

Πληθυσμιακές μελέτες Population Surveys

Σχεδιασμός, δειγματοληψία και μέθοδοι ανάλυσης

Γενικά χαρακτηριστικά και στόχος

Τί είναι

- Συγχρονικές μελέτες που αφορούν μεγάλες πληθυσμιακές ομάδες
 - Παράδειγμα: Γενικός πληθυσμός χωρών ή εθνοτήτων/μειονοτήτων που κατοικούν σε μια χώρα

Πού χρησιμεύουν

Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας, ο σχεδιασμός εθνικής στρατηγικής πρόληψης των χρόνιων νοσημάτων προϋποθέτει τον **προσδιορισμό του επιπολασμού των παραγόντων κινδύνου σε κάθε χώρα μέσω επιδημιολογικών μελετών σε τυχαία επιλεγμένα δείγματα γενικού πληθυσμού**

Ποιος είναι ο στόχος τους;

Στοχεύουν στην εκτίμηση της συχνότητας των χαρακτηριστικών υγείας

- Νοσημάτων
- Παραγόντων κινδύνου
- Δομών/Υπηρεσιών υγείας

Παράγοντας κινδύνου

Οποιοδήποτε χαρακτηριστικό το οποίο μπορεί να αυξήσει την πιθανότητα εμφάνισης κάποιας ασθένειας.



Κύριοι παράγοντες κινδύνου

- Αρτηριακή πίεση
- Σακχαρώδης διαβήτης
- Κάπνισμα
- Αλκοόλ
- Υψηλή χοληστερίνη
- Παχυσαρκία

- Κοινωνική διαβάθμιση
- Στρες
- Ανεργία
- Διατροφή
- Άσκηση
- Περιβαλλοντική ρύπανση
- Κοινωνικο-οικονομικοί

**ΠΟΥ: Παράγοντες κινδύνου εξηγούν το
60% των θανάτων**

Σκεπτικό επιδημιολογικών μελετών

•Είναι αδύνατον να συμπεριλάβουμε ολόκληρο τον πληθυσμό αναφοράς!

•Χρησιμοποιούμε (κατάλληλο) δείγμα ώστε...

- Να έχουμε μικρότερο κόστος
- Να χρειαστούμε λιγότερο χρόνο
- Να επικεντρωθούμε στην ορθότερη/πιο λεπτομερή συλλογή δεδομένων

Στοιχεία Σχεδιασμού

- Ορισμός πληθυσμού αναφοράς (reference/target population)
- Ορισμός πλυθησμού από τον οποίο θα προέλθει το δείγμα (sampling frame)
- Επιλογή μεθόδου δειγματοληψίας (sampling method)
- Μέγεθος δείγματος (sample size)
- Ορισμός δειγματοληπτικής μονάδας (sampling unit)

Άλλα θέματα

- Προϋπολογισμός
- Ethics-Confidentiality
- Επιλογή μεταβλητών-δεδομένων που θα συλλεχθούν
 - Ερωτηματολόγιο-Εξετάσεις υγείας (SOPs)
- Στατιστική μεθοδολογία (λαμβάνει υπόψιν και στοιχεία της δειγματοληψίας)
- Οργάνωση έρευνας πεδίου
- Συλλογή δεδομένων
 - - Βάση δεδομένων και καταγραφή
- Ανάλυση δεδομένων

Θέματα Δεοντολογίας

- Έγκριση από την Επιτροπή Δεοντολογίας Ιατρικής Σχολής ΕΚΠΑ
- Αρχή προστασίας προσωπικών δεδομένων
- Ενυπόγραφη συναίνεση
 - Συμπλήρωση ερωτηματολογίων
 - Εξετάσεις αίματος
 - Αποθήκευση υπολειπόμενου δείγματος

Σχεδιασμός

1. Πληθυσμός
2. Δειγματοληπτικό πλαίσιο
3. Δειγματοληπτική μονάδα
4. Μέθοδος δειγματοληψίας
5. Μέγεθος δείγματος

Σχεδιασμός (1)

•Καθορισμός του **πληθυσμού αναφοράς** (target population)

– Είναι ο πληθυσμός στον οποίο θέλουμε να γενικεύονται, δηλαδή να ισχύουν, τα ευρήματά μας

•Είναι σημαντικό να καθορισθεί με ακρίβεια, γιατί καθορίζει την επιλογή του δείγματος

– Παράδειγμα: Μόνιμος πληθυσμός χώρας- Άτομα ηλικίας >18 ετών, που διαμένουν στην χώρα >5έτη

Σχεδιασμός (2)

.Δειγματοληπτικό πλαίσιο (sampling frame)

- Είναι το τμήμα του πληθυσμού από το οποίο θα αντλήσουμε το δείγμα

.Αποφασίζεται βάσει:

- Ευκολίας πρόσβασης, εγγύτητας, διαθεσιμότητας και άλλων πρακτικών λόγων
- Καταλληλότητας- Θα πρέπει να οδηγεί σε γενικεύσιμα αποτελέσματα για την μελέτη (external validity)

Σχεδιασμός (3)

.Καθορισμός δειγματοληπτικής μονάδας

(sampling unit)

- Μπορεί να είναι άτομο, νοικοκυριό, οικοδομικό τετράγωνο κλπ

Σχεδιασμός (3)

**.Πρωταρχική δειγματοληπτικής μονάδας
(primary sampling unit, PSU)**

– Είναι η μονάδα που επιλέγεται από το
δειγματοληπτικό σχέδιο και μπορεί να είναι
άτομο, νοικοκυριό, οικοδομικό τετράγωνο κλπ

**.Υπόκειται σε κριτήρια επιλεξιμότητας/εξαίρεσης
(inclusion/exclusion criteria)**

Σχεδιασμός (4)

- **Δειγματοληπτική Μέθοδος**
- **Επιλογή δείγματος** με συγκεκριμένη μέθοδο ώστε να είναι κατάλληλο
 - Τα αποτελέσματά που θα προκύψουν να μπορούν να γενικευθούν στον πληθυσμό αναφοράς (sampling method)

Σχεδιασμός (5)

.Υπολογισμός μεγέθους δείγματος ώστε να μπορούμε να απαντήσουμε στο ερευνητικό ερώτημα με ακρίβεια (sample size calculation)

Αντιπροσωπευτικότητα

- Αντιπροσωπευτικό **δείγμα** ενός πληθυσμού αναφοράς είναι ένα δείγμα το οποίο **αντανakλά επαρκώς τα χαρακτηριστικά του πληθυσμού αναφοράς**
- Μεγαλύτερα δείγματα τείνουν να είναι αντιπροσωπευτικά (εφόσον δεν υπάρχει μεροληψία επιλογής)

Τυχαιότητα

Τυχαίο δείγμα είναι μια ομάδα ατόμων του πληθυσμού αναφοράς επιλεγμένη με τυχαίο (μή συστηματικό) τρόπο.

• Αυτό σημαίνει ότι κάθε άτομο του πληθυσμού έχει την ίδια πιθανότητα να επιλεγεί

• Τα τυχαία δείγματα δεν είναι απαραίτητα και αντιπροσωπευτικά

Μέθοδοι Δειγματοληψίας

Είδη Δειγματοληπτικών Μεθόδων

➤ Δειγματοληψία μή βασιζόμενη σε πιθανότητα

(Non-probability sampling)

➤ Δειγματοληψία βάσει πιθανότητας (Probability

sampling)

Δειγματοληψία βάσει πιθανότητας (Probability sampling)

Κάθε δειγματοληπτική μονάδα έχει συγκεκριμένη μή μηδενική πιθανότητα να επιλεγεί στο δείγμα της μελέτης

➤ Είδη:

- Απλή τυχαία δειγματοληψία (Simple random sampling)
- Συστηματική τυχαία δειγματοληψία (Systematic random sampling)
- Στρωματοποιημένη τυχαία δειγματοληψία (Stratified random sampling)
- Κατά συστάδες δειγματοληψία (Cluster sampling)

Δειγματοληψία βάσει πιθανότητας (Probability sampling)

–Επίσης μπορεί να γίνει:

•Πολυσταδιακή τυχαία δειγματοληψία (Multi-stage random sampling)

•Ή δειγματοληψία σε πολλαπλές φάσεις (Multi-phase sampling)

Απλή Τυχαία Δειγματοληψία

- Κάθε άτομο/μονάδα του πληθυσμού έχει την ίδια πιθανότητα να επιλεγεί στο δείγμα
- Μπορεί να γίνει με χρήση λίστας τυχαίων αριθμών και αντιστοίχησης με τα άτομα του πληθυσμού ή με κάποιο αυτοματοποιημένο τρόπο (για παράδειγμα επιλογή από αρχείο registry με χρήση αλγορίθμου)
- Προϋποθέτει καταγραφή των ατόμων του πληθυσμού ή τουλάχιστον γνώση του αριθμού των ατόμων του πληθυσμού

