ΠΑΝΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΚΟΙΝΩΝΙΚΩΝ & ΠΟΛΙΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

**ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΗΣ ΚΑΙ ΠΕΡΙΦΕΡΕΙΑΚΗΣ ΑΝΑΠΤΥΞΗΣ**

**Μ.Π.Σ ΣΤΗΝ ΟΙΚΟΝΟΜΙΚΗ ΚΑΙ ΠΕΡΙΦΕΡΕΙΑΚΗ ΑΝΑΠΤΥΞΗ**

# ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ ΜΕ ΘΕΜΑ

## Η ανάλυση μιας πανελλαδικής έρευνας στον σχολικό πληθυσμό.

**ΑΝΝΑ ΜΕΡΚΟ (Α.Μ. 0811Μ003)**

**ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: CLIVE RICHARDSON**

**ΜΕΛΗ ΕΠΙΤΡΟΠΗΣ : ΑΙΚΑΤΕΡΙΝΗ ΜΙΧΑΛΟΠΟΥΛΟΥ**

**ΑΝΑΣΤΑΣΙΟΣ ΤΑΣΟΠΟΥΛΟΣ**

Ευχαριστώ θερμά,

Τον καθηγητή μου κ. Clive Richardson για την προθυμία του να τεθεί επιβλέπων της παρούσας εργασίας, για την συνεργασία μας, την καθοδήγηση και την υπομονή του. Η ανταπόκριση του ήταν άμεση διευκολύνοντας με πολύ σε οτιδήποτε προέκυπτε.

# Abstract

The design effect of a survey estimate can be used as a tool for measuring sample efficiency and for survey planning. It is defined as the ratio of the variance of an estimator under the actual sample design to that of the estimator under simple random sampling with the same sample size. The purpose of this dissertation is to present the sizes of design effects for the ESPAD survey of the use of substances (alcohol, tobacco, illegal drugs) among the population of high school students. We start with a brief presentation of this Panhellenic survey, which was conducted in 2011, its methodology and its results. Sampling methods are discussed with emphasis on cluster sampling as this was employed in the survey, the unit of sampling being the school class. The cost of using cluster sampling rather than simple random sampling is the design effect. Therefore we study the effect of this survey design in this sample.

Using the statistical program STATA, we estimate the design effect using two methods. The first is the linearization method and the second is the Jackknife method. Initially we estimate the proportion of positive answers to the questions from the survey in which we are interested, and then the corresponding standard errors. The first estimate assumes simple random sampling specifying only the sampling weights and this is followed by estimates for the cluster sampling using the linearization and jackknife methods respectively. Then in logistic regression analyses we estimate the odds ratios and the standard error, first for simple random sampling and then for cluster sampling. Results are compared between the linearization and jackknife methods. Moreover we see by how much the design effect differs among the questions.

# Περιεχόμενα

# Εισαγωγή

Σκοπός της παρούσας εργασίας είναι να υπολογίσει το μέγεθος της επίδρασης του δειγματοληπτικού σχεδιασμού (ΕΔΣ) της Πανελλήνιας έρευνας στο μαθητικό πληθυσμό για τη χρήση εξαρτησιογόνων ουσιών και το πως αυτό επιδρά στον δειγματοληπτικό σχεδιασμό των ερευνών. Γίνεται μια συνοπτική παρουσίαση της έρευνας η οποία υλοποιήθηκε το 2011, της μεθοδολογίας που ακολουθήθηκε για την δειγματοληψία και των ευρημάτων της. Επιπλέον αναφέρονται τέσσερεις μέθοδοι δειγματοληψίας με εκτενέστερη αναφορά στην κατά συστάδα δειγματοληψία καθώς είναι αυτή που έχει χρησιμοποιηθεί στην έρευνα μας. Το δείγμα μας οι μαθητές δηλαδή έχουν χωριστεί σε ομάδες (τάξεις). Το τίμημα στο να ακολουθηθεί η μέθοδος της κατά συστάδες δειγματοληψία έναντι της απλής τυχαίας δειγματοληψίας είναι η ΕΔΣ, για αυτό και θα δούμε τι επίπτωση έχει στο παραπάνω δείγμα.

Σε αυτή την εργασία χρησιμοποιώντας το στατιστικό πρόγραμμα STATA, παρέχουμε δύο μεθόδους για να εκτιμήσουμε την ΕΔΣ. Πρώτη είναι η γραμμική μέθοδος και δεύτερη είναι η Jackknife μέθοδος. Εξετάζουμε την ΕΔΣ σε δύο περιοχές της ανάλυσης. Εξετάζεται πρώτα η εκτίμηση της αναλογίας του πληθυσμού του δείγματος για τις υπό μελέτη ερωτήσεις , αρχικά εκτιμάμε την αναλογία για τη μέθοδο της απλής τυχαίας δειγματοληψίας (ΑΤΔ) καθορίζοντας μόνο τα βάρη και ύστερα την αναλογία για την κατά συστάδες δειγματοληψία (ΚΣΔ) εφαρμόζοντας τη γραμμική και Jackknife μέθοδο αντίστοιχα. Έπειτα ακολουθεί η λογιστική παλινδρόμηση. Αρχικά και αυτή η εκτίμηση γίνεται για την ΑΤΔ και ύστερα για την ΚΣΔ με τη γραμμική μέθοδο. Στη συνέχεια θα γίνει μία σύγκριση της επίδρασης του δειγματοληπτικού σχεδιασμού μεταξύ της γραμμικής και της Jackknife μεθόδου. Επιπλέον θα δούμε αν η ΕΔΣ διαφέρει πολύ μεταξύ των ερωτήσεων.

# 1 Παρουσίαση της έρευνας ESPAD 2011

Η Πανελλήνια έρευνα στο μαθητικό πληθυσμό για τη χρήση εξαρτησιογόνων ουσιών υλοποιήθηκε για πρώτη φορά το 1984 όταν οι ενδείξεις για τη διάδοση της χρήσης ναρκωτικών στους νέους είχαν πολλαπλασιαστεί και η ανησυχία της κοινωνίας είχε κλιμακωθεί.

Το Ευρωπαϊκό πρόγραμμα ESPAD (European School Survey Project on Alcohol and Other Drugs) υλοποιείται από το 1993, ανά τέσσερα χρόνια, σε μαθητές ηλικίας 16 ετών. Στην τελευταία, πιο πρόσφατη έρευνα ESPAD το 2011 συμμετείχαν 39 ευρωπαϊκές χώρες. Στη χώρα μας η Πανελλήνια έρευνα στο μαθητικό πληθυσμό του 1984 επαναλήφθηκε από την Ψυχιατρική Κλινική του Πανεπισημίου Αθηνών το 1993 και από το Ερευνητικό Πανεπιστημιακό Ινστιτούτο Ψυχικής Υγιεινής (ΕΠΙΨΥ) το 1998. Από το 1999 έως σήμερα η έρευνα επαναλαμβάνεται σταθερά από το ΕΠΙΨΥ ανά τετραετία στο πλαίσιο της Πανευρωπαϊκής Έρευνας ESPAD. Ενώ όμως η Πανευρωπαϊκή έρευνα ESPAD περιορίζεται σε δείγμα μαθητικού πληθυσμού ηλικίας 16 ετών, το δείγμα της Πανελλήνιας έρευνας περιελάμβανε πάντα μεγαλύτερο φάσμα ηλικιών (14-18 ετών).

Στόχοι και συμβολή της έρευνας - Η έρευνα εξετάζει:

- Την έκταση και τα χαρακτηριστικά της χρήσης νόμιμων και παράνομων εξαρτησιογόνων ουσιών (καπνός, οινοπνευματώδη, ναρκωτικά, ψυχοδραστικά φάρμακα) στο μαθητικό πληθυσμό εφηβικής ηλικίας, καθώς και τις αντιλήψεις των μαθητών απέναντι στη χρήση ουσιών.
- Τους παράγοντες (προστατευτικούς, κινδύνου) που σχετίζονται με τη χρήση ουσιών.
- Τις διαχρονικές μεταβολές μέσω της επανάληψης της έρευνας ανά τετραετία.
- Τα χαρακτηριστικά του τρόπου ζωής των εφήβων όπως οι οικογενειακές και φιλικές σχέσεις, το σχολικό περιβάλλον, ο ελεύθερος χρόνος, κτλ.
- Τη σύγκριση των στοιχείων σε επίπεδο περιφερειών και νομών της χώρας.
- Τη σύγκριση της κατάστασης στη χώρα μας με αυτήν στις άλλες ευρωπαϊκές χώρες που συμμετέχουν στο Πρόγραμμα ESPAD με το οποίο ακολουθείται η ίδια μεθοδολογία.

Το 2011 η έρευνα ESPAD υλοποιήθηκε από το ΕΠΙΨΥ σε συνεργασία και με την υποστήριξη του ΟΚΑΝΑ και των Κέντρων Πρόληψης ΟΚΑΝΑ και της Τοπικής Αυτοδιοίκησης. Περιέλαβε 676 σχολικές μονάδες και συνολικά 37.000 μαθητές και μαθήτριες ηλικίας 13-19 ετών. Περιελήφθησαν στο δείγμα και μικρότερες ηλικίες με τη συμμετοχή σε αυτήν –εκτός της Γ΄ Γυμνασίου και των τριών τάξεων του Λυκείου- οι Α΄ και Β΄ τάξεις του Γυμνασίου. Η συλλογή των στοιχείων έγινε με τη χρήση ανώνυμου, αυτοσυμπληρούμενου ερωτηματολογίου που χορηγήθηκε μέσα στην τάξη, χωρίς την παρουσία εκπαιδευτικού. Η συμμετοχή των σχολείων και των μαθητών στην έρευνα ήταν προαιρετική. Τα αποτελέσματα της έρευνας είναι αντιπροσωπευτικά και στο επίπεδο των 13 περιφερειών της χώρας καθώς και στο επίπεδο των 49 νομών.

## 1.1 Μεθοδολογία

### 1.1.1 Δειγματοληψία

Η δειγματοληψία των σχολείων που συμμετείχαν στην έρευνα έγινε από τους καταλόγους του Υπουργείου Παιδείας. Πρωτογενής δειγματοληπτική μονάδα (PSU)[1] απετέλεσε το σχολικό τμήμα. Οι PSU επιλέχθηκαν με ίσες πιθανότητες. Για την παρούσα έρευνα απαιτήθηκε δείγμα μαθητών όλων των τάξεων του Γυμνασίου και του Λυκείου. Αρχικά ελήφθησαν τυχαία δείγματα τμημάτων από τις τάξεις Γ' Γυμνάσιου και Α' Λυκείου και στη συνέχεια τα τμήματα για τις υπόλοιπες τάξεις επελέγησαν από τα ίδια Γυμνάσια και Λύκεια, αντίστοιχα.

Οι κατάλογοι του Υπουργείου Παιδείας δεν περιείχαν τον αριθμό των μαθητών για κάθε τμήμα των τάξεων για αυτό ο αριθμός αυτός εκτιμήθηκε επί τη βάσει των στοιχείων των προηγουμένων ερευνών. Από τα στοιχεία των προηγούμενων ερευνών εκτιμήθηκε και ο αριθμός των απόντων και αρνήσεων. Με αυτά τα δεδομένα, το απαιτούμενο μέγεθος του δείγματος υπολογίστηκε στους 20 μαθητές ανά τμήμα.

Βάσει των προδιαγραφών της έρευνας ESPAD του 2011, το μέγεθος του δείγματος στην Αττική και στη Θεσσαλονίκη προσδιορίστηκε ως εξής: περίπου 6.000 μαθητές (όλων των 6 τάξεων) στην Αττική και 2.100 στη Θεσσαλονίκη. Με την προϋπόθεση των 20 μαθητών ανά τάξη, θα συμμετείχαν: 51 Γυμνάσια και 51 Λύκεια στην Αττική (102 σχολεία x 3 τμήματα x 20 μαθητές = 6.120), και 17 Γυμνάσια και 17 Λύκεια στη Θεσσαλονίκη (34 x 3 τμήματα x 20 μαθητές = 2.040). Στο επίπεδο των νομών, προκειμένου να εξασφαλιστεί ένα δείγμα ικανοποιητικού μεγέθους για κάθε νομό της χώρας, έγινε στρωματοποίηση ως προς το νομό, δηλαδή, έγινε ξεχωριστή δειγματοληψία των σχολικών τμημάτων σε κάθε νομό. Λαμβάνοντας υπόψη την επίδραση του σχεδιασμού αλλά και τη διόρθωση πεπερασμένου πληθυσμού , το δείγμα ανά νομό (εκτός Αττικής και Θεσσαλονίκης) ορίστηκε στους 600 μαθητές. Αυτό απαιτούσε τη δειγματοληψία 5 Γυμνασίων και 5 Λυκείων σε κάθε νομό (10 σχολεία x 3 τμήματα x 20 μαθητές = 600). Η εκτίμηση του μεγέθους του συνολικού δείγματος ήταν: 6.120 Αττική + 2.040 Θεσσαλονίκη + (49 άλλοι νομοί x 600) = 37.560 μαθητές.

Το ερωτηματολόγιο της έρευνας ήταν ανώνυμο και οι ερωτήσεις αφορούσαν το κάπνισμα, την κατανάλωση οινοπνευματωδών και τη χρήση άλλων νόμιμων και παράνομων εξαρτησιογόνων ουσιών, τις αντιλήψεις σχετικά με την πρόσβαση, τη διαθεσιμότητα και τους κινδύνους από τη χρήση ουσιών, καθώς άλλα θέματα όπως την ψυχοκοινωνική υγεία, δημογραφικά χαρακτηριστικά, αποκλίνουσα συμπεριφορά, σχέσεις στην οικογένεια, σχέσεις με ομότιμους, σχολικό περιβάλλον, δραστηριότητες, ελεύθερο χρόνο και άλλα θέματα σχετικά με τον εφηβικό τρόπο ζωής. Για τους μαθητές της Α' και της Β' Γυμνασίου, δεδομένης της ηλικίας τους και του γεγονότος ότι συμμετείχαν για πρώτη φορά σε έρευνα σχετική με τη χρήση παράνομων ουσιών, το ερωτηματολόγιο διαμορφώθηκε με τέτοιο τρόπο ώστε να είναι πιο σύντομο.

[1] Από τον αγγλικό όρο Primary Sample Unit

### 1.1.2 Δείγμα πληθυσμού στο όποιο υλοποιήθηκε η έρευνα

Σχολικές μονάδες και τμήματα: Στην έρευνα του 2011 συμμετείχαν συνολικά 676 σχολικές μονάδες: 348 Γυμνάσια και 328 Λύκεια από όλη τη χώρα. Σε 62 σχολεία (9,1%) οι Διευθυντές αρνήθηκαν τη συμμετοχή. Ο συνολικός αριθμός των τμημάτων στα οποία χορηγήθηκαν ερωτηματολόγια ήταν 2.050: 1.059 τμήματα Γυμνασίου και 991 τμήματα Λυκείου.

Μαθητές: Το σύνολο των ενεργών μαθητών και μαθητριών του τελικού δείγματος των σχολείων που έλαβαν μέρος στην έρευνα ήταν 42.885: 22.309 μαθητές Γυμνασίου και 20.575 μαθητές Λυκείου . Από αυτούς συμπλήρωσαν το ερωτηματολόγιο 37.040 μαθητές (86,4%). Το υπόλοιπο 13,6% είτε απουσίαζαν την ημέρα της χορήγησης (8,6% ποσοστό επί των ενεργών μαθητών), είτε αρνήθηκαν να συμμετάσχουν (5,5% ποσοστό επί των παρόντων μαθητών).

### 1.1.3 Σύνοψη ευρημάτων.

**Κάπνισμα**

Σύνολο μαθητών 13-19 ετών

- Ένας στους 5 (20%) έχει καπνίσει μέσα στον τελευταίο μήνα.
- Ένας στους 7 (14%) είναι συστηματικός καπνιστής (καπνίζει καθημερινά)
- Καπνίζουν περισσότερα αγόρια από κορίτσια και οι διαφορές αυξάνονται με τη βαρύτητα του καπνίσματος.

Συστηματικό κάπνισμα σε σχέση με την ηλικία

- Τα 13-14 έτη αποτελούν την ηλικία έναρξης καπνίσματος για την μειονότητα (<2%) ενώ ακολουθεί αλματώδης αύξηση έως την ηλικία των 17-18 ετών όπου καπνίζει ένας στους 4 (24,5%) και στην ηλικία των 19 πάνω από τους μισούς (51%).

**Αλκοόλ**

Σύνολο μαθητών 13-19 ετών

- Έξι στους 10 (61%) ήπιαν μέσα στον τελευταίο μήνα.
- Ένας στους 10 (11%) ήπιε με συχνότητα πάνω από 2 φορές την εβδομάδα
- Ένας στους 3 (34%) έχει μεθύσει τουλάχιστον μία φορά στη ζωή του.
- Το 13% έχουν μεθύσει πάνω από 2 φορές στη ζωή τους.
- Το κρασί, η μπύρα, τα ισχυρά ποτά, και τα αλκοολούχα αναψυκτικά είναι τα ποτά με την ευρύτερη κατανάλωση.
- Πίνουν περισσότεροι μαθητές σε Θεσσαλονίκη και άλλα αστικά και ημιαστικά κέντρα σε σύγκριση με την Αθήνα.

Αλκοόλ σε σχέση με την ηλικία

- Κατανάλωση αλκοόλ και μέθη αυξάνονται πολύ με την ηλικία : πάνω από δυο φορές την εβδομάδα πίνει σχεδόν ένας στους 5 εφήβους 17-18 ετών

(18%), ένας στους 4 ηλικίας 19 ετών (26,8%) και έχει μεθύσει πάνω από δύο φορές ένας στους 4 μαθητές 17-18 ετών (22,7%) και ένας στους 3 μαθητές 19 ετών (32,8%).

**Ναρκωτικά**

Σύνολο μαθητών 15-19 ετών

- Ένας στους 7 (15,2%) έχει κάνει χρήση (έστω και μία φορά) κάποιας παράνομης ουσίας, κυρίως κάνναβης. Η πλειονότητά τους έχει επαναλάβει τη χρήση.
- Έχουν κάνει χρήση περισσότερα αγόρια από κορίτσια (αναλογία ~ 2,5 προς 1).
- Υπερτερούν στη χρήση η Αθήνα και η Θεσσαλονίκη σε σύγκριση με τις άλλες περιοχές.

Χρήση κάνναβης σε σχέση με την ηλικία

- Με την ηλικία αυξάνεται σημαντικά ο αριθμός όσων έχουν κάνει χρήση : Στην ηλικία των 17-18 ετών χρήση κάνναβης τον τελευταίο χρόνο έχει κάνει ένας στους 7 (13,6%) και τον τελευταίο μήνα το 8%.
- Χρήση κάνναβης έστω και μία φορά έχει κάνει στην ηλικία των 13-14 ετών το 2,6% των αγοριών και το 0,9% των κοριτσιών (EPIPSI, 2012).

# 2 Δειγματοληψία

Η ιδέα της δειγματοληψίας οφείλετε κυρίως στον Νορβηγό Kiaer (1895) ο οποίος ισχυρίστηκε ότι η μέχρι τότε τακτική της καθολικής απογραφής έπρεπε να αντικατασταθεί με τη μελέτη επιλεγμένων μονάδων του πληθυσμού, μη διασαφηνίζοντας όμως τον τρόπο που θα γινόταν η επιλογή των μονάδων αυτών. Η πιθανότερη εκδοχή είναι αυτής της καθοδηγούμενης επιλογής και όχι της τυχαίας επιλογής που αποτελεί πλέον τη βασική μέθοδο δειγματοληψίας. Η δειγματοληψία πρέπει να γίνεται βάση κάποιων κριτηρίων επιλογής και να υπάρχει κάποιος βαθμός ελέγχου της επιλογής του δείγματος (Ρίτσαρντσον και Βασίλαινας, 1999).

## 2.1 Βασικές έννοιες για τη δειγματοληψία

Η δειγματοληψία αφορά τη λήψη ενός τμήματος από κάποιο ευρύτερο σύνολο. Θεωρείται επιτυχής όταν η επιλογή του δείγματος παράγει αποτελέσματα, δείκτες και μετρήσεις που είναι γενικεύσιμα και όσο το δυνατόν ακριβέστερα σε σχέση με το ευρύτερο σύνολο. Διακρίνονται δύο είδη δειγματοληψίας: Η δειγματοληψία με πιθανότητα και η δειγματοληψία χωρίς πιθανότητα. Η πρώτη γίνεται σύμφωνα με τους νόμους των πιθανοτήτων, είναι ελεγχόμενη ως προς τους παραμέτρους της και δίνει τη δυνατότητα να γενικευτούν τα συμπεράσματα που εξάγονται από ένα δείγμα για αυτό και δίνει επιπλέον τη δυνατότητα να υπολογίσουμε το σφάλμα εκτίμησης της γενίκευσης. Η δεύτερη γίνεται σε περιπτώσεις που δεν είναι εφικτή η δειγματοληψία με πιθανότητα ή όταν ενδιαφέρει να γίνει γρήγορα μια εφαρμογή της έρευνας, λόγου χάρη μια πιλοτική έρευνα. Τα αποτελέσματα δεν είναι γενικεύσιμα ούτε μπορεί να υπολογιστεί το σφάλμα εκτίμησης.

Είναι χρήσιμο να δώσουμε την ορολογία που χρησιμοποιείται στη δειγματοληψία. *Στοιχείο ή μονάδα ή υποκείμενο:* Πρόκειται για τη βασική μονάδα της δειγματοληψίας που αποτελεί και το υποκείμενο της έρευνας. *Πληθυσμός* είναι το ευρύ σύνολο των υποκειμένων για το οποίο εξάγουμε συμπεράσματα. *Ο πληθυσμός* της έρευνας είναι το τμήμα του ευρύτερου πληθυσμού που μπορεί να συμπεριληφθεί στην έρευνα, δηλαδή τα στοιχεία που είναι υποψήφια στην έρευνα. Αποτελείται δηλαδή από στοιχεία που είναι υποψήφια για να επιλεγούν στο σχηματισμό δείγματος. *Δειγματοληπτικό πλαίσιο* είναι ένας κατάλογος ολόκληρου ή σχεδόν ολόκληρου του πληθυσμού της έρευνας. *Δειγματοληπτική μονάδα* είναι το στοιχείο ή η ομάδα στοιχείων που μπορεί να επιλεγεί σε κάποιο στάδιο της δειγματοληψίας. Αν η δειγματοληψία εκτελείται σε στάδια και επιλέγουμε πρώτα μια ομάδα στοιχείων και κατόπιν κάποια στοιχεία μέσα από την ομάδα, τότε η ομάδα είναι η δειγματοληπτική μονάδα. *Μέγεθος δείγματος:* Πρόκειται για το πλήθος των στοιχείων που διαμορφώνουν το δείγμα και συμβολίζεται με το αγγλικό

γράμμα n. Το μέγεθος δεν χρειάζεται να είναι κάποιο συγκεκριμένο ποσοστό του μεγέθους του πληθυσμού. Αντίθετα, προσδιορίζεται σε σχέση με το μέγεθος του σφάλματος εκτίμησης που παράγεται εξαιτίας της χρήσης δειγματοληψίας, της μεθόδου δειγματοληψίας, της σχετικής ομοιογένειας ή ανομοιογένειας του πληθυσμού και του κόστους δειγματοληψίας.

## 2.2 Απλή τυχαία δειγματοληψία

Στην απλή τυχαία δειγματοληψία κάθε μέλος του πληθυσμού έχει την ίδια πιθανότητα να επιλεγεί για το σχηματισμό του δείγματος με κάθε άλλο μέλος του πληθυσμού. Για να εφαρμοστεί η απλή τυχαία δειγματοληψία τα στοιχεία του στατιστικού πληθυσμού θα πρέπει να είναι καταγραμμένα σε έναν κατάλογο, που χρησιμεύει ως δειγματοληπτικό πλαίσιο. Με δεδομένη αυτή την πληροφορία, γίνεται αντιστοίχιση αριθμών στα μέλη του καταλόγου, όταν δεν υπάρχει ήδη. Κατόπιν επιλέγονται με τυχαίο τρόπο μέλη από τον κατάλογο μέχρι να σχηματιστεί πλήθος ίσο με το μέγεθος του δείγματος που επιθυμούμε να έχουμε. Όλα τα μέλη της λίστας έχουν την ίδια πιθανότητα να επιλεγούν επιπλέον απαιτείται η ανεξαρτησία των επιλογών. Αν ένα στοιχείο εκλεγεί, δεν μπορεί να επανατεθεί στη λίστα, δηλαδή δεν επιτρέπεται να επανεκλεγεί γνωστό και σαν δειγματοληψία χωρίς αντικατάσταση. Εάν η δειγματοληψία γίνεται με αντικατάσταση, μπορεί να εμφανιστεί ένα στοιχείο περισσότερες από μία φορές σε ένα συγκεκριμένο δείγμα. Στην πράξη όλες οι δειγματοληψίες γίνονται χωρίς αντικατάσταση (Levy και Lemeshow, 1991).

Η χρήση απλής τυχαίας δειγματοληψίας δεν οδηγεί αυτομάτως στη δημιουργία αντιπροσωπευτικών δειγμάτων. Το δείγμα μπορεί να αφήνει περιοχές του πληθυσμού ακάλυπτες και τίποτε δεν εξασφαλίζει ότι υπάρχει αντιπροσωπευτικότητα ως προς τα χαρακτηριστικά που μας ενδιαφέρουν. Η χρήση απλής τυχαίας δειγματοληψίας παρουσιάζει ευκολία, ενώ τα περισσότερα προγράμματα στατιστικής επεξεργασίας στον Η/Υ προϋποθέτουν ότι τα προς επεξεργασία δεδομένα προέρχονται από απλή τυχαία δειγματοληψία, και με αυτή την προϋπόθεση υπολογίζονται τα σφάλματα εκτίμησης. Όταν οριστεί το δειγματοληπτικό πλαίσιο, δεν είναι απαραίτητη καμιά άλλη πληροφορία. Χρειαζόμαστε όμως πάντα ένα αναλυτικό δειγματοληπτικό πλαίσιο, πράγμα που προϋποθέτει γνώση και αναλυτική καταγραφή του πληθυσμού (Πασχαλούδης και Ζαφειρόπουλος, 2002). Η απλή τυχαία δειγματοληψία δεν έχει το μικρότερο σφάλμα εκτίμησης άρα δεν δίνει τις πιο ακριβείς προβλέψεις, σε αντίθεση με τη στρωματοποιημένη δειγματοληψία.

## 2.3 Στρωματοποιημένη δειγματοληψία

Όταν ο πληθυσμός μπορεί να διαιρεθεί εκ των προτέρων σε στρώματα ομοιογενή ως προς το χαρακτηριστικό που μας ενδιαφέρει τότε μπορούμε να πάρουμε ένα τυχαίο δείγμα από κάθε στρώμα, οπότε το συνολικό δείγμα που

προκύπτει αντιπροσωπεύει όλα τα στρώματα. Η μέθοδος αυτή ονομάζεται στρωματοποιημένη δειγματοληψία. Η στρωματοποίηση μπορεί να γίνει κατά φύλο, κατά ηλικία, κατά επάγγελμα κτλ. Η στρωματοποιημένη δειγματοληψία συνίσταται ιδιαίτερα στις περιπτώσεις που υπάρχει διαφορετικό πλαίσιο για κάθε στρώμα. Όταν το μέγεθος του τυχαίου δείγματος από κάθε στρώμα είναι κατ' αναλογία με το σχετικό μέγεθος του στρώματος προς το μέγεθος του πληθυσμού τότε έχουμε την περίπτωση της αναλογικά στρωματοποιημένης δειγματοληψίας. Σε μερικές περιπτώσεις η στρωματοποιημένη δειγματοληψία επιβάλλεται να γίνει μη αναλογική. Τέτοια είναι η περίπτωση στην οποία υπάρχουν σημαντικές διαφορές στη διασπορά των στρωμάτων, οπότε φυσικά επιθυμούμε να πάρουμε μεγαλύτερο δείγμα από το στρώμα με τη μεγαλύτερη διασπορά, στην περίπτωση αυτή επιβάλλεται διαφορετική στάθμιση των πληροφοριών ανά στρώμα. Όταν οι διαφορές μεταξύ των στρωμάτων είναι μικρές συγκρινόμενες με τις διαφορές μέσα στα στρώματα τότε η στωματοποιημένη δειγματοληψία δεν αναμένεται να μας δώσει ακριβέστερη εικόνα του πληθυσμού από την απλή τυχαία δειγματοληψία (Ζαχαροπούλου, 2009).

## 2.4 Συστηματική δειγματοληψία

"... Η συστηματική δειγματοληψία, είτε από μόνη της είτε σε συνδυασμό με κάποια άλλη μέθοδο, μπορεί να θεωρηθεί ως η πιο ευρέως διαδεδομένη μέθοδος δειγματοληψίας. " (Levy και Lemeshow, 1999)

Στην περίπτωση που ένα δειγματοληπτικό πλαίσιο είναι διαθέσιμο σε μορφή λίστας, μπορούμε να εφαρμόσουμε συστηματική δειγματοληψία. Έστω ότι ο υπό μελέτη πληθυσμός έχει μέγεθος N που είναι καταγραμμένα σε μια λίστα και φέρουν αρίθμηση με αύξοντα αριθμό. Διαιρούμε το σύνολο των στοιχείων του δειγματοληπτικού πλαισίου, δηλαδή το μέγεθος του πληθυσμού, με το μέγεθος του δείγματος n. Το αποτέλεσμα στρογγυλοποιημένο είναι το βήμα επιλογής των υποκειμένων της έρευνας. Ξεκινάμε επιλέγοντας έναν τυχαίο αριθμό ανάμεσα στο ένα και το N/n, έστω x. Το άτομο που αντιστοιχεί στο συγκεκριμένο αύξοντα αριθμό είναι το πρώτο στοιχείο του δείγματος. Στη συνέχεια επιλέγεται το άτομο με αύξοντα αριθμό x+N/n, μετά το άτομο με αριθμό x+2N/n κ.ο.κ. Το τελευταίο άτομο που θα επιλεγεί θα είναι x+(n-1)N/n. Επιλέγονται έτσι n στοιχεία από το δειγματοληπτικό πλαίσιο. Για να επιτευχθεί αναλογική αντιπροσώπευση του πληθυσμού στο δείγμα, καλό είναι να έχει προηγηθεί ταξινόμηση της λίστας του δειγματοληπτικού πλαισίου ως προς το χαρακτηριστικό που θεωρείται καίριο για τη στρωματοποίηση του πληθυσμού. Αυτό συνήθως είναι κάποιο δημογραφικό χαρακτηριστικό. Η συστηματική δειγματοληψία επιπλέον μοιάζει με την απλή τυχαία δειγματοληψία όταν δεν υπάρχει συγκεκριμένη σειρά στο δειγματοληπτικό πλαίσιο ή δεν υπάρχει κάποιου είδους εποχικότητα στα δεδομένα που απεικονίζεται και στο δειγματοληπτικό πλαίσιο. Η συστηματική δειγματοληψία

μπορεί να αποτελέσει στάδιο ευρύτερης δειγματοληπτικής διαδικασίας, συνήθως το τελευταίο στάδιο. Μοιάζει με την απλή τυχαία δειγματοληψία γιατί δίνει δείγμα με ίσες πιθανότητες όχι όμως με ανεξάρτητες επιλογές. Με τη σωστή εφαρμογή της μεθόδου επιτυγχάνεται η επιλογή αντιπροσωπευτικού δείγματος και γενικά τηρούνται οι ποσοστώσεις διαφόρων χαρακτηριστικών που ενδιαφέρουν τον ερευνητή.

