλους τους μηχανισμούς που έχουν εμφανισθεί στη βιβλιογραφία.

| | Συγκριτική μελέτη | | | | |
|---|---|---|---|---|---|
| Έλεγχος | Jamshidian & Yuan | Li & Yu | Aleksic | Chen et al. | Rouzinov & Berchtold |
| Little (1988) | ✓ | | ✓ | ✓ | ✓ |
| Jamshidian & Jalal (2010) | ✓ | ✓ | | | ✓ |
| Jamshidian & Yuan (2014) | ✓ | | | | |
| Li & Yu (2015) | | ✓ | | | |
| Aleksic (2024) | | | ✓ | | |
| Chen et al. (2023) | | | | ✓ | |
| Rouzinov & Berchtold (2022) | | | | | ✓ |

Πίνακας 4.1: Συγκρίσεις ελέγχων για την υπόθεση MCAR.

# ΚΕΦΑΛΑΙΟ 5

# Επιλογος

Σε πολλές μελέτες οι ερευνητές έρχονται, συχνά, αντιμέτωποι με την ανάλυση συνόλων δεδομένων που δεν είναι πλήρη. Ο Rubin (1976) θέλοντας να εξηγήσει τον μηχανισμό εμφάνισης των ελλειπουσών τιμών και να δώσει απάντηση στο ερώτημα αν το γεγονός ότι κάποιες μεταβλητές εμφανίζουν ελλιπείς τιμές σχετίζεται με την τιμή των μεταβλητών αυτών στο σύνολο των δεδομένων περιέγραψε τρεις μηχανισμούς που οδηγούν σε ελλιπή δεδομένα. Μεταξύ αυτών, για λόγους που παρουσιάστηκαν στο πρώτο κεφάλαιο αυτής της μεταπτυχιακής διατριβής, ξεχωριστή θέση κατέχουν τα MCAR δεδομένα και για αυτόν τον λόγο είναι απαραίτητο όταν υπάρχουν ελλιπή δεδομένα να επιβεβαιώνεται ότι αυτά μπορούν να θεωρηθούν ότι έχουν προέλθει από έναν τέτοιον μηχανισμό. Για τον λόγο αυτό στη βιβλιογραφία, ιδιαίτερα τα τελευταία έτη, παρατηρείται ιδιαίτερο ενδιαφέρον για την εισαγωγή μεθόδων για τον έλεγχο αν ο μηχανισμός έλλειψης των δεδομένων είναι MCAR ή όχι. Σε αυτό το πλαίσιο, οι έλεγχοι που αναλύθηκαν εκτενώς στα Κεφάλαια 2 και 3, προσφέρουν ένα κατάλληλο στατιστικό εργαλείο στη μελέτη του ερευνητικού μας προβλήματος. Επισημαίνεται ότι οι έτοιμες συναρτήσεις της R και όσες δημιουργήθηκαν σε αυτήν τη μεταπτυχιακή διατριβή επιτρέπουν την υλοποίηση των ελέγχων σε πρακτικές εφαρμογές. Τέλος, όπως αναδείχθηκε στο Κεφάλαιο 4 της μεταπτυχιακής διατριβής παραμένει ανοικτό θέμα η υπολογιστική συγκριτική μελέτη όλων των προτεινόμενων μεθοδολογιών.

Κλείνοντας θα ήταν παράλειψη να μην αναφέρουμε ότι στη διατριβή αυτή το ενδιαφέρον επικεντρώθηκε σε ελέγχους της υπόθεσης MCAR για συνεχή δεδομένα. Ωστόσο, στη βιβλιογραφία έχουν εμφανισθεί έλεγχοι και για άλλου τύπου δεδομένα. Πιο συγκεκριμένα, ο Fuchs (1982) πρότεινε έναν τρόπο ελέγχου της MCAR υπόθεσης για διακριτού τύπου δεδομένα. Ουσιαστικά αποτελεί έναν έλεγχο ομοιογένειας για τις μέσες τιμές, κατά πλήρη αναλογία με τους ελέγχους που περιγράφηκαν στο δεύτερο κεφάλαιο. Ο έλεγχος του Fuchs (1982), ενδέχεται να έχει επηρεάσει σε μεγάλο βαθμό το άρθρο του Little (1988), καθώς