Συστηματική Τυχαία Δειγματοληψία

- Κατά τη **συστηματική** δειγματοληψία επιλέγεται τυχαία το πρώτο άτομο (για παράδειγμα σε λίστα του πληθυσμού) και κατόπιν κάθε n -οστό άτομο, μέχρι να συμπληρωθεί το δείγμα.
 - Ιδιαίτερα χρήσιμη μέθοδος για μεγάλους πληθυσμούς.
- Παράδειγμα: Αν επιλέγουμε άτομα από τον τηλεφωνικό κατάλογο, επιλέγεται τυχαία ένας αριθμός από το 1 έως το 20 και έκτοτε επιλέγονται τα άτομα ανά 50 θέσεις (Δηλαδή ο 70κοστός, 12οστός κ.ο.κ).
- Δειγματοληπτικό κλάσμα (sampling fraction) είναι το πηλίκο του μεγέθους δείγματος προς το μέγεθος του πληθυσμού

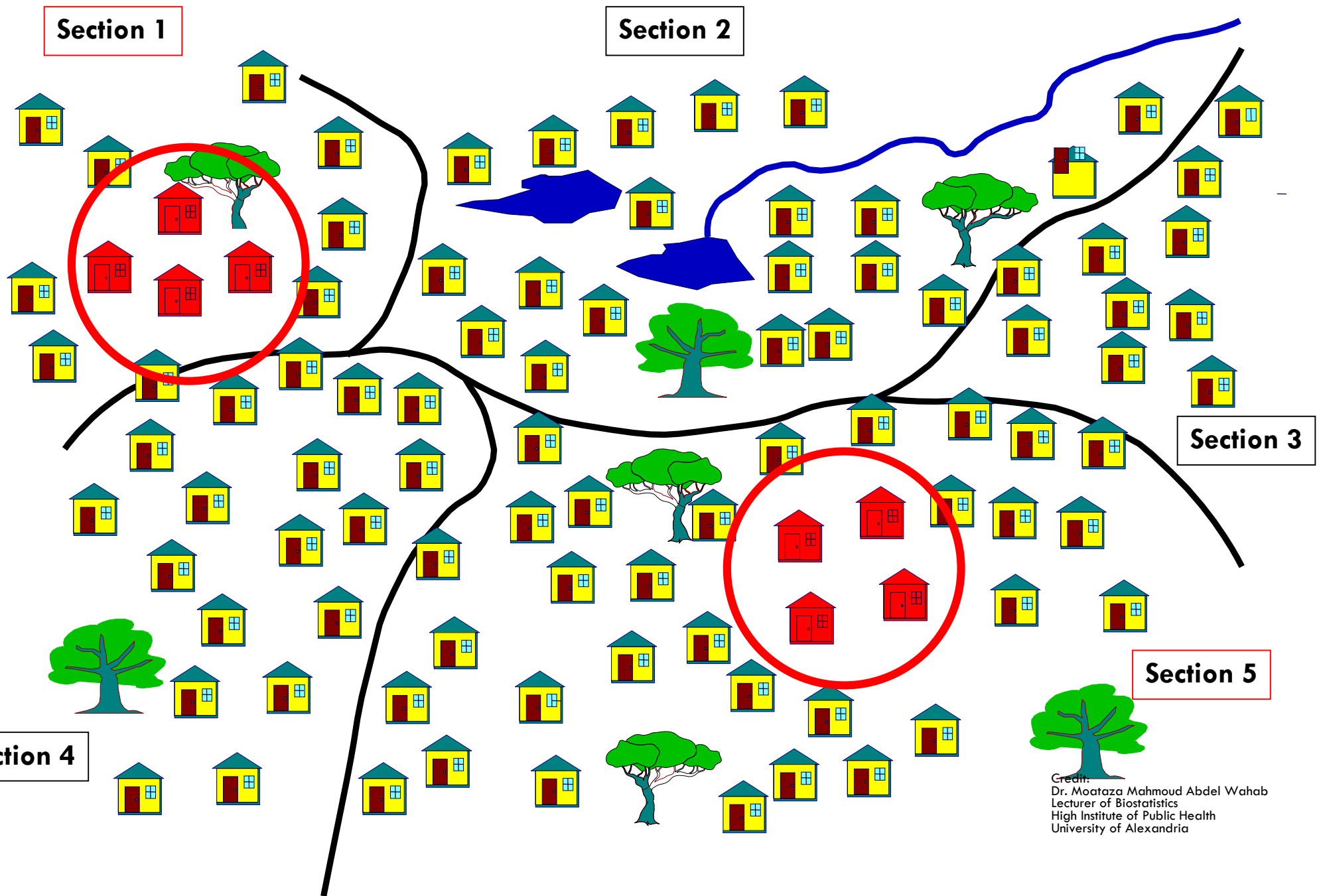
Στρωματοποιημένη Δειγματοληψία

- Η δειγματοληψία πραγματοποιείται ανά στρώματα ή υποπληθυσμούς, είτε ανάλογα με το μέγεθός τους, είτε σε ίσους αριθμούς.
 - Ο σκοπός είναι να εκπροσωπούνται επαρκώς αυτές οι υποομάδες του πληθυσμού στο δείγμα
- Για παράδειγμα, αν τα στρώματα αφορούν σε πόλεις της Ελλάδας, τότε η στρωματοποιημένη δειγματοληψία εξασφαλίζει ότι θα έχουμε άτομα από όλες τις πόλεις.

Κατά συστάδες δειγματοληψία

- Συστάδα είναι μια ομάδα ατόμων/δειγματοληπτικών μονάδων που βρίσκονται κοντά, για παράδειγμα στην ίδια περιοχή ή γειτονιά
- Επιλέγουμε τυχαίο δείγμα στρωμάτων και κατόπιν συμπεριλαμβάνουμε στο δείγμα όλα τα άτομα/μονάδες των στρωμάτων αυτών
- Για παράδειγμα, αν τα στρώματα ήταν σχολεία μιας πόλης, θα επιλέγαμε τυχαία κάποια σχολεία και θα συμπεριλάβαναμε στη μελέτη όλους τους μαθητές των σχολείων αυτών