## 2.5 Κατά συστάδες δειγματοληψία

Οι μικρότερες μονάδες εντός του οποίου ένας πληθυσμός μπορεί να διαιρεθεί ονομάζονται στοιχεία του πληθυσμού, και ομάδες των στοιχείων αυτών συστάδες. Η μέθοδος δειγματοληψίας κατά συστάδες ενδείκνυται για έρευνες που αφορούν ευρείες γεωγραφικές περιοχές με διασπορά. Αφορούν επίσης περιπτώσεις στις οποίες δεν είναι απαραίτητα γνωστός και καταγραμμένος ο πληθυσμός αλλά υπάρχουν διαθέσιμοι κατάλογοι-λίστες με ομάδες που καλύπτουν τον πληθυσμό, δηλαδή με μονάδες του πληθυσμού. Αυτές συνήθως είναι ισοπληθείς γιατί εκφράζουν μονάδες ομοίων κατηγοριών για παράδειγμα σχολικές τάξεις. Έτσι, από τη λίστα των ομάδων του πληθυσμού επιλέγουμε δείγμα κάποιων ομάδων και, στη συνέχεια, όλα τα μέλη των ομάδων που επιλέχτηκαν απαντούν στο ερωτηματολόγιο και άρα συνιστούν το τελικό δείγμα της έρευνας.

> Το πρόβλημα με τη μέθοδο της τυχαίας δειγματοληψίας είναι ότι κατά τη δειγματοληψία ενός πληθυσμού που είναι κατανεμημένο σε μια ευρεία γεωγραφική περιοχή έγκειται στο να καλύψει ένα μεγάλο γεωγραφικό τμήμα, προκειμένου να πάρει από κάθε μία από τις μονάδες του δείγματος. Αυτή η γεωγραφική κάλυψη είναι δαπανηρή υπόθεση στο να μπορεί να γίνει η συλλογή των δειγμάτων. Αλλά, εάν δε λαμβάνονται δείγματα από ολόκληρο τον πληθυσμό ανά περιοχή, μπορεί τα συμπεράσματα που θα βγουν από την έρευνα να μην είναι σωστά. Το δύσκολο είναι να προσδιοριστεί το βέλτιστο μέγεθος της συστάδας όταν το κόστος της έρευνας να είναι συγκεκριμένο (Raj, 1972: 144–145).

Αυτή η δυσχέρεια μπορεί να λυθεί εάν το κόστος της έρευνας και η διακύμανση της εκτίμησης μπορούν να εκφραστούν ως συνάρτηση του μεγέθους της συστάδας. Σύμφωνα με τον Trochim η δειγματοληψία κατά συστάδες περιλαμβάνει :

- Το διαχωρισμό του πληθυσμό σε συστάδες.
- Τυχαία δείγματα συστάδων.
- Τη μετρήση.

Η δειγματοληψία κατά συστάδες υλοποιείται συνήθως για πρακτικούς λόγους και για να μειωθεί το κόστος της έρευνας. Σε σύγκριση με τις παραπάνω μεθόδους παρουσιάζει το μεγαλύτερο δειγματοληπτικό σφάλμα, δηλαδή τη χαμηλότερη ακρίβεια για δεδομένο μέγεθος του δείγματος. Ο λόγος που συμβαίνει αυτό είναι επειδή οι μονάδες που ανήκουν στη ίδια ομάδα παρουσιάζουν πολλά κοινά χαρακτηριστικά (Levy και Lemeshow 1991: 180). Οπότε αν οι μονάδες μέσα σε

μια συστάδα μοιάζουν περισσότερο μεταξύ τους από ότι οι μονάδες που ανήκουν σε διαφορετικές ομάδες, το τυπικό σφάλμα της εκτίμησης θα είναι μεγαλύτερο. Συμπεραίνουμε έτσι ότι όσο μικρότερη είναι η συσχέτιση εντός μιας ομάδας, τόσο το καλύτερο (Raj, 1972: 144).

Αν, κατά τα διάφορα στάδια της δειγματοληψίας, οι σύνθετες μονάδες επιλέγονται με απλή τυχαία δειγματοληψία, το δειγματοληπτικό σχήμα ονομάζεται απλή δειγματοληψία κατά ομάδες σε ένα ή περισσότερα στάδια. Όταν οι σύνθετες μονάδες είναι ομάδες στοιχείων βασισμένες σε γεωγραφικές περιοχές, ο σχεδιασμός ονομάζεται δειγματοληψία κατά περιοχές.

Δειγματοληψία κατά ομάδες σε ένα στάδιο ονομάζεται η δειγματοληπτική τεχνική, η οποία διαιρεί τις στοιχειώδεις μονάδες του πληθυσμού σε ομάδες, επιλέγει ένα δείγμα των ομάδων αυτών και περιλαμβάνει στο τελικό δείγμα των στοιχειωδών μονάδων όλες τις στοιχειώδεις μονάδες που ανήκουν στις ομάδες αυτές.

Δειγματοληψία κατά ομάδες σε δύο στάδια. Αρχικά γίνεται καταγραφή όλων των συστάδων του πληθυσμού, συνήθως η επιλογή γίνεται με απλή τυχαία δειγματοληψία, οι μονάδες που βρίσκονται στις επιλεγμένες ομάδες του πρώτου σταδίου περνάνε στο δεύτερο στάδιο, συνήθως με απλή τυχαία δειγματοληψία ή με τη χρήση της συστηματικής δειγματοληψίας.

Δειγματοληψία κατά ομάδες σε πολλά στάδια ονομάζεται η δειγματοληπτική τεχνική όταν, μετά το πρώτο στάδιο, επιλέγονται δείγματα μικρότερων και μικρότερων ομάδων με τελικό στάδιο την επιλογή δείγματος στοιχειωδών μονάδων ή την περίληψη όλων των στοιχειωδών μονάδων της τελευταίας κατηγορίας σύνθετων ομάδων (Saifuddin, 2009). Στην πράξη, οι συστάδες είναι επίσης στωματοποιημένη δειγματοληψία.

Επιπλέον πρωτοβάθμιες μονάδες δειγματοληψίας θεωρούνται οι ομάδες (PSU) και δευτεροβάθμιες δειγματοληπτικές μονάδες θεωρούνται τα νοικοκυριά / επιμέρους στοιχεία (SSU)[2]. Μπορούμε να επιλέξουμε τη PSU, χρησιμοποιώντας ειδικές τεχνικές δειγματοληψίας, όπως η απλή τυχαία δειγματοληψία ή συστηματική δειγματοληψία ή PPS. Οι SSU επιλέγονται για λόγους ευκολίας ή μερικά με τη χρήση συγκεκριμένου τεχνικών δειγματοληψίας όπως τη PSU.

[2] Από τον αγγλικό όρο Secondary Sample Unit

# 3 Εκτίμηση για δειγματοληψία κατά συστάδες.

Θωρούμε ότι τα στοιχεία του πληθυσμού είναι συγκεντρωμένα σε N ομάδες (PSU). Το μέγεθος μιας ομάδας είναι $M_i$ για $i$ ομάδα. Και ο αντίστοιχος αριθμός των ομάδων σε ένα δείγμα είναι n, και ο αριθμός των στοιχείων μιας $i$ είναι $m_i$. Εάν το $y_{ij}$ είναι η εκτίμηση του j για στοιχείο (SSU) και του $i$ για συστάδα (PSU). Στην απλή περίπτωση όπου οι συστάδες είναι ίσου μεγέθους (αν και μπορεί να είναι μη ρεαλιστική), ο συνολικός αριθμός των στοιχείων του πληθυσμού, $K = N * M$ , όπου $M_i$ = M (σταθερά για όλες τις συστάδες). Εάν οι συστάδες είναι άνισες, ο συνολικός αριθμός των στοιχείων του πληθυσμό είναι :

$$K = \sum_{i=1}^{N} M_i$$

- Σύνολο του πληθυσμού i :

$$t_i = \sum_{j=1}^{M_i} y_{ij}$$

- Εκτιμώμενο συνολικό δείγμα για το i-PSU:

$$\hat{t}_i = \sum_{j \in S_i} M_i \frac{y_{ij}}{m_i} = \sum_{j \in S_i} M_i \, \bar{y}_i$$

- Συνολικός πληθυσμός :

$$t = \sum_{i=1}^{N} t_i = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}$$

- Εκτιμώμενο συνολικό δείγμα για τον πληθυσμό

$$\hat{t} = \sum_{j \in S_i} t_i$$

- Αμερόληπτη (unbiased) εκτίμηση για τον συνολικό πληθυσμό :

$$\hat{t}_{unb} = \frac{N}{n} \sum_{j \in S_i} t_i$$

- Μέση τιμή του πληθυσμού στο cluster i :

$$\bar{Y}_{i,clu} \sum_{j=1}^{M_i} \frac{y_{ij}}{M_i} = \frac{t_i}{M_i}$$

- Μέση τιμή δείγματος για i PSU:

$$\bar{y}_{i,clu} = \sum_{j \in S_i} \frac{y_{ij}}{m_i} = \frac{\hat{t}_i}{m_i}$$

- Ο μέσος του πληθυσμού :

$$\bar{y}_{clu} = \frac{1}{K} \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}$$

- Μέση τιμή δείγματος ( unbiased)

$$\hat{\bar{y}}_{clu} = \frac{\hat{t}}{\sum_{i \in S} m_i}$$

## 3.1 Εκτίμηση της διακύμανσης.

$$\hat{t}_{unb} = \frac{N}{n} \sum_{j \in S_i} t_i = N \frac{\sum_{j \in S_i} t_i}{n}$$

$$= N\bar{y}_{total}$$ , όπου $\bar{y}$ είναι ο μέσος ''total''για το cluster

- Οπότε η διακύμανση είναι[3] :

$$var(\hat{t}_{unb}) = N^2 \frac{S_t^2}{n} \left(1 - \frac{n}{N}\right) \text{ όπου,}$$

$$S_t^2 = \frac{\sum_{i=1}^{N} \left(t_i - \frac{t}{N}\right)^2}{N-1}$$

- Ο μέσος για συστάδες ίσου μεγέθους:

$$\hat{\bar{y}}_{clu} = \frac{\hat{t}}{NM}$$, επειδή στα ίδιου μεγέθους $M_i = m_i = M$

- Οπότε η διακύμανση του μέσου είναι :

$$var(\hat{\bar{y}}) = \frac{1}{N^2 M^2} var(\hat{t}) = \frac{N^2}{N^2} \frac{S_t^2}{nM^2} \left(1 - \frac{n}{N}\right) = \frac{S_t^2}{nM^2} \left(1 - \frac{n}{N}\right)$$

[3] Η διακύμανση του συνόλου ενδέχεται να είναι μεγαλύτερη όταν οι συστάδες είναι άνισες μεταξύ τους.

## 3.2 Η συσχέτιση εντός μιας ομάδας

Η συσχέτιση των μονάδων που υπάρχει μέσα σε μία ομάδα αντανακλά την ομοιογένεια του δείγματος. Σε μια κατά συστάδες δειγματοληψία επιλέγοντας μια επιπλέον μονάδα από την ίδια συστάδα προσθέτει λιγότερες νέες πληροφορίες από το αν η επιλογή γινόταν τυχαία. Για παράδειγμα σε μια δειγματοληψία κατά ομάδες σε ένα στάδιο, το δείγμα δεν παρουσιάζει τόσο μεγάλη ποικιλομορφία όσο θα παρουσίαζε σε μια απλή τυχαία δειγματοληψία, αυτό σημαίνει ότι το πραγματικό μέγεθος του δείγματος από το οποίο θα εξάγουμε αποτελέσματα είναι μειωμένο (Levy και Lemeshow, 1999). Το γεγονός ότι χάνεται η αποτελεσματικότητα λόγω του ότι ακολουθήθηκε η μέθοδος της κατά συστάδες δειγματοληψία αντί της απλής τυχαίας δειγματοληψίας δημιουργεί το πρόβλημα της επίδρασης του δειγματοληπτικού σχεδιασμού. Το οποίο βασικά είναι ο λόγος της πραγματικής διακύμανσης για τη δειγματοληπτική μέθοδο που έχουμε χρησιμοποιήσει προς τη διακύμανση η οποία έχει υπολογιστεί με βάση την παραδοχή της απλής τυχαίας δειγματοληψίας (CENSUS, 2002). Στην πράξη, οι μονάδες εντός μιας (PSU) ομάδας τείνουν να είναι κάπως παρόμοιες η μια με την άλλη για όλες σχεδόν τις μεταβλητές, αν και ο βαθμός ομοιότητας είναι συνήθως χαμηλός. Ως εκ τούτου, η συσχέτιση εντός μιας ομάδας (ρ) είναι σχεδόν πάντα θετική και παίρνει μικρή τιμή (Kalton, Brick και Le, 2005: 105). Δεδομένου αυτού μπορούμε να αναλύσουμε τη διακύμανση σε : διακυμανση (κατά συστάδες δειγματοληψια)

$$\boldsymbol{\sigma^2 = \sigma_w^2 + \sigma_b^2}, \ \boldsymbol{\tau o\, o\pi o\acute{\iota} o\, \sigma\eta\mu a\acute{\iota} \nu \varepsilon\iota\, \Sigma \upsilon \nu o\lambda \iota \kappa \acute{\eta}\, \Delta \iota a \kappa \acute{\upsilon} \mu a \nu \sigma \eta}$$
$$\boldsymbol{= \delta \iota a \kappa \acute{\upsilon} \mu a \nu \sigma \eta\, \mu \acute{\varepsilon} \sigma a (\, within) \sigma \tau \eta \nu\, o \mu \acute{a} \delta a}$$
$$\boldsymbol{+ \delta \iota a \kappa \acute{\upsilon} \mu a \nu \sigma \eta\, \mu \varepsilon \tau a \xi \acute{\upsilon} (between\,) \tau \omega \nu\, o \mu a \delta \acute{\omega} \nu}$$

- Η συσχέτιση εντός μιας ομάδας ορίζεται ως εξής:

$$\boldsymbol{\rho = 1 - \frac{\sigma_w^2}{\sigma^2} = \frac{\sigma_b^2}{\sigma^2} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}}$$

- Πιο συγκεκριμένα :

$$\boldsymbol{\rho = 1 - \frac{n}{n-1} \frac{\sigma_w^2}{\sigma^2}}$$

Ελάχιστη όταν $\sigma_b^2 = 0$ , $\rho = -1 \frac{}{n-1}$

Μέγιστη όταν $\sigma_w^2 = 0$ , $\rho = 1$

## 3.3 Η επίδραση του δειγματοληπτικού σχεδιασμού ( design effect )

Τα δεδομένα που συλλέγονται από σύνθετες δειγματοληπτικές έρευνες που περιλαμβάνουν πολλαπλά στάδια όπως η στρωματοποίηση των πληθυσμών και η ομαδοποίηση των δειγματοληπτικών μονάδων συνήθως απαιτούν από τον ερευνητή να ενσωματώσει στην ανάλυση συντελεστές στάθμισης, τα δειγματοληπτικά βάρη και σχεδιαστικές μεταβλητές οι οποίες θα βοηθήσουν στο να μη βγαίνουν λάθος συμπεράσματα, ιδίως στην εκτίμηση της διακύμανσης. Στις περισσότερες περιπτώσεις, ο δειγματοληπτικός σχεδιασμός αντανακλά στις στρατηγικές που είναι κάτι περισσότερο από την απλή τυχαία δείγματα του πληθυσμό, εν μέρει επειδή οι τελικές δειγματοληπτικές μονάδες είναι ομαδοποιημένες σε πρωτογενής δειγματοληπτικές μονάδες που μπορεί να είναι στρωματοποιημένες εντός του ενδιαφερόμενου πληθυσμό. Η χρήση σύνθετων δειγματοληπτικών μεθόδων έχει ως συνεπεία να αυξάνεται η επίδραση του δειγματοληπτικού σχεδιασμού.

$$\boldsymbol{Deff} = \frac{\textbf{εκτίμηση της διακύμανσης (συνθετη δειγματοληπτική μέθοδος)}}{\textbf{εκτίμηση της διακύμανση (απλής τυχαίας δειγματοληψίας)}}$$

Η επίδραση του δειγματοληπτικού σχεδιασμού μιας έρευνας μπορεί να χρησιμοποιηθεί ως εργαλεία για να υπολογιστεί η αποτελεσματικότητα του δείγματος και για να πραγματοποιηθεί ο σωστότερος σχεδιασμός μιας έρευνας(Park και Lee, 2001). Σύμφωνα με τον Kish (1965) η επίδραση του δειγματοληπτικού σχεδιασμού ορίζεται ως ο λόγος της εκτίμησης της διακύμανσης εφαρμόζοντας κάποια πολύπλοκη δειγματοληπτική μέθοδο προς την εκτίμησης της διακύμανση ένος δείγματος του ίδιου μεγέθους στην οποία όμως εφαρμόζετε η μέθοδος της απλής τυχαίας δειγματοληψίας. Πολύπλοκη δειγματοληπτική μέθοδος μπορεί να θεωρηθεί η διαστρωμάτωση, η ομαδοποίηση και η άνιση στάθμιση. Η αποδοτικότητα κάποιας σύνθετης δειγματοληπτικής μεθόδου μπορεί να αξιολογηθεί για κάθε χαρακτηριστικό που αφορά σε επίπεδο σχεδιασμού μέσω της αποσύνθεσης της επίδρασης του δειγματοληπτικού σχεδιασμού. Αν για παράδειγμα η επίδραση του δειγματοληπτικού σχεδιασμού από την ομαδοποίηση είναι πολύ μεγάλη για ορισμένες εκτιμήσεις μιας έρευνας, τότε μπορούμε να εξετάσουμε τι επιλογές έχουμε για να μειωθεί η επίδραση αυτή. Αν τα οφέλη που προσδίδει η στρωματοποίηση του δείγματος στην ακρίβεια των εκτιμήσεων είναι αμελητέα, τότε μπορούμε να ενισχύσουμε περαιτέρω το ισχύον σύστημα στρωματοποίησης προκειμένου να αποκτήσουν μεγαλύτερο όφελος από τη στρωματοποιημένη δειγματοληψία. Ο τελικός στόχος είναι να μειωθεί η επίδραση του δειγματοληπτικού σχεδιασμού στις βασικές εκτιμήσεις και να διατηρηθούν τα ακριβή αποτελέσματα της δειγματοληψίας έχοντας βέβαια το δείγμα που έχει μικρότερο δυνατό μέγεθος (Park κ.α, 2003).

- Παραγώγιση της διακύμανσης για κατά συστάδες δειγματοληψία.

$$\rho = 1 - \frac{n}{n-1}\frac{\sigma_w^2}{\sigma^2}$$

$$\rho = \frac{(n-1)\sigma^2 - n\sigma_w^2}{(n-1)\sigma^2}$$

$$=> n\sigma^2 - \sigma^2 - n\left(\sigma^2 - \sigma_b^2\right) = \sigma^2(n-1)\rho$$

$$=> n\sigma_b^2 = \sigma^2 + \sigma^2(n-1)\rho$$

$$=> \sigma_b^2 = \frac{\sigma^2}{n}\left[1 + (n-1)\rho\right]$$

$$\boldsymbol{var(\bar{x}) = \sigma_b^2 = \frac{\sigma^2}{n}\left[1 + (n-1)\rho\right]}$$

Ας εξετάσουμε την περίπτωση δειγματοληψίας κατά ομάδες σε ένα στάδιο, όπου n μονάδες του δείγματος επιλέγονται από N ομάδες και το (μέσο) μέγεθος της ομάδας είναι M, τότε η διακύμανση του y είναι:

$$\boldsymbol{Var_{clu}(\bar{y}) = \frac{\sigma_x^2}{nM}\left[1 + (M-1)\rho\right]}$$

Επίδραση δειγματοληπτικού σχεδιασμού

$$\boldsymbol{Deff = 1 + (M-1)\rho}$$

Στην κατά συστάδες δειγματοληψία, το μέγεθος του ρ, μπορεί να είναι αρκετά μεγάλο, αυτό μπορεί να επηρεάσει σημαντικά την ακρίβεια των εκτιμήσεων (Saifuddin 2009) . Σε γενικές γραμμές, καθώς αυξάνεται το μέγεθος της ομάδας το ρ μειώνεται, αλλά το deff εξαρτάται και από το M, το μέγεθος του δείγματος εντός των επιλεγμένων PSU και από το ρ, τη συσχέτιση εντός ομάδας. Καθώς το ρ είναι γενικά θετικό, η επίδραση του δειγματοληπτικού σχεδιασμού από την ομαδοποίηση είναι, κατά κανόνα, μεγαλύτερη από το 1 (Kalton κ.α, 2005¨: 105). Οπότε στη κατά συστάδες δειγματοληψία το να μεγαλώνεις το μέγεθος της ομάδας για να εξομαλυνθεί η ομοιογένεια δεν είναι λύση αντίθετα καθιστά τη δειγματοληψία πιο αναποτελεσματική. Για παράδειγμα, για μια ομάδα 20 μονάδων, αν ρ = 0,1, το $\text{deff} = 1 + (20-1) * 0{,}1 = 2{,}9$ υποδηλώνοντας ότι η πραγματική διακύμανση είναι 2.9 φορές μεγαλύτερη από τη διακύμανση της απλής τυχαίας δειγματοληψίας (SRS) για το ίδιο μέγεθος δείγματος. Ωστόσο, εάν το μέγεθος της ομάδας είναι μεγάλο, δηλαδή m = 200, deff = 1 + (200-1) * 0,1 = 20,9. Οπότε όταν ρ = 0.0, deff = 1 (Saifuddin 2009).

Για να γίνει σωστός δειγματοληπτικός σχεδιασμός οι εκτιμήσεις του ρ σχετικά με τις βασικές μεταβλητές της έρευνας είναι αναγκαίες. Οι εκτιμήσεις αυτές συνήθως βασίζονται σε εκτιμήσεις από προηγούμενες έρευνες για ίδιες ή

παρόμοιες μεταβλητές και PSUs, και η πεποίθηση ότι τα ίδια συμπεράσματα ισχύουν για παρόμοιες μεταβλητές και PSUs . Σε πραγματικές συνθήκες, το PSU δεν μπορεί να έχουν το ίδιο μεγέθος και δεν πραγματοποιείται δειγματοληψία με τη μέθοδο της απλής τυχαίας δειγματοληψίας. Στα περισσότερα δειγματοληπτικά σχέδια που πραγματοποιούνται σε εθνικό επίπεδο επιλέγεται η σωματοποιημένη δειγματοληψία PSU που επιλέγονται χρησιμοποιώντας PPES (ανάλογη πιθανότητα προς το εκτιμώμενο μέγεθος) δειγματοληψία. Ως αποτέλεσμα, η εξίσωση $Deff = 1 + (M - 1)\rho$ δεν εφαρμόζεται άμεσα. Ωστόσο, εξακολουθεί η ερμηνεία του ρ να χρησιμεύει ως ένα χρήσιμο μοντέλο για την επίδραση του δειγματοληπτικού σχεδιασμού από την ομαδοποίηση του δείγματος για μια ποικιλία δειγματοληπτικών σχεδίων εφαρμόζοντας βέβαια κατάλληλές τροποποιήσεις (Kalton κ.α, 2005).

Μπορούμε να πούμε ότι η επίδραση του δειγματοληπτικού σχεδιασμού μεγαλώνει καθώς μεγαλώνει το μέγεθος μιας ομάδας, και καθώς μεγαλώνει η συσχέτιση των μονάδων εντός μιας ομάδας. Η συσχέτιση μέσα σε μία ομάδα αντιπροσωπεύει την πιθανότητα δύο στοιχεία (μονάδες) εντός της ίδιας ομάδας να έχουν την ίδια τιμή για ένα δεδομένο στατιστικό στοιχείο σε σχέση με δύο στοιχεία που επιλέγονται εντελώς τυχαία. Μια τιμή της τάξεως του 0,05 ερμηνεύεται ως εξής: δύο στοιχεία εντός της ομάδας έχουν 5% μεγαλύτερη πιθανότητα να έχουν την ίδια τιμή από ό,τι αν τα δύο στοιχεία έχουν επιλεγεί τυχαία για την έρευνα. Όσο μικρότερη είναι η τιμή, τόσο καλύτερη θα είναι η συνολική αξιοπιστία της εκτίμησης του δείγματος (Turner, 1996).

Το design effect διαφέρει από έρευνα σε έρευνα, ακόμα και μέσα στην ίδια έρευνα, μπορεί να διαφέρει από ερώτηση σε ερώτηση. Για παράδειγμα, "οι ερωτηθέντες που ζουν κοντά μεταξύ τους είναι πιθανό να έχουν παρόμοια χαρακτηριστικά στο βιοτικό επίπεδο, αλλά δεν είναι πιθανό να έχουν παρόμοια χαρακτηριστικά στην αναπηρία'' (Alecxih, Corea και Marker, 1998). Οι ερευνητές χρησιμοποιούν επίσης τον όρο design factor (DEFT) [4], ο οποίος όρος αντιπροσωπεύει τη τετραγωνική ρίζα του design effect. Παίρνοντας την τετραγωνική ρίζα του DEFF και το μέσο όρο αυτών των τιμών, η μεταβλητότητα μειώνεται κάπως. Το DEFT μπορεί επίσης να χρησιμοποιηθεί για την άμεση εκτίμηση των διαστημάτων εμπιστοσύνης (Flores-Cervantes, Brick, και DiGaetano, 1999). Δείχνει ωστόσο πόσο μεγάλο είναι το τυπικό σφάλμα του δείγματος, και, κατά συνέπεια, πόσο αυξάνονται τα διαστήματα εμπιστοσύνης. Για παράδειγμα, εάν το DEFT είναι 3, το διαστήματα εμπιστοσύνης θα είναι 3 φορές ευρύτερο από όσο θα ήταν για ένα απλό τυχαίο δείγμα.

Εν ολίγοις, χρησιμοποιώντας τη κατά συστάδες δειγματοληψία γενικά απαιτεί είτε ένα μεγαλύτερο σε μέγεθος δείγμα από ότι στην απλή τυχαία δειγματοληψία, είτε ένα ευρύτερο διάστημα εμπιστοσύνης. Το design effect

[4] Υπάρχει κάποια σύγχυση σχετικά με την ορολογία. Για παράδειγμα, Ukoumunne et al το ονομάζουν DEFF, ενώ ο Turner ονομάζει DEFT.

χρησιμοποιείται για να καθορίσει πόσο μεγάλο πρέπει είναι το μέγεθος του δείγματος ή ποια πρέπει να τα διαστήματα εμπιστοσύνης (Shackman, 2001)

# 4 Taylor linearization – Jackknife μέθοδος για τον υπολογισμό δειγματοληπτικών σφαλμάτων και του design effect.

Η εκτίμηση της διακύμανσης μιας στατιστικής έρευνας περιπλέκεται όχι μόνο από το πόσο πολύπλοκος είναι ο τρόπος που σχεδιάστηκε η δειγματοληψία αλλά και από το στατιστικό μοντέλο που θα εξεταστεί. Ακόμα και όταν ο σχεδιασμός της έρευνα που υλοποιείται με τη μέθοδο της απλής τυχαίας δειγματοληψίας η εκτίμηση της διακύμανσης κάποιων στατιστικών μοντέλων απαιτεί εξειδικευμένες τεχνικές εκτιμήσεων. Για παράδειγμα, η διακύμανση του μέσου είναι εμφανώς ότι απουσιάζει από τη βασική βιβλιογραφία όπως επίσης και η εκτίμηση του λόγου/αναλογίας του δειγματοληπτικού σφάλματος είναι περίπλοκη, διότι τόσο ο αριθμητής όσο και ο παρονομαστής είναι τυχαίες μεταβλητές. Ορισμένες τεχνικές εκτίμησης της διακύμανσης που δε βρίσκονται στη βασική βιβλιογραφία έχουν επαρκής ευελιξία στο να ανταποκριθούν τόσο στην πολυπλοκότητα του δειγματοληπτικού σχεδιασμού όσο και στα διάφορα στατιστικά μοντέλα. Δύο τέτοιες τεχνικές θα εξετάσουμε παρακάτω.

Στο κομμάτι αυτό θα ασχοληθούμε με τη μεθοδολογία που χρησιμοποιήσαμε για να εκτιμήσουμε τα δειγματοληπτικά σφάλματα της έρευνας και το design effect. Δυστυχώς πληροφορίες σχετικά με τα δειγματοληπτικά σφάλματα και το design effect συχνά δεν είναι διαθέσιμα ή δεν αναφέρονται με αποτέλεσμα να μην χρησιμοποιούνται όταν αναλύονται τα ουσιαστικά αποτελέσματα μίας έρευνας. Και αυτό συμβαίνει παρά το γεγονός ότι υπάρχουν ορισμένα εργαλεία λογισμικού γενικής χρήσης από τα οποία θα μπορούσαν να διεξαχθούν τέτοια αποτελέσματα. Εμείς στη συγκεκριμένη ανάλυση θα κάνουμε χρήση των εργαλείων του προγράμματος STATA. Για να εκτιμήσουμε όμως το design effect πρέπει πρώτα να εκτιμήσουμε τη διακύμανση. Υπάρχει μεγάλη ποικιλία όσον αναφορά στο πως μπορούμε να εκτιμήσουμε τη διακύμανση για τους παραπάνω λόγους. Για κοινωνικές έρευνες, οι οποίες έχουν μεγάλο μέγεθος δείγματος και παράλληλα έχουν σχετικά σύνθετο σχεδιασμό υπάρχει η δυνατότητα εφαρμογής τουλάχιστον δύο γενικών προσεγγίσεων οι οποίες βιβλιογραφικά είναι γενικά καλά τεκμηριωμένες. Αυτές οι προσεγγίσεις βασίζονται στην :

- Γραμμική μέθοδο – σειρές Taylor
- Επαναδειγματοληψία όπως η μέθοδος Bootstrap, στην ισορροπημένη επαναλαμβανόμενη αντιγραφή (BRR) και στη Jackknife επαναλαμβανόμενη αντιγραφή.

Η γραμμική μέθοδος δεν είναι φυσικά ούτε η μοναδική μέθοδος ούτε και πάντα η πιο πρακτική διαδικασία για να εκτιμήσουμε τη διακύμανση για όλες τις έρευνες. Μια εναλλακτική προσέγγιση είναι αυτή της Jackknife μεθόδου (JRR) η οποία θα χρησιμοποιηθεί και στην περίπτωση μας για να δούμε και τη διαφορά της ΕΔΣ μεταξύ των δύο μεθόδων. Η JRR έχει επιλεγεί λόγω της ευρέως διαδεδομένης χρήσης της και κυρίως επειδή η μέθοδος αυτή έχει εγκριθεί επισήμως από την Eurostat για τις στατιστικές της ΕΕ για έρευνες που σχετίζονται με το εισόδημα και τις συνθήκες διαβίωσης που σημαίνει διαχείριση μεγάλου δείγματος. Δεν θα γίνει πλήρη και λεπτομερή περιγραφή της μεθόδου JRR, αλλά θα παρέχουμε μια βάση για να μπορεί να γίνει σύγκριση με τη γραμμική μέθοδο. Αφού γίνει η περιγραφή μετά θα ακολουθήσουν εμπειρικές συγκρίσεις μεταξύ των αποτελεσμάτων από τις δυο παραπάνω μεθόδους. Κάθε μέθοδος βέβαια έχει τα πλεονεκτήματα και τους περιορισμούς της.

## 4.1 Εκτίμηση της διακύμανσης με βάση τη γραμμική μέθοδο.

> Η τεχνική της γραμμικής μεθόδου αναπτύχθηκε για δειγματοληπτικές έρευνες από τον Keyfitz, Woodruff, και άλλους και βασίζεται στη μέθοδο των σειρών Taylor. Προς τα τέλη της δεκαετίας του 1970 δημιουργήθηκαν εύχρηστα λογισμικά προγράμματα από ερευνητές του ερευνητικού ιδρύματος Triangle τα οποία βασίζονται στη γραμμική μέθοδο. Το συγκεκριμένο πρόγραμμα λέγεται SUDAAN (έρευνα για την ανάλυση δεδομένων). Η διαθεσιμότητα αυτού και η εμφάνιση και άλλων τέτοιων λογισμικών προγραμμάτων έχουν κάνει τη γραμμική μέθοδο ίσως την πιο ευρέως χρησιμοποιούμενη για την εκτίμηση της διακύμανσης σε σύνθετες έρευνες (Kalton, 1983: 281)

Η Taylor γραμμικοποίηση είναι μια δημοφιλής μέθοδος εκτίμησης της διακύμανσης για πολύπλοκες στατιστικές, όπως είναι η εκτίμηση του λόγου, της παλινδρόμησης και της λογιστικής παλινδρόμησης .

