5. ΣΥΣΤΗΜΑΤΙΚΗ ΔΕΙΓΜΑΤΟΛΗΨΙΑ (Systematic Sampling)

Συχνά, είναι ταχύτερη και ευκολότερη η επιλογή των μονάδων του πληθυσμού, αν αυτή γίνεται από κάποιο κατάλογο ξεκινώντας από κάποιο τυχαίο αρχικό σημείο και επιλέγοντας μια μονάδα κάθε k (k>0) μονάδες μέχρι να κατασκευασθεί το δείγμα με το δοθέν μέγεθος. Για παράδειγμα, αν πρόκειται να επιλεγούν 1000 καρτέλες από έναν φοριαμό που περιέχει 10000 καρτέλες, είναι ταχύτερο να επιλεγεί ένας τυχαίος αριθμός μεταξύ 1 και 10 και να περιληφθεί στο δείγμα η καρτέλα που αντιστοιχεί σ' αυτόν τον αριθμό καθώς και κάθε δέκατη καρτέλα από εκεί και πέρα, από το να επιλεγούν 1000 τυχαίοι αριθμοί και να περιληφθούν οι καρτέλες που αντιστοιχούν σε αυτούς. Η δειγματοληπτική αυτή τεχνική, η οποία εισάγει ένα συστηματικό στοιχείο στην διαδικασία επιλογής των μονάδων του πληθυσμού, είναι μια μορφή δειγματοληπτικής τεχνικής που είναι γνωστή ως **συστηματική δειγματοληψία (systematic sampling)**.

5.1 Περιγραφή της Διαδικασίας Λήψης ενός Συστηματικού Δείγματος

Έστω ότι οι μονάδες ενός πληθυσμού μεγέθους N είναι αριθμημένες από το 1 μέχρι το N. Έστω k ένας θετικός ακέραιος. Για την επιλογή ενός 1-ανά-k συστηματικού δείγματος μεγέθους n, διαλέγουμε τυχαία μια μονάδα από τις k πρώτες μονάδες και περιλαμβάνουμε στο δείγμα αυτήν και κάθε μονάδα του πληθυσμού που απέχει από αυτήν κατά κάποιο πολλαπλάσιο του k. Η επιλογή της πρώτης μονάδας καθορίζει ολόκληρο το δείγμα. Για παράδειγμα, αν ο πληθυσμός

αποτελείται από τις τιμές $y_1, y_2, \ldots, y_N$ και από τις πρώτες k μονάδες του επιλεγεί η k_0, τότε το δείγμα θα αποτελείται από τις μονάδες

$$y_{k_0},\ y_{k_0+k},\ y_{k_0+2k},\ \ldots,\ y_{k_0+(n-1)k}.$$

Ορισμός: Ένα δείγμα μεγέθους n ονομάζεται 1-ανά-k συστηματικό δείγμα, αν περιλαμβάνει κάθε k μονάδα του πληθυσμού με αρχικό σημείο επιλεγόμενο τυχαία από τις πρώτες k μονάδες του πληθυσμού. Αν οι μονάδες του πληθυσμού εμφανίζονται με τυχαία σειρά στην λίστα από την οποία επιλέγουμε το δείγμα, τότε αυτό ονομάζεται **ψευδοτυχαίο δείγμα (pseudorandom sample)**.

Παραλλαγές της συστηματικής δειγματοληψίας μπορούν να ορισθούν με βάση διαφορετικές διαδικασίες επιλογής του αρχικού σημείου. Για παράδειγμα, είναι δυνατόν το αρχικό σημείο να είναι η μονάδα του πληθυσμού με δείκτη

$$\begin{cases} (k+1)/2\ , & \text{αν ο } k \text{ είναι περιττός} \\ k/2 \text{ ή } (k+2)/2, & \text{διαφορετικά}. \end{cases}$$

Δηλαδή, αντί να ξεκινήσουμε από μια τυχαία μονάδα, ξεκινούμε από το κεντρικό σημείο (ή από ένα σημείο κοντά στο κέντρο) του στρώματος των πρώτων k μονάδων του πληθυσμού.