Cluster Samples of Households



Δειγματοληψία NHANES

National Center for Health Statistics

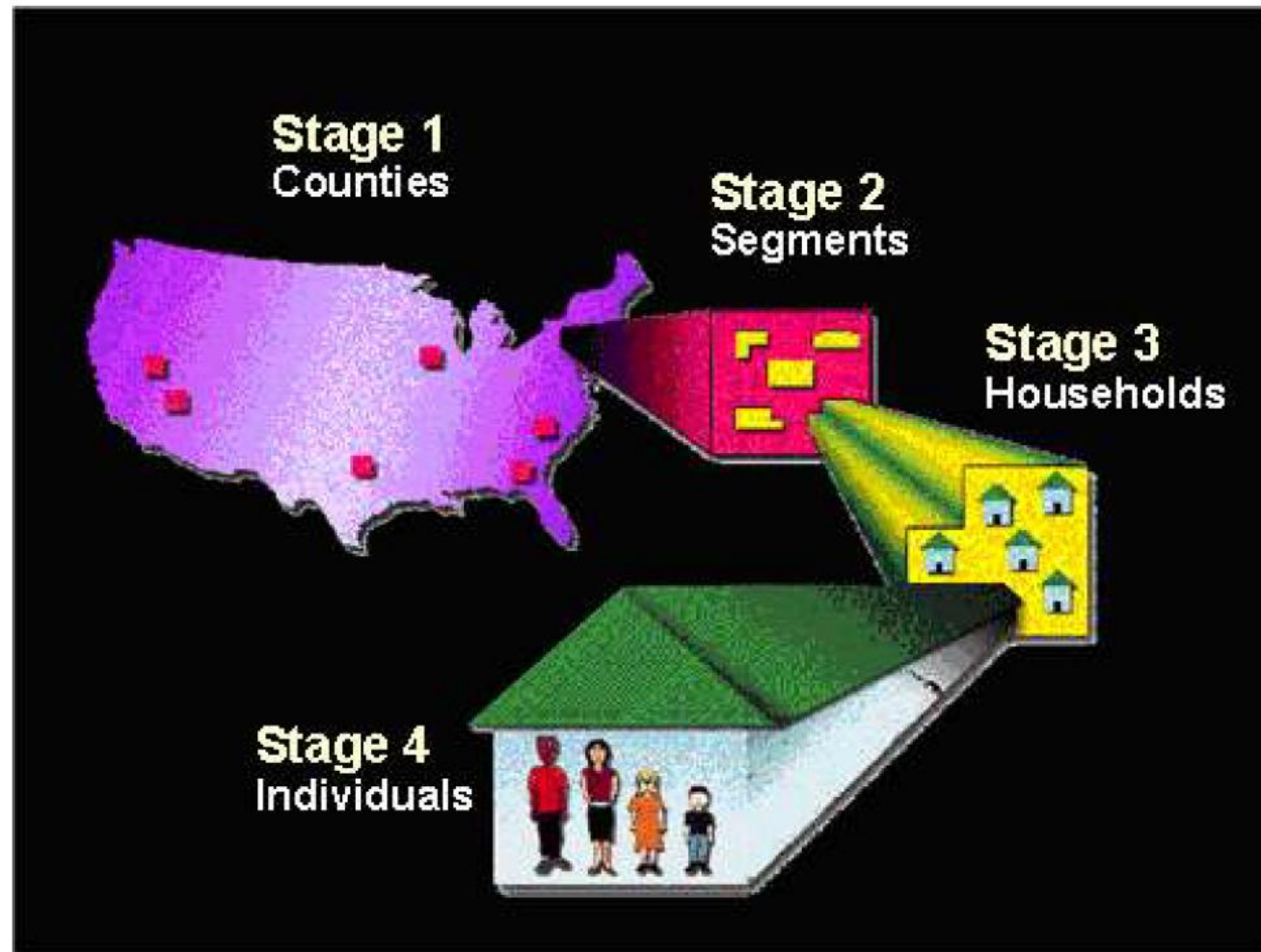


National Health and Nutrition Examination Survey

Module 2: Sample Design

The NHANES samples are not simple random samples. Rather, a complex, multistage, probability sampling design is used to select participants representative of the civilian, non-institutionalized US population. Oversampling of certain population subgroups is also done to increase the reliability and precision of health status indicator estimates for these particular subgroups. Researchers need to take this into account in their analyses by appropriately specifying the sampling design parameters. This can be done using any statistical software that can analyze complex survey designs, such as SAS Survey procedures, SUDAAN, Stata, or R. This module provides an overview of the sample design parameters in NHANES.

NHANES: Δειγματοληπτικό σχέδιο 4 σταδίων



Four Stages of NHANES Sampling Procedure



National Center for Health Statistics

CDC > NCHS

National Health and Nutrition Examination Survey

About NHANES +

What's New +

Webinar

Questionnaires, Datasets, and
Related Documentation +

Survey Participants +

Biospecimen Program +

New Content and Proposal
Guidelines



National Health and Nutrition Examination Survey

Survey Participants



If you were selected,
learn more about
participating

Survey Data and Documentation



Access data,
documentation &
response rates

Publications and Products



View health and
nutrition reports & CDC
Growth Charts

Data Analysis Tutorials



Review step-by-step
guidance on using
NHANES data



Analytic Guidance and 2024 NHANES Survey Content Propo...
NHANES Webinar



Copy link

What's New

NHANES

The National Health Survey Act of 1956 provided for a continuing survey and special studies “ . . . **to secure accurate and current statistical information on the amount, distribution, and effects of illness and disability in the United States . . .** ” and to provide methods and processes to secure this statistical information (1).

Vital and Health Statistics, Series 1, Number 56

National Health and Nutrition Examination Survey: Plan and Operations, 1999–2010

Program and Collection Procedures

Δειγματοληπτικό σχέδιο NHANES

Στάδιο 1

Στάδιο 1: Επιλέχθηκαν οι πρωταρχικές δειγματοληπτικές μονάδες (Primary sampling units, PSUs). Αυτές ήταν πολιτείες ή σε ελάχιστες περιπτώσεις ομάδες πολιτειών.

Οι PSUs επιλέχθηκαν με πιθανότητα ανάλογη ενός μέτρου του μεγέθους τους (PPS).

Πολιτείες με μεγαλύτερο πληθυσμό ήταν πιθανότερο να επιλεγούν συγκρητικά με άλλες με μικρότερο πλυθησμό.

Δειγματοληπτικό σχέδιο NHANES

Στάδιο 1

Για τη NHANES χρησιμοποιήθηκε ένας σταθμισμένος μέσος του πληθυσμού:

Βάρη υπολογισμένα ώστε να δίνουν υψηλότερη πιθανότητα επιλογής σε πολιτείες με μεγαλύτερα ποσοστά ατόμων που βάσει σχεδιασμού είχαν επιλεγεί να υπερ-επιλεχθούν (oversampling)

Δειγματοληπτικό σχέδιο NHANES

Στάδια 2-4

2ο Στάδιο: Κάθε επιλεγμένη πολιτεία χωρίστηκε σε οικοδομικά τετράγωνα και εφαρμόστηκε δειγματοληψία με πιθανότητα ανάλογη μέτρου του μεγέθους τους

3ο Στάδιο: Σε κάθε ένα από τα επιλεγμένα οικοδομικά τετράγωνα επιλέχθηκαν νοικοκυριά από λίστα, επιτρέποντας και πάλι oversampling

4ο Στάδιο: Από κάθε επιλεγμένο νοικοκυριο, επιλέχθηκαν 2 άτομα με τυχαία δειγματοληψία

Oversampling

- Η NHANES σχεδιάστηκε ώστε να παίρνει μεγαλύτερα (αναλογικά) δείγματα από κάποιες συγκεκριμένες υποομάδες που έχουν ιδιαίτερο ενδιαφέρον από πλευράς δημόσιας υγείας
- Στόχος του oversampling να έχουμε ένα αρκετά μεγάλο δείγμα σε μικρές ομάδες του πληθυσμού, που θα επιτρέψει αξιόπιστες εκτιμήσεις σε αυτούς τους πληθυσμούς.

Δειγματοληπτικά βάρη

- Χρειαζόμαστε δειγματοληπτικά βάρη, ώστε να λάβουμε υπόψη μας την άνιση πιθανότητα επιλογής
- Επιτρέπουν οι εκτιμήσεις που βασίζονται σε δείγματα μή ανάλογα του πραγματικού πληθυσμού να χρησιμοποιούνται και να δίνουν αμερόληπτες εκτιμήσεις για τον πληθυσμό.

Δειγματοληπτικά βάρη

• Αν όλα τα άτομα ενός πληθυσμού 1000 ατόμων επιλέγονται με την ίδια πιθανότητα σε δείγμα 100 ατόμων, τότε κάθε ένας από τους συμμετέχοντες στο δείγμα αντιπροσωπεύει 10 άτομα του πληθυσμού με τα ίδια χαρακτηριστικά:

- Αν ο πληθυσμός αποτελείται από γυναίκες 550 γυναίκες και 450 άνδρες και το δείγμα από 55 γυναίκες και 45 άνδρες, τότε κάθε γυναίκα εκπροσωπεί 10 γυναίκες και κάθε άνδρας 10 άνδρες
- Σε αυτή την περίπτωση δεν υπάρχει λόγος να γίνει στάθμιση του δείγματος

Δειγματοληπτικά βάρη

- Έστω ότι μια συγκεκριμένη κοινότητα (π.χ θρησκευτική) της Αμερικής έχει πληθυσμό 500 άτομα
 - Τους εισάγουμε όλους στη μελέτη ώστε να μπορούμε να εκτιμήσουμε με καλή ακρίβεια τα χαρακτηριστικά τους
 - Πιθανότητα επιλογής: 100%
 - Εκπροσωπούνται όλοι

Δειγματοληπτικά βάρη

- Ο γενικός πληθυσμός της πολιτείας Α είναι 5000000.
- Για την ίδια μελέτη, επιλέγουμε από αυτούς 500 άτομα
 - Πιθανότητα επιλογής: $500/5000000=0.0001$
 - Κάθε ένας από τους 500 συμμετέχοντες, εκπροσωπεί 10000 κατοίκους της πολιτείας του
 - Θα πρέπει η «συνεισφορά» του καθενός να πολλαπλασιαστεί επί 1000

Παράδειγμα

Έστω 3 άτομα της μελέτης από διαφορετικές περιοχές, με βάρη 1, 2 και 5

Άτομο	Τιμή χαρακτηριστικού	Πιθανότητα επιλογής	Βάρος	Συνεισφορά
1	10	1.00	1	10
2	20	0.50	2	40
3	30	0.20	5	150
3	60/3=20		8	200 / 8 = 25

Δειγματοληπτικά βάρη

- Εθνική μελέτη χώρας 5 000 000 κατοίκων
- Θέλουμε να έχουμε ένα αρκετά μεγάλο δείγμα κατοίκων μιας απομακρυσμένης περιοχής με 1000 κατοίκους
- Θα επιλέξουμε συμμετέχοντες από αυτή την περιοχή με μεγαλύτερη πιθανότητα από αυτή θα αντιστοιχούσε σε τυχαίο δείγμα (π.χ. 0.5).
- Αν μια πόλη έχει πλυθυσμό 1 000 000 τότε μας αρκεί πιθανότητα επιλογής συμμετεχόντων από την πόλη ίση με 0.0005.