Μια πρώιμη εφαρμογή της μεθόδου ήταν να εκτιμήσει μια συνάρτηση που είναι δύσκολο να υπολογιστεί, για παράδειγμα, μια εκθετική $e^x$ ή μια λογαριθμική $[\log(x)]$συνάρτηση. Το ανάπτυγμα των σειρών Taylor για την $e^x$ περιλαμβάνει τη λήψη της πρώτης και της μεγαλύτερης τάξεως παραγώγων της $e^x$ ως προς το x, την εκτίμηση των παραγώγων σε κάποια βαθμό, συνήθως μηδέν, και την οικοδόμηση των σειρών από όρους με βάση τα παράγωγα. Το ανάπτυγμα για την $e^x$ είναι :

$$1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\frac{x^4}{4!}+..... \quad (4.1)$$

Αυτή είναι μια ειδική εφαρμογή του παρακάτω γενικού τύπου επεκτεινόμενο στο $a$

$$f(x)=f(a)+f'(a)(x-a)+\frac{f''(a)(x-a)^2}{2!}+\frac{f'''(a)(x-a)^3}{3!}+..... \quad (4.2)$$

Σε στατιστικές έρευνες, οι σειρές Taylor χρησιμοποιούνται για να γίνει εκτίμηση μιας μη γραμμικής συνάρτησης και στη συνέχεια, η διακύμανση της συνάρτησης βασίζεται στη συνάρτηση η οποία έχει προσεγγιστεί από τις σειρές Taylor. Συχνά, η προσέγγιση παρέχει μια λογική εκτίμηση της συνάρτησης, και μερικές φορές η προσέγγιση είναι ακόμη και μια γραμμική συνάρτηση. Αυτή η ιδέα της εκτίμησης της διακύμανσης έχει πολλά ονόματα στη βιβλιογραφία, συμπεριλαμβανομένης της γραμμικής μεθόδου, της μεθόδου δέλτα (Kalton, 1983: 44 ), και της διάδοσης της διακύμανσης (Kish, 1965: 583). Σε στατιστικές εφαρμογές, η ανάπτυξη εκτιμάται στη μέση ή αναμενόμενη τιμή του $x$ , το οποίο γράφεται ως $E(x)$. Εάν χρησιμοποιούμε το $E(x)$ για τον παραπάνω γενικό τύπο τότε έχουμε:

$$f(x) = f\left[E(x)\right] + f'\left[E(x)\right]\left[x - E(x)\right] + \frac{f''\left[E(x)\right]\left[x - E(x)\right]^2}{2!} + \ldots\ldots \quad (4.3)$$

Η διακύμανση του $f(x)$ είναι $V\left[f(x)\right] = E\left[f^2(x)\right] - E^2\left[f(x)\right]$ εξ ορισμού και χρησιμοποιώντας το ανάπτυγμα σε σειρές Taylor, έχουμε

$$V\left[f(x)\right] = \left\{f'\left[E(x)\right]\right\}^2 V(x) + \ldots. \qquad (4.4)$$

Το θεώρημα αυτό αποδίδεται και όταν μια συνάρτηση έχει περισσότερες από μία τυχαία μεταβλητή. Στην περίπτωση μιας συνάρτησης των δύο διακυμάνσεων το ανάπτυγμα σε σειρές Taylor αποδίδεται :

$$V\left[f(x_1, x_2)\right] \cong \left(\frac{\partial f}{\partial x_1}\right)\left(\frac{\partial f}{\partial x_2}\right)\mathrm{cov}(x_1, x_2) \qquad (4.5)$$

Εφαρμόζοντας στην παραπάνω εξίσωση, το λόγο δύο μεταβλητών $x$ και $y$ δηλαδή, $r = \frac{y}{x}$ παίρνουμε τον τύπο της διακύμανσης για την εκτίμηση του λόγου/αναλογίας

$$\begin{aligned} V(r) &= \frac{V(y) + r^2 V(x) - 2r\,\mathrm{cov}(x, y)}{x^2} + \ldots\ldots \\ &= r^2\left(\frac{V(y)}{y^2} + \frac{V(x)}{x^2} - \frac{2\,\mathrm{cov}(x, y)}{xy}\right) + \ldots\ldots \end{aligned} \qquad (4.6)$$

Αναπτύσσοντας την εξίσωση (4.4) στην περίπτωση c τυχαίων μεταβλητών, η εκτίμηση της διακύμανση των $\theta = f(x_1, x_2 \ldots., x_c)$ γίνεται από τον τύπο:

$$V(\theta) \cong \Sigma\Sigma\left(\frac{\partial f}{\partial x_1}\right)\left(\frac{\partial f}{\partial x_j}\right)\mathrm{cov}(x_i, x_j) \qquad (4.7)$$

Αναπτύσσοντας την εξίσωση (4.7) για έναν σταθμισμένο εκτιμητή $f(Y) = \hat{Y}_1 = \Sigma w_i y_{ij}, j = 1, 2, ....., c$ περιλαμβάνοντας c μεταβλητές σε ένα δείγμα n παρατηρήσεων Woodruff (1971) οδηγούμαστε στον τύπο :

$$V(\theta) \cong V\left[\Sigma w_1 \Sigma\left(\frac{\partial f}{\partial y_j}\right) y_{ij}\right] \qquad (4.8)$$

Αυτός ο εναλλακτικός τύπος της γραμμικοποιημένης διακύμανση μιας μη γραμμικής εκτίμησης προσφέρει πλεονεκτήματα επειδή παρακάμπτει τον υπολογισμό της c × c μήτρας συνδιασποράς της εξίσωσης (4.7). Αυτή η ευκολία της μετατροπής του προβληματος που σχετίζεται με την εκτίμηση πολλών σταδίων σε πρόβλημα μονοπαραγοντικής εκτίμησης γίνεται με απλή ανταλλαγή των αθροισμάτων. Αυτή η γενική υπολογιστική διαδικασία μπορεί να εφαρμοστεί σε μια ποικιλία από μη γραμμικές εκτιμήσεις, συμπεριλαμβανομένων των συντελεστών παλινδρόμησης (Fuller, 1975).

Για μια σύνθετη έρευνα, αυτή η μέθοδος προσέγγισης εφαρμόζεται σε σύνολα PSU εντός ενός στρώματος. Δηλαδή, η εκτίμηση της διακύμανσης είναι ένας σταθμισμένος συνδυασμός των διακυμάνσεων μεταξύ PSUs εντός του ιδίου στρώματος στην εξίσωση (4.8). Αυτοί οι τύποι είναι πολύπλοκοι, αλλά μπορεί να απαιτούν πολύ λιγότερο χρόνο από ότι απαιτεί η μέθοδο της επαναλαμβανόμενης αντιγραφής που θα συζητηθεί παρακάνω. Η μέθοδος αυτή μπορεί να εφαρμοστεί σε οποιαδήποτε στατιστική εφαρμογή που εκφράζεται με μαθηματικό τρόπο για παράδειγμα, στο μέσο ή στον συντελεστή παλινδρόμησης αλλά όχι σε μη στατιστικές συναρτήσεις όπως είναι η διάμεσος.

Πίνακας 1.

Υπολογισμός τυπικού σφάλματος για την απλή τυχαία δειγματοληψία και με την γραμμική μέθοδο για το ποσοστό των μαθητών που απάντησαν ‘’όχι΄΄ στο κάπνισμα.

ESPAD 2011

| | | Proportion | Std. Err. | Linearized Std. Err. Class + Nomos | Design Effect | Number of obs | Number of obs after Design Effect |
|---|---|---|---|---|---|---|---|
| c08 | Κάπνισες ποτέ τσιγάρο | 0,6087 | 0,003338 | 0,007181 | 4,6276 | 35.462 | 7.663 |
| c09 | Πόσο κάπνισες τις τελευταίες 30 ημέρες | 0,8590 | 0,002464 | 0,004673 | 3,5966 | 35.553 | 9.885 |

Το 60% τον μαθητών δεν έχουν καπνίσει ποτέ τους τσιγάρο το υπόλοιπο ποσοστό έχουν δοκιμάσει έστω και μια φορά. Επιπλέον περίπου το 86% των μαθητών κάπνισαν λιγότερο από ένα τσιγάρο την ημέρα της τελευταίες 30 μέρες. Υπολογίζουμε αρχικά το τυπικό σφάλμα έχοντας απλά το σύνολο του πληθυσμού του δείγματος και της απαντήσεις για την εκάστοτε ερώτηση. Βλέπε αποτελέσματα που διεξήχθησαν από το στατιστικό πρόγραμμα STATA.[5]

Παρατηρούμε ωστόσο ότι το τυπικό σφάλμα αυξάνεται περίπου στο διπλάσιο όταν υπολογίζεται με τη γραμμική μέθοδο για την κατά συστάδες δειγματοληψία, στρωματοποιημένη κατά νομό και ομαδοποιημένη ανά σχολική τάξη που είναι και η βασική μονάδα δειγματοληψίας.

Και για της υπόλοιπες υπό εξέταση ερωτήσεις η διαδικασία υπολογισμού των δύο τυπικών σφαλμάτων έγινε με την ίδια διαδικασία. Έχοντας υπολογίσει τα δύο τυπικά σφάλματα μπορούμε να υπολογίσουμε το design effect σύμφωνα με τους τύπους που αναλύθηκαν στο 3$^{ο}$ κεφάλαιο. Εφόσον η επίδραση του δειγματοληπτικού σχεδιασμού (ΕΔΣ) είναι μεγαλύτερη της μονάδας αυτό σημαίνει ότι η κατά συστάδες μέθοδος δειγματοληψίας που ακολουθήθηκε στην περίπτωση της έρευνας μας είναι λιγότερο αποτελεσματική από ό,τι εάν ακολουθούσαμε τη μέθοδο της απλής τυχαία δειγματοληψίας χωρίς αντικατάσταση. Μάλιστα στις δύο αυτές ερωτήσεις η ΕΔΣ είναι 4,6 και 3,5 αντίστοιχα δείχνοντας μας ότι το δείγμα λόγω του ότι είχε χωριστεί σε ομάδες που αφορούσαν τάξεις είχε ταύτιση απόψεων στις ερωτήσεις αυτές οπότε επιλέγοντας με τη μέθοδο της απλής τυχαίας δειγματοληψίας αντί 35.462 μονάδες 7.663 θα είχαμε επιτύχει το ίδιο αποτέλεσμα προφανώς με λιγότερο κόστος.

[5] Τα αποτελέσματα από το πρόγραμμα STATA βρίσκονται στο Παραρτήμα 1.

Πίνακας 2.

Υπολογισμός τυπικού σφάλματος για την απλή τυχαία δειγματοληψία και με τη γραμμική μέθοδο για το ποσοστό των μαθητών που απάντησαν όχι στην κατανάλωση αλκοόλ .

ESPAD 2011

| | | Proportion | Std. Err. | Linearized Std. Err. Class + Nomos | Design Effect | Number of obs | Number of obs after Design Effect |
|---|---|---|---|---|---|---|---|
| c12 | Ήπιες ποτέ αλκοολούχο ποτό | | | | | | |
| c12a | Σε όλη σου τη ζωή μέχρι και σήμερα | 0,2604 | 0,002922 | 0,006445 | 4,8656 | 34.632 | 7.118 |
| c12b | Στη διάρκεια των 12 τελευταίων μηνών μέχρι και σήμερα | 0,3980 | 0,003299 | 0,007710 | 5,4613 | 34.945 | 6.399 |
| c12c | Στη διάρκεια των 30 τελευταίων ημερών μέχρι και σήμερα | 0,3922 | 0,003296 | 0,006803 | 4,2595 | 35.040 | 8.226 |
| c18 | Τις 30 τελευταίες ημέρες πόσες φορές ήπιες στη σειρά 5 ή περισσότερα αλκοολούχα ποτά | 0,5486 | 0,004175 | 0,005782 | 1,9176 | 23.034 | 12.012 |
| c19 | Πόσες φορές μέθυσες από αλκοολούχα ποτά | | | | | | |
| c19a | Σε όλη σου τη ζωή μέχρι και σήμερα | 0,9916 | 0,000598 | 0,000688 | 1,3228 | 35.191 | 26.603 |
| c19b | Στη διάρκεια των 12 τελευταίων μηνών μέχρι και σήμερα | 0,9964 | 0,000512 | 0,000548 | 1,1482 | 23.000 | 20.031 |
| c19c | Στη διάρκεια των 30 τελευταίων ημερών μέχρι και σήμερα | 0,9983 | 0,000344 | 0,000341 | 0,9878 | 23.105 | 23.390 |

Στον Πίνακα 2 φαίνεται ότι είναι λιγότερο το ποσοστό των μαθητών που δεν καταναλώνουν αλκοολούχα ποτά, μάλιστα μόλις το 26% των μαθητών έχουν καταναλώσει λιγότερα από 1-2 ποτά. Από το 74% όμως των μαθητών που καταναλώνουν αλκοολούχα ποτά φαίνεται ότι το περίπου το 45% κατανάλωσαν περισσότερα από 5 ποτά στη σειρά. Και στην κατηγορία της κατανάλωσης αλκοόλ η ΕΔΣ είναι > 1. Μάλιστα στη ερώτηση του αν ήπιαν ποτέ αλκοολούχο ποτό η ΕΔΣ αγγίζει το 5,4 που σημαίνει ότι υπάρχει μεγάλη ταύτιση ανάμεσα στους μαθητές των τάξεων αποδεικνύοντας ότι το δείγμα των 34.945 μαθητών στη πραγματικότητα είναι σαν να είναι 6.399 καθώς επηρεάζονται στη συμπεριφορά μεταξύ τους. Μόνο

για την τελευταία ερώτηση του Πίνακα 2 ο δειγματοληπτικός σχεδιασμός είναι σωστός καθώς η ΕΔΣ είναι 0,98.

Πίνακας 3.

Υπολογισμός τυπικού σφάλματος για την απλή τυχαία δειγματοληψία και με τη γραμμική μέθοδο για το ποσοστό των μαθητών που απάντησαν ΄΄όχι΄΄ στη χρήση μαριχουάνας η χασίς.

ESPAD 2011

| | | Proportion | Std. Err. | Linearized Std. Err. Class + Nomos | Design Effect | Number of obs | Number of obs after Design Effect |
|---|---|---|---|---|---|---|---|
| c25 | Δοκίμασες ή πήρες ποτέ μαριχουάνα ή χασίς | | | | | | |
| c25a | Σε όλη σου τη ζωή μέχρι και σήμερα | 0,9802 | 0,001069 | 0,001502 | 1,9748 | 35.548 | 18.001 |
| c25b | Στη διάρκεια των 12 τελευταίων μηνών μέχρι και σήμερα | 0,9846 | 0,001150 | 0,001462 | 1,6166 | 23.127 | 14.306 |
| c25c | Στη διάρκεια των 30 τελευταίων ημερών μέχρι και σήμερα | 0,9917 | 0,000856 | 0,001074 | 1,5740 | 23.158 | 14.713 |
| | Δοκίμασες ή πήρες ποτέ έκσταση | | | | | | |
| c29a | Σε όλη σου τη ζωή μέχρι και σήμερα | 0,9979 | 0,000326 | 0,000323 | 0,9817 | 35.343 | 36.004 |

Στον Πίνακα 3 φαίνεται ότι το ποσοστό των μαθητών που δεν κάνουν χρήση μαριχουάνας η χασίς αγγίζει το ποσοστό του 99% και σε αυτή την ερώτηση εξαιρουμένης του τελευταίου σκέλους της η ΕΔΣ είναι > από 1,5 αποδεικνύοντας ότι και σε αυτή τη κατηγορία η μέθοδος της δειγματοληψίας κατά συστάδες που ακολουθήθηκε κατά το δειγματοληπτικό σχεδιασμό δεν ήταν η καταλληλότερη επιλογή.

Πίνακας 4.

Υπολογισμός τυπικού σφάλματος για την απλή τυχαία δειγματοληψία και με τη γραμμική μέθοδο για το ποσοστό των μαθητών που απάντησαν ‘’όχι’’ στη χρήση διαφόρων ουσιών.

ESPAD 2011

| | | Proportion | Std. Err. | Linearized Std. Err. Class + Nomos | Design Effect | Number of obs | Number of obs after Design Effect |
|---|---|---|---|---|---|---|---|
| c31 | Έχεις χρησιμοποιήσει ποτέ κάποια από τις παρακάτω ουσίες | | | | | | |
| c31a | Ηρεμιστικά ή υπνωτικά χωρις τη σύσταση γιατρού | 0,9962 | 0,000426 | 0,000454 | 1,1363 | 35.489 | 31.233 |
| c31b | Αμφεταμίνες | 0,9987 | 0,000303 | 0,000299 | 0,9751 | 23.155 | 23.747 |
| c31c | LSD ή κάποιο άλλο παραισθησιογόνο | 0,9983 | 0,000385 | 0,000453 | 1,3820 | 23.093 | 16.710 |
| c31d | Κράκ | 0,9985 | 0,000327 | 0,000379 | 1,3413 | 23.086 | 17.211 |
| c31e | Κοκαινη | 0,9969 | 0,000514 | 0,000564 | 1,2044 | 23.155 | 19.225 |
| c31f | Ρελιβίνη | 0,9989 | 0,000279 | 0,000277 | 0,9914 | 23.101 | 23.301 |
| c31g | Ηρωίνη | 0,9980 | 0,000320 | 0,000320 | 0,9988 | 23.142 | 23.171 |
| c31h | 'Μαγικά μανιτάρια'' | 0,9975 | 0,000391 | 0,000411 | 1,1039 | 23.076 | 20.904 |
| c31i | GHB | 0,9992 | 0,000232 | 0,000254 | 1,1986 | 23.018 | 19.205 |
| c31j | Αναβολικά χωρίς τη σύσταση γιατρού | 0,9984 | 0,000318 | 0,000312 | 0,9577 | 23.097 | 24.116 |
| c31k | Ναρκωτικά με ένεση | 0,9986 | 0,000245 | 0,000245 | 1,0024 | 23.142 | 23.085 |
| c31l | Κάποιο αλκοολούχο ποτό μαζι με φάρμακα για να αλλάξεις τη διάθεση σου | 0,9983 | 0,000368 | 0,000372 | 1,0235 | 23.187 | 22.654 |

Στον Πίνακα 4 που σχετίζεται με τη χρήση διάφορων ουσιών το ποσοστό των μαθητών που κάνουν χρήση αυτών των ουσιών δε φτάνει ούτε το 1%. Σε αυτή την περίπτωση από τις τιμές της ΕΔΣ που μπορούμε να παρατηρήσουμε βλέπουμε ότι ο δειγματοληπτικός σχεδιασμός που ακολουθήθηκε είναι αρκετά αποτελεσματικός αφού στις περισσότερες κατηγορίες οι τιμές της ΕΔΣ κυμαίνονται κοντά στο 1. Με μια εξαίρεση τη χρήση Αμφεταμινών, LSD ή κάποιο άλλο παραισθησιογόνο και Κράκ, που παρατηρείται υψηλότερη συσχέτιση.

Όπως προαναφέρθηκε είναι γενικά εφαρμόσιμη σε οποιοδήποτε σχέδιο δειγματοληψίας που επιτρέπει αμερόληπτες εκτίμηση της διακύμανσης για γραμμικές εκτιμήσεις, και είναι υπολογιστικά απλούστερη από μια μέθοδο

αναδειγματοληψίας όπως η Jackknife. Ωστόσο, αυτό μπορεί να οδηγήσει σε πολλαπλές εκτιμήσεις της διακύμανσης που είναι ασυμπτωτικά αμερόληπτα σχεδιασμένα υπό την επαναλαμβανόμενη δειγματοληψία. Η επιλογή μεταξύ των μεθόδων εκτίμησης της διακύμανση, απαιτεί την εξέταση και άλλων παραγόντων, όπως την αμερόληπτη προσέγγιση για τη μέθοδο εκτίμησης της διακύμανσης που θα ακολουθηθεί σύμφωνα με ένα υποτιθέμενο μοντέλο, και την εγκυρότητα, γινομένης υπό όρους ενός επαναλαμβανόμενου δειγματοληπτικού πλαισίου.

Για παράδειγμα, στο πλαίσιο της απλής τυχαίας δειγματοληψίας και της εκτίμησης της αναλογίας, $\hat{Y}_R = \left(\frac{\bar{y}}{\bar{x}}\right) X$ του συνολικού πληθυσμού Y, οι Royall και Cumberland (1981) έδειξαν ότι η ευρέως χρησιμοποιούμενη γραμμική μέθοδος εκτιμητής της διακύμανση $v_{LI} = N^2 (n^{-1} - N^{-1}) S_z^2$ δεν ακολουθεί την υπό συνθήκη διακύμανση του $\hat{Y}_R$ έχοντας $\bar{\chi}$ σε αντίθεση με την εκτίμηση της διακύμανσης με τη μέθοδο Jackknife $v_j$. Το $\bar{y}$ και $\bar{\chi}$ είναι τα μέσα του δείγματος, το X είναι ο συνολικός πληθυσμός μιας βοηθητικής μεταβλητής x ,το $S_z^2$ είναι η διακύμανση των καταλοίπων του δείγματος $z_i = y_i - \left(\frac{\bar{y}}{\bar{x}}\right) x_i$ και (n,N) υποδηλώνουν το δείγμα και το μέγεθος του πληθυσμού.

### 4.1.1 Η μέθοδος

Για την παρουσίαση της μεθόδου ξεκινάμε με τη γενική παραδοχή ότι ο εκτιμητής του $\widehat{\theta}$ μιας παραμέτρου θ μπορεί να εκφραστεί ως μια διαφορίσιμη συνάρτηση $g(\hat{\underset{\sim}{Y}})$ των εκτιμώμενων συνόλων $\hat{\underset{\sim}{Y}} = \left(\hat{Y}_1 \ldots \hat{Y}_m\right)^T$ όπου $\hat{Y}_j = \Sigma_{ieU} d_i(S) y_{ij}, j = 1, \ldots, m$, $\theta = g(\underset{\sim}{Y})$ και $\underset{\sim}{Y} = (Y_1 \ldots . Y_m)^T$ μπορούμε να το γράψουμε το $\widehat{\theta}$ ως $\widehat{\theta} = f\left(\underset{\sim}{d}(s), \underset{\sim}{A}_y\right)$ και $\theta = f\left(\underset{\sim}{1}, \underset{\sim}{A}_y\right)$ όπου $A_y$ είναι m x N πίνακας με $j^{th}$ στήλη. $\underset{\sim}{y}_j = \left(y_{j1}, \ldots . y_{jm}\right)^T, j = 1, \ldots \ldots N$, $\underset{\sim}{d}(s) = (d_1(s), \ldots, d_N(s))^T$ και $\underset{\sim}{1}$ είναι το N-διάνυσμα της $1^{ης}$ . Για παράδειγμα εάν το $\hat{\theta}$ δηλώνει τον εκτιμητή της αναλογίας/λόγου $\hat{Y}_R = \left[\frac{\left(\Sigma d_i(s) y_i\right)}{\left(\Sigma d_i(s) x_i\right)}\right] X$, τότε $m = 2, y_{1i} = y_i, y_{2i} = x_i$ και μειώνει στο σύνολο το Y, σημειώνοντας ότι $(Y / X) X = Y$ . Σημείωση ότι $\hat{Y}_R$ είναι μια συνάρτηση του $d(s)$, $y$ και $x$ και το γνωστό σύνολο X, αλλά απορρίψαμε το X για λόγους απλούστευσης οπότε $\hat{Y}_R = f(\underset{\sim}{d}(s), y, x)$.

Η Taylor γραμμικοποίηση του $\hat{\theta}$ σε σχέση με το $\underset{\sim}{Y}$ μας δίνει :

$$\hat{\theta} - \theta = g(\hat{\underset{\sim}{Y}}) - g(\underset{\sim}{Y}) \approx (\partial g(\underset{\sim}{a}) / \partial \underset{\sim}{a})^T \big|_{\underset{\sim}{a} = \underset{\sim}{y}} (\hat{\underset{\sim}{Y}} - \underset{\sim}{Y}) \qquad (4.9)$$

όπου $\partial g(a)/\partial g = (\partial g(a)\partial a_1, ...., \partial g(a)/\partial a_m)^T$ . Υποθέτουμε ότι το (4.9) ικανοποιείται όταν ο δειγματοληπτικός σχεδιασμός πραγματοποιείται κανονικά υπό κατάλληλες συνθήκες.

Ας κάνουμε $\breve{\underset{\sim}{Y}} = \Sigma b_i y_i$ για τυχαίους κανονικούς αριθμούς $\underset{\sim}{b} = (b_1, ....b_N)^T$, και $g(\breve{\underset{\sim}{Y}}) = f(\underset{\sim}{b}, \underset{\sim}{A}_y) = f(\underset{\sim}{b})$ . Σημειώνοντας $\hat{\underset{\sim}{Y}} = \underset{\sim}{A}_y d(s)$ και $\underset{\sim}{Y} = \underset{\sim}{A}_y \underset{\sim}{1}$ μπορούμε να εκφράσουμε το (4.9) σαν:

$$\hat{\theta} - \theta \approx (\partial g(\breve{\underset{\sim}{Y}})/\partial \breve{\underset{\sim}{Y}})^T\big|_{\breve{\underset{\sim}{Y}}=\underset{\sim}{Y}} \underset{\sim}{A}_y(\underset{\sim}{d}(s) - 1) = \Sigma(\partial f(\underset{\sim}{b})/\partial \breve{\underset{\sim}{Y}})^T\big|_{\underset{\sim}{b}=\underset{\sim}{1}} \sum_{k=1}^{N} y_k(d_k(s) - 1) \quad (4.10)$$

Σημειώνοντας ότι το $\breve{\underset{\sim}{Y}} = \underset{\sim}{Y}$ είναι ισοδύναμο με $\underset{\sim}{b} = 1$ μπορούμε να αντικαταστήσουμε το $\underset{\sim}{y}_k = \partial \breve{\underset{\sim}{Y}}/\partial b_k\big|_{\underset{\sim}{b}=\underset{\sim}{1}}$ στο (4.10) και προκύπτει :

$$\hat{\theta} - \theta \approx \Sigma(\partial f(\underset{\sim}{b})/\partial b_k)\big|_{\underset{\sim}{b}=\underset{\sim}{1}} (\sum_{k=1}^{N} d_k(s) - \underset{\sim}{1}) = \underset{\sim}{\tilde{z}}^T(\underset{\sim}{d}(s) - \underset{\sim}{1}) \qquad (4.11)$$

όπου $\underset{\sim}{\tilde{z}} = (\tilde{z}_1, ....\tilde{z}_N)^T$ με $\tilde{z}_k = \partial f(\underset{\sim}{b})/\partial b_k\big|_{\underset{\sim}{b}=\underset{\sim}{1}}$. Συνεπώς από την (4.11) η εκτίμηση της διακύμανσης του $\hat{\theta}$ δίνεται από την εκτίμηση της διακύμανσης του εκτιμώμενου συνόλου $\Sigma d_i(s)\tilde{z}_i = \hat{Y}(\tilde{z})$ , αυτό είναι $\mathrm{var}(\hat{\theta}) \approx v(\tilde{z})$. Τώρα αντικαθιστούμε το $\tilde{z}_k$ με $z_k = \partial f(\underset{\sim}{b})/\partial b_k\big|_{\underset{\sim}{b}=\underset{\sim}{d}(s)}$ , καθώς το $\tilde{z}_k$ είναι άγνωστο, για να πάρουμε τη γραμμική εκτίμηση της διακύμανσης

$$v_L(\hat{\theta}) = v(z) \qquad (4.12)$$

Το $v_L(\hat{\theta})$ που δίνεται από το (4.12) απλός λαμβάνεται από τον τύπο $v_{(y)}$ για $\hat{Y}$ αντικαθιστώντας το $y_i$ με το $z_i$ για $i \in s$. Επιπλέον δεν αξιολογούμε τη μερική παράγωγο $\partial f(\underset{\sim}{b})/\partial b_k$ στο $\underset{\sim}{b} = \underset{\sim}{1}$ για να πάρουμε το $\underset{\sim}{\tilde{z}}$ έτσι ώστε στη συνέχεια να αντικαταστήσουμε τις εκτιμήσεις για τις άγνωστες συνιστώσες του $\underset{\sim}{\tilde{z}}$ . Η μέθοδός μας, ως εκ τούτου, είναι στο ίδιο πνεύμα με την προσέγγιση Binder. Η εκτίμηση της διακύμανσης $v_L$ είναι έγκυρη επειδή $z_i$ είναι μια σταθερά της εκτίμησης του $\tilde{z}_i$ . Εάν υποθέσουμε ότι το $\hat{\theta}$ είναι η εκτίμηση του λόγου $\hat{Y}_R = X\left[\left(\Sigma d_i(s)y_i\right)/\left(\Sigma d_i(s)x_i\right)\right]$ . Τότε $f(\underset{\sim}{b}) = X\left[\left(\Sigma b_i y_i\right)/\left(\Sigma b_i x_i\right)\right] = X\hat{Y}(\underset{\sim}{b})$ και $\partial f(\underset{\sim}{b})/\partial b_k = X\,\dfrac{y_k \Sigma b_i x_i - x_k \Sigma b_i y_i}{\left(\Sigma b_i x_i\right)^2}$.

Άρα, $z_k = \partial f(\underset{\sim}{b}) / \partial b_k \big|_{\underset{\sim}{b}=\underset{\sim}{d}(s)} = \frac{X}{\hat{X}}\left(y_k - \hat{R}x_k\right)$. Η εκτίμηση της διακύμανσης $v_L(\hat{Y}_R)$ είναι ίδια με την εκτίμηση της διακύμανσης του Binder(1996).

### 4.1.2 Εκτιμήτρια της βαθμονόμησης

Η εκτίμηση του λόγου μπορεί να θεωρηθεί και ως εκτιμήτρια της βαθμονόμησης $\hat{Y}_R = \Sigma w_i(s) y_i$, με ρητά βάρη (weights) $w_i(s) = (X / \hat{X}) d_i(s)$ και ικανοποιώντας τους περιορισμού $\Sigma w_i(s) x_i = X$. Η εκτιμήτρια της βαθμονόμησης του συνολικού Υ από τον τύπο $\hat{Y}_w = \Sigma w_i(s) y_i$ με ρητό βάρος $w_i(s)$ και ικανοποιώντας τους περιορισμούς $\Sigma w_i(s) \underset{\sim}{x}_i = \underset{\sim}{X}$ χρησιμοποιείται ευρέως όταν $x_i = (x_{1i}, \ldots . x_{qi})^T$ και $\underset{\sim}{X} = (X_1, \ldots . . X_q)^T$ είναι το διάνυσμα των συνόλων των βοηθητικών μεταβλητών $x_j$, $j = 1, \ldots . . q$.

#### *4.1.2.1 Γενικευμένη εκτίμηση της παλινδρόμησης*

Η γενικευμένη εκτίμηση της παλινδρόμησης του συνόλου Υ δίνεται από το $\hat{Y}_w$ με συντελεστές βαθμονόμησης $w_i(s) = d_i(s) g_i(\underset{\sim}{d}(s))$ όπου,

$$g_i(\underset{\sim}{d}(s)) = 1 + (\underset{\sim}{X} - \hat{\underset{\sim}{X}})^T (\Sigma d_i(s) c_i \underset{\sim}{x}_i \underset{\sim}{x}_i^{T\rightarrow})^{-1} c_i \underset{\sim}{x}_i , \qquad (4.13)$$

με συγκεκριμένες σταθερές $c_i$ και $\hat{\underset{\sim}{X}} = \Sigma d_i(s) \underset{\sim}{x}_i$. Ο εκτιμητής του λόγου , $\hat{Y}_R$ , είναι μια ειδική περίπτωση με q=1 και $c_i = x_i^{-1}$ , και $g_i(\underset{\sim}{d}(s))$, δίνεται από τον παραπάνω τύπο μειωμένο κατά $X / \hat{X}$.

