3. ΣΤΡΩΜΑΤΟΠΟΙΗΜΕΝΗ ΤΥΧΑΙΑ ΔΕΙΓΜΑΤΟΛΗΨΙΑ (Stratified Random Sampling)

Είναι προφανές από τα τυπικά σφάλματα των εκτιμητριών των προηγούμενων παραγράφων, ότι ένας τρόπος να αυξηθεί η ακρίβεια τους είναι να αυξηθεί το μέγεθος n του δείγματος. Αυτός δεν είναι και ο μοναδικός τρόπος. Είναι δυνατόν να αυξηθεί η ακρίβεια των εκτιμήσεων με την γνώση επιπρόσθετων πληροφοριών για την δομή του πληθυσμού. Ενδέχεται για παράδειγμα, ο πληθυσμός να μπορεί να θεωρηθεί ως αποτελούμενος από υποπληθυσμούς μέσα στους οποίους να υπάρχει μεγαλύτερη ομοιογένεια ως προς κάποιο χαρακτηριστικό από ό,τι σε ολόκληρο τον πληθυσμό. Αν λοιπόν, με κάποιο δειγματοληπτικό σχήμα, αντιπροσωπευθούν κατάλληλα όλοι οι υποπληθυσμοί στο δείγμα, τότε είναι δυνατόν να επιτύχουμε μικρότερη διασπορά για τις εκτιμήσεις των παραμέτρων του πληθυσμού από αυτή που έχουμε από ένα απλό τυχαίο δείγμα του ίδιου μεγέθους που δεν λαβαίνει υπ'όψη την δυνατότητα διάκρισης του πληθυσμού σε υποπληθυσμούς.

Ορισμός 3.1: Έστω πληθυσμός μεγέθους N και έστω ότι αυτός μπορεί να διαιρεθεί σε k εσωτερικά ομοιογενείς υποπληθυσμούς μεγέθους N_1, N_2, …, N_k. Αν αυτοί είναι ξένοι μεταξύ τους ώστε να ισχύει ότι $N_1+N_2+\ldots+N_k=N$, οι υποπληθυσμοί αυτοί ονομάζονται **στρώματα (strata)**.

Ορισμός 3.2: Έστω ότι από κάθε ένα από τα στρώματα ενός πληθυσμού επιλέγεται ένα απλό τυχαίο δείγμα μεγέθους n_i, i=1, 2, …, k ανεξάρτητα από τα άλλα. Το δείγμα μεγέθους $n=n_1+n_2+\ldots+n_k$ που προκύπτει από την ένωση των k ανεξαρτήτων απλών τυχαίων δειγμάτων ονομάζεται **στρωματοποιημένο τυχαίο δείγμα (stratified random sample)** και η

διαδικασία επιλογής του ονομάζεται **στρωματοποιημένη τυχαία δειγματοληψία (stratified random sampling)**.

Οι παράγραφοι που ακολουθούν ασχολούνται με τις ιδιότητες των εκτιμητριών των διαφόρων παραμέτρων του πληθυσμού και την καλύτερη δυνατή επιλογή των τιμών των μεγεθών $n_1, n_2, \ldots, n_k$ των υποδειγμάτων για την επίτευξη μέγιστης ακρίβειας.

3.1 Εκτίμηση Μέσης Τιμής Στρωματοποιημένου Πληθυσμού

Σύμφωνα με τους ορισμούς 3.1 και 3.2, τα δεδομένα του προβλήματος συνοψίζονται στον πίνακα 3.1.1.

Πίνακας 3.1.1

	Πληθυσμός				Δείγμα		
Στρώμα	Μέγεθος στρώματος	Μέση τιμή	Διασπορά		Μέγεθος απλού τυχ. δείγματος	Μέσος	Διασπορά
1	N_1	μ_1	σ_1^2		n_1	$\overline{X}_{n_1}$	$\frac{\sigma_1^2}{n_1}\left(1-\frac{n_1}{N_1}\right)$
2	N_2	μ_2	σ_2^2		n_2	$\overline{X}_{n_2}$	$\frac{\sigma_2^2}{n_2}\left(1-\frac{n_2}{N_2}\right)$
.	.	.	.		.	.	.
.	.	.	.		.	.	.
.	.	.	.		.	.	.
i	N_i	μ_i	σ_i^2		n_i	$\overline{X}_{n_i}$	$\frac{\sigma_i^2}{n_i}\left(1-\frac{n_i}{N_i}\right)$
.	.	.	.		.	.	.
.	.	.	.		.	.	.
.	.	.	.		.	.	.
k	N_k	μ_k	σ_k^2		n_k	$\overline{X}_{n_k}$	$\frac{\sigma_k^2}{n_k}\left(1-\frac{n_k}{N_k}\right)$

Εδώ $\bar{X}_{n_i} = \sum_{j=1}^{n_i} X_j^{(i)}$, i = 1, 2, …, k, όπου $X_j^{(i)}$ είναι η j μονάδα του i δείγματος.

Είναι προφανές ότι η μέση τιμή μ του πληθυσμού και οι μέσες τιμές $\mu_1, \mu_2, \ldots, \mu_k$ των υποπληθυσμών συνδέονται με την σχέση

$$\mu = \frac{1}{N}\sum_{i=1}^{k} N_i \mu_i \,. \qquad (3.1.1)$$

Τα αντίστοιχα δειγματικά μεγέθη συνδέονται με μια παρόμοια σχέση:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{X}_{n_i} \,,$$

όπου $\bar{X}_n$ είναι ο μέσος του συνολικού δείγματος.

