

## Lab 4: Solutions

### 1. Θεωρητική κατανόηση του δειγματοληπτικού σχεδιασμού EMENO

1. Ποια ήταν τα κύρια στάδια της δειγματοληψίας (π.χ. επιλογή πρωτογενών μονάδων, νοικοκυριών, ατόμων);

#### Απάντηση

Η μελέτη EMENO χρησιμοποίησε **πολυσταδιακό (τριών σταδίων) στρωματοποιημένο τυχαίο δειγματοληπτικό σχεδιασμό**, ώστε να επιλεγεί αντιπροσωπευτικό δείγμα του ενήλικου πληθυσμού της Ελλάδας ( $\geq 18$  ετών).

- **1ο στάδιο (Πρωτογενείς Μονάδες Δειγματοληψίας, PSUs):** τυχαία επιλογή **απογραφικών περιοχών (enumeration areas)** μέσα σε κάθε stratum.
- **2ο στάδιο:** συστηματική δειγματοληψία **νοικοκυριών** μέσα σε κάθε απογραφική περιοχή, με δειγματοληπτικό διάστημα ανάλογο της πυκνότητας πληθυσμού (π.χ. κάθε 2ο–4ο νοικοκυριό).
- **3ο στάδιο:** τυχαία επιλογή **ενός ενήλικου ατόμου** από κάθε επιλεγμένο νοικοκυριό.

2. Πόσα strata χρησιμοποιήθηκαν και με ποιο κριτήριο καθορίστηκαν;

#### Απάντηση

Χρησιμοποιήθηκαν συνολικά **33 strata**, τα οποία προέκυψαν από τον συνδυασμό:

- **11 ευρύτερων γεωγραφικών περιοχών** της χώρας (με την Αττική και τη Θεσσαλονίκη ως ξεχωριστές ενότητες) ×
- **3 βαθμών αστικότητας:** αστική, ημιαστική και αγροτική περιοχή.

Η στρωματοποίηση αυτή εξασφαλίζει αντιπροσωπευτικότητα γεωγραφικά και ως προς τον βαθμό αστικότητας, ώστε να καλύπτονται επαρκώς τόσο οι αστικές όσο και οι αγροτικές περιοχές.

3. Πώς επιλέχθηκε το άτομο εντός κάθε νοικοκυριού και γιατί χρησιμοποιήθηκε η μέθοδος “τελευταία γενέθλια”;

**Απάντηση**

Από κάθε νοικοκυριό επιλέχθηκε **ένα άτομο** ηλικίας  $\geq 18$  ετών με τη μέθοδο των «**τελευταίων γενεθλίων**» (“last birthday method”).

Η μέθοδος αυτή είναι απλή και αντικειμενική: εξασφαλίζει τυχαία επιλογή μεταξύ όλων των μελών του νοικοκυριού, αποφεύγοντας μεροληψία (π.χ. επιλογή του πιο πρόθυμου ή διαθέσιμου ατόμου). Έτσι, κάθε ενήλικος του νοικοκυριού έχει ίση πιθανότητα να συμμετάσχει.

4. Ποια είναι η πιθανότητα ένταξης ενός ατόμου στο δείγμα και πώς ορίζεται το `base_weight`;

**Απάντηση**

Η **πιθανότητα ένταξης** ενός ατόμου στο δείγμα είναι το **γινόμενο των πιθανοτήτων επιλογής** σε κάθε στάδιο:

$$\pi_i = \pi_{1i} \times \pi_{2i} \times \pi_{3i}$$

όπου:

$\pi_{1i}$ : πιθανότητα επιλογής της απογραφικής περιοχής,

$\pi_{2i}$ : πιθανότητα επιλογής του νοικοκυριού εντός της απογραφικής περιοχής,

$\pi_{3i}$ : πιθανότητα επιλογής του ατόμου εντός του επιλεγμένου νοικοκυριού.

Το **`base_weight`** ορίζεται ως το **αντίστροφο της συνολικής πιθανότητας ένταξης**:

$$base_{weight} = \frac{1}{\pi_i}$$

Έτσι, κάθε συμμετέχων αντιπροσωπεύει έναν αριθμό ατόμων του πληθυσμού με ίδια πιθανότητα επιλογής.