τόσο η μεθοδολογία όσο και ο τρόπος παρουσίασης φαίνεται να ακολουθείται πιστά από τον τελευταίο συγγραφέα. Ο έλεγχος υλοποιείται στη γλώσσα R, στο πακέτο `MCARtest`. Για την περίπτωση των διακριτών δεδομένων παρουσιάστηκε πρόσφατα στη βιβλιογραφία και ο έλεγχος των Berrett and Samworth (2023), ο οποίος υλοποιείται στη γλώσσα R μέσω του πακέτου `MCARtest`. Η κατασκευή και η θεωρία του ελέγχου βασίζεται, κυρίως, στα περιθώρια πολύτοπα (marginal polytopes), τα οποία αποτελούν αντικείμενο μελέτης της Κυρτής Γεωμετρίας. Στόχος του ελέγχου είναι να χαρακτηρίσει όλες τις ανιχνεύσιμες εναλλακτικές υποθέσεις, για τη μηδενική MCAR, παρέχοντας έτσι ελέγχους που επιδεικνύουν ασυμπτωτική ισχύ 100%, ενώ παράλληλα διατηρούν το σφάλμα τύπου Ι. Η προσπάθειά τους χαρακτηρίζεται επιτυχής, καθώς αποδεικνύουν, ότι για τον έλεγχο της υπόθεσης MCAR σε διακριτά δεδομένα, είναι ανάγκη να ελεγχθεί η υπόθεση της συμβατότητας, που ορίζεται στο άρθρο του Kellerer (1984). Επιπλέον, οι Berrett and Samworth (2023) διεξήγαγαν μία συγκριτική μελέτη μεταξύ του ελέγχου που πρότειναν και αυτού που αναφέρεται στο άρθρο του Fuchs (1982), με τα αποτελέσματα να δείχνουν, ότι ο έλεγχος που προτείνουν είναι πιο ισχυρός και διατηρεί το επίπεδο σημαντικότητας για $\alpha = 5\%$. Στην περίπτωση τώρα που τα δεδομένα μας αποτελούν επαναληπτικές μετρήσεις παραπέμπουμε τον/την ενδιαφερόμενο αναγνώστη/στρια στους ελέγχους που έχουν προταθεί στις εργασίες των Diggle (1989), Ridout and Diggle (1991) και Listing and Schlittgen (1998), για συνεχή δεδομένα, και στην εργασία των Park and Davis (1993) για κατηγορικά δεδομένα. Για τον έλεγχο της υπόθεσης MCAR σε διαχρονικά δεδομένα (longitudinal data) παραπέμπουμε στις εργασίες των Park and Lee (1997) και Chassan and Concordet (2023).