Επειδή το μέγεθος N του πληθυσμού δεν είναι εν γένει πολλαπλάσιο του k, διαφορετικά συστηματικά δείγματα από τον ίδιο πληθυσμό ενδέχεται να έχουν διαφορετικό μέγεθος. Αν, για παράδειγμα, N=17, k=5 τα δυνατά 1-ανά-5 συστηματικά δείγματα είναι τα εξής πέντε:

1	2	3	4	5
y_1	y_2	y_3	y_4	y_5
y_6	y_7	y_8	y_9	y_{10}
y_{11}	y_{12}	y_{13}	y_{14}	y_{15}
y_{16}	y_{17}			

Για την αποφυγή αυτού του προβλήματος και την επίτευξη σταθερού δειγματικού μεγέθους, χρησιμοποιείται η εξής μέθοδος:

Τα N στοιχεία του πληθυσμού θεωρούνται τοποθετημένα στην περιφέρεια ενός κύκλου. Έστω k ο πλησιέστερος προς τον λόγο N/n ακέραιος. Ένα στοιχείο από τα N επιλέγεται τυχαία ως αρχικό σημείο και περιλαμβάνεται στο δείγμα μαζί με όλα τα στοιχεία πάνω στην περιφέρεια του κύκλου, που απέχουν από το αρχικό κατά πολλαπλάσια του k, καθώς αυτή διατρέχεται κατά την φορά των δεικτών του ρολογιού. Αυτό συνεχίζεται μέχρι να επιτευχθεί το επιθυμητό δειγματικό μέγεθος. Ουσιαστικά, η διαδικασία αυτή ισοδυναμεί με την διαδικασία που επιλέγει στο δείγμα την μονάδα y_i –N, αν i>N. Επιπλέον, είναι εύκολο να διαπιστωθεί ότι η μέθοδος αυτή εξασφαλίζει την ίδια πιθανότητα επιλογής σε κάθε μονάδα του πληθυσμού και συνεπώς οδηγεί σε αμερόληπτες εκτιμήσεις της μέσης τιμής του πληθυσμού.

Επομένως, αν N=10 και n=4, ακολουθώντας την τεχνική αυτή τα δυνατά 1-ανά-3 συστηματικά δείγματα μεγέθους 4 (k=N/n=10/4=2.5 ≈ 3) είναι τα εξής:

1	2	3	4	5	6	7	8	9	10
y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}
y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_1	y_2	y_3
y_7	y_8	y_9	y_{10}	y_1	y_2	y_3	y_4	y_5	y_6
y_{10}	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9

Όπως αναφέρθηκε προηγουμένως, ένα από τα προφανή πλεονεκτήματα της μεθόδου αυτής είναι η ταχύτητα. Επιπλέον, διαισθητικά φαίνεται να είναι ακριβέστερη από την απλή τυχαία δειγματοληψία. Στην πραγματικότητα, στρωματοποιεί τον πληθυσμό σε n στρώματα που αποτελούνται από τις πρώτες k μονάδες, τις δεύτερες k μονάδες κ.ο.κ. Αναμένεται, επομένως, ότι ένα συστηματικό τυχαίο δείγμα θα έχει την ίδια περίπου ακρίβεια με ένα στρωματοποιημένο τυχαίο δείγμα με n_i =1, i=1,2,…,n. Η διαφορά βρίσκεται στο ότι, στο συστηματικό δείγμα, οι μονάδες έχουν την ίδια σχετική θέση στο

στρώμα, ενώ, στο στρωματοποιημένο τυχαίο δείγμα, η θέση των μονάδων στο στρώμα καθορίζεται τυχαία. Είναι, λοιπόν, το συστηματικό δείγμα πιο ομοιόμορφα κατανεμημένο στον πληθυσμό και αυτό συμβάλλει στο να παρέχει πολύ συχνά ακριβέστερες εκτιμήσεις από ένα στρωματοποιημένο τυχαίο δείγμα.

Υπάρχουν όμως και κίνδυνοι στην χρησιμοποίηση συστηματικής δειγματοληψίας. Οι πιο σημαντικοί αναφέρονται στην περίπτωση περιοδικότητας στις τιμές των μονάδων του πληθυσμού, όσον αφορά την σειρά εμφάνισής τους στην λίστα, αν το k είναι ίσο με την περίοδο ή ένα πολλαπλάσιό της. Για παράδειγμα, μια συστηματική επιλογή μονάδων από μια ″περιοδική″ λίστα οικοδομικών τετραγώνων μιας πόλης μπορεί να οδηγήσει σ' ένα δείγμα που περιέχει τετράγωνα που ανήκουν σε μια γραμμή και επομένως σε αύξηση του σφάλματος των εκτιμητριών.

