4.ΣΤΡΩΜΑΤΟΠΟΙΗΜΕΝΗ ΤΥΧΑΙΑ ΔΕΙΓΜΑΤΟΛΗΨΙΑ (STRATIFIED RANDOM SAMPLING)

Στην τυχαία δειγματοληψία κατά στρώματα ο πληθυσμός των N μονάδων (πρόκειται για τον στατιστικό πληθυσμό και τις στατιστικές μονάδες) χωρίζεται σε υπο-πληθυσμούς N_1, N_2, ..., N_L , μονάδων αντίστοιχα. Αυτοί οι υπο-πληθυσμοί δεν επικαλύπτονται (ένα άτομο ή μονάδα ανήκει μόνο σε ένα υπο-πληθυσμό) και όλοι μαζί περιέχουν το σύνολο των μονάδων του πληθυσμού. Κάθε υπο-πληθυσμός καλείται "στρώση" (*stratum*). Αφού οι στρώσεις έχουν ορισθεί από κάθε μία παίρνουμε ένα τυχαίο δείγμα. Το μέγεθος αυτών των δειγμάτων είναι n_1, n_2, ..., n_L αντίστοιχα.

Η στρωματοποίηση είναι κοινή στρατηγική. Πρακτικοί αλλά και θεωρητικοί λόγοι οδηγούν σ'αυτή:

- Για κάποιο μέρος του πληθυσμού απαιτείται ιδιαίτερη ακρίβεια στις εκτιμήσεις.
- Ο πληθυσμός είναι ήδη στρωματοποιημένος (π.χ. ένα σύνολο λιμνών, το σύνολο των δήμων μιας πόλης).
- Ο πληθυσμός είναι ετερογενής αλλά στο εσωτερικό του περιέχει μέρη (ομάδες, συνιστώσες) τα οποία δείχνουν μια σχετική ομοιογένεια. Αυτές οι συνιστώσες θα αποτελέσουν τις στρώσεις.

Όσο πιο ομοιογενείς στο εσωτερικό τους είναι οι στρώσεις και όσο περισσότερο διαφέρουν μεταξύ τους, τόσο πιο αποδοτική είναι η στρωματοποίηση. Απόδοση εδώ σημαίνει ότι με το ίδιο κόστος μελέτης η εκτίμησή είναι πιο ακριβής. Σε γενικές γραμμές η τυχαία δειγματοληψία κατά στρώματα είναι πιο ακριβής από την τυχαία δειγματοληψία.

4.1. ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΟΥ ΤΟΥ ΠΛΗΘΥΣΜΟΥ

ΕΚΤΙΜΗΤΕΣ

Ο πληθυσμός είναι χωρισμένος σε L στρώσεις

Δεδομένα

στρώση 1: $y_{11}, y_{12}, y_{13}, ..., y_{1i}, ..., y_{1n}$

στρώση 2: $y_{21}, y_{22}, y_{23}, ..., y_{2i}, ..., y_{2n}$

στρώση h: $y_{h1}, y_{h2}, y_{h3}, ..., y_{hi}, ..., y_{hn}$

στρώση L: $y_{L1}, y_{L2}, y_{L3}, ..., y_{Li}, ..., y_{Ln}$

Ο δείκτης h υποδεικνύει τη στρώση και ο δείκτης i τη μονάδα μέσα σε κάθε στρώση.

Η υπέρμετρη αύξηση του αριθμού των στρώσεων δεν οδηγεί σε ακριβέστερες εκτιμήσεις. Ο αριθμός των 6 στρώσεων αναφέρεται από τον Cochran (1977) σαν ένα λογικό όριο.

Παράμετροι του δείγματος

Τά ακόλουθα σύμβολα αναφέρονται στη στρώση h. Ανάλογοι τύποι ισχύουν και για τις υπόλοιπες στρώσεις του πληθυσμού.

