6. ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΚΑΤΑ ΟΜΑΔΕΣ (Cluster Sampling)

Από την θεωρία που αναπτύχθηκε στα προηγούμενα κεφάλαια, φαίνεται ότι μια αλλαγή στον σχεδιασμό της δειγματοληψίας και, κατά συνέπεια, στην μέθοδο εκτίμησης των παραμέτρων ενός πληθυσμού μπορεί να αυξήσει την ακρίβεια των αποτελεσμάτων, χωρίς απαραίτητα να προϋποθέτει αύξηση του μεγέθους του δείγματος. Το συμπέρασμα είναι, επομένως, ότι μεταξύ δυο δειγματοληπτικών τεχνικών, που χρησιμοποιούν ισομεγέθη δείγματα, προτιμητέα είναι αυτή που οδηγεί στο μικρότερο τυπικό σφάλμα. Φυσικά, σημαντικό ρόλο στην επιλογή της κατάλληλης τεχνικής παίζει και το κόστος ανά δειγματοληπτική μονάδα. Αυτό, που ουσιαστικά επιδιώκεται, είναι η μέγιστη δυνατή ακρίβεια με το ελάχιστο δυνατό κόστος. Η επιδίωξη χαμηλότερου κόστους οδηγεί στην λεγόμενη **δειγματοληψία κατά ομάδες (cluster sampling).**

Για την ευκολότερη κατανόηση αυτής της δειγματοληπτικής τεχνικής, ας θεωρήσουμε, για παράδειγμα, τους εξής δυο ενδεχόμενους τρόπους επιλογής ενός δείγματος 200 νοικοκυριών από ένα πληθυσμό 20000 νοικοκυριών μιας πόλης:

1. Αν υπάρχει κατάλογος αυτών των 20000 νοικοκυριών, επιλέγουμε ένα απλό τυχαίο δείγμα μεγέθους 200 από τον κατάλογο.

2. **(α)** Διαιρούμε την πόλη σε 400 περιοχές 50 νοικοκυριών. Επιλέγουμε ένα απλό τυχαίο δείγμα 4 περιοχών και περιλαμβάνουμε στο δείγμα όλα τα νοικοκυριά που ανήκουν σ' αυτές.

(β) Εναλλακτικά, διαιρούμε την πόλη σε 400 περιοχές N_1, N_2,…, N_{400} νοικοκυριών, αντίστοιχα και επιλέγουμε με

αναλογική ή βέλτιστη στρωματοποιημένη δειγματοληψία ένα δείγμα 200 νοικοκυριών από τις περιοχές αυτές.

Γενικά, όπως θα δούμε στις εφαρμογές, η μέθοδος 1 οδηγεί σε μικρότερο τυπικό σφάλμα από αυτό της μεθόδου 2. Όμως, η πρώτη συνεπάγεται μεγαλύτερο κόστος γιατί το δείγμα θα είναι περισσότερο ″απλωμένο″ γεωγραφικά και, επομένως, το κόστος μετάβασης για τον εντοπισμό των μονάδων που επιλέγονται και το κόστος εποπτείας είναι υψηλότερο. Στην δεύτερη περίπτωση, το δείγμα είναι λιγότερο ″απλωμένο″ μια και οι δειγματοληπτούμενες μονάδες είναι σύνθετες. Ουσιαστικά, η δειγματοληψία γίνεται σε πρώτο στάδιο από ένα σύνθετο πληθυσμό, του οποίου οι μονάδες είναι ομάδες μονάδων του αρχικού πληθυσμού. Ακόμη και στην υποπερίπτωση (β), το δεύτερο στάδιο της δειγματοληψίας διεξάγεται σε γεωγραφικά μικρότερη έκταση, με αποτέλεσμα την μείωση του κόστους διεξαγωγής και εποπτείας.

Από την στιγμή που οι ″ομάδες″ (clusters) επιλέγονται, ο ερευνητής μπορεί να περιλάβει στο δείγμα όλες τις στοιχειώδεις μονάδες του αρχικού πληθυσμού που ανήκουν σ' αυτές (περίπτωση 2(α)) ή να επιλέξει ένα δείγμα μικρότερων σύνθετων ομάδων ή και στοιχειωδών μονάδων από τις αρχικές σύνθετες μονάδες (περίπτωση 2(β)).

Ορισμός 6.1: **Δειγματοληψία κατά ομάδες σε ένα στάδιο (single-stage cluster sampling)** ονομάζεται η δειγματοληπτική τεχνική, η οποία διαιρεί τις στοιχειώδεις μονάδες του πληθυσμού σε **ομάδες (clusters)**, επιλέγει ένα δείγμα των ομάδων αυτών και περιλαμβάνει στο τελικό δείγμα των στοιχειωδών μονάδων όλες τις στοιχειώδεις μονάδες που ανήκουν στις ομάδες αυτές.