Δειγματοληπτικά βάρη

- Αυτό σημαίνει ότι κάθε συμμετέχοντας από την απομακρυσμένη περιοχή αντιπροσωπεύει 2 άτομα ($sw=2$) ενώ κάθε συμμετέχοντας από την πόλη 2000 άτομα ($sw=2000$)
- Επομένως για να αποκατασταθεί η δομή του αρχικού πληθυσμού, τα δεδομένα κάθε ατόμου της απομακρυσμενης περιοχής θα διπλασιαστούν, ενώ τα δεδομένα κάθε ατόμου της πόλης θα πολλαπλασιαστούν επί 2000

The weighting effect

Τα weights αυξάνουν τη μεταβλητότητα των εκτιμήσεων, ειδικά αν έχουν μεγάλη διασπορά:

weighting effect: $DEFF_w = 1 + CV^2(w_i)$

CV ο συντελεστής μεταβλητότητας των weights

Σύνθετες Δειγματοληπτικές μέθοδοι

Πολυσταδιακή δειγματοληψία

Γεωγραφικά διαμερίσματα



Πόλεις



Νοικοκυριά



Άτομα

□ Δειγματοληψία σε πολλαπλές φάσεις

Πληθυσμός



Δείγμα 1

→ 1ος έλεγχος



Δείγμα 2

→ 2ος έλεγχος

Δειγματοληψία σε πολλαπλές φάσεις

• Εφαρμόζεται όταν:

• Δεν είναι πρακτικό ή οικονομικό να μετρήσεις ακριβώς όλα τα ενδιαφέροντα χαρακτηριστικά σε όλες τις μονάδες του δείγματος.

• Συνήθως:

• Στην 1η φάση, παίρνεις ένα μεγάλο δείγμα και συλλέγεις φθηνές ή απλές μεταβλητές.

• Στη 2η φάση, παίρνεις υποσύνολο αυτού του δείγματος και συλλέγεις πιο ακριβές ή δύσκολες μετρήσεις.

• Η 2η φάση συνήθως εξαρτάται από τα αποτελέσματα της 1ης, π.χ. επιλέγεις υποσύνολο από ομάδες ενδιαφέροντος.

Παράδειγμα-1: Επιπολασμός του σακχαρώδους διαβήτη στον ενήλικο πληθυσμό.

- *1η φάση (screening phase):* Μεγάλο δείγμα (π.χ. 10.000 άτομα).
- Συλλέγονται απλές, φθηνές μεταβλητές: ηλικία, φύλο, BMI, αυτοαναφερόμενος διαβήτης, οικογενειακό ιστορικό, κ.λπ.
- *2η φάση (diagnostic phase):* Επιλέγεται υποσύνολο (π.χ. 1.000 άτομα) για αιματολογικές εξετάσεις (π.χ. γλυκόζη νηστείας, HbA1c).
- Το υποσύνολο μπορεί να είναι: *τυχαίο δείγμα*, ή *εμπλουτισμένο δείγμα* (π.χ. υπερεκπροσώπηση όσων είχαν υψηλό BMI στην 1η φάση).
- *Στόχος:* Συνδυάζοντας δεδομένα και από τις δύο φάσεις, μπορούμε να εκτιμήσουμε με ακρίβεια τον επιπολασμό του διαβήτη σε όλο τον πληθυσμό, διορθώνοντας για το σχέδιο επιλογής της 2ης φάσης (με βάρη).

Παράδειγμα-2: Επιδημιολογική μελέτη για καρκίνο

1η φάση → στέλνεις ερωτηματολόγιο σε 20.000 άτομα σχετικά με παράγοντες κινδύνου.

2η φάση → καλείς υποσύνολο 2.000 ατόμων (π.χ. εκείνους με οικογενειακό ιστορικό ή τυχαίο δείγμα χωρίς ιστορικό) για γενετικές εξετάσεις.

Η ανάλυση πρέπει να λάβει υπόψη ότι η 2η φάση δεν είναι τυχαίο υποσύνολο αλλά *επιλεγμένο βάσει χαρακτηριστικών της 1ης φάσης*.

	Multistage	Multiphase
Τι μεταβάλλεται σε κάθε βήμα	Η μονάδα δειγματοληψίας (π.χ. πρώτα επιλέγεις σχολεία, μετά μαθητές)	Η ποσότητα ή το είδος των δεδομένων που συλλέγεις από τις ίδιες ή υποσύνολο των ίδιων μονάδων
Στόχος	Να φτάσεις από μεγάλες σε μικρές μονάδες	Βελτίωση αποτελεσματικότητας ή μειώνοντας το κόστος
Κύριο παράδειγμα	Δείγμα: Νομοί → Δήμοι → Νοικοκυριά → Άτομα	1η φάση: βασικά στοιχεία από όλους → 2η φάση: λεπτομερή μέτρηση σε υποσύνολο

Πολυσταδιακή Δειγματοληψία

The Design Effect (DEFF)

Το *design effect* μετρά πόσο “λιγότερο αποδοτικό” είναι ένα σύνθετο δείγμα (με στρώματα, συστάδες, άνισες πιθανότητες επιλογής κ.λπ.) σε σχέση με ένα απλό τυχαίο δείγμα ίδιου μεγέθους.

Ορισμός:

$$DEFF = \frac{Var_{design}(\hat{\theta})}{Var_{SRS}(\hat{\theta})}$$

SRS= Simple Random Sampling

Design: Stratified, Cluster, etc

The Design Effect (DEFF)

Ορισμός: $DEFF = \frac{Var_{design}(\hat{\theta})}{Var_{SRS}(\hat{\theta})}$

Αριθμητής: $Var_{design}(\hat{\theta})$ Πραγματική διακύμανση της εκτιμήτριας $\hat{\theta}$ υπό το συγκεκριμένο σχέδιο δειγματοληψίας που χρησιμοποιήσαμε (design-based variance).

$$Var_{design}(\hat{\theta}) = Var(\hat{\theta} | \text{sampling design})$$

The Design Effect (DEFF)

Ορισμός: $DEFF = \frac{Var_{design}(\hat{\theta})}{Var_{SRS}(\hat{\theta})}$

Παρονομαστής: $Var_{SRS}(\hat{\theta})$ Υποθετική διακύμανση της εκτιμήτριας $\hat{\theta}$, αν είχαμε πάρει ένα απλό τυχαίο δείγμα (SRS) ίδιου μεγέθους από τον ίδιο πληθυσμό.

$$Var_{SRS}(\hat{\theta}) = Var(\hat{\theta} | \text{Simple Random Sampling})$$

The Design Effect (DEFF)

Αν $DEFF=1$: το σχέδιο είναι εξίσου αποδοτικό με SRS.

Αν $DEFF>1$: μεγαλύτερη διακύμανση λόγω σχεδίασης (π.χ. ομαδοποίηση).

Αν $DEFF<1$: πιο αποδοτικό σχέδιο (π.χ. στρωματοποίηση).

Παράδειγμα:

Σε έρευνα κατά συστάδες, οι μονάδες μέσα στην ίδια συστάδα είναι συχνά πιο όμοιες $\rightarrow$ *intraclass correlation* (ρ) $\rightarrow$

$DEFF \approx 1 + (m-1)\rho$ όπου m το μέσο μέγεθος συστάδας.

Άρα, όσο μεγαλύτερη η ομοιότητα εντός συστάδας, τόσο αυξάνεται το DEFF.

Sampling Variance vs. Model Variance

- **Sampling variance:**

Η διακύμανση που προκύπτει **λόγω του σχεδίου δειγματοληψίας**

- Αντικατοπτρίζει την αβεβαιότητα επειδή επιλέχθηκε ένα συγκεκριμένο δείγμα από τον πληθυσμό.
- Υπολογίζεται λαμβάνοντας υπόψη weights, clustering και stratification.

- **Model variance:**

Η διακύμανση που προκύπτει **από το στατιστικό μοντέλο**

- Αφορά την αβεβαιότητα των εκτιμήσεων που προκύπτει από το μοντέλο, δεδομένου του δείγματος.

Sampling Variance vs. Model Variance

- **Κύρια διαφορά:**
- *Sampling variance* → στοχαστικότητα προέρχεται από τη δειγματοληψία.
- *Model variance* → στοχαστικότητα προέρχεται από τις υποθέσεις του μοντέλου.
- Στις αναλύσεις δεδομένων surveys, πρέπει να συνδυάζουμε και τα δύο
- Η εκτίμηση της διακύμανσης πρέπει να είναι design-based, όχι απλώς model-based.

