ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΜΕΛΕΤΗ ΤΟΥ ΠΑΡΑΔΟΞΟΥ BERKSON ΣΤΗ ΒΙΟΣΤΑΤΙΣΤΙΚΗ

Δημήτριος Κυριάκου

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Μάϊος 2023

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΜΕΛΕΤΗ ΤΟΥ ΠΑΡΑΔΟΞΟΥ BERKSON ΣΤΗ ΒΙΟΣΤΑΤΙΣΤΙΚΗ

Δημήτριος Κυριάκου

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Μάϊος 2023

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ……. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Γ. Τζαβελάς, Αναπληρωτής Καθηγητής (Επιβλέπων)
- Χ. Ευαγγελάρας, Αναπληρωτής Καθηγητής
- Κ. Πολίτης, Αναπληρωτής Καθηγητής

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

A STUDY OF BERSKON'S PARADOX IN BIOSTATISTICS

By

Dimitrios Kyriakou

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
May 2023

Στους γονείς μου

Κυριάκο και Αναστασία

Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε στα πλαίσια του μεταπτυχιακού προγράμματος σπουδών «Εφαρμοσμένη Στατιστική» του τμήματος Στατιστικής κι Ασφαλιστικής επιστήμης υπό την επίβλεψη του αναπληρωτή καθηγητή Γεώργιου Τζαβελά τον οποίο ευχαριστώ θερμά για την αποδοχή και τη στήριξη που μου παρείχε κατά τη διάρκεια της συγγραφής. Επίσης, θα ήθελα να ευχαριστήσω τους Κυρίους Χ.Ευαγγελάρα και Κ.Πολίτη για τη συμμετοχή τους στην τριμελή επιτροπή. Τέλος, ένα μεγάλο «ευχαριστώ» στους γονείς μου.

Περίληψη

Η παρούσα διπλωματική έχει ως σκοπό τη μελέτη ενός φαινομένου που είναι γνωστό στη βιβλιογραφία ως παράδοξο ή μεροληψία του Berkson και αποτελεί ένα είδος μεροληψίας λόγω κακής δειγματοληψίας. Εμφανίζεται συχνά σε μελέτες ομάδων (cohort studies) σε δείγματα που λαμβάνονται σε μονάδες υγείας και έχει σαν αποτέλεσμα η ομάδα μελέτης από το νοσοκομείο να εμφανίζει υψηλότερο ποσοστό έκθεσης στο ρίσκο ή ποσοστό ασθένειας σε σύγκριση με τον γενικό πληθυσμό λόγω υπερεκπροσώπευσης. Θα γίνει βιβλιογραφική ανασκόπηση για το παράδοξο του Berkson ενώ θα δοθεί ιδιαίτερη μνεία σε διάφορες μελέτες που πραγματεύθηκαν το φαινόμενο αυτό. Τέλος, θα γίνει μελέτη πραγματικών δεδομένων που παρουσιάζουν μεροληψία Berkson μέσω των σταθμισμένων κατανομών, των οποίων η θεωρία θα μελετηθεί και θα παρουσιαστεί αναλυτικά.

Abstract

This dissertation aims to study a phenomenon known in the literature as Berkson's paradox or bias and is a type of bias due to sampling errors. It often occurs in cohort studies, when samples should be taken in health units and results in the study group from the hospital to show a higher rate of exposure to risk or disease rate compared to the general population due to overrepresentation. A literature review of Berkson's paradox will be made, while special emphasis will be given to various studies that dealt with this phenomenon. Finally, real data showing Berkson bias will be studied with the help of weighted distributions, whose theory will be presented in detail.

Περιεχόμενα

ΛΙΣΤΑ ΠΙΝΑΚΩΝ

ΛΙΣΤΑ ΣΧΗΜΑΤΩΝ

ΚΕΦΑΛΑΙΟ 1

Το παραδοξο του Berkson

1.1 Εισαγωγή

Στη φύση της στατιστικής επιστήμης, οι τρόποι και οι συνθήκες επιλογής των ατόμων ή των μονάδων που συμμετέχουν σε ένα είδος δειγματοληψίας έχουν κρίσιμο χαρακτήρα. Συγκεκριμένα, η περίπτωση της λανθασμένης επιλογής των συνθηκών ή και του τρόπου δειγματοληψίας μπορεί να οδηγήσει σε σφάλματα μέσω των οποίων να εξάγουμε αποτελέσματα τα οποία δε αντικατοπτρίζουν την πραγματική εικόνα που έχει το υπό μελέτη χαρακτηριστικό μέσα στο πλαίσιο του γενικού πληθυσμού.

Γενικά, η επιλογή ενός μη κατάλληλου τρόπου δειγματοληψίας αναφέρεται στην περίπτωση της μη τυχαίας επιλογής των περιπτώσεων. Δηλαδή, οι μεταβλητές απαντήσεων που εκφράζουν το αποτέλεσμα της εκάστοτε επιστημονικής έρευνας να μην προέρχονται από ένα γενικό πληθυσμό αναφοράς, αλλά από κάποια υποομάδα του. Η τελευταία διαθέτει κάποια χαρακτηριστικά τα οποία μπορεί να οδηγήσουν σε μεγαλύτερα σφάλματα και επομένως σε μη έγκυρη συμπερασματολογία.

Σε πολλές περιπτώσεις δειγματοληψίας παρατηρείται το φαινόμενο συσχετίσεων μεταξύ δύο μεταβλητών οι οποίες θεωρητικά είναι ασυσχέτιστες (ή και ανεξάρτητες) είτε αντίθετα συσχετισμένες. Η φύση του προβλήματος, θεωρητικά έγκειται στο ότι η δεσμευμένη πιθανότητα ενός ενδεχομένου δεδομένου ότι έχει πραγματοποιηθεί το ενδεχόμενο της ένωσης του με ένα άλλο ενδεχόμενο, ανεξάρτητο προς αυτό, είναι διαφορετική από την πιθανότητα πραγματοποίησης του σε ένα γενικό πλαίσιο. Τέτοια φαινόμενα

μπορούν να παρατηρηθούν στην επιλογή ενδεχομένων με δυσανάλογες συχνότητες.

Όσον αφορά το μαθηματικό φορμαλισμό του προβλήματος, έχουμε την ακόλουθη περίπτωση: Θεωρούμε δύο ενδεχόμενα A και B τα οποία είναι ανεξάρτητα. Θεωρούμε και ένα ενδεχόμενο C. Υπάρχει η περίπτωση αυτό το ενδεχόμενο να "καταστρέψει" την ανεξαρτησία των A και B στο γενικό πληθυσμό. Δηλαδή, τα A και B να είναι δεσμευμένα ανεξάρτητα ως προς ένα ενδεχόμενο C.

Για παράδειγμα, θεωρούμε τον ακόλουθο πίνακα συχνοτήτων των δύο ενδεχομένων A και B. Με βάση τα παρακάτω στοιχεία, μπορούμε να παρατηρήσουμε εύκολα ότι τα δύο ενδεχόμενα είναι ανεξάρτητα.

Ενδεχόμενο	B	$\overline{B}$
A	500	500
$\overline{A}$	500	500

Table 1.1: Παρατηρήσεις δύο ανεξαρτήτων ενδεχομένων

Αν όμως λάβουμε υπόψη ένα ενδεχόμενο C, τότε αντί του Πίνακα 1.1 θα πάρουμε τα στοιχεία που δίνονται στον Πίνακα 1.2. Παρατηρούμε τα παρακάτω:

- Η πιθανότητα πραγματοποίησης του ενδεχομένου A δεδομένου του C είναι 60
- Η πιθανότητα πραγματοποίησης του ενδεχομένου A δεδομένης της πραγματοποίησης της τομής των C και B είναι 46.66

Αυτό σημαίνει ότι με βάση τα παραπάνω πως τα ενδεχόμενα A και B παρότι είναι ανεξάρτητα με βάση το γενικό πληθυσμό (Πίνακας 1), στην περίπτωση που εισαχθεί ένα άλλο ενδεχόμενο C είναι δεσμευμένα εξαρτημένα. Επιπλέον,

Ενδεχόμενα	C		$\overline{C}$		
	B	$\overline{B}$	B	$\overline{B}$	Σύνολο
A	350	400	150	100	1000
$\overline{A}$	400	100	100	400	1000
Σύνολο	750	500	250	500	2000

Table 1.2: Παρατηρήσεις δύο ανεξαρτήτων ενδεχομένων

παρατηρούμε ότι η πιθανότητα μειώνεται και επομένως συμπεραίνουμε ότι με τα δύο ενδεχόμενα είναι αρνητικά συσχετισμένα δεσμεύοντας ως προς ένα άλλο γεγονός. Ακόμη, φαινόμενα αρνητικής συσχέτισης μέσω δέσμευσης μπορεί να εντοπιστεί και σε ενδεχόμενα τα οποία μεταξύ τους είναι θετικά συσχετισμένα.

Το συμπέρασμα που προκύπτει είναι ότι πρέπει να δοθεί ιδιαίτερη προσοχή στις ιδιότητες των πληθυσμών που μελετάμε. Τέτοιου είδους προβλήματα εντοπίζονται σε πολλές στατιστικές έρευνες.

Στο παρακάτω διάγραμμα διασποράς απεικονίζονται δύο τυχαίες μεταβλητές X και Υ (πλήρες δείγμα). Παρατηρούμε ότι δεν υπάρχει γραμμική τάση μεταξύ αυτών των δύο καθώς "διασκορπίζονται" σε κάθε πιθανή θέση και επομένως είναι ασυσχέτιστες.

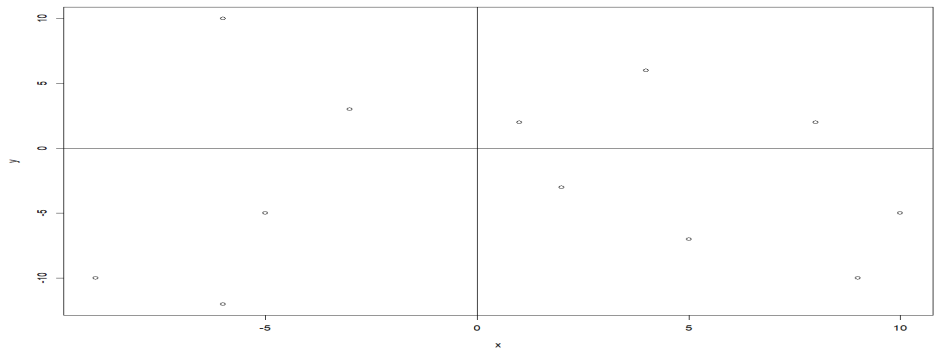

Figure 1.1: Ασυσχέτιστες Παρατηρήσεις

Αν κάνουμε περικοπή ορισμένων δεδομένων από το παραπάνω δείγμα και διατηρήσουμε μόνο τις περιπτώσεις που οι X και Υ δε μπορούν να είναι ταυτόχρονα αρνητικές (δηλαδή, να εξαφανισουμε τις τιμές που απεικονίζονται στο τρίτο τεταρτημόριο του γραφήματος), το διάγραμμα διασποράς, με την εμφάνιση της γραμμής παλινδρόμησης, θα γίνει

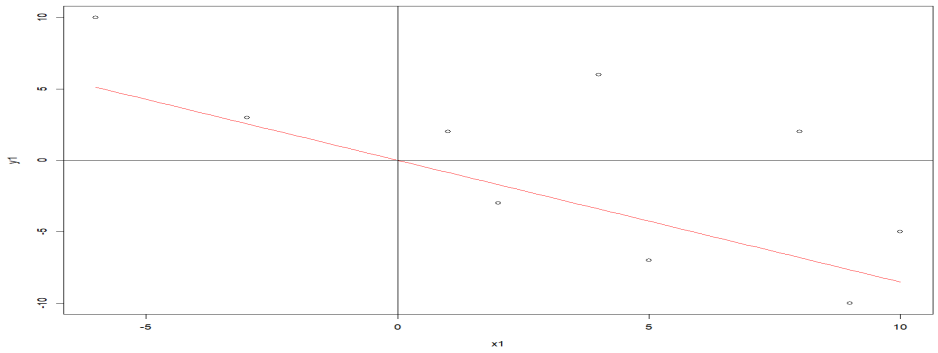

Figure 1.2: Λανθάνουσα συσχέτιση μέσω περικοπής

Παρατηρούμε ότι τα υπάρχοντα δεδομένα δίνουν δύο μεταβλητές αρνητικά συσχετισμένες ενώ η αντίστοιχη γραμμή παλινδρόμησης δείχνει να είναι καλύτερα προσαρμοσμένη σε αυτά. Ουσιαστικά, περιορίσαμε τον τρόπο επιλογής τιμών αυτών των τυχαίων μεταβλητών με αποτέλεσμα να φαίνονται συσχετισμένες ενώ στην πραγματικότητα ήταν ανεξάρτητες με αποτέλεσμα να οδηγηθούμε σε σφάλμα δειγματοληψίας.

1.2 Το παράδοξο

Μία τέτοια κατάσταση δειγματοληψίας εντοπίστηκε από τον J. (1946) η οποία αναφέρεται στην εργασία με τίτλο “Limitations of the Application of Fourfold Table Analysis to Hospital Data”. Στη συγκεκριμένη μελέτη, ο σκοπός του συγγραφέα ήταν να τονίσει την περίπτωση δύο ασθενειών οι οποίες στο γενικό πληθυσμό είναι ανεξάρτητες αλλά μπορούν να γίνουν “πλασμαστικά” συσχετισμένες, αν το δείγμα περιοριστεί σε μία διαδικασια ανάλυσης ασθενών μαρτύρων χρησιμοποιώντας τους ασθενείς ενός νοσοκομείου.

Η συγκεκριμένη “πλασματική” συσχέτιση μεταξύ δύο ανεξάρτητων (στο γενικό πληθυσμό) μεταβλητών αποκαλείται “Παράδοξο του Berkson”, “Μεροληψία του Berkson” ή “σφάλμα του Berkson”. Ο ίδιος ο συγγραφέας προτίμησε την πρώτη ονομασία για το φαινόμενο που μελέτησε.

Συγκεκριμένα, το πρόβλημα το οποίο τέθηκε ήταν το εξής: Έστω ότι ένα νοσοκομειακό ίδρυμα σκοπεύει να πραγματοποιήσει την εκτίμηση μεταξύ των περιστατικών με σακχαρώδη διαβήτη και εκείνων με χολοκυστίτιδα.

Επισημαίνεται, ότι το παράδειγμα έγινε μέσω θεωρητικών τιμών και δε βασίστηκε σε πραγματικά δεδομένα.

Για να διεξαχθεί η εν λόγω έρευνα, χρησιμοποιήθηκε μία ανάλυση ασθενών-μαρτύρων στην οποία νοσηλευόμενοι αναφέρονται ως ασθενείς όταν νοσούν από σακχαρώδη διαβήτη και ως μάρτυρες όσοι έχουν διαθλαστικό σφάλμα. Η συσχέτιση του διαβήτη και της χολοκυστίδας εκτιμάται συγκρίνοντας τον επιπολασμό της χολοκυστίτιδας μεταξύ ασθενών με διαβήτη και μαρτύρων με διαθλαστικό σφάλμα.

Το παράδειγμα που δόθηκε από τον J.Berkson βασίστηκε στην υπόθεση ότι η χολοκυστίτιδα είναι ανεξάρτητη τόσο από το σακχαρώδη διαβήτη όσο και από το θλαστικό σφάλμα στα πλαίσια του πληθυσμού. Όσον αφορά τους νοσηλευόμενους, το ζεύγος των μεταβλητών Χολοκυστίδα-Διαβήτης θεωρείται συσχετισμένο.

Αρχικά, ξεκίνησε με τον παρακάτω πίνακα βάσει του οποιου ο έλεγχος χ^2 έδωσε p-value=0.037<0.05. Στα δεδομένα εξετάζεται η συσχέτιση μεταξύ των δύο ασθενειών στην περίπτωση που ο υπό εξέταση πληθυσμός αναφέρεται στους νοσηλευόμενους. Από τον παραπάνω πίνακα έχουμε:

Ασθένεια	Χολοκυστίτιδα	¬Χολοκυστίτιδα	Σύνολο
Διαβήτης	28	548	576
¬Διαβήτης	1326	39036	40362
Σύνολο	1354	39584	40938

Table 1.3: Κατανομή με βάση την ασθένεια

- Το 4.86% των ατόμων με διαβήτη έχει χολοκυστίτιδα
- Από τα άτομα που δεν έχουν διαβήτη το 3.28% δεν έχει χολοκυστίτιδα

Παρατηρούμε μία διαφορά μεταξύ των δύο ποσοστών, κάτι το οποίο επιβεβαιώνεται και από τον έλεγχο χ^2.

1.3 Μη εμπειρικά Παραδείγματα

Στην παρούσα παράγραφο θα παραθέσουμε μερικά μη εμπειρικά παραδείγματα που ανέπτυξε ο Berkson (1946) για να υποστηρίξει το

επιχείρημα του περί δειγματοληπτικού σφάλματος. Ένας λόγος που μπορεί να εντοπιστεί αυτή η αλλαγή στη σχέση μεταξύ των μεταβλητών οφείλεται στον τρόπο που έγινε η δειγματοληψία. Για παράδειγμα, είναι λογική αυτή η μεταβολή στον υποπληθυσμό όταν το ποσοστό των ατόμων με την ασθένεια Α που νοσηλεύονται να είναι υψηλή. Στην περίπτωση αυτή, ο Berkson έδωσε την παρακάτω αριθμητική περίπτωση: Θεωρούμε έναν πληθυσμό με μέγεθος N=1000000 στον οποίο η πιθανότητα εμφάνισης διαβητη να είναι p_d=0.01, η πιθανότητα εμφάνισης χολοκυστίτιδας να είναι p_c=0.03 και η πιθανότητα εμφάνισης θλαστικού σφάλματος να είναι p_r=0.1. Ακόμη με q_d, q_c και q_r συμβολίζουμε τις αντίστοιχες συμπληρωματικές πιθανότητες ενώ θεωρούμε ότι οι ασθένειες αυτές στον πληθυσμό είναι ανεξάρτητες.

Η διαδικασία εξαγωγής συμπερασμάτων είναι η εξής: Έστω ότι σε έναν πληθυσμό μελετάμε τις ασθένειες i,j και k με πιθανότητες εμφάνισης p_i, p_j και p_k αντίστοιχα και q_i, q_j και q_k τα επιμέρους συμπληρωματικά τους. Τότε, στο σύνολο του πληθυσμού θα ορίσουμε:

- Έστω N_i το πλήθος των ατόμων που έχουν την ασθένεια i στο σύνολο του πληθυσμού: $N_i = p_i \cdot N$,
- Έστω N_{ij} το σύνολο των ατόμων που εμφανίζουν τις ασθένειες i,j αλλά όχι την k: $N_{ij} = p_i \cdot p_j \cdot q_k N$
- Έστω N_{ijk} το σύνολο των ατόμων που εμφανίζουν και τις 3 ασθένειες: $N_{ijk} = p_i \cdot p_j \cdot p_k N$
- Έστω N_0 το σύνολο των ατόμων που δεν εμφανίζουν καμμία από τις 3 ασθένειες: $N_0 = q_i \cdot q_j \cdot q_k N$.

Στον παρακάτω πίνακα, με εφαρμογή των παραπάνω τύπων υπολογίζουμε το μέγεθος για τον εκάστοτε υποπληθυσμό:

Ασθένεια	Ποσοστό	Μεγεθος
Μόνο d	$p_d \cdot q_c \cdot q_r$	87300
Μόνο r	$q_d \cdot q_c \cdot p_r$	267300
Μόνο c	$q_d \cdot p_c \cdot q_r$	960300
d και c	$p_d \cdot p_c \cdot q_r$	29700
d και r	$p_d \cdot q_c \cdot p_r$	9700
c και r	$q_d \cdot p_c \cdot p_r$	2970
όλες	$p_d \cdot p_c \cdot p_r$	300
καμία	$q_d \cdot q_c \cdot q_r$	8642700

Table 1.4: Κατανομή των ασθενειών στον πληθυσμό

θα ελέγξουμε κατά πόσο ή όχι εντοπίζόνται τυχόν διαφορές των ασθενών με χολοκυστίτιδα που αντιμετωπίζουν κάποιο επιπλέον πρόβλημα διαβήτη ή θλαστικού σφάλματος στο πλαίσιο του γενικού πληθυσμούς. Ως ομάδα μαρτύρων βάζουμε εκείνους που έχουν θλαστικό σφάλμα αλλά όχι διαβήτη. Όσοι έχουν διαβήτη και θλαστικό σφάλμα θα τοποθετηθούν στην ομάδα των ατόμων με διαβήτη. Ο πίνακας που δίνεται παρακάτω κάνει σύγκριση τους πληθυσμούς που νοσούν από χολοκυστίτιδα ή διαβήτη.

- Τα άτομα με διαβήτη και χολοκυστίτιδα θα είναι: 2700+300=3000
- Τα ατομα με θλαστικό σφάλμα και χολοκυστίτιδα είναι Nrc=29700
- Τα άτομα με διαβήτη που δεν έχουν χολοκυστίτιδα θα είναι 87300+9700=97000
- Τα άτομα με θλαστικό σφάλμα και δεν εχουν χολοκυστίτιδα Nr=960300

Λαμβάνοντας τα παραπάνω, ο πίνακας συνάφειας γίνεται:

Ασθένεια	Χολοκυστίτιδα	¬Χολοκυστίτιδα	Σύνολο
Διαβήτης	3000	9700	10000
Θλαστικό σφάλμα	29700	960300	990000
Σύνολο	30000	1057300	1090000

Table 1.5: Κατανομή με βάση την ασθένεια

Επομένως, τα αποτελέσματα που λαμβάνουμε, είναι:

- Το ποσοστό των ατόμων με χολοκυστίτιδα στα άτομα με διαβήτη είναι $\frac{3000}{100000} = 3\%$.
- Το ποσοστό των ατόμων που δεν έχουν χολοκυστίτιδα στα άτομα με διαβήτη είναι $\frac{3000}{100000} = 3\%$.