Ορισμός 6.2: **Δειγματοληψία κατά ομάδες σε πολλά στάδια (multi-stage cluster sampling)** ονομάζεται η δειγματοληπτική τεχνική όταν, μετά το πρώτο στάδιο, επιλέγονται δείγματα μικρότερων και μικρότερων ομάδων με τελικό στάδιο την επιλογή δείγματος στοιχειωδών μονάδων (ή την περίληψη όλων των στοιχειωδών μονάδων της τελευταίας κατηγορίας σύνθετων ομάδων).

Παρατήρηση 1: Η περίπτωση 2(α) αναφέρεται σε δειγματοληψία κατά ομάδες σε ένα στάδιο (single-stage cluster sampling), ενώ η 2(β) σε δειγματοληψία κατά ομάδες σε δυο στάδια (two-stage cluster sampling).
Παρατήρηση 2: Αν, κατά τα διάφορα στάδια της δειγματοληψίας, οι σύνθετες μονάδες επιλέγονται με απλή τυχαία δειγματοληψία, το δειγματοληπτικό σχήμα ονομάζεται **απλή δειγματοληψία κατά ομάδες (simple cluster sampling)** σε ένα ή περισσότερα στάδια.

Παρατήρηση 3: Όταν οι σύνθετες μονάδες είναι ομάδες στοιχείων βασισμένες σε γεωγραφικές περιοχές, ο σχεδιασμός ονομάζεται **δειγματοληψία κατά περιοχές (area sampling)**.

Παράδειγμα: Έστω ότι πρόκειται να εκτιμηθεί ο μέσος μισθός (σε ευρώ) του πληθυσμού του πίνακα 6.1 με βάση ένα δείγμα μεγέθους 4.

Πίνακας 6.1

Άτομο	**Μισθός (y)**
Α	1300
Β	6300
Γ	3100
Δ	2000
Ε	3600
Ζ	2200
Η	1800
Θ	2700
Ι	1500
Κ	900
Λ	4800
Μ	1900
Σύνολο	32100
Μέση τιμή μ	2675

Να υπολογισθεί το τυπικό σφάλμα του μέσου $\bar{X}_4$ του δείγματος στην περίπτωση απλής τυχαίας δειγματοληψίας και να συγκριθεί με το τυπικό σφάλμα του μέσου $\hat{\mu}_c$ ενός δείγματος που έχει ληφθεί με δειγματοληψία κατά ομάδες σε ένα στάδιο θεωρώντας την εξής διαίρεση του πληθυσμού σε ομάδες (clusters):

Ομάδα (cluster)	Άτομο	Μισθός (y)
1	Α	1300
	Δ	2000
	Κ	900
	Μ	1900
2	Ζ	2200
	Η	1800
	Θ	2700
	Ι	1500
3	Β	6300
	Γ	3100
	Ε	3600
	Λ	4800

Λύση:
Περίπτωση δειγματοληψίας κατά ομάδες: Προφανώς, θα πρέπει να επιλέξουμε με απλή τυχαία δειγματοληψία μια ομάδα (ένα cluster) και να περιλάβουμε στο δείγμα μας όλα τα άτομα της ομάδας αυτής. Υπάρχουν 3 δυνατά τέτοια δείγματα (1,2 και 3) και οι μέσοι $\bar{U}_i$, i = 1, 2, 3 των δειγμάτων αυτών είναι

Δείγμα	1	2	3
Μέσος $\bar{U}_i$	1525	2050	4450

Αν, λοιπόν, επιλεγεί μια ομάδα με απλό τυχαίο τρόπο και θεωρήσουμε ως δείγμα μας τα τέσσερα στοιχεία που την απαρτίζουν, ο μέσος του δείγματος αυτού θα είναι η ζητούμενη εκτιμήτρια $\hat{\mu}_c$ του μέσου μισθού μ του πληθυσμού. Πιο συγκεκριμένα, αν επιλεγεί η i_0 ομάδα, τότε $\hat{\mu}_c = \bar{U}_{i_0}$. Ισχύει, επομένως, ότι

$$E(\hat{\mu}_c) = (1525+2050+4450)/3 = 2675.$$

Δηλαδή, η αναμενόμενη τιμή της θεωρηθείσας εκτιμήτριας συμπίπτει με την τιμή του μέσου μισθού που θέλουμε να εκτιμήσουμε. Άρα, η εκτιμήτρια $\hat{\mu}_c$ είναι αμερόληπτη εκτιμήτρια του μέσου μισθού μ. η διασπορά του μέσου $\hat{\mu}_c$ είναι ίση με

$$\sigma^2_{\hat{\mu}_c} = \sum_{i=1}^{3} (\bar{U}_i - 2675)^2 / 3 = 1621250$$

και, επομένως, το τυπικό σφάλμα είναι

$$\sigma_{\hat{\mu}_c} = \sqrt{1621250} = 1273.$$

Περίπτωση απλής τυχαίας δειγματοληψίας: Η διασπορά του μέσου $\bar{X}_4$ ενός απλού τυχαίου δείγματος μεγέθους 4 είναι ίση με