Στην πράξη, το βάρος μπορεί να ερμηνευθεί ως «πλήθος πληθυσμού που εκπροσωπεί» η μονάδα, αλλά αυτή η ερμηνεία είναι αυστηρά ορθή μόνο για μη κανονικοποιημένα design weights που ισούνται με το ακριβές  $\frac{1}{\pi_i}$  (πριν από τυχόν rescaling ή calibration). Για κανονικοποιημένα ή αναπροσαρμοσμένα βάρη, η ερμηνεία παραμένει αναλογική (relative) και όχι απόλυτη σε επίπεδο head-count.

## 2. Επισκόπηση & Ορισμός Δειγματοληπτικού Σχεδιασμού

### 2.1 Επισκόπηση dataset

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

#### Απάντηση

```
. use analysis.dta, clear  
. describe
```

Contains data from analysis\_stata13.dta

```
obs:      6,006  
vars:      9  
size:     276,276  
3 Nov 2025 20:55
```

variable name	storage type	display format	value label	variable label
emenoid	double	%12.0g		Identifier (emenoid)
blockid	double	%12.0g		PSU (blockid)
strata	float	%9.0g		Stratum
urban	float	%20.0g	urban_lbl	Degree of urbanization
area	float	%14.0g	region_lbl	Region
base_weight	int	%8.0g		Sampling weight
gender	long	%12.0g	gender_lbl	Gender
age	double	%12.0g		Age (years)
age_gr	float	%9.0g	ilikiaki_lbl	Age group

Sorted by:

### 2.2 Περιγραφή δειγματοληπτικών βαρών

6. Εξετάστε την κατανομή της μεταβλητής **base\_weight**, η οποία αντιστοιχεί στα βάρη δειγματοληψίας. Τι παρατηρείτε;

### Απάντηση

```
. summarize base_weight
```

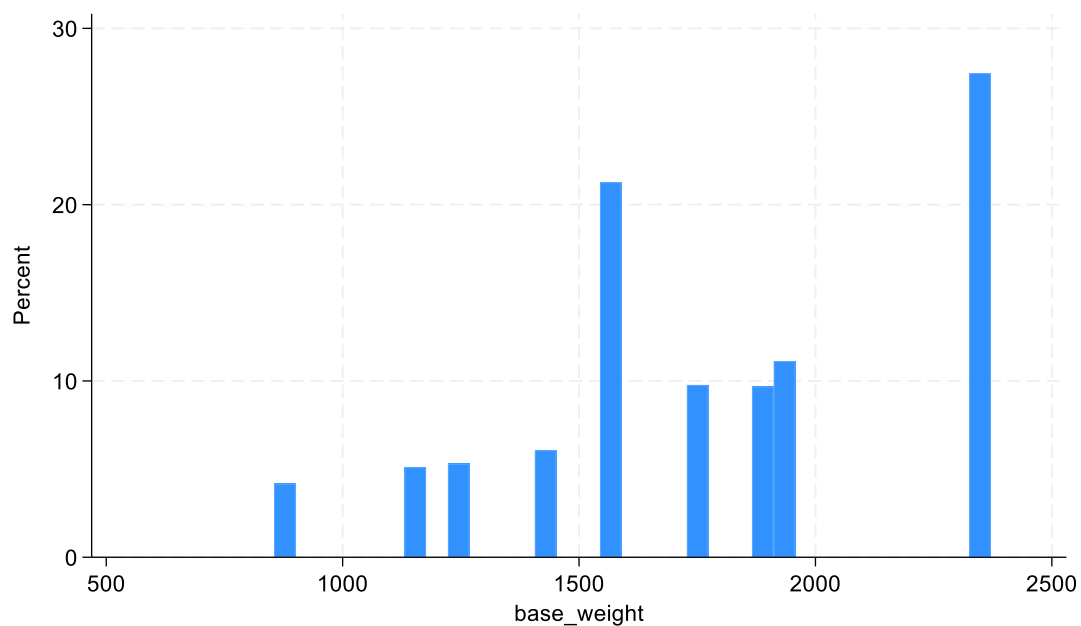
Variable	Obs	Mean	Std. Dev.	Min	Max
base_weight	5993	1804.293	432.3026	855	2371

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

variable	N	mean	sd	p25	p50	p75	min	max
base_weight	5993	1804.293	432.3026	1561	1763	2352	855	2371

Η μεταβλητή **base\_weight** παρουσιάζει σχετικά μεγάλη διακύμανση, με τιμές που κυμαίνονται περίπου από 855 έως 2.371.

```
. histogram base_weight, percent
```



Η κατανομή της μεταβλητής **base\_weight** είναι εμφανώς **διακριτή** και **δεξιά-λοξή**, με σχετικά λίγες μοναδικές τιμές, όπως αναμενόταν για πολυσταδιακό στρωματοποιημένο σχέδιο.