N: συνολικός αριθμός μονάδων του πληθυσμού, N_h: συνολικός αριθμός μονάδων της στρώσης, n_h: αριθμός μονάδων στο δείγμα της στρώσης (τα $n_1, n_2, ..., n_h, ..., n_L$ μπορούν να είναι διαφορετικά), y_{hi}: τιμή της μονάδας i στη συγκεκριμένη στρώση

$W_h = \frac{N_h}{N}$ βάρος της στρώσης $f_h = \frac{n_h}{N_h}$ δειγματοληπτικό κλάσμα της στρώσης

$\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h}$ μέση τιμή στρώσης $s_h^2 = \frac{\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2}{n_h - 1}$ διασπορά στρώσης

Εκτίμηση της μέσης τιμής ($\hat{\bar{Y}}$) του πληθυσμού

$$\hat{\bar{Y}} = \bar{y} = \frac{\sum_{h=1}^{L} N_h \cdot \bar{y}_h}{N} = \sum_{h=1}^{L} W_h \cdot \bar{y}_h$$

$$v(\bar{y}) = s_{\bar{y}}^2 = \frac{1}{N^2} \cdot \sum_{h=1}^{L} \frac{N_h^2 \cdot s_h^2}{n_h}(1 - f_h) = \sum_{h=1}^{L} \frac{W_h^2 \cdot s_h^2}{n_h}(1 - f_h)$$

το τυπικό σφάλμα είναι $s_{\bar{y}} = \sqrt{v(\bar{y})}$

$$P\{\bar{y} - t_{\alpha/2} s_{\bar{y}} < \bar{Y} < \bar{y} + t_{\alpha/2} s_{\bar{y}}\} = 1 - \alpha$$

Εκτίμηση του συνόλου ($\hat{Y}$) του πληθυσμού

$$\hat{Y} = N \cdot \bar{y} \qquad v(\hat{Y}) = s_{\hat{Y}}^{\,2} = N^2 \cdot v(\bar{y}) = N^2 \cdot s_{\bar{y}}^{\,2}$$

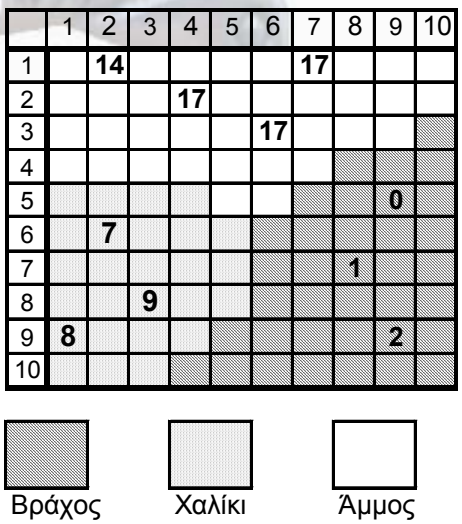
$$P\{\hat{Y} - t_{\alpha/2}\sqrt{v(\hat{Y})} < Y < \hat{Y} + t_{\alpha/2}\sqrt{v(\hat{Y})}\} = 1 - \alpha$$

το t ακολουθεί την κατανομή του Student με n_e βαθμούς ελευθερίας. Σύμφωνα με τον Satterthwaite, 1946 (αναφορά Cochran, 1977) το n_e είναι περίπου ίσο με

$$n_e = \frac{(\sum_{h=1}^{L} g_h \cdot s_h^{\,2})^2}{\sum_{h=1}^{L} \frac{g_h^{\,2} \cdot s_h^{\,4}}{n_h - 1}} \qquad \text{με} \qquad g_h = \frac{N_h \cdot (N_h - n_h)}{n_h}$$

ΠΑΡΑΔΕΙΓΜΑ 4.1

Μια προκαταρτική μελέτη έδειξε ότι η συγκέντρωση ατόμων συγκεκριμένου είδους ψαριού είναι άμεσα συνδεδεμένη με την φύση του βυθού (αριστερό διάγραμμα). Κατά τη διάρκεια δειγματοληψίας στην περιοχή που παρουσιάζει το δεξί διάγραμμα συλλέχθηκαν 10 δείγματα. Οι συγκεντρώσεις που ανεβρέθηκαν ανά μονάδα επιφάνειας καθώς και η φύση του βυθού στην περιοχή φαίνονται στο ίδιο διάγραμμα. Με βάση αυτή την πληροφορία εκτιμήστε την συνολική αφθονία του είδους στην περιοχή του δεξιού διαγράμματος (εκτίμηση του συνολικού αριθμού και υπολογισμός του διαστήματος εμπιστοσύνης της εκτίμησης).