Στην πράξη, είναι ασφαλής η χρησιμοποίηση της συστηματικής δειγματοληψίας, αν δεν υπάρχουν ενδείξεις περιοδικότητας. Βέβαια, αυτό δεν είναι εύκολο να ελεγχθεί. Ο κίνδυνος, όμως, εσφαλμένης χρησιμοποίησης της μεθόδου μπορεί να ελαττωθεί, αν το δείγμα είναι αποτέλεσμα ενός αριθμού συστηματικών επιλογών από διαφορετικά στρώματα.

5.2 Η Διασπορά της Εκτιμήτριας του Μέσου

Είναι προφανές ότι, αν N=nk, τα k δυνατά 1-ανά-k συστηματικά δείγματα είναι οι στήλες του πίνακα 5.2.1.

Από τον πίνακα αυτό, έυκολα μπορεί να δει κανείς ότι με την συστηματική δειγματοληψία, ο πληθυσμός χωρίζεται σε k σύνθετες μονάδες, και η διαδικασία επιλογής ενός συστηματικού δείγματος ισοδυναμεί με την διαδικασία τυχαίας επιλογής μιας σύνθετης μονάδας. Επομένως, ένα 1-ανά-k συστηματικό τυχαίο δείγμα n μονάδων από ένα πληθυσμό μεγέθους N=nk, είναι ένα απλό τυχαίο δείγμα μιας σύνθετης μονάδας από τον πληθυσμό με (σύνθετες) μονάδες τις k στήλες του πίνακα 5.2.1.

Πίνακας 5.2.1

Σύνθεση των k δυνατών συστηματικών δειγμάτων μεγέθους n από ένα πληθυσμό μεγέθους N=nk

Δείγμα					
1	2	. . .	i	. . .	k
Y_1	y_2	. . .	y_i	. . .	y_k
Y_{k+1}	y_{k+2}	. . .	y_{k+i}	. . .	y_{2k}
.	.	. . .	.	. . .	.
.	.	. . .	.	. . .	.
.	.	. . .	.	. . .	.
$Y_{(n-1)k+1}$	$Y_{(n-1)k+2}$	. . .	$y_{(n-1)k+i}$	. . .	y_{nk}

Η διαίρεση του αρχικού πληθυσμού των N=nk μονάδων σε k ομάδες έχει ως αποτέλεσμα την δυνατότητα έκφρασης της διασποράς σ^2 του πληθυσμού μέσω της διασποράς "μεταξύ" των k ομάδων και της διασποράς "μέσα" στις ομάδες. Αυτό γίνεται ευκολότερα αντιληπτό, αν το j στοιχείο του i συστηματικού δείγματος συμβολισθεί με y_{ij}. (Συγκεκριμένα, $y_{ij}=y_{(j-1)k+i}$). Τότε, αν μ είναι η μέση τιμή του πληθυσμού και $\bar{X}_n^{(i)}$ ο μέσος του i δείγματος, ισχύει ότι

$$(N-1)\sigma^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\mu)^2 = (\text{προσθαφαιρώντας το } \bar{X}_n^{(i)}) =$$

$$= n\sum_{i=1}^{k}(\bar{X}_n^{(i)}-\mu)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{X}_n^{(i)})^2 . \qquad (5.2.1)$$

Αλλά, επειδή, κάθε ένα από τα k συστηματικά δείγματα συνεισφέρει $n-1$ βαθμούς ελευθερίας, ο δεύτερος προσθετέος είναι ίσος με $k(n-1)\sigma_w^2$, όπου

$$\sigma_w^2 = \frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{X}_n^{(i)})^2 \qquad (5.2.2)$$

είναι η διασπορά μεταξύ μονάδων του πληθυσμού που ανήκουν στο ίδιο συστηματικό δείγμα. Επίσης, αν $\bar{X}_n^*$ συμβολίζει τον μέσο ενός συστηματικού δείγματος, το άθροισμα στον πρώτο προσθετέο της (5.2.1) είναι ίσο με $k\sigma^2_{\bar{X}_n^*}$, αφού

$$\sigma^2_{\bar{X}_n^*} = \frac{1}{k} \sum_{i=1}^{k} [\bar{X}_n^{(i)} - E(\bar{X}_n^{(i)})]^2 = \frac{1}{k} \sum_{i=1}^{k} (\bar{X}_n^{(i)} - \mu)^2 . \qquad (5.2.3)$$

Είναι, δηλαδή, ο πρώτος προσθετέος ανάλογος της διασποράς μεταξύ των συστηματικών δειγμάτων. Τότε, μπορεί να αποδειχθεί το εξής θεώρημα.