Οι περισσότερες παρατηρήσεις συγκεντρώνονται γύρω από τιμές 1.500–2.000, ενώ υπάρχει δεύτερη ισχυρή κορυφή κοντά στα 2.400–2.500.

Οι χαμηλότερες τιμές (π.χ. <1.200) αντιστοιχούν σε άτομα από strata με **μεγαλύτερη πιθανότητα επιλογής** (πυκνοκατοικημένες αστικές περιοχές), ενώ οι υψηλότερες σε **αραιοκατοικημένες αγροτικές περιοχές** με μικρότερη πιθανότητα επιλογής.

Το γεγονός ότι παρατηρούνται λίγες μοναδικές τιμές αντικατοπτρίζει την ύπαρξη **ενιαίων δειγματοληπτικών λόγων εντός stratum**. Συνεπώς, η διακύμανση των βαρών προέρχεται κυρίως από διαφορές **μεταξύ** και όχι **εντός** strata.

**7.** Τι αντιπροσωπεύει το base\_weight; Τι αντιπροσωπεύει μία τιμή, π.χ. 1.500 ή 2.000;

#### **Απάντηση**

Η μεταβλητή **base\_weight** είναι το **αντίστροφο της συνολικής πιθανότητας επιλογής** ενός συμμετέχοντα. Εκφράζει **πόσα άτομα του γενικού πληθυσμού** εκπροσωπεί το συγκεκριμένο άτομο στην ανάλυση.

**Ερμηνεία τιμών:** Μια τιμή π.χ. **1.500** σημαίνει ότι το άτομο αυτό αντιπροσωπεύει περίπου **1.500 άτομα του πληθυσμού** με τα ίδια χαρακτηριστικά στο συγκεκριμένο stratum. Μια μεγαλύτερη τιμή, π.χ. **2.000** σημαίνει **μικρότερη πιθανότητα ένταξης** (π.χ. αραιοκατοικημένη/αγροτική περιοχή) και άρα «εκπροσωπεί» **περισσότερους** ανθρώπους.

- **Μικρό βάρος** → **υψηλή πιθανότητα επιλογής** (π.χ. αστικές περιοχές)
- **Μεγάλο βάρος** → **χαμηλή πιθανότητα επιλογής** (π.χ. αγροτικές ή αραιοκατοικημένες περιοχές)

**Μεταβολή βαρών:** Στην EMENO, τα base weights έχουν περιορισμένο εύρος διακύμανσης και συχνά εμφανίζονται ίδιες τιμές εντός strata, κάτι που αντικατοπτρίζει τον συγκεκριμένο μηχανισμό επιλογής του δείγματος. Αυτό όμως δεν αποτελεί γενικό χαρακτηριστικό των στρωματοποιημένων δειγματοληψιών.

**Χρήση στην ανάλυση:** Τα βάρη αυτά χρησιμοποιούνται για να **διορθώσουν τις διαφορές στις πιθανότητες επιλογής**, ώστε οι εκτιμήσεις (μέσοι όροι, ποσοστά, λόγοι πιθανοτήτων κ.λπ.) να είναι **αντιπροσωπευτικές του ενήλικου πληθυσμού** της Ελλάδας.

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;

#### Απάντηση

```
. codebook base_weight
```

```
base_weight                                     Sampling weight

      type:  numeric (int)

      range:  [855,2371]          units:  1
unique values: 31                missing .: 13/6006

      mean:  1804.29
      std. dev: 432.303

percentiles:      10%      25%      50%      75%      90%
                  1224      1561      1763      2352      2371
```

Θεωρητικά, επειδή η μελέτη περιλαμβάνει 33 strata, θα περιμέναμε έως και 33 διαφορετικές τιμές του `base_weight` (μία ανά stratum).

Στην πράξη, εμφανίζονται **31 μοναδικές τιμές**, κάτι που δείχνει ότι δύο ζεύγη strata (19, 21 και 14, 5) είχαν ίδιες πιθανότητες επιλογής και συνεπώς ίδιο βάρος,

γεγονός που είναι αναμενόμενο όταν περιοχές έχουν παρόμοιο πληθυσμιακό μέγεθος ή χαρακτηριστικά.