Sampling Variance

Δειγματοληπτική διακύμανση: Αβεβαιότητα που προκύπτει επειδή επιλέγουμε ένα από τα πολλά πιθανά δείγματα που θα μπορούσαν να προκύψουν από τον πληθυσμό σύμφωνα με το δειγματοληπτικό σχέδιο.

$$Var_{design}(\hat{\theta}) = E_{design} \left[(\hat{\theta} - E_{design}[\hat{\theta}])^2 \right]$$

Πληθυσμός είναι σταθερός, όμως αν επαναλαμβάναμε την ίδια δειγματοληψία θα επιλέγαμε (ενδεχομένως) άλλα άτομα του πληθυσμού στο δείγμα της μελέτης.

Η **sampling variance** δείχνει **πόσο θα μεταβάλλεται η εκτιμήτρια** αν επαναλαμβάναμε τη διαδικασία δειγματοληψίας πολλές φορές με το ίδιο σχέδιο.

Απλό Τυχαίο Δείγμα (SRS)

Αν από πληθυσμό μεγέθους N παίρνουμε δείγμα μεγέθους n (απλή τυχαία δειγματοληψία χωρίς επανάθεση):

$$Var_{SRS}(\bar{y}) = \frac{S^2}{n} \left(1 - \frac{n}{N}\right)$$

όπου: $S^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{Y})^2$ η διακύμανση του πληθυσμού

και $(1 - n/N)$ η finite population correction

Αν N είναι πολύ μεγάλο σε σχέση με το n , ο δεύτερος όρος προσεγγίζει το 1.

Σύνθετοι σχεδιασμοί

Σε σύνθετα σχέδια (π.χ. με στρωματοποίηση, συστάδες, άνισες πιθανότητες επιλογής), η **sampling variance** εξαρτάται από τη δομή του σχεδίου.

Γενικά:
$$Var_{design}(\hat{\theta}) = \sum_i \sum_j (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$$

Όπου π_i η πιθανότητα επιλογής του ατόμου i και π_{ij} η πιθανότητα επιλογής των ατόμων i και j μαζί

Εκτίμηση δειγματοληπτικής διασποράς

- Στην πράξη, η sampling variance **εκτιμάται** (δεν υπολογίζεται αναλυτικά) με:
- **Taylor linearization, ή Replication methods** (Jackknife, Balanced Repeated Replication, Bootstrap)
 - Προσομοιώνουν την επίδραση της τυχαιότητας του δειγματοληπτικού σχεδίου.

Δειγματοληψία μή
βασιζόμενη σε
πιθανότητα

Δειγματοληψία μή βασιζόμενη σε πιθανότητα

➤ **Δείγμα ευκολίας (Convenience sampling):**

➤ Επιλογή ατόμων/δειγματοληπτικών μονάδων που είναι εύκολα προσβάσιμα (μη καταγεγραμμένοι πληθυσμοί, π.χ Ρομά στην μελέτη EMENO)

➤ **Δειγματοληψία οδηγούμενη από τους συμμετέχοντες (Respondent driven sampling):**

➤ Επιλογή των πλέον προσβάσιμων ατόμων ενός πληθυσμού και των ατόμων του κοινωνικού κύκλου τους κ.ο.κ (XEN στην μελέτη Αριστοτέλης)

Δειγματοληψία μή βασιζόμενη σε πιθανότητα

➤ **Επιλογή ατόμων από τον ερευνητή, συμμετοχή όσων θεωρεί ότι πρέπει να συμμετέχουν (purposeful sampling):**

➤ Επιλέγονται άτομα που θα συνεισφέρουν περισσότερες πληροφορίες.

➤ **Επιλογή ατόμων με συγκεκριμένα χαρακτηριστικά ή ιδιότητες (Quota sampling):**

➤ Είναι μέθοδος αντίστοιχη της στρωματοποιημένης δειγματοληψίας, αλλά εδώ τα “στρώματα” είναι φτιαγμένα από τον ερευνητή

➤ Ορίζονται εκ των προτέρων ποσοτώσεις (quotas) για συγκεκριμένες υποομάδες του πληθυσμού (π.χ. φύλο, ηλικία, εκπαίδευση) έτσι ώστε το δείγμα να μοιάζει αναλογικά με τον πληθυσμό σε αυτές τις κατηγορίες.

Χαρακτηριστικό	Quota Sampling	Purposeful Sampling
Τύπος έρευνας	Ποσοτική	Ποιοτική
Τρόπος επιλογής	Με βάση προκαθορισμένες ποσοστώσεις (π.χ. φύλο, ηλικία)	Με βάση σκοπιμότητα/κριτήρια γνώσης ή εμπειρίας
Κριτήριο επιλογής	Αντιπροσωπευτικότητα σε επιλεγμένα χαρακτηριστικά	Συνάφεια και ποιότητα πληροφορίας
Τυχαιότητα	Όχι (μη τυχαίο)	Όχι (μη τυχαίο, αλλά σκόπιμο)
Στόχος	Να μοιάζει το δείγμα με τον πληθυσμό σε κάποια χαρακτηριστικά	Να εμβαθύνει σε εμπειρίες ή αντιλήψεις
Συνηθισμένο πεδίο χρήσης	Έρευνες αγοράς, δημοσκοπήσεις	Ποιοτική έρευνα, μελέτες περίπτωσης

Δειγματοληψία μή βασιζόμενη σε πιθανότητα

- Τα είδη δειγματοληψίας μή βασιζόμενης σε πιθανότητα θεωρούνται γενικά μεροληπτικές μέθοδοι επιλογής δείγματος,
- Σε κάποιες περιπτώσεις δεν είναι δυνατή άλλου τύπου δειγματοληψία
- Τα quota και purposeful sampling χρησιμοποιούνται κυρίως στην έρευνα αγοράς και είναι μή κατάλληλα για επιδημιολογικές μελέτες
- Τα convenience sampling και RDS με κατάλληλες μεθόδους ανάλυσης μπορούν να οδηγήσουν σε αμερόληπτα αποτελέσματα



ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΙΑΤΡΙΚΗ ΣΧΟΛΗ

ΕΡΓΑΣΤΗΡΙΟ ΥΓΙΕΙΝΗΣ ΕΠΙΔΗΜΙΟΛΟΓΙΑΣ ΚΑΙ ΙΑΤΡΙΚΗΣ ΣΤΑΤΙΣΤΙΚΗΣ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΡΟΜΑ ΚΑΙ ΥΓΕΙΑ ΣΤΗΝ ΕΛΛΑΔΑ:

**ΧΡΗΣΗ ΚΑΙ ΠΡΟΣΒΑΣΗ ΣΤΙΣ ΥΠΗΡΕΣΙΕΣ ΥΓΕΙΑΣ, ΧΡΟΝΙΑ ΚΑΙ ΑΙΜΑΤΟΓΕΝΩΣ
ΜΕΤΑΔΙΔΟΜΕΝΑ ΝΟΣΗΜΑΤΑ (ΙΟΓΕΝΕΙΣ ΗΠΑΤΙΤΙΔΕΣ Β, C ΚΑΙ HIV ΛΟΙΜΩΞΗ)**

Roma Health in Greece:

**Use of and access to health services, chronic and blood transmitted diseases (viral Hepatitis B, C and
HIV infection)**

**ΙΩΑΝΝΑ ΠΕΤΡΑΚΗ
ΔΙΑΙΤΟΛΟΓΟΣ - ΔΙΑΤΡΟΦΟΛΟΓΟΣ**

Δείγμα Ευκολίας

- Πληθυσμός Ρομά: Μή καταγεγραμμένος και δύσκολα προσβάσιμος
- Μή δυνατή η εφαρμογή probability sampling
- Χρησιμοποιήθηκε δείγμα ευκολίας
- Κριτήρια:
 - Ικανοποιητική γεωγραφική διασπορά,
 - Από διαφορετικές κατηγορίες οικισμού
 - Αριθμητικά επαρκές σε κάθε οικισμό

Δειγματοληψία

1ο στάδιο: Συλλογή πληροφοριών σχετικά με τον αριθμό των οικισμών και τον αριθμό των ατόμων που διαμένουν σε αυτούς, καθώς και τη γενική κατάσταση των οικισμών.

2ο στάδιο Επιλογή αντιπροσωπευτικών οικισμών από τέσσερις περιφέρειες της Ελλάδας

3ο στάδιο Δείγμα των 540 ατόμων μη αναλογικό του πληθυσμού των περιφερειών, ώστε υπάρχει ικανός αριθμός για ανά περιοχή συγκρίσεις, να επιλεγούν ισόποσα δείγματα (135 άτομα) σε κάθε μια από τις τέσσερις περιφέρειες

Πιθανά προβλήματα

Ποια πιστεύετε ότι είναι τα προβλήματα σε τέτοιου είδους μελέτες;

Πιθανά προβλήματα

1. Έλλειψη αντιπροσωπευτικότητας (representativeness)

Το δείγμα δεν αντικατοπτρίζει τον πληθυσμό στον οποίο θέλουμε να γενικεύσουμε τα αποτελέσματα.