$$\sigma^2_{\bar{X}_4} = \frac{\sigma^2}{4}\left(1 - \frac{4}{12}\right),$$

όπου σ^2 είναι η διασπορά ολόκληρου του πληθυσμού και είναι ίση με

$$\sigma^2 = \sum_{i=1}^{12} (y_i - 2675)^2 / 11 = 2469318.2.$$

Δηλαδή, τελικά,

$$\sigma^2_{\bar{X}_4} = 411553,$$

και, επομένως,

$$\sigma_{\bar{X}_4} = 641.52 .$$

Είναι, δηλαδή, το τυπικό σφάλμα της εκτιμήτριας $\hat{\mu}_c$ 1.98 φορές μεγαλύτερο από το τυπικό σφάλμα του μέσου $\bar{X}_4$.

Παρατήρηση: Είναι δυνατόν να χωρισθεί ο πληθυσμός σε κατάλληλες ομάδες και να επιλεγεί ένα δείγμα με διασπορά μικρότερη ή ίση της διασποράς ενός απλού τυχαίου δείγματος του αυτού μεγέθους. Αν, για παράδειγμα, οι ομάδες είχαν ορισθεί ως εξής:

Ομάδα	Άτομα	Μέσος μισθός $\bar{U}_i$
1	ΕΓΗΚ	2350
2	ΑΘΛΜ	2675
3	ΒΖΔΙ	3000

θα είχαμε και πάλι $E(\hat{\mu}_c) = 2675$, αλλά $\sigma^2_{\hat{\mu}_c} = 70417$, δηλαδή, $\sigma_{\hat{\mu}_c} = 265$.

Εδώ, οι ομάδες είναι λιγότερο ομοιογενείς από αυτές της προηγούμενης διαίρεσης. Γενικά, όσο λιγότερο ομοιογενείς είναι οι ομάδες, τόσο μικρότερο τυπικό σφάλμα επιτυγχάνεται. Στην πράξη βέβαια, αυτό είναι αδύνατο να ελεγχθεί. Πάντως, ο βασικός στόχος της δειγματοληψίας κατά ομάδες δεν είναι η επίτευξη του πιο αξιόπιστου δείγματος στοιχειωδών μονάδων, αλλά η επίτευξη των πιο αξιόπιστων αποτελεσμάτων ανά μονάδα κόστους.

6.1 Εκτίμηση της Μέσης Τιμής

Από τα προηγούμενα, προκύπτει ότι η δειγματοληψία κατά ομάδες προϋποθέτει ότι ο πληθυσμός διαιρείται σε υποπληθυσμούς. Μερικοί

από αυτούς εκπροσωπούνται στο δείγμα είτε εξ ολοκλήρου (ένα στάδιο) είτε εν μέρει μέσω κάποιου δείγματος (δύο ή περισσότερα στάδια). Η διαφορά της απλής δειγματοληψίας κατά ομάδες από την απλή στρωματοποιημένη δειγματοληψία έγκειται στο ότι, στην πρώτη, **μόνο ορισμένα** από τα στρώματα εκπροσωπούνται στο δείγμα, ενώ, στην δεύτερη, εκπροσωπούνται **όλα** τα στρώματα και πάντα μέσω απλών τυχαίων υποδειγμάτων.

Στην περίπτωση της απλής δειγματοληψίας κατά ομάδες, ο πληθυσμός $\{y_1, y_2, \ldots, y_N\}$ διαιρείται σε M ομάδες (υποπληθυσμούς, clusters) $\upsilon_1, \upsilon_2, \ldots, \upsilon_M$ μεγέθους $N_1, N_2, \ldots, N_M$, αντίστοιχα, όπου $\sum_{i=1}^{M} N_i = N$. Δηλαδή, αν με y_{ij} συμβολίσουμε την τιμή της j μονάδας της i ομάδας, τότε

$$\upsilon_i = \{y_{i1}, y_{i2}, \ldots, y_{iN_i}\}, \; i = 1, 2, \ldots, M$$

και, επομένως, ο αρχικός πληθυσμός είναι η ένωση $\upsilon_1 \cup \upsilon_2 \cup \ldots \cup \upsilon_M$. Προφανώς,

$$\bar{N} = \sum_{i=1}^{M} N_i / M = N/M \qquad (6.1.1)$$

συμβολίζει το μέσο μέγεθος των ομάδων αυτών.