### 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;

#### Απάντηση

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

```
pweight: base_weight  
VCE: linearized  
Single unit: missing  
Strata 1: strata  
SU 1: blockid  
FPC 1: <zero>
```

Η δήλωση `svyset` καθορίζει τη δομή του δειγματοληπτικού σχεδιασμού, ώστε οι αναλύσεις που ακολουθούν να λαμβάνουν υπόψη τη **στρωματοποίηση, την ομαδοποίηση και τα δειγματοληπτικά βάρη**:

#### **i. Πρωτογενείς Μονάδες Δειγματοληψίας (Primary Sampling Units – PSUs)**

Οι PSU είναι οι μονάδες πρώτου σταδίου του σχεδίου.

Στην EMENO αντιστοιχούν στις **απογραφικές περιοχές ή οικοδομικά τετράγωνα** (`blockid`) που επιλέχθηκαν τυχαία μέσα σε κάθε stratum.

#### **ii. Στρώματα (Strata)**

Η μεταβλητή `strata` αντιπροσωπεύει τον συνδυασμό **γεωγραφικής περιοχής × βαθμού αστικότητας** (33 strata συνολικά).

Ο σωστός ορισμός των strata επιτρέπει τον υπολογισμό της διακύμανσης, λαμβάνοντας υπόψη την ομοιογένεια εντός και την ετερογένεια μεταξύ των στρωμάτων.

#### **iii. Δειγματοληπτικά Βάρη (Sampling Weights)**

Η μεταβλητή `base_weight` είναι το **αντίστροφο της πιθανότητας επιλογής**, εξασφαλίζοντας ότι οι εκτιμήσεις (`svy: mean, proportion, logistic` κ.λπ.) είναι **αντιπροσωπευτικές του συνολικού πληθυσμού**.

Οι σταθμισμένες εκτιμήσεις που προκύπτουν από τις εντολές `svy`: υπό τον σωστό ορισμό `svyset` είναι design-consistent για τα αντίστοιχα πληθυσμιακά μεγέθη, υπό τον δηλωμένο δειγματοληπτικό σχεδιασμό.

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

**Απάντηση**

```
. svydescribe
```

Stratum	#Units	#Obs	#Obs per Unit		
			min	mean	max
1	125	1494	9	12.0	13
2	11	144	11	13.1	21
3	1*	8	8	8.0	8
4	15	170	2	11.3	15
5	8	96	12	12.0	12
6	16	128	8	8.0	8
7	36	436	11	12.1	16
8	10	120	12	12.0	12
9	3	24	8	8.0	8
10	13	156	12	12.0	12
11	6	72	12	12.0	12
12	11	90	8	8.2	10
13	20	246	3	12.3	18
14	6	71	11	11.8	12
15	20	160	8	8.0	8
16	20	249	11	12.4	16
17	9	85	6	9.4	13
18	30	252	5	8.4	16
19	9	107	11	11.9	12
20	5	60	12	12.0	12
21	17	140	8	8.2	12
22	5	50	2	10.0	12
23	5	56	8	11.2	12
24	18	144	8	8.0	8
25	10	120	10	12.0	13
26	10	119	8	11.9	15
27	21	165	3	7.9	9
28	22	261	4	11.9	16
29	12	143	11	11.9	12
30	33	263	7	8.0	9
31	11	132	12	12.0	12
32	8	96	12	12.0	12
33	17	136	8	8.0	8
33	563	5993	2	10.6	21

13 = #Obs with missing values in the  
survey characteristics  
6006



Η εντολή παρέχει συνοπτικά τα χαρακτηριστικά του design. Στην EMENO έχουμε:

- Strata: 33 (γεωγραφία × αστικότητα)
- PSUs: ~563 (δειγματοληπτικά σημεία)
- Μέγεθος δείγματος: 5.993 άτομα
- Εκτιμώμενος πληθυσμός: ~10,8 εκατομμύρια ενήλικες.

Η σωστή δήλωση των **PSU**, **strata** και **weights** διασφαλίζει ότι οι αναλύσεις με svy: υπολογίζουν σωστά τα **τυπικά σφάλματα** και τα **διαστήματα εμπιστοσύνης**, λαμβάνοντας υπόψη τη σύνθετη δομή του δείγματος.

### Περίπτωση strata με μοναδικό PSU

Αν ένα stratum περιέχει **μόνο μία PSU (singleton PSU)**, το Stata δεν μπορεί να εκτιμήσει τη διακύμανση εντός του strata (variance = 0).