Από το παραπάνω διάγραμμα φαίνεται καθαρά ότι το συγκεκριμμένο είδος προτιμά τους αμμώδεις βυθούς. Η πυκνότητα του σ'αυτές τις περιοχές είναι κατά πολύ μεγαλύτερη απ'ότι σε βραχώδεις ή σε περιοχές με χαλίκια. Από τη στιγμή που η πληροφορία αυτή είναι

γνωστή είναι λογικό να προσπαθήσουμε να την εκμεταλλευτούμε σχεδιάζοντας μια δειγματοληψία κατά στρώματα. Χρησιμοποιούμε τον τοπογραφικό χάρτη της περιοχής για να ορίσουμε 3 στρώσεις ανάλογα με τη φύση του βυθού. Περιμένουμε λοιπόν τα δείγματα που θα συλλεχθούν από κάθε στρώση να μοιάζουν μεταξύ τους ενώ τα δείγματα από διαφορετικές στρώσεις να διαφέρουν μεταξύ τους εμφανώς. Αφού καθορίσουμε τις τρείς στρώσεις, τις χωρίζουμε σε βασικές δειγματοληπτικές μονάδες και επιλέγουμε τυχαία κάποιες από αυτές. Στο συγκεκριμένο παράδειγμα επιλέγουμε 3 μονάδες από τη βραχώδη και την περιοχή με τα χαλίκια και 4 από την αμμώδη στρώση. Έτσι έχουμε:

Συνολικός αριθμός δειγματοληπτικών μονάδων N=	
Αριθμός στρώσεων L=	

	Στρώση 1 (βράχος)	Στρώση 2 (χαλίκι)	Στρώση 3 (άμμος)
Δεδομένα y_{hi}=			
Αριθμός μονάδων στο δείγμα κάθε στρώσης n_h=			
Αριθμός μονάδων ανά στρώση N_h=			
Βάρος των στρώσεων $W_h=N_h/N$=			
Δειγματοληπτικό κλάσμα στρώσης $f_h=n_h/N_{h=}$			

Μέση τιμή στρώσεων $\bar{y}_h$			
Διασπορά στρώσεων $s^2_{\bar{y}h}$			

g_h= $g_h\ s_h^2$= $(g_h^2\ s_h^4)/(n_h-1)$			
Βαθμοί ελευθερίας (n_e)=			
Τιμή $t_{(5\%)}$=			

Μέση τιμή στον πληθυσμό $\bar{y}$	

Διασπορά μέσης τιμής $s_{\bar{y}}^2$			
Τυπικό σφάλμα μέσης τιμής $s_{\bar{y}}$			
Διάστημα εμπιστοσύνης (95%)	Κατώτερο όριο		Ανώτερο όριο
		<$\bar{Y}$<	

Σύνολο πληθυσμού $\hat{Y}$			
Διασπορά συνόλου			
Τυπικό σφάλμα συνόλου			
Διάστημα εμπιστοσύνης (95%)	Κατώτερο όριο		Ανώτερο όριο
		<Y<	

ΠΑΡΑΔΕΙΓΜΑ 4.1

4.2. Η ΣΤΡΩΜΑΤΟΠΟΙΗΣΗ ΚΑΙ ΟΙ ΕΠΙΠΤΩΣΕΙΣ ΤΗΣ

ΠΑΡΑΔΕΙΓΜΑ 4.2

ΠΙΝΑΚΑΣ 4.1 *Πληθυσμός μικρών τρωκτικών αποτελούμενος από 9 άτομα που χαρακτηρί-ζονται από τις τιμές y που αντιπροσωπευουν το ολικό ύψος των ατόμων (cm).* Α: αρσενικά, Θ: θηλυκά. (πρόκειται για τα δεδομένα του παραδείγματος 1.1)

