

Δειγματοληψία κατά συστάδες

Η ιδέα της δειγματοληψίας κατά Συστάδες (ΔκΣ) (cluster sampling):

1. Οι μονάδες του πληθυσμού χωρίζονται σε ομάδες ξένες μεταξύ τους, τις **συστάδες**.
 2. Επιλέγεται ΑΤΔ συστάδων.
 3. Ακολουθούμε μία από τις δύο προσεγγίσεις
 - α. γίνεται μέτρηση του χαρακτηριστικού U σε όλες τις μονάδες κάθε επιλεγμένης συστάδας (**μονοσταδιακή δειγματοληψία κατά συστάδες**), ή
 - β. επιλέγεται ΑΤΔ μονάδων από κάθε επιλεγμένη συστάδα και γίνεται μέτρηση του χαρακτηριστικού U μόνο σε αυτές τις επιλεγμένες μονάδες του δείγματος (**δισταδιακή δειγματοληψία κατά συστάδες**)
- **συστάδα**: πρωτογενής μονάδα
 - **μονάδα** (όπου γίνεται η μέτρηση του U): δευτερογενής μονάδα

Δειγματοληψία κατά συστάδες

Παραδείγματα:

1. συστάδα: σχολείο
μονάδα: μαθητής
2. συστάδα: ιχθυοτροφείο
μονάδα: ψάρι
3. συστάδα: πολυκατοικία
μονάδα: ένοικος

Πλεονεκτήματα:

Η ΔκΣ μπορεί να εφαρμοστεί

- παρέχοντας ευκολία στη λήψη δείγματος λόγω κοντινής απόστασης μεταξύ των στοιχείων (μονάδων) του πληθυσμού
- συνήθως όταν δεν υπάρχει διαθέσιμη λίστα για τον πληθυσμό, ή το πλήθος των απλών στοιχείων του πληθυσμού είναι άγνωστο, οπότε ΑΤΔ και ΣΤΔ δεν μπορούν να εφαρμοστούν

Μονοσταδιακή δειγματοληψία κατά συστάδες: Συμβολισμός

Πληθυσμός:

1. N : αριθμός συστάδων
2. m_j : αριθμός μονάδων της j συστάδας, $j = 1, \dots, N$
3. τ_j : ολική τιμή του U στην j συστάδα, $j = 1, \dots, N$
4. $M = \sum_{j=1}^N m_j$: ολικός αριθμός μονάδων του πληθυσμού
5. $\mu = \frac{\sum_{j=1}^N \tau_j}{M} = \frac{\sum_{j=1}^N \tau_j}{\sum_{j=1}^N m_j}$: μέση τιμή του U στον πληθυσμό
6. $\tau = \sum_{j=1}^N \tau_j$: ολική τιμή του U στον πληθυσμό

Δείγμα:

1. n : μέγεθος δείγματος συστάδων
2. m_i^* : αριθμός μονάδων της i συστάδας του δείγματος, $i = 1, \dots, n$
3. τ_i^* : ολική τιμή του U στην i συστάδα του δείγματος, $i = 1, \dots, n$

Μονοσταδιακή δειγματοληψία κατά συστάδες

Εκτίμηση μέσης τιμής μ :

$$\bar{X}_c = \frac{\sum_{i=1}^n \tau_i^*}{\sum_{i=1}^n m_i^*}$$

Παρατηρήσεις:

1. Το $\sum_i m_i^*$ είναι **τυχαία μεταβλητή**, αφού εξαρτάται από το μέγεθος των επιλεγμένων συστάδων κάθε φορά. Για το λόγο αυτό, αν και ο $\bar{X}_c$ στην ουσία εκφράζει «την ολική τιμή του χαρακτηριστικού στο δείγμα διά το μέγεθος του δείγματος», δεν είναι ο «κλασικός» δειγματικός μέσος όπου το μέγεθος του δείγματος είναι σταθερό.
2. Ένας εναλλακτικός τρόπος για να προκύψει ο $\bar{X}_c$ είναι να θεωρήσουμε κάθε συστάδα ως «μονάδα» όπου μετράμε τα (υποθετικά) χαρακτηριστικά