2. Μεροληψία επιλογής (selection bias)

Ο τρόπος επιλογής εισάγει συστηματική προκατάληψη. Οι συμμετέχοντες που είναι εύκολα διαθέσιμοι μπορεί να έχουν κοινά χαρακτηριστικά (π.χ. ηλικία, μόρφωση, κοινωνική τάξη) που επηρεάζουν τις απαντήσεις τους.

3. Περιορισμένη γενικευσιμότητα (limited generalizability)

Τα αποτελέσματα ισχύουν μόνο για το συγκεκριμένο δείγμα και όχι για τον ευρύτερο πληθυσμό. Συχνά, οι μελέτες αυτές χαρακτηρίζονται ως πιλοτικές ή διερευνητικές.

Πιθανά προβλήματα

4. Κίνδυνος για χαμηλή εξωτερική εγκυρότητα (external validity)

Η δυνατότητα να εφαρμοστούν τα αποτελέσματα σε άλλες ομάδες ή συνθήκες είναι χαμηλή.

5. Πιθανότητα αυτοεπιλογής (self-selection bias)

Οι συμμετέχοντες μπορεί να επιλέγουν να συμμετάσχουν επειδή έχουν κάποιο ιδιαίτερο ενδιαφέρον για το θέμα ή συγκεκριμένα χαρακτηριστικά, κάτι που αλλοιώνει τα αποτελέσματα.

6. Δυσκολία σύγκρισης με άλλες μελέτες

Αν άλλες μελέτες χρησιμοποιούν άλλες μεθόδους επιλογής δείγματος, είναι δύσκολο να συγκριθούν τα αποτελέσματα.

Respondent Driven Sampling

- Χρήσιμο για μελέτες “κρυφών πληθυσμών”
- Προσεγγίζονται αρχικά άτομα του πληθυσμού εύκολα προσβάσιμα
- Στη συνέχεια, συστήνουν μέλη του κύκλου τους για να συμμετέχουν στη μελέτη
 - Ποιο είναι το πρόβλημα με το δείγμα που έχει συλλεχθεί με αυτό τον τρόπο;
 - Μπορεί να διορθωθεί και πώς;

Πιθανά προβλήματα

1. Μη τυχαία αρχική επιλογή (seed bias)

Αν τα πρώτα άτομα («seeds») δεν είναι αντιπροσωπευτικά, μπορεί να προκαλέσουν μεροληψία που μεταδίδεται σε όλο το δείγμα.

2. Ανισότητες στα κοινωνικά δίκτυα

Ορισμένοι συμμετέχοντες έχουν μεγαλύτερα ή πιο “κλειστά” δίκτυα, πράγμα που επηρεάζει τις πιθανότητες να στρατολογηθούν ή να στρατολογήσουν άλλους.

3. Ανεπαρκής ανάμειξη (network homophily)

Τα άτομα τείνουν να στρατολογούν άτομα που τους μοιάζουν (ίδιο φύλο, ηλικία, εθνικότητα κ.λπ.), οδηγώντας σε μεροληψία ομοφιλίας.

Πιθανά προβλήματα

4. Εξάρτηση από την ακρίβεια αυτοαναφορών

Οι συμμετέχοντες πρέπει να αναφέρουν με ακρίβεια το μέγεθος του κοινωνικού τους δικτύου, κάτι που συχνά είναι ανακριβές ή αυθαίρετο.

5. Προβλήματα εγκυρότητας (validity issues)

Αν η αλυσίδα στρατολόγησης δεν προχωρήσει αρκετά “κύματα”, το δείγμα δεν “σταθεροποιείται” και τα αποτελέσματα εξαρτώνται υπερβολικά από τα αρχικά άτομα.

6. Δυσκολία υπολογισμού πραγματικού σφάλματος

Παρότι υπάρχουν στατιστικές διορθώσεις, το περιθώριο σφάλματος είναι δύσκολο να εκτιμηθεί με ακρίβεια.

Υπολογισμός μεγέθους δείγματος

Μέγεθος δείγματος

- Θέλουμε να εκτιμήσουμε το ποσοστό των ενηλίκων στην Ελλάδα που καπνίζουν
 - Πόσα άτομα θα έπρεπε να ρωτήσουμε;
- Τι σημαίνει ‘αρκετά μεγάλο δείγμα’;
- Από τι εξαρτάται;

Από τί εξαρτάται το μέγεθος δείγματος

- Από το μέγεθος του πληθυσμού;
- Από την ακρίβεια που θέλουμε;
- Από τη μεταβλητότητα;
- Από το ποσοστό απόκρισης;
- Από την μέθοδο της δειγματοληψίας;

Μέγεθος δείγματος

- Ο αριθμός των ατόμων που πρέπει να συμπεριληφθούν στη μελέτη εξαρτάται **πάντα** από τις παρακάτω παραμέτρους:
 - I. Διαφορά που θέλουμε να εντοπίσουμε
 - II. Μεταβλητότητα τους μεγέθους που μελετάμε
 - III. Επίπεδο σημαντικότητας
 - IV. Ισχύ της μελέτης
- Λαμβάνονται υπόψιν και άλλα χαρακτηριστικά της μελέτης κατά περίπτωση

I. Διαφορά που θέλουμε να εντοπίσουμε

Μια **μεγάλη διαφορά** εντοπίζεται και σε ένα **μικρό δείγμα**, ενώ όσο πιο **μικρή διαφορά** θέλουμε να εντοπίσουμε, τόσο **μεγαλύτερο δείγμα** θα χρειαστούμε.

➤ Αυτό σημαίνει πως το **μέγεθος του δείγματος** είναι **αντιστρόφως ανάλογο** με της διαφοράς που θέλουμε να εντοπίσουμε

Μεταβλητότητα του χαρακτηριστικού που μελετάμε

- Η μεταβλητότητα των μετρήσεων αντανακλάται μέσω της τυπικής απόκλισης ή της διασποράς
- Όσο **μεγαλύτερη** είναι η **μεταβλητότητα** (τυπική απόκλιση και διασπορά) τόσο **μεγαλύτερο δείγμα** απαιτείται.
- Συνεπώς το **μέγεθος του δείγματος** είναι **ανάλογο** με την **μεταβλητότητα**

Επίπεδο σημαντικότητας

- Συνδέεται με το σφάλμα τύπου I που είναι η πιθανότητα απόρριψης της μηδενικής υπόθεσης ενώ είναι αληθής και συνήθως επιλέγεται ίσο με 0.05.
- Αν το **επίπεδο σημαντικότητας** επιλεγεί να είναι **μικρότερο**, τότε το απαιτούμενο μέγεθος δείγματος (δεδομένων των υπολοίπων παραμέτρων) θα είναι **μεγαλύτερο**.
- Συνεπώς το **μέγεθος του δείγματος συνδέεται αντίστροφα με το επίπεδο σημαντικότητας**.

Προσοχή! Το επίπεδο σημαντικότητας καθορίζεται κατά το σχεδιασμό μιας μελέτης και είναι μας ενδιαφέρει μόνο εφόσον εντοπιστεί μια στατιστικώς σημαντική διαφορά

Ισχύς της μελέτης

- Η ισχύς της μελέτης είναι η πιθανότητα να εντοπιστεί μια αληθής στατιστικώς σημαντική διαφορά.
- Η ισχύς είναι ίση με $1 - \beta$, όπου β το σφάλμα τύπου II, δηλαδή η πιθανότητα να μην απορρίψουμε τη μηδενική υπόθεση, ενώ η μηδενική υπόθεση δεν ισχύει.
- Χρειαζόμαστε **μεγαλύτερο δείγμα για να ελαχιστοποιήσουμε το σφάλμα τύπου II**
- **Συνεπώς το μέγεθος του δείγματος είναι μεγαλύτερο για μικρότερο β ή αντίστοιχα για μεγαλύτερη ισχύ.**

Προσοχή! Το β καθορίζεται κατά το σχεδιασμό και είναι σημαντικό μόνο στην περίπτωση που δεν απορριφθεί η μηδενική υπόθεση

Η ισχύς λαμβάνεται υπόψιν μόνο στις συγκριτικές μελέτες και όχι σε όσες στοχεύουν στην εκτίμηση ποσοστών (π.χ. Επιπολασμό). Σε αυτού του είδους τις μελέτες, εκτιμάται διάστημα εμπιστοσύνης με συγκεκριμένο περιθώριο λάθους.