Η γενικευμένη εκτίμηση της παλινδρόμησης μπορεί να εκφραστεί ως μια διαφορίσιμη συνάρτηση του εκτιμώμενου σύνολα. Καθώς η γενική θεωρία που αφορά τη μέθοδο εφαρμόζεται και παραμένει να αξιολογήσει το $z_k = \partial f(\underset{\sim}{b}) / \partial b_k \big|_{\underset{\sim}{b}=\underset{\sim}{d}(s)}$ όπου $f(\underset{\sim}{b}) = \Sigma(b_i g_i(\underset{\sim}{b})) y_i$ έχει προκύψει αντικαθιστώντας το $\underset{\sim}{d}(s)$ με το $\underset{\sim}{b}$ για τον τύπο $\hat{Y}_w$ . Σημειώνοντας ότι $\partial \underset{\sim}{A}(\underset{\sim}{b})^{-1} / \partial b_k = -\underset{\sim}{A}(\underset{\sim}{b})^{-1} (\partial \underset{\sim}{A}(\underset{\sim}{b}) / \partial b_k) \underset{\sim}{A}(\underset{\sim}{b})^{-1}$ όπου $\underset{\sim}{A}(\underset{\sim}{b}) = \Sigma b_i c_i \underset{\sim}{x}_i x_i$ παίρνουμε

$$\frac{\partial(b_k g_k(\underset{\sim}{b}))}{\partial b_k} = g_k(\underset{\sim}{b}) - x_k{}^T \underset{\sim}{A}(\underset{\sim}{b})^{-1} b_k c_k \underset{\sim}{x}_k - (\underset{\sim}{X} - \hat{\underset{\sim}{X}}(\underset{\sim}{b}))^T \underset{\sim}{A}(\underset{\sim}{b})^{-1} (c_k \underset{\sim}{x}_k \underset{\sim}{x}_k{}^T) \underset{\sim}{A}(\underset{\sim}{b})^{1} (b_k c_k \underset{\sim}{x}_k)$$

(4.14)

και για $i \neq k$ $\frac{\partial(b_i g_i(\underset{\sim}{b}))}{\partial b_k} = -\underset{\sim}{x}_k{}^T \underset{\sim}{A}(\underset{\sim}{b})^{-1} (b_i c_i \underset{\sim}{x}_i) - (\underset{\sim}{X} - \hat{\underset{\sim}{X}}(\underset{\sim}{b}))^T \underset{\sim}{A}(\underset{\sim}{b})^{-1} (b_i c_i \underset{\sim}{x}_i)$ (4.15)

Επομένως τώρα από τους παραπάνω τύπους προκύπτει $\partial f(\underset{\sim}{b}) / \partial b_k = g_k(\underset{\sim}{b}) e_k(\underset{\sim}{b})$

όπου $e_k(\underset{\sim}{b}) = y_k - \underset{\sim}{x}_k^T \underset{\sim}{B}(\underset{\sim}{b})$ με $\underset{\sim}{B}(\underset{\sim}{b}) = \underset{\sim}{A}^{-1}(\underset{\sim}{b})(\Sigma_i b_i c_i \underset{\sim}{x}_i y_i)$. Άρα το $z_k = \partial f(\underset{\sim}{b}) / \partial b_k \big|_{\underset{\sim}{b}=\underset{\sim}{d}(s)}$ περιορίζεται στο $z_k = g_k(\underset{\sim}{d}(s))e_k$, όπου $e_k = y_k - \underset{\sim}{x}_k^T \hat{\underset{\sim}{B}}$ με $\hat{\underset{\sim}{B}} = \underset{\sim}{b}(\underset{\sim}{d}(s))$.

Η εκτίμηση της διακύμανσης γίνεται λαμβάνοντας υπόψη τα $g$ - βάρη, $g_k(\underset{\sim}{d}(s))$ σε αντίθεση με την εκτίμηση της διακύμανσης με την απλή γραμμική μέθοδο. Επιπλέον συμφωνεί με το μοντέλο εκτίμησης της διακύμανσης του Sarndal (1989).

#### *4.1.2.2 Εκτιμώντας την εξίσωση*

Επιστρέφουμε στον εκτιμητή του $\hat{\underset{\sim}{\theta}} = (\hat{\theta}_1, ...., \hat{\theta}_p)^T$ ενός διανυσματικού παραμέτρου $\underset{\sim}{\theta}$ το οποίο ορίζεται είτε ρητά είτε έμμεσα ως λύση για την εκτίμηση εξισώσεων με συντελεστές βαθμονόμησης $w_i(s) = d_i(s) g_i(\underset{\sim}{d}(s))$:

$$\hat{\underset{\sim}{S}}(\hat{\underset{\sim}{\theta}}) = \Sigma w_i(s) \underset{\sim}{u}_i(\hat{\underset{\sim}{\theta}}) = \underset{\sim}{0} \qquad (4.16)$$

όπου $\underset{\sim}{u}_i(\hat{\underset{\sim}{\theta}})$ και $\hat{\underset{\sim}{S}}(\hat{\underset{\sim}{\theta}})$ είναι (px1) διανύσματα. Σε γενικές γραμμές, η λύση $\hat{\underset{\sim}{\theta}}$ στις παραπάνω εξισώσεις δεν μπορεί να εκφραστεί ως συνάρτηση των εκτιμώμενων συνόλων. Ως εκ τούτου, ακολουθούν την Binder(1983) προσέγγιση και η εκτίμηση της μήτρας συνδιακύμανσης του $\hat{\underset{\sim}{\theta}}$ με τη γραμμική μέθοδο γράφεται ως :

$$\underset{\sim}{v}_L(\hat{\underset{\sim}{\theta}}) = \left[ \hat{\underset{\sim}{J}}(\hat{\underset{\sim}{\theta}}) \right]^{-1} \hat{\underset{\sim}{\Sigma}}_S(\hat{\underset{\sim}{\theta}}) \left[ \hat{\underset{\sim}{J}}(\hat{\underset{\sim}{\theta}}) \right]^{-1} \qquad (4.17)$$

όπου , $\hat{\underset{\sim}{J}}(\underset{\sim}{\theta}) \equiv \partial \hat{\underset{\sim}{S}}(\theta) / \partial \theta$ και $\hat{\underset{\sim}{\Sigma}}_S(\hat{\underset{\sim}{\theta}})$ είναι η εκτιμώμενη μήτρα συνδιακύμανσης $\underset{\sim}{v}_L(\hat{\underset{\sim}{S}}(\underset{\sim}{\theta})) = \hat{\underset{\sim}{\Sigma}}_S(\underset{\sim}{\theta})$ εκτιμώμενη στο $\underset{\sim}{\theta} = \hat{\underset{\sim}{\theta}}$. Ο Binder (1983) έδωσε συνθήκες κανονικότητας για την εγκυρότητα της $\Sigma(b_i g_i(\underset{\sim}{b})) \underset{\sim}{u}_i(\hat{\underset{\sim}{\theta}}(\underset{\sim}{b})) = \underset{\sim}{0}$. Υπογραμμίζοντας ότι $\hat{\underset{\sim}{S}}(\underset{\sim}{\theta})$ είναι ένα διάνυσμα των εκτιμώμενων συνόλων με $d_i(s) g_i(\underset{\sim}{d}(s))$, από τα προηγούμενα προκύπτει:

$$\underset{\sim}{v}_L(\hat{\underset{\sim}{\theta}}) = \underset{\sim}{v}(\underset{\sim}{z}) \qquad (4.18)$$

όπου $\underset{\sim}{z}_k = \left[ \hat{\underset{\sim}{J}}(\hat{\underset{\sim}{\theta}}) \right]^{-1} g_k(\underset{\sim}{d}(s)) \underset{\sim}{e}_k^*$ , με $\underset{\sim}{e}_k^* = (e_{kl}^*, ..., e_{kp}^*)^T$ και $e_{kj}^* = u_{kj}(\hat{\theta}) - \underset{\sim}{x}_k^T \hat{\underset{\sim}{B}}_{ju}; \; j = 1, ...., p$

Επιπλέον, το $\hat{\underset{\sim}{B}}_{ju}$ το βρίσκουμε από $\hat{\underset{\sim}{B}}_j$ αλλάζοντας το $y_i$ σε $u_{ij}(\hat{\underset{\sim}{\theta}})$ και $\underset{\sim}{v}(\underset{\sim}{z})$ είναι η εκτιμώμενη μήτρα συνδιακύμανσης του διανύσματος των εκτιμώμενων συνόλων $\hat{\underset{\sim}{Z}} = \Sigma d_i(s) \underset{\sim}{z}_i$. Τα παραπάνω αποτελέσματα μπορούμε να τα βρούμε πιο άμεσα εάν

γράψουμε το $\hat{\underset{\sim}{\theta}}$ ως $f(\underset{\sim}{d}(s))$ και εκτιμώντας το $z_k = \partial f(\underset{\sim}{b}) / \partial b_k \big|_{\underset{\sim}{b}=\underset{\sim}{d}(s)}$. Δηλώνουμε το $\hat{\underset{\sim}{\theta}}(\underset{\sim}{b}) = f(\underset{\sim}{b})$ ως τη λύση του $\Sigma(b_i g_i(\underset{\sim}{b}))\underset{\sim}{u}_i(\underset{\sim}{\theta}) = \underset{\sim}{0}$. Παίρνουμε τώρα την παράγωγο του $\Sigma(b_i g_i(\underset{\sim}{b}))\underset{\sim}{u}_i(\hat{\underset{\sim}{\theta}}(\underset{\sim}{b})) = \underset{\sim}{0}$ σε σχέση με το $b_k$ για να προκύψει:

$$\Sigma\left[\partial(b_i g_i(\underset{\sim}{b})) / \partial b_k\right] u_i(\hat{\underset{\sim}{\theta}}(\underset{\sim}{b})) + \Sigma(b_i g_i(\underset{\sim}{b}))\left[\partial \underset{\sim}{u}_i(\hat{\underset{\sim}{\theta}}(b)) / \partial(\hat{\theta}(\underset{\sim}{b}))\right]\partial(\hat{\underset{\sim}{\theta}}(\underset{\sim}{b})) / \partial b_k \quad (4.19)$$

Αντικαθιστώντας τους τύπους (4.14) και (4.15) για $\partial(b_i g_i(\underset{\sim}{b})) / \partial b_k$ στην ακριβώς παραπάνω εξίσωση μας προκύπτει $\underset{\sim}{z}_k = \left[\hat{\underset{\sim}{J}}(\hat{\underset{\sim}{\theta}})\right]^{-1} g_k(\underset{\sim}{d}(s))\underset{\sim}{e}_k^*$ (4.20) μετά από απλοποιήσεις (Demnati και Rao, 2002).

### 4.1.3 Πλεονεκτήματα της γραμμικοποίησης :

Συγκρίνοντας τη γραμμικοποίηση με τις εναλλακτικές μεθόδους που βασίζονται στην αντιγραφή, οι μέθοδοι γραμμικοποίησης προσφέρουν αρκετά πλεονεκτήματα .

- Οι μέθοδοι γραμμικοποίησης είναι υπολογιστικά αποδοτικοί. Καθώς δεν απαιτούν την επανειλημμένη εφαρμογή της διαδικασίας εκτίμησης, αλλά μάλλον μόνο έναν υπολογισμό, οπότε απαιτούνται πολλοί λιγότεροι υπολογισμοί.
- Με την προσέγγιση της γραμμικοποίησης είναι σε γενικές γραμμές πιο εφικτό να επιτύχει τους μέγιστους βαθμούς ελευθερίας για την εκτίμηση της διακύμανσης (δηλαδή, τη μέγιστη ακρίβεια στην εκτίμηση της διακύμανσης δεδομένου του σχεδιασμού και του εκτιμητή). Συχνά οι διαδικασίες που βασίζονται στην αντιγραφή πρέπει να συμβιβαστούν με την ακρίβεια της εκτίμησης της διακύμανσης, προκειμένου να επιτευχθεί μια υπολογιστικά εφικτή μορφή. Αυτό συμβαίνει ιδιαίτερα σε περιπτώσεις που τα δείγματα έχουν συλλεχτεί από ένα στάδιο με το μέγεθος των δειγμάτων να αποτελείται από χιλιάδες μονάδες.
- Η γραμμικοποίηση μπορεί να παρέχει εκτιμητές διακύμανσης που αντιμετωπίζουν με τον πιο ενδεδειγμένο τρόπο όλα τα ενδεχόμενα που παρουσιάζει ένας σύνθετος σχεδιασμός. Αυτό είναι πιο δύσκολο για τις μεθόδους που βασίζονται στη αντιγραφη. Για τη γραμμικοποίηση ένα μόνο πράγμα είναι απαράιτητο να γίνει, πρέπει να γραμμικοποίηθεί ο εκτιμητής κατάλληλα με αμεροληψία, ή τουλάχιστον συνεπεία, ο εκτιμητής διασποράς είναι κατάλληλος για χρήση όταν οι γραμμικοί εκτιμητές αντανακλούν πλήρως στα χαρακτηριστικά που φέρει ο σχεδιασμός. Αυτό είναι δύσκολο με τις διαδικασίες αντιγραφής, ιδίως για πολλαπλών σταδίων σχεδιασμό χρησιμοποιώντας δειγματοληψία χωρίς αντικατάσταση. Συχνά, η λύση είναι να μεταχειριζόμαστε τις μονάδες του πρώτου σταδίου όπως επιλέγονται με την αντικατάσταση, και σε ορισμένες περιπτώσεις αυτό μπορεί να οδηγήσει σε σημαντική μεροληψία στον εκτιμητή της διακύμανσης.

### 4.1.4 Μειονεκτήματα της γραμμικοποίησης

Ωστόσο, υπάρχουν λόγοι για τους οποίους οι γραμμικοποιημένοι μέθοδοι δεν είναι οι μόνοι στη χρήση, και γιατί συνεχίζονται και διεξάγονται έρευνες της έρευνας που σχετίζονται με την εκτίμηση της διακύμανσης γενικότερα και με τη γραμμικοποίηση ειδικότερα. Αυτές περιλαμβάνουν τα ακόλουθα:

- Η μορφή του εκτιμητή της διασποράς διαφέρει για τους διαφορετικούς εκτιμητές των παραμέτρων. Ενώ με την αντιγραφή μια ενιαία προσέγγιση μπορεί να χρησιμοποιηθεί για μια ευρεία ποικιλία εκτιμητών, με τη γραμμικοποίηση πρέπει κανείς να αντλήσει το γραμμικοποιημένο εκτιμητή για κάθε διαφορετική μορφή. Αυτό είναι κουραστικό, αν κάποιος έχει μια έρευνα για την οποία χρησιμοποιούνται πολλοί διαφορετικοί τύποι εκτιμητών των παραμέτρων, και μπορεί να μην είναι πρακτικό για πολλούς δευτερεύοντες χρήστες.
- Κάποιος μπορεί να δυσκολευτεί κατά τον υπολογισμό του σωστού γραμμικοποίημένου εκτιμητή διακύμανσης για κάθε εκτιμητή παράμετρου που τον ενδιαφέρει . Ο εκτιμητής της παραμέτρου μπορεί να λάβει μια πολύ σύνθετη μορφή, σε πολλές περιπτώσεις, και μπορεί να ορίζεται μέσω μια πεπλεγμένης συνάρτησης. Το να βρίσκεις τον κατάλληλο εκτιμητή διασποράς μπορεί να αποτελεί μια πρόκληση, και είναι δύσκολο να επιβεβαιωθεί αν κάποιος έχει τη σωστή μορφή. Και πάλι, αυτή η δυσκολία μπορεί να είναι απαγορευτική για τους δευτερεύοντες χρήστες.
- Συχνά για ένα συγκεκριμένο εκτιμητή παραμέτρου είναι διαθέσιμες περισσότερες από μία μορφή γραμμικού εκτιμητή διασποράς. Πώς μπορεί κάποιος να ξέρει ποια είναι η καλύτερη για να τη χρησιμοποιήσει;
- Σε πολλές εφαρμογές της έρευνας ορισμένα από τα δεδομένα είναι τεκμαρτές τιμές λόγω των δεδομένων που λείπουν κατά τη συλλογή δεδομένων. Είναι γνωστό ότι αγνοώντας το γεγονός ότι τα δεδομένα είναι τεκμαρτά οδηγεί σε μια προκατάληψη στην εκτίμηση της διακύμανσης. Η προσπάθεια να βρεθούν μέθοδοι για να εκτιμηθεί σωστά η διακύμανση της δειγματοληψίας με την παρουσία των τεκμαρτών δεδομένων είναι μια συνεχής ερευνητική πρόκληση, ειδικά στην περίπτωση της γραμμικοποίησης των εκτιμητών της διακύμανσης (Rust, 2007).

## 4.2 Εκτίμηση της διακύμανσης με βάση τη μέθοδο της επαναλαμβανόμενης αντιγραφής Jackknife.

Η ιδέα της μεθόδου Jackknife εισήχθη από τον Quenouille (1949) ως μια τεχνική περιορισμού της μεροληψίας μιας εκτίμησης και αργότερα ο Tukey (1958) πρότεινε πως η ίδια διαδικασία θα μπορούσε να χρησιμοποιηθεί για την εκτίμηση της διακύμανσης. Ο Durbin (1959) πρώτος χρησιμοποίησε αυτή τη μέθοδο στην πρωτοποριακή εργασία του σχετικά με την εκτίμηση του λόγου/αναλογίας.

Αργότερα, εφαρμόστηκε για τον υπολογισμό της διακύμανσης σε πολύπλοκες έρευνες από τον Frankel (1971) με τον ίδιο τρόπο όπως η μέθοδος της ισορροπημένης επαναλαμβανόμενης αντιγραφής (BRR) και ονομάστηκε Jackknife επαναλαμβανόμενη αντιγραφή (JRR). Όπως στη μέθοδο BRR, η JRR μέθοδος γενικά εφαρμόζεται σε PSU εντός στρωμάτων. Οι βασικές αρχές της μεθόδου JRR απεικονίζονται εκτιμώντας τη δειγματική διασπορά του μέσου του δείγματος, ενός απλού τυχαίου δείγματος. Υποθέτουμε ότι n=5 και οι τιμές του δείγματος για y είναι 3,5,2,1,και 4. Η μέση τιμή δείγματος τότε είναι y = 3, και η διακύμανση της δειγματοληψίας αγνοώντας το FCP είναι :

$$u(\bar{y}) = \frac{\Sigma(y_i - \bar{y})^2}{n(n-1)} = 0.5 \qquad (4.21)$$

Η εκτίμηση της διακύμανσης του μέσου με τη Jackknife μέθοδο επιτυγχάνεται ως εξής.

1. Υπολογίζουμε ένα ψευδό-μέσο του δείγματος διαγράφοντας την πρώτη τιμή του δείγματος, με αποτέλεσμα : $\bar{y}_{(1)} = \frac{(5+2+1+4)}{4} = \frac{12}{4}$. Τώρα, με τη διαγραφή της δεύτερης τιμής του δείγματος, παίρνουμε το δεύτερο ψευδό-μέσο $\bar{y}_{(2)} = \frac{10}{4}$, ομοίως, $\bar{y}_{(3)} = \frac{13}{4}, \bar{y}_{(4)} = \frac{14}{4}, \bar{y}_{(5)} = \frac{11}{4}$.
2. Υπολογίζουμε τη μέση τιμή των πέντε ψευδό-τιμών : $\bar{\bar{y}} = \frac{\Sigma \bar{y}_{(i)}}{n} = \frac{\frac{60}{4}}{5} = 3$
3. Η διασπορά μπορεί στη συνέχεια να υπολογιστεί από τη μεταβλητότητα μεταξύ των πέντε ψευδό-μέσων, καθένα από τα οποία περιέχει τέσσερις παρατηρήσεις,

$$u(\bar{\bar{y}}) = \frac{(n-1)\Sigma(\bar{y}_{(i)} - \bar{\bar{y}})^2}{n} = 0.5 \qquad (4.22)$$

Η εξίσωση αυτή δίνει το ίδιο αποτέλεσμα με την εξίσωση (4.21).

Οι μέθοδοι που βασίζονται στην αντιγραφή έχουν ένα σαφές πλεονέκτημα: Μπορούν να εφαρμοστούν σε εκτιμήσεις που δεν εκφράζονται σε όρους που σχετίζονται με τύπους όπως είναι η διάμεσος του δείγματος, καθώς και για εκτιμήσεις που βασίζονται καθαρά σε τύπους. Κανένας τύπος δεν είναι διαθέσιμος για την εκτίμηση της διακύμανσης της διαμέσου μιας δειγματοληψίας, αλλά η Jackknife μέθοδος μπορεί να προσφέρει μια εκτίμηση. Χρησιμοποιώντας το ίδιο παράδειγμα όπως παραπάνω, η διάμεση τιμή του δείγματος είναι 3 και οι πέντε ψευδό-διάμεσοι είναι 3, 2.5, 3.5, 3.5, και 2.5 (ο μέσος αυτών των ψευδό-διαμέσων είναι 3). Η διακύμανση του μέσου υπολογίζεται ως 0.8, χρησιμοποιώντας την εξίσωση (4.23).

Με τον ίδιο τρόπο, η Jackknife μέθοδος μπορεί επίσης να εφαρμοστεί στην επαναλαμβανόμενη δειγματοληψία. Μπορούμε να αφαιρέσουμε ένα αντίγραφο κάθε φορά και να υπολογίζουμε ψευδό-τιμές για την εκτίμηση της Jackknife διακύμανσης, αν και σε αυτή την περίπτωση αυτή η διαδικασία δεν προσφέρει κανένα υπολογιστικό πλεονέκτημα. Επιπλέον όμως μπορεί να εφαρμοστεί σε οποιεσδήποτε τυχαίες ομάδες που σχηματίζονται από οποιοδήποτε δείγμα πιθανότητας. Για παράδειγμα, ένα συστηματικό δείγμα μπορεί να διαιρεθεί σε τυχαίες ή συστηματικές υποομάδες για την Jackknife μέθοδο. Για άλλες περιπτώσεις δειγματοληπτικού σχεδιασμού μπορούν να σχηματιστούν τυχαίες ομάδες μετά από τους πρακτικούς κανόνες που προτείνονται από τον Wolter (1985). Η βασική ιδέα είναι να σχηματίσουμε τυχαίες ομάδες κατά τέτοιο τρόπο ώστε κάθε τυχαία ομάδα να έχει τον ίδιο τρόπο δειγματοληπτικού σχεδιασμού όπως το αρχικό δείγμα. Αυτό απαιτεί λεπτομερείς πληροφορίες σχετικά με τον τρόπο που έγινε ο δειγματοληπτικός σχεδιασμός, οι πληροφορίες αυτές όμως συνήθως δεν είναι διαθέσιμες για της περισσότερες έρευνες οι προσφέρονται για δημόσια χρήση. Η Jackknife μέθοδος, ως εκ τούτου, εφαρμόζεται συνήθως σε PSU και όχι σε τυχαίες ομάδες.

Ας υποθέσουμε ότι $\hat{\theta}$ είναι ένας εκτιμητής μιας παραμέτρου $\theta$ ενός πληθυσμού, ο οποίος βασίζεται σε όλα τα δεδομένα $y_1, y_2 \ldots, y_n$ ή αν έχουμε ($k$) ομάδες $(g_1, g_2 \ldots, g_k)$ παρατηρήσεων $y_i$

$$\hat{\theta} = \theta(g_1, g_2, \ldots, g_k)$$

Έστω ότι $\hat{\theta}_{(i)}$ είναι ένας εκτιμητής της παραμέτρου θ ο οποίος παράγεται με την ίδια μαθηματική έκφραση όπως ο εκτιμητής $\theta$ , αλλά βασίζεται σε όλα τα δεδομένα εκτός των δεδομένων στην ($i$) ομάδα

$$\hat{\theta}_{(i)} = \theta(g_1, g_2, \ldots, g_{i-1}, g_{i+1}, \ldots, g_k),$$

Οι δειγματικοί συντελεστές $w_i$ των δεδομένων που μένουν στο δείγμα αν παραληφθεί η ($i$) ομάδα πολλαπλασιάζονται επί $\left(\frac{k}{k-1}\right)$, όπου $k$ είναι ο αριθμός των ομάδων, όταν υπολογίζεται ο εκτιμητής $\hat{\theta}_{(i)}$… Αυτός ο επαναπροσδιορισμός των βαρών $(w_i)$ είναι απαραίτητος για παραμέτρους που περιλαμβάνουν σύνολα (Τ), αλλά όχι για παραμέτρους που περιλαμβάνουν μέσους $(\bar{X})$ στη στρωματοποιημένη δειγματοληψία. Ο εκτιμητής Jackknife της διασποράς του εκτιμητή $\hat{\theta}_{JK}$ αλλά και της διασποράς του αρχικού εκτιμητή $\hat{\theta}$ της παραμέτρου θ δίνεται από τον τύπο του Efron (1982):

$$\hat{\text{var}}_{JK}(\hat{\theta}) = \frac{k-1}{k}\sum_{i=1}^{\kappa}(\hat{\theta}_{(\iota)} \hat{-} \hat{\theta})^2$$

ή

$$\hat{\text{var}}_{JK}(\hat{\theta}) = \frac{1}{k(k-1)}\sum_{i=1}^{\kappa}\hat{(}\hat{\theta}_{\iota} - \bar{\hat{\theta}})^2$$

όπου $\hat{\theta}_i = k\hat{\theta} - (k-1)\hat{\theta}_{(i)}$,

$\hat{\theta}_{(i)} = \theta(g_1, g_2, \ldots . g_{i+1}, \ldots, g_k), \quad i = 1, 2, \ldots . , k$

$$\bar{\hat{\theta}} = \frac{1}{k}\Sigma^{k}\hat{\theta}_i$$

Ο εκτιμητής Jackknife της παραμέτρου θ δίνεται από τον τύπο

$$\bar{\hat{\theta}} = \frac{1}{k}\Sigma^{k}\hat{\theta}_i$$

### 4.2.1 Πλεονεκτήματα της Μεθόδου Jackknife

Τα πλεονεκτήματα της μεθόδου εκτιμητικής Jackknife είναι τα εξής:

- Αν υπάρχει ένας εκτιμητής $\hat{\theta}$ , της παραμέτρου θ, ο οποίος έχει αμεροληψία τάξης $\frac{1}{n}$ , δηλαδή ισχύει:

  $$E\hat{(}\hat{\theta}) = \theta + \frac{\text{A}}{n} + o\left(\frac{1}{n^2}\right)$$

  τότε ο εκτιμητής Jackknife της παραμέτρου θ, το οποίο θα συμβολίσουμε με $\bar{\hat{\theta}}$ ή $\hat{\theta}_{JK}$ έχει αμεροληψία μικρότερη από την αμεροληψία του εκτιμητή $\hat{\theta}$, δηλαδή

  $$E\hat{(}\hat{\theta}_{JK}) = \theta + \frac{\text{B}}{n^2} + o\left(\frac{1}{n^3}\right)$$ (Quenouille 1956)

- Η έκφραση

  $$\hat{\text{var}}_{JK}(\hat{\theta}) = \frac{1}{k(k-1)}\hat{\sum}_{i=1}^{k}\hat{(}\hat{\theta}_{\iota} - \bar{\hat{\theta}})^2$$

  $$\hat{\text{var}}_{JK}\hat{(}\hat{\theta}) = \frac{k-1}{k}\hat{\sum}_{i=1}^{k}(\hat{\theta}_{(\iota)} \hat{-} \hat{\theta})^2$$

  Όπου $\hat{\theta}_{\iota} = k\hat{\theta} - (k-1)\hat{\theta}_{(\iota)}$

$$\bar{\hat{\theta}} = \frac{1}{k}\sum^{k}\hat{\theta}_t$$

Είναι ένας μη παραμετρικός εκτιμητής της διασποράς του Jackknife εκτιμητή $\bar{\hat{\theta}}$, αλλά και του αρχικού εκτιμητή $\hat{\theta}$ της παραμέτρου θ (Tukey, 1958). Σύμφωνα με την θεμελιώδη εικασία του Tukey (1958) αλλά και με αναλυτική απόδειξη άλλων ερευνητών (Durbin, 1959, Frangos 1980, 1983, 1984, 1986, 1990, 1995), οι εκφράσεις

$$\hat{\theta}_t = \hat{k\theta} - (k-1)\hat{\theta}_{(i)}$$

όπου $\hat{\theta}_{(i)} = \theta(g_1, g_2, ...., g_{i-1}, g_{i+1}, ....g_k)$, $i = 1, 2, ...., k$

οι οποίες λέγονται ψευδοτιμές του εκτιμητή Jackknife, είναι προσεγγιστικά ανεξάρτητες μεταξύ τους και έχουν ως Κατανομή Πιθανότητας την Κατανομή Student's t (Φράγκος , 1998). Το αποτέλεσμα αυτό μπορεί να εφαρμοστεί για τις τιμές $\hat{\theta}_{(i)}$ , έτσι ώστε να κατασκευασθεί το ακόλουθο ανθεκτικό διάστημα εμπιστοσύνης 100(1-α)% για τη παράμετρο $\theta$ (για $k \geq 30$ ).

$$\bar{\hat{\theta}} - z_{a/2}\sqrt{\hat{\theta}_{JK}(\hat{\theta})} \leq \theta \leq \bar{\hat{\theta}} + z_{a/2}\sqrt{\hat{\text{var}}_{JK}(\hat{\theta})}$$

Η JRR δεν περιορίζεται στην επιλογή ενός συνδυασμένου σχεδίου, αλλά μπορεί να εφαρμοστεί σε οποιοδήποτε αριθμό PSUs ανά στρώμα. Εάν η εκτίμηση του U γίνεται από το $u_{hi}$ ( όπου $h$ είναι το στρώμα και $i$ η επανάληψη), $n_h$ είναι ο αριθμός των PSUs του δείγματος στο στρώμα $h$ και $r_h$ είναι ο αριθμός των επαναλήψεων που σχηματίζονται στο στρώμα $h$ τότε η διακύμανση υπολογίζεται από :

$$\upsilon(\bar{u}) = \sum_{h}^{L}\left(\frac{n_h - 1}{r_h}\right)\sum_{i}^{r_h}\left(u_{hi} - \bar{u}\right)^2$$

Εάν κάθε ένα από τα PSUs στο στρώμα $h$ αφαιρείται για να σχηματίσει ένα αντίγραφο, το $r_h = n_h$ σε κάθε στρώμα. Όταν ο αριθμός των στρωμάτων είναι μεγάλος και τα $n_h$ είναι δυο ή περισσότερα, ο υπολογισμός μπορεί να μειωθεί με τη χρήση μόνο μίας αντιγραφής σε κάθε στρώμα. Ωστόσο, ένας επαρκής αριθμός αντιγραφών θα πρέπει να γίνεται σε αναλυτικές μελέτες για να εξασφαλιστούν επαρκείς βαθμοί ελευθερίας.

Η μέθοδος jackknife μπορεί να χρησιμοποιηθεί για να εκτιμηθεί η μεροληψία ενός εκτιμητή εκτιμώμενο πάνω σε ολόκληρο το δείγμα.

$$\bar{\theta}_{Bias} = k\bar{\theta} - (k-1)\bar{\theta}_{JK}$$

Αυτό μειώνει τη μεροληψία κατά μία τάξη μεγέθους, από $O(k^{-1}) \Rightarrow \text{O}(k^{-2})$

Αυτό παρέχει μια εκτιμώμενη διόρθωση της μεροληψίας λόγω της μεθόδου εκτίμησης. Η jackknife μέθοδος δεν είναι σωστή για αμερόληπτο δείγμα (Cameron και Trivedi, 2005).