Παρατηρούμε, ότι δεν υπάρχει κάποια ουσιαστική διαφορά μεταξύ των δύο ομάδων στα ποσοστά. Αν για παράδειγμα χρησιμοποιήσουμε τον έλεγχο χ^2 θα πάρουμε ότι p_{value}= 0.99224>0.05 και επομένως δεν εντοπίζονται στατιστικά σημαντικές διαφορές.

Θα πραγματοποιήσουμε την αντίστοιχη ανάλυση περιοριζόμενοι στον υποπληθυσμό των νοσηλευόμενων εξετάζοντας κατά πόσο ή όχι οι επαγόμενες ομάδες εμφανίζουν διαφορές. Οι τύποι που θα χρησιμοποιηθούν για τη συγκεκριμένη διαδικασία είναι:

- Έστω n_i, s_i το πλήθος των νοσηλευόμενων με την ασθένεια και το ποσοστό τους μέσα στον πληθυσμό αυτών με πλήθος N_i, τότε: $n_i = s_i \cdot N_i$.
- Έστω $t.$ το ποσοστό των ατόμων που έχουν ακριβώς μία ασθένεια αλλά δε νοσηλεύονται. Π.χ $t_i = 1s_i$
- Έστω n_{ij} το πλήθος των νοσηλευόμενων με τις ασθένειες i και j. Τότε $n_{ij} = (1 - t_i \cdot t_j) \cdot N_{ij}$
- Έστω $n_{ijκ}$ των νοσηλευόμενων που φέρουν και τις 3 ασθένειες, τότε: $n_{ijk} = (1 - t_i \cdot t_j \cdot t_k) \cdot N_{ijk}$

Μπορούμε να αναλύσουμε την περίπτωση συσχέτισης των ασθενειών χρησιμοποιώντας τον πληθυσμό των νοσηλευομένων. Αρχικά, θα υποθέσουμε ότι τα ποσοστά των ατόμων που έχουν ακριβώς μία ασθένεια θα είναι στο 5%, δηλαδή:

- Έστω s_d , s_c, $s_r = 0.05$ τα ποσοστά των ατόμων που έχουν ακριβώς μία ασθένεια και t_c, t_d, $t_r = 0.95$ τα αντίστοιχα συμληρωματικά τους
- Έστω $s_{d\ c}$, s_{dr}, s_{cr} τα ποσοστά των ατόμων στον πληθυσμό που έχουν ακριβώς δύο ασθένειες και νοσηλεύονται
- Έστω s_{dcr} το ποσοστό των νοσηλευόμενων που έχουν ακριβώς και τις 3 ασθένειες.

Από τα παραπάνω έχουμε:

- $s_{dc} = 1 - t_d \cdot t_c = 1 - 0,95^2 = 0.0975$ και $s_{dr} = s_{cr} = 0.0975$
- $s_{dcr} = 1 - 0.95^3 = 0.142625$.

Συνοψίζουμε τα αποτελέσματα στον ακόλουθο πίνακα

Ασθένεια	Ποσοστό	Μεγεθος
Μόνο d	0.05	4365
Μόνο r	0.05	13365
Μόνο c	0.05	48015
d και c	0.0975	263
d και r	0.0975	946
c και r	0.0975	2896
όλες	0.142625	43

Table 1.6: Κατανομή των νοσηλευομένων με βάση την ασθένεια

Τώρα, θα γίνει ο έλεγχος για το αν εντοπίζεται κάποια διαφορά στην ομάδα νοσηλευομένων με χολοκυστίτιδα και αντιμετωπίζουν κάποιο επιπλέον πρόβλημα διαβήτη ή θλαστικού σφάλματος. Ως ομάδα μαρτύρων βάζουμε εκείνους που έχουν θλαστικό σφάλμα αλλά όχι διαβήτη. Όσοι έχουν διαβήτη και θλαστικό σφάλμα θα τοποθετηθούν στην ομάδα των ατόμων με διαβήτη. Ο πίνακας **??**κάνει σύγκριση τους πληθυσμούς που νοσούν από χολοκυστίτιδα ή διαβήτη και νοσηλεύονται.

Αρχικά, ορίζουμε τις παρακάτω τιμές για κάθε υποπληθυσμό

- Το σύνολο των ατόμων που έχουν διαβήτη αλλά όχι χολοκυστίτιδα είναι $n_d = 4365 + 946 = 5311$
- Το σύνολο των ατόμων που έχουν διαβήτη και χολοκυστίτιδα είναι $n_{dc} = 23 + 263 = 306$
- Το σύνολο των ατόμων που έχουν θλαστικό σφάλμα και χολοκυστίτιδα είναι $n_{rc} = 2896$
- Το σύνολο των ατόμων που έχουν θλαστικό σφάλμα αλλά όχι χολοκυστίτιδα είναι $n_r = 48015$

Table 1.7: Κατανομή νοσηλευόμενων

Ασθένεια	Χολοκυστίτιδα	¬Χολοκυστίτιδα	Σύνολο
Διαβήτης	306	5311	5617
Θλαστικό σφάλμα	2896	48015	50911
Σύνολο	3202	53326	56528

Ασθένεια	Ποσοστό	Μεγεθος
Μόνο d	0.05	4365
Μόνο r	0.15	40095
Μόνο c	0.2	192060
d και c	0.1925	520
d και r	0.24	2328
c και r	0.32	9504
όλες	0.354	106

Table 1.8: Ποσοστά νοσηλευομένων ανά ασθένεια

Και έχουμε τα εξής:

- Το ποσοστό των ατόμων με χολοκυστίτιδα στο σύνολο των ατόμων με διαβήτη είναι ίσο με $\frac{306}{5617} = 5.45\%$
- Το ποσοστό των ατόμων με χολοκυστίτιδα στο σύνολο των ατόμων με θλαστικό σφάλμα είναι ίσο με $\frac{2896}{50911} = 5.69\%$

Αν πραγματοποιήσουμε τον έλεγχο χ^2, θα μας δώσει $p-value$=0.45>0.05 και επομένως αποδεχόμαστε την περίπτωση ανεξαρτησίας των δεδομένων. Άρα, εδώ βλέπουμε ότι η ανεξαρτησία στο σύνολο του πληθυσμού παρατηρείται και στο υποσύνολο των νοσηλευομένων. Αυτό συνέβη λόγω της "δίκαιης" αναλογίας των νοσηλειών με βάση την εκάστοτε ασθενεια.

Παρακάτω, θα δοθεί ένα αντίστοιχο πρόβλημα αλλά θα αλλάξουμε τα ποσοστά των νοσηλευομένων με βάση την ασθένεια. Θεωρούμε ότι s_d=0.05, s_c=0.15 και s_r=0.2.

- Το σύνολο των ατόμων που έχουν διαβήτη αλλά όχι χολοκυστίτιδα είναι $n_d = 4365 + 2328 = 6693$

- Το σύνολο των ατόμων που έχουν διαβήτη και χολοκυστίτιδα είναι $n_{dc} = 520 + 106 = 626$
- Το σύνολο των ατόμων που έχουν θλαστικό σφάλμα και χολοκυστίτιδα είναι $n_{cr} = 9504$
- Το σύνολο των ατόμων που έχουν θλαστικό σφάλμα αλλά όχι χολοκυστίτιδα είναι $n_r = 192060$.

Με βάση τις προηγούμενες τιμές κατασκευάζουμε τον πίνακα συνάφειας που δινεται ακολούθως: Τα Ποσοστά των ατόμων με χολοκυστίτιδα:

Ασθένεια	Χολοκυστίτιδα	¬Χολοκυστίτιδα	Σύνολο
Διαβήτης	626	6693	7319
Θλαστικό σφάλμα	9504	192060	201564
Σύνολο	10130	198753	208883

Table 1.9: Κατανομή νοσηλευόμενων

- Στο σύνολο των ατόμων με διαβήτη είναι $\frac{626}{7319} = 8.55\%$
- Ενώ στο σύνολο των ατόμων με θλαστικό σφάλμα είναι $\frac{9504}{201564} = 4.72\%$

Παρατηρούμε ότι η υπάρχει διαφορά στα ποσοστά των δύο ομάδων. Κάτι το οποίο επιβεβαιώνεται και από τον έλεγχο χ^2 το οποίο δίνει $p_{value}<0.00001$ και επομένως η διαφορά είναι στατιστικά σημαντική.

Βλέπουμε ουσιαστικά τη μεταβολή των συμπερασμάτων ανάλογα με τα ποσοστά νοσηλείας των ασθενών. Είναι λογική η περίπτωση διαφοράς των ποσοστών των επί μέρους ομάδων διότι το ποσοστό των ατόμων με θλαστικό σφάλμα που θα νοσηλευτούν είναι πολύ μεγαλύτερο σε σχέση με το αντίστοιχο των ατόμων που πάσχουν από διαβήτη. Μάλιστα, η αλλαγή του συμπεράσματος σε σχέση με το προηγούμενο παράδειγμα ήταν σημαντική λόγω των ίσων ποσοστών. Άρα, το συμπέρασμα που προκύπτει από τον τελευταίο πίνακα είναι μεροληπτικό καθώς έχουμε μεγάλο πλήθος ατόμων στη συγκεκριμένη ομάδα.

ΚΕΦΑΛΑΙΟ 2

Μελετες του Παραδοξου

2.1 Βιβλιογραφική ανασκόπηση

Στο παρόν κεφάλαιο θα γίνει αναφορά σε διάφορες αναλύσεις και μελέτες εργασιών που πραγματεύθηκαν το παράδοξο του Berkson. Οι συγκεκριμένες προσεγγίσεις είτε είχαν εμπειρικό είτε μαθηματικό χαρακτήρα απόδειξης. Ορισμένες από αυτές που ειχαν εμπειρικό τρόπο ήταν οι εργασίες των N. (1977), Roberts et al. (1978), C. A.(1979), M. A.(1982).

Το 1954, ο A. έγραψε ότι η μεροληψία του Berkson υπήρχε μόνο για συσχετίσεις μεταξύ ασθενειών και όχι για τη συσχέτιση έκθεσης-ασθένειας. Αυτή η γνώμη, η οποία υποστηρίχθηκε από τους S.(1980), J.(1982), O.(2002) παρέβλεψε την έμμεση μορφή της πλάνης του Berkson και της διάκρισης μεταξύ επικρατών και περιστατικών συνθηκών, οι οποίες είναι βασικές για την κατανόηση του συλλογισμού του Berkson.

Ιδιαίτερη μνεία θα δοθεί στην εργασιά των Roberts et al.(1978) η οποία αναλύει τις θεωρητικές αιτίες της μεροληψίας του Berkson ενώ αποδεικνύουν με εμπειρικό τρόπου την ύπαρξη της χρησιμοποιώντας επαρκή δεδομένα από νοικοκυριά, χωρίζοντας σε 8 ομάδες κλινικής κατάστασης και 6 κατηγορίες φαρμακευτικών αγωγών. Επίσης, εξετάζουν κάθε πιθανό συνδυασμό συσχετίσεων μεταξύ δύο ομάδων είτε σε ολόκληρο το δείγμα είτε στις επί μέρους υποομάδες αυτού, εντοπίζοντας στατιστικά σημαντικές διαφορές στο σχετικό κίνδυνο. Η εργασία θα αναλυθεί στην επόμενη παράγραφο.

Ο D.(1979) εξέτασε την ύπαρξη της μεροληψίας μέσω τυχαίων δειγμάτων

του γενικού πληθυσμού για την παρουσία ή όχι αναπνευστική ή κινητικής νόσου. Στη συνέχεια, έπραξε το ίδιο για όσους είχαν νοσηλευτεί τους προηγούμενους 6 μήνες. Στο δείγμα των νοσηλευόμενων, τα άτομα με αναπνευστική νόσο είναι πολύ πιο πιθανό να πάσχουν από κινητική νόσο (Σχετικές πιθανότητες 4,06). Μπορεί να συμπεράνουμε (λανθασμένα) ότι υπάρχουν συσχετισμοί μεταξύ αυτών των δύο ασθενειών. Αν κοιτάζαμε τον γενικό πληθυσμό, θα συμπεράναμε ότι δεν υπάρχει συσχέτιση μεταξύ των δύο ασθενειών (Σχετικές πιθανότητες 1,06) – το σωστό συμπέρασμα. Το εσφαλμένο συμπέρασμα προκύπτει επειδή τα άτομα που έχουν και τις δύο διαταραχές είναι πιο πιθανό να νοσηλευτούν. Επισημαίνει ότι οι σχετικές πιθανότητες μπορεί να αυξηθούν ή να μειωθούν ψευδώς λόγω του ποσοστού αποδοχής της μεροληψίας.

Το 1986, οι R. H. πρότεινε ότι για να συμβεί η μεροληψία Berkson, η ομάδα ελέγχου δεν χρειαζόταν να νοσηλευτεί και έγραψε πως η υπόθεση ότι η έκθεση δεν έχει καμία επίδραση στη νοσηλεια και δεν είναι βάσιμη σχεδόν ποτέ. Περισσότερες λεπτομέρειες για τη συγκεκριμένη εργασία θα δοθούν στην τελευταία παράγραφο.

Το 2003, οι M. F. επανεξέτασαν τη μεροληψία του Berkson και το περιέγραψαν ως πρωταρχικό πρόβλημα μελέτης ασθενών μαρτύρων βασισμένη σε νοσοκομειακά δεδομένα. Προσέγγισαν μια ιδέα παρόμοια με εκείνη του Berkson. Πρότειναν τη χρήση ελέγχων με ασθένειες με την ίδια συχνότητα εισαγωγής που μπορούν να αποφύγουν το πρόβλημα της συγκεκριμένης μεροληψίας.

Άλλο παράδειγμα αποτελεί η μελέτη των παραγόντων κινδύνου του καρκίνου της ουροδόχου κύστης (B. N. I., 2003), της οποίας ο κίνδυνος εμφάνισης αυξάνεται μέσω του καπνίσματος. Με τη χρήση νοσοκομειακών δεδομένων, η μελέτη ασθενών-μαρτύρων έδειξε πολύ μικρή σχέση μεταξύ καπνίσματος και του συγκεκριμένου καρκίνου. Μελετώντας προσεκτικά το πρόβλημα, παρατήρησαν ότι το ποσοστό των καπνιστών του δείγματος ήταν πολύ μεγαλύτερο σε σχέση με το αντίστοιχο ποσοστό στο γενικό πληθυσμό. Αυτό μπορεί να αλλοίωσε τη φύση της σχέσης μεταξύ του καπνίσματος και του καρκίνου της ουροδόχου κύστης.

Στο άρθρο των P. M. L. (2005) εξετάστηκε η αρτηριακή πίεση μεταξύ ατόμων που νοσηλεύονταν με επαναλαμβανόμενες κρίσεις ημικρανίας ή πονοκεφάλους και διαπιστώθηκε ότι ο επιπολασμός της υψηλής αρτηριακής πίεσης ήταν περίπου 38% σε σύγκριση με το ποσοστό του γενικού πληθυσμού 11%. Υπέδειξαν ότι εντός του νοσοκομείου είναι πιθανό να υπάρχει

Ενδεχόμενο	Σύνολο	Συχνότητα	Συχνότητα (%)
Αλλεργική, διατροφική και μεταβολική νόσος	149	22	14.7
Ψυχικές νευρώσεις	164	21	12.8
Παθήσεις του κυκλοφορικού	193	36	18.7
Παθήσεις του αναπνευστικού	224	20	8.9
Κινησιολογικά προβλήματα	201	23	11.4
Επιτώσεις από χημικά	101	30	29.7
Αρθριτικά	222	28	12.6
κόπωση	140	28	20

Table 2.1: Πίνακας συχνοτήτων των Νοησηλευόμενων επί των ασθενών (a) Roberts et al. (1978)

ισχυρότερη συσχέτιση μεταξύ των δύο ιατρικών καταστάσεων από αυτή που εντοπίζεται στο γενικό πληθυσμό.

2.2 Η εργασία των Roberts RS, Spitzer WO, Delmore

Τα δεδομένα προέρχονται από n=2784 ενήλικες άνω των 25. Όσον αφορά τα δεδομένα, τις κλινικές δοκιμες, τις υπηρεσίες υγείας και τη λήψη φαρμάκων πάρθηκαν από τυχαία δειγματοληψία στην οποία περιλαμβάνονται νοικοκυριά του Νοτίου Οντάριο, εκ των οποίων σε ποσοστό 94.5% συνεργάστηκε πλήρως με τους ειδικούς. Ο ακόλουθος πίνακας αναλύει τα δεδομένα της έρευνας και διατυπώθηκε στην εν λόγω εργασία (όπως και οι υπόλοιποι αυτής της παραγράφου)

Στην παρουσίαση και την ανάλυση δεδομένων οι συγγραφείς προτείνουν διαφοροποίηση με τη μέθοδο του Berkson. Για την ακρίβεια, οριοθετούν την ομάδα μαρτύρων ως τα ατομα που δε νοσούν από την ασθένεια 2, σε αντίθεση με τον Berkson που την ορίζει ως την ομάδα ατόμων που νοσούν από μία ασθένεια 3. Αν νοσούν από την ασθένεια 2 αποτελούν ένα κρούσμα, αλλιώς όχι, ενώ το άτομο μπορεί να εκτίθεται στον αιτιολογικό παράγοντα (ασθένεια 1) ή και όχι. Η δημιουργία προϋποθέσεων των χαρακτηρισμών “ασθένεια 1”

Αγωγή	Σύνολο	Συχνότητα	Συχνότητα (%)
ακετυλοσαλικυλικό οξύ	476	51	10.7
Υπακτικά	153	20	13.1
Υπνωτικά χάπια	60	16	26.7
Βιταμίνες	359	20	5.6
Ηρεμιστικά	105	18	17.1
Χάπια για την καρδιά	84	12	14.3

Table 2.2: Πίνακας συχνοτήτων των Νοσηλευόμενων επί των αγωγών

Ασθένεια Γενικού πληθυσμού	Ναι	Όχι	Σύνολο
Παθήσεις του αναπνευστικού	8	216	224
	93	2467	2560
	101	2683	2784

Table 2.3: Συσχέτιση μεταξύ ασθένειας γενικού πληθυσμού και παθήσεις του αναπνευστικού [3]

(a) Roberts et al. (1978)

και "ασθένεια 2" έχει οριστεί αυθαίρετα. Ακόμη, τα δυνατά ζεύγη, τα οποία είναι n=28 λόγω των 8 ανα 2 συνδυασμών χωρίς επανάληψη, αναλύθηκαν μαζί χρησιμοποιώντας όλα τα φαρμακευτικά και κλινικά ζεύγη διαφορών, στο σύνολο είναι 48=8x6= (πλήθος ασθενειών) x (πλήθος φαρμάκων), χρησιμοποιώντας τον πολλαπλασιαστικό νόμο.

Στους παρακάτω πίνακες παραθέτουμε τα δεδομένα με σκοπό τον έλεγχο συσχετίσεων για τις ομάδες των ασθενειων 4(πάθηση του αναπνευστικού) και 6 (τραυματισμοί) πρώτα στο πλαίσιο του γενικού πληθυσμού και στο τέλος για τον υποπληθυσμο των νοσηλευόμενων.

Από τον πίνακα 2.3a, υπολογίζουμε το λόγο πιθανοτήτων (odds ratio) και το σχετικό κίνδυνο (relative risk).

1. odds ratio $= \frac{\frac{8}{224}}{\frac{93}{2560}} = 0.98$
2. relative risk$= \frac{8\times 2467}{94\times 216}$=0.98.

Ασθένεια Νοσηλευόμενων	Ναι	Όχι	Σύνολο
	3	17	20
Παθήσεις του αναπνευστικού	27	210	237
	30	227	257

Table 2.4: Συσχέτιση μεταξύ ασθένειας νοσηλευομενων και παθήσεις του αναπνευστικού [3]

(a) Roberts et al. (1978)

Από τον παραπάνω πίνακα, υπολογίζουμε το λόγο πιθανοτήτων (odds ratio) και το σχετικό κίνδυνο (relative risk).

1. odds ratio $= \frac{\frac{3}{20}}{\frac{2}{237}} = 1.32$
2. relative risk$= \frac{3\times 210}{27\times 17}$=1.37.

Χρησιμοποιώντας το θεώρημα Fieller (Θεώρημα B.1) για το 95% διάστημα εμπιστοσύνης για τις διαφορές μεταξύ των δύο περιπτώσεων τόσο για το λόγο πιθανοτήτων, τόσο για το σχετικό κίνδυνο, παίρνουμε:

- Σχετικός κίνδυνος: (-1.82, 2.23)
- Λόγος πιθανοτήτων: (-2.09, 3)

Και τα δύο διαστήματα εμπιστοσύνης περιέχουν το μηδέν με αποτέλεσμα να θεωρήσουμε ότι στην περίπτωση αυτή, δεν υπάρχουν στατιστικά σημαντικές διαφορές

Στον παρακάτω πίνακα διδονται όλες οι δυνατές συγκρίσεις μεταξύ πληθυσμού και νοσηλευομενων για κάθε δυνατό ζεύγος ασθενειών πραγματοποιώντας όλους τους απαιτούμενους ελέγχους. Στην περίπτωση του σχετικού κινδύνου χρησιμοποιούμε το θεώρημα του Fieler (Θεώρημα B.1) ενώ για τους λόγους πιθανοτήτων παίρνουμε την προσέγγιση του Bartlett δίνοντας για κάθε έλεγχο το αντίστοιχο p-value.

