# Προχωρημένες Μέθοδοι Δειγματοληψίας

Τ. Μερκούρης

Οικονομικό Πανεπιστήμιο Αθηνών

Προπτυχιακό Πρόγραμμα Στατιστικής

# Περιεχόμενα

- Βασική στατιστική θεωρία πεπερασμένων πληθυσμών
  - Πληθυσμοί, υποπληθυσμοί, μεταβλητές, παράμετροι
  - Τυχαία δειγματοληψία, πιθανότητες επιλογής μονάδων
  - Δειγματοληψία με άνισες πιθανότητες επιλογής
  - Δειγματικά βάρη, αυτοβαρής και μη αυτοβαρής δειγματοληψία
  - Εκτίμηση παραμέτρων πεπερασμένων πληθυσμών
  - Υπολογισμός διακύμανσης εκτιμητών
  - Εκτιμητική υποπληθυσμών
  - Εκτίμηση συνάρτησης κατανομής
  - Γραφική παράσταση δεδομένων δειγματοληψίας

# Περιεχόμενα

- Χρήση βοηθητικών πληροφοριών στην εκτιμητική
  - Μέθοδος γενικευμένης παλινδρόμησης (εκτιμητής λόγου, εκτιμητής παλινδρόμησης, μεταστρωματικός εκτιμητής)
  - Calibration.
- Εκτίμηση διακύμανσης σε περιπλεγμένες δειγματοληψίες. Μέθοδοι επαναληπτικής δειγματοληψίας (τυχαίες ομάδες, jackknife, bootstrap).
- Διαχείριση μή δειγματοληπτικών σφαλμάτων
  Μέθοδοι ρύθμισης για μη απόκριση, imputation.

## Πληθυσμοί, υποπληθυσμοί, μεταβλητές

Πεπερασμένος πληθυσμός $U$ μεγέθους $N$:
είναι σύνολο $N$ διακριτών μονάδων

$$U = \{1, \ldots, N\},$$

όπως, άτομα, νοικοκυριά, επιχειρήσεις, σχολεία κλπ.

Σε πεπερασμένο πληθυσμό $U$

- ορίζονται υποπληθυσμοί $U_d \subset U$, ως υποσύνολα
- δεν υπάρχει εγγενής τυχαιότητα

# Πληθυσμοί, υποπληθυσμοί, μεταβλητές

Ένα ερευνώμενο χαρακτηριστικό του πληθυσμού ορίζει μια μεταβλητή (study/target variable) $y$, με τιμή $y_i$ για την μονάδα $i \in U$.

Η μεταβλητή $y$ είναι **μή τυχαία** — δεν ορίζεται συνάρτηση κατανομής πιθανοτήτων. Οι τιμές $y_i$, $i \in U$, είναι άγνωστες αλλά σταθερές.

Στις δειγματοληπτικές έρευνες ορίζονται, σχεδόν πάντα, πολλές μεταβλητές $y$, $z$, κλπ, (συνεχείς ή κατηγορικές),
π.χ. εισόδημα, απασχόληση, βαθμίδα εκπαίδευσης, κλπ.

# Παράμετροι πεπερασμένων πληθυσμών

*Παράμετρος* $\theta$ του πληθυσμού ορίζεται ως συνάρτηση των τιμών μιας μεταβλητής $y$

$$\theta = \theta(y_1, \ldots, y_N)$$

ή πολλών μεταβλητών, π.χ., $y$ και $z$

$$\theta = \theta(y_1, \ldots, y_N, z_1, \ldots, z_N)$$

# Παράμετροι πεπερασμένων πληθυσμών

Οι κυριότερες παράμετροι είναι:

*Ο ολικός (population total)*

$$Y = \sum_{i=1}^{N} y_i \quad \text{ή} \quad \sum_{U} y_i$$

Π.χ., ολικός αριθμός εργαζομένων, ολική παραγωγή ελαιολάδου

Σε μή πεπερασμένους πληθυσμούς δεν ορίζονται ολικοί

$$Y = \sum_{i=1}^{N} y_i = N, \quad \text{όταν } y_i = 1, i \in U$$

# Παράμετροι πεπερασμένων πληθυσμών

*Ο Μέσος* (*mean*)

$$\bar{Y} = \frac{Y}{N} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

Π.χ., μέσο ατομικό εισόδημα

*Ο λόγος ολικών* (*ratio*)

$$R = \frac{Y}{Z} = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} z_i}$$

Π.χ., ολική παραγωγή ελαιολάδου προς ολική καλλιεργήσιμη έκταση

*Ποσοστό* (*proportion*)

$$P = \frac{N_d}{N} = \frac{1}{N}\sum_{i=1}^{N} y_i, \qquad y_i = \begin{cases} 1, & i \in U_d \\ 0, & i \notin U_d \end{cases}$$

Π.χ., ποσοστό ανέργων, ποσοστό καπνιστών.

# Παράμετροι πεπερασμένων πληθυσμών

*Διακύμανση του y* (*population variance*)

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} y_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} y_i\right)^2\right]$$

Το $S^2$ χρησιμοποιείται στον σχεδιασμό δειγματοληψίας και στον υπολογισμό τυπικού σφάλματος εκτιμητών

Τυπική απόκλιση (standard deviation)

$$S = \sqrt{S^2}$$

*Συντελεστής μεταβλητότητας* (*coefficient of variation*, cv)

$$cv = \frac{S}{\bar{Y}}$$

## Τυχαίο δείγμα (random/probability sample)

Ένα τυχαίο δείγμα είναι ένα υποσύνολο του πληθυσμού $s \subset U$, μεγέθους $n$, που επιλέγεται έτσι ώστε

- Το σύνολο $S = \{s_1, \ldots, s_M\}$ όλων των δυνατών διακριτών δειγμάτων $s$ είναι καλά ορισμένο
- Μία γνωστή πιθανότητα επιλογής $p(s)$ αντιστοιχείται σε κάθε δείγμα $s \in S$
- Κάθε μονάδα $i \in U$ έχει μία μή-μηδενική πιθανότητα επιλογής στο δείγμα $s$
- Κάθε δείγμα $s$ επιλέγεται με την πιθανότητα $p(s)$, με συγκεκριμένο μηχανισμό τυχαιότητας που ονομάζεται *σχέδιο* ή *τεχνική δειγματοληψίας*

Η συνάρτηση $p(s)$ ορίζει μία κατανομή πιθανότητας στο $S$

$$p(s) \geq 0, \quad \sum_{s \in S} p(s) = 1$$

Η πιθανότητα $p(s)$ καθορίζει την πιθανότητα επιλογής κάθε μονάδας $i \in U$, καθώς και τις στατιστικές ιδιότητες των εκτιμητών παραμέτρων.

## Πιθανότητες επιλογής των μονάδων (Inclusion probabilities)

Η πιθανότητα $P(i \in s)$ να περιληφθεί η μονάδα $i$ σέ ένα δείγμα $s$ συμβολίζεται με $\pi_i$ και δίνεται απο την σχέση

$$\pi_i = \sum_{s \ni i} p(s), \quad i = 1, \ldots, N$$

Παράδειγμα 1:

Έστω πληθυσμός $U = \{1, 2, 3, 4\}$ και δείγμα μεγέθους $n = 2$.

Όλα τα δυνατά δείγματα είναι

$$s_1 = \{1, 2\}, s_2 = \{1, 3\}, s_3 = \{1, 4\}, s_4 = \{2, 3\}, s_5 = \{2, 4\}, s_6 = \{3, 4\}$$

Τότε

$$\begin{aligned} \pi_1 &= p(s_1) + p(s_2) + p(s_3) \\ \pi_2 &= p(s_1) + p(s_4) + p(s_5) \\ \pi_3 &= p(s_2) + p(s_4) + p(s_6) \\ \pi_4 &= p(s_3) + p(s_5) + p(s_6) \end{aligned}$$

Αν $p(s_1) = p(s_2) = p(s_3) = p(s_4) = p(s_5) = p(s_6) = 1/6$, τότε $\pi_1 = \pi_2 = \pi_3 = \pi_4 = 1/2$

## Πιθανότητες επιλογής των μονάδων (Inclusion probabilities)

Η πιθανότητα $P(i \in s)$ να περιληφθεί η μονάδα $i$ σέ ένα δείγμα $s$ συμβολίζεται με $\pi_i$ και δίνεται απο την σχέση

$$\pi_i = \sum_{s \ni i} p(s), \quad i = 1, \ldots, N$$

Παράδειγμα 1:

Έστω πληθυσμός $U = \{1, 2, 3, 4\}$ και δείγμα μεγέθους $n = 2$.

Όλα τα δυνατά δείγματα είναι

$$s_1 = \{1, 2\}, s_2 = \{1, 3\}, s_3 = \{1, 4\}, s_4 = \{2, 3\}, s_5 = \{2, 4\}, s_6 = \{3, 4\}$$

Τότε

$$\begin{aligned} \pi_1 &= p(s_1) + p(s_2) + p(s_3) \\ \pi_2 &= p(s_1) + p(s_4) + p(s_5) \\ \pi_3 &= p(s_2) + p(s_4) + p(s_6) \\ \pi_4 &= p(s_3) + p(s_5) + p(s_6) \end{aligned}$$

Αν $p(s_1) = 1/3, p(s_2) = 1/6, p(s_6) = 1/2, p(s_3) = p(s_4) = p(s_5) = 0$,
$\pi_1 = 1/2, \pi_2 = 1/3, \pi_3 = 2/3, \pi_4 = 1/2$

# Πιθανότητες επιλογής των μονάδων (Inclusion probabilities)

Η διαδικασία της τυχαίας δειγματοληψίας απαιτεί $\pi_i > 0$ για κάθε $i \in U$

Η πιθανότητα $P(i \in s, j \in s)$ να περιληφθούν απο κοινού οι μονάδες $i$ και $j$ σε ένα δείγμα $s$ συμβολίζεται με $\pi_{ij}$ και δίνεται απο την σχέση

$$\pi_{ij} = \sum_{s \ni i,j} p(s), \quad i, j = 1, \ldots, N$$

Στο Παράδειγμα 1, έχουμε $\pi_{13} = p(s_2) = 1/6$

## Πιθανότητες επιλογής των μονάδων (Inclusion probabilities)

Η συμπερίληψη μιας μονάδας $i$ σε ένα τυχαίο δείγμα $s$ εκφράζεται απο την *δείκτρια* τυχαία μεταβλητή

$$I_i(s) = \begin{cases} 1, & i \in s \\ 0, & i \notin s \end{cases}$$

Για τα δείγματα του Παραδείγματος 1, $I_3(s_4) = 1$ και $I_3(s_5) = 0$

Η συμπερίληψη δύο μονάδων $i$ και $j$ στο ίδιο τυχαίο δείγμα $s$ εκφράζεται απο το γινόμενο $I_i(s)I_j(s)$

Η μεταβλητή $I_i(s)$ είναι η μόνη **τυχαία** μεταβλητή που ορίζεται για κάθε μονάδα $i \in U$

Ιδιότητες:

$$E(I_i(s)) = P(I_i(s) = 1) = \pi_i$$

$$V(I_i(s)) = \pi_i(1 - \pi_i), \quad C(I_i(s), I_j(s)) = \pi_{ij} - \pi_i\pi_j$$

# Πιθανότητες επιλογής των μονάδων (Inclusion probabilities)

Στις δειγματοληπτικές έρευνες πεπερασμένων πληθυσμών οι μονάδες του πληθυσμού μπορεί να έχουν **άνισες** πιθανότητες $\pi_i$.

Αυτό είναι αποτέλεσμα δειγματοληψίας που στηρίζεται σε γνώση της δομής του πληθυσμού για περιορισμό του δειγματοληπτικού σφάλματος στις εκτιμήσεις παραμέτρων.

Άνισες πιθανότητες επιλογής συνεπάγονται διαφορετική κατανομή του δείγματος από την κατανομή του πληθυσμού.

Παράδειγμα 2:
Απλή τυχαία δειγματοληψία απο πληθυσμό $U$, με $N = 50000$ $n = 3000$, και με ίδια πιθανότητα $\pi_i = n/N = 3/50$, $i \in U$
Η κατανομή (ιστόγραμμα) μιας μεταβλητής $y$ στο δείγμα μοιάζει με την κατανομή της στον πληθυσμό (βλέπε $1^o$ και $2^o$ γράφημα)

## Πιθανότητες επιλογής των μονάδων (Inclusion probabilities)

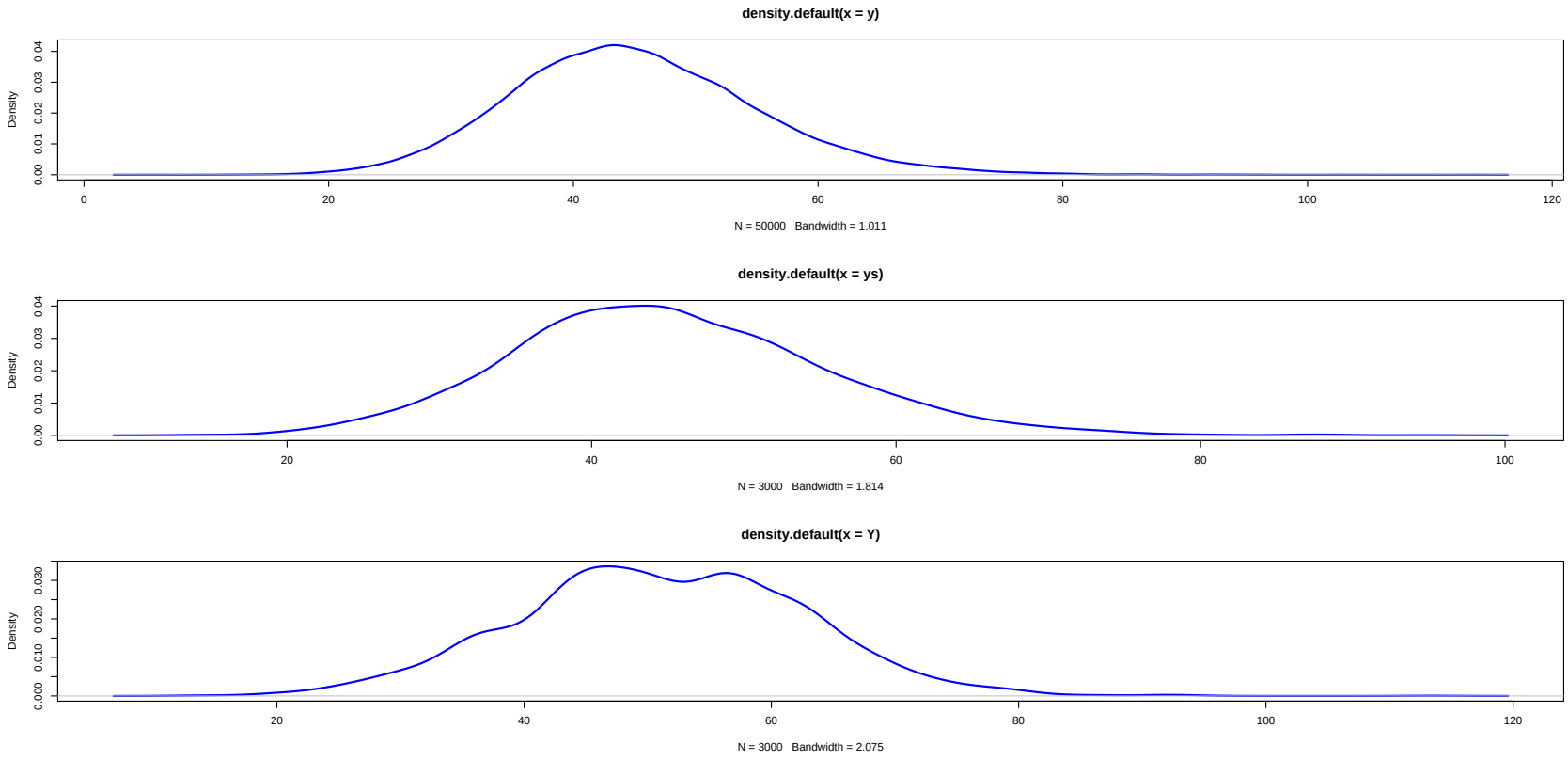
Στις δειγματοληπτικές έρευνες πεπερασμένων πληθυσμών οι μονάδες του πληθυσμού μπορεί να έχουν **άνισες** πιθανότητες $\pi_i$.

Αυτό είναι αποτέλεσμα δειγματοληψίας που στηρίζεται σε γνώση της δομής του πληθυσμού για περιορισμό του δειγματοληπτικού σφάλματος στις εκτιμήσεις παραμέτρων.

Άνισες πιθανότητες επιλογής συνεπάγονται διαφορετική κατανομή του δείγματος από την κατανομή του πληθυσμού.

Παράδειγμα 2:
Στρωματική απλή τυχαία δειγματοληψία απο πέντε στρώματα οριζόμενα απο τις τιμές του $y$ σε αύξουσα διάταξη, με στρωματικά μεγέθη $N1 = 20000, N2 = 12000, N3 = 10000, N4 = 5000, N5 = 3000$, δειγματικά μεγέθη $n_1 = \cdots = n_5 = n/5 = 600$ και με πιθανότητες $n_i/N_i, i = 1, \ldots, 5$.
Η κατανομή τη μεταβλητής $y$ στο δείγμα είναι διαφορετική απο την κατανομή της στον πληθυσμό (βλέπε $3^o$ γράφημα).

# Πιθανότητες επιλογής των μονάδων (Inclusion probabilities)



# Πιθανότητες επιλογής των μονάδων (Inclusion probabilities)

Στις δειγματοληπτικές έρευνες πεπερασμένων πληθυσμών οι μονάδες του πληθυσμού μπορεί να έχουν **άνισες** πιθανότητες $\pi_i$.

Αυτό είναι αποτέλεσμα δειγματοληψίας που στηρίζεται σε γνώση της δομής του πληθυσμού για περιορισμό του δειγματοληπτικού σφάλματος στις εκτιμήσεις παραμέτρων.

΄νισες πιθανότητες επιλογής συνεπάγονται διαφορετική κατανομή του δείγματος από την κατανομή του πληθυσμού.

Η αντιπροσωπευτικότητα του δείγματος αποκαθίσταται με την χρήση των αναγωγικών συντελεστών.

## Δειγματικά βάρη (sampling/design weights)

Το βάρος (weight), ή αναγωγικός συντελεστής, της μονάδας $i \in U$ ορίζεται ως

$$w_i = \frac{1}{\pi_i} I_i(s), \quad i \in U$$

$$E(w_i) = \frac{1}{\pi_i} E(I_i(s)) = \frac{1}{\pi_i} \pi_i = 1$$

Το βάρος μονάδας που δεν έχει επιλεγεί στο δείγμα είναι εξ' ορισμού ίσο με το μηδέν.

Το βάρος μιας επιλεγμένης μονάδας είναι αντιστρόφως ανάλογο της πιθανότητας επιλογής της.

# Δειγματικά βάρη (sampling/design weights)

Η σημασία του $w_i$:

Το βάρος $w_i$ της δειγματικής μονάδας $i$, μπορεί να ερμηνευτεί ως ο αριθμός των μονάδων του πληθυσμού (συμπεριλαμβανομένης της μονάδας $i$) που "αντιπροσωπεύονται" από την δειγματική μονάδα $i$.
Έτσι, μπορούμε να θεωρήσουμε ότι η δειγματική μονάδα $i$ "αντιπροσωπεύει" τον εαυτό της συν $w_i - 1$ μή επιλεγμένες μονάδες του πληθυσμού, και όλες μαζί οι δειγματικές μονάδες "αντιπροσωπεύουν" όλον τον πληθυσμό.
Ένα δείγμα κάθε μονάδα του οποίου έχει το ίδιο δειγματικό βάρος ονομάζεται αυτοβαρές δείγμα.

Έστω ότι στο παράδειγμα 1, επιλέγεται το δείγμα $s_3 = \{1, 4\}$ με απλή τυχαία δειγματοληψία ώστε $w_1 = 2$ και $w_4 = 2$. Τότε το αυτοβαρές δείγμα $s_3$ φτιάχνει τον ψευδοπληθυσμό $\{1, 1, 4, 4\}$.

# Δειγματικά βάρη (sampling/design weights)

Έστω ο ακόλουθος πληθυσμός 12 μονάδων με αντίστοιχες τιμές μιας μεταβλητής $y$

$$U = \{y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8, y_9, y_{10}, y_{11}, y_{12}\}.$$

Έστω οτι ένα απλό τυχαίο δείγμα μεγέθους $n = 2$ απο το $U$ είναι το $s = \{y_3, y_8\}$. Σε αυτή την περίπτωση έχουμε $\pi = n/N = 2/12$ για όλες τις μονάδες του $U$, και $w = 6$ για τις δύο μονάδες του $s$.

Με βάση την ερμηνεία των βαρών, σε σχέση με την αντιπροσωπευτικότητα του δείγματος, μπορεί θεωρητικά το δείγμα να αναχθεί στον πληθυσμό συνθέτοντας τον ψευδοπληθυσμό

$$\{y_3, y_3, y_3, y_3, y_3 y_3, y_8, y_8, y_8, y_8, y_8, y_8\},$$

όπου κάθε μία απο τις μονάδες 3 και 8 του δείγματος αντιπροσωπεύει 6 μονάδες του πληθυσμού, με την ίδια τιμή της μεταβλητής $y$.