Για να αντιμετωπιστεί αυτό, χρησιμοποιείται η επιλογή `singleunit(centered)`:

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

```
pweight: base_weight  
VCE: linearized  
Single unit: centered  
Strata 1: strata  
SU 1: blockid  
FPC 1: <zero>
```

Η επιλογή `singleunit(centered)` είναι μία από τις καθιερωμένες επιλογές για την αντιμετώπιση singleton PSUs. Αποτελεί approximating method και δεν ισοδυναμεί με πλήρη variance identification.

## 3. Σταθμισμένες κατανομές & σύγκριση δείγματος–πληθυσμού

### 3.1 Ηλικιακές ομάδες

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

**Απάντηση**

```
. tab age_gr
```

Age group	Freq.	Percent	Cum.
18-29	634	10.58	10.58
30-39	819	13.67	24.24
40-49	1,020	17.02	41.26
50-59	1,113	18.57	59.84
60-69	1,111	18.54	78.37
70+	1,296	21.63	100.00
Total	5,993	100.00	

Παρατηρούμε ότι στο δείγμα τα ποσοστά των νεότερων ηλικιών είναι χαμηλότερα από των μεγαλύτερων.

Αυτό μπορεί να οφείλεται είτε στον τρόπο επιλογής των μονάδων είτε σε χαμηλότερη συμμετοχή των νεότερων ατόμων.

Η σταθμισμένη κατανομή αποτυπώνει την εκτίμηση για τον πληθυσμό.

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

Number of strata	=	33	Number of obs	=	5993
Number of PSUs	=	563	Population size	=	10813129
			Design df	=	530

Age group	percentages	obs
18-29	10.65	634
30-39	13.77	819
40-49	17.12	1020
50-59	18.29	1113
60-69	18.34	1111
70+	21.83	1296
Total	100	5993

Key: percentages = cell percentages  
obs = number of observations

Μετά τη στάθμιση, οι κατανομές εξισορροπούνται και προσεγγίζουν καλύτερα τη δημογραφική δομή του πληθυσμού, όπως αυτή προκύπτει από τα απογραφικά δεδομένα (ΕΛΣΤΑΤ 2011).

### 3.2 Φύλο & Γεωγραφική περιοχή

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

#### Απάντηση

```
. tab gender
```

Gender	Freq.	Percent	Cum.
Male	2,550	42.46	42.46
Female	3,455	57.54	100.00
Total	6,005	100.00	

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

Number of strata	=	33	Number of obs	=	5993
Number of PSUs	=	563	Population size	=	10813129
			Design df	=	530

Gender	percentages	obs
Male	42.36	2546
Female	57.64	3447
Total	100	5993

Key: percentages = cell percentages  
obs = number of observations

Οι σταθμισμένες και οι μη σταθμισμένες κατανομές φύλου είναι σχεδόν ίδιες (42.4% άνδρες, 57.6% γυναίκες).

Αυτό δείχνει ότι η πιθανότητα επιλογής δεν διαφοροποιήθηκε ουσιαστικά μεταξύ ανδρών και γυναικών.

Επομένως, τα βάρη δεν επηρεάζουν ουσιαστικά τη συγκεκριμένη μεταβλητή, γεγονός που υποδηλώνει ότι το φύλο δεν συνδέεται με άνισες πιθανότητες επιλογής. Αντίθετα, σε άλλες μεταβλητές η στάθμιση μπορεί να επιφέρει μεγαλύτερες αλλαγές λόγω ανισομερούς δειγματοληψίας.

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

**Απάντηση**

```
. tab area
```

Region	Freq.	Percent	Cum.
Athens	1,655	27.56	27.56
Crete	394	6.56	34.12
Thessaloniki	581	9.67	43.79
Thrace	318	5.29	49.08
Thessaly	477	7.94	57.03
Peloponnese	586	9.76	66.78
Epirus	308	5.13	71.91
Corfu	250	4.16	76.07
Central Greece	404	6.73	82.80
Macedonia	669	11.14	93.94
Lesvos-Rhodes	364	6.06	100.00
Total	6,006	100.00	

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

Region	percentages	obs
Athens	36.06	1646
Crete	5.68	394
Thessalo	10.18	580
Thrace	3.603	318
Thessaly	6.986	477
Peloponn	9.543	586
Epirus	3.286	307
Corfu	1.986	250
Central	5.914	404
Macedoni	11.95	667
Lesvos-R	4.81	364
Total	100	5993

Key: percentages = cell percentages  
obs = number of observations

```
. svy: proportion area
```

Survey: Proportion estimation

Number of strata =	33	Number of obs =	5993
Number of PSUs =	563	Population size =	10813129
		Design df =	530