Ζεύγη	Relative risk	Relative risk	p-value	Odds ratio	Odds ratio	p-value
1-2	1.15	0.53	0.32	1.16	0.51	0.36
1-3	1.49	1.72	0.63	1.55	1.94	0.55
1-4	2.22	2.67	0.71	2.47	3.04	0.72
1-5	1.63	2.97	0.13	1.72	3.55	0.16
1-6	0.92	1.64	0.03	0.92	1.79	0.07
1-7	1.19	1.78	0.32	1.21	1.95	0.36
1-8	1.81	0.40	0.01	1.89	0.37	0.04
2-3	1.95	1.40	0.32	2.09	1.50	0.57
2-4	1.82	0.00	0.04	2.09	0.00	0.01
2-5	1.57	1.69	0.89	1.64	1.8	0.87
2-6	1.18	0.8	0.29	1.2	0.78	0.58
2-7	1.58	1.87	0.68	1.66	2.08	0.67
2-8	2.06	1.34	0.25	2.18	1.41	0.48
3-4	1.46	3.31	0.05	1.52	3.86	0.05
3-5	1.32	1.29	0.96	1.35	1.33	1.00
3-6	1.00	0.68	0.20	3.54	0.65	0.70
3-7	2.02	2.91	0.23	2.20	3.54	0.22
3-8	1.25	2.05	0.06	1.28	2.30	0.07
4-5	1.06	3.29	0.01	1.06	4.06	0.02
4-6	0.98	1.32	0.36	0.98	1.37	0.57
4-7	1.07	1.98	0.14	1.08	2.22	0.22
4-8	2.99	1.42	0.04	3.28	1.50	0.16
5-6	1.9	2.03	0.78	1.96	2.32	0.65
5-7	4.54	4.07	0.72	5.98	5.71	0.92
5-8	0.68	0.78	0.75	0.67	0.76	0.81
6-7	2.06	0.91	0.05	2.2	0.90	0.10
6-8	1.83	1.26	0.24	1.91	1.30	0.74
7-8	2.15	3.27	0.10	2.28	4.18	0.13

Table 2.5: Συγκρίσεις ανά ζεύγη

Μία αντίστοιχη μεθοδολογία ακολουθήθηκε για κάθε δυνατό ζεύγος θεραπειών και ασθενειών ως προς τη σύγκριση νοσηλευόμενων και γενικού πληθυσμού. Στη συγκεκριμένη προσέγγιση είχαμε ότι 9 από τις 48 δυνατές συγκρίσεις εμφάνισαν στατιστικά σημαντική διαφορά στο σχετικό κίνδυνο. Στις 5 από αυτές εμφανίστηκε στατιστικά σημαντική διαφορά μεταξύ των επιμέρους λόγων των πιθανοτήτων.

Τα ζητούμενα αποτελέσματα παρατίθενται στον ακόλουθο πίνακα, ο οποίος περιλαμβάνει τις τιμές των σχετικών κινδύνων, των λόγων των πιθανοτήτων καθώς και τα αντίστοιχα p-values για τους επαγόμενους ελέγχους υποθέσεων. Για την περίπτωση των σχετικών κινδύνων ο έλεγχος υποθέσεων είναι

Ο παραπάνω έλεγχος πραγματοποιήθηκε μέσω του θεωρήματος του Fieller. Όσον αφορά την περίπτωση των λόγων των συμπληρωματικών πιθανοτήτων ανάμεσα στο γενικό πληθυσμό και τους νοσηλευόμενους, χρησιμοποιήθηκε ο έλεγχος του Bartlett με τη διατύπωση της μηδενικής και ενναλακτικής υπόθεσης να είναι:

- H_0: ο λόγος των συμπληρωματικών πιθανοτήτων δεν επηρεάζεται από τη νοσηλεία
- H_1: ο λόγος των συμπληρωματικών πιθανοτήτων επηρεάζεται από τη νοσηλεία

Ομάδα θεραπείας	Ασθένεια	RR	RR	P-value	OR	OR	P-value
Ασπιρίνη	Αλλεργία	1.14	0.2	0.02	1.15	0.18	0.2
Ασπιρίνη	κόπωση	1.99	0.76	<0.01	2.09	0.72	0.02
Υπακτικά	Κινησιολογικά	1.48	4.76	0.04	1.53	5.07	0.06
Υπακτικά	Αρθριτικά	1.42	3.4	<0.01	1.48	5	0.01
Υπνωτικά	Κυκλοφορικό	4.95	22.56	0.03	6.38	3.27	0.32
Βιταμίνες	Αλλεργία	1.69	0.00	0.02	1.76	0.00	0.01
Βιταμίνες	τραυματισμός	0.61	1.79	<0.01	0.61	1.92	0.11
Χάπια για την καρδιά	Κυκλοφορικό	14.77	7.06	<0.01	30.65	19.17	0.47
Χάπια για την καρδιά	Αρθριτικά	2.88	8,82	<0.01	3.46	47.92	<0.01

Table 2.6: Συγκρίσεις ανα ζεύγος Θεραπείας-Ασθένειας [3]

(a) Roberts et al. (1978)

Οι συγγραφείς, αναφέρουν ότι το επιχείρημα που χρησιμοποιεί ο Berkson για να υποστηρίξει τη συγκεκριμένη μεροληψία που εμφανίζεται, στηρίζεται στις διαφορετικές βαθμιδες νοσηλείας μεταξύ των ομάδων που αναφέρονται σε

έναν 2x2 πίνακα συνάφειας. Ακόμη, υποστηρίζουν ότι από τη στιγμή που τα δεδομένα περιλαμβάνουν τόσο στους αριθμητές τους νοσηλευόμενους οσο και στους παρανομαστές το σύνολο του πληθυσμού, δύναται να υπολογιστούν οι συγκεκριμένες βαθμίδες για κάθε ζεύγος ασθενειών.

Στον ακόλουθο πινακα παραθέτουμε τα παρατηρούμενα ποσοστά των νοσηλευόμενων που νοσούν με δύο ασθένειες, το αντίστοιχο προβλεπόμενο ποσοστό ενώ το r αναφέρεται στη συσχέτιση των κατατάξεων μεταξύ προβλεπόμενων και παρατηρούμενων ποσοστών

Ζεύγη	Ποσοστό	Προβλεπόμενο ποοστό	p-value(F)
1-2	10%	26.1%	0.02
1-3	33.3%	27.9%	0.001
1-4	16%	22.7%	0.111
1-5	29.4%	21.4%	0.013
1-6	80%	36.2%	0.001
1-7	28.6%	23.3%	0.044
1-8	7.7%	33.4%	0.274
2-3	19%	28.3%	0.001
2-4	0%	23.4%	0.101
2-5	6.7%	21.9%	0.001
2-6	28.6%	38.3%	0.001
2-7	20%	22.3%	0.001
2-8	18.8%	29.9%	0.001
3-4	31,8%	22.3%	0.001
3-5	22.2%	26.8%	0.001
3-6	42.9%	41.3%	0.001
3-7	31%	24.7%	0.001
3-8	58.3%	29.8%	0.001
4-5	29.4%	16.3%	0.001
4-6	37.5%	34.6%	0.001
4-7	21.1%	18.7%	0.145
4-8	10.3%	29.2%	0.001
5-6	38.5%	35.3%	0.001
5-7	13.8%	21.4%	0.2
5-8	28,6%	28.2%	0.001
6-7	18.8%	40.1%	0.001
6-8	44.4%	41.4%	0.001
7-8	36.4%	25.2%	0.001
Συσχέτιση	r=0.39		

Table 2.7: Συγκρίσεις μεταξύ ζευγών ασθενειών

(a) Roberts et al. (1978)

Από τον παραπάνω πίνακα συμπεραίνουμε ότι μόνο σε 5 ζεύγη δεν εντοπίζονται στατιστικά σημαντικές διαφορές. Επίσης, η βαθμίδα νοσηλείας είναι χαμηλότερη στις περιπτώσεις που δεν υπάρχει καμία από τις δύο ασθένειες ενώ υψηλότερη στην περίπτωση των ατόμων που νοσούν και με τις δύο ασθένειες, όπως είναι λογικό.

Επίσης, οι συγγραφείς χρησιμοποίησαν το γεγονός ότι ο Berkson στηρίχτηκε στην πιθανότητα της ένωσης δύο ενδεχομένων για να περιγράψει τις βαθμίδες νοσηλείας για άτομα με δύο ή και παραπάνω ασθένειες.Έτσι, υπολόγισαν την προβλεπόμενη βαθμίδα για τους ασθενείς που νοσούν και με τις δύο ασθένειες, λαμβάνοντας υπόψιν τους ασθενειες που έχουν ακριβώς τη μία από τις δύο ασθένειες.

Αυτό που υπολογίστηκε, είναι ότι οι παρατηρούμενες και οι προβλεπόμενες βαθμιδες, δεν είναι ισχυρά συσχετισμένες αλλά έχουν θετική συσχέτιση γύρω στο 0.39.

Προκειμένου να αποδοθούν οι παρατηρούμενες μεροληψίες ειδικά στο φαινόμενο Berkson μάλλον παρά στην επιλογή μεροληψίας της κλινικής συνθήκης, χρησιμοποιήθηκε το επιχειρήμα του Berkson για να προβλεφθεί η αναμενόμενη μεροληψία σε καθεμία από τις 28 δυνατές επιλογές. Αυτό απαιτεί τροποποίηση της θεωρίας στον αντικατοπτρισμό των δύο ασθενειών και τότε στη χρήση της παρατηρούμενης βαθμίδας νοσηλείας στις περιπτώσεις των μη εκτεθειμενων κρουσμάτων και μαρτύρων. Η τιμή της συσχέτισης των κατατάξεων μεταξύ των προβλεπόμενων και των παρατηρούμενων μεροληψιών ως προς τις 28 συγκρίσεις εντοπίζεται στο 0.40. Αυτή χαμηλή τιμή προέρχεται από την αδύναμη εγκυρότητα της ένωσης δύο ενδεχομένων. Παρόλα αυτά, η συγκεκριμένη συσχέτιση είναι στατιστικά σημαντική δίνοντας p=value <0.025 σε μονόπλευρο έλεγχο.

Τα συμπεράσματα της συγκεκριμένης εργασίας είναι ότι περίπου το 25% των συγκρίσεων έδειξαν στατιστικά σημαντικές διαφορές στο σχετικό κίνδυνο όταν η ανάλυση περιορίστηκε στον πληθυσμό των νοσηλευόμενων. Τα συγκεκριμένα αποτελέσματα αποτελούν μία υποστηρικτή βάση για τα επιχειρήματα του Berkson αλλά υπάρχουν και συγκεκριμένα μειονεκτήματα.

Αρχικά, από τη στιγμή που και οι 76 (48 στα γκρουπ θεραπεία-ασθένεια και 28 στα ζεύγη ασθενειών) στατιστικοί έλεγχοι πραγματοποιήθηκαν ταυτόχρονα, κάποιος θα ανέμενε ότι τουλάχιστον πάνω από 3 έλεγχοι για τη σύγκριση των σχετικών κινδύνων, θα έδειχναν στατιστικά σημαντικό αποτέλεσμα κατά τύχη (συγκεκριμένα 3.8 έλεγχοι , λόγω του επιπέδου σημαντικότητας που ορίστηκε στο 5

Επιπλέον, οι 76 σχέσεις που αναλύθηκαν αφορούσαν μόνο 14 ξεχωριστές κατηγορίες ασθενειών ή κατηγορίες φαρμάκων και επομένως δεν είναι ανεξάρτητες. Αν και οι δύο αυτοί παράγοντες μετριάζουν την πραγματική σημασία αυτών των αποτελεσμάτων, οι συγγραφείς αναφέρουν ότι οι διαφορές

που παρατηρούνται εδώ ξεπερνούν αυτά που θα μπορούσαν να αποδοθούν στην τύχη.

Επίσης, Οι διαφορές που παρατηρούνται μεταξύ των σχετικών κινδύνων στο γενικό πληθυσμό και στους νοσηλευόμενους είναι απίθανο να οφείλεται αποκλειστικά στη μεροληψία Berkson. Ουσιαστικά, υποστηρίζουν ότι αυτό οφείλεται στην ολική μεροληψία η οποία ορίζεται από τον ακόλουθο τύπο :

$$B_T = B_B + B_{CS} + B_U + e,$$

όπου:

- B_T: Είναι η ολική μεροληψία
- B_B: Είναι η μεροληψία Berkson
- B_{CS}: Η μεροληψία η οποια αναφέρει ότι ένα άτομο το οποίο πάσχει από δύο ασθένειες είναι πιο πιθανό να εισαχθεί στο νοσοκομείο
- B_U: Σύνολο μεροληψιών που οφείλεται σε άγνωστους παράγοντες
- e: Τυχαίο σφάλμα

Η συγκεκριμένη στατιστική ανάλυση λαμβάνει υπόψιν το e. Το B_U αναφέρουν ότι είναι πιο σπάνια περίπτωση ενώ η τιμή του πιθανόν να είναι ελάχιστη. Ουσιαστικά, οι ποσότητες B_B και B_{CS} είναι οι δύο τύποι μεροληψίας που συνεισφέρουν περισσότερο στη συνολική μεροληψία. Αν συνεισφέρουν τις τιμές τους σε αντίθετες κατευθύνσεις, η οποία αποτελεί μία επιθυμητή περίπτωση λόγω μείωσεις του σχετικού κινδύνου, η μεροληψία μπορεί να καμουφλαριστεί. Η συγκεκριμένη περίπτωση θα μπορούσε να εξηγήσει τη χαμηλή συσχέτιση που εντοπίζεται μεταξύ των παρατηρούμενων και των προβλεπόμενων μεροληψιών.

2.3 Η εργασία του Felstein

Στο άρθρο του Feilstein σκοπός είναι η επέκταση της έρευνας με συγκεκριμένο τρόπο με τον οποίο η μεροληψία μπορεί να λειτουργήσει σε αναδρομικές έρευνες περιπτώσεων ελέγχου των αιτιών της χρόνιας νόσου. Χρησιμοποιήθηκε ένα απλό αλγεβρικό μοντέλο που και με τη μελέτη

συγκεκριμένων ιδιαίτερων συνθηκών μέσω των οποίων η μεροληψία να είναι είτε ουσιαστική είτε αμελητέα.

Η παρουσίαση θα αφορά άτομα που έχουν ή όχι τρία κύρια Χαρακτηριστικά: Η έκθεση σε ύποπτο αιτιολογικό παράγοντα, η κύρια ασθένεια που προκαλείται από αυτόν τον παράγοντα και μια σύγκριση ή μια συγκεκριμένη συνθήκη (ή συνθήκες ελέγχου). Οι υποομάδες που σχηματίζονται από την επικάλυψη αυτών των τριών χαρακτηριστικών, και την ύπαρξη άλλων ατόμων που δεν έχουν κανένα από τα αναφερόμενα φαινόμενα, μπορούν να υποδειχθούν με δύο εναλλακτικές κλάσεις συμβόλων.

Επίσης, ορίζεται η ακόλουθη ομάδα ποσοτήτων:

- p_1: Η βαθμίδας εμφάνισης της κύριας ασθένειας στον πληθυσμό των μη εκτεθειμένων ατόμων
- p_2:Η βαθμίδα εμφάνισης της κύριας ασθένειας στον πληθυσμό των εκτεθειμένων ατόμων
- p_c: Η βαθμίδα εμφάνισης της συγκρινόμενης συνθήκης
- e: Η βαθμίδα εμφάνισης εκτεθειμένων ατόμων
- h_e, h_d, h_c, h_p Οι βαθμίδες νοσηλείας για τα εκτεθειμένα άτομα, ασθενείς με την κύρια ασθένεια, ασθενείς με τη συγκρινόμενη νόσο και για το γενικό πληθυσμό ανεξάρτητα από τις προαναφερόμενες καταστάσεις αντίστοιχα.
- $\overline{h}_e, \overline{h}_d, \overline{h}_c, \overline{h}_p$ τα επαγόμενα συμπληρωματικά.

Το αλγεβρικό μοντέλο που θα συζητηθεί εδώ, όπως άλλωστε ισχύει και στη μελέτη του Berkson, στηρίζεται στην υπόθεση ότι οι παράγοντες που οδηγούν σε η νοσηλεία δρουν ανεξάρτητα. Επίσης, γίνεται εφαρμογή της κλασικής εξίσωσης της Άλγεβρας Boole:

$$P\left(A \cup B\right) = P\left(A\right) + P\left(B\right) - P\left(A \bigcap B\right).$$

όπου P είναι κάποιος ειδικός τελεστής, όπως η πιθανότητα, ο πληθάριθμος, η βαθμίδα νοσηλείας κλπ.

Κάθε εξεταζόμενη περίπτωση θα αναπαρίσταται από μία τριάδα (E, D, C), όπου στις συνιστώσες βρίσκονται η έκθεση στον κίνδυνο (ή όχι), η προσβολή

Κατάσταση	Ποσοστό εμφάνισης	Βαθμίδα νοσηλείας
(0,0,0)	$(1-e)\cdot(1-p_1)\ \ (1-p_c)$	$1-\overline{h}_p$
(0,0,1)	$(1-e)\cdot(1-p_1)\ \ p_c$	$1-\overline{h}_p\cdot\overline{h}_c$
(0,1,0)	$(1-e)\cdot p_1\cdot(1-p_c)$	$1-\overline{h}_p\cdot\overline{h}_d$
(0,1,1)	$(1-e)\cdot p_1\cdot p_c$	$1-\overline{h}_p\cdot\overline{h}_c\cdot\overline{h}_d$
(1,0,0)	$e\cdot(1-p_2)\cdot(1-p_c)$	
(1,0,1)	$e\cdot(1-p_2)\cdot p_c$	
(1,1,0)	$e\cdot p_2\cdot(1-p_c)$	
(1,1,1)	$e\cdot p_2\cdot p_c$	

Table 2.8: Χαρακτηριστικά καταστάσεων ασθενειών στους νοσηλευόμενους

(ή όχι) από την κύρια ασθένεια και η υπό σύγκριση συνθήκη. Κάθε μία από τις 3 συνιστώσες μπορεί να λάβει τις τιμές 0 (όχι) και 1 (Ναι). Με βάση τον κλασικό πολλαπλασιαστικό νόμο, έχουμε 8 διαφορετικές περιπτώσεις στην οποία αντιστοιχεί μια συγκεκριμένη ποσότητα. Στον ακόλουθο πίνακα, συνοψίζονται οι περιπτώσεις αυτές, ενώ η πιθανότητα εμφάνισης στον πληθυσμό των νοσηλευομένων θα υπολογίζεται από το γινόμενο ποσοστό εμφάνισης x Βαθμίδα νοσηλείας.