# Δειγματικά βάρη (sampling/design weights)

Έστω τώρα οτι δείγμα μεγέθους 4 απο το $U$ είναι το $s = \{y_2, y_5, y_8, y_{11}\}$. Εδώ έχουμε $\pi = n/N = 4/12$, και $w = 3$ για τις τέσσερις μονάδες του δείγματος.

Ο ψευδοπληθυσμός που μπορεί να συντεθεί με βάση αυτό το δείγμα και τα βάρη των τεσσάρων δειγματικών μονάδων είναι

$$\{y_2, y_2, y_2, y_5, y_5, y_5, y_8, y_8, y_8, y_{11}, y_{11}, y_{11}\}$$

Προφανώς, μεγαλύτερο δείγμα, με μεγαλύτερη πιθανότητα των μονάδων του πληθυσμού να περιληφθούν σε αυτό (άρα με μικρότερο βάρος) έχει καλύτερη αντιπροσωπευτικότητα.

# Δειγματικά βάρη (sampling/design weights)

Οι ¨ανηγμένες¨ δειγματικές τιμές $w_i y_i$ μιας μεταβλητής $y$ διορθώνουν την δυσαναλογικότητα του δείγματος, ως προς τον πληθυσμό δειγματοληψίας, που είναι αποτέλεσμα άνισων πιθανοτήτων επιλογής των δειγματικών μονάδων.

Οι αναγωγικοί συντελεστές (βάρη) χρησιμεύουν στην αναγωγή δειγματικών χαρακτηριστικών στα αντίστοιχα πληθυσμιακά χαρακτηριστικά.

# Στατιστική θεωρία πεπερασμένων πληθυσμών

Η στατιστική για πεπερασμένους πληθυσμούς είναι κατά κανόνα **περιγραφική**, δηλαδή ενδιαφέρεται για την εκτίμηση παραμέτρων του πληθυσμού.

Η εκτίμηση παραμέτρων του πληθυσμού βασίζεται σε δείγμα $s$ μεγέθους $n$ που επιλέγεται τυχαία με πιθανότητα $p(s)$.

Η αβεβαιότητα για την εκτίμηση οφείλεται στο γεγονός ότι μόνο μέρος του πληθυσμού ερευνάται. Ενώ τα χαρακτηριστικά του πληθυσμού παραμένουν σταθερά, οι εκτιμήσεις αυτών εξαρτώνται από ποιό δείγμα επιλέγεται.

Ο εκτιμητής μιας παραμέτρου $\theta = \theta(y_1, \ldots, y_N)$ είναι συνάρτηση του τυχαίου δείγματος, και συμβολίζεται με

$$\hat{\theta} = \hat{\theta}(s) = \hat{\theta}(y_1, \ldots, y_n)$$

Ο εκτιμητής $\hat{\theta}(s)$ είναι τυχαία μεταβλητή. Το μόνο τυχαίο στοιχείο είναι το σύνολο $s$ που ορίζει ποιές μονάδες απαρτίζουν το δείγμα. Η διαφορά τιμών του $\hat{\theta}(s)$ από δείγμα σε δείγμα συνιστά την δειγματική διακύμανση του $\hat{\theta}(s)$.

Η συμπερασματολογία για πεπερασμένους πληθυσμούς βασίζεται στην έννοια της υποθετικής επανάληψης της τυχαίας δειγματοληψίας, με δειγματοληπτικό σχέδιο $p(s)$, που έχει ως αποτέλεσμα την επιλογή διαφορετικών δειγμάτων.

Σύμφωνα με αυτή την αρχή της υποθετικής επανάληψης, η αναμενόμενη τιμή $E(\hat{\theta})$ του $\hat{\theta}(s)$ δίνεται από την

$$E(\hat{\theta}) = \sum_{s \in S} p(s)\hat{\theta}(s)$$

Το $E(\hat{\theta})$ είναι ανισοβαρής (ή σταθμικός) μέσος όρος των δυνατών τιμών $\hat{\theta}(s)$ του $\hat{\theta}$, με τις πιθανότητες $p(s)$ ως βάρη.

Όταν το $p(s)$ είναι σταθερό για όλα τα δείγματα $s \in S$, ιδίου μεγέθους $n$, τότε

$$E(\hat{\theta}) = \frac{1}{M} \sum_{s \in S} \hat{\theta}(s),$$

όπου $M$ είναι ο συνολικός αριθμός δειγμάτων μεγέθους $n$.

Ο εκτιμητής $\hat{\theta}$ είναι αμερόληπτος αν $E(\hat{\theta}) = \theta$, δηλαδή αν είναι ¨κατά μέσο όρο¨ ίσος με την εκτιμώμενη παράμετρο $\theta$.

Ένα μέτρο της δειγματικής διακύμανσης του $\hat{\theta}$, που συμβολίζεται με $V(\hat{\theta})$, δίνεται από την

$$V(\hat{\theta}) = \sum_{s \in S} p(s) \left[ \hat{\theta}(s) - E(\hat{\theta}(s)) \right]^2$$

Το τυπικό σφάλμα (standard error) του $\hat{\theta}$ ορίζεται ως $\sqrt{V(\hat{\theta})}$

Ως δείκτης αξιοπιστίας ενός αμερόληπτου εκτιμητή χρησιμοποιείται το σχετικό τυπικό της σφάλμα, γνωστό ως συντελεστής μεταβλητότητας (coefficient of variation), που ορίζεται ως $CV(\hat{\theta}) = \sqrt{V(\hat{\theta})}/\theta$, και εκφράζεται ποσοστιαία.

Ως μέρος της διαδικασίας εκτιμήσεων, η εκτίμηση της διακύμανσης $V(\hat{\theta})$, που συμβολίζεται $\hat{V}(\hat{\theta})$, υπολογίζεται από τα δεδομένα της δειγματοληψίας. Μία εκτίμηση του συντελεστή μεταβλητότητας είναι η $\sqrt{\hat{V}(\hat{\theta})}/\hat{\theta}$.

## Εκτίμηση παραμέτρων (parameter estimation)

Η συνάρτηση $p(s)$ έχει θεωρητικό ενδιαφέρον, ως μαθηματικό εργαλείο που θεμελιώνει την πιθανοθεωρία της δειγματοληψίας και καθορίζει τις στατιστικές ιδιότητες των εκτιμητριών, αλλά είναι συνήθως πολύπλοκη στην χρήση της.
Στην πράξη είναι πολύ πιο εύκολος ο καθορισμός της αναμενόμενης τιμής και της διακύμανσης συγκεκριμένων εκτιμητών γνωρίζοντας μόνο τις πιθανότητες $\pi_i$ και $\pi_{ij}$.

Επειδή οι κύριες παράμετροι που σχετίζονται με μια μεταβλητή $y$ είναι συναρτήσεις του πληθυσμιακού ολικού $Y = \sum_1^N y_i$, η μεθοδολογία εκτίμησης ασχολείται πρωταρχικά με αυτή την βασική παράμετρο.

# Εκτίμηση παραμέτρων (parameter estimation)

Για ένα δείγμα $s = \{y_1, \ldots, y_n\}$, ο εκτιμητής Horvitz-Thompson του $Y$ ορίζεται ως ο γραμμικός συνδυασμός (weighted sum)

$$\hat{Y} = \sum_{i=1}^{N} w_i y_i = \sum_{i=1}^{n} \frac{1}{\pi_i} y_i$$

Ο εκτιμητής $\hat{Y}$ είναι το άθροισμα των "ανηγμένων" δειγματικών τιμών $w_i y_i$ της μεταβλητής $y$.

Για την ειδική περίπτωση ολικού $Y = N$,

$$\hat{N} = \sum_{i=1}^{N} w_i = \sum_{i=1}^{n} \frac{1}{\pi_i}$$

Παράδειγμα: Σε απλή τυχαία δειγματοληψία, $\pi_i = n/N$, ώστε $w_i = N/n$, $\hat{N} = \sum_{i=1}^{n} \frac{1}{\pi_i} = N$ και $\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} y_i$.

# Εκτίμηση παραμέτρων (parameter estimation)

Ο εκτιμητής $\hat{Y}$ είναι αμερόληπτος, δηλαδή ισχύει $E(\hat{Y}) = Y$.

$$E(\hat{Y}) = \sum_{i=1}^{N} E(w_i) y_i = \sum_{i=1}^{N} y_i = Y$$

Η διακύμανση του $\hat{Y}$ δίνεται από την

$$V(\hat{Y}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i y_j$$

Ένας αμερόληπτος εκτιμητής του $V(\hat{Y})$ που υπολογίζεται από το δείγμα $s = \{y_1, \ldots, y_n\}$ δίνεται από την

$$\hat{V}(\hat{Y}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\pi_{ij}} \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i y_j$$

# Εκτίμηση παραμέτρων

$$\hat{V}(\hat{Y}) = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{\pi_{ij}} \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i y_j$$

Λόγω του διπλού αθροίσματος ο τύπος αυτός είναι δύσχρηστος στην πράξη. Για μερικά δειγματοληπτικά σχέδια $p(s)$ μπορεί να γίνει κατάλληλη απλοποίηση του τύπου για γρήγορους υπολογισμούς.

Επίσης, για μερικά δειγματοληπτικά σχέδια $p(s)$ είναι πολύ δύσκολο να υπολογιστούν τα $\pi_{ij}$. Τότε χρησιμοποιούνται άλλες προσεγγιστικές μέθοδοι εκτίμησης του $V(\hat{Y})$, που θα περιγραφούν αργότερα.

Στην απλή τυχαία δειγματοληψία, $\pi_i = n/N$, $\pi_{ij} = n(n-1)/N(N-1)$, και ο τύπος του $\hat{V}(\hat{Y})$ παίρνει την απλή μορφή

$$\hat{V}(\hat{Y}) = \frac{N(N-n)}{n} \frac{1}{(n-1)} \sum_{i=1}^{n} (y_i - \bar{y})^2, \qquad \text{όπου } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

# Εκτίμηση παραμέτρων

Ο εκτιμητής του πληθυσμιακού μέσου $\bar{Y}$ δίνεται από την

$$\hat{\bar{Y}} = \frac{\hat{Y}}{N} = \frac{1}{N}\sum_{i=1}^{N} w_i y_i = \frac{1}{N}\sum_{i=1}^{n} \frac{1}{\pi_i} y_i$$

Ένας εναλλακτικός εκτιμητής του $\bar{Y}$ που χρησιμοποιείται όταν το $N$ είναι άγνωστο δίνεται από την

$$\tilde{\bar{Y}} = \frac{\hat{Y}}{\hat{N}} = \frac{\sum_1^N w_i y_i}{\sum_1^N w_i} = \frac{\sum_1^n (1/\pi_i) y_i}{\sum_1^n 1/\pi_i}$$

Για κάποια δειγματοληπτικά σχέδια $p(s)$ οι εκτιμητές $\hat{\bar{Y}}$ και $\tilde{\bar{Y}}$ ταυτίζονται. Ακόμη και όταν το $N$ είναι γνωστό και οι δύο εκτιμητές διαφέρουν, ο εκτιμητής $\tilde{\bar{Y}}$ προτιμάται γιατί συνήθως έχει μικρότερη διακύμανση.

# Εκτίμηση παραμέτρων

Ο μη γραμμικός εκτιμητής $\tilde{\tilde{Y}}$ είναι **προσεγγιστικά** (για μεγάλα δείγματα) αμερόληπτος. Η **προσεγγιστική** διακύμανση του $\tilde{\tilde{Y}}$ δίνεται από την

$$\begin{aligned} V(\tilde{\tilde{Y}}) &= \frac{1}{N^2} V\left[\sum_{i=1}^{n} w_i(y_i - \bar{Y})\right] \\ &= \frac{1}{N^2} \sum_{i=1}^{N}\sum_{j=1}^{N} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right)(y_i - \bar{Y})(y_j - \bar{Y}). \end{aligned}$$

Ένας εκτιμητής του $V(\tilde{\tilde{Y}})$ δίνεται απο την

$$\begin{aligned} \hat{V}(\tilde{\tilde{Y}}) &= \frac{1}{\hat{N}^2} \hat{V}\left[\sum_{i=1}^{n} w_i(y_i - \tilde{\tilde{Y}})\right] \\ &= \frac{1}{\hat{N}^2} \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{\pi_{ij}} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right)(y_i - \tilde{\tilde{Y}})(y_j - \tilde{\tilde{Y}}) \end{aligned}$$

## Εκτίμηση παραμέτρων

Ο εκτιμητής ενός πληθυσμιακού ποσοστού $P$ δίνεται γενικά από την

$$\tilde{P} = \frac{\hat{N}_d}{\hat{N}} = \frac{\sum_1^{N_d} w_i}{\sum_1^N w_i} = \frac{\sum_1^{n_d} 1/\pi_i}{\sum_1^n 1/\pi_i},$$

όπου $n_d$ είναι το υποσύνολο του δείγματος που αντιστοιχεί στον υποπληθυσμό $U_d$, στον οποίο αναφέρεται το ποσοστό.

Για την προσεγγιστική αμεροληψία και την διακύμανση του $\tilde{P}$ ισχύουν τα ίδια όπως για τον $\tilde{\bar{Y}}$. Ειδικότερα, παρατηρώντας ότι $\tilde{P} = \tilde{\bar{Y}}$ ορίζοντας $y_i = 1$ αν $i \in U_d$ και $y_i = 0$ αν $i \notin U_d$, προκύπτει ότι

$$\hat{V}(\tilde{P}) = \frac{1}{\hat{N}^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\pi_{ij}} \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) (y_i - \tilde{P})(y_j - \tilde{P}).$$

Στην απλή τυχαία δειγματοληψία, ο εκτιμητής είναι $\hat{P} = \hat{N}_d/N$, και

$$\hat{V}(\hat{P}) = \frac{N-n}{N(n-1)} \hat{P}(1-\hat{P})$$

# Εκτίμηση παραμέτρων

Ο εκτιμητής ενός πληθυσμιακού λόγου $R = Y/Z$ δίνεται απο την

$$\hat{R} = \hat{Y}/\hat{Z}$$

Ο μη γραμμικός εκτιμητής $\hat{R}$ είναι προσεγγιστικά (για μεγάλα δείγματα) αμερόληπτος, με προσεγγιστική διακύμανση που δίνεται από την

$$\begin{aligned} V(\hat{R}) &= \frac{1}{Z^2} V\left[\sum_{i=1}^{n} w_i(y_i - Rz_i)\right] \\ &= \frac{1}{Z^2} \sum_{i=1}^{N}\sum_{j=1}^{N} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right)(y_i - Rz_i)(y_j - Rz_j). \end{aligned}$$

# Εκτίμηση παραμέτρων

Ένας εκτιμητής του $V(\hat{R})$ δίνεται απο την

$$\begin{aligned}\hat{V}(\hat{R}) &= \frac{1}{\hat{\bar{Z}}^2}\hat{V}\left[\sum_{i=1}^{n} w_i(y_i - \hat{R}z_i)\right] \\ &= \frac{1}{\hat{\bar{Z}}^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{\pi_{ij}}\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right)(y_i - \hat{R}z_i)(y_j - \hat{R}z_j).\end{aligned}$$

Στην απλή τυχαία δειγματοληψία ο τύπος του $\hat{V}(\hat{R})$ παίρνει την απλή μορφή

$$\hat{V}(\hat{R}) = \frac{N}{\hat{\bar{Z}}^2}\frac{N-n}{n}\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{R}z_i)^2.$$

# Εκτίμηση παραμέτρων

Έστω τώρα ο απλός γραμμικός συνδιασμός (διαφορά) εκτιμητών δύο ολικών $\hat{Y} - \hat{Z}$. Η διακύμανση αυτής της διαφοράς δίνεται απο την

$$V(\hat{Y} - \hat{Z}) = V(\hat{Y}) + V(\hat{Z}) - 2Cov(\hat{Y}, \hat{Z}),$$

όπου

$$Cov(\hat{Y}, \hat{Z}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i z_j$$

είναι η συνδιακύμανση των $\hat{Y}$ και $\hat{Z}$.

## Εκτίμηση παραμέτρων

Αμερόληπτος εκτιμητής του $V(\hat{Y} - \hat{Z})$ που υπολογίζεται από το δείγμα $s = \{y_1, \ldots, y_n\}$ δίνεται από την

$$\hat{V}(\hat{Y} - \hat{Z}) = \hat{V}(\hat{Y}) + \hat{V}(\hat{Z}) - 2\hat{Cov}(\hat{Y}, \hat{Z}),$$

όπου

$$\hat{Cov}(\hat{Y}, \hat{Z}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\pi_{ij}} \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i z_j$$

Ευκολότερος υπολογισμός του $\hat{V}(\hat{Y} - \hat{Z})$ στηρίζεται στην παρατήρηση ότι η διαφορά $\hat{Y} - \hat{Z} = \sum_s w_i y_i - \sum_s w_i z_i$ ορίζει τον εκτιμητή ολικού $\hat{D} = \sum_s w_i d_i$, με $d_i = y_i - z_i$. Τότε

$$\hat{V}(\hat{D}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\pi_{ij}} \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) d_i d_j.$$

Αυτή η διαδικασία υπολογισμού διακύμανσης μπορεί να γενικευθεί με προφανή τρόπο στην περίπτωση οποιασδήποτε γραμμικής συνάρτησης ολικών.

# Εκτίμηση παραμέτρων

Η εκτίμηση της πληθυσμιακής διακύμανσης μιας μεταβλητής $y$

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$$

βασίζεται στην ισοδύναμη έκφραση του $S^2$ ως συνάρτηση ολικών:

$$S^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} y_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} y_i\right)^2\right].$$

Τότε

$$\hat{S}^2 = \frac{1}{\hat{N}-1}\left[\sum_{i=1}^{n} w_i y_i^2 - \frac{1}{\hat{N}}\left(\sum_{i=1}^{n} w_i y_i\right)^2\right] = \frac{1}{\hat{N}-1}\sum_{i=1}^{n} w_i (y_i - \tilde{\tilde{Y}})^2,$$

όπου $\tilde{\tilde{Y}} = \hat{Y}/\hat{N}$.

# Εκτίμηση παραμέτρων

Με το ιδιο τρόπο μπορεί να γίνει εκτίμηση της συνδιακύμανσης δύο μεταβλητών $y$ και $z$

$$\begin{aligned} S_{yz} &= \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{Y})(z_i-\bar{Z}) \\ &= \frac{1}{N-1}\left[\sum_{i=1}^{N} y_i z_i - \frac{1}{N}\left(\sum_{i=1}^{N} y_i\right)\left(\sum_{i=1}^{N} z_i\right)\right]. \end{aligned}$$

Τότε

$$\begin{aligned} \hat{S}_{yz} &= \frac{1}{\hat{N}-1}\left[\sum_{i=1}^{n} w_i y_i z_i - \frac{1}{\hat{N}}\left(\sum_{i=1}^{n} w_i y_i\right)\left(\sum_{i=1}^{n} w_i z_i\right)\right] \\ &= \frac{1}{\hat{N}-1}\sum_{i=1}^{n} w_i(y_i-\tilde{\tilde{Y}})(z_i-\tilde{\tilde{Z}}). \end{aligned}$$

# Εκτίμηση παραμέτρων

Οι εκτιμητές $\hat{S}^2$ και $\hat{S}_{yz}$ δεν είναι αμερόληπτοι, αλλά τείνουν στις αντίστοιχες πληθυσμιακές παραμέτρους $S^2$ και $S_{yz}$ όταν το $n$ τείνει στο $N$.

Οι εκτιμητές $\hat{S}^2$ και $\hat{S}_{yz}$ είναι αμερόληπτοι όταν $\hat{N} = N$.

## Εκτιμητική υποπληθυσμών

Σε μεγάλης κλίμακας δειγματοληπτικές έρευνες, ενδιαφερώμαστε συχνά για εκτιμήσεις παραμέτρων διαφόρων υποπληθυσμών (domains) του πληθυσμού, τα οποία **δεν** συνιστούν στρώματα.

Έστω υποπληθυσμός $U_d \subset U$ με μέγεθος (αριθμό μονάδων) $N_d$. Το σχετικό μέγεθος του $U_d$, δηλαδή το ποσοστό μονάδων του $U$ που ανήκουν στο $U_d$, είναι $P_d = N_d/N$.

Έστω δείγμα $s$ απο το $U$ και $s_d = s \cap U_d$, (δηλαδή $s_d$ είναι το υποσύνολο του $s$ που ανήκει στον υποπληθυσμό $U_d$). Το μέγεθος $n_{s_d}$ του $s_d$ είναι τυχαίο, και μπορεί να γραφτεί ως

$$n_{s_d} = \sum_U I(i \in s_d) = \sum_{U_d} I(i \in s),$$

και προφανώς ισχύει $E(n_{s_d}) = \sum_{U_d} \pi_i$.

# Εκτιμητική υποπληθυσμών

Παράδειγμα:
Στην περίπτωση απλής τυχαίας δειγματοληψίας, οπότε $\pi_i = n/N$, προκύπτει ότι $E(n_{s_d}) = n\,N_d/N$, δηλαδή το αναμενόμενο μέγεθος του δείγματος $s_d$ είναι αναλογικό του μεγέθους του $U_d$.