- U' με τιμές $\tau_1, \dots, \tau_N$, δηλαδή $U'_j = \tau_j, j = 1, \dots, N$
- V' με τιμές $m_1, \dots, m_N$, δηλαδή $V'_j = m_j, j = 1, \dots, N$

και η εκτίμηση του μ ανάγεται στην εκτίμηση του λόγου

$$r' = \frac{\mu_{U'}}{\mu_{V'}} = \frac{\tau_{U'}}{\tau_{V'}} = \frac{\sum_j \tau_j}{\sum_j m_j} = \mu.$$

Μονοσταδιακή δειγματοληψία κατά συστάδες

Ιδιότητες του $\bar{X}_c$:

Για $n \rightarrow N$ είναι

1. $E\bar{X}_c \approx \mu$
2. $\text{Var}(\bar{X}_c) \approx \frac{N^2}{M^2} (1-f) \frac{1}{n} \frac{1}{N-1} \sum_{j=1}^N (\tau_j - \mu m_j)^2$ (απαιτεί γνώση του M)
3. $\hat{\text{Var}}(\bar{X}_c) \approx \frac{N^2}{M^2} (1-f) \frac{1}{n} \frac{1}{n-1} \sum_{i=1}^n (\tau_i^* - \bar{X}_c m_i^*)^2$ (απαιτεί γνώση του M)

Μονοσταδιακή δειγματοληψία κατά συστάδες

Εκτίμηση ολικής τιμής τ :

- Αν M γνωστό τότε

$$\hat{\tau}_c = M\bar{X}_c$$

- Αν M άγνωστο τότε

$$\hat{\tau} = N \frac{1}{n} \sum_{i=1}^n \tau_i^*$$

(επειδή το πρόβλημα μπορεί να αναχθεί σε πρόβλημα εκτίμησης ολικής τιμής σε ΑΤΔ για το (υποθετικό) χαρακτηριστικό

$$U' \text{ με τιμές } U'_j = \tau_j, j = 1, \dots, N,$$

$$\text{δηλαδή } \tau = \sum_{j=1}^N \tau_j = \tau_{U'} = N\mu_{U'} \text{ εκτιμώντας το } \mu_{U'} \text{ με } \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{n} \sum_{i=1}^n \tau_i^*)$$

Μονοσταδιακή δειγματοληψία κατά συστάδες

Ιδιότητες των $\hat{\tau}_c$ και $\hat{\tau}$:

1. Οι ιδιότητες του $\hat{\tau}_c$ προκύπτουν άμεσα από αυτές του $\bar{X}_c$.
2. Ο $\hat{\tau}$ είναι ΑΕ του τ , αφού προέρχεται από ΑΤΔ.
3. $\text{Var}(\hat{\tau}) = \text{Var}(N\bar{X}) = N^2\text{Var}(\bar{X}) = N^2(1-f)\frac{\sigma_0^2}{n}$, όπου $\sigma_0^2 = \frac{1}{N-1} \sum_{j=1}^N (\tau_j - \frac{1}{N} \sum_{j=1}^N \tau_j)^2$
4. $\hat{\text{Var}}(\hat{\tau}) = N^2(1-f)\frac{S^2}{n}$, όπου $S^2 = \frac{1}{n-1} \sum_{i=1}^n (\tau_i^* - \frac{1}{n} \sum_{i=1}^n \tau_i^*)^2$

Παρατήρηση:

Για M γνωστό, προκύπτει ο εναλλακτικός εκτιμητής του μ (ο οποίος είναι ΑΕ διότι προκύπτει από ΑΤΔ)

$$\hat{\mu} = \frac{\hat{\tau}}{M} = \frac{N \frac{1}{n} \sum_{i=1}^n \tau_i^*}{M}$$