# ΠΑΡΑΡΤΗΜΑ Α΄

# Συναρτησεισ στην R

Στο παρόν παράρτημα παρατίθενται οι απαραίτητες συναρτήσεις στη γλώσσα προγραμματισμού R για την υλοποίηση των ελέγχων του Aleksić (2024), των Chen et al. (2023), των Li and Yu (2015) και των Jamshidian and Yuan (2014), αντίστοιχα. Επισημαίνεται ότι η συνάρτηση που αφορά στον έλεγχο των Jamshidian and Yuan (2014) βασίζεται στη συνάρτηση που έχει δοθεί στην εργασία τους με το όνομα `NPV test`.

```
library(expm)
library(MVN)

Aleksic_test<-function(data,alpha)
{
  fullmat<-t(na.omit(t(data)))
mismat<-data[, !(names(data) %in% colnames(fullmat))]
  indmat<-matrix(1-as.numeric(is.na(mismat)),ncol = ncol(
     ↪ mismat))
  n<-nrow(data);u<-ncol(fullmat);v<-ncol(mismat)
  #We do this for faster computation than using for loops
  #We subtract the elements with step n to create the
     ↪ first part of equation (9) of Aleksic's paper
  d<-(1/(n*(n-1)))*colSums(kronecker(fullmat,indmat)[-(seq
     ↪ (from=1,to=n*n,by=n)),])

  #now the second part
  c<-(1/n)*t(fullmat)%*%indmat;c1<-as.vector(t(c))
  #equation (9)
  Tn<-(d-c1)
  Tn_new<-Tn*sqrt(n)
  #equation (11) this is kronecker product. Aleksic just
     ↪ doesnt mention it
  S<-kronecker(var(fullmat),var(indmat))

    S1<-solve(S)
```

```
  #Results
    An<-sum((sqrtm(S1)%*%Tn_new)^2)
    cat(paste0("Statistic:"),An,"\n")
    cat(paste0("Critical Value:"),qchisq(p=1-alpha,df=u*v)
       ↪ ,"\n")
    cat(paste0("Statistic>Critical Value:"), An>qchisq(p
       ↪ =1-alpha,df=u*v),"\n")
    cat(paste0("p-value:"),pchisq(q= An,df=u*v,lower.tail=
       ↪ FALSE),"\n")

}
```

```
library(kSamples)
library(stats)
testChengChungBasu<-function(data)
{

X<-as.data.frame(data)



  class_variable<-function(x) {
  if (class(x)=="factor")
  {
    return("Categorical")
  }
   else if (length(unique(x)) < 10) {
      return("Discrete")
    }
    else {
      return("Continuous")
    }
  }
  variable_types<-sapply(X,class_variable)

  categorical_cols<-which(variable_types == "Categorical")
  discrete_cols<-which(variable_types == "Discrete")
  continuous_cols<-which(variable_types == "Continuous")

  #p-values
  p_continuous<-c()
  names_continuous<-c()
  p_discrete<-c()
  names_discrete<-c()
  p_categorical<-c()
```

```
  names_categorical<-c()



#this works only if group>2 and number of observation in
   ↪ each group>2 for ad.test. Otherwise it skips the
   ↪ calculation
 #Continuous
 for (col_idx in continuous_cols) {
   y_var<-X[, col_idx]
   missing_pattern<-apply(X, 1, function(row) paste(is.na
      ↪ (row), collapse = "_"))
   groups<-split(y_var, missing_pattern)

   valid_groups<-lapply(groups, function(g) na.omit(as.
      ↪ numeric(g)))
   valid_groups<-Filter(function(x) length(x) > 1,
      ↪ valid_groups)

   if (length(valid_groups) > 1) {
     result<-do.call(ad.test, valid_groups)
     p_val<-result$ad[1,3]
     p_continuous<-c(p_continuous, p_val)
     names_continuous<-c(names_continuous,names(X)[
        ↪ col_idx])
   }
 }

 #Discrete
 for (col_idx in discrete_cols) {
   y_var<-X[, col_idx]
   missing_pattern<-apply(X, 1, function(row) paste(is.na
      ↪ (row), collapse = "_"))
   groups<-split(y_var, missing_pattern)

   valid_groups<-lapply(groups, function(g) na.omit(as.
      ↪ numeric(g)))
   valid_groups<-Filter(function(x) length(x) > 1,
      ↪ valid_groups)

   if (length(valid_groups) > 1) {
     result<-do.call(ad.test, valid_groups)
     p_val<-result$ad[1,3]
     p_discrete<-c(p_discrete, p_val)
     names_discrete<-c(names_discrete,names(X)[col_idx])
```

```
    }
  }

  #Categoricals. Creation of contingency table and apply X
     ↪ ^2 of pearson
  for (col_idx in categorical_cols) {
    y_var<-X[, col_idx]
    missing_pattern<-apply(X, 1, function(row) paste(is.na
       ↪ (row), collapse = "_"))
    contingency_data<-data.frame(Var = y_var, Pattern =
       ↪ missing_pattern)
    contingency_data<-contingency_data[!is.na(
       ↪ contingency_data$Var), ]

    table_cat<-table(contingency_data$Var,
       ↪ contingency_data$Pattern)
    if (nrow(table_cat) > 1 && ncol(table_cat) > 1) {
      test_result<-chisq.test(table_cat)
      p_val<-test_result$p.value
      p_categorical<-c(p_categorical, p_val)
      names_categorical<-c(names_categorical,names(X)[
         ↪ col_idx])
    }
  }


  if (length(p_continuous) > 0) {
    adj_p_cont<-p.adjust(p_continuous, method = "fdr")
  }


  if (length(p_discrete) > 0) {