1. Η ομάδα ελέγχου αποτελείται από νοσηλευόμενους ασθενείς χωρίς την κύρια ασθένεια:

   Αναφερόμαστε στην περίπτωση των τριάδων όπου η δεύτερη συνιστώσα παίρνει την τιμή 1. Οι διαθέσιμες τριάδες στη συγκεκριμένη περίπτωση είναι οι εκτεθειμένοι (οι τριάδες (1,0,0) και (1,0,1)) και οι μη εκτεθειμένοι (οι τριάδες (0,0,1) και (0,0,0)). Στην περίπτωση των εκτεθειμένων, η συνολική πιθανότητα είναι:

$$\begin{aligned} a_1 &= e\cdot p_2\cdot(1-p_c)\cdot(1-\overline{h_d}\cdot\overline{h_c}\cdot\overline{h_p}) \\ &+ e\cdot p_2\cdot p_c\cdot(1-\overline{h_c}\cdot\overline{h_d}\cdot\overline{h_e}\cdot\overline{h_p}) \\ &= e\cdot p_2\left(1-\overline{h_d}\cdot\overline{h_e}\cdot\overline{h_p}\cdot(1-p_c+\overline{h_c}\cdot p_c)\right) \\ &\overset{\overline{h_c}=1-h_c}{=} e\cdot p_2\left(1-\overline{h_d}\cdot\overline{h_e}\cdot\overline{h_p}\cdot(1-h_c\cdot p_c)\right) \end{aligned}$$

   Ενώ στην περίπτωση των μη εκτεθειμένων, η αντίστοιχη πιθανότητα θα

είναι

$$\begin{aligned} a_2 &= (1-e)\cdot p_1\cdot(1-p_c)\cdot(1-\overline{h_d}\cdot\overline{h_p}) \\ &+ +(1-e)\cdot p_1\cdot p_c\cdot(1-\overline{h_c}\cdot\overline{h_d}\cdot\overline{h_p}) \\ &= (1-e)\cdot p_1\left(1-\overline{h_d}\cdot\overline{h_p}\cdot(1-p_c+\overline{h_c}\cdot p_c)\right) \\ &\overset{\overline{h_c}=1-h_c}{=} (1-e)\cdot p_1\left(1-\overline{h_d}\cdot\overline{h_p}\cdot(1-h_c\cdot p_c)\right) \end{aligned}$$

Επειδή, η ποσότητα $h_c\cdot p_c$ θα είναι πολύ μικρή θα θεωρήσουμε ότι $1-h_c\cdot p_c\approx 1$ και θα πάρουμε:

$$\frac{a_1}{a_2}=\frac{e\cdot p_2\left(1-\overline{h_d}\cdot\overline{h_e}\cdot\overline{h_p}\right)}{(1-e)\cdot p_1\left(1-\overline{h_d}\cdot\overline{h_p}\right)}$$

Όσον αφορά την ομάδα ελέγχου, παίρνουμε τις περιπτώσεις (1,0,0) και (1,0,1) με συνολική πιθανότητα

$$\begin{aligned} b_1 &= e\cdot(1-p_2)\cdot(1-p_c)\cdot(1-\overline{h_c}\cdot\overline{h_p}) \\ &+ +e\cdot(1-p_2)\cdot p_c\cdot(1-\overline{h_c}\cdot\overline{h_e}\cdot\overline{h_p}) \\ &= e\cdot(1-p_2)\cdot\left(1-\overline{h_e}\cdot\overline{h_p}\cdot(1-p_c+\overline{h_c}\cdot p_c)\right) \\ &\approx e\cdot(1-p_2)\left(1-\overline{h_c}\cdot\overline{h_p}\right) \end{aligned}$$

και τις περιπτώσεις (0,0,0) και (0,0,1) με πιθανότητα

$$\begin{aligned} b_2 &= (1-e)\cdot(1-p_1)\cdot p_c\cdot(1-\overline{h_c}\cdot\overline{h_p}) \\ &+ +(1-e)\cdot(1-p_1)\cdot(1-p_c)\cdot(1-\overline{h_p}) \\ &= (1-e)\cdot(1-p_1)\cdot\left(1-\overline{h_p}\cdot(1-p_c+\overline{h_c}\cdot p_c)\right) \\ &\approx (1-e)\cdot(1-p_1)\left(1-\overline{h_p}\right) \end{aligned}$$

Τότε, θα πάρουμε άμεσα:

$$\frac{b_1}{b_2}=\frac{e\cdot(1-p_2)\cdot\left(1-\overline{h_c}\cdot\overline{h_p}\right)}{(1-e)\cdot(1-p_1)\cdot\left(1-\overline{h_p}\right)}.$$

Από τις παραπάνω περιπτώσεις παίρνουμε ότι ο επαγόμενος λόγος πιθανοτήτων, θα είναι:

$$\begin{aligned} \frac{\frac{a_1}{a_2}}{\frac{b_1}{b_2}} &= \frac{p_2}{p_1}\times\frac{\left[1-\overline{h_d}\cdot\overline{h_c}\cdot\overline{h_p}\right]\cdot\left[1-\overline{h_p}\right]}{\left[1-\overline{h_d}\cdot\overline{h_p}\right]\cdot\left[1-\overline{h_c}\cdot\overline{h_p}\right]} \\ &= \frac{p_2}{p_1}\times\left[1-\frac{h_d\cdot h_e\cdot(1-h_p)}{\left[1-\overline{h_d}\cdot\overline{h_p}\right]\cdot\left[1-\overline{h_c}\cdot\overline{h_p}\right]}\right] \end{aligned}$$

Στην προηγούμενη αναπαράσταση, το δεξί μέλος θα είναι πάντα μικρότερο από 1, εκτός από τις περιπτώσεις που $h_c = 0$ ή $h_d = 0$ είτε $1 - h_p = 0$ Οι τελευταίες δύο περιπτώσεις είναι μη δυνατές, επειδή $h_d = 0$, το οποίο σημαίνει πρακτικά ότι κανένας ασθενής δε θα νοσηλευτεί ενώ η τελευταία περίπτωση στην ουσία αναφέρει ότι όλος ο πληθυσμός νοσηλεύεται. Πρακτικά, έχουμε ότι ο λόγος των πιθανοτήτων θα υποεκτιμά, πάντα, το λόγο κινδύνου, εκτός αν $h_e = 0$, δηλαδή ότι η έκθεση στον κίνδυνο δεν έχει καμια επίδραση στη νοσηλεία. Ένα αριθμητικό παράδειγμα είναι να θεωρήσουμε τις ποσοτητες

- $h_e = 0.06$,
- $h_p = 0.04$,
- $h_d = 0.2$,

άρα

- $1 - h_e = 0.94$,
- $1 - h_p = 0.96$,
- $1 - h_d = 0.8$

και να πάρουμε άμεσα ότι:

$$\begin{aligned}\left[1 - \frac{h_d \cdot h_e \cdot (1 - h_p)}{\left[1 - \overline{h_d} \cdot \overline{h_p}\right] \cdot \left[1 - \overline{h_c} \cdot \overline{h_p}\right]}\right] &= \left[1 - \frac{0.2 \cdot 0.06 \cdot 0.96}{[1 - 0.8 \cdot 0.96] \cdot [1 - 0.94 \cdot 0.96]}\right] \\ &= 0.49\end{aligned}$$

το οποίο σημαίνει πως ο λόγος των πιθανοτήτων θα είναι μισός σε σχέση με τη σωστή τιμή του σχετικού κίνδύνου.

2. Η ομάδα ελέγχου αποτελείται από ασθενείς που δε νοσούν με συγκρινόμενη ασθένεια:

   Σε αυτή την περίπτωση, πρέπει να ληφθεί απόφαση σχετικά με την κατάσταση των ασθενών που έχουν τόσο τη νόσου όσο και τη συνθήκη ελέγχου. Εάν τέτοιοι ασθενείς θεωρούνται πάντα ως προσελημένπο - που είναι η συνηθισμένη προσέγγιση - οι δύο ομάδες ασθενών είναι ίδιες με αυτές στην Κατάσταση 1. Οι δύο νοσηλευόμενες ομάδες ελέγχου, που αποτελούν τις τριάδες (1,0,1) και (0,0,1), επάγουν τις ακόλουθες αναλογίες:

- Εκτεθιμένοι αλλά όχι ασθενείς: $c_1 = e \cdot (1 - p_3) \cdot p_c \cdot \left[1 - \overline{h_e} \cdot \ \overline{h_c} \cdot \overline{h_p}\right]$
- Μη εκτεθειμένοι ασθενεις: $c_2 = (1 - e) \cdot (1 - p_1) \cdot p_c \cdot \left[1 - \overline{h_c} \cdot \overline{h_p}\right]$

Και ο επαγόμενος λόγος πιθανοτήτων, θα είναι:

$$\frac{\frac{a_1}{a_2}}{\frac{c_1}{c_2}} = \frac{p_2}{p_1} \times \left(\frac{\left(1 - \overline{h_d} \cdot \overline{h_e} \cdot \overline{h_p}\right) \cdot \left(1 - \overline{h_c} \cdot \overline{h_p}\right)}{\left(1 - \overline{h_d} \cdot \overline{h_p}\right) \cdot \left(1 - \overline{h_c} \cdot \overline{h_e} \cdot \overline{h_p}\right)} \right)$$

Στην παραπάνω αναπαράσταση, για να έχουμε αμερόληπτη εκτίμηση θα πρέπει είτε $\overline{h_e} = 0$, είτε $\overline{h_c} = \overline{h_d}$, δηλαδή η συγκρινόμενη συνθήκη και η ομάδα ελέγχου να έχουν ίδιους ρυθμούς νοσηλείας. Διαφορετικά, ο λόγος πιθανοτήτων δίνει μία μεροληπτική εκτίμηση τους σχετικούς κινδύνου. Το είδος και το μέγεθος της αμεροληψίας εξαρτώνται από τις τιμές των $\overline{h_c}$, $\overline{h_d}$.

Για παράδειγμα, θεωρούμε ότι $h_p = 0.04$, $h_e = 0.06$, $h_d = 0.2$ και θα εξετάσουμε για διαφορετικές τιμές του h_c. Αν $h_c = 0.1$ το δεξί μέλος της ποσότητας του λόγου πιθανοτήτων θα είναι 0.87, η οποία υποεκτιμά το σχετικό κίνδυνο, ενώ για $h_c = 0.3$ η ίδια ποσότητα θα είναι 1.07 και οδηγούμαστε σε υπερεκτίμηση.

3. Η ομάδα ελέγχου αποτελείται από ασθενείς που δε νοσούν με την κύρια ασθένεια Στη συγκεκριμένη περίπτωση, στην ομάδα ελέγχου βρίσκονται ασθενείς που δε νοσηλεύονται και απεικονίζονται από τις τριάδες (0,0,0), (0,0,1), (1,0,0) και (1,0,1). Οι ποσότητες που αντιστοιχούν στην ομάδα της μη κύριας ασθένειας αποτελούνται από

   - Εκτεθειμένους και μη προσβεβλημένους:

     $$d_1 = e\,(1 - p_2) \cdot (1 - p_c) + e \cdot (1 - p_2) \cdot p_c = e \cdot (1 - p_2)$$

   - Μη Εκτεθειμένους και μη προσβεβλημένους:

     $d_2 = (1 - e) \cdot (1 - p_1) \cdot p_c + (1 - e) \cdot (1 - p_1) \cdot (1 - p_c) = (1 - e) \cdot (1 - p_1)$

     Από τα παραπάνω, θα πάρουμε τον ακόλουθο λόγο πιθανοτήτων:

     $$\frac{\frac{a_1}{a_2}}{\frac{d_1}{d_2}} = \frac{p_2 \cdot (1 - p_1)}{p_1 \cdot (1 - p_2)} \ \times \left(\frac{1 - \overline{h_d} \cdot \overline{h_e} \cdot \overline{h_p}}{1 - \overline{h_d} \cdot \overline{h_p}} \right)$$

     και έτσι η ποσότητα $b_h = \left(\frac{1 - \overline{h_d} \cdot \overline{h_e} \cdot \overline{h_p}}{1 - \overline{h_d} \cdot \overline{h_p}} \right)$ αποτελεί τον παράγοντα μεροληψίας. Αν θεωρήσουμε ότι $h_e = 0$ θα πάρουμε μία αμερόληπτη

εκτίμηση. Το ίδιο θα συμβεί και στις περιπτώσεις που $h_p = 0$ (μη εφικτή περίπτωση, καθως θεωρούμε ότι κανένας από τον πληθυσμό δε θα νοσηλευτεί) ή $h_e = 0$. Αν θεωρήσουμε ότι $h_d = 1$, σημαίνει ότι κάθε ασθενής (κύρια ασθένεια) θα νοσηλευτεί, η οποία είναι μία περίπτωση απίθανη όπως και το να μη νοσηλευτεί κανείς. Θα θεωρήσουμε ότι όλες οι ποσότητες h_e, h_p, h_d θα βρίσκονται στο διάστημα (0,1). $\overline{h_d}\cdot\overline{h_p}\cdot\overline{h_e} < \overline{h_d}\cdot\overline{h_p}$ και τότε $1-\overline{h_d}\cdot\overline{h_p}\cdot\overline{h_e} > 1-\overline{h_d}\cdot\overline{h_p}$ και επομένως θα πάρουμε ότι $b_h > 1$. Πρακτικά, αυτό έχει ως συνέπεια μία λανθάνουσα αύξηση του λόγου κινδύνων. Σε μια προσπάθεια να αποφευχθεί αυτή η μεροληψία, μπορεί να επιλεγεί η ομάδα ελέγχου του πληθυσμού από άτομα που έχουν τη συνθήκη σύγκρισης, αν και η ανακάλυψη και η αναγνώριση τους θα ήταν μια ιδιαίτερα δύσκολη πράξη πληθυσμιακής έρευνας. Ακόμα και να βρεθεί αυτή η ομάδα, το σφάλμα θα παραμείνει. Οι σχετικές τους αναλογίες στην ομάδα με τη συγκρινόμενη συνθήκη θα είναι $e \cdot (1 - p_2) \cdot p_c$ και $(1 - e) \cdot (1 - p_1) \cdot p_c$ για τους εκτεθειμένους και μη αντίστοιχα. Ο επαγόμενος λόγος στη ομάδα ελέγχου θα είναι $\frac{e\cdot(1-p_2)}{(1-e)\cdot(1-p_1)}$ και ο λόγος πιθανοτήτων θα είναι ο ίδιος με πριν. Στη συγκεκριμένη κατάσταση, είδαμε να συμβαίνει το αντίστροφο (υπερεκτίμηση) σε σχέση με την πρώτη (υποεκτίμηση). Αν θεωρήσουμε ότι $h_e \neq 0$, δε θα επιτευχθεί αμερόληπτη εκτίμηση του λόγου πιθανοτήτων μέσω της επιλογής της ομάδας ελέγχου από άτομα που δεν έχουν προσβληθεί από την κύρια ασθένεια. Ο λόγος κινδύνου θα είναι λανθασμένα μικρότερος αν η ομάδα ελέγχου προέρχεται μόνο από νοσηλευόμενους και λανθασμένα μεγαλύτερος αν προέρχεται από τον πληθυσμό. Η καλύτερη ελπίδα για την απόκτηση μιας αμερόληπτης εκτίμησης της αναλογίας πιθανοτήτων, είναι όταν οι ασθενείς νσηλεύονται να νοσηλευόμενος είναι να επιλέξετε μια ομάδα σύγκρισης νοσηλευόμενων που έχει μια πάθηση για την οποία ισχύει $h_c \approx h_d$.

Όπως σχολιάζουν και οι συγγραφείς, τα αποτελέσματα της συγκεκριμένης εργασίας έκαναν τις εξής συνεισφορές:

- Για την υπόδειξη της σχέσης μεταξύ της αναλογίας κινδύνου για έκθεση/μη έκθεση σε μια μελέτη της αναλογίας πιθανοτήτων της περίπτωσης ασθενών μαρτύρων, χρησιμοποιήθηκε ένα αλγεβρικό μοντέλο που είναι σχετικά έυκολα κατανοητό.
- Το μοντέλο έχει αναπτυχθεί ειδικά για τους τρεις τύπους μεθόδων δειγματοληψίας που χρησιμοποιούνται συνήθως σε αναδρομικές μελέτες περιπτώσεων ελέγχου. Έτσι, οι καταστάσεις στις οποίες

τα περιστατικά είναι όλα νοσηλευόμενοι ασθενείς, με τις ομάδες ελέγχου να προέρχονται από (α) ασθενείς που νοσηλεύονται με άλλες ασθένειες, (β) ασθενείς που νοσηλεύονται μεσυγκεκριμένη συγκριτική κατάσταση και (γ) μη νοσηλευόμενα άτομα του πληθυσμού.

- Το μοντέλο λαμβάνει υπόψη, μέσω του παράγοντα h_p τη νοσηλεία που συμβαίνει για λόγους που δεν σχετίζονται με τη νόσο, την κατάσταση ελέγχου ή την έκθεση. Αυτή η γενίκευση δεν έχει χρησιμοποιηθεί ξανά στη βιβλιογραφία.
- Τα αποτελέσματα καταδεικνύουν τα σχετικά μεγέθη της μεροληψίας που θα προκύψουν στα τρία σενάρια που αναφέρονται καταστάσεις και υποδεικνύουν τις συνθήκες που απαιτούνται για την αποφυγή μεροληψίας κατά τον έλεγχο υποθέσεων. Ο λόγος πιθανοτήτων χρησιμοποιείται για την εκτίμηση του δείκτη κινδύνου. Η αναλογία πιθανοτήτων θα είναι πάντα αμερόληπτη αν $h_e = 0$.

Όσον αφορά, τη μεροληψία του Berkson, εξαγονται τα ακόλουθα συμπεράσματα:

- Η μεροληψία Berkson θα είχε αποφευχθεί στο παρελθόν, αν τα περιστατικά και οι ομάδες ελέγχουν είχαν όλες επιλεχθεί από τον πληθυσμό.
- Εάν οι περιπτώσεις επιλέγονται από έναν νοσηλευόμενο πληθυσμό, οι μαθηματικές επιπτώσεις της μεροληψίας του Berkson δεν θα αποφευχθούν επιλέγοντας την ομάδα ελέγχου από τον πληθυσμό. Στην πραγματικότητα, μια τέτοια επιλογή πάντα θα ανεβάζει ψευδώς την αναλογία πιθανοτήτων. Αντίθετα, θα μειώσουν λανθασμένα τον λόγο πιθανοτήτων εάν επιλεγεί η ομάδα ελέγχου όλους τους νοσηλευόμενους ασθενείς που δεν έχουν την κύρια ασθένεια.
- Με μια ομάδα ασθενών που νοσηλεύονται, η καλύτερη επιλογή για την ελπίδας αποφυγής της μεροληψίας του Berkson είναι η επιλογή μιας νοσηλευόμενης ομάδας ελέγχου, που έχει μια κατάσταση σύγκρισης για την οποία το ποσοστό των νοσηλεία ισούται με το ανάλογο ποσοστό στην ομάδα ασθενών (περιπτώσεων). Ακόμα και ίσες να είναι, ο λόγος πιθανοτήτων θα είναι μεροληπτικός προς τα πάνω εάν $h_e = 0$ $h_c > h_d$ ή $h_c < h_d$

2.4 Συμπεράσματα

Με βάση τη συνολική βιβλιογραφία, μερικά συμπεράσματα δίνονται ακολούθως:

- Η επιλογή των κατάλληλων ελέγχων για μια μελέτη ασθενών-μαρτύρων είναι κρίσιμη για την απόκτηση των καλύτερων δυνατών πληροφοριών. Η μεροληψία του Berkson θα πρέπει να ληφθεί υπόψη και να γίνουν σχέδια για να αποφευχθεί αυτό όπου είναι δυνατόν. Σε μια μελέτη που χρησιμοποιεί δεδομένα από νοσοκομειακά περιστατικά, μπορεί να είναι προτιμότερο να χρησιμοποιηθούν ταιριαστοί έλεγχοι που βρίσκονται επίσης στο νοσοκομείο, αλλά η χρήση ελέγχων από τον γενικό πληθυσμό είναι καλύτερη, για να αποφευχθεί η μεροληψία του Berkson.

- Η έμμεση εμφάνιση της μεροληψίας του Berkson (συσχετίσεις έκθεσης-ασθένειας που προκύπτουν επειδή μια άλλη ασθένεια σχετίζεται με την έκθεση που μελετάται) μετριάζεται κυρίως με τη χρήση περιστατικών ασθένειας (δηλ. μη επικρατούσες, ήδη υπάρχουσες περιπτώσεις). Μπορεί επίσης να αποφευχθεί με τον αποκλεισμό ασθενών που έχουν νοσηλευτεί λόγω άλλης ασθένειας.

- Όταν είναι γνωστό ότι μια συσχέτιση μεταξύ μιας έκθεσης και ενός αποτελέσματος επηρεάζει την επιλογή των περιπτώσεων και των ελέγχων σε μια μελέτη (π.χ. σε ένα νοσοκομειακό περιβάλλον), είναι προτιμότερο να προσαρμόζεται η ανάλυση για να προσπαθήσουμε να αντιμετωπίσουμε την εν λόγω μεροληψία.

ΚΕΦΑΛΑΙΟ 3

Σταθμισμενες Κατανομές

3.1 Εισαγωγή

Στο παρόν κεφάλαιο θα προβούμε στην εισαγωγή και την ανάλυση των σταθμισμένων κατανομών. Είναι μία κλάση κατανομών οι οποίες χρησιμοποιούνται στην επιλογή δειγματοληπτικού δείγματος. Για την ακρίβεια, σε πολλές περιπτώσεις εφαρμογών παρατηρείται μία λανθάνουσα συσχέτιση, οποιουδήποτε προσήμου, μεταξύ δύο τυχαίων μεταβλητών X και Y, οι οποίες μπορεί να μην είναι ούτε συσχετισμένες ούτε αντίθετα συσχετισμένες. Για την ακρίβεια, μπορεί να υπάρχει μία κρυφή μεταβλητή, έστω Z, η οποία μπορεί να μην περιλαμβάνεται στην εκάστοτε μελέτη ή να είναι δύσκολο να βρεθεί, με αποτέλεσμα να δημιουργείται μία σχέση είτε πιο ισχυρή είτε πιο αδύναμη από την πραγματική της ισχύ. Άλλος ένας λόγος που μπορεί οδηγεί στην παρατήρηση ψευδούς συσχέτισης μεταξύ δύο χαρακτηριστικών είναι η υπερεκπροσώπηση στο δείγμα των ατόμων που ικανοποιούν συγκεκριμένες ιδιότητες. Μία τέτοια περίπτωση εντοπίζεται στη μεροληψία του Berkson την οποία αναλύουμε στην παρούσα εργασία.

Εφόσον το παράδοξο του Berkson συμβαίνει στην πραγματικότητα από την υπερεκπροσώπηση στοιχείων του δείγματος που ικανοποιούν συγκεκριμένες ιδιότητες, το παράδοξο του Berkson μπορεί να ερμηνευθεί ως πρόβλημα επιλογής μεροληπτικού δείγματος , με αποτέλεσμα ένα δείγμα που δεν είναι αντιπροσωπευτικό του υπό μελέτη πληθυσμού.

Ένας λόγος που συμβαίνει η μεροληψία του Berkson είναι ότι το δείγμα που λαμβάνεται, στην ουσία είναι ένα μη τυχαία δείγμα από τον πληθυσμό. Σε

πολλές περιπτώσεις τα μεροληπτικά δείγματα προκύπτουν λόγω της φύσης του προβλήματος. Έτσι, για τη μελέτη της μεροληψίας του Berkson θα χρησιμοποιήσουμε την έννοια των πολυμεταβλητών σταθμισμένων κατανομων οι οποίες αναλύονται για παράδειγμα στις εργασίες των Mahfoud et al. (1982) και Sarabia J.M et al.(2008), με σκοπό να προσμαρμόσουμε τη μεροληψία σε ένα δείγμα. Ο Rao.CR (1965) εισήγαγε την έννοια των μονομεταβλητών σταθμισμένων κατανομών για να περιγράψει τη μεροληψία σε ένα δείγμα.