## Εκτιμητική υποπληθυσμών

Έστω ότι θέλουμε να εκτιμήσουμε τον ολικό $Y_d = \sum_{U_d} y_i$ του υποπληθυσμού για μια μεταβλητή $y$. Ορίζουμε νέα μεταβλητή $y_d$ έτσι ώστε

$$y_{di} = \begin{cases} y_i, & i \in U_d \\ 0, & i \notin U_d \end{cases}$$

Τότε $Y_d$ είναι ο ολικός (για όλο το $U$ ) για την νέα μεταβλητή $y_d$, δηλαδή

$$Y_d = \sum_{U_d} y_i = \sum_{U} y_{di}.$$

Τώρα μπορούμε να χρησιμοποιήσουμε την γενική θεωρία για την εκτίμηση του ολικού $Y_d$.

Ο εκτιμητής HT του $Y_d$ είναι

$$\hat{Y}_d = \sum_{s_d} w_i y_i = \sum_{s} w_i y_{di} \quad (= \sum_{s} \frac{1}{\pi_i} y_{di}).$$

# Εκτιμητική υποπληθυσμών

Με βάση την γενική θεωρία

$$E(\hat{Y}_d) = \sum_U E(w_i) y_{di} = Y_d$$

$$\begin{aligned} V(\hat{Y}_d) &= \sum_U \sum_U \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_{di} y_{dj} \\ &= \sum_{U_d} \sum_{U_d} \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i y_j \end{aligned}$$

$$\begin{aligned} \hat{V}(\hat{Y}_d) &= \sum_s \sum_s \frac{1}{\pi_{ij}} \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_{di} y_{dj} \\ &= \sum_{s_d} \sum_{s_d} \frac{1}{\pi_{ij}} \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i y_j \end{aligned}$$

# Εκτιμητική υποπληθυσμών

Στην περίπτωση απλής τυχαίας δειγματοληψίας, $\hat{Y}_d = (N/n)\sum_{s_d} y_i$ και

$$V(\hat{Y}_d) = \frac{N(N-n)}{n}\frac{1}{N-1}\sum_U (y_{di} - \bar{y}_d)^2, \qquad \bar{y}_d = \frac{1}{N}\sum_U y_{di}$$

$$\hat{V}(\hat{Y}_d) = \frac{N(N-n)}{n}\frac{1}{n-1}\sum_s (y_{di} - \hat{\bar{y}}_d)^2, \qquad \hat{\bar{y}}_d = \frac{1}{n}\sum_s y_{di}$$

# Εκτιμητική υποπληθυσμών

Ενδιαφέρουσα είναι η περίπτωση εκτίμησης της διαφοράς ολικών $Y_{d1} - Y_{d2}$ για δύο υποπληθυσμούς $U_{d1}$ και $U_{d2}$, π.χ., η διαφορά ολικού τζίρου για δύο διαφορετικούς τύπους επιχειρήσεων. Τότε

$$\hat{Y}_{d1} - \hat{Y}_{d2} = \sum_{s_{d1}} w_i y_i - \sum_{s_{d2}} w_i y_i,$$

$$V(\hat{Y}_{d1} - \hat{Y}_{d2}) = V(\hat{Y}_{d1}) + V(\hat{Y}_{d2}) - 2Cov(\hat{Y}_{d1}, \hat{Y}_{d2}),$$

όπου

$$\begin{aligned} Cov(\hat{Y}_{d1}, \hat{Y}_{d2}) &= \sum_{U} \sum_{U} \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_{d1i} y_{d2j} \\ &= \sum_{U_{d1}} \sum_{U_{d2}} \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i y_j \end{aligned}$$

Η αμερόληπτη εκτίμηση του $Cov(\hat{Y}_{d1}, \hat{Y}_{d2})$ είναι

$$\hat{Cov}(\hat{Y}_{d1}, \hat{Y}_{d2}) = \sum_{i \in s_{d1}} \sum_{j \in s_{d2}} \frac{1}{\pi_{ij}} \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i y_j.$$

Στην απλή τυχαία δειγματοληψία,

$$Cov(\hat{Y}_{d1}, \hat{Y}_{d2}) = -\frac{N(N-n)}{n} \frac{1}{N-1} \sum_{U} (y_{d1i} - \bar{y}_{d1})(y_{d2i} - \bar{y}_{d2}),$$

με αμερόληπτη εκτίμηση

$$\hat{Cov}(\hat{Y}_{d1}, \hat{Y}_{d2}) = -\frac{N(N-n)}{n} \frac{1}{n-1} \sum_{s} (y_{d1i} - \hat{\bar{y}}_{d1})(y_{d2i} - \hat{\bar{y}}_{d2}).$$

# Εκτιμητική υποπληθυσμών

Ο εκτιμητής HT του $N_d$ είναι

$$\hat{N}_d = \sum_{s_d} w_i \quad (= \sum_{s_d} \frac{1}{\pi_i})$$

Ο εκτιμητής HT του μέσου $\bar{Y}_d = Y_d/N_d$ του υποπληθυσμού $U_d$ είναι

$$\hat{\bar{Y}}_d = \hat{Y}_d/\hat{N}_d.$$

Η προσεγγιστική διακύμανση του $\hat{\bar{Y}}_d$ δίνεται απο την

$$\begin{aligned} V(\hat{\bar{Y}}_d) &= \frac{1}{N_d^2} V\left[\sum_{s_d} w_i (y_i - \bar{Y}_d)\right] \\ &= \frac{1}{N_d^2} \sum_{U_d} \sum_{U_d} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right)(y_i - \bar{Y}_d)(y_j - \bar{Y}_d) \end{aligned}$$

## Εκτιμητική υποπληθυσμών

Ένας εκτιμητής του $V(\hat{\bar{Y}}_d)$ δίνεται απο την

$$\begin{aligned}\hat{V}(\hat{\bar{Y}}_d) &= \frac{1}{\hat{N}_d^2}\hat{V}\left[\sum_{s_d} w_i(y_i - \hat{\bar{Y}}_d)\right] \\ &= \frac{1}{\hat{N}_d^2}\sum_{s_d}\sum_{s_d}\frac{1}{\pi_{ij}}\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right)(y_i - \hat{\bar{Y}}_d)(y_j - \hat{\bar{Y}}_d)\end{aligned}$$

Στην απλή τυχαία δειγματοληψία ο τύπος του $\hat{V}(\hat{\bar{Y}}_d)$ παίρνει την απλή μορφή

$$\hat{V}(\hat{\bar{Y}}_d) = \frac{n(N-n)}{n_d^2 N}\frac{n_d - 1}{n - 1}\frac{1}{n_d - 1}\sum_{s_d}(y_i - \bar{y}_d)^2,$$

όπου $\bar{y}_d = \frac{1}{n_d}\sum_{s_d} y_i$.
Ο προσεγγιστικός τύπος της διακύμανσης ισχύει όταν το αναμενόμενο μέγεθος του δείγματος στον υποπληθυσμό είναι επαρκώς μεγάλο.

# Εκτιμητική υποπληθυσμών

Πολύ χρήσιμη στις δειγματοληπτικές έρευνες είναι η σύγκριση των μέσων δύο διαφορετικών υποπληθυσμών, π.χ., διαφορά στο μέσο ατομικό εισόδημα για γυναίκες και άνδρες. Έστω λοιπόν η διαφορά των εκτιμημένων μέσων

$$\hat{D} = \hat{\bar{Y}}_{d1} - \hat{\bar{Y}}_{d2}.$$

Η διακύμανση του $\hat{D}$ είναι

$$V(\hat{D}) = V(\hat{\bar{Y}}_{d1}) + V(\hat{\bar{Y}}_{d2}) - 2Cov(\hat{\bar{Y}}_{d1}, \hat{\bar{Y}}_{d1}),$$

όπου, προσεγγιστικά,

$$Cov(\hat{\bar{Y}}_{d1}, \hat{\bar{Y}}_{d2}) = \frac{1}{N_{d1}N_{d2}} \sum_{U_{d1}} \sum_{U_{d2}} \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) (y_i - \bar{Y}_{d1})(y_j - \bar{Y}_{d2}).$$

Ένας εκτιμητής του $Cov(\hat{\bar{Y}}_{d1}, \hat{\bar{Y}}_{d2})$ δίνεται απο την

$$\hat{Cov}(\hat{\bar{Y}}_{d1}, \hat{\bar{Y}}_{d2}) = \frac{1}{\hat{N}_{d1}\hat{N}_{d2}} \sum_{s_{d1}} \sum_{s_{d2}} \frac{1}{\pi_{ij}} \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) (y_i - \hat{\bar{Y}}_{d1})(y_j - \hat{\bar{Y}}_{d2}).$$

Παρατήρηση: Προκύπτει εύκολα ότι η προσεγγιστική συνδιακύμανση $Cov(\hat{\bar{Y}}_{d1}, \hat{\bar{Y}}_{d2})$ είναι ίση με το μηδέν όταν $\pi_{ij}/\pi_i\pi_j$ είναι σταθερό για όλα τα $i, j \in U$, όπως ισχύει στην περίπτωση της απλής τυχαίας δειγματοληψίας. Τότε

$$V(\hat{D}) = V(\hat{\bar{Y}}_{d1}) + V(\hat{\bar{Y}}_{d2}).$$

# Εκτίμηση της συνάρτησης κατανομής

Εκτίμηση της συνάρτησης κατανομής μίας μεταβλητής επιτρέπει εκτίμηση παραμέτρων που δεν είναι συναρτήσεις ολικών, όπως η διάμεσος, τα τεταρτημόρια, εκατοστημόρια κ.λπ.

Για μία μεταβλητή $y$, η αθροιστική συνάρτηση κατανομής (cumulative distribution function) ορίζεται ως

$$F(y) = \frac{\sum_{i=1}^{N} z_i(y)}{N}, \qquad \text{όπου} \quad z_i(y) = \begin{cases} 1, & \text{αν} \quad y_i \leq y \\ 0, & \text{αν} \quad y_i > y \end{cases}$$

Για οποιοδήποτε αριθμό $y \in (-\infty, +\infty)$, το $F(y)$ είναι το ποσοστό των μονάδων του πληθυσμού με τιμή $y_i$ μικρότερη ή ίση του $y$.

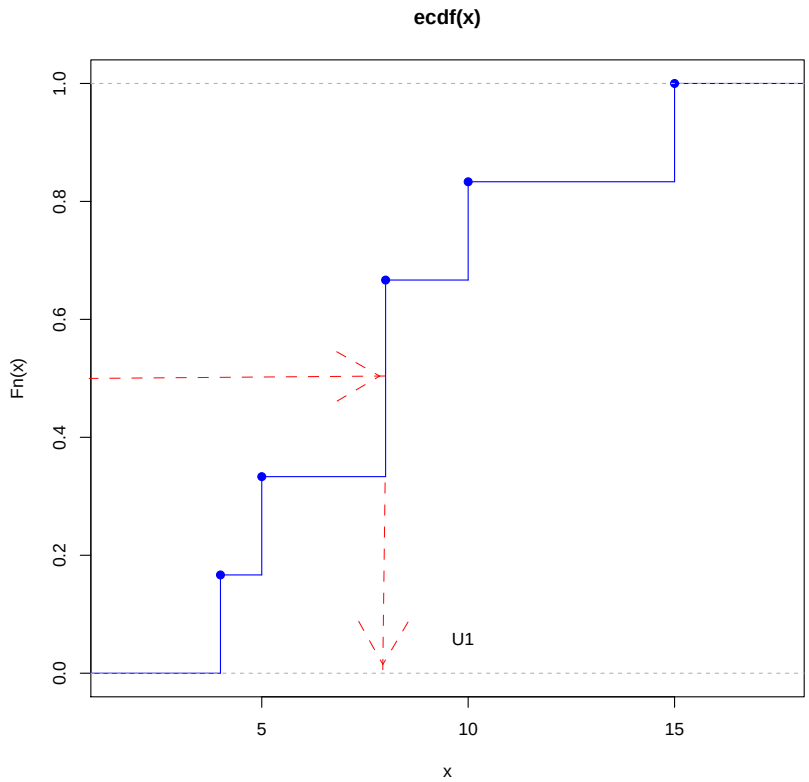
# Εκτίμηση της συνάρτησης κατανομής

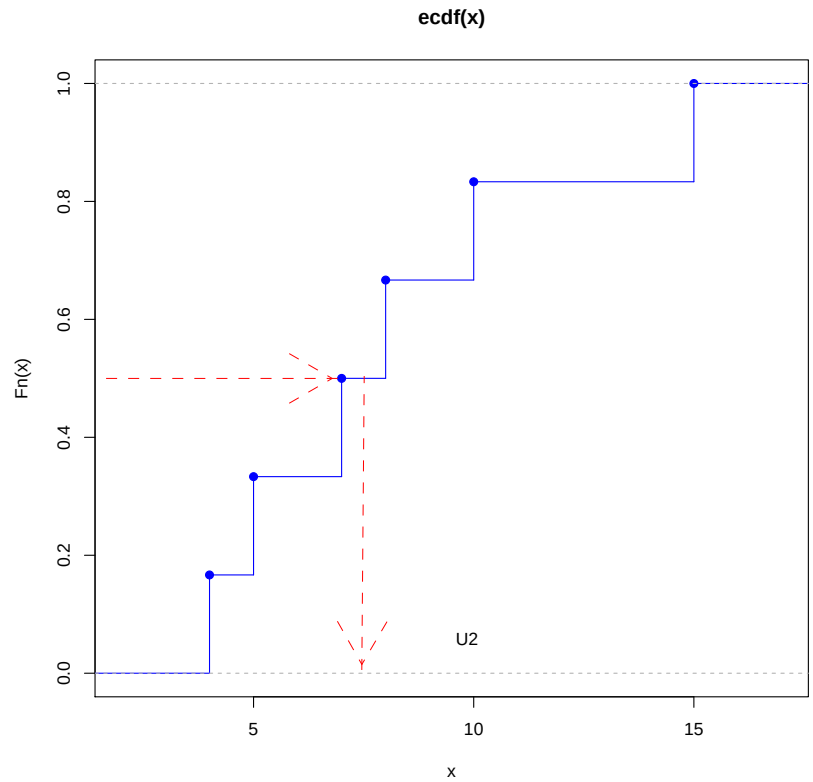
Αν η συνάρτηση $F$ ήταν συνεχής, η διάμεσος $m$ του πληθυσμού θα οριζόταν ως η τιμή για την οποία $F(m) = 0.5$, ή $m = F^{-1}(0.5)$. Όμως, επειδή η $F$ είναι αύξουσα συνάρτηση βήματος (με βήματα στις τιμές του $y$ στον πληθυσμό) είναι δυνατόν η συνάρτηση $F(y)$ να μήν παίρνει ακριβώς την τιμή 0.5.

Ορίζουμε ως διάμεσο του πληθυσμού την τιμή για την οποία $F(m) = 0.5$, αν υπάρχει τέτοια μοναδική τιμή, αλλιώς η διάμεσος είναι οποιαδήποτε τιμή στο διάστημα $[m_1, m_2]$, όπου $m_1$ είναι η μεγαλύτερη τιμή του $y$ στον πληθυσμό με $F(y) < 0.5$ και $m_2$ είναι η μικρότερη τιμή του $y$ με $F(y) > 0.5$. (Συνήθως η διάμεσος ορίζεται ως $m = (m_1 + m_2)/2$).

Οι δύο περιπτώσεις απεικονίζονται στα ακόλουθα διαγράμματα για τα δεδομένα $U_1 = \{4, 5, 8, 8, 10, 15\}$ και $U_2 = \{4, 5, 7, 8, 10, 15\}$, όπου $m = 8$ και $m = 7.5$ για τα $U_1$ και $U_2$, αντίστοιχα.

# Εκτίμηση της συνάρτησης κατανομής



# Εκτίμηση της συνάρτησης κατανομής

Η διάμεσος του πληθυσμού για την μεταβλητή $y$ υπολογίζεται πρακτικά ως εξής:
Με τις τιμές του $y$ διατεταγμένες κατά αύξον μέγεθος, όπου $y_j$ είναι η τιμή με θέση μεγέθους $j$, η διάμεσος $m$ δίνεται απο τον τύπο

$$m = \begin{cases} \frac{y_j + y_{j+1}}{2} & \text{αν} \quad \frac{j}{N} = \frac{1}{2} \\ \\ y_{j+1} & \text{αν} \quad \frac{j}{N} < \frac{1}{2} < \frac{j+1}{N} \end{cases}$$

Ας σημειωθεί ότι το $j/N$ είναι το ποσοστό των τιμών του $y$ απο την μικρότερη μέχρι και την διατεταγμένη τιμή $y_j$.

# Εκτίμηση της συνάρτησης κατανομής

Γενικά, $\theta_q$ είναι το $q$ εκατοστημόριο αν $F(\theta_q) = q$, αν υπάρχει τέτοια μοναδική τιμή $\theta_q$, αλλιώς $\theta_q \in [a, b]$, όπου $a$ είναι η μεγαλύτερη τιμή του $y$ με $F(y) < q$ και $b$ είναι η μικρότερη τιμή του $y$ με $F(y) > q$.
(Συνήθως το $\theta_q$ ορίζεται ως $\theta_q = (a + b)/2$).

Με γενίκευση του τύπου υπολογισμού της διαμέσου, το $\theta_q$ υπολογίζεται ως εξής:
Με τις τιμές του $y$ διατεταγμένες κατά αύξον μέγεθος, όπου $y_j$ είναι η τιμή με θέση μεγέθους $j$, το $\theta_q$ δίνεται απο τον τύπο

$$\theta_q = \begin{cases} \frac{y_j + y_{j+1}}{2} & \text{αν} \quad \frac{j}{N} = \frac{1}{q} \\ \\ y_{j+1} & \text{αν} \quad \frac{j}{N} < \frac{1}{q} < \frac{j+1}{N} \end{cases}$$

# Εκτίμηση της συνάρτησης κατανομής

Η εμπειρική αθροιστική συνάρτηση κατανομής είναι εκτίμηση της συνάρτησης $F(y)$ που υπολογίζεται απο ένα δείγμα, και δίνεται απο την

$$\hat{F}(y) = \frac{\sum_{i=1}^{n} w_i z_i(y)}{\sum_{i=1}^{n} w_i}.$$

Για αυτοβαρές δείγμα (με ίσα βάρη),

$$\hat{F}(y) = \frac{\sum_{i=1}^{n} z_i(y)}{n},$$

δηλαδή το $\hat{F}(y)$ είναι το ποσοστό των μονάδων του δείγματος με τιμή $y_i$ μικρότερη ή ίση του $y$.

Τότε η εκτίμηση $\hat{m}$ της διαμέσου και των εκατοστημορίων $\hat{\theta}_q$ καθορίζεται απο το $\hat{F}(y)$, με τον ίδιο τρόπο που το $m$ και το $\theta_q$ καθορίζεται απο το $F(y)$.

# Εκτίμηση της συνάρτησης κατανομής

Παράδειγμα:

Έστω ότι δείγμα μεγέθους $n = 5$ επιλέγεται απο πληθυσμό μεγέθους $N = 50$, στο οποίο οι τιμές του $y$ για τις επιλεγμένες μονάδες είναι $\{8, 15, 12, 4, 8\}$ και οι αντίστοιχες πιθανότητες επιλογής είναι $\{0.10, 0.10, 0.05, 0.10, 0.20\}$. Αναδιατάσσοντας τις τιμές του $y$ κατά αύξον μέγεθος έχουμε τα δεδομένα του δείγματος

| $i$ | $y_i$ | $\pi_i$ | $w_i$ |
|---|---|---|---|
| 1 | 4 | 0.10 | 10 |
| 2 | 8 | 0.10 | 10 |
| 3 | 8 | 0.20 | 5 |
| 4 | 12 | 0.05 | 20 |
| 5 | 15 | 0.10 | 10 |

Το άθροισμα των βαρών δίνει $\hat{N} = 55$, και η συνάρτηση $\hat{F}$ υπολογίζεται ως εξής:

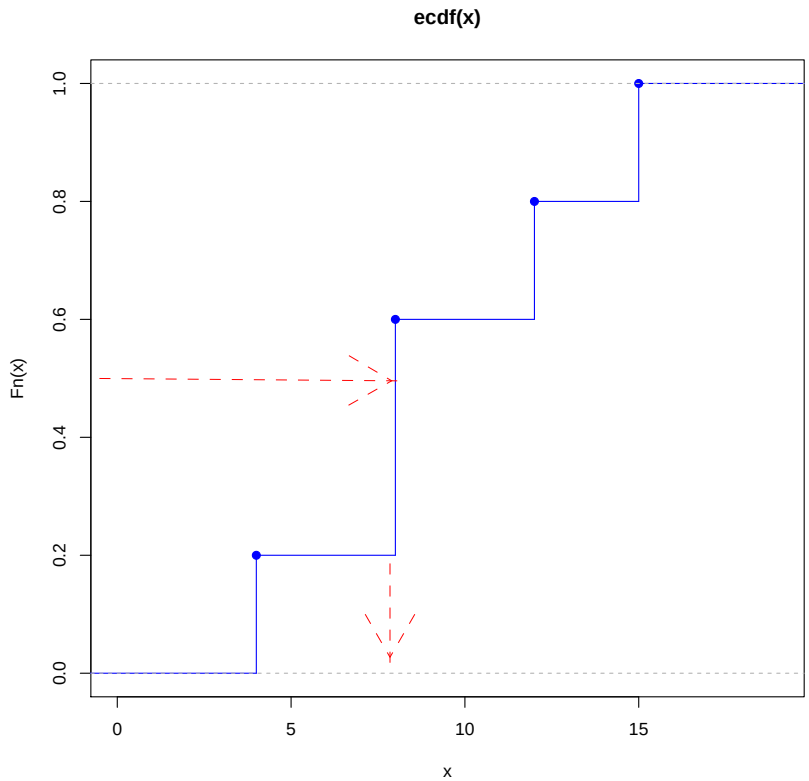
# Εκτίμηση της συνάρτησης κατανομής

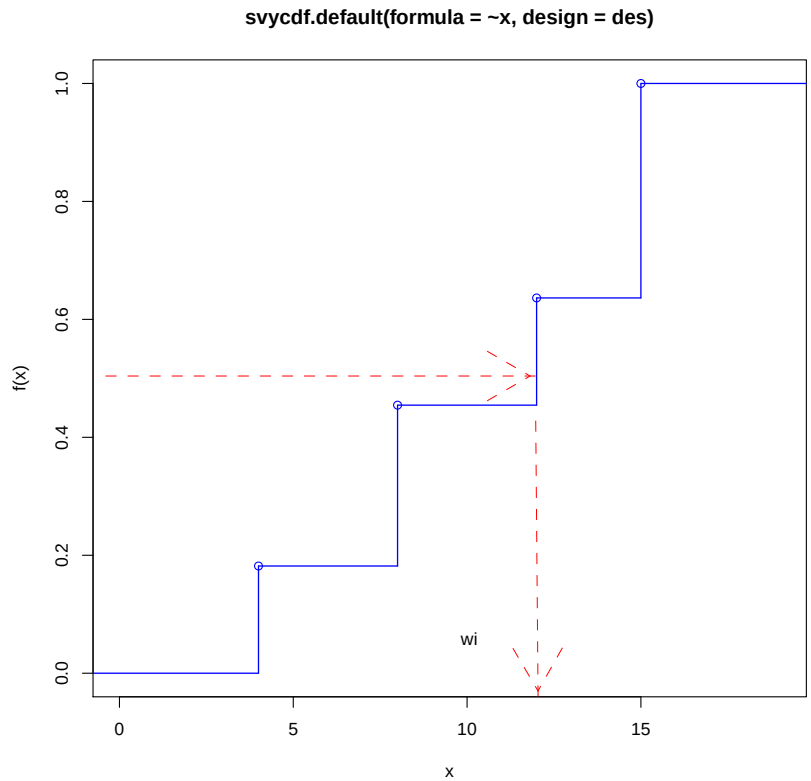
| $y$ | $\hat{F}(y)$ |
|---|---|
| $y < 4$ | 0 |
| $4 \leq y < 8$ | 10/55 |
| $8 \leq y < 12$ | (10+10+5)/55=25/55 |
| $12 \leq y < 15$ | (10+10+5+20)/55=45/55 |
| $15 \leq y$ | 1 |

Προκύπτει ότι η εκτίμηση της διαμέσου είναι $\hat{m} = 12$. Υπολογισμός της διαμέσου χωρίς την χρήση των βαρών (ή, ισοδύναμα, υποθέτοντας ίσα βάρη) δίνει την λανθασμένη εκτίμηση $\hat{m} = 8$.