Έστω $\{U_1, U_2, \ldots, U_m\}$ ένα απλό τυχαίο δείγμα m ομάδων από τον πληθυσμό M ομάδων. Το j στοιχείο της i επιλεγείσας ομάδας θα συμβολίζεται με U_{ij}, $j = 1, 2, \ldots, N_i'$, $i = 1, 2, \ldots, m$. Δηλαδή, $U_i = \{U_{i1}, U_{i2}, \ldots, U_{iN_i'}\}$. Εδώ, N_i' συμβολίζει το μέγεθος της i επιλεγείσας ομάδας. Αυτό αντιστοιχεί στο πρώτο στάδιο της δειγματοληψίας. Στην περίπτωση που επακολουθεί και δεύτερο στάδιο,

το σύνολο $\{U_{i1}, U_{i2}, ..., U_{in_i}\}$, i = 1, 2, …, m θα συμβολίζει ένα απλό τυχαίο δείγμα n_i μονάδων από την ομάδα U_i που επελέγη κατά το πρώτο στάδιο (i= 1, 2, …, m). Τότε,

$$\bar{n} = \sum_{i=1}^{m} n_i / m = \text{ μέσο μέγεθος των m υποδειγμάτων} \qquad (6.1.2)$$

$$\bar{U}_i = \sum_{j=1}^{n_i} U_{ij} / n_i = \text{μέσος του i υποδείγματος, i = 1, 2, …, m} \qquad (6.1.3)$$

$$S_i^2 = \sum_{j=1}^{n_i} (U_{ij} - \bar{U}_i)^2 / (n_i - 1)$$

$$= \frac{1}{(n_i - 1)} \times \text{ (άθροισμα τετραγωνικών αποκλίσεων των παρατηρήσεων του i δείγματος από τον μέσο του)}$$

$$= \text{διασπορά του i δείγματος, i = 1, 2, …, m} \qquad (6.1.4)$$

Περιοριζόμενοι στην περίπτωση διεξαγωγής της δειγματοληψίας σε ένα στάδιο μόνο ($N_i' \leftrightarrow n_i$), έχουμε

$$\mu = \sum_{i=1}^{M} \sum_{j=1}^{N_i} y_{ij} / N = \text{ μέση τιμή του αρχικού πληθυσμού (μέση τιμή του υπό εξέταση χαρακτηριστικού ανά στοιχειώδη μονάδα)} \qquad (6.1.5)$$

$$\mu_i = \sum_{j=1}^{N_i} y_{ij} / N_i = \text{ μέση τιμή της i ομάδας (του i υποπληθυσμού), i=1,2,…,M} \qquad (6.1.6)$$

$$\bar{\mu} = \sum_{i=1}^{M} \sum_{j=1}^{N_i} y_{ij} / M$$

$$= \sum_{i=1}^{M} N_i \, \mu_i / M \; = \text{μέση τιμή του χαρακτηριστικού ανά σύνθετη μονάδα (ανά ομάδα)} \qquad (6.1.7)$$

Προφανώς,

$$\mu = \sum_{i=1}^{M} N_i \, \mu_i \; / \sum_{i=1}^{M} N_i = \bar{\mu} / \bar{N} \qquad (6.1.8)$$

Για την εκτίμηση της μέσης τιμής μ του αρχικού πληθυσμού θα περιορισθούμε στην περίπτωση ενός σταδίου ($n_i=N_i$ και $\bar{U}_i=\mu_i$, i=1, 2, …, M) και θα διακρίνουμε τις εξής υποπεριπτώσεις:

Περίπτωση (i): $N_i=N_j$, $i \neq j$ (**Ισομεγέθεις ομάδες,** δηλαδή $N_i=\bar{N}$, i=1, 2, …, M).

Στην περίπτωση αυτή, τα μεγέθη N_i' των ομάδων που θα επιλεγούν θα είναι ίσα με $\bar{N}$, δηλαδή, $N_i' = \bar{N}$, i=1, 2, …, m.

Θεώρημα 6.1.1: Η στατιστική συνάρτηση

$$\hat{\mu}_c = \sum_{i=1}^{m} \sum_{j=1}^{\bar{N}} y_{ij} / (m\bar{N})$$

ή, ισοδύναμα, η στατιστική συνάρτηση

$$\hat{\mu}_c = \sum_{i=1}^{m} \bar{U}_i / m = \sum_{i=1}^{m} \mu_i / m \qquad (6.1.9)$$

είναι η αμερόληπτη εκτιμήτρια της μέσης τιμής μ με διασπορά

$$V(\hat{\mu}_c) \cong \frac{M-m}{M} \frac{1}{m} \frac{1}{\bar{N}^2} \sum_{i=1}^{M} \left(\sum_{j=1}^{\bar{N}} y_{ij} - \bar{\mu} \right)^2 \Big/ (M-1)$$

$$= \frac{M-m}{M} \frac{1}{m} \sum_{i=1}^{M} (\mu_i - \mu)^2 / (M-1) \qquad (6.1.10)$$

(Η δεύτερη ισότητα προκύπτει από το γεγονός ότι $\bar{\mu} = \bar{N}\mu$ και

$$\sum_{j=1}^{N_i} y_{ij} = \sum_{j=1}^{\bar{N}} y_{ij} = N_i \mu_i = \bar{N} \mu_i).$$

Πρόταση 6.1.1: Η στατιστική συνάρτηση

$$S^2_{\hat{\mu}_c} = \frac{M-m}{M} \frac{1}{m} \sum_{i=1}^{m} (\bar{U}_i - \hat{\mu}_c)^2 / (m-1) \qquad (6.1.11)$$

αποτελεί μια αμερόληπτη εκτιμήτρια της $V(\hat{\mu}_c)$.

