

Στατιστικές Μέθοδοι στην Επιδημιολογία

Lab 4

Μελέτη EMENO

Θα χρησιμοποιήσουμε δεδομένα από τη μελέτη EMENO (Εθνική Μελέτη Νοσηρότητας και Παραγόντων Κινδύνου), μια συγχρονική επιδημιολογική μελέτη υγείας που πραγματοποιήθηκε στην Ελλάδα την περίοδο **2013–2016**. Η EMENO χρησιμοποίησε **πολυσταδιακό (τριών σταδίων) στρωματοποιημένο τυχαίο δειγματοληπτικό σχεδιασμό** για την επιλογή αντιπροσωπευτικού δείγματος ενηλίκων (≥ 18 ετών), με βάση τα στοιχεία της Απογραφής 2011.

Στο πλαίσιο της στρωματοποίησης, η χώρα χωρίστηκε σε **γεωγραφικά διαμερίσματα** (11 ευρύτερες γεωγραφικές περιοχές \times 3 βαθμοί αστικοποίησης: αστική, ημιαστική, αγροτική), δημιουργώντας **33 στρώματα (strata)** συνολικά. Από αυτά προέκυψαν **577 απογραφικές περιοχές (enumeration areas)**, στις οποίες εφαρμόστηκε η διαδικασία επιλογής νοικοκυριών και ατόμων σύμφωνα με το πρωτόκολλο. Στη συνέχεια υλοποιήθηκαν τα **3 στάδια επιλογής**:

- (1) τυχαία επιλογή **απογραφικών περιοχών (PSUs)** μέσα στα strata,
- (2) συστηματική δειγματοληψία **νοικοκυριών** ανά PSU,
- (3) τυχαία επιλογή **ενός** ενηλίκου ανά νοικοκυριό με το κριτήριο της πιο πρόσφατης ημερομηνίας γέννησης.

Οι **δειγματοληπτικοί λόγοι (sampling fractions)** διέφεραν μεταξύ των strata, ώστε να διασφαλιστεί επαρκής κάλυψη και ακρίβεια σε όλους τους συνδυασμούς **γεωγραφικής περιοχής \times βαθμού αστικοποίησης**. Τυχόν **υπερ-/υπο-εκπροσώπηση** στο ακατέργαστο δείγμα **διορθώνεται** στην ανάλυση με τα βάρη. Τα **δειγματοληπτικά βάρη (sampling weights)** υπολογίστηκαν ως **αντίστροφα των πιθανοτήτων επιλογής** (inverse of inclusion probabilities), ώστε κάθε συμμετέχων να εκπροσωπεί τον κατάλληλο αριθμό ατόμων στον πληθυσμό.

Σκοπός: Να κατανοηθεί ο τρόπος με τον οποίο ο δειγματοληπτικός σχεδιασμός (στρωματοποίηση, στάδια επιλογής, βάρη) επηρεάζει την ανάλυση των δεδομένων επιδημιολογικών ερευνών και να εφαρμοστούν βασικές τεχνικές στάθμισης (*survey weights* και *post-stratification*).

1. Θεωρητική κατανόηση του δειγματοληπτικού σχεδιασμού EMENO

1. Ποια ήταν τα κύρια στάδια της δειγματοληψίας (π.χ. επιλογή πρωτογενών μονάδων, νοικοκυριών, ατόμων);
2. Πόσα strata χρησιμοποιήθηκαν και με ποιο κριτήριο καθορίστηκαν;
3. Πώς επιλέχθηκε το άτομο εντός κάθε νοικοκυριού και γιατί χρησιμοποιήθηκε η μέθοδος “τελευταία γενέθλια”;
4. Ποια είναι η πιθανότητα ένταξης ενός ατόμου στο δείγμα και πώς ορίζεται το `base_weight`;

2. Επισκόπηση & Ορισμός Δειγματοληπτικού Σχεδιασμού

2.1 Επισκόπηση *dataset*

5. Ανοίξτε το αρχείο `analysis.dta`. Ποιες μεταβλητές περιλαμβάνει;

```
. use analysis.dta, clear  
. desc
```

2.2 Περιγραφή δειγματοληπτικών βαρών

6. Εξετάστε την κατανομή της μεταβλητής `base_weight`, η οποία αντιστοιχεί στα βάρη δειγματοληψίας. Τι παρατηρείτε;

```
. codebook base_weight  
. summarize base_weight  
. tabstat base_weight, stat(n mean sd p25 p50 p75 min max)  
. histogram base_weight, percent
```

7. Τι αντιπροσωπεύει το `base_weight`; Τι αντιπροσωπεύει μία τιμή, π.χ. 1.500 ή 2.000;

8. Πόσες διαφορετικές τιμές θα περιμένατε θεωρητικά στα `base_weight`, αν η πιθανότητα επιλογής ήταν ίδια για όλα τα άτομα μέσα σε κάθε `stratum`; Γιατί ο πραγματικός αριθμός των τιμών μπορεί να διαφέρει;