		Πραγματικότητα	
		H_0 αληθής	H_0 όχι αληθής
Απόφαση βάσει μελέτης	Δεν απορρίπτουμε H_0 $\Delta = 0$ $\ln(HR) = 0, HR=1$ $\ln(\lambda_T) - \ln(\lambda_S)=0$	 $(1-\alpha)$	Σφάλμα τύπου II $(\beta=20\%)$
	Απορρίπτουμε H_0 $\Delta < 0, HR<1$	Σφάλμα τύπου I $(\alpha=2.5\%)$	 $(\text{Ισχύς}=1-\beta=0.80)$

Μέγεθος δείγματος και είδος ερευνητικού ερωτήματος

Κύριος στόχος μελετών:

- Διερεύνηση, περιγραφή και εκτίμηση δεικτών (Περιγραφικές μελέτες)
- Έλεγχο υπόθεσης, για παράδειγμα σύγκριση τιμών δείκτη σε διαφορετικές ομάδες (Αναλυτικές)
- Στατιστικά μοντέλα για τη σχέση έκβασης με περισσότερες από μία μεταβλητές (Πολυπαραγοντικές μελέτες)
- Μέγεθος δείγματος και είδος του ερευνητικού ερωτήματος:
 - Περιγραφικές μελέτες: Διάστημα εμπιστοσύνης, α
 - Συγκριτικές μελέτες: Ισχύς μελέτης, β
 - Επιπλέον μεταβλητές απαιτούν μεγαλύτερο δείγμα, ώστε να διατηρείται η ισχύς και στις υποομάδες

Μέγεθος δείγματος και σχεδιασμός

•Ο υπολογισμός του μεγέθους του δείγματος πρέπει να λαμβάνει υπόψιν στοιχεία της δειγματοληπτικής μεθόδου

- Στρωματοποίηση
- Συστάδες
- Oversampling

Έλεγχος υπόθεσης και μέγεθος δείγματος

Έχουμε μια δίτιμη μεταβλητή (π.χ. νόσημα/όχι νόσημα). Έστω (p) η πραγματική επίπτωση (prevalence). Θέλουμε να ελέγξουμε:

$$H_0 : p = p_0$$

$$H_a : p \neq p_0$$

(p_0) = θεωρητική/υπόθεση prevalence

(p) = αναμενόμενη prevalence που θέλουμε να ανιχνεύσουμε

Θέλουμε να εντοπίσουμε διαφορά: $\delta = |p - p_0|$

Καθορίζουμε τα σφάλματα ως:

$$|\hat{p} - p_0| = z_{1-\alpha/2} \cdot \sqrt{\frac{p_0(1-p_0)}{N}}$$

$$\delta = z_{1-\beta} \sqrt{p(1-p)/N}$$

Έλεγχος υπόθεσης και μέγεθος δείγματος

Εάν ενδιαφερόμαστε να εντοπίσουμε μια διαφορά δ προς οποιαδήποτε κατεύθυνση (δηλαδή κάνουμε δίπλευρο έλεγχο), το αντίστοιχο διάστημα εμπιστοσύνης για το p μπορεί να γραφτεί ως:

$$\delta = z_{1-\alpha/2} \sqrt{\frac{p_0(1-p_0)}{N}} + z_{1-\beta} \sqrt{\frac{p_1(1-p_1)}{N}}.$$

Ακρίβεια και μέγεθος δείγματος

Σε μια περιγραφική (descriptive) μελέτη, θέλουμε να εκτιμήσουμε μια αναλογία (π.χ. επιπολασμό) στον πληθυσμό, δηλαδή μια τιμή:

p = πραγματικός επιπολασμός στον πληθυσμό

Χρησιμοποιούμε ένα δειγματικό ποσοστό $\hat{p}$ για να εκτιμήσουμε το p .

Θέλουμε η εκτίμηση αυτή να είναι αρκετά ακριβής, δηλαδή το διάστημα εμπιστοσύνης του $\hat{p}$ να έχει συγκεκριμένο εύρος ($\pm E$).

Έλεγχος υπόθεσης και μέγεθος δείγματος

Αν υποθέσουμε **τυχαία δειγματοληψία** και αρκετά μεγάλο δείγμα (ώστε να ισχύει το Κεντρικό Οριακό Θεώρημα), τότε:

$$\hat{p} \sim N \left(p, \frac{p(1-p)}{n} \right)$$

- Δηλαδή το $\hat{p}$ ακολουθεί κανονική κατανομή με μέση τιμή p διασπορά $\frac{p(1-p)}{n}$

Έλεγχος υπόθεσης και μέγεθος δείγματος

Το $(1-\alpha)\%$ διάστημα εμπιστοσύνης για το p θα είναι:

$$\hat{p} \pm Z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

Η ακρίβεια (E) είναι το **μέγιστο αποδεκτό σφάλμα εκτίμησης**, δηλαδή το μισό του εύρους του διαστήματος εμπιστοσύνης:

$$E = Z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

Λύνοντας ως προς n :

$$n = \frac{Z_{1-\alpha/2}^2 p(1-p)}{E^2}$$

Υπολογισμός Μεγέθους δείγματος

□ Ακρίβεια εκτίμησης:

□ Παράδειγμα: Δεδομένα συγχρονικής μελέτης
(Περιγραφικά)

□ Ποσοτική μεταβλητή:

–1 δείγμα: $n = (z\sigma/E)^2$

σ^2 : διασπορά,
E: Επιθυμητό περιθώριο σφάλματος
p: Πιθανότητα παρουσίας της έκβασης

–2 δείγματα: $n_i = 2(z\sigma^*/E)^2$

□ Κατηγορική μεταβλητή:

–1 δείγμα $n = p(1-p)(z/E)^2$

–2 δείγματα $n = [p_1(1-p_1) + p_2(1-p_2)](z/E)^2$

$$*\sigma^{pooled} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$$

Παράδειγμα: Επιπολασμός καπνίσματος- Β1

- Στόχος: Να εκτιμήσουμε το ποσοστό ενηλίκων που καπνίζουν σε μια περιοχή

–Να έχουμε ακριβείς εκτιμήσεις τόσο για άνδρες όσο και για γυναίκες

Βήμα 1: Ορισμός πληθυσμού και στρωμάτων
Ο πληθυσμός ενηλίκων είναι 100.000 άτομα:

Στρώμα	Ποσοστό στον πληθυσμό	Πληθυσμός
Άνδρες	40%	40.000
Γυναίκες	60%	60.000

Παράδειγμα: Επιπολασμός καπνίσματος-B2

Βήμα 2: Θέλουμε 95% εμπιστοσύνη και $\pm 5\%$ ακρίβεια

Για μια απλή αναλογία (π.χ. ποσοστό καπνιστών), ο τύπος για το μέγεθος δείγματος είναι:

$$n = p(1 - p)(z / E)^2$$

όπου:

($Z = 1.96$) (95% επίπεδο εμπιστοσύνης)

($p = 0.5$) (μέγιστη διακύμανση – συντηρητική επιλογή)

($E = 0.05$) (επιθυμητό περιθώριο σφάλματος)

$$n = 0.5(1 - 0.5)1.96^2 / 0.05^2 = 384.16$$

Άρα, για τον συνολικό πληθυσμό, ένα απλό τυχαίο δείγμα $n = 385$ θα έδινε περιθώριο $\pm 5\%$.

Παράδειγμα: Επιπολασμός καπνίσματος-B3

Βήμα 3: Παίρνουμε το δείγμα ανάλογο του πληθυσμού (κοινπιθανότητα επιλογής για κάθε άτομο, ανεξαρτήτως φύλου):

Στρώμα	Ποσοστό	Δείγμα
Άνδρες	40%	$0.40 \times 385 = 154$
Γυναίκες	60%	$0.60 \times 385 = 231$

Οι 154 άνδρες είναι λίγοι, θα είχαμε μεγάλο περιθώριο σφάλματος στους άνδρες.

Παράδειγμα: Επιπολασμός καπνίσματος-B4

Βήμα 4: Εφαρμόζουμε oversampling των ανδρών

Ας υποθέσουμε ότι θέλουμε περίπου **ίσα δείγματα** για τα δύο φύλα (ώστε κάθε φύλο να έχει παρόμοια ακρίβεια).

Κάνουμε oversampling των ανδρών (αντί για 40% άνδρες, θα πάρουμε 50% άνδρες).