\_prop\_9: area = Central Greece  
\_prop\_11: area = Lesvos-Rhodes

	Linearized			
	Proportion	Std. Err.	[95% Conf. Interval]	
area				
Athens	.3606329	.0023782	.3559743	.3653179
Crete	.0567956	.0015109	.0538993	.0598377
Thessaloniki	.1017866	.0008361	.1001559	.1034409
Thrace	.0360318	.000282	.0354819	.0365898
Thessaly	.0698616	.0017496	.0665019	.0733778
Peloponnese	.0954345	.0018514	.0918591	.099134
Epirus	.0328619	.0004576	.0319748	.0337727
Corfu	.0198579	.0008406	.0182721	.0215783
_prop_9	.0591365	.0011402	.0569359	.0614167
Macedonia	.1195006	.0016321	.1163313	.1227442
_prop_11	.0481	.0002386	.0476334	.048571

Note: Strata with single sampling unit centered at overall mean.

Μετά τη στάθμιση με τα αρχικά βάρη (base\_weight), η αναλογία των συμμετεχόντων από την Αθήνα αυξάνεται (27,6% → 36,1%), ενώ μικρότερες περιοχές μειώνονται.

Οι διαφορές αυτές προκύπτουν από τις άνισες πιθανότητες επιλογής στα διάφορα strata του δειγματοληπτικού σχεδίου.

Η στάθμιση εξισορροπεί αυτές τις διαφορές, χωρίς να σημαίνει ακόμη ότι οι εκτιμήσεις είναι πλήρως αντιπροσωπευτικές του πληθυσμού.

**Παράδειγμα (από το dataset)**

Στο δείγμα, η Κέρκυρα αντιπροσωπεύει το **4,16% (250/6.006)** των συμμετεχόντων, ενώ στον πληθυσμό εκτιμάμε πως αντιστοιχεί σε μόλις **1,99%**.

Άρα, η Κέρκυρα έχει **υπερεκπροσωπηθεί κατά περίπου 2,09 φορές** ( $4,16\% / 1,99\%$ ) στο δείγμα.

Στη στάθμιση, κάθε συμμετέχοντας από την Κέρκυρα θα λάβει βάρος περίπου **1/2,09** (0,478469), ώστε η συμβολή του στις εκτιμήσεις να μειωθεί και να αντικατοπτρίζει τη σωστή πληθυσμιακή αναλογία

### **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
<b>Σύνολο</b>	<b>48,6</b>	<b>51,4</b>

### **Απάντηση**

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

**Lab 4. Survey Design και Στάθμιση  
στην Ανάλυση Επιδημιολογικών Ερευνών**

Number of strata = 33  
Number of PSUs = 563

Number of obs = 5993  
Population size = 10813129  
Design df = 530

Age group	Gender		Total
	Male	Female	
18-29	5.061 298	5.584 336	10.65 634
30-39	5.918 350	7.853 469	13.77 819
40-49	6.788 406	10.33 614	17.12 1020
50-59	6.823 425	11.47 688	18.29 1113
60-69	7.997 484	10.34 627	18.34 1111
70+	9.777 583	12.05 713	21.83 1296
Total	42.36 2546	57.64 3447	100 5993

Key: cell percentages  
number of observations

Pearson:

Uncorrected chi2(5) = 25.6045  
Design-based F(4.96, 2629.62) = 4.9013 P = 0.0002

Note: Strata with single sampling unit centered at overall mean.

Από τη σύγκριση των ποσοστών του δείγματος με τα απογραφικά στοιχεία της ΕΛΣΤΑΤ (2011) παρατηρούμε ότι:

- Οι νεότερες ηλικιακές ομάδες (18–39 ετών) εμφανίζονται λιγότερο στο δείγμα σε σχέση με τον πληθυσμό.
- Οι μεγαλύτερες ηλικίες (50+) εμφανίζονται περισσότερο στο δείγμα.
- Οι γυναίκες αποτελούν μεγαλύτερο ποσοστό στο δείγμα (περίπου 57–58%) σε σχέση με την απογραφή (51.4%), ενώ οι άνδρες αντίστροφα μικρότερο ποσοστό (42–43% έναντι 48.6%).

Επομένως, το αρχικό –ακόμη και το design-weighted– δείγμα δεν είναι πλήρως αντιπροσωπευτικό του πληθυσμού ως προς την ηλικία και το φύλο.