Υπολογίζοντας το τυπικό σφάλμα με τη Jackknife μέθοδο εξάγουμε σχεδόν τα ίδια αποτελέσματα με τη γραμμική μέθοδο. Συγκρίνοντας τα παρακάτω αποτελέσματα που αφορούν την ερώτηση για το αν οι μαθητές έχουν δοκιμάσει η όχι τσιγάρο κατανοούμε ότι οι δυο μέθοδοι μας δίνουν με μια ελάχιστη απόκλιση ίδιες τιμές όσο αναφορά την επίδραση του δειγματοληπτικού σχεδιασμού, παρόλο το γεγονός ότι η Jackknife μέθοδος αντιπροσωπεύει μια διαφορετική στρατηγική η οποία χρησιμοποιεί μια διαφορετική μέθοδο για την εκτίμηση της διακύμανσης. Αποτελέσματα από το πρόγραμμα STATA εφαρμόζοντας τη Jackknife μέθοδο βρίσκονται στο παράρτημα 2.

Παρακάτω βρίσκονται οι αντίστοιχοι πίνακες για τη Jackknife μέθοδο. Εφόσον οι τιμές είναι ίδιες δεν κρίνεται απαραίτητο να γίνει κάποιος σχολιασμός παρά μόνο ότι ο αριθμός των αντιγραφών ήταν 2030 που αντιστοιχεί στον αριθμό των βασικών μονάδων της έρευνας οι οποίες μονάδες αυτές είναι οι σχολικές τάξεις. Δηλαδή κάθε φορά στον υπολογισμό αφαιρείται μια σχολική τάξη και όχι μεμονωμένα ένας μαθητής.

Πίνακας 5.

Υπολογισμός τυπικού σφάλματος για την απλή τυχαία δειγματοληψία και με την Jackknife μέθοδο για το ποσοστό των μαθητών απάντησαν ‘’όχι’’ στο κάπνισμα

ESPAD 2011

| | | Proportion | Std. Err. | jackknife Std. Err. Class + Nomos | Design Effect | Number of obs | Number of obs after Design Effect |
|---|---|---|---|---|---|---|---|
| c08 | Κάπνισες ποτέ τσιγάρο | 0,6087 | 0,003338 | 0,007182 | 4,6292 | 35.462 | 7.661 |
| c09 | Πόσο κάπνισες τις τελευταίες 30 ημέρες | 0,8590 | 0,002464 | 0,004673 | 3,5977 | 35.553 | 9.882 |

Πίνακας 6.

Υπολογισμός τυπικού σφάλματος για την απλή τυχαία δειγματοληψία και με την Jackknife μέθοδο για το ποσοστό των μαθητών που απάντησαν ‘’όχι’’ στην κατανάλωση αλκοόλ .

ESPAD 2011

| | | Proportion | Std. Err. | jackknife Std. Err. Class + Nomos | Design Effect | Number of obs | Number of obs after Design Effect |
|---|---|---|---|---|---|---|---|
| c12 | Ήπιες ποτέ αλκοολούχο ποτό | | | | | | |
| c12a | Σε όλη σου τη ζωή μέχρι και σήμερα | 0,2604 | 0,002922 | 0,006445 | 4,8668 | 34.632 | 7.116 |
| c12b | Στη διάρκεια των 12 τελευταίων μηνών μέχρι και σήμερα | 0,3980 | 0,003299 | 0,007711 | 5,4627 | 34.945 | 6.397 |
| c12c | Στη διάρκεια των 30 τελευταίων ημερών μέχρι και σήμερα | 0,3922 | 0,003296 | 0,006804 | 4,2606 | 35.040 | 8.224 |
| c18 | Τις 30 τελευταίες ημέρες πόσες φορές ήπιες στη σειρά 5 ή περισσότερα αλκοολούχα ποτά | 0,5486 | 0,004175 | 0,005782 | 1,9180 | 23.034 | 12.009 |
| | Πόσες φορές μέθυσες από αλκοολούχα ποτά | | | | | | |
| c19a | Σε όλη σου τη ζωή μέχρι και σήμερα | 0,9916 | 0,000598 | 0,000688 | 1,3232 | 35.191 | 26.595 |
| c19b | Στη διάρκεια των 12 τελευταίων μηνών μέχρι και σήμερα | 0,9964 | 0,000512 | 0,000548 | 1,1486 | 23.000 | 20.024 |
| c19c | Στη διάρκεια των 30 τελευταίων ημερών μέχρι και σήμερα | 0,9983 | 0,000344 | 0,000341 | 0,9878 | 23.105 | 23.390 |

Πίνακας 7.

Υπολογισμός τυπικού σφάλματος για την απλή τυχαία δειγματοληψία και με την Jackknife μέθοδο για το ποσοστό των μαθητών που απάντησαν ‘’όχι’’ στη χρήση μαριχουάνας η χασίς.

ESPAD 2011

| | | Proportion | Std. Err. | jackknife Std. Err. Class + Nomos | Design Effect | Number of obs | Number of obs after Design Effect |
|---|---|---|---|---|---|---|---|
| c25 | Δοκίμασες ή πήρες ποτέ μαριχουάνα ή χασίς | | | | | | |
| c25a | Σε όλη σου τη ζωή μέχρι και σήμερα | 0,9802 | 0,001069 | 0,001502 | 1,9750 | 35.548 | 17.999 |
| c25b | Στη διάρκεια των 12 τελευταίων μηνών μέχρι και σήμερα | 0,9846 | 0,001150 | 0,001462 | 1,6168 | 23.127 | 14.304 |
| c25c | Στη διάρκεια των 30 τελευταίων ημερών μέχρι και σήμερα | 0,9917 | 0,000856 | 0,001075 | 1,5752 | 23.158 | 14.702 |
| | Δοκίμασες ή πήρες ποτέ έκσταση | | | | | | |
| c29a | Σε όλη σου τη ζωή μέχρι και σήμερα | 0,9979 | 0,000326 | 0,000323 | 0,9817 | 35.343 | 36.004 |

Πίνακας 8.

Υπολογισμός τυπικού σφάλματος για την απλή τυχαία δειγματοληψία και με την Jackknife μέθοδο για το ποσοστό των μαθητών που απάντησαν ‘’όχι στη χρήση διάφορων ουσιών.

ESPAD 2011

| | | Proportion | Std. Err. | jackknife Std. Err. Class + Nomos | Design Effect | Number of obs | Number of obs after Design Effect |
|---|---|---|---|---|---|---|---|
| c31 | Έχεις χρησιμοποιήσει ποτέ κάποια από τις παρακάτω ουσίες | | | | | | |
| c31a | Ηρεμιστικά ή υπνωτικά χωρις τη σύσταση γιατρού | 0,9962 | 0,000426 | 0,000454 | 1,1363 | 35.489 | 31.233 |
| c31b | Αμφεταμίνες | 0,9987 | 0,000303 | 0,000299 | 0,9751 | 23.155 | 23.747 |
| c31c | LSD ή κάποιο άλλο παραισθησιογόνο | 0,9983 | 0,000385 | 0,000453 | 1,3832 | 23.093 | 16.695 |
| c31d | Κράκ | 0,9985 | 0,000327 | 0,000379 | 1,3413 | 23.086 | 17.211 |
| c31e | Κοκαινη | 0,9969 | 0,000514 | 0,000564 | 1,2040 | 23.155 | 19.231 |
| c31f | Ρελιβίνη | 0,9989 | 0,000279 | 0,000277 | 0,9914 | 23.101 | 23.301 |
| c31g | Ηρωίνη | 0,9980 | 0,000320 | 0,000320 | 0,9994 | 23.142 | 23.156 |
| c31h | 'Μαγικά μανιτάρια'' | 0,9975 | 0,000391 | 0,000411 | 1,1039 | 23.076 | 20.904 |
| c31i | GHB | 0,9992 | 0,000232 | 0,000254 | 1,1986 | 23.018 | 19.205 |
| c31j | Αναβολικά χωρίς τη σύσταση γιατρού | 0,9984 | 0,000318 | 0,000312 | 0,9577 | 23.097 | 24.116 |
| c31k | Ναρκωτικά με ένεση | 0,9986 | 0,000245 | 0,000245 | 1,0024 | 23.142 | 23.085 |
| c31l | Κάποιο αλκοολούχο ποτό μαζι με φάρμακα για να αλλάξεις τη διάθεση σου | 0,9983 | 0,000368 | 0,000372 | 1,0235 | 23.187 | 22.654 |

# 5 Παλινδρόμηση

Ο όρος παλινδρόμηση εισήχθη το 1877 από τον Francis Galton [6] στην προσπάθειά του να ερμηνεύσει τη σχέση του ύψους μεταξύ γονιών και παιδιών. Ο όρος ‘’παλινδρόμηση’’[7] σήμαινε αρχικά την παλινδρόμηση προς τη μέση τιμή, οπότε τα ‘’μοντέλα παλινδρόμησης’’ και η ‘’ανάλυση παλινδρόμησης’’ αναφέρονται στη σχέση μεταξύ μιας εξαρτημένης παραμέτρου π.χ. του μέσου μιας ποσότητας και ενός συνόλου ποσοτικών ανεξάρτητων μεταβλητών. Σήμερα, πάντως, ο όρος ‘’παλινδρόμηση’’ αναφέρεται στη σχέση μιας παραμέτρου με ένα σύνολο προσδιοριστών ανεξάρτητα από τη φύση τους. Η παλινδρόμηση είναι μια στατιστική τεχνική μοντελοποίησης για την έρευνα της συσχέτισης μεταξύ μίας εξαρτημένης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών. Τα μοντέλα παλινδρόμησης περιλαμβάνουν τις ακόλουθες μμεταβλητές:

- Οι άγνωστες παράμετροι συσχέτισης που δηλώνονται ως $\beta$ (διάνυσμα).
- Οι ανεξάρτητες μεταβλητές $X$ (διάνυσμα).
- Η εξαρτημένη μεταβλητή $Y$ .

Ένα μοντέλο παλινδρόμησης συσχετίζει το Y σε μία συνάρτηση παλινδρόμησης των $X$ και $\beta$ . $Y \simeq F(X,\beta)$ . Ο συνήθης τύπος είναι $E(Y/X) = f(X,\beta)$.Η Ανάλυση παλινδρόμησης μας βοηθά να κατανοήσουμε τη μεταβολή της εξαρτημένης μμεταβλητής Y όταν μεταβάλλεται μία από τις ανεξάρτητες μεταβλητές X, ενώ οι άλλες ανεξάρτητες μεταβλητές παραμένουν σταθερές.

## 5.1 Απλή - Πολλαπλή γραμμική παλινδρόμηση

Το βασικό μαθηματικό μοντέλο που περιγράφει τη σχέση μεταξύ δύο μεταβλητών είναι η ευθεία γραμμή. Το γραμμικό μοντέλο για δύο μεταβλητές αποτελεί τη βάση για τη δημιουργία πιο σύνθετων μοντέλων μεταξύ περισσότερων μεταβλητών. Το μαθηματικό μοντέλο που χρησιμοποιείται είναι η απλή γραμμική παλινδρόμηση, καθώς η σχέση μεταξύ των δύο μεταβλητών περιγράφεται από μια ευθεία γραμμή σύμφωνα με την ακόλουθη ισότητα:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Στην παραπάνω ισότητα , το Y είναι η εξαρτημένη μεταβλητή της απλής γραμμικής παλινδρόμησης, ενώ το $X_1$ είναι η ανεξάρτητη μεταβλητή. Ουσιαστικά,

[6] Ο Francis Galton το 1885 προέβη στην εξήγηση του όρου ‘’παλινδρόμηση’’ στηριζόμενος στις αρχές της κανονικής κατανομής. Ο Galton το 1889 εξέδωσε το ‘’natural inheritance’’ συμπεριλαμβάνοντας την εξήγηση του για τον όρο ‘’παλινδρόμηση’’.
[7] Ο όρος ‘’παλινδρόμηση’’ συχνά αποδίδεται με τον όρο ‘’εξάρτηση’’.

το Υ αντιστοιχεί στη μελετώμενη έκβαση, ενώ το $X_1$ αντιστοιχεί στο μελετώμενο προσδιοριστή. Το $\beta_0$ είναι η σταθερά της απλής γραμμικής παλινδρόμησης και είναι η μέση τιμή που λαμβάνει η μεταβλητή Υ, όταν η μεταβλητή $X_1$ ισούται με 0. Το $\beta_1$ περιγράφει την κλίση της ευθείας γραμμής που συσχετίζει το $X_1$ με το Υ. Το $\beta_1$ είναι ο αναμενόμενος αριθμός των μονάδων που μεταβάλλεται το Υ κάθε φορά που η τιμή του $X_1$ μεταβάλλεται κατά μία μονάδα. Το $\varepsilon$ είναι το τυχαίο σφάλμα που αντιπροσωπεύει την τυχαία απόκλιση από την αναμενόμενη τιμή της εξαρτημένης μεταβλητής Υ. Η μέση τιμή του τυχαίου σφάλματος, γενικά, θεωρείται ίση με 0. Οι τιμές των $\beta_0$ και $\beta_1$ κυμαίνονται θεωρητικά από -∞ έως +∞. Είναι σπάνιο η πραγματική σχέση μεταξύ δύο μεταβλητών να είναι τελείως γραμμική. Η υπόθεση που ισχύει για την εφαρμογή ενός μαθηματικού μοντέλου σε μια ανάλυση είναι ότι το μοντέλο αυτό αποτελεί μια απλοποιημένη περιγραφή της σχέσης μεταξύ δύο μεταβλητών και δεν συμμορφώνεται αναγκαστικά με την πραγματική τους σχέση. Προσεγγίζει, ωστόσο, αρκετά την πραγματική σχέση μεταξύ των δύο μεταβλητών, με αποτέλεσμα να δικαιολογείται η χρήση του.

Το μοντέλο, ωστόσο, της γραμμικής παλινδρόμησης μπορεί να επεκταθεί, περιλαμβάνοντας περισσότερες από μία ανεξάρτητες μεταβλητές, οπότε προκύπτει το μοντέλο της πολλαπλής γραμμικής παλινδρόμησης.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + .....$$

Η ισότητα αυτή αντιστοιχεί και πάλι σε μια ευθεία γραμμή, αλλά η γραμμή αυτή διαγράφει στο χώρο ένα διάστημα με περισσότερες από δύο διαστάσεις, καθώς σε κάθε μεταβλητή αντιστοιχεί μία διάσταση. Όταν υπάρχουν δύο ανεξάρτητες μεταβλητές και μία εξαρτημένη, τότε απαιτείται η τρισδιάστατη απεικόνιση των δεδομένων στο χώρο, καθώς απαιτούνται δύο διαστάσεις για τις δύο ανεξάρτητες μεταβλητές και μία διάσταση για την εξαρτημένη μεταβλητή. Μαθηματικά, τουλάχιστον, δεν υπάρχει περιορισμός στον αριθμό των ανεξάρτητων μεταβλητών που μπορούν να χρησιμοποιηθούν στην πολλαπλή γραμμική παλινδρόμηση, αν και μικρός αριθμός δεδομένων επιτρέπει να χρησιμοποιηθούν λίγες μόνο ανεξάρτητες μεταβλητές (MEDNET, 2009).

## 5.2 Λογιστική παλινδρόμηση.

Η λογιστική παλινδρόμηση είναι μια μέθοδος πολυπαραγοντικής στατιστικής ανάλυσης που χρησιμοποιεί ένα σύνολο ανεξαρτήτων μεταβλητών για τη διερεύνηση της κίνησης μιας κατηγορικής εξαρτημένης μεταβλητής. Είναι χρήσιμη σε καταστάσεις στις οποίες επιθυμούμε την πρόβλεψη της ύπαρξης ή της απουσίας ενός χαρακτηριστικού ή ενός συμβάντος. Η πρόβλεψη αυτή βασίζεται στην

κατασκευή ενός γραμμικού μοντέλου και συγκεκριμένα στον προσδιορισμό των τιμών που παίρνουν οι συντελεστές ενός συνόλου ανεξάρτητων μεταβλητών που χρησιμοποιούνται ως μεταβλητές πρόβλεψης. Εκτός από την πρόβλεψη ένα μοντέλο λογιστικής παλινδρόμησης δίνει τη δυνατότητα να εκτιμήσουμε την επίδραση κάθε ανεξάρτητης μεταβλητής στη εξαρτημένη μεταβλητή. Η πιο διαδεδομένη βιβλιογραφικά έκφραση της λογιστικής παλινδρόμησης είναι:

$$\ln(odds) = a + b_1 x_1 + b_2 x_2 + .....b_k x_k$$

Το δεξί μέρος της εξίσωσης δημιουργείται από ένα γραμμικό συνδυασμό των ανεξάρτητων μεταβλητών που συμμετέχουν στο μοντέλο παλινδρόμησης. Το αριστερό μέρος περιέχει τις τιμές της εξαρτημένης μεταβλητής με τη μορφή του λογαρίθμου των odds, δηλαδή του λογαρίθμου της σχέσης $odds = \frac{\mathrm{P}}{(1-P)}$. Το odds και το P εκφράζει την πιθανότητα του συμβάντος του γεγονότος. Οι συντελεστές των ανεξάρτητων μεταβλητών στην εξίσωση παλινδρόμησης εκτιμούνται με βάση τη μεθόδου Μεγίστης Πιθανοφανείς. Σύμφωνα με τη μέθοδο αυτή η τιμή των συντελεστών των ανεξάρτητων μεταβλητών είναι αυτή που κάνει τις παραχωρηθείσες τιμές της εξαρτημένης μεταβλητής πιο πιθανές, βάσει του συνόλου των ανεξαρτήτων μεταβλητών.

Σε ένα γραμμικό μοντέλο η ανεξάρτητη μεταβλητή, μπορεί να χρησιμοποιηθεί ως συνεχής, ενώ ταυτόχρονα μπορούν να συμπεριληφθούν στο στατιστικό μοντέλο και άλλες μεταβλητές που πιθανών να επηρεάζουν την εξαρτημένη μεταβλητή. Σε ένα απλό μοντέλο γραμμικής παλινδρόμησης ο μέσος μιας συνεχούς απόκρισης y μπορεί να περιγραφεί ως μια σχέση με μια ανεξάρτητη μεταβλητή X (συνεχής ή κατηγορική) με την ακόλουθη σχέση :

$$E(y \mid X) = \beta_0 + \beta_1 X$$

Η βασική υπόθεση που πρέπει να ισχύει για να έχει νόημα το παραπάνω γραμμικό μοντέλο είναι ότι η εξαρτημένη μεταβλητή $y$ ακολουθεί κανονική κατανο. Το ερώτημα που ανακύπτει, είναι πως διατυπώνεται και πως εκτιμάται ένα υπόδειγμα στο οποίο η εξαρτημένη μεταβλητή είναι δίτιμη[8] ή όταν έχει παραπάνω τιμές. Αυτό που μας ενδιαφέρει όπως προαναφέρθηκε είναι η πρόβλεψη της ύπαρξης ή της απουσίας ενός χαρακτηριστικού, με άλλα λόγια η δεσμευομένη πιθανότητα $P(Y = 1 \mid X)$ να υπάρχει ή να απουσιάζει κάποιο χαρακτηριστικό με βάση μια οποιαδήποτε έκθεση X. Συνεπώς, θα χρειαζόταν ένα μοντέλο της μορφής :

$$P(Y = 1 \mid X) = \beta_0 + \beta_1 X$$

[8] Όταν οι επιλογές είναι δύο, παίρνει δηλαδή δύο τιμές 0 ή 1.

Στη περίπτωση αυτή δύο είναι τα προβλήματα που πρέπει να αντιμετωπιστούν. Πρώτον, το στατιστικό προβλήμα ότι η εξαρτημένη μεταβλητή δεν ακολουθεί κανονική κατανομή, και δεύτερον, το αριθμητικό πρόβλημα ότι το δεξί μέρος της εξίσωσης θα πρέπει να περιοριστεί να δίνει τιμές στο διωστήρα (0,1). Και τα δύο αυτά θέματα αντιμετωπίζονται με τη χρήση της λογιστικής παλινδρόμησης, μιας μεθοδολογίας που ανήκει στα Γενικευμένα Γραμμικά Μοντέλα.

### 5.2.1 Λογιστική συνάρτηση

Όπως είδαμε, το μοντέλο παλινδρόμησης που περιγράφει τον κίνδυνο εξαφάνισης μιας ασθένειας θα πρέπει να δίνει τιμές μέσα στο διάστημα (0,1). Θα πρέπει λοιπόν να χρησιμοποιηθεί ο κατάλληλος μαθηματικός μετασχηματιστές της δεξιάς πλευράς της εξίσωσης:

$$E\left(y \mid X\right) = \beta_0 + \beta_1 X$$

έτσι ώστε οποιαδήποτε εκτόνωση και αν προκύπτει για κάποια πρόβλεψη αυτό ποτέ να μη βρίσκεται κάτω από το μηδέν ή πάνω από το ένα. έστω $\eta = \beta_0 + \beta_1 X$ το συστηματικό μέρος της εξίσωσης. Η λογιστική συνάρτηση του $\eta$ ορίζεται ως :

$$f(\eta) = \frac{1}{1+\exp(-\eta)}$$

Η λογιστική συνάρτηση έχει μερικές πολύ χρήσιμες ιδιότητες για το λόγο αυτό κυριαρχεί στην ανάλυση κατηγορικών δεδηγμένων. Η συνάρτηση έχει σιγμοειδή μορφή, όταν το $\eta = -\infty$ τότε η $f(\eta) = 0$ ενώ όταν $\eta = \infty$ τότε $f(\eta) = 1$.

### 5.2.2 Το λογιστικό μοντέλο

Είδαμε τη παραπάνω λογιστική συνάρτηση (Ντζούφρας, 2009). Η τελευταία σχέση είναι ισοδύναμη με:

$$P(Y = 1 \mid X) = \frac{\exp(\eta)}{1+\exp(\eta)}$$

Η σχέση αυτή ορίζει το μοντέλο της λογιστικής παλινδρόμησης. Το μοντέλο μπορεί να γραφεί ως λόγος σχετικών πιθανοτήτων στη μορφή :

$$\frac{P(Y=1|X)}{1-P(Y=1|X)} = \exp(\beta_0 + \beta_1 X)$$

Παίρνοντας το λογάριθμο της αριστερής πλευράς της εξίσωσης καταλήγουμε στη σχέση:

$$\log\left[\frac{P(Y=1|X)}{1-P(Y=1|X)}\right] = \log it(P(Y=1|X) = \beta_0 + \beta_1 X = \eta$$

Το μοντέλο αυτό συνδέει το λογάριθμο του λόγου σχετικών πιθανοτήτων εμφάνισης η όχι ενός χαρακτηριστικού γραμμικά με την ανεξάρτητη μεταβλητή Χ. Για να δημιουργηθεί το μοντέλο αυτό της λογιστικής παλινδρόμησης το συστατικό μέρος της εξίσωσης $\eta$ συνδέεται με τη χρήση της λογιστικής συνάρτησης για το μετασχηματισμό της εξαρτημένης μεταβλητής. Για το λόγο αυτό η λογιστική συνάρτηση ονομάζεται εδώ συνάρτηση δεσμού και η $f(.) = \frac{\exp(.)}{(1+\exp(.))}$ είναι η αντίστροφη της συνάρτησης δεσμού.

Το αριστερό μέρος του μοντέλου $P(Y=1|X) = \frac{\exp(\eta)}{1+\exp(\eta)}$ ορίζει το στοχαστικό κομμάτι του υποδείγματος, την κατανομή της εξαρτημένης μεταβλητής $y$ δεδομένης μιας παρατηρηθείσας τιμής για την ανεξάρτητη μεταβλητή $x$. Εφόσον η $y$ εδώ είναι δίτιμη τότε ισχύουν οι ακόλουθες υποθέσεις για το μοντέλο λογιστικής παλινδρόμησης :

- Το $y_i$ ακολουθεί διωνυμική κατανομή (όπου ο δείκτης $i$ υποδηλώνει την $i$-οστή παρατήρηση)
- Ο μέσος $E(Y|X) = P(Y=1|X)$ δίνεται από τη λογιστική συνάρτηση
- Οι τιμές της εξαρτημένης μεταβλητής είναι στατιστικά ανεξάρτητες

## 5.3 Εκτίμηση εύρωστου τυπικού σφάλματος κατά Huber

Στην έρευνα μας η μέθοδος που χρησιμοποιήθηκε ήταν η δειγματοληψία κατά συστάδες. Τον πληθυσμό της έρευνας αποτελούν οι μαθητές οι οποίοι χωρίστηκαν σε ομάδες- τάξεις. Αυτό το είδος των δεδομένων που έχουν εντός μια τάξεως συσχετισμούς ενσωματωμένα στη δομή δεδομένων, πρέπει να ληφθούν υπόψη στην εκτίμηση των παραμέτρων. Θα περιγράψουμε τη μέθοδο του Huber (1967), επίσης γνωστή ως White ή Sandwich μέθοδος, για λογιστικά μοντέλα όσον αφορά τον υπολογισμό της εύρωστης τυπικής εκτίμησης σφαλμάτων για δεδομένα που ακολουθήθηκε η μέθοδος της δειγματοληψίας κατά συστάδες. Το παραδοσιακό πρότυπο εκτίμησης σφάλματος με τη μέθοδο της λογιστικής

παλινδρόμησης που βασίζεται στη μέγιστη πιθανοφάνεια από ανεξάρτητες παρατηρήσεις δεν είναι πλέον κατάλληλη για δεδομένα που είναι δομημένα με την κατά συστάδες μέθοδο καθώς οι παρατηρήσεις στην ίδια ομάδα τείνουν να έχουν παρόμοια χαρακτηριστικά και είναι πιο πιθανό να συσχετίζονται η μια με την άλλη. Οι εύρωστες τυπικές εκτιμήσεις σφάλματων είναι απαραίτητες να ληφθούν υπόψη για τη συσχέτιση εντός μίας ομάδα. Ο Huber (1967) πρότεινε ένα μοντέλο, για τον υπολογισμό εύρωστων τυπικών σφαλμάτων αν υπάρχει ετεροσκεδαστικότητα στη δειγματοληψία κατά συστάδες. Εάν :

$$p_i = \frac{e^{x_i\beta}}{1-e^{x_i\beta}} \qquad i = 1,.....n$$

είναι η πιθανότητα ενός γεγονότος, όπου $x_i$ είναι οι μεταβλητές που σχετίζονται με την πιθανότητα εκδήλωσης αυτού του γεγονότος και $\beta$ είναι οι συντελεστές παλινδρόμησης.

Έστω,

$$L = \log(\Pi f(x_i)) = \Sigma \log(f(x_i)) = \Sigma l(x_i) \qquad i = 1,.....n$$

είναι η λογαριθμική πιθανοφάνεια, τότε ορίζουμε το αποτέλεσμα της συνάρτησης ως:

$$S_i = \frac{\partial(L)}{\partial(x_i\beta)}$$

και του Hessian ως:

$$H_j = \frac{\partial^2 L}{\partial(x_i\beta)^2}$$

για $i-οστη$ παρατήρηση $i = 1,.....n$ Ας υποθέσουμε ότι αφαιρούμε της παρατήρησης $i-οστη$ από το μοντέλο, τότε οι εκτιμήσεις θα διαφοροποιηθούν από το ποσό των $-D^{-1}S_i x_i^T$ όπου η μήτρα $D = \sum_i H_i(x_i^T x_i)$.

Υποθέτουμε ότι καμία μεμονωμένη παρατήρηση δεν έχει πολύ μεγάλη επίδραση στην πράξη, τότε το αποτέλεσμα της αφαίρεσης δύο παρατηρήσεων είναι περίπου το ίδιο μα το αποτέλεσμα που θα είχαμε εάν αφαιρούσαμε κάθε παρατήρηση ξεχωριστά. Η ίδια λογική ισχύει και για τις παρατηρήσεις σε ένα σύμπλεγμα, αφαιρώντας όλα τα μέλη μιας ομάδας η τιμή που θα μας προκύψει θα είναι περίπου ισοδύναμη με την τιμή που θα είχαμε αν αφαιρούσαμε κάθε μέλος με τη σειρά. Σύμφωνα με το μοντέλο του Huber, η εύρωστη τυπική εκτίμηση της διακύμανσης είναι:

$$Var(\beta) = D^{-1}(\sum_i S_i x_i^T x_i S_i) D^{-1}$$

Μπορούμε να δούμε από τη φυσική εμφάνιση του ανωτέρω τύπου τον λόγο που είναι ευρέως γνωστός ως «Sandwich εκτίμηση".

Για το λογιστικό μοντέλο, μπορούμε, μετά από κάποιες άλγεβρα πράξεις να έχουμε το εξής:

$$S_i = \frac{\partial(L)}{\partial(x_i\beta)} = y_i - p_i$$

και του Hessian αντίστοιχα:

$$H_j = \frac{\partial^2 L}{\partial(x_i\beta)^2} = p_i(1-p_i)\,.$$

Στη συνέχεια, εάν το συνδέσουμε με το $Var(\beta) = D^{-1}(\sum_i S_i x_i^T x_i S_i) D^{-1}$ , έχουμε:

$$\begin{aligned} Var(\beta) = & (\sum_i (p_i(1-p_i)(x_i^T x_i)))^{-1} \\ & *(\sum_i (y_i - p_i) x_i^T x_i (y_i - p_i)) \\ & *(\sum_i (p_i(1-p_i)(x_i^T x_i)))^{-1} \end{aligned}$$

Αν για παράδειγμα έχουμε C συστάδες (κάθε συστάδα έχει $g_j$ παρατηρήσεις, $j = 1,.....C$ ), και κάθε συστάδα είναι ανεξάρτητη. Τότε το εύρωστο τυπικό σφάλμα είναι:

$$Var(\beta) = \tilde{D}^{-1}(\sum_j U_j^T U_j)\tilde{D}^{-1}$$

όπου

$$\tilde{D} = \sum_{j=1}^{C} \sum_{k=1}^{g_j} (H_{jk} x_{jk}^T x_{jk})$$

$$U_j = \sum_{k=1}^{g_j} x_{jk} S_{jk},\ j = 1,.....C.,$$

η συνεισφορά της κάθε συστάδας στο αποτέλεσμα.

Για τα δεδομένα με βάρη, θα έχουμε $X^T X$ , $X^T y$ και $y^T y$ αντικαθίστανται από $X^T W X$ , $X^T W y$ και $y^T W y$ όπου $W$ είναι ένας διαγώνιος πίνακας των οποίων τα διαγώνια στοιχεία είναι τα στοιχεία του $w$ , το διάνυσμα των βαρών.