Απεικόνιση της εμπειρικής συνάρτησης $\hat{F}$ υπολογισμένης χωρίς βάρη και με βάρη δίνεται στα ακόλουθα διαγράμματα.

# Εκτίμηση της συνάρτησης κατανομής



# Εκτίμηση της συνάρτησης κατανομής

Πρακτικά, για τιμές του $y$ διατεταγμένες κατά αύξον μέγεθος, όπου $y_j$ είναι η τιμή με θέση μεγέθους $j$, η εκτιμημένη διάμεσος $\hat{m}$ δίνεται απο τον τύπο

$$\hat{m} = \begin{cases} \frac{y_j + y_{j+1}}{2} & \text{αν} \quad \frac{\sum_{i=1}^{j} w_i}{\sum_{i=1}^{n} w_i} = \frac{1}{2} \\ y_{j+1} & \text{αν} \quad \frac{\sum_{i=1}^{j} w_i}{\sum_{i=1}^{n} w_i} < \frac{1}{2} < \frac{\sum_{i=1}^{j+1} w_i}{\sum_{i=1}^{n} w_i} \end{cases}$$

Αν τα βάρη όλων των δειγματικών μονάδων είναι ίσα, τότε η διάμεσος $\hat{m}$ δίνεται από τον απλοποιημένο τύπο

$$\hat{m} = \begin{cases} \frac{y_j + y_{j+1}}{2} & \text{αν} \quad \frac{j}{n} = \frac{1}{2} \\ y_{j+1} & \text{αν} \quad \frac{j}{n} < \frac{1}{2} < \frac{j+1}{n} \end{cases}$$

που δίνει την διάμεσο του $y$ στο δείγμα.

# Γραφική απεικόνιση κατανομής δεδομένων: ιστόγραμμα

Για την κατασκευή ιστογράμματος για ένα δείγμα μεγέθους $n$, χωρίζουμε το εύρος των παρατηρήσεων σε $k$ κλάσεις που έχουν όλες πλάτος $b$. Τότε το ύψος του ιστού στην κλάση $j$ είναι το εκτιμημένο ποσοστό των παρατηρήσεων του πληθυσμού στην κλάση αυτή διαιρεμένο με το πλάτος $b$, δηλαδή

$$\text{ύψος}(j) = \frac{\sum_{i=1}^{n} w_i z_i(j)}{b \sum_{i=1}^{n} w_i}, \quad \text{όπου} \quad z_i(j) = \begin{cases} 1, & \text{αν} \quad i \text{ είναι στην κλάση } j \\ 0, & \text{αλλιώς} \end{cases}$$

Για αυτοβαρή δείγματα, το ύψος κάθε ιστού $j$ δίνεται απο τον συνήθη τύπο

$$\text{ύψος}(j) = \frac{\sum_{i=1}^{n} z_i(j)}{bn}.$$

Η κατασκευή ιστογράμματος γίνεται εύκολη με την χρήση κατάλληλου προγράμματος, που λαμβάνει υπόψη τα βάρη των παρατηρήσεων (βλέπε βιβλιογραφία: T. Lumley).

## Χρήση βοηθητικών μεταβλητών στην εκτίμηση παραμέτρων

Στις δειγματοληπτικές έρευνες είναι πολύ χρήσιμο να ορίζονται **βοηθητικές** μεταβλητές (auxiliary variables), συνεχείς ή κατηγορικές.

Παραδείγματα: γεωγραφική τοποθεσία, φύλλο, ηλικία, έκταση αγροκτήματος, κλάδος επιχείρησης.

Οι τιμές μίας βοηθητικής μεταβλητής μπορεί να είναι γνωστές πριν από τη δειγματοληψία, π.χ., γεωγραφική τοποθεσία, ή γίνονται γνωστές μόνο για τις μονάδες του επιλεγμένου δείγματος, π.χ. ηλικία ατόμων.

Οι βοηθητικές μεταβλητές χρησιμοποιούνται:

- στον σχεδιασμό της δειγματοληψίας
- στον ορισμό υποπληθυσμών
- στην βελτίωση των εκτιμητών παραμέτρων του ερευνώμενου πληθυσμού.

# Χρήση βοηθητικών μεταβλητών στην εκτίμηση παραμέτρων

Ο εκτιμητής Horvitz-Thompson $\hat{Y}$ βασίζεται στις δειγματικές τιμές του $y$.

Έστω ότι διαθέτουμε πληροφορία για μία βοηθητική (πολυ)μεταβλητή (auxiliary vector) $\mathbf{x}$ με $p$ συνιστώσες, δηλαδή $\mathbf{x} = \left(x_1, \ldots, x_p\right)'$, είτε με τις τιμές $\mathbf{x}_i = \left(x_{i1}, \ldots, x_{ip}\right)', i \in U$ είτε με τον ολικό $\mathbf{X} = \sum_U \mathbf{x}_i$.

Αυτή η πληροφορία (για όλο τον πληθυσμό) μπορεί να χρησιμοποιηθεί για την βελτίωση της εκτίμησης του $\hat{Y}$ αν το $y$ συσχετίζεται με το $\mathbf{x}$.

# Χρήση βοηθητικών μεταβλητών στην εκτίμηση παραμέτρων

Έστω ότι ο συσχετισμός του $y$ με το $\mathbf{x}$ είναι τέτοιος ώστε για κάθε $i \in U$ το $y_i$ να προσεγγίζεται ("προβλέπεται") απο τον γραμμικό συνδιασμό $\mathbf{x}_i^{'}\mathbf{B} = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$, δηλαδή $y_i \approx \mathbf{x}_i^{'}\mathbf{B}$.

Κατάλληλος συντελεστής $\mathbf{B}$ (με τις $p$ συνιστώσες) υπολογίζεται με την μέθοδο ελαχίστων τετραγώνων (least squares), που ελαχιστοποιεί το άθροισμα των τετραγώνων $\sum_U (y_i - \mathbf{x}_i^{'}\mathbf{B})^2$. Αυτό το $\mathbf{B}$ δίνεται απο την

$$\mathbf{B} = \left( \sum_U \frac{\mathbf{x}_i \mathbf{x}_i^{'}}{q_i} \right)^{-1} \sum_U \frac{\mathbf{x}_i y_i}{q_i},$$

όπου $q_i$ είναι γνωστοί παράγοντες (συνήθως $q_i = 1$).

Για μονομεταβλητή $x$ (και $q_i = 1$) , το μονοδιάστατο $B$ έχει την μορφή

$$B = \frac{\sum_U x_i y_i}{\sum_U x_i^2}.$$

# Χρήση βοηθητικών μεταβλητών στην εκτίμηση παραμέτρων

Έστω ότι ο συσχετισμός του $y$ με το $\mathbf{x}$ είναι τέτοιος ώστε για κάθε $i \in U$ το $y_i$ να προσεγγίζεται ("προβλέπεται") απο τον γραμμικό συνδιασμό $\mathbf{x}_i^{'}\mathbf{B} = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$, δηλαδή $y_i \approx \mathbf{x}_i^{'}\mathbf{B}$.

Κατάλληλος συντελεστής $\mathbf{B}$ υπολογίζεται με την μέθοδο ελαχίστων τετραγώνων (least squares), που ελαχιστοποιεί το άθροισμα των τετραγώνων $\sum_U (y_i - \mathbf{x}_i^{'}\mathbf{B})^2$. Αυτό το $\mathbf{B}$ δίνεται απο την

$$\mathbf{B} = \left( \sum_U \frac{\mathbf{x}_i \mathbf{x}_i^{'}}{q_i} \right)^{-1} \sum_U \frac{\mathbf{x}_i y_i}{q_i},$$

όπου $q_i$ είναι γνωστοί παράγοντες (συνήθως $q_i = 1$).

Με αυτό το $\mathbf{B}$ ορίζεται σχέση γραμμικής παλινδρόμησης (linear regression) μεταξύ $y$ και $\mathbf{x}$, το $\mathbf{B}$ είναι ο συντελεστής παλινδρόμησης (regression coefficients), και οι διαφορές $e_i = y_i - \mathbf{x}_i^{'}\mathbf{B}$ είναι τα κατάλοιπα (residuals) της παλινδρόμησης.

# Χρήση βοηθητικών μεταβλητών στην εκτίμηση παραμέτρων

Ο συντελεστής $\mathbf{B}$

$$\mathbf{B} = \left( \sum_U \frac{\mathbf{x}_i \mathbf{x}_i^{'}}{q_i} \right)^{-1} \sum_U \frac{\mathbf{x}_i y_i}{q_i},$$

είναι παράμετρος του πληθυσμού. Εκτίμηση του $\mathbf{B}$ βάσει δείγματος $s$ δίνεται απο την

$$\hat{\mathbf{B}} = \left( \sum_s \frac{w_i \mathbf{x}_i \mathbf{x}_i^{'}}{q_i} \right)^{-1} \sum_s \frac{w_i \mathbf{x}_i y_i}{q_i},$$

Για τις δειγματικές τιμές του $y$ τα κατάλοιπα είναι $\hat{e}_i = y_i - \mathbf{x}_i^{'} \hat{\mathbf{B}}$.

Ασυμπτωτικά (δηλαδή για μεγάλο δείγμα) το $\hat{\mathbf{B}}$ είναι προσεγγιστικά ίσο με το $\mathbf{B}$.

# Εκτιμητής παλινδρόμησης (Regression estimator)

Ο εκτιμητής παλινδρόμησης του ολικού $Y$ ορίζεται ως

$$\begin{aligned}
\hat{Y}^{GR} &= \hat{Y} + \hat{\mathbf{B}}^{'}(\mathbf{X} - \hat{\mathbf{X}}) \\
&= \sum_s w_i y_i + \sum_s \frac{w_i y_i \mathbf{x}_i^{'}}{q_i} \left( \sum_s \frac{w_i \mathbf{x}_i \mathbf{x}_i^{'}}{q_i} \right)^{-1} (\mathbf{X} - \sum_s w_i \mathbf{x}_i)
\end{aligned}$$

# Εκτιμητής παλινδρόμησης (Regression estimator)

Ιδιότητες του $\hat{Y}^{GR}$:

Ασυμπτωτικά, το $\hat{Y}^{GR}$ δίνεται προσεγγιστικά απο την

$$\hat{Y}^{GR} \approx \hat{Y} + \mathbf{B}^{'}(\mathbf{X} - \hat{\mathbf{X}}).$$

Προκύπτει ότι $E(\hat{Y}^{GR}) \approx Y$, ($\hat{Y}^{GR}$ είναι προσεγγιστικά αμερόληπτος εκτιμητής του $Y$).

Εναλλακτικά,

$$\begin{aligned} \hat{Y}^{GR} &\approx \mathbf{B}^{'}\mathbf{X} + (\hat{Y} - \mathbf{B}^{'}\hat{\mathbf{X}}) \\ &= \mathbf{B}^{'}\mathbf{X} + \sum_{s} w_i(y_i - \mathbf{x}_i^{'}\mathbf{B}) \\ &= \mathbf{B}^{'}\mathbf{X} + \sum_{s} w_i e_i \end{aligned}$$

# Εκτιμητής παλινδρόμησης (Regression estimator)

Προκύπτει ότι ασυμπτωτικά $V(\hat{Y}^{GR}) \approx V(\sum_s w_i e_i)$

Παρατήρηση: Το $\sum_s w_i e_i$ είναι HT εκτιμητής του ολικού $\sum_U e_i$.
΄ρα

$$V(\hat{Y}^{GR}) \approx \sum_U \sum_U \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) e_i e_j$$

Ένας ασυμπτωτικά αμερόληπτος εκτιμητής του $V(\hat{Y}^{GR})$ είναι

$$\hat{V}(\hat{Y}^{GR}) \approx \sum_s \sum_s \frac{1}{\pi_{ij}} \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) \hat{e}_i \hat{e}_j$$

Ο εκτιμητής $\hat{Y}^{GR}$ έχει ασυμπτωτικά μικρότερη διακύμανση απο τον HT εκτιμητή $\hat{Y}$ αν τα κατάλοιπα $e_i = y_i - \mathbf{x}_i^{'}\mathbf{B}$ έχουν μικρότερη διακύμανση απο τις τιμές $y_i$.

# Ειδικές περιπτώσεις εκτιμητή παλινδρόμησης

Εκτιμητής λόγου (ratio estimator):

Έστω η μονομεταβλητή $x_i$, ώστε προσεγγιστικά $y_i \approx Bx_i$, και έστω $q_i = x_i$.

Τότε $B = Y/X$ (γιατί;), $\hat{B} = \hat{Y}/\hat{X}$, και απο τον γενικό τύπο του $\hat{Y}^{GR}$ προκύπτει ο εκτιμητής

$$\hat{Y}^R = \hat{Y}\frac{X}{\hat{X}} \quad (= X\hat{B})$$

Η διακύμανση του $\hat{Y}^R$ προκύπτει απο τον γενικό τύπο του $V(\hat{Y}^{GR})$.

Ο εκτιμητής $\hat{Y}^R$ έχει μικρότερη διακύμανση απο τον εκτιμητή $\hat{Y}$ όταν οι τιμές $y_i$ είναι διεσπαρμένες σε μικρή απόσταση απο ευθεία γραμμή που περνάει απο το μηδέν (μικρά κατάλοιπα $e_i = y_i - x_iB$).

# Ειδικές περιπτώσεις εκτιμητή παλινδρόμησης

Όταν $x_i = 1$, ώστε $y_i \approx B$, $X = N$ και $B = Y/N$, τότε

$$\hat{Y}^R = \hat{Y}\frac{N}{\hat{N}} \quad (= N\hat{B}).$$

Αυτή η παραλλαγή του εκτιμητή λόγου μπορεί να χρησιμοποιηθεί σε δειγματοληπτικά σχέδια $p(s)$ για τα οποία $\hat{N} \neq N$.

Γενικεύσεις του εκτιμητή $\hat{Y}^R$ ορίζονται όταν διαφορετικές γραμμικές παλινδρομήσεις $y_i \approx Bx_i$ ορίζονται για διαφορετικούς υποπληθυσμούς που απαρτίζουν τον πληθυσμό $U$ (βλέπε "μεταστρωμάτωση" κατωτέρω).

# Ειδικές περιπτώσεις εκτιμητή παλινδρόμησης

*Μεταστρωματικός εκτιμητής* (poststratified estimator):

Μεταστρωμάτωση είναι ο διαμερισμός ενός δείγματος σε υποσύνολα (μεταστρώματα) που αντιστοιχούν σε συγκεκριμένους υποπληθυσμούς. Η μεταστρωμάτωση γίνεται μόνο μετά την συλλογή των στοιχείων, οπότε οι δειγματικές μονάδες αναγνωρίζονται ως μέλη των υποπληθυσμών.

Παράδειγμα: Σε δειγματοληψία ανθρώπινου πληθυσμού, είναι δυνατή η μεταστρωμάτωση κατά συγκεκριμένες ομάδες ηλικίας αν η ηλικία είναι μία απο τις βοηθητικές πληροφορίες που συλλέγονται απο το δείγμα.

# Ειδικές περιπτώσεις εκτιμητή παλινδρόμησης

Ας θεωρήσουμε μεταστρώματα $U_1, \ldots, U_G$ με αντίστοιχα γνωστά μεγέθη $N_1, \ldots, N_G$ και δείγματα $s_1, \ldots, s_G$.
Ας υποθέσουμε διφορετικές γραμμικές προσεγγίσεις κατά μετάστρωμα, $y_i \approx B_g$ για $i \in U_g$, ώστε $B_g = Y_g/N_g$ και $\hat{B}_g = \hat{Y}_g/\hat{N}_g$.

Ο μεταστρωματικός εκτιμητής ορίζεται ως

$$\hat{Y}^{PS} = \sum_{g=1}^{G} \hat{Y}_g \frac{N_g}{\hat{N}_g} \qquad \left(= \sum_{g=1}^{G} N_g \hat{\bar{Y}}_g\right).$$

Η διακύμανση του $\hat{Y}^{PS}$ προκύπτει απο τον γενικό τύπο του $V\left(\hat{Y}^{GR}\right)$.

Ο εκτιμητής $\hat{Y}^{PS}$ είναι αποτελεσματικότερος του $\hat{Y}^{R} = \hat{Y}\frac{N}{\hat{N}}$ όταν τα $B_g$ διαφέρουν σημαντικά.

Οι ασυμπτωτικές ιδιότητες του $\hat{Y}^{PS}$ απαιτούν αρκετά μεγάλα μεγέθη $n_g$ ή μικρό αριθμό μεταστρωμάτων.

# Σημαντικές ιδιότητες του $\hat{Y}^{GR}$

Ο εκτιμητής $\hat{Y}^{GR}$ μπορεί να γραφτεί εναλλακτικά ως

$$\begin{aligned}
\hat{Y}^{GR} &= \sum_s w_i y_i + \sum_s \frac{w_i y_i \mathbf{x}_i'}{q_i} \left( \sum_s \frac{w_i \mathbf{x}_i \mathbf{x}_i'}{q_i} \right)^{-1} (\mathbf{X} - \sum_s w_i \mathbf{x}_i) \\
&= \sum_s w_i \left[ 1 + \frac{\mathbf{x}_i'}{q_i} \left( \sum_s \frac{w_i \mathbf{x}_i \mathbf{x}_i'}{q_i} \right)^{-1} (\mathbf{X} - \sum_s w_i \mathbf{x}_i) \right] y_i \\
&= \sum_s c_i y_i
\end{aligned}$$

όπου $c_i = w_i g_i$, με $g_i = 1 + \frac{\mathbf{x}_i'}{q_i} \left( \sum_s \frac{w_i \mathbf{x}_i \mathbf{x}_i'}{q_i} \right)^{-1} (\mathbf{X} - \sum_s w_i \mathbf{x}_i)$

Ο εκτιμητής $\hat{Y}^{GR}$ έχει γραμμική μορφή, ως προς τις τιμές $y_i$, όπως και ο HT εκτιμητής $\hat{Y} = \sum_s w_i y_i$. Τα βάρη $c_i$ δεν εξαρτώνται απο το $y$.

# Σημαντικές ιδιότητες του $\hat{Y}^{GR}$

Με αντικατάσταση του $y_i$ με το $\mathbf{x}_i$ προκύπτει απο τον τύπο

$$\hat{Y}^{GR} = \sum_s w_i y_i + \sum_s \frac{w_i y_i \mathbf{x}_i^{'}}{q_i} \left( \sum_s \frac{w_i \mathbf{x}_i \mathbf{x}_i^{'}}{q_i} \right)^{-1} (\mathbf{X} - \sum_s w_i \mathbf{x}_i)$$

ότι ο εκτιμητής παλινδρόμησης του ολικού $\mathbf{X}$ είναι $\hat{\mathbf{X}}^{GR} = \mathbf{X}$.