Οι αποκλίσεις μπορεί να οφείλονται:

- σε **διαφορετικές πιθανότητες επιλογής** μεταξύ στρωμάτων.
- σε **άνισα ποσοστά απόκρισης**: οι νεότεροι εργαζόμενοι είναι συχνά λιγότερο διαθέσιμοι. Επιπλέον, οι γυναίκες και οι μεγαλύτεροι ηλικιακά εμφανίζουν μεγαλύτερη προθυμία συμμετοχής, πιθανώς λόγω αυξημένης ευαισθητοποίησης σε θέματα υγείας.
- Κοινωνικοοικονομικοί παράγοντες ενδέχεται να επηρεάζουν και τους δύο μηχανισμούς (τόσο το design, όσο και τη συμμετοχή)

Το αποτέλεσμα είναι ένα δείγμα που υπερεκπροσωπεί γυναίκες και άτομα μεγαλύτερης ηλικίας και υποεκπροσωπεί άνδρες και άτομα νεότερης ηλικίας.

Η διαδικασία **post-stratification** διορθώνει αυτές τις ανισορροπίες, προσαρμόζοντας τα βάρη ώστε οι σταθμισμένες κατανομές να ταυτίζονται με τις απογραφικές (ΕΛΣΤΑΤ 2011).

Εάν δεν εφαρμοζόταν στάθμιση, οι εκτιμήσεις επιπολασμού θα ήταν μεροληπτικές. Για παράδειγμα:

- (1) θα υπερεκτιμούνταν η **επίπτωση ή ο επιπολασμός ηλικιο-εξαρτώμενων εκβάσεων** (όπως υπέρταση, διαβήτης, αρθρίτιδα), και
- (2) θα υπήρχε **μετατόπιση στην εκτίμηση εκβάσεων με διαφοροποίηση φύλου** (όπως οστεοπόρωση).

#### 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.

**Απάντηση**

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

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

Result	# of obs.	
not matched	0	
matched	5,993	( _merge==3)

```
. rename _merge _merge_census
```

Η εντολή *merge* συγχωνεύει τις παρατηρήσεις του δείγματος με τα στοιχεία της απογραφής ΕΛΣΤΑΤ 2011 με βάση τα τρία κλειδιά (*area*, *gender*, *age\_gr*).

Όλες οι παρατηρήσεις χωρίς ορισμένες τιμές σε αυτά τα πεδία απομακρύνονται, ώστε η αντιστοίχιση να είναι ακριβής.

Ο έλεγχος της μεταβλητής *\_merge\_census* δείχνει αν όλες οι εγγραφές του δείγματος βρήκαν αντίστοιχη κατηγορία στην απογραφή.

Στην EMENO, όλες σχεδόν οι παρατηρήσεις αντιστοιχούν επιτυχώς ( $\approx 100\%$ ), κάτι που επιβεβαιώνει τη συμβατότητα του κωδικοποιημένου dataset με την απογραφή.

**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*;

**Απάντηση**

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

Η μεταβλητή **total\_we** εκφράζει το **εκτιμώμενο πληθυσμιακό μέγεθος** (συχνότητα) κάθε κελιού της ταξινόμησης *περιοχή × φύλο × ηλικιακή ομάδα*, όπως αυτό προκύπτει από τα αρχικά βάρη δειγματοληψίας (*base\_weight*).

Με άλλα λόγια, αν αθροίσουμε όλα τα *base\_weight* για τα άτομα μιας κατηγορίας (π.χ. γυναίκες 50–59 ετών στην Αττική), παίρνουμε **την εκτίμηση του πληθυσμού** που αυτή η κατηγορία αντιπροσωπεύει πριν τη μετα-στρωμάτωση.

Είναι **count-type estimate**, όχι ποσοστό.

**17.** Υπολογίστε τον προσαρμοσμένο συντελεστή μεταστρωμάτωσης (*adj\_factor*):

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

#### Απάντηση

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

Ο συντελεστής **adj\_factor** εκφράζει **πόσο πρέπει να προσαρμοστεί το αρχικό βάρος** ώστε το εκτιμώμενο πλήθος του δείγματος να ταιριάζει με την απογραφή.

Αν *adj\_factor* > 1 → η ομάδα υποεκπροσωπείται στο δείγμα (τα βάρη αυξάνονται).

Αν *adj\_factor* < 1 → η ομάδα υπερεκπροσωπείται (τα βάρη μειώνονται).