Δεδομένου ότι η εκτίμηση του εύρωστου τυπικού σφάλματος βασίζεται σε δείγμα δεδομένων, μια πεπερασμένη διόρθωση του δείγματος είναι απαραίτητη για την προσαρμογή των εκτιμήσεων πιο κοντά στην τιμή του πραγματικού

πληθυσμού. Υπάρχουν δυο δημοφιλείς προσαρμογές, η μία είναι σαν τον τύπο της παλινδρόμησης,

$$q_c = \frac{N-1}{(N-p-1)} * \frac{C}{C-1}$$

όπου $N$ είναι ο αριθμός του συνόλου των παρατηρήσεων και $p$ είναι ο αριθμός των παραγόντων πρόβλεψης στο μοντέλο και C είναι ο αριθμός των συστάδων στα δεδομένα. Η δεύτερη προσαρμογή είναι ο τύπος:

$$q_c = \frac{C}{C-1}$$

Για δεδομένα με πολύ μεγάλο αριθμό παρατηρήσεων, η επίδραση αυτών των δύο προσαρμογών είναι περίπου ίδια. Λαμβάνοντας υπόψη την παραπάνω διόρθωση του δείγματος, ο τύπος του Huber για την εύρωστη διακύμανση είναι:

$$Var(\beta) = q_c \tilde{D}^{-1} (\sum_j U_j^T U_j) \tilde{D}^{-1}$$

Χρησιμοποιώντας το πρόγραμμα STATA βγήκαν τα εξής αποτελέσματα για την εκτίμηση των τυπικών σφαλμάτων με τη μέθοδο της λογιστικής παλινδρόμησης. Αναλυτικότερα στο παράρτημα 3 και 4. Αρχικά εφαρμόστηκε η μέθοδος της λογιστικής παλινδρόμησης με την υπόθεση ότι το δείγμα μας ήταν δομημένο με τη μέθοδο της απλής τυχαίας δειγματοληψίας. Στη δεύτερη στήλη των παρακάτω πινάκων παρατηρούμε τις τιμές του τυπικού σφάλματος που προέκυψαν. Στη συνέχεια εκτελέστηκε η λογιστική παλινδρόμηση για τη δειγματοληψία κατά συστάδες (στρωματοποιημένη κατά νομούς με ομάδες ανά σχολική τάξη) βασισμένη στη γραμμική μέθοδο εκτίμησης τυπικών σφαλμάτων. Για την εφαρμογή της η Λογιστική Παλινδρόμηση απαιτεί κατηγορική εξαρτημένη μεταβλητή και κατηγορικές ή ποσοτικές ανεξάρτητες μεταβλητές. Όπως και στην γραμμική παλινδρόμηση οι ποσοτικές μεταβλητές εισάγονται στην αρχική τους μορφή. Οι κατηγορικές μεταβλητές προκειμένου να εισαχθούν στο μοντέλο, πρέπει να μετασχηματισθούν σε ψευδομεταβλητές. Για τη δημιουργία των ψευδομεταβλητών θα πρέπει μετά τον ορισμό όλων των ανεξάρτητων μεταβλητών να προσδιορίσουμε εκείνες που είναι κατηγορικές και για τις οποίες θα πρέπει να δημιουργηθούν ψευδομεταβλητές. Στην περίπτωση μας ορίσαμε τις κατηγορικές μεταβλητές το φύλο με τιμή 1 τα αγόρια και τιμή 2 τα κορίτσια. Η απάντηση στο κάπνισμα όχι = 0 και ναι =1. Στο παράρτημα 3 γίνεται αναφορά σχετικά με την κωδικοποιήση των κατηγορικών μεταβλητών. Επιπλέον οι μεταβλητές που σχετίζονται με την ηλικιακή ομάδα έχουν ως εξής: Ηλικιακή ομάδα 13-14 = 1 , 14-16=2, 16-18 =3, η ηλιακή ομάδα 19+ αφαιρέθηκε για τον υπολογισμό. Και τελευταία αφορά τη βαθμολογία. Με 1 ορίσαμε τους μαθητές με βαθμολογία 10-13, με 2 βαθμολογία 13-18, και με 3 βαθμολογία 18-20. Οπότε θα εξετάσουμε το κάπνισμα, τη κατανάλωση αλκοόλ, τη χρήση χασίς και τη χρήση διαφόρων άλλων ναρκωτικών ουσιών σε σχέση με το φύλο, την ηλικιακή ομάδα και τη βαθμολογία των μαθητών.

Πίνακας 9.

Υπολογισμός τυπικού σφάλματος των συντελεστών της λογιστικής παλινδρόμησης για την απλή τυχαία δειγματοληψία και την κατά συστάδες δειγματοληψία για τις ερωτήσεις που αφορούν το κάπνισμα.

ESPAD 2011

| | | | Odds ratio | Std. Err. | Linearized Std. Err. Class + Nomos | Design Effect | Number of obs | Number of obs after Design Effect |
|---|---|---|---|---|---|---|---|---|
| c08 | Κάπνισες ποτέ τσιγάρο | φύλο | 0,8758 | 0,024572 | 0,030972 | 1,5887 | 35.462 | 22.321 |
| | | Ηλικιακή ομάδα_13-14 | 0,0499 | 0,004759 | 0,005724 | 1,4467 | 35.462 | 24.513 |
| | | Ηλικιακή ομάδα_15-16 | 0,1911 | 0,017633 | 0,021548 | 1,4934 | 35.462 | 23.746 |
| | | Ηλικιακή ομάδα_17-18 | 0,4400 | 0,040787 | 0,049786 | 1,4899 | 35.462 | 23.801 |
| | | Βαθμολογία_ 10-13 | 2,1173 | 0,088074 | 0,098953 | 1,2623 | 34.942 | 27.681 |
| | | Βαθμολογία_ 18_20 | 0,4026 | 0,013545 | 0,016069 | 1,4075 | 34.942 | 24.826 |
| c09 | Πόσο κάπνισες τις τελευταίες 30 ημέρες | φύλο | 0,7444 | 0,030599 | 0,037997 | 1,5420 | 35.553 | 23.057 |
| | | Ηλικιακή ομάδα_13-14 | 0,0182 | 0,002028 | 0,002243 | 1,2236 | 35.553 | 29.056 |
| | | Ηλικιακή ομάδα_15-16 | 0,1145 | 0,009765 | 0,010510 | 1,1583 | 35.553 | 30.694 |
| | | Ηλικιακή ομάδα_17-18 | 0,3102 | 0,025475 | 0,025834 | 1,0283 | 35.553 | 34.573 |
| | | Βαθμολογία_ 10-13 | 2,7592 | 0,133526 | 0,146821 | 1,2091 | 35.029 | 28.972 |
| | | Βαθμολογία_ 18_20 | 0,3131 | 0,019248 | 0,021592 | 1,2584 | 35.029 | 27.836 |

Παρατηρούμε από τη στήλη odds ratio ότι τα αγόρια έχουν 0,8 φορές περισσότερη odds να καπνίζουν από ότι τα κορίτσια. Επιπλέον ότι η ηλικιακή ομάδα 17-18 έχει 0,4 φορές περισσότερη odd (Fuller κ.α, 1986), να καπνίζει και όπως φαίνεται έχει και τη μεγαλύτερη σε σχέση με της υπόλοιπες ομάδες. Όπως είναι αναμενόμενο μαθητές με χαμηλή βαθμολογία έχουν 2,1 φορές μεγαλύτερη odds να καπνίζουν από οι υπόλοιποι μαθητές. Με τον ίδιο τρόπο μπορούμε να αναλύσουμε και τα άλλα αποτελέσματα.

Πίνακας 10.

Υπολογισμός τυπικού σφάλματος των συντελεστών της λογιστικής παλινδρόμησης για την απλή τυχαία δειγματοληψία και την κατά συστάδες δειγματοληψία για τις ερωτήσεις που αφορούν την κατανάλωση αλκοόλ.

ESPAD 2011

| | | | Odds ratio | Std. Err. | Linearized Std. Err. Class + Nomos | Design Effect | Number of obs | Number of obs after Design Effect |
|---|---|---|---|---|---|---|---|---|
| | Ήπιες ποτέ αλκοολούχο ποτό | | | | | | | |
| | Σε όλη σου τη ζωή - | | | | | | | |
| c12a | σήμερα | φύλο | 0,7157 | 0,021851 | 0,025563 | 1,3686 | 34.632 | 25.304 |
| | | Ηλικιακή ομάδα_13-14 | 0,1247 | 0,014804 | 0,016221 | 1,2005 | 34.632 | 28.848 |
| | | Ηλικιακή ομάδα_15-16 | 0,4917 | 0,058813 | 0,064244 | 1,1932 | 34.632 | 29.024 |
| | | Ηλικιακή ομάδα_17-18 | 1,4182 | 0,176037 | 0,190405 | 1,1699 | 34.632 | 29.602 |
| | | Βαθμολογία_ 10-13 | 1,0288 | 0,049768 | 0,053712 | 1,1647 | 34.122 | 29.296 |
| | | Βαθμολογία_ 18_20 | 0,6703 | 0,022041 | 0,025410 | 1,3291 | 34.122 | 25.674 |
| | Τις 30 τελευταίες ημέρες πόσες φορές ήπιες στη σειρά ≥5 | | | | | | | |
| c18 | αλκοολ.ποτά | φύλο | 0,5316 | 0,018157 | 0,020628 | 1,2906 | 23.034 | 17.848 |
| | | Ηλικιακή ομάδα_13-14 | 0,3321 | 0,470376 | 0,470522 | 1,0006 | 23.034 | 23.020 |
| | | Ηλικιακή ομάδα_15-16 | 0,5132 | 0,042440 | 0,048449 | 1,3032 | 23.034 | 17.674 |
| | | Ηλικιακή ομάδα_17-18 | 0,7084 | 0,058882 | 0,065272 | 1,2288 | 23.034 | 18.744 |
| | | Βαθμολογία_ 10-13 | 1,6065 | 0,079751 | 0,082481 | 1,0697 | 22.629 | 21.155 |
| | | Βαθμολογία_ 18_20 | 0,4421 | 0,017542 | 0,018306 | 1,0890 | 22.629 | 20.780 |

Όπως φαίνεται οι μαθητές που ανήκουν στην τρίτη ηλικιακή ομάδα έχουν 1,4 φορές περισσότερες πιθανότητες να έχουν πιεί αλκοολούχα ποτά μειώνοντας όμως αυτή την πιθανότητα στα μισά να έχουν πιει περισσότερα από 5 ποτά στη σειρά τον τελευταίο μήνα. Παρατηρείται επιπλέον ότι η κατανάλωση αλκοόλ συσχετίζεται και με τους μαθητές που έχουν χαμηλή βαθμολογική απόδοση οι οποίοι έχουν κατά 2,9 μεγαλύτερες πιθανότητες να έχουν μεθύσει σε σχέση με τους υπόλοιπους μαθητές.

Συνέχεια πίνακα 10.

| | | | Odds ratio | Std. Err. | Linearized Std. Err. Class + Nomos | Design Effect | Number of obs | Number of obs after Design Effect |
|---|---|---|---|---|---|---|---|---|
| | Πόσες φορές μέθυσες από αλκοολ.ποτά | | | | | | | |
| | Σε όλη σου τη ζωή - σήμερα | | | | | | | |
| c19a | | φύλο | 0,2289 | 0,039537 | 0,040186 | 1,0331 | 35.191 | 34.065 |
| | | Ηλικιακή ομάδα_13-14 | 0,0464 | 0,013472 | 0,014867 | 1,2178 | 35.191 | 28.897 |
| | | Ηλικιακή ομάδα_15-16 | 0,2016 | 0,047435 | 0,044623 | 0,8849 | 35.191 | 39.767 |
| | | Ηλικιακή ομάδα_17-18 | 0,4207 | 0,091331 | 0,087228 | 0,9122 | 35.191 | 38.580 |
| | | Βαθμολογία_ 10-13 | 2,9347 | 0,471290 | 0,463469 | 0,9671 | 34.658 | 35.837 |
| | | Βαθμολογία_ 18_20 | 0,3597 | 0,077857 | 0,075674 | 0,9447 | 34.658 | 36.687 |

Πίνακας 11.

Υπολογισμός τυπικού σφάλματος των συντελεστών της λογιστικής παλινδρόμησης για την απλή τυχαία δειγματοληψία και την κατά συστάδες δειγματοληψία για τις ερωτήσεις που αφορούν τη χρήση χασίς.

ESPAD 2011

| | | | Odds ratio | Std. Err. | Linearized Std. Err. Class + Nomos | Design Effect | Number of obs | Number of obs after Design Effect |
|---|---|---|---|---|---|---|---|---|
| | Δοκίμασες ή πήρες ποτέ μαριχουάνα ή χασίς | | | | | | | |
| | Σε όλη σου τη ζωή - σήμερα | | | | | | | |
| c25a | | φύλο | 0,2359 | 0,031015 | 0,040391 | 1,6960 | 35.548 | 20.960 |
| | | Ηλικιακή ομάδα_13-14 | 0,0187 | 0,004612 | 0,005233 | 1,2874 | 35.548 | 27.612 |
| | | Ηλικιακή ομάδα_15-16 | 0,0664 | 0,011445 | 0,012575 | 1,2072 | 35.548 | 29.446 |
| | | Ηλικιακή ομάδα_17-18 | 0,2224 | 0,031635 | 0,036296 | 1,3164 | 35.548 | 27.004 |
| | | Βαθμολογία_ 10-13 | 3,4907 | 0,427288 | 0,467225 | 1,1957 | 35.008 | 29.279 |
| | | Βαθμολογία_ 18_20 | 0,4583 | 0,078169 | 0,079499 | 1,0343 | 35.008 | 33.847 |

Πίνακας 12.

Υπολογισμός τυπικού σφάλματος των συντελεστών της λογιστικής παλινδρόμησης για την απλή τυχαία δειγματοληψία και την κατά συστάδες δειγματοληψία για τις ερωτήσεις που αφορούν τη χρήση έκστασης.

ESPAD 2011

| | | | Odds ratio | Std. Err. | Linearized Std. Err. Class + Nomos | Design Effect | Number of obs | Number of obs after Design Effect |
|---|---|---|---|---|---|---|---|---|
| | Δοκίμασες ή πήρες ποτέ έκσταση | | | | | | | |
| c29a | Σε όλη σου τη ζωή - σήμερα | φύλο | 0,1487 | 0,064662 | 0,064716 | 1,0017 | 35.343 | 35.284 |
| | | Ηλικιακή ομάδα_13-14 | 0,1266 | 0,067102 | 0,064980 | 0,9378 | 35.343 | 37.689 |
| | | Ηλικιακή ομάδα_15-16 | 0,1301 | 0,063232 | 0,061071 | 0,9328 | 35.343 | 37.888 |
| | | Ηλικιακή ομάδα_17-18 | 0,1766 | 0,088739 | 0,086109 | 0,9416 | 35.343 | 37.535 |
| | | Βαθμολογία_10-13 | 1,4285 | 0,556138 | 0,556113 | 0,9999 | 34.811 | 34.814 |
| | | Βαθμολογία_18_20 | 1,2987 | 0,487704 | 0,493622 | 1,0244 | 34.811 | 33.981 |

Πίνακας 13.

Υπολογισμός τυπικού σφάλματος των συντελεστών της λογιστικής παλινδρόμησης για την απλή τυχαία δειγματοληψία και την κατά συστάδες δειγματοληψία για τις ερωτήσεις που αφορούν τη χρήση διάφορων ουσιών. ESPAD 2011

| | | | Odds ratio | Std. Err. | Linearized Std. Err. Class + Nomos | Design Effect | Number of obs | Number of obs after Design Effect |
|---|---|---|---|---|---|---|---|---|
| | Έχεις χρησιμοποιήσει ποτέ κάποια από τις παρακάτω ουσίες | | | | | | | |
| c31a | Ηρεμιστικά ή υπνωτικά χωρις τη σύσταση γιατρού | φύλο | 0,6827 | 0,155754 | 0,154791 | 0,9877 | 35.489 | 35.932 |
| | | Ηλικιακή ομάδα_13-14 | 0,0293 | 0,018230 | 0,018030 | 0,9782 | 35.489 | 36.281 |
| | | Ηλικιακή ομάδα_15-16 | 0,2570 | 0,095741 | 0,095096 | 0,9866 | 35.489 | 35.973 |
| | | Ηλικιακή ομάδα_17-18 | 0,3272 | 0,121105 | 0,111098 | 0,8416 | 35.489 | 42.170 |
| | | Βαθμολογία_10-13 | 2,4069 | 0,695743 | 0,757096 | 1,1841 | 34.949 | 29.514 |
| | | Βαθμολογία_18_20 | 1,1134 | 0,296916 | 0,301191 | 1,0290 | 34.949 | 33.964 |
| c31k | Ναρκωτικά με ένεση | φύλο | 0,2313 | 0,114890 | 0,115333 | 1,0077 | 23.142 | 22.965 |
| | | Ηλικιακή ομάδα_13-14 | | | | | | |
| | | Ηλικιακή ομάδα_15-16 | 0,5083 | 0,285934 | 0,286627 | 1,0048 | 23.140 | 23.028 |
| | | Ηλικιακή ομάδα_17-18 | 0,1990 | 0,116611 | 0,116664 | 1,0009 | 23.140 | 23.119 |
| | | Βαθμολογία_10-13 | 1,5952 | 0,636900 | 0,647637 | 1,0340 | 22.738 | 21.990 |
| | | Βαθμολογία_18_20 | 1,0719 | 0,500684 | 0,497985 | 0,9893 | 22.738 | 22.985 |

Συνέχεια Πίνακα 13.

| | | | Odds ratio | Std. Err. | Linearized Std. Err. Class + Nomos | Design Effect | Number of obs | Number of obs after Design Effect |
|---|---|---|---|---|---|---|---|---|
| c31l | Κάποιο αλκοολούχο ποτό μαζι με φάρμακα για να αλλάξεις τη διάθεση σου | φύλο | 0,4826 | 0,224152 | 0,222271 | 0,9833 | 23.187 | 23.581 |
| | | Ηλικιακή ομάδα_13-14 | | | | | | |
| | | Ηλικιακή ομάδα_15-16 | 0,1256 | 0,062872 | 0,063059 | 1,0059 | 23.185 | 23.048 |
| | | Ηλικιακή ομάδα_17-18 | 0,1409 | 0,076443 | 0,076333 | 0,9971 | 23.185 | 23.252 |
| | | Βαθμολογία_10-13 | 1,0110 | 0,567518 | 0,579291 | 1,0419 | 22.782 | 21.865 |
| | | Βαθμολογία_18_20 | 1,2376 | 0,616075 | 0,611795 | 0,9862 | 22.782 | 23.102 |

Παρατηρώντας τη στήλη του design effect στον πίνακα 9 ο οποίος σχετίζεται με τη συνήθεια των μαθητών σε σχέση με το κάπνισμα διαπιστώνουμε ότι παίρνει τιμές μεγαλύτερες της μονάδας οπότε θα ήταν αποτελεσματικότερη αν η δειγματοληψία είχε γίνει με τη μέθοδο της απλής τυχαίας δειγματοληψίας χωρίς αντικατάσταση. Σε κάποιες ερωτήσεις ξεπερνάει το 1,5 δείχνοντας μας ότι το 1/3 του δείγματος μας είναι περιττό καθώς υπάρχει συσχέτιση μεταξύ των μαθητών, οπότε προσεγγιστικά με 10000 λιγότερους μαθητές και ακολουθώντας τη μέθοδο της απλής τυχαίας δειγματοληψίας χωρίς αντικατάσταση θα επιτυγχάναμε το ίδιο αποτέλεσμα. Σε όλες τις ερωτήσεις διαπιστώνουμε οι μαθητές με χαμηλή βαθμολογία είναι περισσότερο επιρρεπής ιδιαίτερα στη δοκιμή μαριχουάνας. Στη συσχέτιση του αλκοόλ και της δοκιμής μαριχουάνας η χασίς με το φύλο, την ηλικία και την βαθμολογία των μαθητών η επίδραση του δειγματοληπτικού σχεδιασμού είναι > 1. Φαίνεται πως υπάρχει ταύτιση απόψεων και πιθανός ακόμα και συμπεριφοράς μεταξύ των μαθητών. Στα υπόλοιπα αποτελέσματα είτε είναι κοντά στη μονάδα είτε μικρότερης πιθανός γιατί οι θετικές απαντήσεις που αφορούσαν αυτές τις ερωτήσεις ήταν μόλις 1%.

# Συμπεράσματα

Χρησιμοποιήθηκε το στατιστικό πρόγραμμα STATA και τα δεδομένα από την Πανελλήνια έρευνα που έγινε στο μαθητικό πληθυσμό για τη χρήση εξαρτησιογόνων ουσιών. Η ανάγκη της δειγματοληψίας γεννήθηκε από την δυσκολία, από πλευράς κόστους και μέσων, της εξέτασης όλων των μονάδων ενός πληθυσμών ως προς ένα ιδιαίτερο χαρακτηριστικό τους. Στην έρευνα συμμετείχαν 37000 μαθητές από όλη την Ελλάδα. Αρχικά οι μονάδες της δειγματοληψίας ομαδοποιήθηκαν και βασική μονάδα δειγματοληψίας απετέλεσε το σχολικό τμήμα. Για να εξασφαλιστεί ένα δείγμα ικανοποιητικού μεγέθους για κάθε νομό της χώρας έγινε στρωματοποίηση ως προς το νομό, δηλαδή, έγινε ξεχωριστή δειγματοληψία των σχολικών τμημάτων σε κάθε νομό.

Η αποτελεσματική στρωματοποίηση παράγει εκτιμητές παραμέτρων που έχουν μικρότερες διασπορές από τους εκτιμητές που απορρέουν από την απλή τυχαία δειγματοληψία. Πλεονεκτήματα της στρωματοποιημένης δειγματοληψίας και της κατά συστάδες δειγματοληψίας τυχαίνουν να είναι και η μείωση του κόστους και η διοικητική ευκολία ελέγχου και αποπεράτωσης της δειγματοληψίας. Στην κατά συστάδες δειγματοληψία όμως υπάρχει ένα βασικό μειονέκτημα που είναι η συσχέτιση των μονάδων που υπάρχει μέσα σε μία ομάδα η οποία αντανακλά την ομοιογένεια του δείγματος. Η συσχέτιση μέσα σε μία ομάδα αντιπροσωπεύει την πιθανότητα δύο μονάδες εντός της ίδιας ομάδας να έχουν την ίδια τιμή για ένα δεδομένο στατιστικό στοιχείο σε σχέση με δύο στοιχεία που επιλέγονται εντελώς τυχαία. Λόγου αυτού του γεγονότος μπορεί να μας παρουσιαστεί το πρόβλημα δειγματοληπτικού σχεδιασμού. Η επίδραση του δειγματοληπτικού σχεδιασμού (design effect) μιας έρευνας μπορεί να χρησιμοποιηθεί ως εργαλείο για να υπολογιστεί η αποτελεσματικότητας του δείγματος και για να πραγματοποιηθεί ο σωστότερος σχεδιασμός μιας έρευνας. Η επίδραση του δειγματοληπτικού σχεδιασμού από την ομαδοποίηση είναι, κατά κανόνα, μεγαλύτερη από το 1.

Για να εκτιμήσουμε όμως το design effect πρέπει πρώτα να εκτιμήσουμε τη διακύμανση. Ο υπολογισμός της διακύμανσης έγινε αρχικά για την απλή τυχαία δειγματοληψία και ύστερα για την κατά συστάδες δειγματοληψία με τη γραμμική και την Jackknife. Επιπλέον υπολογίστηκε η διακύμανση με τη λογιστική παλινδρόμηση ακολουθώντας τα ίδια βήματα. Στη λογιστική παλινδρόμηση πέρα από τη σχέση των υπό εξέταση ερωτήσεων με το φύλο , την ηλικία και την σχολική απόδοση εξετάσαμε πάλι τις τιμές του design effect. Οι τιμές αυτές στις περισσότερες περιπτώσεις μας υποδεικνύουν ότι θα ήταν αποτελεσματικότερη εάν για την έρευνα είχε προτιμηθεί η μέθοδος της απλής τυχαίας δειγματοληψίας. Σχετικά με τη γραμμική και Jackknife μέθοδο οι τιμές μεταξύ τους έχουν ελάχιστες αποκλίσεις. Αποκλείουν ωστόσο με τη διακύμανση που προέκυψε από την απλή τυχαία δειγματοληψία υποδεικνύοντας ότι η κατά συστάδες δειγματοληψία στην

περίπτωση αυτή δεν ήταν η σωστότερη επιλογή. Το design effect διαφέρει από ερώτηση σε ερώτηση συμπεραίνοντας ότι σε κάποια θέματα δεν υπάρχει ταύτιση απόψεων μεταξύ των μαθητών και σε αυτά τα θέματα η κατά συστάδες δειγματοληψία να είναι αποτελεσματική μέθοδος για μια έρευνα.

Εν κατακλείδι, χρησιμοποιώντας τη κατά συστάδες δειγματοληψία γενικά απαιτεί είτε ένα μεγαλύτερο σε μέγεθος δείγμα από ότι στην απλή τυχαία δειγματοληψία, είτε ένα ευρύτερο διάστημα εμπιστοσύνης. Το design effect χρησιμοποιείται για να καθορίσει πόσο μεγάλο πρέπει είναι το μέγεθος του δείγματος ή ποια πρέπει να τα διαστήματα εμπιστοσύνης και ποια μέθοδος δειγματοληψίας θα ήταν πιο συμφέρουσα.

# Βιβλιογραφία


- Binder, D. (1983) *On the variances of asymptotically normal estimators from complex surveys.* International Statistical Review 51, pp: 279-292.
- Binder, D. (1996) *Linearization methods for single phase and two phase samples: A cookbook approach*. Survey Methodology 22, pp: 17-22.
- Cameron, A Trivedi, P.K. (2005) *Microeconometrics : Methods and Applications.* Cambridge New York: Cambridge University Press.
- Durbin, J. (1959*) A note on the application of Quenouille's method of bias reduction to the estimation of ratios*. Biometrika 46, pp 477-480.
- Efron, B. (1982) *The Jackknife, the Bootstrap, and Other Resampling Plans.* Philadelphia, PA: Society for Industrial and Applied Mathematics.
- Frankel, M.R. (1971) *Inference from Survey Samples.* Institute for Social Research, University of Michigan.
- Fuller, W. A. (1975) *Regression analysis for sample survey.* Sankhya, 37 (3), Series C, 117–132.
- Huber, P. J. (1967) *The behavior of maximum likelihood estimates under non-standard conditions.* Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, pp: 221-233
- Kalton, G. (1983) *Introduction to Survey Sampling*. SAGE University Paper series on Quantitative Applications in the Social Sciences, series no. 07-035, Beverly Hills and London: SAGE Publications, Inc.
- Kalton, G., Brick, M. & Le, T. (2005) *Estimating components of design effects for use in sample design.* In *Household Sample Surveys in Developing and Transition Countries.* New York: United Nations Publication. Chapter 6, pp: 105
- Kiaer, A.W. (1895) *The Representative Method of Statistical Surveys. English translation*, 1976. Oslo: Statistik Sentralbyro.
- Kish, L. (1965) *Survey Sampling*, New York: John Wiley & Sons, Inc.
- Levy, P.S. & Lemeshow, S. (1999) *Sampling of Populations: Methods and Applications*. 3rd edition, New York: John Wiley and Sons, Inc.
- Quenouille, M. (1949) *Approximate tests of correlation in time series.* Journal of the Royal Statistical Society, Series B 11, pp: 18-44.
- Quenouille, M. (1956) Notes on bias in estimation. Biometrika 43,pp:353-360
  Raj, D. (1972) *The Design Of Sample Surveys*. New York: McGraw-Hill Book Company.
- Royall, R.M. & Cumberland, W.G. (1981) *An empirical study of the ratio estimator and estimators of its variance*, Journal of the American Statistical Association 76, pp: 66-88.
- Saifuddin, A. (2009) *Cluster Sampling*. Dept. of Biostatistics School of Hygiene and Public Health Johns Hopkins University

- Sarndal, C.E., Swensson, B., & Wretman, I.H. (l989) *The weighted residual technique for estimating the variance of the general regression estimator of the finite population total.* Biometrika 76, pp: 527-537.
- Tukey, J.W. (1958*) Bias and confidence in not quite large samples.* Annals of Mathematical Statistics 29, 614
- Turner, AG. (1996) *Sampling Topics for Disability Surveys.* United Nations Statistics Division.
- Wolter, K. (1985) *Introduction to Variance Estimation*. New York: Springer-Verlag, p.p 31-33.
- Woodruff, R. S. (1971) *A simple method for approximating the variance of a complicated estimate*. Journal of the American Statistical Association, 66, pp: 411–414.
- Ζαχαροπούλου, Χ. (2009) *Στατιστική Μέθοδοι –Εφαρμογές*. Τόμος Α, 2η Έκδοση, Αθήνα: Εκδόσεις Ζυγός.
- Ρίτσαρντσον, Κ., Βασίλαινας, Α. (1999) *Εισαγωγή στην στατιστική*. Αθήνα: Εκδόσεις Κάκτος.
- Φράγκος, Χ. (1998) *Στατιστικη επιχειρησεων*. Αθήνα: Εκδόσεις Σταμούλης.

Ψαρρού, Μ., Ζαφειρόπουλος, Κ., (2001) *Επιστημονική έρευνα Θεωρία και εφαρμογές στις κοινωνικές επιστήμες*. Αθήνα: Εκδόσεις Τυπωθήτω.