Σε αυτή την περίπτωση, τα βάρη $c_i$ προσαρμόζωνται (calibrated) στον γνωστό πληθυσμιακό ολικό $\mathbf{X}$, δηλαδή $\sum_s c_i \mathbf{x}_i = \sum_U \mathbf{x}_i$.

# Calibration

Calibration (προσαρμογή) είναι μία διαδικασία προσαρμογής των δειγματικών βαρών στην γραμμική μορφή $\sum_s w_i y_i$ ώστε τα νέα βάρη $c_i$ (calibrated weights) να ικανοποιούν την εξίσωση προσαρμογής $\sum_s c_i \mathbf{x}_i = \sum_U \mathbf{x}_i$, δηλαδή η διαδικασία εκτίμησης να αναπαράγει ακριβώς γνωστούς πληθυσμιακούς ολικούς $\mathbf{X}$.

Τα ίδια βάρη $c_i$ παράγουν τον εκτιμητή (calibration estimator)

$$\hat{Y}^C = \sum_s c_i y_i$$

του ολικού $Y$ οποιασδήποτε μεταβλητής $y$.

# Calibration

Εύκολα προκύπτει ότι

$$\hat{Y}^C = \hat{Y} + \sum_s (c_i - w_i) y_i$$

Ο εκτιμητής $\hat{Y}^C$ θα είναι προσεγγιστικά αμερόληπτος αν $E\left[\sum_s (c_i - w_i) y_i\right] \approx 0$, δηλαδή αν οι διαφορές $c_i - w_i$ είναι μικρές.

Τα κατάλληλα βάρη $c_i$ μπορούν να καθοριστούν με την ελαχιστοποίηση της συνάρτησης αποστάσεων $\sum_s q_i (c_i - w_i)^2 / w_i$ υπο τον περιορισμό $\sum_s c_i \mathbf{x}_i = \sum_U \mathbf{x}_i$. Συνήθως $q_i = 1$.
Η ελαχιστοποίηση δίνει

$c_i = w_i g_i$, όπου $g_i = 1 + \frac{\mathbf{x}_i'}{q_i} \left( \sum_s \frac{w_i \mathbf{x}_i \mathbf{x}_i'}{q_i} \right)^{-1} (\mathbf{X} - \sum_s w_i \mathbf{x}_i)$,

δηλαδή

$$\hat{Y}^C = \hat{Y}^{GR}$$

# Calibration

Βασικός σκοπός της διαδικασίας calibration είναι η συμβατότητα συγκεκριμένων εκτιμήσεων με ολικούς του πληθυσμού που είναι ήδη γνωστοί απο άλλες πηγές (π.χ., διοικητικές, απογραφικές).

Το calibration οδηγεί σε εκτιμητή γραμμικής μορφής ταυτοτικά ίδιο με τον εκτιμητή παλινδρόμησης, ενώ δεν χρησιμοποιεί καμία υπόθεση γραμμικής συσχέτισης (παλινδρόμησης) του $y$ με το $\mathbf{x}$.

Οι ρυθμιστικοί παράγοντες $g_i$ στο $c_i = w_i g_i$ εξαρτώνται απο τις παρατηρήσεις $\mathbf{x}_i$, αλλά είναι ανεξάρτητοι του $y$. Μπορεί να θεωρηθούν ως μέτρο της διαφοράς δείγματος και πληθυσμού. Ισχύει ότι $g_i \to 1$ όταν $n \to N$.

Επειδή $\hat{Y}^C = \hat{Y}^{GR}$, ο εκτιμητής $\hat{Y}^C$ μπορεί να γραφεί και ως $\hat{Y}^C = \hat{Y} + \hat{\mathbf{B}}'(\mathbf{X} - \hat{\mathbf{X}})$.

# Calibration

*Ειδική περίπτωση* 1:

Έστω κατηγορική μεταβλητή $\mathbf{x}$ με $p$ κατηγορίες που αντιστοιχούν σε διαμέριση ενός πληθυσμού σε $p$ πληθυσμιακές ομάδες $U_1, \ldots, U_p$. Έστω ότι τα μεγέθη των ομάδων, $N_1, \ldots, N_p$ είναι γνωστά. (Τί θυμίζουν αυτά;)

Η τιμή της μεταβλητής $\mathbf{x}$ για την μονάδα $i \in U$ ορίζεται ως

$$\mathbf{x}_i = (\delta_{i1}, \ldots, \delta_{ip},)', \qquad \text{όπου } \delta_{ij} = \begin{cases} 1, & i \in U_j \\ 0, & i \notin U_j \end{cases}$$

και δηλώνει σε ποιά ομάδα ανήκει η μονάδα $i$.

Τότε ο ολικός του $\mathbf{x}$ είναι

$$\sum_{i \in U} \mathbf{x}_i = (N_1, \ldots, N_p)'.$$

# Calibration

Έστω $s$ είναι ένα τυχαίο δείγμα απο το $U$, και $s_j = s \cap U_j$ είναι το σύνολο των δειγματικών μονάδων που ανήκουν στην κατηγορία (ομάδα) $j$. Τότε

$$\sum_{i \in s} w_i \mathbf{x}_i = (\hat{N}_1, \ldots, \hat{N}_p)', \qquad (\hat{N}_j = \sum_{i \in s_j} w_j)$$

Calibration που ικανοποιεί την εξίσωση

$$\sum_{i \in s} c_i \mathbf{x}_i = \sum_{i \in U} \mathbf{x}_i, \qquad \text{δηλαδή } (\hat{N}_1^c, \ldots, \hat{N}_p^c)' = (N_1, \ldots, N_p)')$$

δίνει $g_i = N_j / \hat{N}_j$ αν $i \in U_j$ (ώστε $c_i = w_i N_j / \hat{N}_j$) και το calibration estimator του ολικού $Y$, οποιασδήποτε μεταβλητής $y$, δίνεται απο την

$$\hat{Y}^C = \sum_{i \in s} c_i y_i = \sum_{j=1}^{p} \sum_{i \in s_j} c_i y_i = \sum_{j=1}^{p} \frac{N_j}{\hat{N}_j} \sum_{i \in s_j} w_i y_i = \sum_{j=1}^{p} \hat{Y}_j \frac{N_j}{\hat{N}_j},$$

όπου $\hat{Y}_j$ είναι ο HT εκτιμητής του ολικού του $y$ για την κατηγορία $j$.

# Calibration

Παρατηρήσεις:

- Σε αυτή την περίπτωση calibration, ο εκτιμητής $\hat{Y}^C$ είναι ο ίδιος με τον εκτιμητή μεταστρωμάτωσης!
- Ο εκτιμητής $\hat{Y}^C$ έχει απλή μορφή (γενικά αυτό δεν ισχύει) και η κατασκευή του είναι απλή.
- Επειδή $\hat{Y}^C = \hat{Y}^{GR}$, ο εκτιμητής $\hat{Y}^C$ μπορεί να γραφεί και ως $\hat{Y}^C = \hat{Y} + \hat{\mathbf{B}}'(\mathbf{N} - \hat{\mathbf{N}})$, όπου $\mathbf{N} = (N_1 \ldots, N_p)'$, $\hat{\mathbf{N}} = (\hat{N}_1 \ldots, \hat{N}_p)'$.

  Αναλυτικά: $\hat{Y}^C = \hat{Y} + \hat{B}_1(N_1 - \hat{N}_1) + \cdots + \hat{B}_p(N_1 - \hat{N}_p)$.
- Συνήθης περίπτωση calibration: π.χ., οι $p$ κατηγορίες της μεταβλητής $\mathbf{x}$ είναι ομάδες ηλικίας (σε δειγματοληψίες ατόμων), ή κλάδοι επιχειρήσεων (σε δειγματοληψίες επιχειρήσεων).

*Ειδική περίπτωση* 2:

Έστω κατηγορική μεταβλητή **x** με $p + q$ κατηγορίες που αντιστοιχούν σε δύο διαφορετικές διαμερίσεις ενός πληθυσμού σε $p$ ομάδες $U_{11}, \ldots, U_{1p}$ και σε $q$ ομάδες $U_{21}, \ldots, U_{2q}$. Έστω ότι τα μεγέθη των ομάδων, $N_{11}, \ldots, N_{1p}$ και $N_{21}, \ldots, N_{2q}$, αντιστοίχως, είναι γνωστά.

Παράδειγμα: Διαμέριση πληθυσμού ατόμων κατά ομάδες ηλικίας και κατά γεωγραφικές περιοχές.

Η τιμή της μεταβλητής **x** για την μονάδα $i \in U$ ορίζεται ως

$$\mathbf{x}_i = \left(\delta_{i11}, \ldots, \delta_{i1p}, \delta_{i21}, \ldots, \delta_{i2q}\right)',$$

όπου $\delta_{i1j} = \begin{cases} 1, & i \in U_{1j} \\ 0, & i \notin U_{1j} \end{cases}$ και $\delta_{i2k} = \begin{cases} 1, & i \in U_{2k} \\ 0, & i \notin U_{2k} \end{cases}$

# Calibration

Ο ολικός του $\mathbf{x}$ είναι

$$\sum_{i \in U} \mathbf{x}_i = (N_{11}, \ldots, N_{1p}, N_{21}, \ldots, N_{2q})'.$$

Έστω $s$ είναι ένα τυχαίο δείγμα απο το $U$, και $s_{1j} = s \cap U_{1j}$, $s_{2k} = s \cap U_{2k}$ είναι οι διαμερίσεις του δείγματος που αντιστοιχούν στις δύο διαμερίσεις του πληθυσμού. Τότε

$$\sum_{i \in s} w_i \mathbf{x}_i = (\hat{N}_{11}, \ldots, \hat{N}_{1p}, \hat{N}_{21}, \ldots, \hat{N}_{2q})', \quad (\hat{N}_{1j} = \sum_{i \in s_{1j}} w_i), \quad (\hat{N}_{2k} = \sum_{i \in s_{2k}} w_i)$$

Calibration που ικανοποιεί την εξίσωση

$$\sum_{i \in s} c_i \mathbf{x}_i = \sum_{i \in U} \mathbf{x}_i$$

εξισώνει τα μεγέθη των $p + q$ ομάδων με τις εκτιμήσεις τους που προκύπτουν απο το δείγμα $s$.

# Calibration

Σε αυτή την περίπτωση, τα (calibrated) βάρη $c_i$ δεν έχουν απλή μορφή, αλλά το $\hat{Y}^C$ μπορεί να γραφτεί ως

$$\hat{Y}^C = \hat{Y} + \hat{\mathbf{B}}_1'(\mathbf{N}_1 - \hat{\mathbf{N}}_1) + \hat{\mathbf{B}}_2'(\mathbf{N}_2 - \hat{\mathbf{N}}_2),$$

όπου $\mathbf{N}_1 = (N_{11} \ldots, N_{1p})'$, $\hat{\mathbf{N}}_1 = (\hat{N}_{11} \ldots, \hat{N}_{1p})'$ και
$\mathbf{N}_2 = (N_{21} \ldots, N_{2q})'$, $\hat{\mathbf{N}}_2 = (\hat{N}_{21} \ldots, \hat{N}_{2q})'$.

# Calibration

Διασταύρωση των δύο διαμερίσεων παράγει $p \times q$ ομάδες $U_{jk}$, $j = 1, \ldots, p$, $k = 1, \ldots, q$ με αντίστοιχα μεγέθη $N_{jk}$. Αυτό είναι ισοδύναμο με μονή διαμέριση του $U$ σε $p \times q$ ομάδες ως προς δύο χαρακτηριστικά ταυτοχρόνως, π.χ., με ταξινόμιση κάθε μονάδας πληθυσμού ατόμων κατά γεωγραφική περιοχή και ομάδα ηλικίας.

Αν τα μεγέθη $N_{jk}$ είναι γνωστά, το calibration που εξισώνει τα $\hat{N}_{jk}$ με τα $N_{jk}$ ($p \times q$ εξισώσεις) ανάγεται στην πρώτη περίπτωση, που παράγει τον απλό εκτιμητή μεταστρωμάτωσης.

# Calibration

Παρατηρήσεις:

Σε δειγματοληπτικές έρευνες μεγάλης κλίμακας γίνεται calibration για πολλαπλή διαμέριση του πληθυσμού, π.χ., κατά φύλλο, κατά ομάδες ηλικίας και κατά γεωγραφική περιοχή. Η διαδικασία είναι γενίκευση της διαδικασίας της *περίπτωσης* 2.

Calibration για πολλαπλή διαμέριση με πολλές συνολικά ομάδες συνεπάγεται μεγάλο αριθμό περιορισμών (εξισώσεων εκτιμήσεων με ολικούς). Αυτό μπορεί να έχει αρνητικές επιπτώσεις:

- μερικούς αρνητικούς παράγοντες προσαρμογής $g_i$, άρα και αρνητικά βάρη $c_i = w_i g_i$.
- το δείγμα στις διάφορες ομάδες να μή είναι επαρκές για τις ασυμπτωτικές ιδιότητες του $\hat{Y}^C$.

# Calibration στην εκτιμητική υποπληθυσμών

Έστω ότι για κάποια βοηθητική πολυμεταβλητή $\mathbf{x}$ έχει γίνει calibration, ώστε $\sum_s c_i \mathbf{x}_i = \sum_U \mathbf{x}_i$.
Τότε τα βάρη $c_i$ μπορούν να χρησιμοποιηθούν για την εκτίμηση του ολικού $Y_d$ οποιασδήποτε μεταβλητής $y$ και για οποιοδήποτε $U_d$.

Η διαδικασία είναι η ίδια όπως για τον εκτιμητή HT $\hat{Y}_d$, αλλά χρησιμοποποιώντας τα $c_i$ αντί τα $w_i$, δηλαδή

$$\hat{Y}_d^C = \sum_s c_i y_{di}.$$

# Calibration στην εκτιμητική υποπληθυσμών

Χρησιμοποιώντας την αναλυτική έκφραση

$$c_i = w_i g_i = w_i \left[ 1 + \frac{\mathbf{x}_i'}{q_i} \left( \sum_s \frac{w_i \mathbf{x}_i \mathbf{x}_i'}{q_i} \right)^{-1} (\mathbf{X} - \sum_s w_i \mathbf{x}_i) \right]$$

ο εκτιμητής μπορεί να πάρει την εναλλακτική μορφή εκτιμητή παλινδρόμησης

$$\hat{Y}_d^C = \hat{Y}_d + \hat{\mathbf{B}}_d'(\mathbf{X} - \hat{\mathbf{X}}),$$

όπου $\hat{\mathbf{B}}_d = \left( \sum_s \frac{w_i \mathbf{x}_i \mathbf{x}_i'}{q_i} \right)^{-1} \sum_s \frac{w_i \mathbf{x}_i y_{di}}{q_i} = \left( \sum_s \frac{w_i \mathbf{x}_i \mathbf{x}_i'}{q_i} \right)^{-1} \sum_{s_d} \frac{w_i \mathbf{x}_i y_i}{q_i}.$

# Calibration στην εκτιμητική υποπληθυσμών

Παρατηρήσεις:

Το calibration δεν έχει γίνει ειδικά για το $U_d$, δηλαδή $\sum_{s_d} c_i \mathbf{x}_i \neq \sum_{U_d} \mathbf{x}_i$.

Με όρους παλινδρόμησης αυτό σημαίνει ότι η βοηθητική μεταβλητή $\mathbf{x}$ έχει χρησιμοποιηθεί για την βελτίωση της εκτίμησης στο επίπεδο του πληθυσμού $U$ και όχι στο επίπεδο του υποπληθυσμού $U_d$.

Συνέπεια αυτού είναι ότι η βελτίωση της εκτίμησης (έναντι του $\hat{Y}_d$) είναι μικρή ή μηδενική (όσο μικρότερο το $U_d$ τόσο μικρότερη η βελτίωση).

# Calibration στην εκτιμητική υποπληθυσμών

Αν θέλουμε calibration για το $U_d$, δηλαδή $\sum_{s_d} c_i \mathbf{x}_i = \sum_{U_d} \mathbf{x}_i$, ή αν θέλουμε βελτίωση της εκτίμησης του $Y_d$, τότε η διαδικασία κατασκευής του $\hat{Y}_d^C$ περιορίζεται στο $s_d$, και τότε

$$\hat{Y}_d^C = \hat{Y}_d + \hat{\mathbf{B}}_d^{'}(\mathbf{X}_d - \hat{\mathbf{X}}_d),$$

όπου $\hat{\mathbf{B}}_d = \left( \sum_{s_d} \frac{w_i \mathbf{x}_i \mathbf{x}_i^{'}}{q_i} \right)^{-1} \sum_{s_d} \frac{w_i \mathbf{x}_i y_i}{q_i}$.

Calibration σε επίπεδο $U_d$ προϋποθέτει ότι οι ολικοί $\mathbf{X}_d$ είναι διαθέσιμοι, πράγμα που ίσως να μη ισχύει (ειδικά για πολύ μικρό $U_d$).

Επίσης, για μικρό $U_d$ ή/και για πολλές βοηθητικές μεταβλητές, το δείγμα ίσως δεν είναι αρκετά μεγάλο για να ισχύουν οι ασυμπτωτικές ιδιότητες του $\hat{Y}_d^C$.

# Εκτίμηση διακύμανσης σε περιπλεγμένες δειγματοληπτικές έρευνες (Variance estimation for complex surveys)

Όπως είδαμε σε προηγούμενο κεφάλαιο, για οποιοδήποτε δειγματοληπτικό σχέδιο $p(s)$ η διακύμανση του ολικού $\hat{Y}$ δίνεται από την

$$V(\hat{Y}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i y_j$$

Αν $\pi_{ij} > 0$ για όλα τα $i, j \in U$, ένας αμερόληπτος εκτιμητής του $V(\hat{Y})$ που υπολογίζεται από το δείγμα $s = \{y_1, \ldots, y_n\}$ δίνεται από την

$$\hat{V}(\hat{Y}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\pi_{ij}} \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i y_j$$

Όπως έχει ήδη δειχτεί, η διακύμανση των σημαντικών μη γραμμικών συναρτήσεων εκτιμητών ολικών $\hat{\bar{Y}} = \hat{Y}/\hat{N}$, $\hat{P} = \hat{N}_d/\hat{N}$, και $\hat{R} = \hat{Y}/\hat{Z}$ υπολογίζεται προσεγγιστικά (για μεγάλα δείγματα) με κατάλληλη παραλαγή του βασικού τύπου διακύμανσης εκτιμητή ολικού.

# Εκτίμηση διακύμανσης σε περιπλεγμένες δειγματοληπτικές έρευνες

Στην πράξη, η γενική χρήση του βασικού τύπου

$$\hat{V}(\hat{Y}) = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{\pi_{ij}} \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i y_j$$

είναι προβληματική για τους εξής λόγους:

- Λόγω του διπλού αθροίσματος ο τύπος αυτός είναι υπολογιστικά δύσχρηστος για μεγάλα δείγματα.
- Για πολλά δειγματοληπτικά σχέδια $p(s)$ είναι πολύ δύσκολο να υπολογιστούν οι πιθανότητες $\pi_{ij}$.
- Ο τύπος δεν μπορεί να εφαρμοστεί στην περίπτωση μή ομαλών συναρτήσεων ολικών (π.χ., διαμέσου, τεταρτημορίων).

Σε τέτοιες προβληματικές περιπτώσεις χρησιμοποιούνται άλλες, προσεγγιστικές, μέθοδοι εκτίμησης της διακύμανσης.

**Μία απλή προσεγγιστική μέθοδος εκτίμησης του $V(\hat{Y})$:**

Ένας προσεγγιστικός εκτιμητής του $V(\hat{Y})$ δίνεται απο τον τύπο

$$\tilde{V}(\hat{Y}) = \frac{1}{n(n-1)} \sum_s \left( \frac{y_i}{\pi_i/n} - \hat{Y} \right)^2$$

Αυτός ο απλοποιημένος εκτιμητής υπολογίζεται ως εάν η δειγματοληψία έχει γίνει με επανάθεση, ενώ στην πραγματικότητα έχει γίνει χωρίς επανάθεση, και έτσι παρακάμπτει τις πιθανότητες $\pi_{ij}$ και το διπλό άθροισμα.

Στην περίπτωση απλής τυχαίας δειγματοληψίας έχουμε $\tilde{V}(\hat{Y}) = \frac{N^2}{n(n-1)} \sum_s (y_i - \bar{y})^2$, όπου $\bar{y} = \sum_s y_i/n$. Αν το κλάσμα δειγματοληψίας $f = n/N$ είναι πολύ μικρό, ώστε $1 - f \approx 1$, τότε $\tilde{V}(\hat{Y}) = \hat{V}(\hat{Y})$ (δηλαδή η διακύμανση που υπολογίζεται με τον γενικό τύπο).

# Εκτίμηση διακύμανσης σε περιπλεγμένες δειγματοληπτικές έρευνες

Η προσεγγιστική εκτίμηση εφαρμόζεται και στην στρωματική δειγματοληψία, με $H$ στρώματα και στρωματικά δειγματικά μεγέθη $n_h$:

$$\tilde{V}(\hat{Y}) = \sum_{h=1}^{H} \frac{1}{n_h(n_h-1)} \sum_{s_h} \left( \frac{y_i}{\pi_i/n_h} - \hat{Y}_h \right)^2$$

Η απλοποίηση των υπολογισμών για το $\tilde{V}(\hat{Y})$, σε σύγκριση με το $\hat{V}(\hat{Y})$, είναι σημαντική. Όμως, το $\tilde{V}(\hat{Y})$ δεν είναι αμερόληπτος εκτιμητής του $V(\hat{Y})$.