Μονοσταδιακή δειγματοληψία κατά συστάδες

Συνοπτικά

εκτίμηση μέσης τιμής μ

$$\bar{X}_c = \frac{\sum_{i=1}^n \tau_i^*}{\sum_{i=1}^n m_i^*} \quad (\text{δεν είναι AE})$$

$$\hat{\mu} = \frac{N \frac{1}{n} \sum_{i=1}^n \tau_i^*}{M} \quad (\text{AE})$$

εκτίμηση ολικής τιμής τ

$$\hat{\tau}_c = M \bar{X}_c \quad (M \text{ γνωστό, δεν είναι AE})$$

$$\hat{\tau} = N \frac{1}{n} \sum_{i=1}^n \tau_i^* \quad (\text{AE})$$

Μονοσταδιακή δειγματοληψία κατά συστάδες

Πρόταση 1

Αν $m_j = m, \forall j = 1, \dots, N$, τότε $\bar{X}_c = \hat{\mu}$ και $\hat{\tau}_c = \hat{\tau}$.

Απόδειξη.

Δισταδιακή δειγματοληψία κατά συστάδες

Στην δισταδιακή ΔκΣ, για την i συστάδα του δείγματος

- m_i^* είναι γνωστό
- τ_i^* είναι **άγνωστο**, αφού δεν καταγράφουμε την ολική τιμή του U σε όλη τη συστάδα του δείγματος, αλλά μόνο σε μέρος (δείγμα) αυτής

Εκτίμηση των μ και τ :

Προκύπτουν από τους τύπους της μονοσταδιακής ΔκΣ αντικαθιστώντας το άγνωστο τ_i^* με τον εκτιμητή (ολικής τιμής σε ΑΤΔ)

$$\hat{\tau}_i^* = m_i^* \bar{X}_i$$

Δισταδιακή δειγματοληψία κατά συστάδες

Συνοπτικά

εκτίμηση μέσης τιμής μ

$$\bar{X}_c = \frac{\sum_{i=1}^n m_i^* \bar{X}_i}{\sum_{i=1}^n m_i^*} \quad (\text{δεν είναι AE})$$

$$\hat{\mu} = \frac{N \frac{1}{n} \sum_{i=1}^n m_i^* \bar{X}_i}{M} \quad (\text{AE})$$

εκτίμηση ολικής τιμής τ

$$\hat{\tau}_c = M \bar{X}_c \quad (M \text{ γνωστό, δεν είναι AE})$$

$$\hat{\tau} = N \frac{1}{n} \sum_{i=1}^n m_i^* \bar{X}_i \quad (\text{AE})$$

Στρωματοποιημένη δειγματοληψία κατά συστάδες

Η ιδέα της στρωματοποιημένης ΔκΣ:

1. Οι μονάδες του πληθυσμού χωρίζονται σε **στρώματα** (στρωματοποίηση).
2. Ομαδοποίηση των μονάδων κάθε στρώματος σε **συστάδες**.
3. Από κάθε στρώμα επιλέγεται ΑΤΔ συστάδων.
4. Ακολουθούμε μία από τις δύο προσεγγίσεις
 - α. γίνεται μέτρηση του χαρακτηριστικού U σε όλες τις μονάδες κάθε επιλεγμένης συστάδας (**στρωματοποιημένη μονοσταδιακή δειγματοληψία κατά συστάδες**), ή
 - β. επιλέγεται ΑΤΔ μονάδων από κάθε επιλεγμένη συστάδα και γίνεται μέτρηση του χαρακτηριστικού U μόνο σε αυτές τις επιλεγμένες μονάδες του δείγματος (**στρωματοποιημένη δισταδιακή δειγματοληψία κατά συστάδες**)

Στρωματοποιημένη μονοσταδιακή ΔκΣ: Συμβολισμός

Πληθυσμός:

1. N_ℓ : αριθμός συστάδων του ℓ στρώματος, $\ell = 1, \dots, L$
2. $N = \sum_{\ell=1}^L N_\ell$: συνολικός αριθμός συστάδων
3. $\tau_{\ell j}$: ολική τιμή του U της j συστάδας του ℓ στρώματος, $\ell = 1, \dots, L$,
 $j = 1, \dots, N_\ell$
4. $m_{\ell j}$: αριθμός μονάδων της j συστάδας του ℓ στρώματος, $\ell = 1, \dots, L$,
 $j = 1, \dots, N_\ell$
5. $M = \sum_{\ell=1}^L \sum_{j=1}^{N_\ell} m_{\ell j}$: συνολικός αριθμός μονάδων του πληθυσμού