Ιδιότητες όλων των κλάσεων των σταθμισμένων κατανομών αναλύονται στις εργασίες των Arnold B.C and Nagaraja HN. (1991), Jain et al. (1995), Y. A. (2006). Ορισμένες από αυτές θα αναφερθούν στην ανάπτυξη της απαιτούμενης θεωρίας.

Η προσέγγιση του παράδοξου του Berkson που θα παρουσιαστεί στη συνέχεια βασίστηκε στην εργασία των Economou et al.(2020).

Ορισμός 3.1. *Έστω (X, Y) ένα τυχαίο διάνυσμα δύο διαστάσεων με από κοινού συνάρτηση πυκνότητας πιθανότητας (σ.π.π) $f(x, y; \theta)$, όπου το θ είναι ένα στοιχείο ενός παραμετρικού χώρου $\Theta \subset \mathbb{R}^s$. Το διδιάστατο τυχαίο διάνυσμα με από κοινού σ.π.π $(X_w,\ Y_w)$ με από κοινού σ.π.π*

$$f_w(x, y; \theta) = \frac{w(x, y; \theta)}{\mathbb{E}[w(X, Y; \theta)]} \cdot f(x, y; \theta)$$

Καλείται το διάνυσμα της σταθμισμένων τυχαίων μεταβλητών που αντιστοιχούν στο (X,Y), σε σύνδεση με τη μη αρνητική συνάρτηση δύο μεταβλητών $w(x, y; \theta)$ για την οποία ισχύει $\mathbb{E}[w(X, Y; \theta)] < \infty$. Η σ.π.π $f_w(x, y; \theta)$ είναι η σταθμισμένη σπ δύο μεταβλητών η οποία αντιστοιχεί στη $f(x, y; \theta)$ μέσω της w, ενώ κάθε σ.π.π f_w καθορίζεται από το $(X_w,\ Y_w)$.

Από τον ορισμό της από κοινού σταθμισμένης σπ μπορουμε να ορίσουμε τις επαγόμενες περιθώριες και δεσμευμένες σπ.

Η περιθώρια σταθμισμένη σππ της X, είναι:

$$\begin{aligned} f_{X,w}(x) &= \int_0^{oo} f_w(x,y;\theta)\,dy = \int_0^{oo} \frac{w(x,y;\theta)}{E[w(X,Y;)]} f(x,y;\theta)\ dy \\ &= \int_0^{oo} f_X(x;\theta)\frac{w(x,y;\theta)}{E[w(X,Y;\theta)]} f(y;\theta \mid x)\ dy \\ &= f_X(x;\theta)\int_0^{oo} \frac{w(x,y;\theta)}{E[w(X,Y;\theta)]} f(y;\theta \mid x)\ dy \\ &= f_X(x;\theta)\frac{E[w(x,Y;\theta \mid X)]}{E[w(X,Y;\theta)]} \\ &= f_X(x;\theta)\frac{w_1(x;\theta)}{E[E[w(X,Y;\theta) \mid X]]} \\ &= f_X(x;\theta)\frac{w_1(x;\theta)}{E[w_1(X;\theta)]}, \end{aligned}$$

όπου $w_1(X;\theta) = \mathbb{E}[w(X,Y;\theta) \mid X]$.

Αντίστοιχα, η περιθώρια σταθμισμένη σππ της τμ Y είναι:

$$f_{Y,w}(y) = f_Y(y;\theta)\frac{w_2(y;\theta)}{E[w_2(Y;\theta)]}$$

Από τα παραπάνω, έχουμε ότι οι δεσμευμένες σταθμισμένες κατανομές θα υπολογίζονται από τους ακόλουθους τύπους

$$\begin{aligned} f_{Y|X,w}(y|X) &= \frac{f_w(x,y;\theta)}{f_{X,w}(x;\theta)} = \frac{w(x,y;\theta)f_{Y|X}(y \mid x)}{w_1(x;\theta)}, \\ f_{X|Y,w}(x|Y) &= \frac{f_w(x,y;\theta)}{f_{Y,w}(y;\theta)} = \frac{w(x,y;\theta)f_{X|Y}(x \mid y)}{w_2(y;\theta)}. \end{aligned}$$

Μία άμεση ιδιότητα των σταθμισμένων κατανομών δίδεται στην παρακάτω πρόταση.

Πρόταση 3.1. *Για μία τυχούσα σταθμισμένη συνάρτηση w, όταν ισχύουν οι δύο από τις ακόλουθες ιδιότητες τότε ικανοποιείται και η τρίτη.*

1. *Οι τυχαίες μεταβλητές X, Y είναι ανεξαρτητες*
2. *Οι τυχαίες μεταβλητές X_w, Y_w είναι ανεξαρτητες*

3. Για κάθε $(x,y) \in S_1 \times S_2$, έχουμε $w(x,y) = w_1(x) \cdot w_2(y)$„ όπου οι συναρτήσεις $w_1.w_2$ είναι μη αρνητικές και $S_1 \times S_2$ είναι το καρτεσιανό γινόμενο των S_1 και S_2 των στηριγμάτων των X και Y αντίστοιχα.

Απόδειξη 1. *1, 3 Στην περίπτωση που ισχύουν ο υποθέσεις 1. και 3. έχουμε ως δεδομένα ότι:*

$$\begin{aligned} f(x,y) &= f(x) \cdot f(y), \\ w(x,y) &= w(x) \cdot w(y). \end{aligned}$$

Μπορούμε να πάρουμε

$$f_w(x,y;\theta) = \frac{w(x,y;\theta)}{E[w(X,Y;\theta)\,]} \cdot f(x,y) = w_1(x) \cdot w_2(y) \; \frac{f(x)\, f(y)}{E[w_1(X)\, w_2(Y)]}.$$

Από τη στιγμή που οι τμ θεωρούνται ανεξάρτητες θα έχουμε:

$$E[w_1(X) \cdot w_2(Y)] = E[w_1(X)\,] \cdot E[w_2(Y)\,].$$

Από τα παραπάνω λαμβάνουμε

$$f_w(x,y;\theta\,) = \frac{w_1(x)}{E[w_1(X)]} f(x) \cdot \frac{w_2(y)}{E[w_2(Y)\,]} f(y) = f_{X,w}(x) \cdot f_{Y,w}(y)$$

που είναι το ζητούμενο.

2,3 Αν υποθέσουμε ότι ισχύουν οι υποθέσεις 2 και 3 μπορούμε να αποδείξουμε ανάλογα με το προηγούμενο ότι θα ισχύει η 1.

1,2 Αν ισχύουν οι 1 και 2 τότε θα έχουμε:

$$\begin{aligned} f_{Y,w}(y) = f_{Y|X,w}(y|X) &= \frac{f_w(x,y;\theta)}{f_{X,w}(x;\theta)} \\ &= \frac{w(x,y;\theta) f_{Y|X}(y \mid x)}{w_1(x;\theta)} \\ &= \frac{w(x,y;\theta) f_Y(y)}{w_1(x;\theta)}, \\ f_{X,w}(x) = f_{X|Y,w}(x|Y) &= \frac{f_w(x,y;\theta)}{f_{Y,w}(y;\theta)} \\ &= \frac{w(x,y;\theta) f_{X|Y}(x \mid y)}{w_2(y;\theta)} \\ &= \frac{w(x,y;\theta) f_X(x)}{w_2(y;\theta)}. \end{aligned}$$

Από τα οποία, θα ισχύουν:

$$\begin{aligned} f(x,y)_{X,Y;w} &= f_{X,w}(x)\cdot f_{Y,w}(y) \Longleftrightarrow \\ \frac{w(x,y;\theta)}{\mathbb{E}[w(X,Y;\theta)]}\cdot f(x,y;\theta) &= \frac{f_Y(y;\theta)\cdot w_2(y;\theta)}{E\left[w_2(Y;\theta)\right]}\frac{f_X(x;\theta)\cdot w_1(x;\theta)}{E\left[w_1(X;\theta)\right]} \Longleftrightarrow \\ \frac{w(x,y;\theta)}{\mathbb{E}[w(X,Y;\theta)]}\cdot f(x,y;\theta) &= \frac{f(x,y)_{X,Y;w}\cdot w_1(x;\theta)\cdot w_2(y;\theta)}{E\left[w_1(X;\theta)\right]\cdot E\left[w_2(Y;\theta)\right]} \Longleftrightarrow \\ \frac{w(x,y;\theta)}{\mathbb{E}[w(X,Y;\theta)]} &= \frac{w_1(x;\theta)\cdot w_2(y;\theta)}{E\left[w_1(X;\theta)\right]\cdot E\left[w_2(Y;\theta)\right]}. \end{aligned}$$

Με βάση το παραπάνω, έχουμε ότι μπορεί δύο τυχαίες μεταβλητές να είναι ανεξάρτητες και οι επαγόμενες σταθμισμένες να μην είναι όταν δεν ικανοποιείται η τρίτη συνθήκη της προηγούμενης πρότασης, όπως επίσης δύο εξαρτημένες τυχαίες μεταβλητές να επάγουν ανεξάρτητες περιθώριες σταθμισμένες. Τέτοιες περιπτώσεις εξετάζουμε στα ακόλουθα παραδείγματα

Παράδειγμα 3.1. *Θεωρούμε τις τυχαίες μεταβλητές X και Y με από κοινού πυκνότητα*

$$f(x,y) = 2\cdot 3\cdot e^{-2x-3y},\ x,y>0.$$

και επαγόμενες περιθώριες

$$\begin{aligned} f(x) &= 2e^{-2x},\ x>0, \\ f(y) &= 3e^{-3y},\ y>0, \end{aligned}$$

Επομένως, θα έχουμε $f(x,y) = f(x)\cdot f(y)$ και τότε οι X και Y είναι ανεξάρτητες. Αν θεωρήσουμε ότι $w(x,y) = x+y$, τότε θα πάρουμε:

$$f_w(x,y) = \frac{x+y}{E[X+Y]}\cdot f(x,y;\theta) = \frac{x+y}{\frac{1}{2}+\frac{1}{3}}e^{-2x-3y},$$

το οποίο δε μπορεί να γραφεί στην απαιτούμενη μορφή για να συμπεράνουμε ότι οι X_w, Y_w είναι ανεξαρτητες. Εδώ δε μπορούμε να επικαλεστούμε την τρίτη συνθήκη της προηγούμενης πρότασης λόγω της μορφής που έχει η συνάρτηση $w(x,y)$ η οποία δε μπορεί να διαχωριστεί μέσω των επαγόμενων συναρτήσεων μίας μεταβλητής.

Παράδειγμα 3.2. *Θεωρούμε τις τυχαίες μεταβλητές (X,Y) με την ακόλουθη από κοινού σππ:*

$$f(x,y) = x+y,\ 0<x,\ y<1.$$

Παίρνουμε άμεσα ότι οι συγκεκριμένες τυχαίες μεταβλητές, δεν είναι ανεξάρτητες, επειδή οι επαγόμενες περιθώριες συναρτήσεις πυκνότητας δίνονται στην ακόλουθη μορφή

$$\begin{aligned} f_Y(y) &= \int_0^1 x+y\ dx = y+\frac{1}{2},\ 0<y<1, \\ f_X(x) &= \int_0^1 x+y\ dy = x+\frac{1}{2},\ 0<x<1. \end{aligned}$$

Άρα, έχουμε ότι η από κοινού πυκνότητα δε μπορεί να γραφεί ως γινόμενων των επιμέρους περιθωρίων.

Αν ορίσουμε τη συνάρτηση

$$w(x,y) = \frac{1}{x+y},\ 0<x,\ y<1,$$

θα πάρουμε

$$\begin{aligned} \mathbb{E}[X+Y] &= \int_0^1\int_0^1 (x+y)dydx = 1 \\ f_w(x,y) &= 1,\ 0<x,\ y<1, \\ f_w(x) &= 1,\ 0<x<1, \\ f_w(y) &= 1,\ 0<y<1. \end{aligned}$$

Επομένως, έχουμε ότι δύο εξαρτημένες τυχαίες μεταβλητές, μπορούν να ορίσουν ανεξάρτητες σταθμισμένες κατανομές.

Με βάση τα παραπάνω συνοψίζουμε τα ακόλουθα

1. Η ανεξαρτησία δύο τυχαίων μεταβλητών δεν είναι αναγκαία συνθήκη για την ανεξαρτησία των επαγόμενων σταθμισμένων τυχαίων μεταβλητών
2. Δύο εξαρτημένες μεταβλητές μπορούν να επάγουν δύο σταθμισμένες κατατομές, οι οποίες μπορεί να είναι ανεξάρτητες.

Με βάση τα παραπάνω, φαίνεται έντονα η αναπαράσταση της σταθμισμένης συνάρτησης που χρησιμοποιούμε, ειδικά όταν είναι της μορφής $w_1(x)\cdot w_2(y)$.Μία πιθανή επιλογή τύπου σταθμισμένων συναρτήσεων που δημιουργούν εξάρτηση, δινεται από

$$w(x,y;\theta,\gamma_1,\gamma_2) = 1-h(x;\theta,\gamma_1)\ \cdot g(y;\theta,\ \gamma_2)$$

Με κατάλληλη επιλογή των συναρτήσεων $h\left(x;\theta,\gamma_1\right),\ g\left(y;\theta,\ \gamma_2\right)$, όπου το θ είναι μία διανυσματική παράμετρος η οποία καθορίζει τη διδιάστατη κατανομή του τυχαίου διανύσματος (X,Y) και $\gamma=(\gamma_1,\ \gamma_2)$ είναι ένα διάνυσμα επιπλέον παραμέτρων με θετικές συνιστώσες.

Επίσης, οι συναρτήσεις h,g έχουν οριστεί έτσι ώστε να εγγυώνται ότι η σταθμισμένη συνάρτηση w θα είναι μη αρνητική. Ταυτόχρονα, θα πρέπει η συνάρτηση $w\left(x,y;\theta,\ \gamma_1,\ \gamma_2\right)$ να δημιουργεί την κατανομή η οποία θα αντικατοπτρίζει την πιθανότητα επιλογής ενός ζεύγους παρατηρήσεων (x,y) σε ένα δείγμα. Για παράδειγμα, αυτό μπορεί να σημαίνει ότι οι συναρτήσεις $h\left(x;\theta,\ \gamma_1\right),\ g\left(y;\theta,\ \gamma_2\ \right)$ θα παίρνουν μικρές (μεγάλες) τιμές σε περιοχές όπου ένα ζεύγος (x,y) λαμβάνει υψηλή (χαμηλή) πιθανότητα επιλογής στο δείγμα.

3.2 Σταθμισμένες κατανομές και είδη σταθμισμένων συναρτήσεων

Στη συνέχεια, θα ασχοληθούμε με 4 ειδικές περιπτώσεις σταθμισμένων συναρτήσεων τις οποίες θα χρησιμοποιήσουμε σε δείγματα των οποίων η μεροληψια να οφείλεται στη μεροληψια Berkson. Αυτές ορίζονται ακολούθως.

A. Η περίπτωση στην οποία είναι πιο πιθανή η επιλογή μονάδων με μεγάλες τιμές στο X ή/και μεγάλες τιμές στο Y. Αυτό σημαίνει ότι ο μηχανισμός μη τυχαίας δειγματοληψίας δίνει ζεύγη (x, y) με μεγάλες τιμές στο X και/ή στο Y μεγάλη πιθανότητα να να παρατηρηθεί. Από την άλλη πλευρά, δίνει μικρή πιθανότητα σε ζεύγη (x, y) με μικρές τιμές τόσο στο X όσο και στο Y . Μια λογική και μαθηματικά βολική επιλογή, ανάμεσα σε άπειρο αριθμό συναρτήσεων με την παραπάνω συμπεριφορά, για τη σταθμισμένη συνάρτηση θα μπορούσε να γίνει η ακόλουθη επιλογή:

$$w_1\left(x,y;\theta,\gamma_1,\gamma_2\right)\ =1-\left\{1-F_X\left(x;\ \theta_x\right)\ \right\}^{\ \frac{1}{\gamma_1}}\left\{1-F_Y\left(y;\ \theta_y\right)\ \right\}^{\ \frac{1}{\gamma_2}}.$$

Όπου F_x, F_Y είναι οι συναρτήσεις κατανομής των X και Y αντίστοιχα, ενώ θ_x και θ_y είναι συναρτήσεις του θ. Στο γράφημα "Case A" απεικονίζονται τα ζεύγη δύο σταθμισμένων τυχαίων μεταβλητών $(X_w,\ Y_w)$ από ένα προσομοιωμένο δείγμα 50 παρατηρήσεων. Οι τμ X,Y είναι ανεξάρτητες τμ που ακολουθούν την κανονική κατανομη με μεσο 0

και τυπική απόκλιση 2 (απεικονίζονται στο γράφημα "initial case"), αμφότερες, ενώ επιλέξαμε $\gamma_1 = \gamma_2 = 20$. Από το γράφημα είναι προφανές ότι υπάρχει πράγματι μια τάση να παρατηρούνται ζεύγη με μεγάλες τιμές στο X και/ή Y και ότι το αριστερό κάτω μέρος του πληθυσμού υποεκπροσωπείται στο δείγμα, ενώ δημιουργήθηκε μία λανθάνουσα αρνητική συσχέτιση.

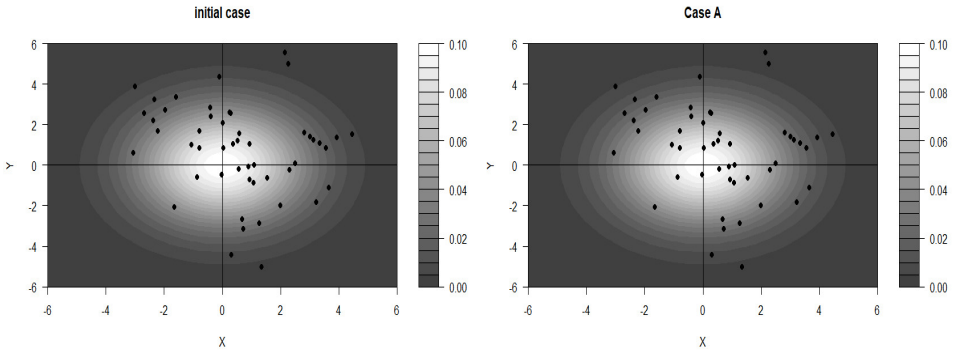

Figure 3.1: Case A

(a) Economou et al. (2020)

B. Στη δεύτερη περίπτωση θα μελετήσουμε την κατάσταση η οποία αντιστοιχεί σε μη τυχαία δειγματοληψία με παρατηρήσεις από τα ζεύγη (x,y) με μικρές τιμές στα X και/ή μικρές τιμές στα Y, που είναι πιο πιθανό να επιλεγούν. Η μαθηματική διατύπωση της σταθμισμένης συνάρτησης για τη συγκεκριμένη περίπτωση, δίνεται από:

$$w_2\left(x, y;\ \theta,\ \gamma_1,\ \gamma_2\right) = 1 - \left[F_X\left(x;\theta_x\right)\right]^{\frac{1}{\gamma_1}} \cdot \left[F_Y\left(y;\theta_y\right)\right]^{\frac{1}{\gamma_2}}$$

Από την ίδια δειγματοληψία των X και Y δημιούργουμε τις αντίστοιχες σταθμισμένες τυχαίες μεταβλητές, και οι παρατηρήσεις δίνονται στο γράφημα “case B”. Από το γράφημα παρατηρούμε εύκολα ότι υπάρχει πράγματι μια τάση να παρατηρούνται ζεύγη με μικρές τιμές στο X και/ή Y και ότι το αριστερό κάτω μέρος του πληθυσμού υποεκπροσωπείται στο δείγμα. Τέλος, παρατηρούμε μία λανθάνουσα αρνητική συσχέτιση.