Πρόταση 6.1.2: Η στατιστική συνάρτηση

$$\hat{Y} = M\bar{N}\hat{\mu}_c \qquad (6.1.12)$$

είναι αμερόληπτη εκτιμήτρια του συνολικού μεγέθους $y = M\bar{\mu} = M\bar{N}\mu$. Προφανώς,

$$V(\hat{Y}) = (M\bar{N})^2 V(\hat{\mu}_c) \qquad (6.1.13)$$

Περίπτωση (ii): $N_i \neq N_j,\ i \neq j$ (**Ανισομεγέθεις ομάδες)**.
Τότε, ισχύει το εξής θεώρημα:

Θεώρημα 6.1.2: Η στατιστική συνάρτηση

$$\hat{\mu}_c = \sum_{i=1}^{m}\sum_{j=1}^{N_i'} U_{ij} \Big/ \sum_{i=1}^{m} N_i' \tag{6.1.14}$$

είναι μια αμερόληπτη εκτιμήτρια της παραμέτρου μ με διασπορά

$$V(\hat{\mu}_c) \cong \frac{M-m}{M}\ \frac{1}{m}\ \frac{1}{\bar{N}^2}\ \frac{1}{M-1}\sum_{i=1}^{M}\left(\sum_{j=1}^{N_i} y_{ij} - N_i\mu\right)^2 .$$

Πόρισμα: Η στατιστική συνάρτηση

$$S^2_{\hat{\mu}_c} = \frac{M-m}{M}\ \frac{1}{m\bar{N}^2}\ \frac{1}{m-1}\sum_{i=1}^{m}\left(\sum_{j=1}^{N_i'} U_{ij} - N_i'\hat{\mu}_c\right)^2 \tag{6.1.15}$$

είναι αμερόληπτη εκτιμήτρια του $V(\hat{\mu}_c)$.

Παρατήρηση: Για την εκτίμηση του συνολικού μεγέθους y ισχύουν και πάλι οι σχέσεις (6.1.12) και (6.1.13).

Παράδειγμα: Το παράδειγμα της ενότητας 6.1 αποτελεί εφαρμογή των σχέσεων (6.1.9) και (6.1.10) με M=3 και m=1.

6.2 Περίπτωση Ποσοστών

Έστω πληθυσμός μεγέθους N, ο οποίος διαιρείται σε M ομάδες (clusters) μεγέθους $N_1, N_2, \ldots, N_M$. Έστω p το ποσοστό των μονάδων του

πληθυσμού, οι οποίες ανήκουν σε μια κατηγορία Α. Το πρόβλημα που θα αντιμετωπισθεί είναι η εκτίμηση του p με βάση ένα δείγμα που θα επιλεγεί με απλή δειγματοληψία κατά ομάδες.

Έστω $N_A^{(i)}$ ο αριθμός των μονάδων της ομάδας i που ανήκουν στην κατηγορία Α και $p_i = N_A^{(i)}/N_i$ το ποσοστό των μονάδων της i ομάδας που ανήκουν στην κατηγορία Α, i=1, 2, …, Μ. Προφανώς, ισχύει ότι

$$p = \sum_{i=1}^{M} N_i p_i \Big/ \sum_{i=1}^{M} N_i \,. \qquad (6.2.1)$$

Έστω ότι, από το σύνολο των Μ ομάδων, επιλέγεται ένα απλό τυχαίο δείγμα m ομάδων $U_1, U_2, \ldots, U_m$ (πρώτο στάδιο) και έστω ότι, από την i επιλεγείσα ομάδα, επιλέγεται ένα απλό τυχαίο δείγμα στοιχειωδών μονάδων μεγέθους n_i (δεύτερο στάδιο). Αν $X^{(i)}$ συμβολίζει τον αριθμό των μονάδων του i δείγματος που ανήκουν στην κατηγορία Α, τότε μια αμερόληπτη εκτιμήτρια του p_i είναι η

$$\hat{p}_i = X^{(i)}/n_i \,, \qquad i=1, 2, \ldots, m \qquad (6.2.2)$$

Περιοριζόμενοι στην περίπτωση δειγματοληψίας σε ένα στάδιο μόνο, μπορούμε να εφαρμόσουμε την θεωρία της προηγούμενης ενότητας για την εκτίμηση του p, αν θεωρήσουμε τις αντιστοιχίες $\mu_i \leftrightarrow p_i$, $\mu \leftrightarrow p$. Συγκεκριμένα, οδηγούμεθα στα εξής συμπεράσματα.