Δείγμα:

1. n_ℓ : μέγεθος δείγματος συστάδων από το ℓ στρώμα, $\ell = 1, \dots, L$
2. $n = \sum_{\ell=1}^L n_\ell$: συνολικό μέγεθος δείγματος συστάδων
3. $\tau_{\ell i}^*$: ολική τιμή του U της i επιλεγμένης συστάδας του ℓ στρώματος στο δείγμα, $\ell = 1, \dots, L$, $i = 1, \dots, n_\ell$
4. $m_{\ell i}^*$: αριθμός μονάδων της i επιλεγμένης συστάδας του ℓ στρώματος στο δείγμα, $\ell = 1, \dots, L$, $i = 1, \dots, n_\ell$

Στρωματοποιημένη μονοσταδιακή ΔκΣ

Εκτίμηση του μ :

$$\text{Επειδή } \mu = \frac{\tau}{M} = \frac{\sum_{\ell=1}^L \sum_{j=1}^{N_{\ell}} \tau_{\ell j}}{\sum_{\ell=1}^L \sum_{j=1}^{N_{\ell}} m_{\ell j}} = \frac{\sum_{\ell=1}^L N_{\ell} \bar{\tau}_{\ell}}{\sum_{\ell=1}^L N_{\ell} \bar{m}_{\ell}} = \frac{\sum_{\ell=1}^L W_{\ell} \bar{\tau}_{\ell}}{\sum_{\ell=1}^L W_{\ell} \bar{m}_{\ell}}$$

- $\bar{\tau}_{\ell}$: μέση τιμή ολικής τιμής του χαρακτηριστικού ανά συστάδα στο ℓ στρώμα στον πληθυσμό
- $\bar{m}_{\ell}$: μέση τιμή του πλήθους μονάδων ανά συστάδα στο ℓ στρώμα στον πληθυσμό

παίρνουμε ως εκτιμητή

$$\hat{\mu}_c = \frac{\sum_{\ell=1}^L W_{\ell} \bar{\tau}_{\ell}^*}{\sum_{\ell=1}^L W_{\ell} \bar{m}_{\ell}^*}$$

όπου

$$\bar{\tau}_{\ell}^* = \frac{1}{n_{\ell}} \sum_{i=1}^{n_{\ell}} \tau_{\ell i}^*, \quad \bar{m}_{\ell}^* = \frac{1}{n_{\ell}} \sum_{i=1}^{n_{\ell}} m_{\ell i}^*$$

Στρωματοποιημένη μονοσταδιακή ΔκΣ

Παρατήρηση:

Ένας εναλλακτικός τρόπος για να προκύψει ο $\hat{\mu}_c$ είναι να θεωρήσουμε κάθε συστάδα ως «μονάδα» όπου μετράμε τα (υποθετικά) χαρακτηριστικά

- U' (κύριο) με τιμές $U'_{\ell j} = \tau_{\ell j}, \ell = 1, \dots, L, j = 1, \dots, N_\ell$
- V' (βοηθητικό) με τιμές $V'_{\ell j} = m_{\ell j}, \ell = 1, \dots, L, j = 1, \dots, N_\ell$

και η εκτίμηση του μ ανάγεται στην εκτίμηση του λόγου

$$r' = \frac{\mu_{U'}}{\mu_{V'}} = \frac{\tau_{U'}}{\tau_{V'}} = \frac{\tau}{M} = \mu \text{ σε ΣΤΔ,}$$

οπότε παίρνουμε

$$\hat{\mu}_c = \hat{r}' = \frac{\bar{X}_{st}}{\bar{Y}_{st}} = \frac{\sum_{\ell=1}^L W_\ell \bar{X}_\ell}{\sum_{\ell=1}^L W_\ell \bar{Y}_\ell} = \frac{\sum_{\ell=1}^L W_\ell \bar{\tau}_\ell^*}{\sum_{\ell=1}^L W_\ell \bar{m}_\ell^*}$$