2.3 Ορισμός δειγματοληπτικού σχεδίου (`svyset`)

9. Ορίστε τις μεταβλητές του δειγματοληπτικού σχεδιασμού με την εντολή `svyset`. Ποια στοιχεία δηλώνετε στο `svyset` και γιατί είναι απαραίτητα;

```
. svyset blockid [pweight = base_weight], strata(strata)
```

10. Πόσα `strata` και πόσες PSUs υπάρχουν; Υπάρχει **strata** με μοναδικό PSU; Αν ναι, ποιο είναι το πρόβλημα και πώς θα το χειριστείτε;

```
. svydescribe  
. svyset blockid [pweight = base_weight], strata(strata)  
singleunit(centered)
```

3. Σταθμισμένες κατανομές & σύγκριση δείγματος–πληθυσμού

3.1 Ηλικιακές ομάδες

11. Υπολογίστε την **κατανομή των ηλικιακών ομάδων** χωρίς και με στάθμιση. Τι παρατηρείτε;

```
. tab age_gr  
. svy: tab age_gr, per obs
```

3.2 Φύλο & Γεωγραφική περιοχή

12. Συγκρίνετε τις μη σταθμισμένες και τις σταθμισμένες κατανομές για το φύλο. Τι αλλάζει;

```
. tab gender  
. svy: tab gender, per obs
```

13. Επαναλάβετε για τη γεωγραφική περιοχή. Ποιες διαφορές παρατηρούνται μεταξύ της μη σταθμισμένης και της σταθμισμένης κατανομής;

```
. tab area  
. svy: tab area, per obs  
. svy: proportion area
```

3.3 Σύγκριση με τις κατανομές του πληθυσμού (ΕΛΣΤΑΤ)

14. Χρησιμοποιώντας τα στοιχεία της ΕΛΣΤΑΤ, συγκρίνετε την κατανομή φύλου και ηλικιακών ομάδων με αυτή που εκτιμήθηκε από το δείγμα. Τι παρατηρείτε; Υπάρχουν διαφορές; Αν ναι, πώς μπορούν να διορθωθούν;

Πανελλαδικά		
Πληθυσμός %		
Ηλικία	Άνδρας	Γυναίκα
18–29	9,0	8,5
30–39	9,2	9,0
40–49	8,8	9,0
50–59	7,6	8,0
60–69	6,1	6,6
70–79	5,8	6,2
80+	2,7	3,9
Σύνολο	48,6	51,4

```
. svy: tab age_gr gender, per obs
```

4. Δημιουργία τελικών βαρών (μετα-στρωμάτωσης)

gender	age_gr	area	freq
0	0	1	296461
0	1	1	317576
0	2	1	282668
0	3	1	235520
0	4	1	182732

Απογραφή ΕΛΣΤΑΤ 2011

15. Εξαιρέστε παρατηρήσεις με άγνωστη γεωγραφική περιοχή, ηλικιακή ομάδα ή φύλο και συγχωνεύστε το dataset με το αρχείο **census2011.dta** χρησιμοποιώντας ως κλειδιά τις μεταβλητές area, gender και age_gr. Ελέγξτε πόσες παρατηρήσεις δεν βρήκαν αντιστοίχιση.

```
. drop if missing(area) | missing(age_gr) | missing(gender)
```

```
. merge m:1 area gender age_gr using census2011.dta  
. rename _merge _merge_census
```

16. Υπολογίστε, για κάθε συνδυασμό περιοχής, φύλου και ηλικιακής ομάδας, το **συνολικό σταθμισμένο μέγεθος** (άθροισμα των αρχικών βαρών `base_weight`). Τι εκφράζει η μεταβλητή `total_we`;

```
. bysort area gender age_gr: egen total_we = total(base_weight)
```

17. Υπολογίστε τον προσαρμοσμένο συντελεστή μεταστρωμάτωσης (`adj_factor`):

$$\text{adj_factor} = \frac{\text{Πληθυσμιακή συχνότητα (από census)}}{\text{Εκτιμώμενη συχνότητα (από base weights)}}$$

```
. gen adj_factor = freq / total_we
```

18. Δημιουργήστε τη μεταβλητή του τελικού βάρους (`weight_final`)

```
. gen weight_final = base_weight * adj_factor  
. label var weight_final "Post-stratified final weight"
```

19. Δηλώστε εκ νέου το δειγματοληπτικό σχέδιο με τα τελικά βάρη μέσω της εντολής **svyset**.

```
. svyset blockid [pweight = weight_final], strata(strata)  
singleunit(centered)
```

20. Υπολογίστε εκ νέου την **κατανομή ηλικίας και φύλου** με τα τελικά βάρη και συγκρίνετε τα αποτελέσματα με τα επίσημα στοιχεία της ΕΛΣΤΑΤ. Είναι πλέον πιο κοντά;

```
. svy: tab age_gr gender, per obs
```