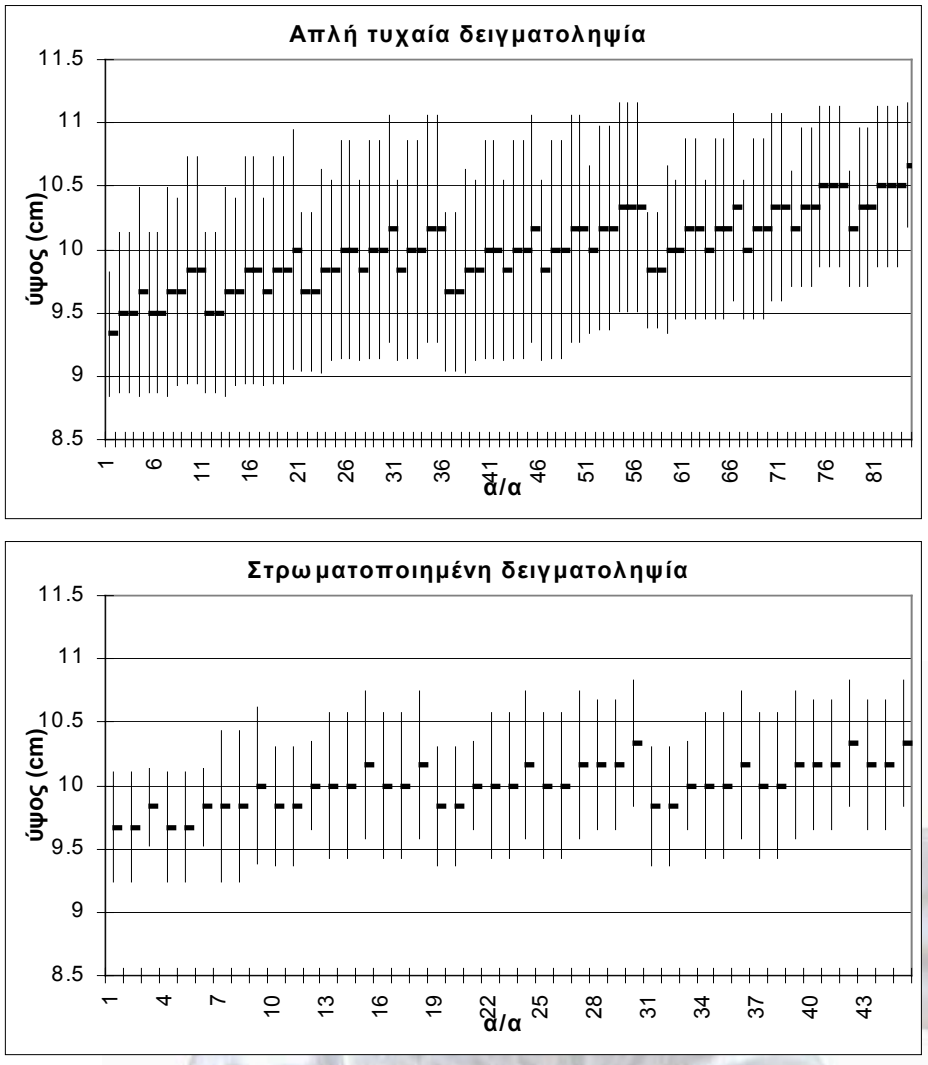
y_i
8,9,9,10,10,10,11,11,12
Α,Α,Α,Θ,Θ,Θ,Θ, Θ, Θ
N=9
$\bar{Y}$=10.0
σ^2=1.333