Θεώρημα 5.2.1: Η διασπορά του μέσου $\bar{X}_n^*$ ενός 1-ανά-k συστηματικού δείγματος μεγέθους n από ένα πληθυσμό N=nk μονάδων δίνεται από τον τύπο

$$\sigma^2_{\bar{X}_n^*} = [(N-1)\sigma^2 - k(n-1)\sigma_W^2] \big/ N . \qquad (5.2.4)$$

Απόδειξη: Προφανώς η (5.2.1) ισοδυναμεί με την σχέση

$$(N-1)\sigma^2 = nk\sigma^2_{\bar{X}_n^*} + k(n-1)\sigma_W^2 ,$$

η οποία οδηγεί στην (5.2.4).

Πόρισμα: Ο μέσος $\bar{X}_n^*$ ενός συστηματικού δείγματος είναι ακριβέστερος από τον μέσο $\bar{X}_n$ ενός απλού τυχαίου δείγματος του ίδιου μεγέθους τότε και μόνο τότε αν $\sigma_W^2 > \sigma^2$.

Απόδειξη: Ισχύει ότι $V(\bar{X}_n) = \frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)$. Τότε, από την (5.2.4), ισχύει ότι

$$V(\bar{X}_n^*) < V(\bar{X}_n)$$
$$\Leftrightarrow \frac{N-1}{N}\sigma^2 - \frac{k(n-1)}{N}\sigma_W^2 < \frac{N-n}{N}\ \frac{\sigma^2}{n}$$
$$\Leftrightarrow k(n-1)\sigma_W^2 > \left(N - 1 - \frac{N-n}{N}\right)\sigma^2 \equiv k(n-1)\sigma^2, \quad \text{ο.ε.δ.}$$

Δηλαδή, η συστηματική δειγματοληψία οδηγεί σε μικρότερο τυπικό σφάλμα, αν η διασπορά μέσα στο δείγμα είναι μεγαλύτερη από την διασπορά ολόκληρου του πληθυσμού. Επομένως, μεγαλύτερη ακρίβεια επιτυγχάνεται, αν οι μονάδες του δείγματος έχουν μεγαλύτερη ετερογένεια σε σχέση με αυτήν που έχουν οι μονάδες όλου του πληθυσμού.

Παρατήρηση: Είναι προφανές ότι στην περίπτωση εκτίμησης του ποσοστού p των μονάδων του πληθυσμού που ανήκουν σε μια κατηγορία Α, εφαρμόζεται η παραπάνω θεωρία, αν η παράμετρος μ αντικατασταθεί από την παράμετρο p και ο μέσος $\bar{X}_n^{(i)}$ αντικατασταθεί από την στατιστική συνάρτηση $\hat{p}^{(i)}$, όπου $\hat{p}^{(i)}$ είναι το ποσοστό των μονάδων του i συστηματικού δείγματος που ανήκουν στην κατηγορία Α. Είναι προφανές ότι ο πληθυσμός θα αποτελείται από μονάδες της μορφής

$$y_{ij} = \begin{cases} 1, & \text{αν η j μονάδα του i δείγματος ανήκει στην A} \\ 0, & \text{διαφορετικά.} \end{cases}$$

Παράδειγμα: Ένας πληθυσμός 360 νοικοκυριών μιας συνοικίας με μικτό πληθυσμό (αριθμημένων από το 1 έως το 360) έχει καταγραφεί σε έναν κατάλογο κατά αλφαβητική σειρά ως προς το επίθετο του αρχηγού του νοικοκυριού. Τα νοικοκυριά των οποίων ο αρχηγός είναι μη λευκός, εμφανίζονται με τους εξής αύξοντες αριθμούς.
28, 31-33, 36-41, 44, 45, 47, 55, 56, 58, 58, 68, 69, 82, 83, 85, 86, 89-94, 98, 99, 101, 107-110, 114, 154, 156, 178, 223, 224, 296, 298-300, 302-304, 306-323, 325-331, 333, 335-339, 341, 342.
Να συγκριθεί η ακρίβεια ενός συστηματικού δείγματος, που περιλαμβάνει ένα άτομο ανά 8 άτομα του πληθυσμού, με την ακρίβεια ενός απλού τυχαίου δείγματος του ίδιου μεγέθους, αν υποτεθεί ότι θέλουμε να εκτιμήσουμε το ποσοστό των νοικοκυριών με μη λευκούς αρχηγούς.