**Παράδειγμα (από το dataset):** Για γυναίκες ≥70 ετών στη Λέσβο/Ρόδο:

Εκτιμώμενος πληθυσμός = 47.184,

Απογραφή = 38.675

*adj\_factor* = 38.675 / 47.184 = **0,82**.

Άρα τα βάρη αυτής της ομάδας μειώνονται κατά 18%.

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

**Απάντηση**

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

Η μεταβλητή **weight\_final** προκύπτει πολλαπλασιάζοντας το αρχικό βάρος (`base_weight`) με τον συντελεστή προσαρμογής (`adj_factor`).

Τα **τελικά βάρη** εξασφαλίζουν ότι η σταθμισμένη κατανομή του δείγματος *αναπαράγει επακριβώς* τις πληθυσμιακές κατανομές κατά περιοχή, φύλο και ηλικία (post-stratification adjustment).

Με αυτό τον τρόπο, διορθώνεται η μικρή απόκλιση του design-weighted δείγματος από τη δημογραφική δομή του πληθυσμού.

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

**Απάντηση**

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

Stratum	#Units	#Obs	#Obs per Unit		
			min	mean	max
1	125	1494	9	12.0	13
2	11	144	11	13.1	21
3	1*	8	8	8.0	8
4	15	170	2	11.3	15
5	8	96	12	12.0	12
6	16	128	8	8.0	8
7	36	436	11	12.1	16
8	10	120	12	12.0	12
9	3	24	8	8.0	8
10	13	156	12	12.0	12
11	6	72	12	12.0	12
12	11	90	8	8.2	10
13	20	246	3	12.3	18
14	6	71	11	11.8	12
15	20	160	8	8.0	8
16	20	249	11	12.4	16
17	9	85	6	9.4	13
18	30	252	5	8.4	16
19	9	107	11	11.9	12
20	5	60	12	12.0	12
21	17	140	8	8.2	12
22	5	50	2	10.0	12
23	5	56	8	11.2	12
24	18	144	8	8.0	8
25	10	120	10	12.0	13
26	10	119	8	11.9	15
27	21	165	3	7.9	9
28	22	261	4	11.9	16
29	12	143	11	11.9	12
30	33	263	7	8.0	9
31	11	132	12	12.0	12
32	8	96	12	12.0	12
33	17	136	8	8.0	8
33	563	5993	2	10.6	21

Με τη νέα δήλωση `svyset`, το Stata αναγνωρίζει τα **post-stratified weights** ως επίσημα δειγματοληπτικά βάρη.

Από αυτό το σημείο και μετά, όλες οι αναλύσεις με πρόθεμα `svy:` θα χρησιμοποιούν το πλήρως σταθμισμένο σχέδιο — λαμβάνοντας υπόψη τόσο τις αρχικές πιθανότητες επιλογής όσο και τη μεταστρωμάτωση.

**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;

#### Απάντηση

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

**Lab 4. Survey Design και Στάθμιση**  
**στην Ανάλυση Επιδημιολογικών Ερευνών**

Number of strata = 33  
Number of PSUs = 563

Number of obs = 5993  
Population size = 8926434  
Design df = 530

Age group	Gender		Total
	Male	Female	
18-29	9.147 298	8.593 336	17.74 634
30-39	9.271 350	9.049 469	18.32 819
40-49	8.751 406	8.962 614	17.71 1020
50-59	7.584 425	8.008 688	15.59 1113
60-69	6.088 484	6.617 627	12.7 1111
70+	7.707 583	10.22 713	17.93 1296
Total	48.55 2546	51.45 3447	100 5993

Key: cell percentages  
number of observations

Pearson:

Uncorrected chi2(5) = 19.4794  
Design-based F(4.53, 2402.12) = 3.2966 P = 0.0076

Note: Strata with single sampling unit centered at overall mean.

Μετά την εφαρμογή των τελικών βαρών (weight\_final), οι σταθμισμένες κατανομές ηλικίας και φύλου **ευθυγραμμίζονται σχεδόν πλήρως** με τα αντίστοιχα ποσοστά της απογραφής ΕΛΣΤΑΤ 2011. Οι μικρές αποκλίσεις που τυχόν παραμένουν οφείλονται σε στρογγυλοποιήσεις και μη απόκριση. Αυτό δείχνει ότι η διαδικασία μεταστροφής **διόρθωσε αποτελεσματικά** τις μικρές ανισορροπίες του δείγματος και ότι τα τελικά βάρη αποδίδουν **πλήρως αντιπροσωπευτικές εκτιμήσεις**.