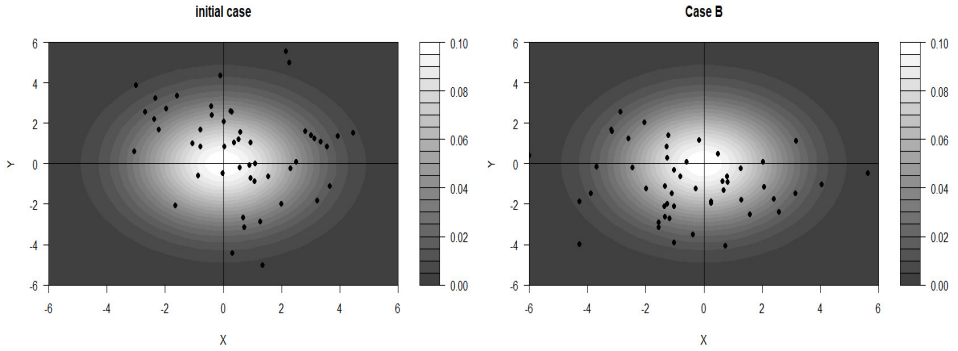

Figure 3.2: Case B

(a) Economou et al. (2020)

C. Η τρίτη περίπτωση έχει να κάνει με την επιλογή της σταθμισμένης συνάρτησης βάσει της οποίας είναι πιο πιθανή η επιλογή παρατηρήσεων με μεγάλες τιμές του X και/ή μικρές τιμές για το Y, και εκφράζεται μαθηματικά με την παρακάτω συνάρτηση

$$w_3\left(x, y;\ \theta,\ \gamma_1,\ \gamma_2\right) = 1 - \left[1 - F_X\left(x; \theta_x\right)\right]^{\frac{1}{\gamma_1}} \cdot \left[F_Y\left(y; \theta_y\right)\right]^{\frac{1}{\gamma_2}}$$

Συνεχίζοντας, στο ίδιο προσομοιωμένο δείγμα παρατηρούμε ότι στο γράφημα "Case C" παρατηρούνται περισσότερο υψηλές τιμές για το X και μικρές τιμές για το Y. Ακόμη, παρατηρούμε μία λανθάνουσα θετική συσχέτιση

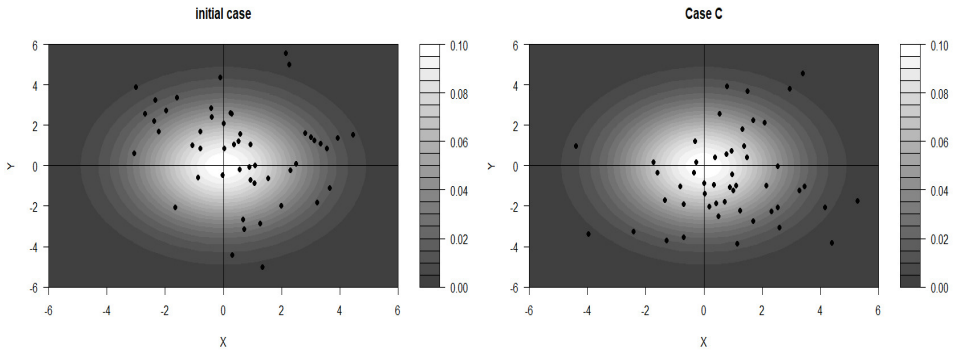

Figure 3.3: Case C

(a) Economou et al. (2020)

D. Η τελευταία περίπτωση είναι η ακριβώς αντίθεση με την τρίτη και με ανάλογο τρόπο παίρνουμε ότι η σταθμευμένη συνάρτηση θα δίνεται από τον ακόλουθο τύπο:

$$w_4\left(x,y;\ \theta,\ \gamma_1,\ \gamma_2\right) = 1 - \left[F_X\left(x;\theta_x\right)\right]^{\frac{1}{\gamma_1}} \cdot \left[1 - F_Y\left(y;\theta_y\right)\right]^{\frac{1}{\gamma_2}}$$

Η προσομοίωση έδωσε το γράφημα "Case D" και παρατηρούμε την περίπτωση την οποία αναλύσαμε. Επίσης, όπως και στην περίπτωση C, έτσι κι εδώ παρατηρούμε μία λανθάνουσα θετική συσχέτιση.

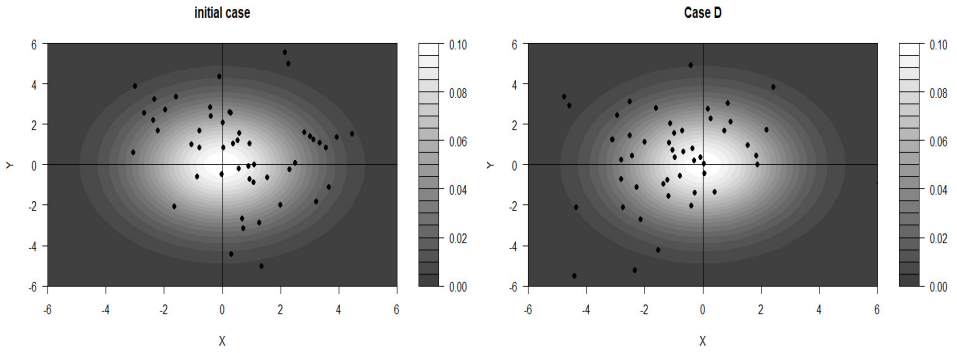

Figure 3.4: Case D

(a) Economou et al. (2020)

Όπως παρατηρούμε και από τις παραπάνω περιπτώσεις, έχουμε ότι οι σταθμισμένες συναρτήσεις επιδρούν ανάλογα με τις καταστάσεις που ορίσαμε στις μοντελοποιήσεις A-D. Προφανώς, ύπαρχουν άπειρες επιλογές τέτοιων τύπων σταθμισμένων συναρτήσεων οι οποίες μπορούν να χρησιμοποιηθούν για τη μοντελοποίηση. Όμως, εδώ θα δοθεί έμφαση στην ιδιότητα μονοτονίας ως προς x και y που διαθέτουν αυτές οι συναρτήσεις βάρους καθώς και στη χρήση των περιθώριων συναρτήσεων κατανομών που μπορούν να απλοποιήσουν την απόδειξη ύπαρξης των εν λόγω ιδιοτήτων.

Για την περαιτέρω αναπτυξη της θεωρίας, θα μελετηθεί η επίδραση των συγκεκριμένων κατανομών στην ποσότητα $cov(X_w, Y_w)$ και συγκεκριμένα στην εξάρτηση που επάγεται στις X_w και Y_w. Αυτή η ιδιότητα θα χρησιμοποιηθεί στη συνέχεια με σκοπό την επιλογή μίας εκ των τεσσάρων προαναφερθέντων σταθμισμένων συναρτήσεων. Για το σκοπό, αυτό δίνουμε τον ακόλουθο ορισμό

Ορισμός 3.2. *Μία συνάρτηση δύο μεταβλητών $\phi(x,y)$ θα ονομάζεται αντίστροφα συμμετρική τάξης 2 (reverse regular of order 2, RR2) όταν*

$$\varphi\left(x_1,y_1\right)\cdot\varphi\left(x_2,\ y_2\ \right) \leq \varphi\left(x_1,\ y_2\ \right)\ \varphi\left(x_2,\ y_1\right),\ x_1 < x_2,\ y_1 < y_2,$$

και ολικά θετική τάξης 2 (totally positive of order 2, TP2) όταν η ανισότητα ισχύει ανάποδα.

Στην ακόλουθη πρόταση θα δείξουμε ποια ιδιότητα ικανοποιούν οι σταθμευμένες συναρτήσεις των περιπτώσεων A-D.

Πρόταση 3.2. *Οι συναρτήσεις $w_1,\ w_2$ είναι RR2 ενώ οι συναρτήσεις w_3, w_4 είναι TP2.*

Απόδειξη 2. *Η απόδειξη θα γίνει για τις συναρτήσεις w_1 και w_3. Για τις υπόλοιπες η διαδικασία είναι ίδια. Αρχικά, επιλέγουμε $x_1 < x_2$ και $y_1 < y_2$.Η συνάρτηση κατανομής μίας τυχαίας μεταβλητής είναι άυξουσα, και επομένως έχουμε*

A. Για τις συναρτήσεις $g(y;\theta,\gamma_2) = [1 - F_Y(y)]^{\frac{1}{\gamma_2}}$ και $h\,(x;\theta,\ \gamma_1) = (1 - F_X\,(x))^{\ \frac{1}{\gamma_1}}$, παίρνουμε:

$$1 - F_X\,(x_2)\ \leq 1 - F_X\,(x_1)\ \Longrightarrow h\,(x_2;\theta,\ \gamma_1)\ \leq h\,(x_1;\theta,\ \gamma_1)$$

και

$$1 - F_Y\,(y_2)\ \leq 1 - F_Y\,(y_1)\ \Longrightarrow g\,(y_2;\theta,\ \gamma_2)\ \leq g\,(y_1;\theta,\ \gamma_2)$$

Άρα, λαμβάνουμε άμεσα

$$\begin{aligned}
& \phi(x_1,y_1)\cdot\phi(x_2,y_2) - \phi(x_1,y_2)\cdot\phi(x_2,y_1) = \\
= \ & -h(x_1;\theta,\gamma_1)\cdot g(y_1;\theta,\gamma_2) - h(x_2;\theta,\gamma_1)\cdot g(y_2;\theta,\gamma_2) \\
+ \ & h(x_1;\theta,\gamma_1)\cdot g(y_2;\theta,\gamma_2) + h(x_2;\theta,\gamma_1)\cdot g(y_1;\theta,\gamma_2) \\
= \ & h(x_1;\theta,\gamma_1)\cdot[g(y_2;\theta,\gamma_2) - g(y_1;\theta,\gamma_2)] \\
- \ & h(x_2;\theta,\gamma_1)\cdot[g(y_2;\theta,\gamma_2) - g(y_1;\theta,\gamma_2)] \\
= \ & [h(x_1;\theta,\gamma_1) - h(x_2;\theta,\gamma_1)]\cdot[g(y_2;\theta,\gamma_2) - g(y_1;\theta,\gamma_2)] \\
\leq \ & 0.
\end{aligned}$$

Επομένως η σταθμισμένη συναρτηση της A είναι RR2.

C. Έχουμε $h\,(x;\theta,\gamma_1) = (1 - F_X\,(x)\ )^{\frac{1}{\gamma_1}}$ και $g\,(y;\theta,\gamma_2) = (F_Y\,(y)\ )^{\frac{1}{\gamma_2}}$. Από αυτές έπονται οι διατάξεις $h\,(x_2;\theta,\gamma_1)\ \leq;\ h\,(x_1;\theta,\ \gamma_1)$ και

$g(y_1;\theta,\gamma_2) \le g(y_2;\theta,\gamma_2)$. *Τότε, θα πάρουμε*

$$\begin{aligned} & \phi(x_1,y_1)\cdot\phi(x_2,y_2) - \phi(x_1,y_2)\cdot\phi(x_2,y_1) = \\ = \;& -h(x_1;\theta,\gamma_1)\cdot g(y_1;\theta,\gamma_2) - h(x_2;\theta,\gamma_1)\cdot g(y_2;\theta,\gamma_2) \\ + \;& h(x_1;\theta,\gamma_1)\cdot g(y_2;\theta,\gamma_2) + h(x_2;\theta,\gamma_1)\cdot g(y_1;\theta,\gamma_2) \\ = \;& h(x_1;\theta,\gamma_1)\cdot[g(y_2;\theta,\gamma_2) - g(y_1;\theta,\gamma_2)] \\ - \;& h(x_2;\theta,\gamma_1)\cdot[g(y_2;\theta,\gamma_2) - g(y_1;\theta,\gamma_2)] \\ = \;& [h(x_1;\theta,\gamma_1) - h(x_2;\theta,\gamma_1)]\cdot[g(y_2;\theta,\gamma_2) - g(y_1;\theta,\gamma_2)] \\ \ge \;& 0. \end{aligned}$$

Από το οποίο παίρνουμε άμεσα ότι η C είναι TP2.

Το επόμενο θεώρημα είναι ένα αποτέλεσμα το οποίο δίνεται στην εργασία των Nanda and Jain (1999) και χρησιμοποιούνται τεχνικές οι οποίες είναι εκτός του σκοπού της παρούσας εργασίας. Για περισσότερες λεπτομέρειες παραπέμπουμε στην εργασία των Ebrahimi and Ghosh, (1981) και Lehmann (1966).

Πρόταση 3.3. *Για δύο ανεξάρτητες τυχαίες μεταβλητές* $X,\ Y$ *ισχύει:*

- $Cov\,(X_{w_i},\ Y_{w_i}) \le 0,\ i = 1,2$
- $Cov\,(X_{w_i},\ Y_{w_i}) \ge 0,\ i = 3,4.$

Στο παραπάνω αποτέλεσμα, η επίδραση της σταθμισμένης συνάρτησης στη δομή εξάρτησης των X_w, Y_w πραγματοποιήθηκε υπό την υπόθεση της ανεξατησίας των X, Y. Στο ακόλουθο αποτέλεσμα δίδεται η αναλοίωτη συμπεριφορά εξαρτήσεων χρησιμοποιώντας τις προαναφερθείσες σταθμισμένες συναρτήσεις.

Στη συνέχεια, θα χρησιμοποιήσουμε τους όρους PLRD (positive likelihood ratio dependence) και NLRD (Negative likelihood ratio dependence)

Πρόταση 3.4. *Για ένα τυχαίο διάνυσμα* (X,Y) *ισχύουν τα παρακάτω:*

1. *Αν το* (X,Y) *είναι PLRD, τότε το* $(X_{w_i},Y_{w_i}),\ i = 3,4$ *είναι επίσης PLRD και έχουμε* $cov\,(X_{w_i},Y_{w_i}) \ge 0,\ i = 3,4$.
2. *Αν το* (X,Y) *είναι NLRD, τότε το* $(X_{w_i},Y_{w_i}),\ i = 1,2$ *είναι επίσης NLRD και έχουμε* $cov\,(X_{w_i},Y_{w_i}) \le 0,\ i = 1,2$.

Η προηγούμενη πρόταση, ουσιαστικά αναδεικνύει το εξής ότι η συνδυακύμανση δύο τυχαιων μεταβλητών θα έχει ακριβώς το ίδιο πρόσημο με τη συνδυακύμανσης της αντιστοιχης σταθμισμένης κατανομής ή θα είναι μηδέν όταν το αρχικό τυχαίο διάνυσμα είναι PLRD(NLRD).

Για παράδειγμα, θεωρούμε ένα τυχαίο διάνυσμα από τη διδιάστατη κανονική με μέσο διάνυσμα μ και πίνακα συνδιακυμάνσεων $\mathbf{\Sigma}$, έχουμε τις ακόλουθες περιπτώσεις

- Αν ο συντελεστής συσχέτισης είναι μη αρνητικός, δηλαδή $\rho \geq 0$, στην περίπτωση του τυχαίου διανύσματος με την PLRD ιδιότητα θα έχουμε ότι για τις περιπτώσεις C-D θα πάρουμε συνδιακύμανση των επαγόμενων σταθμισμένων τυχαίων μεταβλητών με το ίδιο πρόσημο.
- Στην περίπτωση που $\rho \leq 0$ τότε αν έχουμε την NLRD ιδιότητα, στις περιπτώσεις A-B θα λάβουμε ότι ο η συνδιακύμανση των επαγόμενων σταθμισμένων τυχαίων μεταβλητών θα είναι μη αρνητική.

Αυτό όμως, είναι σε ειδικές περιπτώσεις και έτσι η περίπτωση ενός γενικευμένου αποτελέσματος που περιλαμβάνει τα παραπάνω αποτελέσματα, χωρίς να λαμβάνουμε υπόψη κάποια αρχική συνθήκη, παραμένει ένα ανοιχτό πρόβλημα.

Η επιλογή της σταθμισμένης συνάρτησης βασίζεται κυρίως στη φύση του προβλήματος και στα χαρακτηριστικά που κληρονομούνται από το παρατηρούμενο δείγμα ανά σταθμισμένη συνάρτηση. Εάν η μη τυχαία δειγματοληψία είναι γνωστή, τότε το w_1 (w_2) υιοθετείται όταν οι μονάδες με Οι μεγάλες (μικρές) τιμές στο X και/ή οι μεγάλες (μικρές) τιμές στο Y είναι πιο πιθανό να επιλεγούν, ενώ το w_3 (w_4) υιοθετείται όταν οι μονάδες με μεγάλες (μικρές) τιμές στο X και/ή μικρές (μεγάλες) τιμές στο Y είναι πιο πιθανό να επιλεγούν.

Θεωρούμε τώρα την περίπτωση που δεν υπήρχε διαθέσιμη προηγούμενη πληροφορία σχετικά με τη μη τυχαία δειγματοληψία. Ωστόσο, είναι εκ των προτέρων γνωστό ότι τα X και Υ είναι ανεξάρτητα. Τότε στην περίπτωση μίας λανθάνουσας αρνητικής συσχέτισης θα έχουμε να επιλέξουμε μεταξύ των σταθμισμενων συναρτήσεων w_1 και w_2 λόγω του ότι $Cov\left(X_{w_i}, Y_{w_i}\right) < 0,\ i = 1, 2$. Αντίστοιχα, στην περίπτωση μίας λανθάνουσας θετικής συσχέτισης θα επιλέξουμε μεταξύ των σταθμισμένων συναρτήσεων w_3 και w_4 επειδή $Cov\left(X_{w_i}, Y_{w_i}\right) > 0,\ i = 3, 4$.

Για την επιλογή, και στις δύο περιπτώσεις, μεταξύ των δύο υποψηφίων σταθμισμένων συναρτήσεων, θα πρέπει να εξετάσουμε ξανά τη φύση του μηχανισμού δειγματοληψίας και να αναγνωρίσουμε ποια ζεύγη είναι πιο πιθανό να παρατηρηθούν και που όχι, όπως εξηγήθηκε προηγουμένως.

Τέλος, αν τα X και Y δεν είναι ανεξάρτητα, αλλά είναι PLRD τότε θα έχουμε, όπως αποδεικνύεται, ότι οι επαγόμενες σταθμισμένες κατανομές μέσω των w_3 και w_4 θα έχουν ειτε θα έχουν θετική συσχέτιση ειτε θα είναι ασυσχέτιστες. Αντιστοιχα, έχουμε μια λογική κατάληξη στην περίπτωση που οι X και Y είναι NLRD μέσω των σταθμισμένων συναρτήσεων w_1 και w_2 με $Cov\left(X_{w_i},\ Y_{w_i}\right)\ \leq 0,\ i = 1, 2.$

3.3 Στατιστική Συμπερασματολογία

Η συνάρτηση πιθανοφάνειας υπό ένα μεροληπτικό δείγμα n παρατηρήσεων $D_i = (X_i, Y_i)$, $i = 1, \ldots.., n$ από ένα γεννήτορα πληθυσμό με γνωστή παραμετρική μορφή αλλά με άγνωστες παραμέτρους, είναι στις περισσότερες των περιπτώσεων πολύπλοκη. Ως συνέπεια αυτού, η πιθανιοφάνεια μπορεί να είναι υπολογιστικά χρονοβόρα και σχεδόν πάντα αδύνατο να γραφει σε αναλυτική μορφή. Αυτά τα χαρακτηριστικά οδηγούν στην υιοθέτηση μίας μεθόδου που αψηφά την πιθανοφάνεια και είναι σημαντική για τη συμπερασματολογία όπως αναφέρεται και στην εργασία των G.(1984). Μία τέτοια μέθοδος είναι και η Μπεϋζιανή προσεγγιστική υπολογιστική (Approximate Bayesian computation- ABC) που αποτελεί μία κλάση υπολογιστικών μεθόδων που χρησιμοποιούνται στη Μπεϋζιανή Συμπερασματολογία. Ο αλγόριθμος απόρριψης ABC περιγράφεται μέσω των ακόλουθων βημάτων:

1. A. Υποθέτουμε μία εκ των προτέρων κατανομή για κάθε μία από τις παραμέτρους που περιλαμβάνοται στην αναπαράσταση της σταθμισμένης συνάρτησης , δηλαδή για τη διανυσματική παράμετρο θ της διδιάστατης τυχαίας μεταβλητής καθώς και για τις παραμέτρους γ_1, γ_2.

   B. Προσομοιώνουμε παρατηρήσεις $\theta^*, \gamma_1^*,\ \gamma_2^*$ από τις προαναφερθείσες εκ των προτέρων κατανομές και θεωρούμε $\zeta^* = (\theta^*, \gamma_1^*,\ \gamma_2^*)$

2. Προσαρμόζουμε το ζ^* κατάλληλα στη δομή της σταθμισμένης συνάρτησης και προσομοιώνουμε το δείγμα

$D_i^* = (X_i^*,\ Y_i^*)\,,\ i = 1, \ldots.., n$, το οποίο προέρχεται από την κατανομή με συνάρτηση πυκνότητας $f_w\,(x, y;\ \theta^*,\ \gamma_1^*,\ \gamma_2^*)\,.$

3. Υπολογίζουμε την απόκλιση των προσομοιωμένων παρατηρήσεων με τις παρατηρήσεις των πραγματικών δεδομένων $\Delta_{\zeta^*} = d\,(D, D^*)\,,$ όπου το d είναι ένα μέτρο απόκλισης.

4. Αποδεχόμαστε το D^* όταν $\Delta_{\zeta^*} < \varepsilon$ για κάποιο $\varepsilon > 0$. Επιστρέφουμε στο βήμα 1B. και επανάλαβε τη διαδικασία για M φορές.

Ο παραπάνω αλγόριθμος περιγράφει τη διαδικασία λήψης δείγματος από μια κατανομή κοντά στην εκ των υστέρων κατανομή του ζ^* μέσω της σύγκρισης του προσομοιωμένου δείγματος με τα πραγματικά δεδομένα. Στη συνέχεια , θα αναφερθούν σημαντικές λεπτομέρειες του αλγορίθμου με πρακτικό τρόπο.

1. Θα πρέπει να γίνει επιλογή της εκ των προτέρων κατανομής του διανύσματος θ καθώς και των παραμέτρων $\gamma_1.\,\gamma_2$. Οποιαδήποτε εκ των προτέρων γνώση θα πρέπει να εφαρμοστεί σε αυτές τις κατανομές. υτυχώς, τέτοια γνώση για τις συνιστώσες του ϑ αναμένεται να υπάρχει, αφού γνωρίζουμε την πιθανή εμφάνιση του παραδόξου του Berkson. Από την άλλη πλευρά, δεν αναμένεται να υπάρξει προηγούμενη ενημέρωση για τις παραμέτρους $\gamma_1.\,\gamma_2$

   Για το λόγο αυτό, συνίστανται εκ των προτέρων κατανομές με μεγάλες τυπικές αποκλίσεις. Μετά από ένα αριθμός μη απορρίψεων (βλ. Βήμα 4), οι εκ των προτέρων κατανομές όλων των παραμέτρων μπορούν να ενημερωθούν χρησιμοποιώντας τις πληροφορίες που λαμβάνονται από αυτά τα δείγματα για να αυξηθεί το ποσοστό αποδοχής. Για να διατηρήσουμε την προτεινόμενη μέθοδο όσο το δυνατόν πιο απλή, προτείνουμε τη χρήση περικομμένης κανονικής κατανομής από το μηδέν ως εκ των προτέρων κατανομή για οποιαδήποτε θετική παράμετρο, την κανονική κατανομή για οποιαδήποτε πραγματική παράμετρο και η ομοιόμορφη κατανομή για οποιαδήποτε παράμετρο οριοθετημένη σε ένα πεπερασμένο διάστημα [a,b].