Για να γίνουν κατανοητές οι συνέπειες της στρωματοποίησης θα χρησιμοποιήσουμε τα δεδομένα του παραδείγματος 1.1 μόνο που τα άτομα του πληθυσμού εκτός από το ύψος τους χαρακτηρίζονται και από το φύλο (αρσενικά και θηλυκά). Γνωρίζοντας ότι το φύλο παιζει συχνά καθοριστικό ρόλο στο μέγεθος των ατόμων μπορούμε να χωρίσουμε τον πληθυσμό του παραδείγματος 4.2 σε δυο στρώσεις που περιλαμβάνουν τα αρσενικά και τα θηλυκά. Όπως και στο παράδειγμα 1.1 θα προσπαθήσουμε να σχηματίσουμε όλα τα δυνατά δείγματα ολικού μεγέθους 6 ατόμων από τον υποθετικό αυτό πληθυσμό. Μάλιστα για να απλοποιήσουμε την κατάσταση θα κρατήσουμε το ίδιο δειγματοληπτικό κλάσμα και στις δυο στρώσεις. Έτσι από τη στρώση 1 που περιλαμβάνει τα αρσενικά θα πάρουμε ένα δείγμα 2 ατόμων (f=2/3) και από τη στρώση 2 (τα θηλυκά) δείγμα 4 ατόμων (f=4/6=2/3). Οι συνδυασμοί 2 ατόμων από 3 είναι 3 και 4 ατόμων από 6 είναι 15. Οι τελικοί συνδυασμοί των δειγμάτων των δυο στρώσεων είναι 45 (15x3). Παίρνοντας λοιπόν όλα τα δυνατά δείγματα της στρωματοποιημένης δειγματοληψίας και εφαρμόζοντας τους τύπους της προηγουμένης παραγράφου υπολογίζουμε (εκτιμούμε) το μέσο ύψος των ατόμων του πληθυσμού, τη διασπορά του μέσου ύψους καθώς και το διάστημα εμπιστοσύνης. Στο παράδειγμα 1.1 είχαμε πάρει όλα τα δυνατά δείγματα μεγέθους 6 από τον πληθυσμό αυτό με μια απλή τυχαία δειγματοληψία.

Μπορούμε λοιπόν να συγκρίνουμε την στρωματοποιημένη με την απλή τυχαία δειγματοληψία κοιτάζοντας τα διαστήματα εμπιστοσύνης που δίνουν οι δυο αυτές στρατηγικές. Θυμίζουμε ότι το κόστος της δειγματοληψίας είναι το ίδιο και στις δυο περιπτώσεις ($n=6$ και $n=n_1+n_2=2+4=6$) και συνεπώς η στρατηγική που δίνει τα στενότερα διαστήματα εμπιστοσύνης που περιέχουν την πραγματική μέση τιμή είναι η καλύτερη. Η σύγκριση μπορεί να γίνει στο διάγραμμα 4.1. Από το διάγραμμα αυτό φαίνεται ότι:

- η στρωματοποιημένη τυχαία δειγματοληψία δίνει σημαντικά ακριβέστερες εκτιμήσεις από την απλή τυχαία
- όλα τα διαστήματα εμπιστοσύνης της στρωματοποιημένης δειγματοληψίας περιέχουν την πραγαματική μέση τιμή του πληθυσμού (που σ'αυτή την περίπτωση του εικονικού αυτού πληθυσμού είναι γνωστή)

Φαίνεται λοιπόν ότι με το ίδιο κόστος η στρωματοποίηση δίνει καλύτερα αποτελέσματα από την απλή τυχαία δειγματοληψία. Η αύξηση της ακρίβειας μεγαλώνει θεαματικά με την αύξηση της ομοιογένειας στο εσωτερικό των στρώσεων. Θεωρήστε για παράδειγμα στον προηγούμενο πληθυσμό όλα τα αρσενικά άτομα να είχαν ύψος 9 cm και όλα τα θηλυκά 11 cm.. Τότε η διασπορά των στρώσεων θα είναι 0 και συμφωνα με τους τύπους της προηγούμενης παραγράφου η διασπορά της μέσης τιμής του πληθυσμού θα είναι και αυτή μηδενική. Μόνο ένα από τα 84 δυνατά δείγματα 6 ατόμων της απλής τυχαίας δειγματοληψίας θα έδινε μηδενική διασπορά (αυτό που περιλαμβάνει τα 6 θηλυκά άτομα) αλλά δυστυχώς αυτό το δείγμα θα έδινε μια εκτίμηση της μέσης τιμής που θα ήταν μακριά από την πραγματική του πληθυσμού.