## Αναφορές


- Alecxih, L., Corea, J. & Marker, D. (1998) *A statistical assessment and state tabulations.* Department of Health & Human Services. Section II B, Precision of estimates. [Διαδίκτυο] Hhs database. Διαθέσιμο στο: http://aspe.hhs.gov/health/reports/st_est/ . [Πρόσβαση στις 7 Μαΐου 2014]
- Demnati, A. & Rao, J. N. K. (2002) *Linearization variance estimators for survey data.* [Διαδίκτυο] Ssc database. Διαθέσιμο στο: http://www.ssc.ca/survey/documents/SSC2002_A_Demnati.pdf. [Πρόσβαση στις 19 Μαΐου 2014]
- EPIPSI.(2012) *Πανελλήνια έρευνα για την χρήση εξαρτησιογόνων ουσιών στους μαθητές, Έρευνα ESPAD 2011,* [Διαδίκτυο] Διαθέσιμο στο : http://www.epipsi.gr/Tekmiriosi/epid/Epidimiologikes_erevnes/Ekthesi%20Apotelesmatwn%20Ereynas%20ESPAD%202011-EPIPSI.pdf [Πρόσβαση στις 8 Μαρτίου 2014]
- Flores-Cervantes, I., Brick, J.M., & DiGaetano, R. (1997) *NSAF Variance Estimation*. Assessing the New Federalism, An Urban Institute Program to Assess Changing Social Policies. Report No. 4 [Διαδίκτυο] Urban database. Διαθέσιμο στο: http://www.urban.org/UploadedPDF/Methodology_4.pdf . [Πρόσβαση στις 9 Μαΐου 2014]
- MEDNET (2009) *Πολυμεταβλητή ανάλυση επιδημιολογικών δεδομένων .* [Διαδίκτυο] Mednet database. Διαθέσιμο στο:

http://www.mednet.gr/archives/2009-3/pdf/407.pdf [Πρόσβαση στις 20 Σεπτεμβρίου 2014]

- Park, I. et al. (2003) *Design effects and survey planning.* [Διαδίκτυο] Westat database. Διαθέσιμο στο: https://www.amstat.org/sections/SRMS/Proceedings/y2003/Files/JSM2003000820.pdf . [Πρόσβαση στις 3 Μαΐου 2014]
- Park, I., & Lee, H. (2001) *The design effect: Do we know all about it?* [Διαδίκτυο] Westat database. Διαθέσιμο στο: https://www.amstat.org/sections/srms/Proceedings/y2001/Proceed/00144.pdf . [Πρόσβαση στις 3 Μαΐου 2014]
- Rust, K. (2007) *Usage of Linearization Variance Estimators for Survey Estimates – Discussion .* [Διαδίκτυο] Amstat database. Διαθέσιμο στο: *http://www.amstat.org/meetings/ices/2007/proceedings/ICES2007-000215.PDF* [Πρόσβαση στις 14 Ιουνίου 2014]
- Rust, K. (2007) *Usage of Linearization Variance Estimators for Survey Estimates – Discussion .* [Διαδίκτυο] Amstat database. Διαθέσιμο στο: *http://www.amstat.org/meetings/ices/2007/proceedings/ICES2007-000215.PDF* [Πρόσβαση στις 14 Ιουνίου 2014]
- Shackman, G. (2001) *Design and Methodology*. [Διαδίκτυο] Census database. Διαθέσιμο στο: http://www.census.gov/prod/2002pubs/tp63rv.pdf [Πρόσβαση στις 5 Ιουλίου 2014]
- Ντζούφρας, Ι και Περπέρογλου, Α. (2009) *Εισαγωγή στην Βιοστατιστική και την Επιδημιολογία.* [Διαδίκτυο] Actuar.aegean database. Διαθέσιμο στο: *http://www.actuar.aegean.gr/notes/biostatistics-v4-0.pdf* [Πρόσβαση στις 20 Σεπτέμβριου 2014]

# Παραρτήματα

## Παράρτημα 1.

### 1st PROPORTION

```
. proportion c08 [pweight = weight], nolabel
Proportion estimation                 Number of obs     =    35462
```

| | Proportion | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c08** | | | | |
| 0 | .6086507 | .0033382 | .6021077 | .6151936 |
| 1 | .3913493 | .0033382 | .3848064 | .3978923 |

```
. proportion c09 [pweight = weight], nolabel
Proportion estimation                 Number of obs     =    35553
```

| | Proportion | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c09** | | | | |
| 0 | .859001 | .0024639 | .8541716 | .8638304 |
| 1 | .140999 | .0024639 | .1361696 | .1458284 |

```
. proportion c12a [pweight = weight], nolabel
Proportion estimation                 Number of obs     =    34632
```

| | Proportion | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c12a** | | | | |
| 0 | .2603722 | .0029216 | .2546458 | .2660985 |
| 1 | .7396278 | .0029216 | .7339015 | .7453542 |

```
. proportion c12b [pweight = weight], nolabel
Proportion estimation                 Number of obs     =    34945
```

| | Proportion | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c12b** | | | | |
| 0 | .3979769 | .0032991 | .3915106 | .4044432 |
| 1 | .6020231 | .0032991 | .5955568 | .6084894 |

```
. proportion c12c [pweight = weight], nolabel
Proportion estimation                 Number of obs     =    35040
```

| | Proportion | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c12c** | | | | |
| 0 | .3921942 | .0032963 | .3857334 | .398655 |
| 1 | .6078058 | .0032963 | .601345 | .6142666 |

```
. proportion c18 [pweight = weight], nolabel

Proportion estimation               Number of obs    =    23034
```

| | Proportion | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c18** | | | | |
| 0 | .5485991 | .0041751 | .5404155 | .5567826 |
| 1 | .4514009 | .0041751 | .4432174 | .4595845 |

```
. proportion c19a [pweight = weight], nolabel

Proportion estimation               Number of obs    =    35191
```

| | Proportion | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c19a** | | | | |
| 0 | .9916269 | .0005981 | .9904547 | .9927992 |
| 1 | .0083731 | .0005981 | .0072008 | .0095453 |

```
. proportion c19b [pweight = weight], nolabel

Proportion estimation               Number of obs    =    23000
```

| | Proportion | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c19b** | | | | |
| 0 | .9963508 | .0005116 | .9953481 | .9973535 |
| 1 | .0036492 | .0005116 | .0026465 | .0046519 |

```
. proportion c19c [pweight = weight], nolabel

Proportion estimation               Number of obs    =    23105
```

| | Proportion | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c19c** | | | | |
| 0 | .9983404 | .0003435 | .9976671 | .9990136 |
| 1 | .0016596 | .0003435 | .0009864 | .0023329 |

```
. proportion c25a [pweight = weight], nolabel

Proportion estimation               Number of obs    =    35548
```

| | Proportion | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c25a** | | | | |
| 0 | .9801901 | .0010687 | .9780954 | .9822848 |
| 1 | .0198099 | .0010687 | .0177152 | .0219046 |

```
. proportion c25b [pweight = weight], nolabel

Proportion estimation               Number of obs    =    23127
```

| | Proportion | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c25b** | | | | |
| 0 | .9846261 | .0011497 | .9823726 | .9868795 |
| 1 | .0153739 | .0011497 | .0131205 | .0176274 |

```
. proportion c25c [pweight = weight], nolabel

Proportion estimation               Number of obs    =    23158
```

| | Proportion | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c25c** | | | | |
| 0 | .991728 | .0008563 | .9900497 | .9934063 |
| 1 | .008272 | .0008563 | .0065937 | .0099503 |

```
. proportion c29a [pweight = weight], nolabel

Proportion estimation                Number of obs    =    35343
```

| | Proportion | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c29a** | | | | |
| 0 | .9979414 | .0003255 | .9973035 | .9985794 |
| 1 | .0020586 | .0003255 | .0014206 | .0026965 |

```
. proportion c31a [pweight = weight], nolabel

Proportion estimation                Number of obs    =    35489
```

| | Proportion | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31a** | | | | |
| 0 | .996241 | .000426 | .9954061 | .9970759 |
| 1 | .003759 | .000426 | .0029241 | .0045939 |

```
. proportion c31b [pweight = weight], nolabel

Proportion estimation               Number of obs    =   23155
```

| | Proportion | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31b** | | | | |
| 0 | .9987223 | .0003027 | .998129 | .9993155 |
| 1 | .0012777 | .0003027 | .0006845 | .001871 |

```
. proportion c31c [pweight = weight], nolabel

Proportion estimation               Number of obs    =   23093
```

| | Proportion | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31c** | | | | |
| 0 | .9983186 | .000385 | .9975639 | .9990733 |
| 1 | .0016814 | .000385 | .0009267 | .0024361 |

```
. proportion c31d [pweight = weight], nolabel

Proportion estimation               Number of obs    =   23086
```

| | Proportion | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31d** | | | | |
| 0 | .9984555 | .0003269 | .9978148 | .9990963 |
| 1 | .0015445 | .0003269 | .0009037 | .0021852 |

```
. proportion c31e [pweight = weight], nolabel

Proportion estimation               Number of obs    =   23155
```

| | Proportion | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31e** | | | | |
| 0 | .9968884 | .000514 | .9958809 | .997896 |
| 1 | .0031116 | .000514 | .002104 | .0041191 |

```
. proportion c31f [pweight = weight], nolabel

Proportion estimation               Number of obs    =   23101
```

| | Proportion | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31f** | | | | |
| 0 | .9989301 | .0002785 | .9983843 | .999476 |
| 1 | .0010699 | .0002785 | .000524 | .0016157 |

```
. proportion c31g [pweight = weight], nolabel

Proportion estimation               Number of obs    =   23142
```

| | Proportion | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31g** | | | | |
| 0 | .9980264 | .0003201 | .997399 | .9986538 |
| 1 | .0019736 | .0003201 | .0013462 | .002601 |

```
. proportion c31h [pweight = weight], nolabel

Proportion estimation               Number of obs    =   23076
```

| | Proportion | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31h** | | | | |
| 0 | .9974999 | .0003908 | .9967339 | .9982658 |
| 1 | .0025001 | .0003908 | .0017342 | .0032661 |

```
. proportion c31i [pweight = weight], nolabel

Proportion estimation               Number of obs    =   23018
```

| | Proportion | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31i** | | | | |
| 0 | .9991514 | .0002321 | .9986964 | .9996063 |
| 1 | .0008486 | .0002321 | .0003937 | .0013036 |

```
. proportion c31j [pweight = weight], nolabel
```

```
. proportion c31i [pweight = weight], nolabel

Proportion estimation               Number of obs    =    23018
```

| | Proportion | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31i** | | | | |
| 0 | .9991514 | .0002321 | .9986964 | .9996063 |
| 1 | .0008486 | .0002321 | .0003937 | .0013036 |

```
. proportion c31j [pweight = weight], nolabel

Proportion estimation               Number of obs    =    23097
```

| | Proportion | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31j** | | | | |
| 0 | .9983566 | .0003184 | .9977326 | .9989807 |
| 1 | .0016434 | .0003184 | .0010193 | .0022674 |

```
. proportion c31k [pweight = weight], nolabel

Proportion estimation               Number of obs    =    23142
```

| | Proportion | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31k** | | | | |
| 0 | .998644 | .0002451 | .9981636 | .9991245 |
| 1 | .001356 | .0002451 | .0008755 | .0018364 |

```
. proportion c31l [pweight = weight], nolabel

Proportion estimation               Number of obs    =    23187
```

| | Proportion | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31l** | | | | |
| 0 | .9982821 | .0003675 | .9975618 | .9990023 |
| 1 | .0017179 | .0003675 | .0009977 | .0024382 |

# Παράρτημα 2

## 2nd PROPORTION ΓΡΑΜΜΙΚΗ ΜΕΘΟΔΟΣ

```
      pweight: weight
          VCE: linearized
  Single unit: missing
     Strata 1: Nomos
         SU 1: CLASS
        FPC 1: <zero>

. svy linearized : proportion c08, nolabel
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      49          Number of obs    =   35462
Number of PSUs   =    2030          Population size  = 35457.1
                                     Design df        =    1981
```

| | Proportion | Linearized Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c08** | | | | |
| 0 | .6086507 | .0071811 | .5945673 | .6227341 |
| 1 | .3913493 | .0071811 | .3772659 | .4054327 |

```
. svy linearized : proportion c09, nolabel
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      49          Number of obs    =   35553
Number of PSUs   =    2030          Population size  = 35530.4
                                     Design df        =    1981
```

| | Proportion | Linearized Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c09** | | | | |
| 0 | .859001 | .0046727 | .8498372 | .8681648 |
| 1 | .140999 | .0046727 | .1318352 | .1501628 |

```
. svy linearized : proportion c12a, nolabel
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      49          Number of obs    =   34632
Number of PSUs   =    2030          Population size  = 34554.2
                                     Design df        =    1981
```

| | Proportion | Linearized Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c12a** | | | | |
| 0 | .2603722 | .0064445 | .2477334 | .2730109 |
| 1 | .7396278 | .0064445 | .7269891 | .7522666 |

```
. svy linearized : proportion c12b, nolabel
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      49          Number of obs    =   34945
Number of PSUs   =    2029          Population size  = 34906.7
                                     Design df        =    1980
```

| | Proportion | Linearized Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c12b** | | | | |
| 0 | .3979769 | .0077098 | .3828568 | .413097 |
| 1 | .6020231 | .0077098 | .586903 | .6171432 |

```
. svy linearized : proportion c12c, nolabel
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      49          Number of obs    =   35040
Number of PSUs   =    2029          Population size  = 34974.5
                                     Design df        =    1980
```

| | Proportion | Linearized Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c12c** | | | | |
| 0 | .3921942 | .0068031 | .3788522 | .4055362 |
| 1 | .6078058 | .0068031 | .5944638 | .6211478 |

```
. svy linearized : proportion c18, nolabel
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      49          Number of obs    =   23034
Number of PSUs   =    1328          Population size  = 23961.1
                                     Design df        =    1279
```

| | Proportion | Linearized Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c18** | | | | |
| 0 | .5485991 | .0057815 | .5372568 | .5599413 |
| 1 | .4514009 | .0057815 | .4400587 | .4627432 |

```
. svy linearized : proportion c19a, nolabel
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      49          Number of obs    =   35191
Number of PSUs   =    2030          Population size  = 35162.6
                                     Design df        =    1981
```

| | Proportion | Linearized Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c19a** | | | | |
| 0 | .9916269 | .0006879 | .9902778 | .992976 |
| 1 | .0083731 | .0006879 | .007024 | .0097222 |

```
. svy linearized : proportion c19b, nolabel
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      49          Number of obs    =   23000
Number of PSUs   =    1328          Population size  = 23944.6
                                     Design df        =    1279
```

| | Proportion | Linearized Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c19b** | | | | |
| 0 | .9963508 | .0005482 | .9952754 | .9974263 |
| 1 | .0036492 | .0005482 | .0025737 | .0047246 |

```
. svy linearized : proportion c19c, nolabel
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      49          Number of obs    =   23105
Number of PSUs   =    1328          Population size  =   24059
                                    Design df        =    1279
```

| | Proportion | Linearized Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c19c** | | | | |
| 0 | .9983404 | .0003414 | .9976707 | .9990101 |
| 1 | .0016596 | .0003414 | .0009899 | .0023293 |

```
. svy linearized : proportion c25a, nolabel
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      49          Number of obs    =   35548
Number of PSUs   =    2030          Population size  = 35546.6
                                    Design df        =    1981
```

| | Proportion | Linearized Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c25a** | | | | |
| 0 | .9801901 | .0015018 | .9772449 | .9831353 |
| 1 | .0198099 | .0015018 | .0168647 | .0227551 |

```
. svy linearized : proportion c25b, nolabel
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      49          Number of obs    =   23127
Number of PSUs   =    1328          Population size  = 24061.9
                                    Design df        =    1279
```

| | Proportion | Linearized Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c25b** | | | | |
| 0 | .9846261 | .0014618 | .9817584 | .9874938 |
| 1 | .0153739 | .0014618 | .0125062 | .0182416 |

```
. svy linearized : proportion c25c, nolabel
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      49          Number of obs    =   23158
Number of PSUs   =    1328          Population size  = 24110.1
                                    Design df        =    1279
```

| | Proportion | Linearized Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c25c** | | | | |
| 0 | .991728 | .0010743 | .9896203 | .9938357 |
| 1 | .008272 | .0010743 | .0061643 | .0103797 |

```
. svy linearized : proportion c29a, nolabel
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      49          Number of obs    =   35343
Number of PSUs   =    2030          Population size  =   35370
                                     Design df        =    1981
```

| | Proportion | Linearized Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c29a** | | | | |
| 0 | .9979414 | .0003225 | .997309 | .9985739 |
| 1 | .0020586 | .0003225 | .0014261 | .002691 |

```
. svy linearized : proportion c31a, nolabel
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      49          Number of obs    =   35489
Number of PSUs   =    2030          Population size  = 35489.5
                                     Design df        =    1981
```

| | Proportion | Linearized Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31a** | | | | |
| 0 | .996241 | .0004541 | .9953505 | .9971315 |
| 1 | .003759 | .0004541 | .0028685 | .0046495 |

```
. svy linearized : proportion c31b, nolabel
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      49          Number of obs    =   23155
Number of PSUs   =    1328          Population size  = 24109.9
                                     Design df        =    1279
```

| | Proportion | Linearized Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31b** | | | | |
| 0 | .9987223 | .0002989 | .9981359 | .9993087 |
| 1 | .0012777 | .0002989 | .0006913 | .0018641 |

```
. svy linearized : proportion c31c, nolabel
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      49          Number of obs    =   23093
Number of PSUs   =    1328          Population size  = 24044.7
                                     Design df        =    1279
```

| | Proportion | Linearized Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31c** | | | | |
| 0 | .9983186 | .0004526 | .9974306 | .9992066 |
| 1 | .0016814 | .0004526 | .0007934 | .0025694 |

```
. svy linearized : proportion c31d, nolabel
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      49          Number of obs    =   23086
Number of PSUs   =    1328          Population size  = 24035.8
                                     Design df        =    1279
```

| | Proportion | Linearized Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31d** | | | | |
| 0 | .9984555 | .0003786 | .9977128 | .9991982 |
| 1 | .0015445 | .0003786 | .0008018 | .0022872 |

```
. svy linearized : proportion c31e, nolabel
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      49          Number of obs    =   23155
Number of PSUs   =    1328          Population size  = 24106.5
                                     Design df        =    1279
```

| | Proportion | Linearized Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31e** | | | | |
| 0 | .9968884 | .0005641 | .9957818 | .997995 |
| 1 | .0031116 | .0005641 | .002005 | .0042182 |

```
. svy linearized : proportion c31f, nolabel
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      49          Number of obs    =   23101
Number of PSUs   =    1328          Population size  = 24061.9
                                     Design df        =    1279
```

| | Proportion | Linearized Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31f** | | | | |
| 0 | .9989301 | .0002773 | .9983861 | .9994741 |
| 1 | .0010699 | .0002773 | .0005259 | .0016139 |

```
. svy linearized : proportion c31g, nolabel
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      49          Number of obs    =   23142
Number of PSUs   =    1328          Population size  =   24087
                                     Design df        =    1279
```

| | Proportion | Linearized Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31g** | | | | |
| 0 | .9980264 | .0003199 | .9973987 | .998654 |
| 1 | .0019736 | .0003199 | .001346 | .0026013 |

```
. svy linearized : proportion c31h, nolabel
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      49          Number of obs    =   23076
Number of PSUs   =    1328          Population size  = 24031.3
                                    Design df        =    1279
```

| | Proportion | Linearized Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31h** | | | | |
| 0 | .9974999 | .0004106 | .9966943 | .9983054 |
| 1 | .0025001 | .0004106 | .0016946 | .0033057 |

```
. svy linearized : proportion c31i, nolabel
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      49          Number of obs    =   23018
Number of PSUs   =    1328          Population size  = 23970.7
                                    Design df        =    1279
```

| | Proportion | Linearized Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31i** | | | | |
| 0 | .9991514 | .0002541 | .9986529 | .9996499 |
| 1 | .0008486 | .0002541 | .0003501 | .0013471 |

```
. svy linearized : proportion c31j, nolabel
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      49          Number of obs    =   23097
Number of PSUs   =    1328          Population size  = 24048.9
                                    Design df        =    1279
```

| | Proportion | Linearized Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31j** | | | | |
| 0 | .9983566 | .0003116 | .9977454 | .9989679 |
| 1 | .0016434 | .0003116 | .0010321 | .0022546 |

```
. svy linearized : proportion c31k, nolabel
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      49          Number of obs    =   23142
Number of PSUs   =    1328          Population size  = 24085.9
                                    Design df        =    1279
```

| | Proportion | Linearized Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31k** | | | | |
| 0 | .998644 | .0002454 | .9981626 | .9991254 |
| 1 | .001356 | .0002454 | .0008746 | .0018374 |

```
. svy linearized : proportion c31l, nolabel
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      49          Number of obs    =   23187
Number of PSUs   =    1328          Population size  = 24132.2
                                    Design df        =    1279
```

| | Proportion | Linearized Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31l** | | | | |
| 0 | .9982821 | .0003718 | .9975527 | .9990114 |
| 1 | .0017179 | .0003718 | .0009886 | .0024473 |

# Παράρτημα 3

## 3rd JACKKNIFE PROPORTION

Survey: Proportion estimation

| | | | |
|---|---|---|---|
| Number of strata = | 49 | Number of obs = | 35462 |
| Number of PSUs = | 2030 | Population size = | 35457.1 |
| | | Replications = | 2030 |
| | | Design df = | 1981 |

| | Proportion | Jackknife Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c08** | | | | |
| 0 | .6086507 | .0071823 | .594565 | .6227363 |
| 1 | .3913493 | .0071823 | .3772637 | .405435 |

Survey: Proportion estimation

| | | | |
|---|---|---|---|
| Number of strata = | 49 | Number of obs = | 35553 |
| Number of PSUs = | 2030 | Population size = | 35530.4 |
| | | Replications = | 2030 |
| | | Design df = | 1981 |

_prop_2: c09 = **Καθόλου**

| | Proportion | Jackknife Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c09** | | | | |
| 0 | .859001 | .0046734 | .8498358 | .8681662 |
| _prop_2 | .140999 | .0046734 | .1318338 | .1501642 |

Survey: Proportion estimation

| | | | |
|---|---|---|---|
| Number of strata = | 49 | Number of obs = | 34632 |
| Number of PSUs = | 2030 | Population size = | 34554.2 |
| | | Replications = | 2030 |
| | | Design df = | 1981 |

| | Proportion | Jackknife Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c12a** | | | | |
| 0 | .2603722 | .0064453 | .2477319 | .2730124 |
| 1 | .7396278 | .0064453 | .7269875 | .7522681 |

Survey: Proportion estimation

| | | | |
|---|---|---|---|
| Number of strata = | 49 | Number of obs = | 34945 |
| Number of PSUs = | 2029 | Population size = | 34906.7 |
| | | Replications = | 2029 |
| | | Design df = | 1980 |

| | Proportion | Jackknife Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c12b** | | | | |
| 0 | .3979769 | .0077108 | .3828548 | .413099 |
| 1 | .6020231 | .0077108 | .586901 | .6171452 |

Survey: Proportion estimation

| | | | |
|---|---|---|---|
| Number of strata = | 49 | Number of obs = | 35040 |
| Number of PSUs = | 2029 | Population size = | 34974.5 |
| | | Replications = | 2029 |
| | | Design df = | 1980 |

| | Proportion | Jackknife Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c12c** | | | | |
| 0 | .3921942 | .006804 | .3788504 | .405538 |
| 1 | .6078058 | .006804 | .5944621 | .6211495 |

Survey: Proportion estimation

Number of strata = 49 Number of obs = 23034
Number of PSUs = 1328 Population size = 23961.1
Replications = 1328
Design df = 1279

| | Proportion | Jackknife Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c18** | | | | |
| 0 | .5485991 | .0057822 | .5372554 | .5599427 |
| 1 | .4514009 | .0057822 | .4400573 | .4627446 |

Survey: Proportion estimation

Number of strata = 49 Number of obs = 35191
Number of PSUs = 2030 Population size = 35162.6
Replications = 2030
Design df = 1981

| | Proportion | Jackknife Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c19a** | | | | |
| 0 | .9916269 | .000688 | .9902777 | .9929762 |
| 1 | .0083731 | .000688 | .0070238 | .0097224 |

Survey: Proportion estimation

Number of strata = 49 Number of obs = 23000
Number of PSUs = 1328 Population size = 23944.6
Replications = 1328
Design df = 1279

| | Proportion | Jackknife Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c19b** | | | | |
| 0 | .9963508 | .0005483 | .9952752 | .9974265 |
| 1 | .0036492 | .0005483 | .0025735 | .0047248 |

Survey: Proportion estimation

Number of strata = 49 Number of obs = 23105
Number of PSUs = 1328 Population size = 24059
Replications = 1328
Design df = 1279

| | Proportion | Jackknife Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c19c** | | | | |
| 0 | .9983404 | .0003414 | .9976706 | .9990101 |
| 1 | .0016596 | .0003414 | .0009899 | .0023294 |

Survey: Proportion estimation

Number of strata = 49 Number of obs = 35548
Number of PSUs = 2030 Population size = 35546.6
Replications = 2030
Design df = 1981

| | Proportion | Jackknife Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c25a** | | | | |
| 0 | .9801901 | .0015019 | .9772445 | .9831356 |
| 1 | .0198099 | .0015019 | .0168644 | .0227555 |

Survey: Proportion estimation

Number of strata = 49 Number of obs = 23127
Number of PSUs = 1328 Population size = 24061.9
Replications = 1328
Design df = 1279

| | Proportion | Jackknife Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c25b** | | | | |
| 0 | .9846261 | .0014619 | .981758 | .9874941 |
| 1 | .0153739 | .001462 | .0125058 | .0182421 |

Survey: Proportion estimation

Number of strata = 49 Number of obs = 23158
Number of PSUs = 1328 Population size = 24110.1
Replications = 1328
Design df = 1279

| | Proportion | Jackknife Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c25c** | | | | |
| 0 | .991728 | .0010747 | .9896197 | .9938363 |
| 1 | .008272 | .0010746 | .0061637 | .0103802 |

Survey: Proportion estimation

Number of strata = 49 Number of obs = 35343
Number of PSUs = 2030 Population size = 35370
Replications = 2030
Design df = 1981

| | Proportion | Jackknife Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c29a** | | | | |
| 0 | .9979414 | .0003225 | .9973089 | .9985739 |
| 1 | .0020586 | .0003225 | .0014261 | .0026911 |

Survey: Proportion estimation

Number of strata = 49 Number of obs = 35489
Number of PSUs = 2030 Population size = 35489.5
Replications = 2030
Design df = 1981

| | Proportion | Jackknife Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31a** | | | | |
| 0 | .996241 | .0004541 | .9953505 | .9971316 |
| 1 | .003759 | .0004541 | .0028684 | .0046495 |

Survey: Proportion estimation

Number of strata = 49 Number of obs = 23155
Number of PSUs = 1328 Population size = 24109.9
Replications = 1328
Design df = 1279

| | Proportion | Jackknife Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31b** | | | | |
| 0 | .9987223 | .0002989 | .998136 | .9993086 |
| 1 | .0012777 | .0002989 | .0006914 | .0018641 |

Survey: Proportion estimation

Number of strata = 49 Number of obs = 23093
Number of PSUs = 1328 Population size = 24044.7
Replications = 1328
Design df = 1279

| | Proportion | Jackknife Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31c** | | | | |
| 0 | .9983186 | .0004528 | .9974303 | .9992069 |
| 1 | .0016814 | .0004528 | .0007931 | .0025697 |

Survey: Proportion estimation

Number of strata = 49 Number of obs = 23086
Number of PSUs = 1328 Population size = 24035.8
Replications = 1328
Design df = 1279

| | Proportion | Jackknife Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31d** | | | | |
| 0 | .9984555 | .0003786 | .9977129 | .9991982 |
| 1 | .0015445 | .0003786 | .0008018 | .0022871 |

Survey: Proportion estimation

Number of strata = 49 Number of obs = 23155
Number of PSUs = 1328 Population size = 24106.5
Replications = 1328
Design df = 1279

| | Proportion | Jackknife Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31e** | | | | |
| 0 | .9968884 | .000564 | .9957819 | .9979949 |
| 1 | .0031116 | .000564 | .0020051 | .0042181 |

Survey: Proportion estimation

Number of strata = 49 Number of obs = 23101
Number of PSUs = 1328 Population size = 24061.9
Replications = 1328
Design df = 1279

| | Proportion | Jackknife Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31f** | | | | |
| 0 | .9989301 | .0002773 | .9983861 | .9994742 |
| 1 | .0010699 | .0002774 | .0005258 | .001614 |

Survey: Proportion estimation

Number of strata = 49 Number of obs = 23142
Number of PSUs = 1328 Population size = 24087
Replications = 1328
Design df = 1279

| | Proportion | Jackknife Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31g** | | | | |
| 0 | .9980264 | .00032 | .9973987 | .9986541 |
| 1 | .0019736 | .0003199 | .001346 | .0026013 |

Survey: Proportion estimation

Number of strata = 49 Number of obs = 23076
Number of PSUs = 1328 Population size = 24031.3
Replications = 1328
Design df = 1279

| | Proportion | Jackknife Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31h** | | | | |
| 0 | .9974999 | .0004106 | .9966942 | .9983055 |
| 1 | .0025001 | .0004106 | .0016945 | .0033058 |

Survey: Proportion estimation

Number of strata = 49 Number of obs = 23018
Number of PSUs = 1328 Population size = 23970.7
Replications = 1328
Design df = 1279

| | Proportion | Jackknife Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31i** | | | | |
| 0 | .9991514 | .0002541 | .9986528 | .99965 |
| 1 | .0008486 | .0002541 | .0003501 | .0013472 |

.

Survey: Proportion estimation

Number of strata = 49 Number of obs = 23097
Number of PSUs = 1328 Population size = 24048.9
Replications = 1328
Design df = 1279

| | Proportion | Jackknife Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31j** | | | | |
| 0 | .9983566 | .0003116 | .9977453 | .998968 |
| 1 | .0016434 | .0003116 | .0010321 | .0022546 |

Survey: Proportion estimation

Number of strata = 49 Number of obs = 23142
Number of PSUs = 1328 Population size = 24085.9
Replications = 1328
Design df = 1279

| | Proportion | Jackknife Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31k** | | | | |
| 0 | .998644 | .0002454 | .9981626 | .9991255 |
| 1 | .001356 | .0002454 | .0008746 | .0018374 |

Survey: Proportion estimation

Number of strata = 49 Number of obs = 23187
Number of PSUs = 1328 Population size = 24132.2
Replications = 1328
Design df = 1279

| | Proportion | Jackknife Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| **c31l** | | | | |
| 0 | .9982821 | .0003718 | .9975527 | .9990114 |
| 1 | .0017179 | .0003718 | .0009885 | .0024473 |