2. Είναι εύκολα αντιληπτό πως η προσομοιώση ενός τυχαίου δείγματος με πυκνότητα $f_w\,(x, y;\ \zeta^*)$ είναι μία σύνθετη περίπτωση. Το θετικό είναι ότι στις περισσότερες των περιπτώσεων, ένας μεγάλος αριθμός, έστω N, παρατηρήσεων μπορεί να προσομοιωθεί από την πυκνότητα $f\,(x, y;\ \theta^*)\,.$ Αυτό το δείγμα N παρατηρήσεων μπορεί να θεωρηθεί ως «πληθυσμός». Για να λάβουμε ένα δείγμα από το $f_w\,(x, y;\ \zeta^*)$ μπορεί κανείς να

εφαρμόσει μια σταθμισμένη δειγματοληψία χωρίς αντικατάσταση, με σκοπό την επιλογή n παρατηρήσεων από τον «πληθυσμό» με βάρη ανάλογα με τη σταθμισμένη συνάρτηση $w(x, y;\ \zeta^*)$. Προφανώς, αυτή η διαδικασία προσθέτει ένα άλλο επίπεδο προσέγγισης της εκ των υστέρων κατανομής των παραμέτρων, αλλά αυτό δεν θα πρέπει να είναι τόσο σημαντικό εάν το μέγεθος του «πληθυσμού» N είναι σχετικά μεγάλο σε σύγκριση με το μέγεθος του δείγματος n.

3. Ένα μέτρο απόκλισης που μπορεί να χρησιμοποιηθεί μέσω του ολοκληρωτικού τετραγωνικού σφάλματος δίνεται από το ακόλουθο ολοκλήρωμα

$$T = \int_x \int_y \left[f_D(x, y) \ - f_{D^*}(x, y) \right]^2 dydx,$$

οπου $f_D,\ f_{D^*}$ είναι οι συναρτήσεις πυκνότητας βάσει των πραγματικών και προσομοιωμένων δειγμάτων αντίστοιχα. Υπό τη μηδενική υπόθεση ότι τα δείγματα προέρχονται από την ίδια πυκνότητα, το T θα είναι ασυμπτωτικά κανονικό με γνωστή μέση τιμή και τυπική απόκλιση. Ακολούθως, η κανονικοποιημένη απόλυτη τιμή του T μπορεί να θεωρηθεί ως ένα μέτρο απόκλισης $d(D,\ D^*)$. μεταξύ δύο δειγμάτων. Για περαιτέρω λεπτομέρειες παραπέμπουμε τον αναγνώστη στη μελέτη των K. S.(2012).

4. Τιμές του ε κοντά στο 0 εγγυώνται ότι οι αποδεχούμενες τιμές του ζ^* αποτελούνται από ένα δείγμα από μια κατανομή κοντά στην εκ των υστέρων κατανομή του ζ. Ωστόσο, πολύ μικρές τιμές οδηγούν σε υψηλά ποσοστά απόρριψης και επομένως σε μεγάλο αριθμό προσομοιώσεων για να ληφθεί ένας σχετικά μεγάλος αριθμός παρατηρήσεων από την εκ των υστέρων κατανομή. Η χρήση του $\varepsilon = 1.96$ αντιστοιχεί σε ένα σημαντικό επίπεδο 0,05 για τον έλεγχο της μηδενικής υπόθεσης ότι και τα δύο δείγματα προέρχονται από την ίδια πυκνότητα.

Ο αλγόριθμος απόρριψης ABC επιτρέπει το χειρισμό πολύπλοκων μοντέλων με μεγάλο αριθμό παραμέτρων, αν και όσο αυξάνεται ο αριθμός των παραμέτρων, ο ρυθμός αποδοχής των προσομοιωμένων δειγμάτων μειώνεται εκθετικά λόγω του κριτηρίου αποδοχής (F. G. O. [2010]). Κατά συνέπεια, τα πολύπλοκα μοντέλα απαιτούν μεγάλο αριθμό προσομοιώσεων που μπορεί να είναι χρονοβόρες ή να απαιτούν σημαντική υπολογιστική ισχύ.

ΚΕΦΑΛΑΙΟ 4

Εφαρμογη

4.1 Εισαγωγή

Στο παρόν κεφάλαιο θα παρουσιάσουμε την ανάλυση των Initiative. T. G. (2020) στην οποία γίνεται η μελέτη της μεροληψίας του Berkson σε παραμετρικά προβλήματα με τη χρήση των σταθμισμένων συναρτήσεων, με τη χρήση πραγματικών δεδομένων από άτομα με άνοια λόγω της νόσου του Αλτσχάιμερ. Ο υπομεταβολισμός του εγκεφάλου σχετίζεται με τη φυσιολογική γήρανση αλλά και με την πιο κοινή αιτία έναρξης άνοιας, τη με τη νόσο του Αλτσχάιμερ. Από τις πιο πρόσφατες μελέτες πάνω σε αυτό το πρόβλημα εντοπίζονται στη μελέτη των Rasmussen.K.L et al.(2018). Στην άνοια, παρατηρείται η μείωση του μεταβολισμού της αμφίπλευρης γωνιακής έλικας στ οπίσθιο και στον κάτω κροταφικό φλοιό. Το φθόριο 18(18F-FDG, ειναι ένα ραδιοισότοπο το οποίο χρησιμοποείται στην τομογραφία εκπομπής ποζιτρονίων (PET), επιτρέπει την ανίχνευση του τυπικού υπομεταβολικού προτύπου AD όχι μόνο στην άνοια AD, αλλά και σε ήπια γνωστική εξασθένηση (MCI) λόγω αλτσχάιμε. Επιπλέον, στις μελέτες των Ishibashi K. O. K. (2018), Huang.Z et al. (2018) και Jack C. C. M. (2018) τονίζεται ότι η πρόσληψη της 18F-FDG μειώνεται σε μεγάλες περιοχές του εγκεφάλου με την προχωρημένη ηλικία σε γνωστικά φυσιολογικά ηλικιωμένα άτομα με ή χωρίς AD νευροπαθολογικές αλλαγές.

Για τη ζητούμενη ανάλυση θα χρησιμοποιηθούν πραγματικά δεδομένα από την ένωση περίπου 50 ακαδημαϊκών ιδρυμάτων και ιδιωτικών εταιρειών των ΗΠΑ και Καναδά, η οποία ονομάζεται AD Neuroimaging Initiative (ADNI)

και υποστηρίζεται από το Εθνικό Ινστιτούτο για τη γήρανση (NIA), μη κερδοσκοπικούς οργανισμούς και ιδιωτικές φαρμακευτικές εταιρείες. Τα δεδομένα βρίσκονται στη διεύθυνση www.adni-info.org και αποτελούνται από 76 ασθενείς με τη νόσο του Αλτσχάιμερ.

Η εφαρμογή που θα χρησιμοποιηθεί έχει ως σκοπό τη μελέτη για το αν η προτεινόμενη μέθοδος μπορεί να αποκαλύψει χαρακτηριστικά του υπό εξέταση πληθυσμού με βάση ένα μεροληπτικό δείγμα. Για την ακρίβεια, ο πληθυσμός αποτελείται από άτομα με τη νόσο του αλτσχάιμερ και από άτομα με ανησυχητικά σημάδια ασθενούς μνήμης που αναζητούν βοήθεια σε κλινικές μνήμης. Όσον αφορά τα δεδομένα, θα προσαρμοστούν σε μία τυπική περίπτωση κλινικής μελέτης όπου θα βασιστεί κυρίως σε άτομα με τη νόσο του Αλτσχάιμερ.

4.2 Μεθοδολογία και Ανάλυση Δεδομένων

Για κάθε ασθενή του δείγματος έχουμε καταγράψει την ηλικία του X και την τιμή που έλαβε στο FDG PET score Y. Θεωρούμε ότι αμφότερες προέρχονται από έναν πληθυσμό που ακολουθεί την κανονική κατανομή. Πιο συγκεκριμένα, θα έχουμε ότι το επαγόμενο τυχαίο διάνυσμα που ορίζουν θα είναι μία διδιάστατη κανονική κατανομή (Ορισμός Α.1) με μέσο διάνυσμα (μ_x, μ_y) και πίνακα συνδυακυμάνσεων Σ με τυπικές αποκλίσεις $\sigma_x, \sigma_y > 0$ και συντελεστή συσχετισης ρ.

Λόγω της αρνητικής σχέσης μεταξύ της ανάπτυξης ηλικίας και του μεταβολισμού του εγκεφάλου, το αναμενόμενο είναι να ισχύει ότι $\rho \leq 0$. Όσον αφορά την ηλικία, έχουμε ότι τα άτομα με άνοια, με ήπια γνωστική εξασθένιση και υποκείμενο νόσημα μνήμης, ακολουθουν μία μονοκόρυφη κατανομή η οποία κινείται στις ηλικίες 80-84. Για τις συγκεκριμένες πληροφορίες, αναφέρουμε συνοπτικά τις εργασίες των L.(2018), J. W. M. (2014), Prince M., Knapp M., Guerchet M. G. M.(2014) και J. M. S. (2012).

Όσον αφορά την ανάλυση που θα χρησιμοποιήσουμε, θα ξεκινήσουμε με την κλασική προσέγγιση των δεδομένων, αναλύοντας τις εκτιμήτριες μέγιστης πιθανοφάνειας των παραμέτρων των κατανομών των X και Y και θα πραγματοποιήσουμε τον έλεγχο του Spearman για την τιμή της συσχέτισης τους. Η τελευταία θα χρειαστεί για το είδος της σταθμισμένης κατανομής που θα χρησιμοποιήσουμε, με βάση το αποτέλεσμα της Πρότασης 3.4. Ουσιαστικά, λόγω του συγκεκριμένου αποτελέσματος, θα είναι διαθέσιμες

μόνοι οι σταθμισμένες συναρτήσεις των περιπτώσεων $A-B$ ή $C-D$. Ακολούθως, από το καταλληλότερο ζεύγος θα πρέπει να διαλέξουμε ποια από τις δύο θα είναι η σωστή επιλογή, η οποία θα πραγματοποιηθεί με βάση το πως κατανέμονται οι τιμές των X και Y όπως αναλύεται στις περιπτώσεις $A-D$ της παραγράφου 3.2.

Τέλος, για τη χρήση του αλγορίθμου ABC θα θεωρήσουμε ότι οι εκ των προτέρων κατανομές των παραμέτρων μ_x, μ_y θα προέρχονται από κανονικές κατανομές με μέση τιμή τον αντίστοιχο δειγματικό μέσο και τυπική απόκλιση την αντίστοιχη δειγματική απόκλιση. Δηλαδή,

$$\mu_x \sim \mathcal{N}(\overline{x},\, s_x^2),\ \mu_y \sim \mathcal{N}(\overline{y},\, s_y^2)$$

Όσον αφορά τις τυπικές αποκλίσεις των δύο τυχαίων μεταβλητών θα χρησιμοποιήσουμε την Περικεκκομένη από το μηδέν κανονική κατανομή (ορισμός **??**) και συγκεκριμένα θα έχουμε

$$\sigma_x \sim \mathcal{TN}(s_x,\, \frac{s_x^2}{9}),\ \sigma_y \sim \mathcal{TN}(s_y,\, \frac{s_y^2}{9})$$

Τέλος, για τις παραμέτρους γ_1, γ_2 θα θεωρησουμε ότι αμφότερες ακολουθούν την $\mathcal{TN}(20,\, \frac{20^2}{9})$ ενώ για το συντελεστή συσχέτισης θα έχουμε $\rho = 2W - 1$ με $W \sim Beta(3,5)$ (Ορισμός Α.2) . Στη συνέχεια, θα προσομοιώσουμε 50 τιμές από κάθε τμ και θα γίνει μεταβολή των εκ των προτέρων κατανομών. Η τροποποίηση θα γίνει μέσω της αντικατάστασης των παραμέτρων των κατανομών παίρνοντας ως αναμενόμενη τιμή τη μέση τιμή των παρατηρήσεων και ως τυπική απόκλιση τη δειγματική τυπική απόκλιση αυτών, κρατώντας την ίδια οικογένεια κατανομών για κάθε περίπτωση. Συγκεκριμένα, για το συντελεστή συσχέτισης θα πάρουμε

$$\rho \sim Beta\left(3, 3 \cdot \frac{1-\overline{r}}{1+\overline{r}}\right).$$

4.3 Αριθμητικά Αποτελέσματα

Αρχικά, για τα πραγματικά δεδομένα καταγράψαμε το ιστόγραμμα (Γράφημα 4.1) και τα μέτρα θέσης και διασποράς τα οποία απεικονίζονται στον πίνακα 4.1.

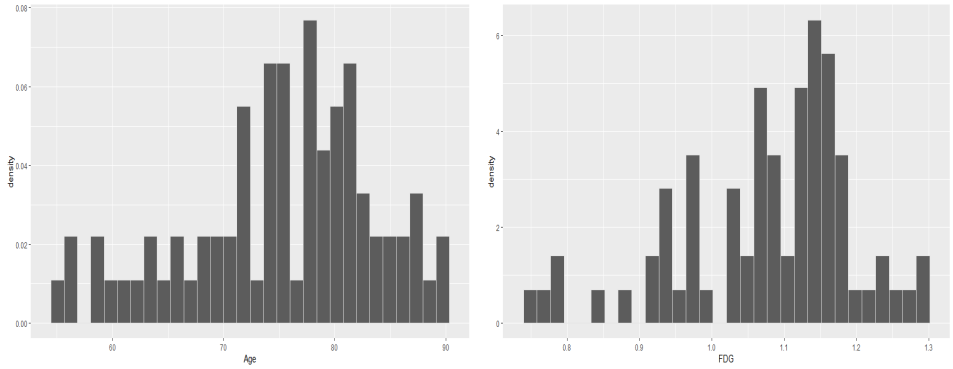

Figure 4.1: Ιστόγραμμα των μεταβλητών Age και FDG

	Age	FDG
Min	55.60	0.753
Q_1	71.05	1.016
Διάμεσος	76.35	1.109
Μέση τιμή	75.25	1.076
Q_3	81.00	1.161
Max	90.30	1.296
Διασπορά	73.7606596	0.0146578
Τυπική απόκλιση	8.588403	0.1210694

Table 4.1: Περιγραφικά στοιχεία των μεταβλητών Age και FDG

Στο παρακάτω διάγραμμα διασποράς των μεταβλητών Age και FDG παρατηρούμε ότι εντοπίζεται μία μέτρια θετική σχέση μεταξύ των δύο μεταβλητών. Η δειγματική συσχέτιση των δύο μεταβλητών υπολογίστηκε στο 0.4721 κάτι το οποίό μπορούμε να το αντιληφθούμε και από το παρακάτω γράφημα. Η μπλε γραμμή απεικονίζει το μοντέλο απλής παλινδρόμησης μεταξύ των δύο μεταβλητών.

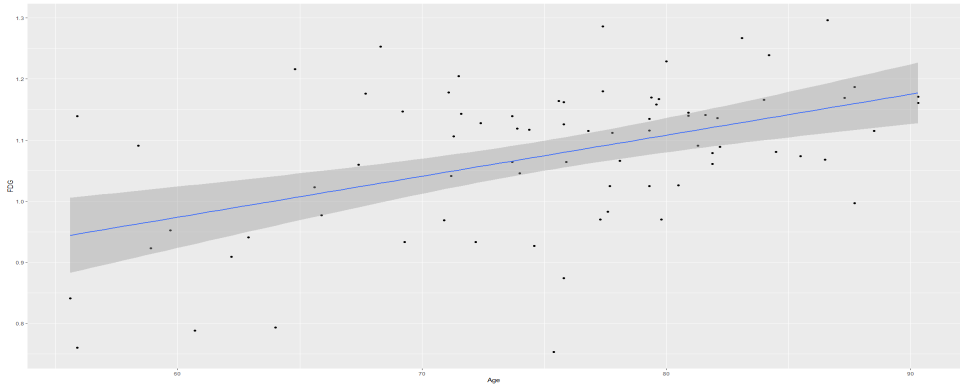

Figure 4.2: Διάγραμμα διασποράς των μεταβλητών Age και FDG

Για τις παραμέτρους του προβλήματος, με τη χρήση της κλασσικής προσέγγισης έχουμε ότι οι εκτιμήτριες μέγιστης πιθανοφάνειας των παραμέτρων της κανονικής κατανομής, θα είναι: $\hat{\mu}_x = 75.25263$, $\hat{\mu}_y = 1.076132$, $\hat{\sigma}_x = 8.531713$ $\hat{\sigma}_y = 0.1203$ και $\hat{\rho} = 0.1203$. Για το συντελεστή συσχέτισης, έχουμε οτι ο έλεγχος Spearman έδωσε $r = 0.392$ με p-value=0.0004611<0.05, το οποίο έρχεται σε αντίφαση με το γνωστό εύρημα της αντίθετης συσχέτισης μεταξύ ηλικίας και του εγκεφαλικού μεταβολισμού, και μπορεί να εξηγηθεί μέσα από το πλαίσιο της μεροληψίας Berkson και της χρήσης μεροληπτικού δείγματος. Προσομοιώνουμε 50 τιμές από τις priors και θα λάβουμε τα παρακάτω μέτρα θέσης και διασποράς των παραμέτρων

Παράμετρος	μέση τιμή	τ.α
μ_x	73.0947801	3.64645609
σ_x	8.9767988	2.57231042
μ_y	1.1584870	0.05419129
σ_y	0.1207818	0.03404304
ρ	-0.2296014	0.26193137
γ_1	18.8501190	6.23740519
γ_2	18.9103741	6.80798534

Table 4.2: Μέση τιμή και τυπική απόκλιση 50 τιμών από την εκ των προτέρων κατανομή

Λαμβάνοντας τα παραπάνω στοιχεία, θα κάνουμε αναπροσαρμογή στις τιμές των μέσων τιμών και τυπικών αποκλίσεων των prior κατανομών των εν λόγω παραμέτρων.

Επίσης, από τον έλεγχο του Spearman και το γεγονός ότι έχουμε κανονικά δεδομένα, θα έχουμε ότι το ζεύγος (X, Y) θα είναι $PLRD$ θα πάρουμε απο την πρόταση 3.4 ii) ότι η τα σταθμισμένα ζευγη για τις συναρτήσεις w_3 και w_4 θα είναι επίσης PLRD και τότε θα κρατήσουμε ως υποψήφιες μόνο τα ζεύγη $C - D$. Όπως έχει ήδη τονιστεί, η προχωρημένη ηλικία σχετίζεται τόσο με τη μείωση του μεταβολισμού του εγκεφάλου όσο και αύξηση της επίπτωσης της AD, η οποία χαρακτηρίζεται από υπομεταβολισμό του εγκεφάλου. Ωστόσο, οι συμμετέχοντες σε ερευνητικές πρωτοβουλίες, όπως το ADNI, που είναι πολύ ευαίσθητο σε ό,τι αφορά τη γνωστική εξασθένηση, δεν είναι πολύ ηλικιωμένοι, ώστε να μπορούν να υπηρετούν ως εθελοντές και να αναζητούν συμβουλές και βοήθεια νωρίς, όταν αντιμετωπίζουν μόνο ήπια έλλειψη μνήμης και ο εγκεφαλικός υπομεταβολισμός τους είναι πολύ ήπιος επίσης Ως αποτέλεσμα, ένα δείγμα μιας ερευνητικής πρωτοβουλίας όπως το ADNI είναι πιο πιθανό να περιλαμβάνει ηλικιωμένα άτομα με μικρότερη ηλικία ή/και άτομα με όχι τόσο σοβαρή μείωση του εγκεφαλικού μεταβολισμού, δηλαδή με σχετικά υψηλές βαθμολογίες FDG PET. Με βάση αυτά τα χαρακτηριστικά έχουμε ότι η καταλληλότερη σταθμισμένη συναρτηση για το συγκεκριμένο πρόβλημα θα είναι η w_4 λόγω της περιγραφής της. Επίσης, θα χρησιμοποιήσουμε και τη σταθμισμενη συνάρτηση w_3 για λόγους σύγκρισης.