Σε πολλές περιπτώσεις η μεροληψία είναι θετική, και το $\tilde{V}(\hat{Y})$ μπορεί να χρησιμεύσει ως άνω όριο εκτίμησης της διακύμανσης $V(\hat{Y})$.

# Μέθοδοι επαναληπτικής δειγματοληψίας (resampling, or replication, methods of variance estimation)

## Η γενική μεθοδολογία

Έστω παράμετρος $\theta$, με εκτιμητή $\hat{\theta}$ που υπολογίζεται μέσω δείγματος $s$ με δειγματοληπτικό σχέδιο $p(s)$.

Θεωρείστε ένα αριθμό (έστω $K$) κατάλληλων υποσυνόλων του δείγματος $s$, και τους διαφορετικούς ομοιότυπους εκτιμητές $\hat{\theta}_1, \ldots, \hat{\theta}_K$ που υπολογίζονται απο τα $K$ διαφορετικά υποσύνολα $s_1, \ldots, s_K$.

Ένας εναλλακτικός εκτιμητής του $\theta$ που στηρίζεται στο πλήρες δείγμα $s$ είναι ο μέσος όρος

$$\hat{\theta}^\star = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}_k$$

# Μέθοδοι επαναληπτικής δειγματοληψίας

Παράδειγμα 1:

$$\theta = Y, \qquad \hat{\theta} = \hat{Y} = \sum_{s} w_i y_i, \qquad \hat{\theta}_k = \hat{Y}_k = \sum_{s_k} w_i y_i, \quad k = 1, \ldots, K$$

$$\hat{\theta}^{\star} = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}_k = \frac{1}{K} \sum_{k=1}^{K} \hat{Y}_k$$

Παράδειγμα 2:

$$\theta = \frac{Y}{Z}, \qquad \hat{\theta} = \frac{\hat{Y}}{\hat{Z}}, \qquad \hat{\theta}_k = \frac{\hat{Y}_k}{\hat{Z}_k}, \quad k = 1, \ldots, K$$

$$\hat{\theta}^{\star} = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}_k = \frac{1}{K} \sum_{k=1}^{K} \frac{\hat{Y}_k}{\hat{Z}_k}$$

Θεωρούμε δύο εκτιμητές διακύμανσης του $\hat{\theta}^\star$:

$$\hat{V}_1 = \frac{1}{K(K-1)} \sum_{k=1}^{K} \left( \hat{\theta}_k - \hat{\theta}^\star \right)^2$$

και

$$\hat{V}_2 = \frac{1}{K(K-1)} \sum_{k=1}^{K} \left( \hat{\theta}_k - \hat{\theta} \right)^2$$

Απο την ταυτότητα

$$\sum_{k=1}^{K} \left( \hat{\theta}_k - \hat{\theta} \right)^2 = \sum_{k=1}^{K} \left( \hat{\theta}_k - \hat{\theta}^\star \right)^2 + K(\hat{\theta}^\star - \hat{\theta})^2$$

προκύπτει ότι $\hat{V}_2 \geq \hat{V}_1$.

Όταν οι εκτιμητές $\hat{\theta}_1, \ldots, \hat{\theta}_K$ είναι ασυσχέτιστοι και έχουν την ίδια αναμενόμενη τιμή, τότε το $\hat{V}_1$ είναι αμερόληπτος εκτιμητής του $V(\hat{\theta}^\star)$.

Αμφότερα τα $\hat{V}_1$ και $\hat{V}_2$ χρησιμεύουν επίσης για την εκτίμηση του $V(\hat{\theta})$, υπο την υπόθεση ότι τα $V(\hat{\theta}^\star)$ και $V(\hat{\theta})$ είναι περίπου ίσα.

Οι προσεγγιστικοί εκτιμητές διακύμανσης στην συνέχεια του κεφαλαίου είναι της μορφής $\hat{V}_1$ και $\hat{V}_2$. Τα $\hat{\theta}_1, \ldots, \hat{\theta}_K$ είναι συνήθως συσχετισμένα, και αμφότερα τα $\hat{V}_1$ και $\hat{V}_2$ είναι τότε μή αμερόληπτοι εκτιμητές του $V(\hat{\theta}^\star)$ και του $V(\hat{\theta})$.

## Η μέθοδος των Τυχαίων Ομάδων (Random Groups)

Έστω ότι ένα τυχαίο $s$ δείγμα απο πληθυσμό $U$ διαμερίζεται σε $K$ μη επικαλυπτόμενες τυχαίες ομάδες (υπο-δείγματα) $s_1, \ldots, s_K$, ώστε $s = \cup_{k=1}^{K} s_k$. Οι ομάδες αυτές δεν είναι (στατιστικά) ανεξάρτητες.

Υποθέτουμε ότι η διαμέριση του $s$ γίνεται με ένα μηχανισμό τυχαιότητας έτσι ώστε κάθε τυχαία ομάδα $s_k$ να έχει τον ίδιο σχεδιασμό δειγματοληψίας με αυτόν του πλήρους δείγματος $s$.

Έστω ότι $\hat{\theta}_1, \ldots, \hat{\theta}_K$ είναι εκτιμητές του $\theta$, όπου $\hat{\theta}_k$ υπολογίζεται μόνο με τα δεδομένα του $s_k$, $k = 1 \ldots, K$.

# Μέθοδοι επαναληπτικής δειγματοληψίας

Θεωρούμε δύο εναλλακτικούς εκτιμητές του $\theta$: τον εκτιμητή $\hat{\theta}_{RG}$, που είναι ο μέσος

$$\hat{\theta}_{RG} = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}_k,$$

και τον εκτιμητή $\hat{\theta}$ που υπολογίζεται απο το πλήρες δείγμα $s$, χωρίς την διαμέρισή του στις $K$ ομάδες.

Θεωρούμε τους εναλλακτικούς εκτιμητές διακύμανσης:

$$\hat{V}_{RG1} = \frac{1}{K(K-1)} \sum_{k=1}^{K} \left( \hat{\theta}_k - \hat{\theta}_{RG} \right)^2$$

και

$$\hat{V}_{RG2} = \frac{1}{K(K-1)} \sum_{k=1}^{K} \left( \hat{\theta}_k - \hat{\theta} \right)^2$$

Αμφότεροι οι $\hat{V}_{RG1}$ και $\hat{V}_{RG2}$ είναι εκτιμητές του $V(\hat{\theta}_{RG})$ και του $V(\hat{\theta})$, αλλά δέν είναι αμερόληπτοι εκτιμητές.

Η μεροληψία του $\hat{V}_{RG1}$ ως εκτιμητή του $V(\hat{\theta}_{RG})$ δίνεται απο τον τύπο

$$E(\hat{V}_{RG1}) - V(\hat{\theta}_{RG}) = -\frac{1}{K(K-1)} \sum_{k=1}^{K} \sum_{l=1 \; k \neq l}^{K} C(\hat{\theta}_k, \hat{\theta}_l).$$

Αν όλα τα ζεύγη έχουν την ίδια συνδιακύμανση (έστω $C$), τότε $E(\hat{V}_{RG1}) - V(\hat{\theta}_{RG}) = -C$.

Ισχύει ότι $\hat{V}_{RG2} \geq \hat{V}_{RG1}$.

Όταν το δείγμα $s$ επιλέγεται με στρωματική δειγματοληψία, τότε η μέθοδος των τυχαίων ομάδων εφαρμόζεται για κάθε στρώμα ξεχωριστά.

# Μέθοδοι επαναληπτικής δειγματοληψίας

Παρατηρήσεις:

Η μέθοδος των τυχαίων ομάδων είναι υπολογιστικά πολύ απλή.

Σε αρκετές περιπτώσεις, η διαμέριση του δείγματος σε τυχαίες ομάδες με ίδιο σχεδιασμό δειγματοληψίας δέν είναι απλή.

Η εκτίμηση της διακύμανσης του $\theta$ είναι "ασταθής" (unstable), δηλαδή έχει μεγάλη διακύμανση, λόγω του μικρού αριθμού τυχαίων ομάδων που χρησιμοποιούνται στην πράξη.

Στην πράξη η χρήση αυτής της μεθόδου δεν είναι όσο συχνή όσο άλλες, πιό εξελιγμένες, μέθοδοι που βασίζονται στην έννοια της επαναληπτικής δειγματοληψίας.

## Η μέθοδος Jackknife

Έστω $s$ είναι τυχαίο δείγμα απο πληθυσμό $U$, και $\hat{\theta}$ είναι εκτιμητής μιας παραμέτρου $\theta$.

Το $s$ διαμερίζεται σε $K$ τυχαίες ομάδες $s_1, \ldots, s_K$ ίσου μεγέθους $m = n/K$. Οι ομάδες αυτές συνιστούν τυχαία δείγματα απο το πλήρες δείγμα.

Υποθέτουμε ότι η επιλογή των υπο-δειγμάτων $s_1, \ldots, s_K$ γίνεται με απλή τυχαία δειγματοληψία, έστω και αν το πλήρες δείγμα $s$ δεν έχει επιλεγεί με απλή τυχαία δειγματοληψία.

# Μέθοδοι επαναληπτικής δειγματοληψίας

Για κάθε ομάδα $k$, $k = 1, \ldots, K$, υπολογίζουμε το $\hat{\theta}_{(k)}$, που είναι εκτιμητής του ιδίου τύπου με το $\hat{\theta}$, αλλά με τα δεδομένα που απομένουν μετά την παράλειψη της ομάδας $k$, δηλαδή με τα δεδομένα στο $s - s_k$.

Σημείωση 1: Τα διαφορετικά σύνολα $s - s_k, \quad k = 1, \ldots, K$ επικαλύπτονται.

Σημείωση 2: Στον υπολογισμό του $\hat{\theta}_{(k)}$, τα βάρη $w_i$ για όλα τα $i \in s - s_k$ πολλαπλασιάζονται με $K/(K-1)$ για αντιστάθμισμα στην έλλειψη του $s_k$.

Για $k = 1, \ldots, K$ ορίζουμε την "ψευδοτιμή"

$$\hat{\theta}_k = k\hat{\theta} - (K-1)\hat{\theta}_{(k)}$$

Ο εκτιμητής Jackknife του $\theta$ ορίζεται ως ο μέσος των ψευδοτιμών $\hat{\theta}_k$

$$\hat{\theta}_{JK} = \frac{1}{K}\sum_{k=1}^{K}\hat{\theta}_k$$

Ο εκτιμητής Jackknife διακύμανσης του $\hat{\theta}_{JK}$ είναι

$$\begin{aligned} \hat{V}_{JK} &= \frac{1}{K(K-1)} \sum_{k=1}^{K} \left( \hat{\theta}_k - \hat{\theta}_{JK} \right)^2 \\ &= \frac{K-1}{K} \sum_{k=1}^{K} \left( \hat{\theta}_{(k)} - \hat{\bar{\theta}} \right)^2 \end{aligned}$$

όπου $\hat{\bar{\theta}} = \sum_{k=1}^{K} \hat{\theta}_{(k)}/K$. Το $\hat{V}_{JK}$ χρησιμοποιείται ως εκτιμητής του $V(\hat{\theta})$, καθώς και του $V(\hat{\theta}_{JK})$.

Για καλή ακρίβεια του εκτιμητή $\hat{V}_{JK}$ απαιτείται ικανός αριθμός ομάδων (μεγάλο $K$). Ο μέγιστος δυνατός αριθμός ομάδων αντιστοιχεί στην ειδική περίπτωση όπου $K = n, m = 1$.

# Μέθοδοι επαναληπτικής δειγματοληψίας

Παράδειγμα: Δείγμα $s$ μεγέθους $n$ με απλή τυχαία δειγματοληψία.

$$\pi = n/N, \quad w_i = N/n, \quad \theta = Y, \quad \hat{\theta} = \hat{Y} = \sum_s w_i y_i = (N/n) \sum_s y_i.$$

Έστω $K$ υπο-δείγματα $s_1, \ldots, s_K$ ίσου μεγέθους $m = n/K$. Ισχύει $K/(K-1) = n/(n-m)$.

$$\hat{\theta}_{(k)} = \frac{K}{K-1} \frac{N}{n} \sum_{s-s_k} y_i = \frac{N}{n-m} \sum_{s-s_k} y_i = N\hat{\bar{Y}}_{s-s_k}$$

$$\hat{\theta}_k = K\hat{\theta} - (K-1)\hat{\theta}_{(k)} = \frac{N}{m} \sum_{s_k} y_i = N\hat{\bar{Y}}_{s_k}$$

$$\hat{\theta}_{JK} = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}_k = \frac{N}{n} \sum_s y_i = N\hat{\bar{Y}}_s \quad (= \hat{\theta})$$

# Μέθοδοι επαναληπτικής δειγματοληψίας

$$\hat{V}_{JK} = \frac{1}{K(K-1)} \sum_{k=1}^{K} \left( \hat{\theta}_k - \hat{\theta}_{JK} \right)^2 = \frac{N^2}{K(K-1)} \sum_{k=1}^{K} \left( \hat{\bar{Y}}_{s_k} - \hat{\bar{Y}}_s \right)^2$$

$$= \frac{N^2(K-1)}{K} \sum_{k=1}^{K} \left( \hat{\bar{Y}}_{s-s_k} - \hat{\bar{Y}}_s \right)^2$$

Όταν $K = n$ (και $m = 1$), τότε

$$\hat{V}_{JK} = \frac{N^2}{n(n-1)} \sum_{s} \left( y_i - \hat{\bar{Y}}_s \right)^2 = \frac{1}{1-f} \hat{V}(\hat{Y}),$$

όπου $\hat{V}(\hat{Y})$ είναι ο αμερόληπτος εκτιμητής του $V(\hat{Y})$ σύμφωνα με τον γενικό τύπο. Σε αυτή την περίπτωση, η προσεγγιστική διακύμανση $\hat{V}_{JK}$ είναι μεγαλύτερη απο την διακύμανση $\hat{V}(\hat{Y})$, με την διαφορά να τείνει στο μηδέν όταν το $f = n/N$ τείνει στο μηδέν.

Σε συνέχεια του παραδείγματος, έστω τώρα ο (εναλλακτικός του $\hat{Y}$) εκτιμητής λόγου $\hat{\theta} = \hat{R} = (\hat{Y}/\hat{X})X = (\sum_s y_i / \sum_s x_i)X$, για απλή τυχαία δειγματοληψία.

Για $K = n$, $m = 1$ έχουμε $\hat{\theta}_{(k)} = (\sum_{s-k} y_i / \sum_{s-k} x_i)X$, $k = 1, \ldots, n$, όπου $s - k$ συμβολίζει το δείγμα χωρίς την μονάδα $k$.
Ο εκτιμητής Jackknife του $V(\hat{\theta})$ είναι

$$\hat{V}_{JK} = \frac{n-1}{n} \sum_{k=1}^{n} (\hat{\theta}_{(k)} - \hat{\bar{\theta}})^2,$$

όπου $\hat{\bar{\theta}} = \sum_{k=1}^{n} \hat{\theta}_{(k)}/n$.

Για εφαρμογή του ανωτέρω παραδείγματος θεωρούμε τα δεδομένα δειγματοληψίας (αρχείο cherry.csv στο e-class) απο ένα πληθυσμό 2967 δέντρων κερασιών. Η δειγματοληψία αυτή, που περιγράφεται στο βιβλίο Lohr (2009), είχε σκοπό την εκτίμηση του ολικού όγκου (σε κυβικά μέτρα) ξυλίας για τον πληθυσμό των κερασιών.

Μετρήσεις ύψους, διαμέτρου και όγκου έγιναν σε δείγμα 31 κερασιών που επιλέγησαν με απλή τυχαία δειγματοληψία. Η ολική διάμετρος $X$ των κερασιών ήταν γνωστή: $X = 41837$ πόδια. Η πολύ ισχυρή συσχέτιση μεταξύ διαμέτρου και όγκου, $\rho = 0,96$, και η γνώση του $X$, δικαιολογεί την χρήση εκτιμητή λόγου για την εκτίμηση του ολικού όγκου $Y$ με βοηθητική μεταβλητή την διάμετρο.

Με κοινό βάρος $N/n = 2967/31 = 95,71$ για όλες τις μονάδες, υπολογίζουμε τις εκτιμήσεις $\hat{Y} = 89517.26$, $\hat{X} = 39307.96$, και $\hat{R} = (\hat{Y}/\hat{X})X = 95272.16$. Επίσης υπολογίζουμε τις διακυμάνσεις $\hat{V}_{JK}(\hat{Y}) = 76729654$ και $\hat{V}_{JK}(\hat{R}) = 30765141$.

Η σχετική διαφορά διακυμάνσεων $(\hat{V}_{JK}(\hat{R}) - \hat{V}_{JK}(\hat{Y}))/\hat{V}_{JK}(\hat{Y}) = -0,599$ δείχνει ότι διακύμανση του εκτιμητή λόγου $\hat{R}$ είναι μικρότερη της διακύμανσης του εκτιμητή $\hat{U}$ σχεδόν κατά 60%.

## Jackknife για δειγματοληψία κατά συστάδες (cluster sampling)

Έστω ότι το δείγμα $s$ αποτελείται απο $K$ clusters, που συνιστούν τυχαίο δείγμα απο ένα πληθυσμό απο $N_c$ clusters.

Στη περίπτωση αυτή τα clusters είναι οι τυχαίες ομάδες του δείγματος στις οποίες εφαρμόζεται η μέθοδος Jackknife.

Τότε αν $\hat{\theta}$ είναι ο εκτιμητής που υπολογίζεται με το πλήρες δείγμα $s$, το $\hat{\theta}_{(k)}$ είναι εκτιμητής που υπολογίζεται χωρίς το cluster $k$.

Σε πολυσταδιακή (multistage) δειγματοληψία η μέθοδος Jackknife εφαρμόζεται στο πρώτο στάδιο δειγματοληψίας, με τις $K$ πρωτογενείς μονάδες δειγματοληψίας (primary sampling units, PSU) να συνιστούν τις τυχαίες ομάδες του δείγματος, άσχετα με τον αριθμό των δευτερογενών ή και τριτογενών μονάδων.
Ο εκτιμητής $\hat{\theta}_{(k)}$ υπολογίζεται με τα δεδομένα που απομένουν μετά την παράλειψη του PSU $k$.

## Jackknife για στρωματική δειγματοληψία (stratified sampling)

Έστω στρωματική δειγματοληψία με $H$ στρώματα, και έστω ότι το δείγμα απο το στρώμα $h$ $(h = 1, \ldots, H)$ διαμερίζεται με τυχαίο τρόπο σε $K_h$ ομάδες (υπο-δείγματα).

Έστω $\hat{\theta}$ είναι ο εκτιμητής του $\theta$ που υπολογίζεται με το πλήρες δείγμα $s$. Τότε $\hat{\theta}_{(hk)}$ είναι ο εκτιμητής του $\theta$ που υπολογίζεται με τα δεδομένα που απομένουν στο δείγμα αφού παραλειφθεί η ομάδα $k$ του στρώματος $h$.

Ο εκτιμητής Jackknife του $V(\hat{\theta})$ είναι

$$\hat{V}_{JK}(\hat{\theta}) = \sum_{h=1}^{H} \frac{K_h - 1}{K_h} \sum_{k=1}^{K_h} \left(\hat{\theta}_{(hk)} - \hat{\theta}\right)^2$$

## Jackknife για στρωματική πολυσταδιακή δειγματοληψία

Στην στρωματική πολυσταδιακή δειγματοληψία η μέθοδος Jackknife εφαρμόζεται ξεχωριστά στο κάθε στρώμα στο πρώτο στάδιο δειγματοληψίας, με τυχαίες ομάδες τα $K_h$ επιλεγμένα PSU στο στρώμα $h$ $(h = 1, \ldots, H)$. Τα $K_h$ PSU , συμβολιζόμενα με $s_{hk}$, συνιστούν το δείγμα $s_h$ του στρώματος $h$, δηλαδή, $s_h = \bigcup_{k=1}^{K_h} s_{hk}$.

Έστω $\hat{\theta}_{(hk)}$ είναι ο εκτιμητής του $\theta$ όταν οι παρατηρήσεις του PSU $k$ του στρώματος $h$ (δηλαδή το $s_{hk}$) παραλείπεται. Στον υπολογισμό του $\hat{\theta}_{(hk)}$ αυξάνονται κατά $K_h/(K_h - 1)$ τα βάρη των παρατηρήσεων στα υπόλοιπα PSU του στρώματος $h$, ενώ τα βάρη των παρατηρήσεων στα υπόλοιπα στρώματα δεν αλλάζουν.