Λύση: N=360, k=8. Άρα, n=N/k=360/8=45. Επομένως, υπάρχουν 8 δυνατά 1-ανά-8 συστηματικά δείγματα μεγέθους 45. Έστω

$$y_{ij} = \begin{cases} 1, & \text{αν το j νοικοκυριό του i δείγματος} \\ & \text{έχει μη λευκό αρχηγό} \\ 0, & \text{διαφορετικά.} \end{cases}$$

Με την βοήθεια του πίνακα 5.2.1, είναι εύκολο να δει κανείς ότι οι αριθμοί $\sum_{j=1}^{45} y_{ij}$, i=1, 2, …, 8 των νοικοκυριών με μη λευκό αρχηγό για τα διάφορα δυνατά δείγματα είναι

Δείγμα i	1	2	3	4	5	6	7	8	Σύνολο
$\sum_{j=1}^{45} y_{ij}$	7	13	10	10	12	9	10	10	81

Έστω $\hat{p}^*$ η εκτιμήτρια του p από ένα συστηματικό τυχαίο δείγμα. Τότε, από την (5.2.3), έχουμε

$$V(\hat{p}^*) = \frac{1}{k}\sum_{i=1}^{k}(\hat{p}^{(i)} - p)^2 ,$$

όπου

$$\hat{p}^{(i)} = \sum_{j=1}^{45} y_{ij}/45, \qquad i = 1, 2, \ldots, k$$

και

$$p = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}/N = \sum_{i=1}^{8}\sum_{j=1}^{45} y_{ij}/360 = \frac{81}{360} = 0.225 .$$

Δηλαδή, τελικά,

$$V(\hat{p}^*) = 0.001412.$$

Για την εκτιμήτρια $\hat{p}$ βασισμένη σε ένα απλό τυχαίο δείγμα μεγέθους 45, έχουμε

$$V(\hat{p}) = \frac{N-n}{n}\ \frac{p(1-p)}{N-1} = \frac{360-45}{45}\ \frac{(0.225)(0.775)}{359} = 0.00340 .$$

Είναι, δηλαδή, η τιμή της διασποράς της $\hat{p}^*$ ίση με το 41.53% της τιμής της διασποράς της $\hat{p}$.

Παρατήρηση: Η συστηματική δειγματοληψία χρησιμοποιείται συχνά λόγω της απλότητάς της, σε πληθυσμούς, στους οποίους η αρίθμηση των μονάδων είναι τυχαία. Αυτό συμβαίνει, για παράδειγμα, στις περιπτώσεις

δειγματοληψίας από ένα αρχείο ονομάτων, τα οποία εμφανίζονται κατά αλφαβητική σειρά, με την προϋπόθεση ότι το χαρακτηριστικό που μελετάται δεν σχετίζεται με το επίθετο των συγκεκριμένων ατόμων. Στην περίπτωση αυτή, δεν θα υπάρχει τάση ή στρωματοποίηση ως προς τις τιμές y_i των μονάδων του πληθυσμού καθώς θα διατρέχεται το αρχείο και δεν θα υπάρχει συσχέτιση μεταξύ γειτονικών τιμών. Επομένως, αναμένεται ότι η συστηματική δειγματοληψία θα είναι ισοδύναμη με την απλή τυχαία δειγματοληψία και θα οδηγεί στην ίδια διασπορά, οποτεδήποτε ο πληθυσμός έχει *τυχαία διάταξη*, με την έννοια της τυχαίας αρίθμησης των μονάδων που τον απαρτίζουν. Για οποιονδήποτε συγκεκριμένο πεπερασμένο πληθυσμό και για οποιεσδήποτε τιμές του n και k, αυτό δεν αληθεύει ακριβώς. Ο λόγος είναι ότι η διασπορά της εκτιμήτριας στην περίπτωση της συστηματικής δειγματοληψίας, η οποία βασίζεται μόνο σε k βαθμούς ελευθερίας, συμπεριφέρεται μάλλον ανορθόδοξα, όταν η τιμή του k είναι μικρή και, επομένως, ενδέχεται να υπερβαίνει ή να είναι μικρότερη από την διασπορά της εκτιμήτριας στην περίπτωση της απλής τυχαίας δειγματοληψίας. Μπορεί, όμως, να αποδειχθεί ότι, κατά μέσο όρο, οι δύο διασπορές ταυτίζονται. Συγκεκριμένα, αποδεικνύεται ότι

$$E(V_S)=V_R,$$

όπου V_S και V_R συμβολίζουν τις διασπορές της εκτιμήτριας στην περίπτωση συστηματικής και απλής τυχαίας δειγματοληψίας, αντίστοιχα.