αφού

$$\begin{aligned} \bar{X}_\ell &= \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} X_{\ell i} = \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} \tau_{\ell i}^* & \text{και} \\ \bar{Y}_\ell &= \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} Y_{\ell i} = \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} m_{\ell i}^* \end{aligned}$$

Στρωματοποιημένη μονοσταδιακή $\Delta\kappa\Sigma$

Εκτίμηση ολικής τιμής τ : $\tau = \sum_{\ell=1}^L \sum_{j=1}^{N_{\ell}} \tau_{\ell j} = \sum_{\ell=1}^L N_{\ell} \bar{\tau}_{\ell}$

- Αν M γνωστό τότε

$$\hat{\tau}_c = M \hat{\mu}_c$$

- Αν M άγνωστο τότε

$$\hat{\tau} = \sum_{\ell=1}^L N_{\ell} \bar{\tau}_{\ell}^*$$

Παρατήρηση:

Για M γνωστό, προκύπτει ο εναλλακτικός εκτιμητής του μ

$$\hat{\mu} = \frac{\hat{\tau}}{M} = \frac{\sum_{\ell=1}^L N_{\ell} \bar{\tau}_{\ell}^*}{M}$$

Στρωματοποιημένη μονοσταδιακή ΔκΣ

Συνοπτικά

εκτίμηση μέσης τιμής μ

$$\hat{\mu}_c = \frac{\sum_{\ell=1}^L W_{\ell} \bar{\tau}_{\ell}^*}{\sum_{\ell=1}^L W_{\ell} \bar{m}_{\ell}^*} \quad (\text{δεν είναι AE})$$

$$\hat{\mu} = \frac{\sum_{\ell=1}^L N_{\ell} \bar{\tau}_{\ell}^*}{M} \quad (\text{AE})$$

εκτίμηση ολικής τιμής τ

$$\hat{\tau}_c = M \hat{\mu}_c \quad (M \text{ γνωστό, δεν είναι AE})$$

$$\hat{\tau} = \sum_{\ell=1}^L N_{\ell} \bar{\tau}_{\ell}^* \quad (M \text{ άγνωστο, AE})$$

Στρωματοποιημένη μονοσταδιακή ΔκΣ

Πρόταση 2

Αν $m_{\ell j} = m_{\ell}, \forall j = 1, \dots, N_{\ell}$, τότε $\hat{\mu}_c = \hat{\mu}$ και $\hat{\tau}_c = \hat{\tau}$.

Απόδειξη.

Στρωματοποιημένη δισταδιακή ΔκΣ

Στη στρωματοποιημένη δισταδιακή ΔκΣ, για την i συστάδα του δείγματος

- $m_{\ell i}^*$ είναι γνωστό
- $\tau_{\ell i}^*$ είναι **άγνωστο**, αφού δεν καταγράφουμε την ολική τιμή του U σε όλη την i συστάδα του ℓ στρώματος του δείγματος, αλλά μόνο σε μέρος (δείγμα) αυτής, οπότε και $\bar{\tau}_{\ell}^* = \frac{1}{n_{\ell}} \sum_{i=1}^{n_{\ell}} \tau_{\ell i}^*$ είναι **άγνωστο**.

Εκτίμηση των μ και τ :

Προκύπτουν από τους τύπους της δισταδιακής ΔκΣ αντικαθιστώντας το άγνωστο $\bar{\tau}_{\ell}^*$ με τον εκτιμητή

$$\hat{\tau}_{\ell}^* = \frac{1}{n_{\ell}} \sum_{i=1}^{n_{\ell}} m_{\ell i}^* \bar{X}_{\ell i}$$

όπου $\bar{X}_{\ell i}$ είναι ο δειγματικός μέσος του U στην i επιλεγμένη συστάδα του ℓ στρώματος.