Στρώμα	Πληθυσμός	Αναλογία στον πληθυσμό	Προτεινόμενο δείγμα	Αναλογία στο δείγμα
Άνδρες	40000	40%	250	50%
Γυναίκες	60000	60%	250	50%
Σύνολο	100000	100%	500	100%

Παράδειγμα: Επιπολασμός καπνίσματος-B5

Βήμα 5: Προκύπτουν δειγματοληπτικά βάρη

Επειδή οι άνδρες υπερεκπροσωπούνται στο δείγμα, κάθε άνδρας θα πρέπει “μετράει λιγότερο” και κάθε γυναίκα θα πρέπει “μετράει περισσότερο”.

$$\text{Βάρος} = \frac{\text{ποσοστό στον πληθυσμό}}{\text{ποσοστό στο δείγμα}}$$

Στρώμα	Αναλογία Πληθυσμού	Αναλογία Δείγματος	Βάρος
Άνδρες	0.40	0.50	0.8
Γυναίκες	0.60	0.50	1.2

Παράδειγμα: Επιπολασμός καπνίσματος-B6

Βήμα 6: Πώς επηρεάζεται το συνολικό μέγεθος δείγματος από τη στρωματοποίηση

Αν θέλαμε ίδιο περιθώριο σφάλματος ($\pm 5\%$) και για άνδρες και για γυναίκες,
τότε κάθε στρώμα θα χρειαζόταν ≈ 385 άτομα, άρα συνολικά
 $385 * 2 = 770$

Παράδειγμα: Επιπολασμός καπνίσματος-B7

Βήμα 7: Προσαρμογή για design effect

Αν το δειγματοληπτικό σχέδιο είναι πολύπλοκο
(π.χ. δειγματοληψία σε συστάδες με μέσο μέγεθος
5 άτομα και συντελεστή συσχέτισης $\rho=0.125$),

$$DEFF \approx 1 + (m-1)\rho = 1 + (5-1)*0.125 = 1.5$$

και υποθέσουμε ($DEFF = 1.5$), τότε:

$$N_{\text{τελικό}} = 770 * 1.5 = 1155$$

ΕΜΕΝΟ μέγεθος δείγματος

1 δείγμα με κύριο στόχο εκτίμηση επιπολασμού νοσημάτων

Βασίστηκε στην εκτίμηση του επιπολασμού της ιογενούς ηπατίτιδας Β- Κατηγορική έκβαση

Εκτίμηση επιπολασμού από προηγούμενες μελέτες: 1.5% δηλαδή $p=0.015$

Σφάλμα τύπου Ι: $\alpha=0.05$

Περιθώριο σφάλματος: $E=0.003$

$$N = z^2 p(1-p)/E^2 \approx 6000$$

ΕΜΕΝΟ μέγεθος δείγματος

On the basis of the ATTICA study results, the overall prevalence of hypercholesterolemia (total serum cholesterol levels >200 mg/dL or use of lipid lowering agents), hypertension (average blood pressure levels $>140/90$ mmHg or receiving antihypertensive medication), and of diabetes mellitus (fasting blood glucose >125 mg/dL or receiving antidiabetic medication) was about 40%, 32%, and 7%, respectively. **Calculating sampling errors using the corresponding formula for stratified sampling and with a target sample size of 6000 individuals**, the above-mentioned prevalence could be estimated with 1.28%, 1.22%, and 0.66% precision, respectively.

Υπολογισμός Μεγέθους δείγματος

□ Έλεγχος Υπόθεσης:

□ Παράδειγμα: Δεδομένα συγκριτικής μελέτης

□ Ποσοτική μεταβλητή:

–1 δείγμα:

$$n = \left(\frac{Z_{1-\alpha/2} + Z_{1-B}}{ES} \right)^2 \quad ES = \frac{[\mu_1 - \mu_0]}{\sigma}$$

–2 δείγματα:

$$n_i = 2 \left(\frac{Z_{1-\alpha/2} + Z_{1-B}}{ES} \right)^2 \quad ES = \frac{[\mu_1 - \mu_2]}{\sigma^*}$$

□ Κατηγορική μεταβλητή:

–1 δείγμα

$$n = \left(\frac{Z_{1-\alpha/2} + Z_{1-B}}{ES} \right)^2 \quad ES = \frac{|p_1 - p_0|}{\sqrt{p_0(1-p_0)}}$$

–2 δείγματα

$$n_i = 2 \left(\frac{Z_{1-\alpha/2} + Z_{1-B}}{ES} \right)^2 \quad ES = \frac{|p_1 - p_2|}{\sqrt{p(1-p)}}$$

Παράδειγμα 1: Εκτίμηση επιπέδων σακχάρου προπτυχιακών φοιτητών

Εκτίμηση δείκτη/ποσοτική μεταβλητή/1 δείγμα

- Αν θέλουμε $\alpha=0.05$, δηλαδή 95% διάστημα εμπιστοσύνης το $z= 1.96$
- Από προηγούμενη πιλοτική μελέτη η τυπική απόκλιση έχει εκτιμηθεί ίση με 25
- Θέλουμε το περιθώριο σφάλματος να είναι το πολύ 5mg/dl

$$n = (z\sigma/E)^2$$
$$n = (1.96 \times 25 / 5)^2$$

Δηλαδή το n είναι 96

Παράδειγμα 2: Ποιο είναι το ποσοστό ανδρών και γυναικών καπνιστών στο ΕΚΠΑ;

Εκτίμηση/κατηγορικά δεδομένα/2 δείγματα

- Για 95% διάστημα εμπιστοσύνης $z = 1.96$
- Από πιλοτική μελέτη για τους άνδρες $p = 0.25$ ενώ για τις γυναίκες $p = 0.2$
- Θέλουμε περιθώριο σφάλματος το πολύ 10% (.1)
$$n = [p_1(1 - p_1) + p_2(1 - p_2)](z / E)^2$$
- Συνεπώς:
$$n = [.25(1 - .25) + .2(1 - .2)](1.96 / .1)^2$$

$$n = 95 \text{ (ανά ομάδα)}$$

Παράδειγμα 3: Διαφέρουν τα ποσοστά καπνιστών μεταξύ των φοιτητών και του γενικού πληθυσμού?

Έλεγχος υπόθεσης/κατηγορικά δεδομένα/1 δείγμα

- Επιλέγουμε $\alpha=0.05$ (άρα $z_{1-\alpha/2}=1.96$); Ισχύ $(1-B)$ 80% (άρα $z_{1-B}=0.84$)
- Από πιλοτική μελέτη μεταξύ των φοιτητών είναι $p = 0.22$ ενώ στο γενικό πληθυσμό είναι $p = 0.3$

➤ Τότε:

$$n = \left(\frac{z_{1-\alpha/2} + z_{1-B}}{ES} \right)^2$$

$$ES = \frac{|p_1 - p_0|}{\sqrt{p_0(1-p_0)}}$$

$$ES = \frac{|.22 - .3|}{\sqrt{.3(1-.3)}}$$

$$n = \left(\frac{1.96 + .84}{.17} \right)^2$$

$$n = 272$$

Παράδειγμα 4: Διαφέρουν τα επίπεδα σακχάρου σε άνδρες και γυναίκες?

Έλεγχος υπόθεσης/συνεχή δεδομένα/2 δείγματα

• Επιλέγουμε $\alpha=0.05$ (άρα $z_{1-\alpha/2}=1.96$); Ισχύ $(1-B)$ 80% (άρα $z_{1-B}=0.84$)

➤ Από πιλοτική μελέτη μεση τιμή σακχάρου στις γυναίκες: 95mg/dl με $sd=10$; στους άνδρες: 100mg/dl με $sd=10$

➤ Συνεπώς:

$$n_i = 2 \left(\frac{z_{1-\alpha/2} + z_{1-B}}{ES} \right)^2$$

$$ES = \frac{[\mu_1 - \mu_2]}{\sigma}$$

$$ES = \frac{|95 - 100|}{10}$$

$$n = 2 \left(\frac{1.96 + .84}{.5} \right)^2$$

$$n = 63 \text{ (ανά ομάδα)}$$

Ανάλυση Surveys Stata

[SVY] svyset -- Declare survey design for dataset

Syntax

Single-stage design

- `svyset [psu] [weight] [, design_options options]`
- Multiple-stage design
- `svyset psu [weight] [, design_options] [|| ssu , design_options] ... [options]`

•Example (NHANES)

- `. webuse nhanes2f`
- `. svyset psuid [pweight=finalwgt], strata(stratid)`

Stata

-

• Αφού δηλωθεί ότι τα δεδομένα προέρχονται από survey με συγκεκριμένο σχεδιασμό, ο σχεδιασμός λαμβάνεται υπόψη στις εκτιμήσεις.

-

```
.webuse nhanes2f
```

- . svyset psuid [pweight=finalwgt], strata(stratid)
- . svy: mean zinc
- . svy: regress zinc age age2 weight female black orace rural

Δυνατότητα παρακολούθησης

- Συχνά ζητείται από τους συμμετέχοντες να συναινέσουν για μελλοντική επικοινωνία
- Για όσους συμμετέχοντες συναινέσουν, γίνεται νέος κύκλος συλλογής δεδομένων, με όσες πληροφορίες πρέπει να ανανεωθούν
- Η μελέτη σε αυτή την περίπτωση γίνεται από
 - συγχρονική, μελέτη κοορτής