# Παράρτημα 4

## 1ST REGRESSION

```
. logistic c08 new_1  new_2 new_3 [pweight = weight]

Logistic regression                                     Number of obs   =      35462
                                                        Wald chi2(3)    =    3031.87
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -20788.223                       Pseudo R2       =     0.1242

------------------------------------------------------------------------------
             |               Robust
         c08 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       new_1 |   .0499401   .0047586   -31.45   0.000     .0414326    .0601944
       new_2 |   .1911125   .0176327   -17.94   0.000     .1594977    .2289939
       new_3 |   .4400422    .040787    -8.86   0.000     .3669421     .527705
------------------------------------------------------------------------------

. logistic c08 c05_1 c05_3 [pweight = weight]

Logistic regression                                     Number of obs   =      34942
                                                        Wald chi2(2)    =    1447.64
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -22061.624                       Pseudo R2       =     0.0559

------------------------------------------------------------------------------
             |               Robust
         c08 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       c05_1 |   2.117299    .088074    18.03   0.000     1.951527    2.297153
       c05_3 |   .4025823   .0135448   -27.04   0.000     .3768912    .4300245
------------------------------------------------------------------------------

. logistic c08 c01 [pweight = weight]

Logistic regression                                     Number of obs   =      35462
                                                        Wald chi2(1)    =      22.34
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -23717.867                       Pseudo R2       =     0.0008

------------------------------------------------------------------------------
             |               Robust
         c08 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         c01 |   .8758106    .024572    -4.73   0.000     .8289505    .9253197
------------------------------------------------------------------------------


. logistic c09 c01 [pweight = weight]

Logistic regression                                     Number of obs   =      35553
                                                        Wald chi2(1)    =      51.58
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -14415.515                       Pseudo R2       =     0.0032

------------------------------------------------------------------------------
             |               Robust
         c09 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         c01 |   .7443612   .0305988    -7.18   0.000      .686741     .806816
------------------------------------------------------------------------------

. logistic c09 new_1 new_2 new_3 [pweight = weight]

Logistic regression                                     Number of obs   =      35553
                                                        Wald chi2(3)    =    1771.12
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -12454.873                       Pseudo R2       =     0.1388

------------------------------------------------------------------------------
             |               Robust
         c09 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       new_1 |   .0182445   .0020279   -36.02   0.000     .0146731    .0226851
       new_2 |   .1145368    .009765   -25.42   0.000     .0969113    .1353677
       new_3 |   .3101856   .0254752   -14.25   0.000     .2640667    .3643592
------------------------------------------------------------------------------

. logistic c09 c05_1 c05_3 [pweight = weight]

Logistic regression                                     Number of obs   =      35029
                                                        Wald chi2(2)    =    1102.97
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -13057.284                       Pseudo R2       =     0.0778

------------------------------------------------------------------------------
             |               Robust
         c09 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       c05_1 |    2.75919   .1335256    20.97   0.000     2.509512    3.033708
       c05_3 |   .3130685   .0192476   -18.89   0.000     .2775283    .3531601
------------------------------------------------------------------------------
```

```
. logistic c12a c01 [pweight = weight]

Logistic regression                                     Number of obs   =      34632
                                                        Wald chi2(1)    =     120.04
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -19766.707                       Pseudo R2       =     0.0047

------------------------------------------------------------------------------
             |               Robust
         c12a | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          c01 |   .7156907   .0218507   -10.96   0.000     .6741204    .7598246
------------------------------------------------------------------------------

. logistic c12a new_1 new_2 new_3 [pweight = weight]

Logistic regression                                     Number of obs   =      34632
                                                        Wald chi2(3)    =    3091.97
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -17022.598                       Pseudo R2       =     0.1429

------------------------------------------------------------------------------
             |               Robust
         c12a | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       new_1 |   .1246746   .0148042   -17.53   0.000     .0987879    .1573447
       new_2 |   .4917048   .0588128    -5.93   0.000     .3889487    .6216079
       new_3 |   1.418241   .1760369     2.82   0.005     1.111977    1.808856
------------------------------------------------------------------------------

. logistic c12a c05_1 c05_3 [pweight = weight]

Logistic regression                                     Number of obs   =      34122
                                                        Wald chi2(2)    =     171.15
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -19451.834                       Pseudo R2       =     0.0066

------------------------------------------------------------------------------
             |               Robust
         c12a | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       c05_1 |   1.028776   .0497682     0.59   0.558     .9357141    1.131094
       c05_3 |   .6702886   .0220411   -12.17   0.000     .6284515    .7149109
------------------------------------------------------------------------------

. logistic c18 c01 [pweight = weight]

Logistic regression                                     Number of obs   =      23034
                                                        Wald chi2(1)    =     342.24
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -15576.173                       Pseudo R2       =     0.0177

------------------------------------------------------------------------------
             |               Robust
         c18 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          c01 |    .531577   .0181573   -18.50   0.000     .4971543     .568383
------------------------------------------------------------------------------

. logistic c18 new_1 new_2 new_3 [pweight = weight]

Logistic regression                                     Number of obs   =      23034
                                                        Wald chi2(3)    =     127.74
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -15751.136                       Pseudo R2       =     0.0067

------------------------------------------------------------------------------
             |               Robust
         c18 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       new_1 |   .3320782   .4703762    -0.78   0.436     .0206798    5.332551
       new_2 |   .5131533   .0424399    -8.07   0.000     .4363643    .6034552
       new_3 |   .7084438   .0588818    -4.15   0.000     .6019472    .8337817
------------------------------------------------------------------------------

. logistic c18 c05_1 c05_3 [pweight = weight]

Logistic regression                                     Number of obs   =      22629
                                                        Wald chi2(2)    =     686.14
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -14988.276                       Pseudo R2       =     0.0375

------------------------------------------------------------------------------
             |               Robust
         c18 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       c05_1 |   1.606455   .0797507     9.55   0.000     1.457511    1.770621
       c05_3 |   .4421232   .0175419   -20.57   0.000     .4090445    .4778769
------------------------------------------------------------------------------
```

```
. logistic c19a c01 [pweight = weight]

Logistic regression                                     Number of obs   =      35191
                                                        Wald chi2(1)    =      72.87
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -1641.0474                       Pseudo R2       =     0.0362

------------------------------------------------------------------------------
             |               Robust
         c19a | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          c01 |   .2288951   .0395371    -8.54   0.000     .1631579    .3211181
------------------------------------------------------------------------------

. logistic c19a new_1 new_2 new_3 [pweight = weight]

Logistic regression                                     Number of obs   =      35191
                                                        Wald chi2(3)    =     131.93
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -1612.4126                       Pseudo R2       =     0.0530

------------------------------------------------------------------------------
             |               Robust
         c19a | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       new_1 |   .0464095   .0134719   -10.58   0.000     .0262734     .081978
       new_2 |   .2016412   .0474353    -6.81   0.000     .1271562    .3197577
       new_3 |   .4206913   .0913312    -3.99   0.000     .2748966    .6438098
------------------------------------------------------------------------------

. logistic c19a c05_1 c05_3 [pweight = weight]

Logistic regression                                     Number of obs   =      34658
                                                        Wald chi2(2)    =     100.11
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -1562.8428                       Pseudo R2       =     0.0451

------------------------------------------------------------------------------
             |               Robust
         c19a | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       c05_1 |   2.934684   .4712898     6.70   0.000     2.142221    4.020299
       c05_3 |   .3596642   .0778574    -4.72   0.000     .2353073    .5497421
------------------------------------------------------------------------------


. logistic c25a c01 [pweight = weight]

Logistic regression                                     Number of obs   =      35548
                                                        Wald chi2(1)    =     120.69
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -3316.7286                       Pseudo R2       =     0.0411

------------------------------------------------------------------------------
             |               Robust
         c25a | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          c01 |   .2359029    .031015   -10.99   0.000     .1823152    .3052416
------------------------------------------------------------------------------

. logistic c25a new_1 new_2 new_3 [pweight = weight]

Logistic regression                                     Number of obs   =      35548
                                                        Wald chi2(3)    =     382.20
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -3067.6527                       Pseudo R2       =     0.1131

------------------------------------------------------------------------------
             |               Robust
         c25a | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       new_1 |   .0186944   .0046124   -16.13   0.000     .0115265    .0303198
       new_2 |   .0663756   .0114447   -15.73   0.000     .0473414    .0930628
       new_3 |   .2223557   .0316348   -10.57   0.000     .1682467    .2938663
------------------------------------------------------------------------------

. logistic c25a c05_1 c05_3 [pweight = weight]

Logistic regression                                     Number of obs   =      35008
                                                        Wald chi2(2)    =     182.96
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -3142.8276                       Pseudo R2       =     0.0554

------------------------------------------------------------------------------
             |               Robust
         c25a | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       c05_1 |   3.490679   .4272879    10.21   0.000     2.746096    4.437148
       c05_3 |   .4583297   .0781694    -4.57   0.000     .3280978    .6402546
------------------------------------------------------------------------------
```

```
. logistic c29a c01 [pweight = weight]

Logistic regression                                     Number of obs   =      35343
                                                        Wald chi2(1)    =      19.20
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -500.68491                       Pseudo R2       =     0.0422

------------------------------------------------------------------------------
             |               Robust
         c29a | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          c01 |   .1486709   .0646616    -4.38   0.000     .0633887    .3486904
------------------------------------------------------------------------------

. logistic c29a new_1 new_2 new_3 [pweight = weight]

Logistic regression                                     Number of obs   =      35343
                                                        Wald chi2(3)    =      19.82
                                                        Prob > chi2     =     0.0002
Log pseudolikelihood = -507.17048                       Pseudo R2       =     0.0298

------------------------------------------------------------------------------
             |               Robust
         c29a | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       new_1 |   .1266193   .0671017    -3.90   0.000     .0448133     .357761
       new_2 |   .1301136   .0632319    -4.20   0.000      .050195    .3372759
       new_3 |   .1766269   .0887387    -3.45   0.001     .0659791     .472832
------------------------------------------------------------------------------

. logistic c29a c05_1 c05_3 [pweight = weight]

Logistic regression                                     Number of obs   =      34811
                                                        Wald chi2(2)    =       0.93
                                                        Prob > chi2     =     0.6293
Log pseudolikelihood = -489.58237                       Pseudo R2       =     0.0015

------------------------------------------------------------------------------
             |               Robust
         c29a | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       c05_1 |   1.428514   .5561379     0.92   0.360     .6660417    3.063851
       c05_3 |   1.298652   .4877041     0.70   0.487     .6220524    2.711181
------------------------------------------------------------------------------


. logistic c31a c01 [pweight = weight]

Logistic regression                                     Number of obs   =      35489
                                                        Wald chi2(1)    =       2.80
                                                        Prob > chi2     =     0.0943
Log pseudolikelihood = -875.64866                       Pseudo R2       =     0.0027

------------------------------------------------------------------------------
             |               Robust
         c31a | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          c01 |   .6826779   .1557539    -1.67   0.094      .436529    1.067625
------------------------------------------------------------------------------

. logistic c31a new_1 new_2 new_3 [pweight = weight]

Logistic regression                                     Number of obs   =      35489
                                                        Wald chi2(3)    =      33.83
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -836.71202                       Pseudo R2       =     0.0470

------------------------------------------------------------------------------
             |               Robust
         c31a | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       new_1 |   .0292855   .0182304    -5.67   0.000     .0086451    .0992046
       new_2 |   .2569923   .0957413    -3.65   0.000     .1238252    .5333733
       new_3 |   .3271961   .1211052    -3.02   0.003     .1583991    .6758706
------------------------------------------------------------------------------

. logistic c31a c05_1 c05_3 [pweight = weight]

Logistic regression                                     Number of obs   =      34949
                                                        Wald chi2(2)    =       9.83
                                                        Prob > chi2     =     0.0073
Log pseudolikelihood = -852.15761                       Pseudo R2       =     0.0096

------------------------------------------------------------------------------
             |               Robust
         c31a | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       c05_1 |   2.406904   .6957427     3.04   0.002      1.36587    4.24139
       c05_3 |   1.113386   .2969155     0.40   0.687     .6601596    1.87777
------------------------------------------------------------------------------
```

```
. logistic c31k c01 [pweight = weight]

Logistic regression                                     Number of obs   =      23142
                                                        Wald chi2(1)    =       8.69
                                                        Prob > chi2     =     0.0032
Log pseudolikelihood = -232.07578                       Pseudo R2       =     0.0272

------------------------------------------------------------------------------
             |               Robust
         c31k | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          c01 |   .2312945   .1148901    -2.95   0.003     .0873677    .6123215
------------------------------------------------------------------------------

. logistic c31k new_1 new_2 new_3 [pweight = weight]

note: new_1 != 0 predicts failure perfectly
      new_1 dropped and 2 obs not used

Logistic regression                                     Number of obs   =      23140
                                                        Wald chi2(2)    =       9.82
                                                        Prob > chi2     =     0.0074
Log pseudolikelihood = -234.28572                       Pseudo R2       =     0.0180

------------------------------------------------------------------------------
             |               Robust
         c31k | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       new_2 |   .5082806   .2859342    -1.20   0.229     .1687557    1.530906
       new_3 |   .1989796   .1166108    -2.76   0.006     .0630909    .6275528
------------------------------------------------------------------------------

. logistic c31k c05_1 c05_3 [pweight = weight]

Logistic regression                                     Number of obs   =      22738
                                                        Wald chi2(2)    =       1.49
                                                        Prob > chi2     =     0.4752
Log pseudolikelihood = -233.84064                       Pseudo R2       =     0.0022

------------------------------------------------------------------------------
             |               Robust
         c31k | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       c05_1 |   1.595231   .6369003     1.17   0.242     .7294221    3.488736
       c05_3 |   1.071871   .5006835     0.15   0.882     .4290783    2.677619
------------------------------------------------------------------------------


. logistic c31l c01 [pweight = weight]

Logistic regression                                     Number of obs   =      23187
                                                        Wald chi2(1)    =       2.46
                                                        Prob > chi2     =     0.1167
Log pseudolikelihood = -290.95446                       Pseudo R2       =     0.0084

------------------------------------------------------------------------------
             |               Robust
         c31l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          c01 |   .4826315   .2241521    -1.57   0.117     .1942164    1.199349
------------------------------------------------------------------------------

. logistic c31l new_1 new_2 new_3 [pweight = weight]

note: new_1 != 0 predicts failure perfectly
      new_1 dropped and 2 obs not used

Logistic regression                                     Number of obs   =      23185
                                                        Wald chi2(2)    =      20.09
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -281.64807                       Pseudo R2       =     0.0401

------------------------------------------------------------------------------
             |               Robust
         c31l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       new_2 |   .1256471   .0628721    -4.15   0.000     .0471219    .3350288
       new_3 |   .1409233   .0764425    -3.61   0.000     .0486695     .408046
------------------------------------------------------------------------------

. logistic c31l c05_1 c05_3 [pweight = weight]

Logistic regression                                     Number of obs   =      22782
                                                        Wald chi2(2)    =       0.20
                                                        Prob > chi2     =     0.9044
Log pseudolikelihood = -287.11466                       Pseudo R2       =     0.0007

------------------------------------------------------------------------------
             |               Robust
         c31l | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       c05_1 |   1.010958    .567518     0.02   0.985     .3364319    3.03787
       c05_3 |    1.23758   .6160749     0.43   0.669     .4664874    3.28327
------------------------------------------------------------------------------
```

# Παράρτημα 5

## 2nd REGRESSION

```
. svyset CLASS [pweight=weight], strata(Nomos) vce(linearized) singleunit(missi
> ng)

      pweight: weight
          VCE: linearized
  Single unit: missing
     Strata 1: Nomos
         SU 1: CLASS
        FPC 1: <zero>

. svy linearized : logistic c08 c01
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     35462
Number of PSUs     =      2030                  Population size    = 35457.064
                                                 Design df          =      1981
                                                 F(   1,   1981)    =     14.06
                                                 Prob > F           =    0.0002

------------------------------------------------------------------------------
             |             Linearized
         c08 | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         c01 |   .8758106   .0309717    -3.75   0.000     .8171284     .938707
------------------------------------------------------------------------------

. svy linearized : logistic c08 agegroup_1 agegroup_2 agegroup_3
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     35462
Number of PSUs     =      2030                  Population size    = 35457.064
                                                 Design df          =      1981
                                                 F(   3,   1979)    =    639.04
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
         c08 | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  agegroup_1 |   .0499401   .0057235   -26.15   0.000     .0398873    .0625264
  agegroup_2 |   .1911125   .0215481   -14.68   0.000     .1531992    .2384085
  agegroup_3 |   .4400422   .0497857    -7.26   0.000     .3524779    .5493597
------------------------------------------------------------------------------

. svy linearized : logistic c08 c05_1 c05_3
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     34942
Number of PSUs     =      2030                  Population size    = 34909.82
                                                 Design df          =      1981
                                                 F(   2,   1980)    =    469.22
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
         c08 | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        c05_1 |   2.117299   .0989527    16.05   0.000     1.931865    2.320533
        c05_3 |   .4025823   .0160692   -22.79   0.000     .3722698     .435363
------------------------------------------------------------------------------
```

```
. svy linearized : logistic c09 c01
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     35553
Number of PSUs     =      2030                  Population size    = 35530.373
                                                 Design df          =      1981
                                                 F(   1,   1981)    =     33.45
                                                 Prob > F           =    0.0000
```

| c09 | Odds Ratio | Linearized Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| c01 | .7443612 | .0379965 | -5.78 | 0.000 | .6734524 | .8227361 |

```
. svy linearized : logistic c09 agegroup_1 agegroup_2 agegroup_3
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     35553
Number of PSUs     =      2030                  Population size    = 35530.373
                                                 Design df          =      1981
                                                 F(   3,   1979)    =    419.75
                                                 Prob > F           =    0.0000
```

| c09 | Odds Ratio | Linearized Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| agegroup_1 | .0182445 | .0022432 | -32.56 | 0.000 | .0143354 | .0232194 |
| agegroup_2 | .1145368 | .0105096 | -23.61 | 0.000 | .0956737 | .1371188 |
| agegroup_3 | .3101856 | .0258337 | -14.06 | 0.000 | .2634428 | .3652221 |

```
. svy linearized : logistic c09 c05_1 c05_3
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     35029
Number of PSUs     =      2030                  Population size    = 34976.776
                                                 Design df          =      1981
                                                 F(   2,   1980)    =    374.79
                                                 Prob > F           =    0.0000
```

| c09 | Odds Ratio | Linearized Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| c05_1 | 2.75919 | .1468214 | 19.07 | 0.000 | 2.485764 | 3.062691 |
| c05_3 | .3130685 | .0215916 | -16.84 | 0.000 | .2734627 | .3584106 |

```
. svy linearized : logistic c12a c01
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     34632
Number of PSUs     =      2030                  Population size    = 34554.177
                                                 Design df          =      1981
                                                 F(   1,   1981)    =     87.71
                                                 Prob > F           =    0.0000
```

| c12a | Odds Ratio | Linearized Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| c01 | .7156907 | .0255629 | -9.37 | 0.000 | .6672733 | .7676214 |

```
. svy linearized : logistic c12a agegroup_1 agegroup_2 agegroup_3
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     34632
Number of PSUs     =      2030                  Population size    = 34554.177
                                                 Design df          =      1981
                                                 F(   3,   1979)    =    661.93
                                                 Prob > F           =    0.0000
```

| c12a | Odds Ratio | Linearized Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| agegroup_1 | .1246746 | .0162205 | -16.00 | 0.000 | .0965977 | .1609123 |
| agegroup_2 | .4917048 | .0642437 | -5.43 | 0.000 | .3805596 | .6353107 |
| agegroup_3 | 1.418241 | .190405 | 2.60 | 0.009 | 1.08994 | 1.845429 |

```
. svy linearized : logistic c12a c05_1 c05_3
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     34122
Number of PSUs     =      2030                  Population size    = 34021.388
                                                 Design df          =      1981
                                                 F(   2,   1980)    =     58.31
                                                 Prob > F           =    0.0000
```

| c12a | Odds Ratio | Linearized Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| c05_1 | 1.028776 | .0537115 | 0.54 | 0.587 | .9286528 | 1.139695 |
| c05_3 | .6702886 | .02541 | -10.55 | 0.000 | .6222629 | .7220209 |

```
. svy linearized : logistic c18 c01
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     23034
Number of PSUs     =      1328                  Population size    = 23961.116
                                                 Design df          =      1279
                                                 F(   1,   1279)    =    265.18
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
         c18 | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         c01 |    .531577   .0206275   -16.28   0.000     .4926115    .5736246
------------------------------------------------------------------------------

. svy linearized : logistic c18 agegroup_1 agegroup_2 agegroup_3
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     23034
Number of PSUs     =      1328                  Population size    = 23961.116
                                                 Design df          =      1279
                                                 F(   3,   1277)    =     25.81
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
         c18 | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  agegroup_1 |   .3320782   .4705222    -0.78   0.437     .0206077    5.351206
  agegroup_2 |   .5131533   .0484493    -7.07   0.000     .4263879    .6175744
  agegroup_3 |   .7084438   .0652724    -3.74   0.000     .5912972    .8487993
------------------------------------------------------------------------------

. svy linearized : logistic c18 c05_1 c05_3
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     22629
Number of PSUs     =      1328                  Population size    = 23516.565
                                                 Design df          =      1279
                                                 F(   2,   1278)    =    297.83
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
         c18 | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       c05_1 |   1.606455   .0824814     9.23   0.000     1.452524    1.776699
       c05_3 |   .4421232   .0183056   -19.71   0.000     .4076307    .4795343
------------------------------------------------------------------------------

.
```

```
. svy linearized : logistic c19a c01
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     35191
Number of PSUs     =      2030                  Population size    = 35162.595
                                                 Design df          =      1981
                                                 F(   1,   1981)    =     70.54
                                                 Prob > F           =    0.0000
```

| c19a | Odds Ratio | Linearized Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| c01 | .2288951 | .0401855 | -8.40 | 0.000 | .1622205 | .3229739 |

```
. svy linearized : logistic c19a agegroup_1 agegroup_2 agegroup_3
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     35191
Number of PSUs     =      2030                  Population size    = 35162.595
                                                 Design df          =      1981
                                                 F(   3,   1979)    =     36.46
                                                 Prob > F           =    0.0000
```

| c19a | Odds Ratio | Linearized Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| agegroup_1 | .0464095 | .0148668 | -9.58 | 0.000 | .0247609 | .0869854 |
| agegroup_2 | .2016412 | .0446225 | -7.24 | 0.000 | .1306459 | .3112165 |
| agegroup_3 | .4206913 | .087228 | -4.18 | 0.000 | .2801327 | .6317762 |

```
. svy linearized : logistic c19a c05_1 c05_3
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     34658
Number of PSUs     =      2030                  Population size    = 34600.262
                                                 Design df          =      1981
                                                 F(   2,   1980)    =     47.75
                                                 Prob > F           =    0.0000
```

| c19a | Odds Ratio | Linearized Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| c05_1 | 2.934684 | .4634693 | 6.82 | 0.000 | 2.153032 | 4.000113 |
| c05_3 | .3596642 | .075674 | -4.86 | 0.000 | .2380638 | .5433767 |

```
.
```

```
. svy linearized : logistic c25a c01
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     35548
Number of PSUs     =      2030                  Population size    = 35546.603
                                                 Design df          =      1981
                                                 F(   1,   1981)    =     71.16
                                                 Prob > F           =    0.0000
```

| c25a | Odds Ratio | Linearized Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| c01 | .2359029 | .0403913 | -8.44 | 0.000 | .1686169 | .330039 |

```
. svy linearized : logistic c25a agegroup_1 agegroup_2 agegroup_3
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     35548
Number of PSUs     =      2030                  Population size    = 35546.603
                                                 Design df          =      1981
                                                 F(   3,   1979)    =    100.91
                                                 Prob > F           =    0.0000
```

| c25a | Odds Ratio | Linearized Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| agegroup_1 | .0186944 | .0052334 | -14.22 | 0.000 | .0107964 | .0323702 |
| agegroup_2 | .0663756 | .0125748 | -14.32 | 0.000 | .0457774 | .0962423 |
| agegroup_3 | .2223557 | .036296 | -9.21 | 0.000 | .1614427 | .3062514 |

```
. svy linearized : logistic c25a c05_1 c05_3
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     35008
Number of PSUs     =      2030                  Population size    = 34982.093
                                                 Design df          =      1981
                                                 F(   2,   1980)    =     66.70
                                                 Prob > F           =    0.0000
```

| c25a | Odds Ratio | Linearized Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| c05_1 | 3.490679 | .4672248 | 9.34 | 0.000 | 2.684772 | 4.538499 |
| c05_3 | .4583297 | .0794988 | -4.50 | 0.000 | .32617 | .6440387 |

```
. svy linearized : logistic c29a c01
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     35343
Number of PSUs     =      2030                  Population size    = 35369.979
                                                 Design df          =      1981
                                                 F(   1,   1981)    =     19.17
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |                 Linearized
         c29a | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          c01 |   .1486709   .0647155    -4.38   0.000     .0633106    .3491204
------------------------------------------------------------------------------

. svy linearized : logistic c29a agegroup_1 agegroup_2 agegroup_3
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     35343
Number of PSUs     =      2030                  Population size    = 35369.979
                                                 Design df          =      1981
                                                 F(   3,   1979)    =      7.13
                                                 Prob > F           =    0.0001

------------------------------------------------------------------------------
             |                 Linearized
         c29a | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  agegroup_1 |   .1266193   .0649799    -4.03   0.000     .0462811    .3464143
  agegroup_2 |   .1301136   .0610709    -4.34   0.000     .0518267    .3266572
  agegroup_3 |   .1766269   .0861085    -3.56   0.000     .0678935    .4594997
------------------------------------------------------------------------------

. svy linearized : logistic c29a c05_1 c05_3
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     34811
Number of PSUs     =      2030                  Population size    = 34814.967
                                                 Design df          =      1981
                                                 F(   2,   1980)    =      0.46
                                                 Prob > F           =    0.6316

------------------------------------------------------------------------------
             |                 Linearized
         c29a | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       c05_1 |   1.428514   .5561125     0.92   0.360     .6657543    3.065173
       c05_3 |   1.298652   .4936218     0.69   0.492     .6162407    2.736749
------------------------------------------------------------------------------

.
```

```
. svy linearized : logistic c31a c01
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     35489
Number of PSUs     =      2030                  Population size    = 35489.529
                                                 Design df          =      1981
                                                 F(   1,   1981)    =      2.83
                                                 Prob > F           =    0.0924

------------------------------------------------------------------------------
             |             Linearized
         c31a | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          c01 |   .6826779   .1547912    -1.68   0.092     .4376182    1.064967
------------------------------------------------------------------------------

. svy linearized : logistic c31a agegroup_1 agegroup_2 agegroup_3
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     35489
Number of PSUs     =      2030                  Population size    = 35489.529
                                                 Design df          =      1981
                                                 F(   3,   1979)    =     11.46
                                                 Prob > F           =    0.0000

------------------------------------------------------------------------------
             |             Linearized
         c31a | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  agegroup_1 |   .0292855   .0180304    -5.73   0.000     .0087552    .0979577
  agegroup_2 |   .2569923   .0950955    -3.67   0.000     .1243813    .5309884
  agegroup_3 |   .3271961   .1110978    -3.29   0.001     .1681165    .6368045
------------------------------------------------------------------------------

. svy linearized : logistic c31a c05_1 c05_3
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     34949
Number of PSUs     =      2030                  Population size    = 34924.614
                                                 Design df          =      1981
                                                 F(   2,   1980)    =      4.04
                                                 Prob > F           =    0.0178

------------------------------------------------------------------------------
             |             Linearized
         c31a | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       c05_1 |   2.406904   .7570959     2.79   0.005     1.298818    4.460355
       c05_3 |   1.113386   .3011914     0.40   0.691     .6549967    1.892571
------------------------------------------------------------------------------

.
```

```
. svy linearized : logistic c31k c01
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     23142
Number of PSUs     =      1328                  Population size    = 24085.875
                                                 Design df          =      1279
                                                 F(   1,   1279)    =      8.62
                                                 Prob > F           =    0.0034
```

| c31k | Odds Ratio | Linearized Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| c01 | .2312945 | .1153329 | -2.94 | 0.003 | .08696 | .615193 |

```
. svy linearized : logistic c31k agegroup_1 agegroup_2 agegroup_3
(running logistic on estimation sample)

note: agegroup_1 != 0 predicts failure perfectly
      agegroup_1 dropped and 2 obs not used

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     23140
Number of PSUs     =      1328                  Population size    = 24083.321
                                                 Design df          =      1279
                                                 F(   2,   1278)    =      4.90
                                                 Prob > F           =    0.0076
```

| c31k | Odds Ratio | Linearized Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| agegroup_2 | .5082806 | .2866266 | -1.20 | 0.230 | .1681297 | 1.536607 |
| agegroup_3 | .1989796 | .1166641 | -2.75 | 0.006 | .0629892 | .6285659 |

```
. svy linearized : logistic c31k c05_1 c05_3
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     22738
Number of PSUs     =      1328                  Population size    = 23644.094
                                                 Design df          =      1279
                                                 F(   2,   1278)    =      0.72
                                                 Prob > F           =    0.4864
```

| c31k | Odds Ratio | Linearized Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| c05_1 | 1.595231 | .6476371 | 1.15 | 0.250 | .7193207 | 3.537728 |
| c05_3 | 1.071871 | .4979853 | 0.15 | 0.881 | .4308287 | 2.66674 |

```
.
```

```
. svy linearized : logistic c31l c01
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     23187
Number of PSUs     =      1328                  Population size    = 24132.214
                                                 Design df          =      1279
                                                 F(   1,   1279)    =      2.50
                                                 Prob > F           =    0.1139
```

| c31l | Odds Ratio | Linearized Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| c01 | .4826315 | .2222714 | -1.58 | 0.114 | .1955382 | 1.191241 |

```
. svy linearized : logistic c31l agegroup_1 agegroup_2 agegroup_3
(running logistic on estimation sample)

note: agegroup_1 != 0 predicts failure perfectly
      agegroup_1 dropped and 2 obs not used

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     23185
Number of PSUs     =      1328                  Population size    =  24129.66
                                                 Design df          =      1279
                                                 F(   2,   1278)    =      9.91
                                                 Prob > F           =    0.0001
```

| c31l | Odds Ratio | Linearized Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| agegroup_2 | .1256471 | .0630585 | -4.13 | 0.000 | .0469414 | .3363174 |
| agegroup_3 | .1409233 | .0763326 | -3.62 | 0.000 | .048695 | .4078324 |

```
. svy linearized : logistic c31l c05_1 c05_3
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =     22782
Number of PSUs     =      1328                  Population size    = 23688.658
                                                 Design df          =      1279
                                                 F(   2,   1278)    =      0.10
                                                 Prob > F           =    0.9038
```

| c31l | Odds Ratio | Linearized Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| c05_1 | 1.010958 | .5792913 | 0.02 | 0.985 | .3284901 | 3.111315 |
| c05_3 | 1.23758 | .6117954 | 0.43 | 0.666 | .4692288 | 3.264087 |

## Παράρτημα 6

Αναλυτικά οι ερωτήσεις:

- c08: Κάπνισες ποτέ τσιγάρο; Αν ναι, πόσες φορές;

Ποτέ 1-2 φορές 3-5 φορές 6-9 φορές 10-19 φορές 20-39 φορές

40 και πάνω

- c09: Πόσο κάπνισες τις τελευταίες 30 ημέρες ;

Καθόλου

Λιγότερο από 1 τσιγάρο την εβδομάδα

Λιγότερο από 1 τσιγάρο την ημέρα

1-5 τσιγάρα την ημέρα

6-10 τσιγάρα την ημέρα

11-20 τσιγάρα την ημέρα

Πάνω από 20 τσιγάρα την ημέρα

- c12a: Ήπιες ποτέ κάποιο αλκοολούχο ποτό ; Αν ναι πόσες φορές

Σε όλη σου τη ζωή έως και σήμερα

Ποτέ 1-2 φορές 3-5 φορές 6-9 φορές 10-19 φορές 20-39 φορές

40 και πάνω

- c12b :Στη διάρκεια των 12 τελευταίων μηνών μέχρι και σήμερα

Ποτέ 1-2 φορές 3-5 φορές 6-9 φορές 10-19 φορές 20-39 φορές

40 και πάνω

- c12c : Στη διάρκεια των τελευταίων ημερών μέχρι και σήμερα

Ποτέ 1-2 φορές 3-5 φορές 6-9 φορές 10-19 φορές 20-39 φορές

40 και πάνω.

- c18: Θυμήσου άλλη μια φορά τις 30 τελευταίες ημέρες. Πόσες φορές ήπιες στη σειρά πέντε η περισσότερα ποτά από το ίδιο η διαφορετικά αλκοολούχα ποτά ;

  Ποτέ

Μία φορά

Δύο φορές

3-5 φορές

6-9 φορές

10 ή περισσότερες φορές

- c19a : Πόσες φορές μέθυσες από αλκοολούχα ποτά ;

Σε όλη σου τη ζωή έως και σήμερα

Ποτέ 1-2 φορές 3-5 φορές 6-9 φορές 10-19 φορές 20-39 φορές

40 και πάνω

- c19b: Στη διάρκεια των 12 τελευταίων μηνών μέχρι και σήμερα

Ποτέ 1-2 φορές 3-5 φορές 6-9 φορές 10-19 φορές 20-39 φορές

40 και πάνω

- c19c: Στη διάρκεια των τελευταίων ημερών μέχρι και σήμερα

Ποτέ 1-2 φορές 3-5 φορές 6-9 φορές 10-19 φορές 20-39 φορές

40 και πάνω

- c25a : Δοκίμασες ή πήρες ποτέ μαριχουάνα ή χασίς

Σε όλη σου τη ζωή έως και σήμερα

Ποτέ 1-2 φορές 3-5 φορές 6-9 φορές 10-19 φορές 20-39 φορές

40 και πάνω

- c25b :Στη διάρκεια των 12 τελευταίων μηνών μέχρι και σήμερα

Ποτέ 1-2 φορές 3-5 φορές 6-9 φορές 10-19 φορές 20-39 φορές

40 και πάνω

- c25c : Στη διάρκεια των 30 τελευταίων ημερών μέχρι και σήμερα

Ποτέ 1-2 φορές 3-5 φορές 6-9 φορές 10-19 φορές 20-39 φορές

40 και πάνω

- c29a : Δοκίμασες ή πήρες ποτέ έκσταση ; Σε όλη σου τη ζωή μέχρι και σήμερα

Ποτέ 1-2 φορές 3-5 φορές 6-9 φορές 10-19 φορές 20-39 φορές

40 και πάνω

- c31b c31c c31d c31e c31f c31g c31h c31i c31j c31k c31l : Έχεις χρησιμοποιήσει ποτέ κάποια από τις παρακάτω ουσίες: Ηρεμιστικά ή υπνωτικά χωρις τη σύσταση γιατρού, Αμφεταμίνες, LSD ή κάποιο άλλο παραισθησιογόνο, Κράκ, Κοκαινη, Ρελιβίνη, Ηρωίνη, Μαγικά μανιτάρια'', GHB, Αναβολικά χωρίς τη σύσταση γιατρού, Ναρκωτικά με ένεση, Κάποιο αλκοολούχο ποτό μαζι με φάρμακα για να αλλάξεις τη διάθεση σου

Ποτέ 1-2 φορές 3-5 φορές 6-9 φορές 10-19 φορές 20-39 φορές

40 και πάνω

Επαναπροσδιορισμός κατηγορικών μεταβλητών

- recode c08 (1 = 0) (2 3 4 5 6 7 = 1)
- recode c09 (1 2 3 = 0) (4 5 6 7 = 1)
- recode c12a (1 2 = 0) (3 4 5 6 7 = 1 )
- recode c12b (1 2 = 0) (3 4 5 6 7 = 1 )
- recode c12c (1 = 0) (2 3 4 5 6 7 = 1 )
- recode c18 (1 = 0) (2 3 4 5 6 = 1)
- recode c19a c19b c19c (1 = 0) (2 3 4 5 6 7 = 1)
- recode c25a c25b c25c (1 = 0) (2 3 4 5 6 7 = 1)
- recode c29a (1 = 0) (2 3 4 5 6 7 = 1)
- recode c31a c31b c31c c31d c31e c31f c31g c31h c31i c31j c31k c31l (1 = 0) (2 3 4 5 6 7 = 1)