    adj_p_dis<-p.adjust(p_discrete, method = "fdr")
  }


  if (length(p_categorical) > 0) {
    adj_p_cat<-p.adjust(p_categorical, method = "fdr")
  }

print_pvalues <- function(names_vec, pvalues, var_type) {
  cat(paste("\n", var_type, "Variables:\n", sep = ""))

  for (i in seq_along(names_vec)) {
    var_name <- names_vec[i]
```

```
    if (!is.null(pvalues) && i <= length(pvalues)) {
      cat(paste0(var_name, ": ",pvalues[i], "\n"))
    } else {
      cat(paste0(var_name, ": NULL\n"))
    }
  }
}
print_pvalues(names_continuous, if (exists("adj_p_cont"))
    ↪ adj_p_cont else NULL, "Continuous")
print_pvalues(names_discrete, if (exists("adj_p_dis"))
    ↪ adj_p_dis else NULL, "Discrete")
print_pvalues(names_categorical, if (exists("adj_p_cat"))
    ↪ adj_p_cat else NULL, "Categorical")


}
testChengChungBasu(data = Y)
```

```
library(MASS)
library(dplyr)
library(proxy)


Li_Yu_test<-function(nboot,alpha,df)
  {
  Q<-c()
  B<-c()
  W<-c()

 d<-function(A,B,n_total) {
    A<-as.matrix(A)
    B<-as.matrix(B)

    nA<-nrow(A)
    nB<-nrow(B)

    if (nA==0 || nB==0) return(d_val=0)

    else{
      dists<-proxy::dist(A,B)
      g_ab<-mean(dists)
      distsA<-as.matrix(proxy::dist(A))
      distsB<-as.matrix(proxy::dist(B))
```

```
    if (nA>=2){
      g_aa<-mean(distsA)
              }
     else
       { g_aa<-0}
    if (nB>=2){
      g_bb<-mean(distsB)
              }
     else
       {g_bb<-0}
     d_val<-max(2*g_ab-g_aa-g_bb,0)*((nA*nB)/(2*n_total))
     return(d_val)
   }
 }


g<-function(A) {
   A<-as.matrix(A)
   nA<-nrow(A)
   dists_A<-as.matrix(proxy::dist(A))
  if (nA >= 2) {
    g_aa<-mean(dists_A)*(nA/2)
              }
   else {
     g_aa<-0
         }
   return(g_aa)
 }

 get_missing_pattern<-function(row) {
   paste(ifelse(is.na(row),1,0), collapse = "")
 }
 missing_patterns<-apply(Y,1,get_missing_pattern)
 df<-as.data.frame(Y)%>%mutate(missing_pattern=
    ↪ missing_patterns)
 groups<-split(df,df$missing_pattern)
 groups<-lapply(groups,function(x) x[,-ncol(x)])

 observed_cols_per_group<-lapply(groups,function(x) {
   complete_vars<-colSums(!is.na(x))==nrow(x)
   names(complete_vars[complete_vars])
 })

 group_names<-names(observed_cols_per_group)
 n_groups<-length(group_names)
```

```
  common_vars_pairs<-list()

  for (i in 1:(n_groups-1)) {
    for (j in (i+1):n_groups) {
      group_i<-group_names[i]
      group_j<-group_names[j]
      common_vars<-intersect(observed_cols_per_group[[
         ↪ group_i]],observed_cols_per_group[[group_j]])
      if (length(common_vars)>0) {
        pair_name<-paste0("Group_",group_i,"-",group_j)
        common_vars_pairs[[pair_name]]<-common_vars

      }
      else
      {
        cat(paste0("No common variables found between
           ↪ groups"))
      }
    }
  }
  new_groups_from_pairs<-list()
  for (pair_name in names(common_vars_pairs)) {
    common_vars<-common_vars_pairs[[pair_name]]
    split_names<-strsplit(pair_name,"-")[[1]]
    group_name_i<-sub("Group_","",split_names[1])
    group_name_j<-split_names[2]

    group_i_data<-groups[[group_name_i]][,common_vars,drop
       ↪ =FALSE]
    group_j_data<-groups[[group_name_j]][,common_vars,drop
       ↪ =FALSE]

    new_groups_from_pairs[[paste0(group_name_i,"
       ↪ _subset_for_",pair_name)]]<-group_i_data
    new_groups_from_pairs[[paste0(group_name_j,"
       ↪ _subset_for_",pair_name)]]<-group_j_data
  }

n_total<-nrow(Y)

  d_result<-list()
  pair_ids<-unique(sub(".*_subset_for_(.*)","\\1", names(
     ↪ new_groups_from_pairs)))
  for (pair_id in pair_ids) {
```

```
    subset_names<-grep(paste0("_subset_for_",pair_id, "$")
       ↪ , names(new_groups_from_pairs),value=TRUE)

    if (length(subset_names) == 2) {
      group1_data<-new_groups_from_pairs[[subset_names
         ↪ [1]]]
      group2_data<-new_groups_from_pairs[[subset_names
         ↪ [2]]]

      d_result[[pair_id]]<-d(A=group1_data,B=group2_data,
         ↪ n_total)
    }
    else
    {
      cat(paste0("Not enough groups to compare"))
    }
  }

  B<-sum(unlist(d_result))

   cleaned_groups<-lapply(groups, function(x) {
    cleaned_x<-t(na.omit(t(x)))
    attributes(cleaned_x)$na.action<-NULL
    class(cleaned_x)<-NULL
    return(cleaned_x)
  })

  W<-sum(unlist(lapply(cleaned_groups,g)))
  Q<-B*(n_total-n_groups)/(W*(n_groups-1))

  Qboot<-numeric(nboot)
  Wboot<-numeric(nboot)
  Bboot<-numeric(nboot)


for(B in 1:nboot)
  {
    df_complete<-na.omit(Y)