(i) Περίπτωση ισομεγεθών ομάδων ($N_i = \overline{N}$, $\hat{p}_i = p_i$, i=1, 2, …, M)

Στην περίπτωση αυτή,

$$N_i' = \overline{N}\,, \; i=1, 2, \ldots, m,$$

οπότε,

$$\hat{p}_i = X^{(i)} \big/ \bar{N}\,, \qquad i=1, 2, \ldots, m \tag{6.2.3}$$

και ισχύει το εξής θεώρημα.

Θεώρημα 6.2.1: Η στατιστική συνάρτηση

$$\hat{p}_c = \sum_{i=1}^{m} \hat{p}_i / m \tag{6.2.4}$$

είναι αμερόληπτη εκτιμήτρια της παραμέτρου p με διασπορά

$$V(\hat{p}_c) = \frac{M-m}{M} \; \frac{1}{m} \; \frac{1}{M-1} \sum_{i=1}^{M} (p_i - p)^2 \,. \tag{6.2.5}$$

Πρόταση 6.2.1: Μια αμερόληπτη εκτιμήτρια της $V(\hat{p}_c)$ είναι η στατιστική συνάρτηση

$$S^2_{\hat{p}_c} = \frac{M-m}{M} \; \frac{1}{m} \; \frac{1}{m-1} \sum_{i=1}^{m} (\hat{p}_i - \hat{p}_c)^2 \,. \tag{6.2.6}$$

(ii) Περίπτωση ανισομεγεθών ομάδων ($N_i \neq N_j$, $i \neq j$).

Θεώρημα 6.2.2: Η στατιστική συνάρτηση

$$\hat{p}_c = \sum_{i=1}^{m} X^{(i)} \Bigg/ \sum_{i=1}^{m} N_i' \tag{6.2.7}$$

είναι μια αμερόληπτη εκτιμήτρια της παραμέτρου p με διασπορά

$$V(\hat{p}_c) = \frac{M-m}{M} \; \frac{1}{m\bar{N}^2} \; \frac{1}{M-1} \sum_{i=1}^{M} (N_A^{(i)} - N_i p)^2 \qquad (6.2.8)$$

Απόδειξη: Η αμεροληψία είναι προφανής. Για την απόδειξη της (6.2.8), αρκεί να παρατηρηθεί ότι, επειδή τα N_i είναι εν γένει άγνωστα, η (6.2.7) είναι ουσιαστικά μια εκτιμήτρια λόγου δύο μεγεθών. Κατά συνέπεια, μπορεί να εφαρμοσθεί η θεωρία της ενότητας 2.6, αν θεωρηθεί η αντιστοιχία $y_i \leftrightarrow N_A^{(i)}$, $R \leftrightarrow p$, $\mu_x \leftrightarrow \bar{N}$ και $x_i \leftrightarrow N_i$. Επομένως, η (6.2.8) είναι άμεση συνέπεια της (2.6.1).

Πρόταση 6.2.2: Η στατιστική συνάρτηση

$$S^2_{\hat{p}_c} = \frac{M-m}{M} \; \frac{1}{m\bar{N}^2} \; \frac{1}{m-1} \sum_{i=1}^{m} (X^{(i)} - N_i' \hat{p}_c)^2$$

ή η υπολογιστικά περισσότερο προσφερόμενη μορφή της

$$S^2_{\hat{p}_c} = \frac{M-m}{M} \; \frac{1}{m\bar{N}^2} \; \frac{1}{m-1} \left(\sum_{i=1}^{m} X^{(i)2} - 2\hat{p}_c \sum_{i=1}^{m} X^{(i)} N_i' + \hat{p}_c^2 \sum_{i=1}^{m} N_i'^2 \right) \qquad (6.2.9)$$

είναι μια αμερόληπτη εκτιμήτρια της $V(\hat{p}_c)$.

Παράδειγμα: Έστω ότι οι 660 φοιτητές του τμήματος Στατιστικής ενός πανεπιστημίου μπορούν να διαιρεθούν σε 110 τάξεις των 6 φοιτητών έτσι, ώστε κάθε φοιτητής να ανήκει σε μια μόνο τάξη. Για να εξετασθεί πώς αντιμετωπίζεται μια μελετώμενη μεταβολή στο πρόγραμμα σπουδών, επιλέγονται τυχαία 11 τάξεις και όλοι οι φοιτητές των τάξεων

αυτών περιλαμβάνονται στο δείγμα. Έστω ότι οι αριθμοί υπέρ της μεταβολής είναι

Τάξη (i)	1	2	3	4	5	6	7	8	9	10	11	Σύνολο
Αριθμός φοιτητών υπέρ ($X^{(i)}$)	3	5	2	3	4	1	4	2	6	1	2	33

Να εκτιμηθεί το πραγματικό ποσοστό p των φοιτητών του τμήματος που είναι υπέρ της μεταβολής.