Σχημα 4.1 *Σύγκριση των εκτιμήσεων της απλής τυχαίας και της στρωματοποιημένης δειγματοληψίας. Παρουσιάζονται εκτιμήσεις της μέσης τιμής και των διαστημάτων εμπιστοσύνης (κατακόρυφα ευθύγραμμα τμήματα) που προέρχονται από όλα τα δυνατά δείγματα μεγέθους 6 ατόμων από τον πληθυσμό του παραδείγματος* 4.2.

ΠΑΡΑΔΕΙΓΜΑ 4.2

4.3. ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΚΑΙ ΜΕΙΟΝΕΚΤΗΜΑΤΑ ΤΗΣ ΣΤΡΩΜΑΤΟΠΟΙΗΜΕΝΗΣ ΤΥΧΑΙΑΣ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ

⇧ Πλεονεκτήματα ⇧

- Συνήθως οδηγεί σε ακριβέστερες εκτιμήσεις από την απλή τυχαία δειγματοληψία.
- Οι εκτιμητές είναι αμερόληπτοι (εκτός από τον εκτιμητή λόγου όταν αυτός χρησιμοποιείται στα πλαίσια αυτής της στρατηγικής).

⇩ Μειονεκτήματα ⇩C

- Ένα λάθος στον υπολογισμό του βάρους των στρώσεων οδηγεί σε μεροληψίες που δεν εξαλείφονται όσο κι αν μεγαλώσει το μέγεθος του δείγματος.
- Για τον παραπάνω λόγο η διπλή δειγματοληψία απαιτεί μια ευρεία πρώτη φάση.

- Είναι μια ευκολοπροσάρμοστη στρατηγική που συνδυάζεται και με άλλες οδηγώντας σε περίπλοκους σχεδιασμούς που όμως επιτρέπουν τον υπολογισμό της ακρίβειας των εκτιμητών και τη δημιουργία διαστημάτων εμπιστοσύνης.
- Επιτρέπει την κατ'επιλογή μεγαλύτερη συμμετοχή στο δείγμα ατόμων του πληθυσμού που προέρχονται από συγκεκριμένες στρώσεις (αυτό μπορεί να εξυπηρετήσει παράλληλες μελέτες).
- Επιτρέπει την ανάλυση της επίδρασης πάνω στα άτομα του πληθυσμού της παραμέτρου που χρησιμοποιήθηκε για τη στρωματοποίηση.
- Επιτρέπει τη διεξαγωγή της δειγματοληψίας ακόμα κι αν διακυμάνσεις στην κατανομή της προσπάθειας στο χώρο ή το χρόνο είναι αναπόφευκτες (σε κάποιες περιοχές η πρόσβαση είναι δύσκολη ή υπάρχουν δυσχέρειες για κάποιες περιόδους π.χ. νύχτα ή αργίες).
- Ακόμα κι αν γίνουν λάθη στη στην κατανομή των ατόμων στις στρώσεις οι εκτιμήσεις παραμένουν αμερόληπτες.
- Ακόμα κι αν δεν υπάρχουν πληροφορίες για μια έστω και στοιχειώδη στρωματοποίηση η στρατηγική αυτή εφαρμόζεται κατόπιν διπλής δειγματοληψίας (μια πρώτη χαλαρή δειγματοληψία για τη μελέτη των χαρακτηριστικών του πληθυσμού και του περιβάλλοντος του και στη συνέχεια με βάση αυτή την πληροφορία μια στρωματοποιημένη δειγματοληψία για τις τελικές εκτιμήσεις). Σ'αυτή την περίπτωση το κέρδος στην ακρίβεια της εκτίμησης φυσικά μειώνεται.
- Λόγω της εκ των προτέρων διαίρεσης του πληθυσμού σε στρώσεις (κατηγορίες, ομάδες) κάποιες από τις στατιστικές αναλύσεις δεν εφαρμόζονται άμεσα.