Σημειώνεται ότι το $\tau_{\ell i}^*$ εκτιμάται με τον εκτιμητή (ολικής τιμής σε ΑΤΔ)

$$\hat{\tau}_{\ell i}^* = m_{\ell i}^* \bar{X}_{\ell i}$$

Στρωματοποιημένη δισταδιακή ΔκΣ

Συνοπτικά

εκτίμηση μέσης τιμής μ

$$\hat{\mu}_c = \frac{\sum_{\ell=1}^L W_{\ell} \hat{\tau}_{\ell}^*}{\sum_{\ell=1}^L W_{\ell} \bar{m}_{\ell}^*} \quad (\text{δεν είναι AE})$$

$$\hat{\mu} = \frac{\sum_{\ell=1}^L N_{\ell} \hat{\tau}_{\ell}^*}{M} \quad (\text{AE})$$

εκτίμηση ολικής τιμής τ

$$\hat{\tau}_c = M \hat{\mu}_c \quad (M \text{ γνωστό, δεν είναι AE})$$

$$\hat{\tau} = \sum_{\ell=1}^L N_{\ell} \hat{\tau}_{\ell}^* \quad (\text{AE})$$

όπου

$$\hat{\tau}_{\ell}^* = \frac{1}{n_{\ell}} \sum_{i=1}^{n_{\ell}} m_{\ell i}^* \bar{X}_{\ell i}$$

Πολυσταδιακή ΔκΣ

Η ιδέα της πολυσταδιακής ΔκΣ:

Η βασική ιδέα της πολυσταδιακής ΔκΣ είναι η γενίκευση της δισταδιακής ΔκΣ εφαρμόζοντας ΑΤΔ σε περισσότερα στάδια, μέχρι και το τελευταίο στάδιο όπου παίρνουμε ΑΤΔ των (στοιχειωδών) μονάδων του πληθυσμού. Το πλήθος των σταδίων όπου εφαρμόζουμε ΑΤΔ χαρακτηρίζει στην ουσία και το ακριβές είδος της πολυσταδιακής ΔκΣ, δηλ. τρισταδιακή, τετρασταδιακή, κτλ.

Τρισταδιακή ΔκΣ:

- Ο πληθυσμός είναι χωρισμένος σε συστάδες (πρωτογενείς δειγματοληπτικές μονάδες) και λαμβάνουμε ΑΤΔ πρωτογενών μονάδων – 1ο στάδιο.
- Κάθε πρωτογενής δειγματοληπτική μονάδα (συστάδα) είναι χωρισμένη περαιτέρω σε δευτερογενείς δειγματοληπτικές μονάδες (συστάδες), και από κάθε επιλεγμένη πρωτογενή μονάδα λαμβάνουμε ΑΤΔ δευτερογενών μονάδων – 2ο στάδιο.
- Κάθε δευτερογενής δειγματοληπτική μονάδα αποτελείται από τις (στοιχειώδεις) μονάδες του πληθυσμού, και από κάθε επιλεγμένη δευτερογενή μονάδα λαμβάνουμε ΑΤΔ (στοιχειδών) μονάδων του πληθυσμού – 3ο στάδιο.

Πολυσταδιακή ΔκΣ

Παράδειγμα τρισταδιακής ΔκΣ:

Ας υποθέσουμε ότι θέλουμε να μελετήσουμε το ποσοστό των μαθητών Λυκείου σε ολόκληρη την Ελλάδα οι οποίοι καπνίζουν, κάνοντας **τρισταδιακή ΔκΣ**.

- Χωρίζουμε την Ελλάδα σε **εκπαιδευτικές περιφέρειες** (πρωτογενείς μονάδες).
- Κάθε εκπαιδευτική περιφέρεια χωρίζεται σε **Λύκεια** (δευτερογενείς μονάδες).
- Κάθε Λύκειο αποτελείται από τους **μαθητές**, τις (στοιχειώδεις) μονάδες του πληθυσμού (τριτογενείς μονάδες).

Επομένως, παίρνουμε

1. ΑΤΔ εκπαιδευτικών περιφερειών – **1ο στάδιο**
2. ΑΤΔ Λυκείων σε κάθε επιλεγμένη εκπαιδευτική περιφέρεια του πρώτου σταδίου – **2ο στάδιο**
3. ΑΤΔ μαθητών σε κάθε επιλεγμένο Λύκειο του δεύτερου σταδίου – **3ο στάδιο**