Λύση: Προφανώς, m=11, M=110, N_i =6, i=1,2,…,110. Τότε, από την (6.2.4) έχουμε

$$\hat{p}_c = \sum_{i=1}^{11} \hat{p}_i / 11 = \sum_{i=1}^{11} X^{(i)} \Big/ (6 \cdot 11) = 0.5 .$$

Επίσης, από την (6.2.6) προκύπτει ότι

$$S^2_{\hat{p}_c} = \frac{110-11}{110} \; \frac{1}{11} \; \frac{1}{10} \sum_{i=1}^{11} (\hat{p}_i - 0.5)^2 = 0.00591 .$$

Άρα,

$$S_{\hat{p}_c} = 0.0769 .$$

Σημείωση: Ένα απλό τυχαίο δείγμα 66 φοιτητών, που ενδεχομένως θα έδινε την ίδια εκτίμηση για το p, θα οδηγούσε σε διασπορά της εκτίμησης ίση με

$$S^2_{\hat{p}} = \frac{\hat{p}(1-\hat{p})}{n-1} \; \frac{N-n}{N} = \frac{(0.5)(0.5)}{65} \; \frac{660-66}{660} = 0.00346$$

και, επομένως, σε τυπικό σφάλμα ίσο με

$$S_{\hat{p}} = 0.0588 .$$

Παρατήρηση: Η 1-ανά-k συστηματική δειγματοληψία είναι μια μορφή δειγματοληψίας κατά ομάδες σε ένα στάδιο. Οι ομάδες είναι τα k δυνατά συστηματικά δείγματα (M=k) και το απλό τυχαίο δείγμα ομάδων που επιλέγεται είναι μεγέθους 1 (m=1).

6.3 Επίδραση του Δειγματοληπτικού Σχήματος (The Design Effect (deff))

Μια χρήσιμη ποσότητα που εισήχθη το 1965 από τον L.Kish, σε σχέση με σύνθετα δειγματοληπτικά σχήματα, είναι η λεγόμενη **επίδραση του δειγματοληπτικού σχήματος (design effect (deff))**. Η ποσότητα αυτή ορίζεται ως ο λόγος της διασποράς της εκτίμησης που προκύπτει από ένα δείγμα, το οποίο ελήφθη σύμφωνα με την χρησιμοποιηθείσα σύνθετη δειγματοληπτική μέθοδο, προς την διασπορά της εκτίμησης που προκύπτει από ένα απλό τυχαίο δείγμα του ίδιου μεγέθους. Ορίζεται από την σχέση

$$\text{deff} = \frac{V_A}{V_B},$$

όπου V_A και V_R συμβολίζουν την διασπορά της εκτίμησης από το δείγμα (του σύνθετου σχήματος) και από το απλό τυχαίο δείγμαν αντίστοιχα. Το μέτρο αυτό χρησιμοποιείται κυρίως στις εξής δύο περιπτώσεις: Στον προσδιορισμό του απαιτούμενου δειγματικού μεγέθους και στον υπολογισμό της αποτελεσματικότητας των σύνθετων δειγματοληπτικών σχημάτων. Για παράδειγμα, για την εκτίμηση του ποσοστού των ατόμων ενός πληθυσμού, τα οποία έχουν κάποιο χαρακτηριστικό, είναι συχνά προτιμότερο να χρησιμοποιείται μια συνθετότερη δειγματοληπτική μονάδα από το άτομο. Ας θεωρήσουμε, για παράδειγμα, την περίπτωση των φοιτητών του παραδείγματος της ενότητας 6.2. Ως δειγματοληπτική μονάδα στο παράδειγμα αυτό, χρησιμοποιήθηκε η τάξη έξι φοιτητών. Για την εκτίμηση του ποσοστού των φοιτητών που ήταν υπέρ της

μελετώμενης μεταβολής στο πρόγραμμα σπουδών του τμήματός τους, ένα απλό τυχαίο δείγμα έντεκα τάξεων των έξι φοιτητών οδήγησε σε εκτίμηση, της οποίας η διασπορά εκτιμήθηκε ίση με

$$\hat{V}_A = S^2_{\hat{p}} = 0.00591.$$

Υπολογίσθηκε επίσης ότι ένα απλό τυχαίο δείγμα 66 φοιτητών, που θα οδηγούσε στην ίδια εκτίμηση του p, θα έδινε διασπορά ίση με

$$\hat{V}_R = S^2_{\hat{p}} = 0.00346.$$

Μια εκτίμηση, επομένως, του μέτρου deff παρέχεται από την τιμή

$$\hat{\text{deff}} = \frac{S^2_{\hat{p}_c}}{S^2_{\hat{p}}} = \frac{0.00591}{0.00346} = 1.708.$$