    bootstrap_indices<-sample(1:nrow(df_complete),size=
       ↪ nrow(Y), replace=TRUE)
    df_boot<-df_complete[bootstrap_indices, ]
    df_boot[is.na(Y)]<-NA
    missing_patterns<-apply(df_boot, 1,
       ↪ get_missing_pattern)
    boot_df<-as.data.frame(df_boot) %>% mutate(
```

```
    ↪ missing_pattern=missing_patterns)
boot_groups<-split(boot_df, boot_df$missing_pattern)
boot_groups<-lapply(boot_groups, function(x) x[, -ncol
    ↪ (x)])
n_boot_groups<-length(boot_groups)

observed_cols_per_boot_group<-lapply(boot_groups,
    ↪ function(x) {
  complete_vars <- colSums(!is.na(x)) == nrow(x)
  names(complete_vars[complete_vars])
})


boot_group_names<-names(observed_cols_per_boot_group)
common_boot_vars_pairs<-list()

for (i in 1:(n_boot_groups-1)) {
  for (j in (i + 1):n_boot_groups) {
    group_i<-boot_group_names[i]
    group_j<-boot_group_names[j]
    common_vars<-intersect(
      ↪ observed_cols_per_boot_group[[group_i]],
      ↪ observed_cols_per_boot_group[[group_j]])
    if (length(common_vars)>0) {
      pair_name<-paste0("bootGroup_", group_i, "-",
        ↪ group_j)
      common_boot_vars_pairs[[pair_name]]<-common_vars
    }
  }
}


  new_boot_groups_from_pairs<-list()
for (pair_name in names(common_boot_vars_pairs))
  {
  common_vars<-common_boot_vars_pairs[[pair_name]]
  split_names<-strsplit(pair_name,"-")[[1]]
  group_name_i<-sub("bootGroup_","", split_names[1])
  group_name_j<-split_names[2]

  group_i_data<-boot_groups[[group_name_i]][,
    ↪ common_vars,drop=FALSE]
  group_j_data<-boot_groups[[group_name_j]][,
    ↪ common_vars,drop=FALSE]
```

```
      new_boot_groups_from_pairs[[paste0(group_name_i, "
         ↪ _subset_for_", pair_name)]]<-group_i_data
      new_boot_groups_from_pairs[[paste0(group_name_j, "
         ↪ _subset_for_", pair_name)]]<-group_j_data
      }
    d_boot_result<-list()
    pair_ids<-unique(sub(".*_subset_for_(.*)", "\\1",
       ↪ names(new_boot_groups_from_pairs)))

    for (pair_id in pair_ids) {
      subset_names<-grep(paste0("_subset_for_", pair_id, "
         ↪ $"),names(new_boot_groups_from_pairs),value=
         ↪ TRUE)
      if (length(subset_names)==2) {
        group1_data<-new_boot_groups_from_pairs[[
           ↪ subset_names[1]]]
        group2_data<-new_boot_groups_from_pairs[[
           ↪ subset_names[2]]]
        d_boot_result[[pair_id]]<-d(group1_data,
           ↪ group2_data,n_total)
      }
    }

    Bboot[B]<-sum(unlist(d_boot_result))

    cleaned_boot_groups<-lapply(boot_groups, function(x) {
      cleaned_x<-t(na.omit(t(x)))
      attributes(cleaned_x)$na.action<-NULL
      class(cleaned_x)<-NULL
      cleaned_x
    })

    Wboot[B]<-sum(unlist(lapply(cleaned_boot_groups,g)))
    Qboot[B]<-Bboot[B]*(n_total-n_boot_groups)/(Wboot[B]*(
       ↪ n_boot_groups-1))


  }
  c_a<-quantile(Qboot,probs=1-alpha)

 cat(paste0(("Critical Value:")),c_a,"\n")
 cat(paste0("Statistic:"),Q,"\n")
 cat(paste0(("Reject MCAR Hypothesis:")),Q>c_a,"\n")

}
```

```
 ADtest<- function (ymiss) {
  library(MissMech)
  p=dim(ymiss)[2]
  yord=OrderMissing(ymiss,del.lesscases=0)
  ymiss=yord$data #ordered data
  patused=yord$patused #observed data patterns
  ngroups=yord$g
  ind_varcomp=which(colSums(patused,na.rm=TRUE)==ngroups)
  onetop=1:p
  ind1=onetop
  if(length(ind_varcomp)!=0){
    ind1=onetop[-ind_varcomp]
  }
  cnt=0;
  pvals=matrix(0,length(ind1)*(p-1),3)
  for (j1 in ind1){
    indo=which(!is.na(ymiss[,j1]))
    indm=which(is.na(ymiss[,j1]))
    ind2=onetop[-j1]
    for (j2 in ind2){
      yo=ymiss[indo,j2]
      yo=yo[!is.na(yo)]
      ym=ymiss[indm,j2]
      ym=ym[!is.na(ym)]
      #not part of the orginal code#
      if (length(yo) < 2 || length(ym) < 2) next
      ########################################
      AD=AndersonDarling(data=c(yo,ym),number.cases=c(
         ↪ length(yo),length(ym)))
      cnt=cnt+1
      pvals[cnt,1]=j1
      pvals[cnt,2]=j2
      pvals[cnt,3]=AD$pn
    }
  }
  #not part of the original code
  pvals<-pvals[1:cnt, ,drop = FALSE]
  ################################
return(pvals)
}

BenHoch<- function (pvals) {
  alpha=0.05
```

```
  pvals1=sort(pvals)
  k=length(pvals1)
  J=seq(1:k)
  L=J*alpha/(k*sum(1/J))
  ind=which(pvals1<=L)
  if (length(ind)==0){cutoff =0}
  else {r=max(ind); cutoff=pvals1[r]}
  reject=pvals<=cutoff
  list(cutoff=cutoff,reject=reject)
}



#not part of the original code
ADtest(ymiss)
BenHoch(ADtest(ymiss)[,3])
ADtest(ymiss)[,3]
####
```