Αναλυτικά, αν το $w_i$ συμβολίζει γενικά το βάρος της παρατήρησης $i$, σε οποιοδήποτε στρώμα και PSU, τότε τα ρυθμισμένα βάρη για τον υπολογισμό του $\hat{\theta}_{(hk)}$ ορίζονται ως εξής:

$$w_{(hk)i} = \begin{cases} w_i, & \text{αν}, i \notin s_h \\ \frac{K_h}{K_h - 1} w_i, & \text{αν}, i \in s_h - s_{hk} \\ 0, & \text{αν} , i \in s_{hk} \end{cases}$$

Ο εκτιμητής Jackknife του $V(\hat{\theta})$ είναι τότε

$$\hat{V}_{JK}(\hat{\theta}) = \sum_{h=1}^{H} \frac{K_h - 1}{K_h} \sum_{k=1}^{K_h} \left(\hat{\theta}_{(hk)} - \hat{\theta}\right)^2$$

Παράδειγμα:

Εκτίμηση ολικού, όπου $\hat{\theta} = \hat{Y} = \sum_s w_i y_i$.
Η επανάληψη $\hat{Y}_{(hk)}$ είναι

$$\hat{Y}_{(hk)} = \sum_s w_{(hk)i} y_i = \sum_{s-s_h} w_i y_i + \sum_{s_h - s_{hk}} \frac{K_h}{K_h - 1} w_i y_i$$

$\hat{V}_{JK}(\hat{Y}) = ?$

# Μέθοδοι επαναληπτικής δειγματοληψίας

Η επαναληπτική διαδικασία παράλειψης ενός υποδείγματος (PSU) $s_{hk}$ για τον υπολογισμό του εκτιμητή $\hat{\theta}_{(hk)}$, διαδοχικά για όλα τα $K_h$ PSU κάθε στρώματος $h$, είναι ανεξάρτητη οποιασδήποτε μεταβλητής και παραμέτρου.

Στην πράξη, η παράλειψη του $s_{hk}$ απο το συνολικό δείγμα $s$ γίνεται έμμεσα με την ρύθμιση $w_i = 0$ για κάθε μονάδα $i \in s_{hk}$, (και με την κατάλληλη ρύθμιση των βαρών για τις υπόλοιπες μονάδες του $s$ όπως περιγράφεται πιο πάνω). Με αυτόν τον τρόπο, δημιουργούνται επαναληπτικά τόσα σύνολα ρυθμισμένων βαρών για όλες τις μονάδες του $s$ όσα είναι συνολικά τα PSU στο $s$.

Τα σύνολα βαρών αυτά θα αποτελέσουν πρόσθετες στήλες στο αρχείο των δεδομένων, και θα χρησιμοποιηθούν για τον υπολογισμό των εκτιμητών $\hat{\theta}_{(hk)}$ οποιασδήποτε παραμέτρου $\theta_{(hk)}$ για οποιαδήποτε μεταβλητή, με το ίδιο τρόπο που υπολογίζεται ο εκτιμητής $\hat{\theta}$.

Παρατηρήσεις:

Η μέθοδος Jackknife είναι χρήσιμη για εκτίμηση διακύμανσης του εκτιμητή κάθε παραμέτρου. Δεν είναι όμως ικανοποιητική όταν η εκτιμώμενη παράμετρος δεν είναι ομαλή συνάρτηση ολικών (π.χ., διάμεσος και quartiles).

Όταν τα στρώματα είναι πολλά, με πολλά PSU έκαστο, η μέθοδος απαιτεί πολλούς υπολογισμούς.

Η μέθοδος μπορεί να εφαρμοστεί και στον εκτιμητή calibration $\hat{\theta}^C$. Προς τούτο, σε κάθε επανάληψη γίνεται πάλι calibration στα ρυθμισμένα βάρη για παραγωγή του εκτιμητή calibration $\hat{\theta}^C_{(k)}$ , που υπολογίζεται πλέον με τα calibrated βάρη.

## Η μέθοδος Bootstrap

Η μέθοδος Bootstrap βασίζεται στην επαναληπτική δειγματοληψία (resampling) για την επιλογή ενός αριθμού αντίτυπων υπο-δειγμάτων (replicates), με επανάθεση, απο το πλήρες δείγμα.
Με αυτά τα υπο-δείγματα υπολογίζονται replicate εκτιμητές του $\theta$, με βάση τους οποίους γίνεται η εκτίμηση της διακύμανσης $V(\hat{\theta})$.

Η διαδικασία περιγράφεται για στρωματική πολυσταδιακή δειγματοληψία.

# Μέθοδοι επαναληπτικής δειγματοληψίας

Έστω $K_h$ ο αριθμός PSU στο δείγμα απο το στρώμα $h$. Σε κάθε επανάληψη της διαδικασίας επιλογής υπο-δειγμάτων, επιλέγονται με απλή τυχαία δειγματοληψία $K_h - 1$ PSU με επανάθεση.

Αυτό γίνεται ανεξάρτητα για κάθε στρώμα $h = 1, \ldots, H$, και έτσι δημιουργείται ένα bootstrap replicate, αποτελούμενο απο $\sum_{h=1}^{H}(K_h - 1)$ PSU.

Η διαδικασία επαναλαμβάνεται $R$ φορές, παράγοντας $R$ bootstrap replicates.

Έστω $m_{hk}(r)$ οι φορές που το PSU $k$ του στρώματος $h$ επιλέγεται στο replicate $r$ $(r = 1, \ldots, R)$. Σημείωση: $0 \leq m_{hk}(r) \leq K_h - 1$.

# Μέθοδοι επαναληπτικής δειγματοληψίας

Για το replicate $r$ γίνεται ρύθμιση των δειγματικών βαρών ως εξής:

$$w_i(r) = w_i \frac{K_h}{K_h - 1} m_{hk}(r), \text{ για την παρατήρηση } i \text{ στο PSU } k \text{ του στρώματος } h.$$

Για κάθε $r$ έστω $\hat{\theta}_{(r)}$ ο replicate εκτιμητής του $\theta$, που υπολογίζεται όπως ο εκτιμητής $\hat{\theta}$ αλλά χρησιμοποιώντας τα βάρη $w_i(r)$ αντί τα αρχικά βάρη $w_i$.

Τότε, ο εκτιμητής bootstrap της διακύμανσης $V(\hat{\theta})$ είναι

$$\hat{V}_B(\hat{\theta}) = \frac{1}{R-1} \sum_{r=1}^{R} \left(\hat{\theta}_{(r)} - \hat{\theta}\right)^2.$$

# Μέθοδοι επαναληπτικής δειγματοληψίας

Όπως στην μέθοδο Jackknife, έτσι και στην μέθοδο Bootstrap η επαναληπτική διαδικασία σχηματισμού υπο-δειγμάτων (replicates) είναι ανεξάρτητη οποιασδήποτε μεταβλητής και παραμέτρου.

Σε κάθε επανάληψη $r$ ρυθμίζονται τα βάρη των μονάδων όλου του δείγματος όπως περιγράφεται πιο πάνω (για τα μή επιλεγμένα PSU τα βάρη είναι μηδέν), και έτσι δημιουργείται ένα νέο σύνολο βαρών για όλο το πλήρες δείγμα που θα χρησιμοποιηθεί για τον υπολογισμό του εκτιμητή $\hat{\theta}_{(r)}$ οποιασδήποτε παραμέτρου $\theta$ για οποιαδήποτε μεταβλητή, με το ίδιο τρόπο που υπολογίζεται ο εκτιμητής $\hat{\theta}$.

# Μέθοδοι επαναληπτικής δειγματοληψίας

Παρατηρήσεις:

Ο αριθμός $R$ είναι αυθαίρετος, αλλά συνήθως είναι $R = 1000$ ή $R = 500$ ή και μικρότερος.

Η μέθοδος Bootstrap δίνει καλή εκτίμηση διακύμανσης και για ομαλές συναρτήσεις ολικών και για μή ομαλές συναρτήσεις (π.χ., quartiles).

Η μέθοδος Bootstrap συνήθως απαιτεί λιγότερους υπολογισμούς απο την μέθοδο Jackknife.

Όταν η διαδικασία εκτίμησης περιλαμβάνει calibration, τότε εκτός απο το calibration που γίνεται στα βάρη του αρχικού (πλήρους) δείγματος για να υπολογιστεί ο εκτιμητής $\hat{\theta}^C$, πρέπει να γίνει calibration στα ρυθμισμένα βάρη $w_i(r)$ για κάθε relicate $r$ για να υπολογιστεί ο replicate εκτιμητής $\hat{\theta}^C_{(r)}$.

# Μή Απόκριση (Nonresponse)

## Μή απόκριση μονάδας (unit nonresponse)

Σε δειγματοληπτικές έρευνες, μερικές δειγματικές μονάδες δεν αποκρίνονται ολικά, με την έννοια ότι δεν συλλέγεται από αυτές καμία από τις ζητούμενες πληροφορίες.

Αιτίες μη-απόκρισης

Οι κυριότερες αιτίες μη-απόκρισης περιλαμβάνουν:

- αδυναμία επικοινωνίας με δειγματικές μονάδες
- απουσία
- αδυναμία απόκρισης (π.χ., γλώσσα, αναλφαβητισμός)
- ασθένεια
- δυσπρόσιτες μονάδες
- άρνηση

# Μή Απόκριση (Nonresponse)

## Επιπτώσεις μη-απόκρισης

Ενδεχόμενη μεροληψία, επειδή παραβιάζεται η βασική αρχή της τυχαιότητας του δείγματος.
Οι μη-αποκρινόμενες μονάδες μπορεί να είναι συστηματικά διαφορετικές από τις αποκρινόμενες, ώστε το αποκριθέν μέρος του δείγματος να μην είναι πλέον αντιπροσωπευτικό του πληθυσμού.
Το αποκριθέν μέρος του δείγματος είναι αντιπροσωπευτικό του μέρους του πληθυσμού που θα αποκρίνετο στην δειγματοληπτική έρευνα, το οποίο σπανίως θα ήταν το ίδιο με ολόκληρο τον πληθυσμό της δειγματοληψίας.

Το μέγεθος της μεροληψίας εξαρτάται από τον συσχετισμό των χαρακτηριστικών των μη αποκρινομένων με τις μεταβλητές της έρευνας, και αυξάνει με το ποσοστό μή-απόκρισης.
Όσο ισχυρότερος ο συσχετισμός μεταξύ της τιμής $y_i$ μιάς μεταβλητής $y$ για την μονάδα $i$ και της πιθανότητας μη-απόκρισης της μονάδας, τόσο μεγαλύτερη η μεροληψία εκτιμήσεων σχετικών με την μεταβλητή αυτή.

# Μή Απόκριση (Nonresponse)

Για παράδειγμα, ας υποθέσουμε ότι σε δειγματοληψία ατομικού εισοδήματος τα άτομα υψηλού εισοδήματος έχουν μεγαλύτερη πιθανότητα μη-απόκρισης από άτομα χαμηλού εισοδήματος.
Το αποτέλεσμα θα είναι ότι για μεταβλητές που συσχετίζονται θετικά με το εισόδημα, τα άτομα με υψηλές τιμές αυτών των μεταβλητών δεν θα αντιπροσωπεύονται επαρκώς στο δείγμα.
Πάντως, ανεξάρτητα από αυτόν τον συσχετισμό, στην περίπτωση εκτίμησης ολικών είναι προφανής η μεροληψία (υπο-εκτίμηση) που προκύπτει από την απώλεια δειγματικών μονάδων.

Ας σημειωθεί, ότι σε οποιαδήποτε περίπτωση μη-απόκρισης το μέγεθος της μεροληψίας δεν μπορεί να εκτιμηθεί.

# Μή Απόκριση (Nonresponse)

Επίπτωση μη απόκρισης ενδέχεται να υπάρχει και για την διακύμανση εκτιμητών. Η απώλεια πληροφορίας λόγω μή απόκρισης (πληροφορία απο λιγότερες δειγματικές μονάδες) θα συνεπάγετο αύξηση της διακύμανσης ενός εκτιμητή αν η διακύμανση της αντίστοιχης μεταβλητής στο αποκριθέν μέρος του δείγματος παρέμενε η ίδια με την διακύμανση στο πλήρες δείγμα (ή ήταν μεγαλύτερη). Ωστόσο αυτό είναι πιό πιθανό να μή συμβαίνει.

Στο ανωτέρω παράδειγμα μεγαλύτερου ποσοστού μή απόκρισης ατόμων υψηλού εισοδήματος, η διακύμανση του εισοδήματος στους αποκριθέντες είναι μικρότερη απο αυτήν στο πλήρες δείγμα. Επομένως, θα υπάρχει μεροληψία στη εκτίμηση (δηλαδή υποεκτίμηση) της διακύμανσης των εκτιμητών που σχετίζονται με το εισόδημα.

Μέτρηση της απόκρισης

Έστω $n_\alpha$ $(n_\alpha < n)$ το μέγεθος του υποσυνόλου του δείγματος για το οποίο υπάρχει απόκριση. Ένα μέτρο απόκρισης δίνεται από το ποσοστό απόκρισης

$$p_a = \frac{n_\alpha}{n}.$$

Το μέτρο αυτό, που συνήθως εκφράζεται *επί τοις εκατό*, δηλώνει τον βαθμό επιτυχίας στην εξασφάλιση απόκρισης από τις μονάδες του επιλεγμένου δείγματος.

# Μή Απόκριση (Nonresponse)

Ένα εναλλακτικό μέτρο απόκρισης δίνεται από το σταθμισμένο ποσοστό απόκρισης

$$\tilde{p}_\alpha = \frac{\sum_{i=1}^{n_\alpha} w_i}{\sum_{i=1}^{n} w_i} = \frac{\hat{N}_\alpha}{\hat{N}},$$

όπου $w_i = 1/\pi_i$ είναι το βάρος της αποκριθείσας μονάδας $i$, και $\hat{N}_\alpha$ είναι εκτίμηση του αριθμού των μονάδων του πληθυσμού που θα αποκρίνονταν δοθείσης της επιλογής τους στο δείγμα.
Το $\tilde{p}_\alpha$ ερμηνεύεται ως μία εκτίμηση της μέσης πιθανότητας απόκρισης στα μέλη του πληθυσμού.

Τα μέτρα απόκρισης $p_\alpha$ και $\tilde{p}_\alpha$ μπορεί να διαφέρουν πολύ μεταξύ τους. Είναι, όμως, ισοδύναμα όταν τα βάρη όλων των μονάδων είναι ίσα.
Τα δύο αυτά μέτρα δεν δίνουν το μέγεθος της μεροληψίας που είναι αποτέλεσμα της μή-απόκρισης. Μικρό μέγεθος μη-απόκρισης μπορεί να προκαλέσει μεγάλη μεροληψία αν ο συσχετισμός μή απόκρισης και μεταβλητών της έρευνας είναι ισχυρός.

# Μή Απόκριση (Nonresponse)

## Αντιμετώπιση του προβλήματος μή απόκρισης

### 1. Πρόληψη μη-απόκρισης

Πρόληψη του προβλήματος κατά τον σχεδιασμό της έρευνας. Παράγοντες που σχετίζονται με ενδεχόμενη μη-απόκριση περιλαμβάνουν:

- αντικείμενο της έρευνας
- σχεδιασμό του ερωτηματολογίου
- επιλογή, εκπαίδευση και εποπτεία των συνεντευκτών
- μέθοδο συλλογής στοιχείων
- χρονική περίοδο και συνθήκες διεξαγωγής της έρευνας
- συχνότητα συλλογής στοιχείων σε περίπτωση επαναληπτικής έρευνας.

# Μή Απόκριση (Nonresponse)

2. Μείωση μη-απόκρισης

Διαδικασίες μείωσης της μη απόκρισης κατά την διεξαγωγή της έρευνας περιλαμβάνουν:

- Call-backs, follow-ups
  Επαναληπτικές προσπάθειες επικοινωνίας
  Διαφορετικές μέρες και ώρες (για προσωπική ή τηλεφωνική συνέντευξη)
  Διαφορετικές μέθοδοι συλλογής (π.χ., τηλεφωνικό follow-up σε ταχυδρομική έρευνα)
- Δειγματοληψία μη-αποκρινομένων
  Επιλέγεται δείγμα μη-αποκρινομένων και καταβάλλεται προσπάθεια να συλλεγούν στοιχεία από όλες τις μονάδες του.
  Αν η διαδικασία είναι επιτυχής μπορεί να εξασφαλίσει αμεροληψία με κατάλληλη μεθοδολογία. Ωστόσο είναι χρονοβόρα και δαπανηρή διαδικασία, και για αυτό σπανίως εφαρμόζεται.

# Μή Απόκριση (Nonresponse)

3. Ρύθμιση βαρών των αποκριθεισών μονάδων

Ο σκοπός μίας τέτοιας ρύθμισης είναι να αυξήσει τα βάρη των αποκριθέντων ώστε αυτοί να αντιπροσωπεύσουν και τους μη αποκριθέντες.
Ας παρατηρήσουμε πρώτα ότι ενώ με πλήρη απόκριση ο εκτιμητής του μεγέθους $N$ του πληθυσμού

$$\hat{N} = \sum_{i=1}^{n} w_i,$$

είναι αμερόληπτος, και για μερικούς δειγματοληπτικούς σχεδιασμούς είναι ακριβώς ίση με το $N$, με το ελλειπές δείγμα η εκτιμήτρια γίνεται

$$\hat{N}_\alpha = \sum_{i=1}^{n_\alpha} w_i,$$

και δίνει υπο-εκτίμηση του $N$ λόγω απώλειας $n - n_\alpha$ μονάδων.

# Μή Απόκριση (Nonresponse)

Η παρατήρηση αυτή υποδείχνει τον τρόπο ρύθμισης των βαρών για αποκατάσταση της αμεροληψίας της εκτιμήτριας του $N$.
Συγκεκριμένα, τα βάρη όλων των αποκριθέντων πολλαπλασιάζονται με τον κοινό παράγοντα $\hat{N}/\hat{N}_\alpha$, που είναι το αντίστροφο του σταθμισμένου ποσοστού απόκρισης $\tilde{p}_\alpha$.
Έτσι, το ρυθμισμένο βάρος της αποκριθείσας μονάδας $i$ είναι

$$\tilde{w}_i = w_i \frac{1}{\tilde{p}_\alpha} = w_i \frac{\hat{N}}{\hat{N}_\alpha} = w_i \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n_\alpha} w_i}.$$

Τότε, ο εκτιμητής που προκύπτει από την χρήση των ρυθμισμένων βαρών είναι

$$\tilde{N} = \sum_{i=1}^{n_\alpha} \tilde{w}_i = \frac{1}{\tilde{p}_\alpha} \sum_{i=1}^{n_\alpha} w_i = \frac{1}{\tilde{p}_\alpha} \hat{N}_\alpha = \hat{N},$$

δηλαδή, ο αμερόληπτος εκτιμητής που θα έδινε το πλήρες δείγμα.

# Μή Απόκριση (Nonresponse)

Η ερμηνεία του $\tilde{p}_\alpha$ ως εκτιμημένης μέσης πιθανότητας απόκρισης των μονάδων του πληθυσμού, οδηγεί στην ενδιαφέρουσα θεώρηση του ρυθμισμένου βάρους κάθε επιλεγμένης και αποκριθείσας μονάδας ως το αντίστροφο της πιθανότητας επιλογής και απόκρισης της μονάδας αυτής, δηλαδή $\tilde{w}_i = 1/\pi_i \tilde{p}_\alpha$.

Τα ρυθμισμένα βάρη χρησιμοποιούνται στην εκτίμηση οποιασδήποτε άλλης παραμέτρου, αλλά δεν εξασφαλίζουν την αμεροληψία της εκτίμησης. Έτσι, ο εκτιμητής του ολικού $Y$

$$\tilde{Y} = \sum_{i=1}^{n_\alpha} \tilde{w}_i y_i = \frac{1}{\tilde{p}_\alpha} \sum_{i=1}^{n_\alpha} w_i y_i$$

δεν είναι αμερόληπτος, αλλά η μεροληψία του ίσως είναι μικρότερη από εκείνη που θα προέκυπτε χωρίς την ρύθμιση των βαρών.

# Μή Απόκριση (Nonresponse)

Η εκτιμημένη μέση πιθανότητα απόκρισης $\tilde{p}_\alpha$ δεν εξαρτάται απο τις μεταβλητές της έρευνας, και είναι η ίδια για όλες τις αποκριθείσες μονάδες του δείγματος.

Αν οι πραγματικές πιθανότητες απόκρισης είναι οι ίδιες για όλες τις μονάδες του πληθυσμού, και η απόκριση κάθε μονάδας είναι ανεξάρτητη απο την απόκριση των άλλων μονάδων, τότε οι μή αποκριθέντες είναι ως να έχουν επιλεγεί τυχαία απο το δείγμα, και οι αποκριθέντες συνιστούν ένα αντιπροσωπευτικό δείγμα.
Η υπόθεση τέτοιου μηχανισμού απόκρισης γίνεται σιωπηρά όταν η μή απόκριση δεν λαμβάνεται υπόψη.

Στη μη ρεαλιστική αυτή περίπτωση, τα ρυθμισμένα βάρη παράγουν αμερόληπτο εκτιμητή ολικού $\tilde{Y} = \sum_{i=1}^{n_\alpha} \tilde{w}_i y_i$ για κάθε μεταβλητή.

## Μή Απόκριση (Nonresponse)

Πιό ρεαλιστική είναι η υπόθεση ότι για ένα διαμερισμό του δείγματος σε διαφορετικές τάξεις σύμφωνα με κάποιο(α) χαρακτηριστικό(ά), οι πιθανότητες απόκρισης είναι (περίπου) ίδιες για μονάδες που ανήκουν στην ίδια τάξη.

Για κάθε τέτοια τάξη μία εκτίμηση αυτής της πιθανότητας είναι το σταθμισμένο ποσοστό απόκρισης των μονάδων της τάξης, που μπορεί να χρησιμοποιηθεί για την ρύθμιση των βαρών όλων των αποκριθεισών μονάδων της τάξης. Με άλλα λόγια, ένα διαφορετικό $\tilde{p}_\alpha$ χρησιμοποιείται για κάθε τάξη.