Όταν η τιμή του λόγου f=n/N είναι μικρή, μπορούμε, επομένως, να εκτιμήσουμε το μέγεθος του απαιτούμενου δείγματος υπολογίζοντας την τιμή του n (του αριθμού των ατόμων) που απαιτούνται με ένα απλό τυχαίο δείγμα ατόμων και πολλαπλασιάζοντας στην συνέχεια με 1.708. Γενικότερα, υπολογίζοντας την τιμή της παραμέτρου deff για τις εκτιμήσεις των σημαντικών παραμέτρων με βάση ένα σύνθετο δειγματοληπτικό σχήμα, μπορούμε να χρησιμοποιήσουμε τους απλούς τύπους υπολογισμού του απαιτούμενου μεγέθους του δείγματος στην περίπτωση απλής τυχαίας δειγματοληψίας για να υπολογίσουμε το απαιτούμενο μέγεθος του δείγματος του συνθετότερου σχήματος, πολλαπλασιάζοντας με την τιμή του deff. Ταυτόχρονα, έχουμε την ευκαιρία να κρίνουμε κατά πόσο το σύνθετο δειγματοληπτικό σχήμα πλεονεκτεί ως προς την αποτελεσματικότητα σε σχέση με το απλό τυχαίο δείγμα. Παραδείγματα τέτοιων υπολογισμών δίνονται για την περίπτωση της στρωματοποιημένης τυχαίας δειγματοληψίας, της συστηματικής δειγματοληψίας και της δειγματοληψίας κατά ομάδες στα παραδείγματα των ενοτήτων 3.2, 5.2 και 6, αντίστοιχα.

ΑΣΚΗΣΕΙΣ

1. Μια ομάδα 61 λεπρών υποβλήθηκε σε μια θεραπεία 48 εβδομάδων. Για τον έλεγχο της αποτελεσματικότητας της θεραπείας, η παρουσία βακίλων ελέγχθηκε βακτηριολογικά σε 6 σημεία του σώματος κάθε ασθενούς. Οι αριθμοί των "αρνητικών" σημείων περιέχονται στον παρακάτω πίνακα:

Αριθμός "αρνητικών" σημείων (y)	0	1	2	3	4	5	6
αριθμός ασθενών με y αρνητικά σημεία	17	11	4	4	7	14	4

Να εκτιμηθεί το ποσοστό των αρνητικών σημείων στον πληθυσμό και να εκτιμηθεί το τυπικό σφάλμα της εκτίμησης θεωρώντας κάθε ασθενή ως "ομάδα" έξι σημείων.

2. Να συγκριθεί η ακρίβεια της εκτίμησης του p της προηγούμενης άσκησης με αυτή που θα είχαμε στην περίπτωση απλής τυχαίας δειγματοληψίας.

3. Ένα απλό τυχαίο δείγμα 30 νοικοκυριών επελέγη και τα μέλη των νοικοκυριών αυτών ταξινομήθηκαν σύμφωνα με το φύλο τους. Ο πίνακας που ακολουθεί δίνει την συχνότητα ν_{ij} των νοικοκυριών με i άνδρες και j γυναίκες.

	j					
i	0	1	2	3	4	Σύνολο
0	0	1	0	0	0	1
1	0	4	7	1	1	13
2	0	5	3	0	0	8
3	0	5	1	1	1	8
Σύνολο	0	15	11	2	2	30

Να εκτιμηθεί το ποσοστό p των ανδρών σε ολόκληρο τον πληθυσμό θεωρώντας κάθε νοικοκυριό ως μια ομάδα μελών.

4. Να συγκριθεί η διασπορά της εκτίμησης της προηγούμενης άσκησης με την διασπορά της εκτίμησης του ποσοστού p, αν αυτή είχε βασισθεί σε ένα απλό τυχαίο δείγμα του ίδιου μεγέθους.

5. Από τον πληθυσμό των 20 οικοδομικών τετραγώνων μιας μικρής πόλης επιλέγεται ένα απλό τυχαίο δείγμα 5 τετραγώνων και οι εργαζόμενοι όλων των καταστημάτων λιανικής πώλησης περιλαμβάνονται στο δείγμα. Να εκτιμηθεί ο συνολικός αριθμός των εργαζομένων σε καταστήματα λιανικής πώλησης της πόλης αυτής, αν τα αποτελέσματα της δειγματοληψίας είναι ως εξής:

οικοδομικό τετράγωνο	1	2	3	4	5
αριθμός εργαζομένων	4	6	2	6	1

Να εκτιμηθεί το τυπικό σφάλμα της εκτίμησης.