# Βιβλιογραφια


Aleksić, D. (2024). A novel test of missing completely at random: U-statistics-based approach. *Statistics*, 58(4):1004–1023.

Aleksić, D. (2025). A generalization of a U-statistics-based MCAR test: utilizing partially observed variables. ArXiv, 2501.05596.

Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Bentler, P., Kim, K., and Yuan, K. (2004). Testing homogeneity of covariances with infrequent missing data patterns. *Unpublished Manuscript*.

Berrett, T. and Samworth, R. (2023). Optimal nonparametric testing of missing completely at random and its connections to compatibility. *The Annals of Statistics*, 51:2170–2193.

Chassan, M. and Concordet, D. (2023). How to test the missing data mechanism in a hidden markov model. *Computational Statistics & Data Analysis*, 182:107723.

Chen, R., Chung, Y.-C., Basu, S., and Shi, Q. (2023). Diagnostic test for realized missingness in mixed-type data. *Sankhya B*, 86:109–138.

David, F. N. (1939). On Neyman's "smooth" test for goodness of fit: I. distribution of the criterion $\psi 2$ when the hypothesis tested is true. *Biometrika*, 31(1-2):191–199.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Diggle, P. J. (1989). Testing for random dropouts in repeated measurement data. *Biometrics*, 45:1255–1258.

Dixon, W. J. and Brown, M. B. (1983). *BMDP statistical software manual: to accompany the 1988 software release*. Berkeley: University of California Press.

Efron, B. and Tibshirani, R. (1994). *An Introduction to the Bootstrap (1st ed.)*. Chapman and Hall/CRC.

Fisher, R. A. (1932). *Statistical methods for research workers*. Oliver and Boyd, Edinburgh.

Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data. *Journal of the American Statistical Association*, 77(378):270–278.

Hawkins, D. M. (1981). A new test for multivariate normality and homoscedasticity. *Technometrics*, 23(1):105–110.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19:308–334.

Jamshidian, M. and Jalal, S. (2010). Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data. *Psychometrika*, 75(4):649–674.

Jamshidian, M. and Mata, M. (2008). Postmodeling sensitivity analysis to detect the effect of missing data mechanisms. *Multivariate Behavioral Research*, 43:432–452.

Jamshidian, M. and Yuan, K.-H. (2013). Data-driven sensitivity analysis to detect missing data mechanism with applications to structural equation modelling. *Journal of Statistical Computation and Simulation*, 83(7):1344–1362.

Jamshidian, M. and Yuan, K.-H. (2014). Examining missing data mechanisms via homogeneity of parameters, homogeneity of distributions, and multivariate normality. *WIREs Computational Statistics*, 6(1):56–73.

Jennrich, R. I. (1970). An asymptotic $\chi 2$ test for the equality of two correlation matrices. *Journal of the American Statistical Association*, 65:904–912.

Kellerer, H. (1984). Duality theorems for marginal problems. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 67:399–432.

Kim, K. and Bentler, P. (2002). Tests of homogeneity of means and covariance matrices for multivariate incomplete data. *Psychometrika*, 67:609–623.

Lee, S.-Y. and Tsui, K.-L. (1982). Covariance structure analysis in several populations. *Psychometrika*, 47(3):297–308.

Li, J. and Yu, Y. (2015). A nonparametric test of missing completely at random for incomplete multivariate data. *Psychometrika*, 80(3):707–726.

Listing, J. and Schlittgen, R. (1998). Tests if dropouts are missed at random. *Biometrical Journal*, 40(8):929–935.

Little, R. and Rubin, D. (2002). *Statistical Analysis with Missing Data, 2nd Edition.* Wiley Series in Probability and Statistics. Wiley, New York.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202.

Marozzi, M. (2009). Some notes on the location–scale Cucconi test. *Journal of Nonparametric Statistics*, 21(5):629–647.

Marozzi, M. (2014). The multisample Cucconi test. *Stat Methods Appl*, 23:209–227.

Mealli, F. and Rubin, D. B. (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika*, 102(4):995–1000.

Nagao, H. (1973). On some test criteria for covariance matrix. *The Annals of Statistics*, 1(4):700 – 709.

Neyman, J. (1937). Smooth test for goodness of fit. *Scandinavian Actuarial Journal*, 1937(3-4):149–199.

Noé, M. (1972). The calculation of distributions of two-sided Kolmogorov-Smirnov type statistics. *The Annals of Mathematical Statistics*, 43(1):58–64.

Park, T. and Davis, C. S. (1993). A test of the missing data mechanism for repeated categorical data. *Biometrics*, 49(2):631–638.

Park, T. and Lee, S.-Y. (1997). A test of missing completely at random for longitudinal data with missing observations. *Statistics in Medicine*, 16(16):1859–1871.

Ridout, M. S. and Diggle, P. J. (1991). Testing for random dropouts in repeated measurement data. *Biometrics*, 47(4):1617–1621.

Rizzo, M. L. and Székely, G. J. (2010). Disco analysis: a nonparametric extension of analysis of variance. *The Annals of Applied Statistics*, 4(2):1034–1055.

Rouzinov, S. and Berchtold, A. (2022). Regression-based approach to test missing data mechanisms. *Data, MDPI*, 7:1–28.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Scholz, F. W. and Stephens, M. A. (1987). K-sample Anderson-Darling tests. *Journal of the American Statistical Association*, 82(399):918–924.

Srivastava, M. S. and Dolatabadi, M. (2009). Multiple imputation and other resampling schemes for imputing missing observations. *Journal of Multivariate Analysis*, 100(9):1919–1937.

Tang, M.-L. and Bentler, P. M. (1998). Theory and method for constrained estimation in structural equation models with incomplete data. *Computational Statistics and Data Analysis*, 27(3):257–270.

van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press, New York.

Van Buuren, S. and Oudshoorn, K. (1999). *Flexible multivariate imputation by MICE*. Leiden: TNO.

Wand, M. and Jones, M. (1994). *Kernel Smoothing (1st ed.)*. Chapman and Hall/CRC.