Έστω $K$ τέτοιες τάξεις μεγέθους $N_k$, $(k = 1, \ldots, K)$, και $n_{k\alpha}$ οι αποκριθέντες στην τάξη $k$. Τότε

$$\tilde{Y} = \sum_{k=1}^{K} \tilde{Y}_k = \sum_{k=1}^{K} \frac{1}{\tilde{p}_{k\alpha}} \sum_{i=1}^{n_{k\alpha}} w_i y_i = \sum_{k=1}^{K} \frac{\hat{N}_k}{\hat{N}_{k\alpha}} \sum_{i=1}^{n_{k\alpha}} w_i y_i.$$

Τα βάρη των αποκριθέντων στην τάξη $k$ αυξάνονται ομοιόμορφα κατά $1/\tilde{p}_{k\alpha} = \hat{N}_k/\hat{N}_{k\alpha}$, ώστε να αντιπροσωπεύονται στο δείγμα και οι μή αποκριθέντες της τάξης αυτής.

# Μή Απόκριση (Nonresponse)

Παράδειγμα: Ας υποθέσουμε ότι για κάθε μονάδα ενός δείγματος ατόμων η ηλικία είναι γνωστή, και ότι το δείγμα διαμερίζεται σε τάξεις ηλικίας όπως φαίνεται στον πίνακα.

| | Ηλικία | | | | | |
|---|---|---|---|---|---|---|
| | 15-24 | 25-34 | 35-44 | 45-64 | 65+ | ολικός |
| $n_k$ | 202 | 220 | 180 | 195 | 203 | 1000 |
| $n_{k\alpha}$ | 124 | 187 | 162 | 187 | 203 | 863 |
| $\sum_{i=1}^{n_k} w_i$ | 30322 | 33013 | 27046 | 29272 | 30451 | 150104 |
| $\sum_{i=1}^{n_{k\alpha}} w_i$ | 18693 | 28143 | 24371 | 28138 | 30451 | 129796 |
| $\tilde{p}_{k\alpha}$ | 0,6165 | 0,8525 | 0,9011 | 0,9613 | 1,000 | |
| $1/\tilde{p}_{k\alpha}$ | 1,622 | 1,173 | 1,110 | 1,040 | 1,000 | |

Το βάρος κάθε αποκρινόμενου ηλικίας μεταξύ 15 και 24 πολλαπλασιάζεται με 1,622, και παρομοίως για τα βάρη αποκρινομένων στις άλλες ηλικιακές τάξεις. Στην τάξη 65+ δεν υπήρξε μή απόκριση, και τα βάρη δεν άλλαξαν. Ας σημειωθεί ότι $1/\tilde{p}_\alpha = \hat{N}/\hat{N}_\alpha = 1{,}156$.

# Μή Απόκριση (Nonresponse)

Στην πράξη, για σημαντική μείωση της μεροληψίας, οι τάξεις πρέπει να καθοριστούν έτσι ώστε οι μονάδες σε κάθε τάξη να είναι όσο δυνατόν όμοιες, ώς προς κύριες μεταβλητές, και τα σταθμισμένα ποσοστά απόκρισης να διαφέρουν όσο δυνατόν περισσότερο μεταξύ των τάξεων.

Συνήθη χαρακτηριστικά καθορισμού των τάξεων είναι γεωγραφικές μεταβλητές, καθώς και άλλες μεταβλητές που υπάρχουν στο πλαίσιο δειγματοληψίας. Για παράδειγμα, σε δειγματοληψία επιχειρήσεων, μεταβλητές όπως τύπος επιχείρησης και μέγεθος επιχείρησης.

# Μή Απόκριση (Nonresponse)

Επίσης, χαρακτηριστικά που καθορίζονται απο *paradata* δηλαδή, στοιχεία που προκύπτουν απο την διαδικασία δειγματοληψίας. Αυτά μπορεί να είναι στοιχεία που καταγράφουν οι συνεντευκτές, όπως ημέρα και ώρα κλήσης για συνέντευξη, τακτική προσέγγισης, αποτέλεσμα κλήσης, κ.λ.π.
΄λλα στοιχεία είναι, παρατηρήσεις για κάθε νοικοκυριό του δείγματος, όπως τύπος κατοικίας, σύστημα ασφαλείας, ενδείξεις παρουσίας παιδιών, παρατηρήσεις για την γειτονιά, κ.λ.π.

Συχνά, για διευκόλυνση της διαδικασίας, ως τάξεις χρησιμοποιούνται τα στρώματα του δείγματος.

# Μή Απόκριση (Nonresponse)

Η ρύθμιση βαρών για μη-απόκριση ίσως προκαλέσει αύξηση της διακύμανσης των εκτιμήσεων όταν οι τάξεις ρύθμισης είναι πολλές και περιέχουν λίγες δειγματικές μονάδες.

Τότε, οι πιθανότητες απόκρισης δεν εκτιμώνται με ακρίβεια, πράγμα που έχει ως αποτέλεσμα αύξηση της διακύμανσης των εκτιμήσεων. Επίσης, άσχετα με τον αριθμό των τάξεων, μερικές από αυτές μπορεί να απαιτούν μεγάλους συντελεστές ρύθμισης, με αποτέλεσμα αύξηση της διακύμανσης των εκτιμήσεων.

Μέθοδοι διαφορετικής ρύθμισης του βάρους κάθε μονάδας ξεχωριστά, με εκτίμηση της πιθανότητας μή απόκρισής της, που βασίζονται σε χρήση βοηθητικών μεταβλητών και σε υπόθεση ισχύος κάποιων μοντέλων, υπάρχουν στην βιβλιογραφία αλλά σπάνια εφαρμόζονται στην πράξη.

# Μή Απόκριση (Nonresponse)

Ο εκτιμητής

$$\tilde{Y} = \sum_{k=1}^{K} \frac{\hat{N}_k}{\hat{N}_{k\alpha}} \sum_{i=1}^{n_{k\alpha}} w_i y_i.$$

είναι της ίδιας μορφής με τον μεταστρωματικό εκτιμητή. Η διαφορά είναι ότι στην μεταστρωμάτωση τα μεγέθη των μεταστρωμάτων $N_j$ είναι γνωστά, ενώ στην ρύθμιση βαρών των αποκριθέντων κατά τάξεις τα μεγέθη των τάξεων $N_k$ είναι άγνωστα και εκτιμώνται με τα $\hat{N}_k$.

Σημείωση: Στην ρύθμιση βαρών των αποκριθέντων κατά τάξεις οι ρυθμιστικοί παράγοντες $\hat{N}_k/\hat{N}_{k\alpha}$ είναι πάντοτε μεγαλύτεροι απο ένα, ενώ στην μεταστρωμάτωση οι ρυθμιστικοί παράγοντες μπορεί να είναι οποιοσδήποτε θετικός αριθμός (γιατί;)

# Μή Απόκριση (Nonresponse)

Η μεταστρωμάτωση είναι μορφή ρύθμισης των βαρών για μή απόκριση, με τα βάρη του κάθε μεταστρώματος $g$ να πολλαπλασιάζονται με $N_g/\hat{N}_{ga}$, όπου $N_g$ είναι το μέγεθος του μεταστρώματος $g$ και $\hat{N}_{ga}$ είναι η εκτίμηση του $N_g$ που προκύπτει απο του αποκριθέντες του μεταστρώματος αυτού, έτσι ώστε το άθροισμα των ρυθμισμένων βαρών του μεταστρώματος να ισούται με $N_g$ (calibration).

Σε μία δειγματοληψία μπορεί να γίνεται ρύθμιση βαρών κατά τάξεις για διόρθωση μή απόκρισης, και μεταστρωμάτωση με μεταστρώματα που είναι διαφορετικά απο τις τάξεις ή επικαλύπτωνται μερικώς με αυτές. Όπως και στην περίπτωση των τάξεων, οι μονάδες του ίδιου μεταστρώματος θα πρέπει να έχουν περίπου την ίδια πιθανότητα να αποκριθούν, ώστε η μεταστρωμάτωση να έχει ως αποτέλεσμα τη μείωση της μεροληψίας.

**Μερική μή απόκριση (item nonresponse)**

Για μερικές δειγματικές μονάδες μπορεί να υπάρχει μερική απόκριση, με την έννοια ότι δεν αποκρίνονται σε όλες τις ερωτήσεις. Αυτό συμβαίνει όταν ο αποκρινόμενος αρνείται ή παραλείπει ή δεν μπορεί να απαντήσει σε κάποιες ερωτήσεις, ή όταν κάποιες λανθασμένες απαντήσεις δεν ήταν δυνατό να διορθωθούν κατά την διαδικασία του ελέγχου των στοιχείων (input editing).

Η διαδικασία αναπλήρωσης μη διαθέσιμων ή λανθασμένων στοιχείων της έρευνας με κατάλληλα στοιχεία για την δημιουργία ενός πλήρους αρχείου δεδομένων είναι γνωστή με τον όρο **imputation**.

# Μή Απόκριση (Nonresponse)

Υπάρχουν διάφορες μέθοδοι imputation. Καλές μέθοδοι imputation μπορούν να διατηρήσουν γνωστές σχέσεις μεταξύ μεταβλητών και να μειώσουν την μεροληψία λόγω μερικής μη απόκρισης.

Το imputation χρησιμοποιείται μόνο για την αναπλήρωση μη διαθέσιμων ή λανθασμένων στοιχείων, όχι για ολική μη απόκριση.

Το imputation ήταν στο παρελθόν κυρίως μία χειρωνακτική διαδικασία, τώρα πιο συχνά χρησιμοποιούνται αυτοματοποιημένα συστήματα.

Το τεχνητό μικρό σύνολο δεδομένων δειγματοληψίας (πηγή: βιβλίο Lohr (2009)) που φαίνεται στον κατωτέρω πίνακα θα χρησιμοποιηθεί για την εξήγηση διαφορετικών μεθόδων imputation.
Ο αριθμός "1" στις δύο τελευταίες στήλες σημαίνει ότι ο αποκρινόμενος απάντησε ναι στην ερώτηση.

# Μή Απόκριση (Nonresponse)

| Person | Age | Sex | Years of Education | Crime Victim | Violent Crime Victim |
|---|---|---|---|---|---|
| 1 | 47 | M | 16 | 0 | 0 |
| 2 | 45 | F | ? | 1 | 1 |
| 3 | 19 | M | 11 | 0 | 0 |
| 4 | 21 | F | ? | 1 | 1 |
| 5 | 24 | M | 12 | 1 | 1 |
| 6 | 41 | F | ? | 0 | 0 |
| 7 | 36 | M | 20 | 1 | ? |
| 8 | 50 | M | 12 | 0 | 0 |
| 9 | 53 | F | 13 | 0 | ? |
| 10 | 17 | M | 10 | ? | ? |
| 11 | 53 | F | 12 | 0 | 0 |
| 12 | 21 | F | 12 | 0 | 0 |
| 13 | 18 | F | 11 | 1 | ? |
| 14 | 34 | M | 16 | 1 | 0 |
| 15 | 44 | M | 14 | 0 | 0 |
| 16 | 45 | M | 11 | 0 | 0 |
| 17 | 54 | F | 14 | 0 | 0 |
| 18 | 55 | F | 10 | 0 | 0 |
| 19 | 29 | F | 12 | ? | 0 |
| 20 | 32 | F | 10 | 0 | 0 |

## Μέθοδοι Imputation

Ο μέθοδοι αυτές χωρίζονται σε εκείνες που χρησιμοποιούν στοιχεία μόνο από τον μη αποκρινόμενο και άλλα βοηθητικά στοιχεία, και εκείνες που χρησιμοποιούν στοιχεία από άλλους αποκρινόμενους της έρευνας.

*Επαγωγικό (deductive) imputation*
Η μέθοδος αυτή προσδιορίζει την τιμή που λείπει με βεβαιότητα κάνοντας χρήση λογικών περιορισμών και άλλων στοιχείων τής ίδιας μονάδας, π.χ., μία συνιστώσα που λείπει από ένα άθροισμα. Είναι ο ιδεατός αλλά λιγότερο συχνός τύπος imputation.
Στο παράδειγμα του πίνακα, για το άτομο 9 δεν υπάρχει απάντηση στην τελευταία ερώτηση. Όμως, η απάντηση στην πρωτελευταία ερώτηση λογικά συνεπάγεται ότι η απάντηση που λείπει πρέπει να είναι 0.

# Μή Απόκριση (Nonresponse)

*Ιστορικό imputation*
Αυτή η μέθοδος είναι περισσότερο χρήσιμη σε διαχρονικές έρευνες, ειδικά για μεταβλητές που είναι διαχρονικά ευσταθείς. Χρησιμοποιεί τιμές που αναφέρθηκαν από την ίδια μονάδα (αποκρινόμενο) σε προηγούμενη μέτρηση της έρευνας.

Σε περιπτώσεις που η απόκριση σε προηγούμενη μέτρηση μπορεί να καθορίσει με βεβαιότητα την τρέχουσα απόκριση, η μέθοδος αυτή είναι ειδική περίπτωση επαγωγικού imputation.

# Μή Απόκριση (Nonresponse)

*Imputation με μέση τιμή*

Με την απλή αυτή μέθοδο γίνεται αναπλήρωση της τιμής που λείπει για κάποια μεταβλητή με τον μέσο όρο τιμών των αποκριθέντων δειγματικών μονάδων. Για την ίδια μεταβλητή, ο μέσος όρος αυτός χρησιμοποιείται για κάθε μονάδα για την οποία γίνεται imputation.

Η μέθοδος χρησιμοποιείται μόνο για ποσοτικές μεταβλητές, και συχνά χρησιμοποιείται ως τελευταία λύση. Διατηρεί ολικούς και μέσους όρους, αλλά αλλοιώνει κατανομές και συσχετισμούς πολυμεταβλητών. Επίσης, προκαλεί τεχνητή συγκέντρωση στον μέσο όρο, με αποτέλεσμα την τεχνητή μείωση της διακύμανσης των τιμών της μεταβλητής για την οποία έχει γίνει imputation.

Στο παράδειγμα, για τα άτομα 2, 4 και 6 λείπει η τιμή για τα έτη εκπαίδευσης. Και για τα τρία αυτά άτομα γίνεται αναπλήρωση της τιμής που λείπει με τον μέσο όρο των 17 αποκριθέντων στην συγκεκριμένη ερώτηση: 12,70. Μετά το imputation ο μέσος όρος για τα 20 άτομα είναι ο ίδιος με τον μέσο όρο των αποκριθέντων.

# Μή Απόκριση (Nonresponse)

*Imputation με μέση τιμή κατά τάξεις*

Το συνολικό δείγμα χωρίζεται σε τάξεις έτσι ώστε μονάδες της ίδιας τάξης να είναι παρόμοιες. Σε κάθε τέτοια τάξη γίνεται imputation χρησιμοποιώντας τον μέσο όρο τιμών των αποκριθέντων της τάξης. Η αλλοίωση της κατανομής των τιμών της μεταβλητής και η τεχνητή μείωση της διακύμανσης, είναι λιγότερο σοβαρές από ό,τι στην προηγούμενη μέθοδο.

Για τα δεδομένα του παραδείγματος, το δείγμα χωρίζεται σε τέσσερες τάξεις κατά ηλικία και φύλο.

| | | Age | |
|---|---|---|---|
| | | $\leq 34$ | $\geq 35$ |
| | | Persons | Persons |
| Sex | | 3,5,10,14 | 1,7,8,15,16 |
| | F | Persons | Persons |
| | | 4,12,13,19,20 | 2,6,9,11,17,18 |

# Μή Απόκριση (Nonresponse)

Για τα άτομα 2 και 6, η τιμή που λείπει για τα έτη εκπαίδευσης αναπληρώνεται με τον μέσο όρο των τεσσάρων γυναικών ηλικίας ίσης ή μεγαλύτερης των 35 ετών που αποκρίθηκαν στην ερώτηση: 12,25. Για το άτομο 4, η αναπλήρωση γίνεται με τον μέσο όρο των τεσσάρων γυναικών ηλικίας ίσης ή μικρότερης των 34 ετών που αποκρίθηκαν στην ερώτηση: 11,25.

Μετά το imputation, σε κάθε τάξη ο μέσος όρος είναι ο ίδιος με τον μέσο όρο των αποκριθέντων.

# Μή Απόκριση (Nonresponse)

*Hot-deck imputation*
Αυτή είναι μια κατηγορία μεθόδων που δημιουργούν πιο αυθεντική διακύμανση των αναπληρωμένων τιμών από ό,τι το imputation με μέσο όρο. Οι μέθοδοι αυτές δίνουν πάντα εφικτές τιμές επειδή οι τιμές ανήκουν σε αποκριθέντες της ίδιας έρευνας.

*Τυχαίο συνολικό hot-deck imputation*
Με αυτή την μέθοδο η τιμή που λείπει αναπληρώνεται με την τιμή ενός "δότη", που επιλέγεται τυχαία από τους αποκριθέντες. Αυτή είναι η απλούστερη μορφή hot-deck imputation.

*Τυχαίο hot-deck imputation κατά τάξεις*
Αυτή είναι παραλλαγή της προηγούμενης μεθόδου, στην οποία σχηματίζονται κατάλληλες τάξεις δειγματικών μονάδων, όπως και στο imputation με μέση τιμή κατά τάξεις. Για μονάδα μίας τάξης, ο δότης επιλέγεται τυχαία από τους αποκριθέντες της ίδιας τάξης.

# Μή Απόκριση (Nonresponse)

Στο παράδειγμά μας, για το άτομο 10 λείπουν οι απαντήσεις στις δύο τελευταίες ερωτήσεις. Στην τάξη του, τα άτομα 3, 5 και 14 έχουν αποκριθεί και στις δύο ερωτήσεις, και έτσι ένα απο τα τρία επιλέγεται τυχαία ως δότης.

*Ακολουθιακό hot deck imputation*
Με την μέθοδο αυτή γίνεται αναπλήρωση της τιμής που λείπει με την αντίστοιχη τιμή από την τελευταία αποκριθείσα μονάδα της ίδιας τάξης που προηγείται στο αρχείο δεδομένων. Το πλεονέκτημα αυτής της μη τυχαίας διαδικασίας είναι η ευκολία της ακολουθιακής επεξεργασίας του αρχείου. Το μειονέκτημά της είναι ότι συχνά κάνει πολλαπλή χρήση του ιδίου δότη.

Στο παράδειγμα, για το άτομο 19 λείπει η απάντηση στην πρωτελευταία ερώτηση. Το άτομο 13 ήταν το τελευταίο της ίδιας τάξης που αποκρίθηκε, και έτσι η αναπλήρωση γίνεται με την τιμή 1.

# Μή Απόκριση (Nonresponse)

*Nearest-Neighbor Hot-Deck Imputation*
Η τιμή που λείπει αναπληρώνεται με την αντίστοιχη τιμή ενός αποκριθέντος που είναι ο "πλησιέστερος", σύμφωνα με κάποια συνάρτηση απόστασης μεταξύ παρατηρήσεων που ορίζεται με γνωστές βοηθητικές πληροφορίες.

Για παράδειγμα, αν ηλικία και φύλο χρησιμοποιηθούν για την συνάρτηση απόστασης, τότε ο αποκριθείς(είσα) του ιδίου φύλου και της πλησιέστερης ηλικίας επιλέγεται ως δότης.

Έτσι στο παράδειγμά μας, οι τιμές που λείπουν για το άτομο 10 αναπληρώνονται με τις αντίστοιχες του ατόμου 3, που είναι του ιδίου φύλου και της πλησιέστερης ηλικίας με το άτομο 10.

# Μή Απόκριση (Nonresponse)

*Regression Imputation*

Αυτή η μέθοδος χρησιμοποιεί παλινδρόμηση (regression) της μεταβλητής για την οποία χρειάζεται imputation σε ένα σύνολο μεταβλητών για τις οποίες υπάρχει απόκριση απο όλες τις μονάδες. Η εξίσωση παλινδρόμησης χρησιμοποιείται μετά για πρόβλεψη των τιμών που λείπουν.

Στο παράδειγμα, έχουμε μόνο 18 συμπληρωμένες αποκρίσεις για την μεταβλητή "crime victim" (μάλλον λίγες για προσαρμογή μοντέλου στα δεδομένα μας), αλλά μία λογιστική παλινδρόμηση της απόκρισης με επεξηγηματική μεταβλητή την ηλικία δίνει το ακόλουθο μοντέλο για την προβλεπόμενη πιθανότητα ένα άτομο να είναι "crime victim", $\hat{p}$,

$$\log \frac{\hat{p}}{1-\hat{p}} = 2,5643 - 0,0896 \times age.$$

Η προβλεφθείσα πιθανότητα για ένα άτομο 17 ετών να είναι "crime victim" είναι 0,74. Επειδή αυτή η πιθανότητα είναι μεγαλύτερη απο το προκαθορισμένο όριο 0,5, η τιμή 1 αναπληρώνει την τιμή που λείπει για το άτομο 10.

# Μή Απόκριση (Nonresponse)

*Cold-deck imputation*

Σε αυτή την μέθοδο, οι αναπληρούμενες τιμές είναι απο προηγούμενη δειγματοληπτική έρευνα ή από ιστορικά στοιχεία.

# Προτεινόμενη βιβλιογραφία


Lohr, S.L. (2009). *Sampling: Design and Analysis*. Second Edition, Brooks/Cole. Cengage Learning.

Särndal, C-E, Swensson, B., Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer.

Lumley, T. (2010). *Complex Surveys. A Guide to Analysis Using R*. Wiley.