Όσον αφορά τη σταθμισμενη συνάρτηση w_3 πήραμε 1746 δείματα από τα 3000 ενώ για τη σταθμισμένη συνάρτηση w_4 έχουμε 2177 από τα 3000. Τα αποτελέσματα των προσομοιώσεων δίνονται στον παρακάτω πίνακα με τα περιγραφικά στοιχεία κάθε παραμέτρου.

w_3	μ_x	σ_x	μ_y	σ_y	ρ	γ_1	γ_2
2.5%	67.4078	5.3911	1.0833	0.0673	-0.6792	7.0594	6.8242
25%	70.8843	7.7185	1.1326	0.1009	-0.3943	14.4482	14.7908
50%	73.1464	9.1066	1.1608	0.1218	-0.1846	18.6157	18.9554
75%	75.2581	10.6457	1.1910	0.1425	0.0180	22.9021	23.4667
97.5%	78.9335	13.6551	1.2406	0.1808	0.4163	30.5665	31.3729
μέσος	73.1329	9.2178	1.1611	0.1223	-0.1775	18.7251	19.0560
τ.α	3.0327	2.1072	0.0410	0.0295	0.2890	6.0619	6.3911
w_4	μ_x	σ_x	μ_y	σ_y	ρ	γ_1	γ_2
2.5%	74.4316	5.5746	0.9699	0.0729	-0.6216	5.6737	7.0066
25%	78.3879	7.8954	1.0198	0.1044	-0.2655	14.5890	15.1421
50%	80.4995	9.2463	1.0486	0.1221	-0.0166	19.2484	19.1468
75%	82.6662	10.5984	1.0755	0.1401	0.2319	24.3185	23.29553
97.5%	86.4059	13.3684	1.1242	0.1742	0.6133	35.1635	32.0239
μέσος	80.5293	9.2938	1.0471	0.1224	-0.0129	19.6309	19.2376
τ.α	3.1071	2.0471	0.0400	0.0263	0.3311	7.3010	6.2705

Table 4.3: Περιγραφικά στοιχεία των εκ των υστέρων κατανομών

Αρχικά, παρατηρούμε ότι και στις δύο περιπτώσεις έχουμε ότι ο δειγματικός μέσος και η δειγματική διάμεσος του συντελεστή συσχέτισης είναι αρνητικοί. Το συγκεκριμένο αποτέλεσμα συμφωνεί με την εκ των προτέρων άποψη πως η τιμη του FDG PET επηρεάζεται από την αύξηση της ηλικίας. Γενικά, εχουμε ότι και τα δύο μοντέλα συμφωνούν σχετικά με τα χαρακτηριστικά των παραμέτρων σ_x, σ_y, ρ, γ_1, γ_2. Εκεί που μπορούμε να εντοπίσουμε σημαντικές διαφορές, είναι στην περίπτωση των μέσων. Αρχικά, όσον αφορά το μέσο του πληθυσμού της μεταβλητής X έχουμε ότι στο μοντέλο w_3 ο μέσος των τιμών της μέσης τιμής ειναι στο 73.1329 ενώ στο μοντέλο w_4 είναι στο 80.52 ξεπερνώντας την τιμή 80. Επίσης, βλέπουμε ότι κάτι παραπάνω από το 50 % των τιμών του μ_x στο μοντέλο w_4 ξεπερνάει την τιμή 80 ενώ αντίθετα στο w_3 έχουμε ότι το 97.5 % βρίσκεται κάτω πό το 78.9335. Μεταξύ των δύο μοντέλων εντοπίζουμε σοβαρές διαφορές στην παράμετρο μ_x. Για τη μέση τιμή της τμ Y, παίρνουμε ότι στο μοντέλο w_3 η μέση τιμή βρισκεται στο 1.1611 ενώ στο w_4 είναι στο 1.0486.

Παρατηρούμε ότι η εκτιμήτρια μέγιστης πιθανοφάνειας του μ_x υποεκτιμά την τιμή της μέσης ηλικίας ενώ βρίσκεται κοντά στα επίπεδα του 2.5 % των τιμώμ που δίνει η εκ των υστέρων κατανομή στο μοντέλο w_4 ενώ είναι σίγουρα μικρότερη από το 75 % αυτών. Επίσης, η εμπ του μ_y υπερεκτιμά την τιμή του μέσου σκορ FDG PET και είναι μεγαλύτερη από το 75 % των τιμών που δίνει

η εκ των υστέρων κατανομή. Ακόμη, για το συντελεστή συσχέτισης παιρνουμε ότι η εμπ του είναι μικρότερη μόνο από το 2.5 % των τιμών που έδωσε το μοντέλο w_4. Όσον αφορά τις τυπικές αποκλίσεις δεν εντοπίζονται προβλήματα ιδιαίτερης σημασίας.

4.4 Συμπεράσματα

Αρχικά, η μοντελοποίηση που χρησιμοποιήθηκε για τη μελέτη της μεροληψίας του Berkson προέρχεται από το γεγονός πως αυτό το παράδοξο είναι στην πραγματικότητα ένα πρόβλημα που βασιζεται στην επιλογή της μεροληψίας βασιζόμενοι στο ότι στο δείγμα υπερεκπροσωπούνται άτομα που ικανοποιούν συγκεκριμένες ιδιότητες. Έτσι, οδηγούμαστε στη άμεση επιλογή μίας σταθμισμένης κατανομής με σκοπό την περιγραφή και μοντελοποίηση δεδομένων αυτής της φύσης. Μέσω των σταθμισμένων συναρτήσεων γίνεται δυνατό να περιγραφούν διαφορετικά σενάρια και βαθμοί μεροληψίας καταφέροντας να βγουν συμπεράσματα για τον υπό μελέτη πληθυσμό, μέσω του αλγορίθμου απόρριψης ABC.

Χρησιμοποιώντας τη Μπεϋζιανή τεχνική, έχουμε ότι ο αλγόριθμος ABC θα επηρεαστεί από την επιλογή των εκ των προτέρων κατανομών. Αυτό δεν είναι κάτι πρωτόγνωρο, μιας και η εκ των υστέρων κατανομή, σε κάθε πρόβλημα Μπεϋζιανής συμπερασματολογίας, καθορίζεται πλήρως από την επιλογή της εκ των προτέρων, κυρίως σε δείγματα μικρού μεγέθος ή όταν η εκ των προτέρων κατανομή είναι ακατάλληλη. Στην περίπτωση που δεν έχουμε γνώση των εκ των προτέρων κατανομών, τότε η εκ των υστέρων επηρεάζεται κυρίως από τα δεδομένα. Αντίθετα, η γνώση της εκ των προτέρων κατανομής, όπως συνέβη στην εφαρμογή που αναλύσαμε, επηρεάζει τη μορφή της εκ των υστέρων.

Όσον αφορά, την εφαρμογή, όπως αναφέρουν και οι ιδιοι συγγραφείς της μελέτης, έχουμε ότι είναι ανοιχτό ως προς διαφορετικές προσεγγίσεις διαφόρων τεχνικών της Βιοστατιστικής.

ΠΑΡΑΡΤΗΜΑ Α

Βασικές Κατανομές

Α.1 Διδιάστατη κανονική Κατανομή

Ορισμός Α.1. *Το τυχαίο διάνυσμα $Z = (X, Y)$ ακολουθεί τη διδιάστατη κανονική κατανομή $\mathcal{N}_2(\mu, \Sigma)$ με μέσο διάνυσμα μ και πίνακα συνδυακυμάνσεων Σ=$\begin{pmatrix} \sigma_x & cov(X,Y) \\ cov(X,Y) & \sigma_y \end{pmatrix}$ όταν η επαγόμενη από κοινού συνάρτηση πυκνοτητας δίνεται από τον ακόλουθο τύπο:*

$$f(z; \mu, \Sigma) = \frac{1}{2\pi} \cdot det(\Sigma)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(z-\mu)'\Sigma^{-1}(z-\mu)\right\},$$

ενώ ο πίνακας Σ είναι θετικά ημιορισμένος.

Πρόταση Α.1. *Έστω ότι το τυχαίο διάνυσμα (X, Y) προέρχεται από τη διδιάστατη κανονικη $\mathcal{N}_2(\mu, \Sigma)$, τότε ισχύουν τα εξής:*

- *Η περιθώρια κατανομή της X είναι η μονοδιάστατη κανονική $\mathcal{N}(\mu_x, \sigma_x^2)$*
- *Η δεσμευμένη κατανομή τηςX δοθέντος $Y = y$ είναι η μονοδιάστατη κανονική $\mathcal{N}(\mu_x + \frac{\sigma_x}{\sigma_y} \cdot \rho(y - \mu_y), (1 - \rho^2) \cdot \sigma_x^2)$*
- *Η περιθώρια κατανομή τηςY είναι η μονοδιάστατη κανονική $\mathcal{N}(\mu_y, \sigma_y^2)$*
- *Η δεσμευμένη κατανομή τηςX δοθέντος $Y = y$ είναι η μονοδιάστατη κανονική $\mathcal{N}(\mu_y + \frac{\sigma_y}{\sigma_x} \cdot \rho(x - \mu_x), (1 - \rho^2) \cdot \sigma_y^2)$*

Με βάση τα παραπάνω, η προσομοίωση μιας πραγματοποίησης από ένα τυχαίο διάνυσμα από τη $\mathcal{N}_2(\mu, \Sigma)$ γίνεται με τον ακόλουθο τρόπο:

- Αν οι X και Y είναι ανεξάρτητες, τότε προσομοιώνουμε κάθε μία τμ από την περιθώρια κατανομή της.
- Αν δεν είναι ανεξάρτητες τότε προσομοιώνουμε τη μία τμ από την περιθώρια της και την άλλη θα την προσομοιώσουμε από τη δεσμευμένη κατανομή της δεσμεύοντας ως προς την τιμή της πρώτης.

Α.2 Κατανομή Βήτα

Ορισμός Α.2. *Μία τυχαία μεταβλητή X ακολουθεί την κατανομή Βήτα με παραμέτρους $p > 0$ και $q > 0$ και η επαγόμενη πυκνότητα δινεται από τη συνάρτηση*

$$f(x; p, q) = \frac{\Gamma(p+q)}{\Gamma(p) \cdot \Gamma(q)} \cdot x^{p-1} \cdot (1-x)^{q-1},\ x \in (0,1),$$

όπου $\Gamma(p) = \int_0^\infty t^{p-1} e^{-t} dt$, είναι η συνάρτηση Γάμμα. Επίσης, εχουμε $\mathbb{E}[X] = \frac{p}{p+q}$.

Α.3 Περικεκκομένη Κανονική Κατανομή

Ορισμός Α.3. *Έστω X μία τυχαία μεταβλητή που προέρχεται από τη μονοδιάστατη κανονική $\mathcal{N}(\mu, \sigma^2)$. Η δεσμευμένη κατανομή της $(X \mid a \geq X \geq b)$, με $a, b \in [-\infty, +\infty]$ θα ακολουθεί την περικκεκομένη κανονική στο $[a, b]$ $\mathcal{TN}(a, b, \mu, \sigma^2)$, με σππ:*

$$f(x, a, b, \mu, \sigma^2) = \frac{1}{\Sigma} \frac{\phi(\frac{x-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})},\ x \in [a, b],$$

με Φ, ϕ να είναι οι συναρτήσεις κατανομής και πυκνότητας της τυποποιημένης κανονικής αντίστοιχα.

Για την προσομοίωση μίας τιμής από την $\mathcal{TN}(a, b, \mu, \sigma^2)$, θα χρησιμοποιήσουμε τον ακόλουθο τύπο:

$$X = \Phi^{-1}\left(\Phi(a) + U \cdot (\Phi(b) - \Phi(a))\right) \cdot \sigma + \mu,$$

όπου U είναι μία τυχαία μεταβλητή από την $\mathcal{U}(0,1)$ και Φ^{-1} είναι η αντίστροφη συνάρτηση της συναρτησης κατανομής της τυποποιημένης κανονικής.

ΠΑΡΑΡΤΗΜΑ Β

Θεωρήματα

Παρακάτω θα δώσουμε το θεώρημα του Fieller(1932) για το διάστημα εμπιστοσύνης του λόγου μεταξύ δύο μέσων τιμών δύο τυχαίων μεταβλητών, οι οποίες πιθανον να είναι συσχετισμένες. Η αναφορά του συγκεκριμένου θεωρήματος θα γίνει με βάση τους συμβολισμούς της εργασίας των Luxburg U., Volker H.F., prefix=von, prefixi=v.(2009).

Θεωρούμε ένα δείγμα παρατηρήσεων $(X_1, Y_1),\ldots,\ (X_n, Y_n)$ από ανεξάρτητα τυχαία διανύσματα (X, Y) με

- $\mathbb{E}[X] = \mu_x$, $\mathbb{V}[X] = \sigma_x^2$
- $\mathbb{E}[Y] = \mu_y$, $\mathbb{V}[Y] = \sigma_y^2$
- $Cov(X, Y) = \sigma$.

Επίσης, θεωρούμε τις ακόλουθες εκτιμήτριες

- $\hat{\mu_x} = \frac{\sum_{i=1}^{n} X_i}{n},\ \hat{\sigma_x^2} = \frac{\sum_{i=1}^{n}(X_i - \hat{\mu_x})^2}{n-1},$
- $\hat{\mu_y} = \frac{\sum_{i=1}^{n} Y_i}{n},\ \hat{\sigma_y^2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{\mu_y})^2}{n-1},$
- $\hat{\sigma} = \frac{\sum_{i=1}^{n}(Y_i - \hat{\mu_y})(X_i - \hat{\mu_x})}{n-1}$

τις ποσότητες:

- $q_1 = \frac{\hat{\mu_x}^2}{\hat{\sigma_x^2}}$,
- $q_2 = \frac{\hat{\mu_x}^2\sigma_y^2+\hat{\mu_y}^2\sigma_x^2-2\hat{\mu_x}hat\mu_y\sigma}{\sigma_x^2\sigma_y^2-\sigma^2}$

και

$$L_{1,2} = \frac{1}{\hat{\mu_x}^2 - v^2\hat{\sigma^2}_x}\left[\hat{\mu_x}\hat{\mu_y} - v^2\hat{\sigma}) \pm \sqrt{\hat{\mu_x}\hat{\mu_y} - v^2\hat{\sigma})^2 - \hat{\mu_x}^2 - v^2\hat{\sigma^2}_x\hat{\mu_y}^2 - v^2\hat{\sigma^2}_y}\right].$$

όπου $v = (t_{n-1,1-\frac{\alpha}{2}})$.

Θεώρημα Β.1. *Ένα (1-α)% διάστημα εμπιστοσύνης για το λόγο* $\frac{\mu_x}{\mu_y}$ *είναι:*

$$R = \begin{cases} (-\infty, +\infty) & q_2^2 \leq v^2 \\ (-\infty, \min(L_1, L_2)) \cap (\max(L_1, L_2), +\infty) & \textit{if } q_1^2 < v^2 < q_2^2 \\ (\min(L_1, L_2), \max(L_1, L_2)) & \textit{αλλιώς.} \end{cases}$$

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Berkson J. “Limitations of the application of fourfold table analysis to hospital data”. In: *Biometrics Bulletin* 2.3 (1946), pp. 47–53.

[2] Snyder N., Atterbury N. “Increased concurrence of cirrhosis and bacterial endocarditis: a clinical and postmortem study”. In: *Gastroenterology* 73.5 (1977), pp. 1107–1113.

[3] Roberts S., Spitzer W., Delmore T., Sackett D. D. T. “An empirical demonstration of Berkson’s bias”. In: *Journal of chronic diseases* 31.2 (1978), pp. 119–128.

[4] Conn H. O., Snyder N., Atterbury C. A. “The Berkson bias in action”. In: *The Yale journal of biology and medicine* 52.1 (1979), p. 141.

[5] Gerber L. M., Wolf A. Braham R., Alderman M. A. “Effects of sample selection on the coincidence of hypertension and diabetes”. In: *JAMA* 247.1 (1982), pp. 43–46.

[6] Kraus A. “The use of hospital data in studying the association between a characteristic and a disease”. In: *Public Health Reports* 69.12 (1954), p. 1211.

[7] Walter S. “Berkson’s bias and its control in epidemiologic studies”. In: *Journal of Chronic Diseases* 33.11-12 (1980), pp. 721–725.

[8] Schlesselman J. *Case-control studies: design, conduct, analysis*. Vol. 2. Oxford university press, 1982.

[9] Miettinen O. “Feinstein and study design”. In: *Journal of clinical epidemiology* 55.12 (2002), pp. 1167–1172.

[10] Sackett D. “Bias in analytic research”. In: *The case-control study consensus and controversy*. Elsevier, 1979, pp. 51–63.

[11] Feinstein A., Walter S., Horwitz R. H. “An analysis of Berkson’s bias in case-control studies”. In: *Journal of chronic diseases* 39.7 (1986), pp. 495–504.

[12] Schwartzbaum, Ahlbom A., Feychting M. F. "Berkson's bias reviewed". In: *European journal of epidemiology* 18.12 (2003), pp. 1109–1112.

[13] Sadetzki S., Bensal D., Novikov I., Modan B. N. I. "The Limitations of Using Hospital Controls in Cancer Etiology: One More Example for Berkson's Bias". In: *European journal of epidemiology* (2003), pp. 1127–1131.

[14] Prudenzano M.P., Monetti C., Merico L., Cardinali V, Genco.S, Lamberti P., Livrea P. M. L. "The comorbidity of migraine and hypertension. A study in a tertiary care headache centre". In: *The journal of headache and pain* 6 (2005), pp. 220–222.

[15] Patil G. Mahfoud M. *On weighted distributions, in statistics and probability: essays in honor of CR Rao (pp. 479-492)*. 1982.

[16] Déniz E.G. Sarabia J.M. "Construction of multivariate distributions: a review of some recent results". In: (2008).

[17] Rao C. "On discrete distributions arising out of methods of ascertainment". In: *Sankhyā: The Indian Journal of Statistics, Series A* (1965), pp. 311–324.

[18] Arnold B., Nagaraja H. "On some properties of bivariate weighted distributions". In: *Communications in Statistics-Theory and Methods* 20.5-6 (1991), pp. 1853–1860.

[19] Jain K., Nanda A. "On multivariate weighted distributions". In: *Communications in Statistics-Theory and Methods* 24.10 (1995), pp. 2517–2539.

[20] Navarro J., Ruiz J., Aguila Y. A. "Multivariate weighted distributions: a review and some extensions". In: *Statistics* 40.1 (2006), pp. 51–64.

[21] Nanda AK., Jain K. "Some weighted distribution results on univariate and bivariate cases". In: *Journal of Statistical planning and Inference* 77.2 (1999), pp. 169–180.

[22] Ghosh M., Ebrahimi N. "Multivariate NBU and NBUE distributions". In: *Egyptian Statist. J* 25 (1981), pp. 36–55.

[23] Lehmann E. "Some concepts of dependence". In: *The Annals of Mathematical Statistics* 37.5 (1966), pp. 1137–1153.

[24] Diggle P., Gratton G. "Monte Carlo methods of inference for implicit statistical models". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 46.2 (1984), pp. 193–212.

[25] Duong T., Goud B., Schauer K. S. "Closed-form density-based framework for automatic detection of cellular morphology changes". In: *Proceedings of the National Academy of Sciences* 109.22 (2012), pp. 8382–8387.

[26] Csilléry K., Blum M., Gaggiotti O., Olivier F. G. O. "Approximate Bayesian computation (ABC) in practice". In: *Trends in ecology & evolution* 25.7 (2010), pp. 410–418.

[27] Economou P., Batsidis A., Tzavelas G., Alexopoulos P., Alzheimer's Disease Neuroimaging Initiative. T. G. "Berkson's paradox and weighted distributions: An application to Alzheimer's disease". In: *Biometrical Journal* 62.1 (2020), pp. 238–249.

[28] Tybjærg-Hansen A., GNordestgaard B., Frikke-Schmidt R. Rasmussen K. F.-S. R. "Absolute 10-year risk of dementia by age, sex and APOE genotype: a population-based cohort study". In: *Cmaj* 190.35 (2018), E1033–E1041.

[29] Onishi A., Fujiwara Y., Oda K., Ishiwata K., Ishii K. Ishibashi K. O. K. "Longitudinal effects of aging on 18F-FDG distribution in cognitively normal elderly individuals". In: *Scientific reports* 8.1 (2018), pp. 1–8.

[30] Jiang J., Sun Y., Zhou H., Li S., Huang Z., Wu P., Shi K.,Zuo C., ADNI. Z. H. "Study of the influence of age in 18F-FDG PET images using a data-driven approach and its evaluation in Alzheimer's disease". In: *Contrast media & molecular imaging* 2018 (2018).

[31] ABennett D., Lennow K.B., Carrillo M., Dunn.B, Budd Haeberlein S., Holtzman D. M., Jagust W., Jessen F., Karlawish J., Liu L., Molinuevo J., Montine T., Phelps C., Rankin K., Rowe C., Scheltens P., Siemers E., Snyder H., Sperling R. Jack C. C. M. "NIA-AA research framework: toward a biological definition of Alzheimer's disease". In: *Alzheimer's & Dementia* 14.4 (2018), pp. 535–562.

[32] Drew L. "An age-old story of dementia." In: *Nature* 559.7715 (2018), S2–S3.

[33] Fritsch T., McClendon M., Wallenda M., Maggie.S, Hyde, Trevor F., Larsen J. W. M. "Prevalence and cognitive bases of subjective memory complaints in older adults: Evidence from a community sample". In: *Journal of neurodegenerative diseases* 2014 (2014).

[34] Prince M., Knapp M., Guerchet M. G. M. "Dementia UK: -overview". In: (2014).

[35] Ward A., Arrighi M., Michels S., Cedarbaum J. M. S. “Mild cognitive impairment: disparity of incidence and prevalence estimates”. In: *Alzheimer’s & Dementia* 8.1 (2012), pp. 14–21.

[36] Fieller.E.C. “The Distribution of the Index in a Normal Bivariate Population”. In: *Biometrika* 24.3/4 (1932), pp. 428–440. ISSN: 00063444. URL: http://www.jstor.org/stable/2331976 (visited on 01/29/2023).

[37] Luxburg U., Volker H.F., prefix=von, prefixi=v. “A GEOMETRIC APPROACH TO CONFIDENCE SETS FOR RATIOS: FIELLER’S THEOREM, GENERALIZATIONS AND BOOTSTRAP”. In: *Statistica Sinica* 19.3 (2009), pp. 1095–1117. ISSN: 10170405, 19968507. URL: http://www.jstor.org/stable/24308947 (visited on 01/29/2023).