Στην πράξη όμως, η στατιστική συνάρτηση που χρησιμοποιείται ως εκτιμήτρια του μ δεν είναι η $\bar{X}_n$, αλλά η

$$\hat{\mu}_n = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{X}_{n_i} \,, \qquad (3.1.2)$$

στην οποία οι μέσοι των δειγμάτων από τα διάφορα στρώματα σταθμίζονται με τους συντελεστές βαρύτητας N_i/N, i=1, 2, …, k των στρωμάτων.

Προφανώς,

$$\hat{\mu}_n = \bar{X}_n \text{ αν } \frac{n_i}{n} = \frac{N_i}{N}, \text{ i=1, 2, …, k}$$

ή, ισοδύναμα,

$$\text{αν } \frac{n_i}{N_i} = \frac{n}{N} \Leftrightarrow f_i = f\ ,\ \ i=1, 2, \ldots, k.$$

Δηλαδή, οι δυο στατιστικές συναρτήσεις συμπίπτουν αν όλα τα δείγματα εκπροσωπούν το ίδιο ποσοστό μονάδων των αντίστοιχων στρωμάτων.

Θεώρημα 3.1.1: Η στατιστική συνάρτηση $\hat{\mu}_n$ είναι αμερόληπτη εκτιμήτρια της μέσης τιμής μ του πληθυσμού και ισχύει ότι

$$V(\hat{\mu}_n) = \sum_{i=1}^{k} \left(\frac{N_i}{N} \right)^2 \frac{\sigma_i^2}{n_i} \left(1 - \frac{n_i}{N_i} \right) \qquad (3.1.3)$$

Απόδειξη:

Αμεροληψία:

Ισχύει εδώ ότι

$$E(\hat{\mu}_n) = \frac{1}{N} \sum_{i=1}^{k} N_i E(\bar{X}_{n_i})\ .$$

Αλλά, οι στατιστικές συναρτήσεις $\bar{X}_{n_1}$, $\bar{X}_{n_2}$, ..., $\bar{X}_{n_k}$ ως μέσοι απλών τυχαίων δειγμάτων είναι αμερόληπτες εκτιμήτριες των μέσων τιμών μ_1, μ_2, …, μ_k αντίστοιχα των (υπο)πληθυσμών, από τους οποίους τα δείγματα αυτά επελέγησαν. Άρα,

$$E\left(\bar{X}_{n_i} \right) = \mu_i$$

και, επομένως,

$$E(\hat{\mu}_n) = \frac{1}{N} \sum_{i=1}^{k} N_i \mu_i = \mu\ .$$

Η τελευταία σχέση αποδεικνύει ότι η $\hat{\mu}_n$ είναι αμερόληπτη εκτιμήτρια του μ. Ισχύει επίσης ότι

$$V(\hat{\mu}_n) = V\left(\frac{1}{N}\sum_{i=1}^{k} N_i \bar{X}_{n_i}\right)$$
$$= \frac{1}{N^2}\left\{\sum_{i=1}^{k} N_i^2 V(\bar{X}_{n_i}) + 2\sum_{1\le i<j\le k} N_i N_j \mathrm{Cov}(\bar{X}_{n_i}\bar{X}_{n_j})\right\}.$$

Αλλά, $\mathrm{Cov}(\bar{X}_{n_i}, \bar{X}_{n_j}) = 0$, αφού τα k απλά τυχαία δείγματα είναι αμοιβαία ανεξάρτητα. Άρα, ισχύει η σχέση

$$V(\hat{\mu}_n) = \frac{1}{N^2}\sum_{i=1}^{k} N_i^2 V(\bar{X}_{n_i}) \ ,$$

η οποία οδηγεί στην (3.1.3) και ολοκληρώνει την απόδειξη του θεωρήματος.

Έστω $y_j^{(i)}$ η j μονάδα του πληθυσμού που ανήκει στο i στρώμα. Τότε ισχύει το εξής πόρισμα.

Πόρισμα: Η στατιστική συνάρτηση

$$\hat{Y}_{st} = N\hat{\mu}_n \qquad (3.1.4)$$

είναι αμερόληπτη εκτιμήτρια του συνολικού μεγέθους

$$y = \sum_{i=1}^{k}\sum_{j=1}^{k} y_j^{(i)}$$

και ισχύει ότι

$$V(\hat{Y}_{st}) = \sum_{i=1}^{k} N_i^2 \frac{\sigma_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right). \qquad (3.1.5)$$

Απόδειξη: Ισχύει προφανώς ότι

$$y = \sum_{i=1}^{k}\sum_{j=1}^{k} y_j^{(i)} = \sum_{i=1}^{k} N_i\mu_i = N\mu .$$

Επομένως, $E(\hat{Y}_{st}) = NE(\hat{\mu}_n) = N\mu = y$ και $V(\hat{Y}_{st}) = N^2V(\hat{\mu}_n)$. Η σχέση αυτή σε συνδυασμό με την (3.1.3) οδηγεί στην (3.1.5).

Επειδή από κάθε στρώμα επιλέγεται ένα απλό τυχαίο δείγμα, έπεται ότι η στατιστική συνάρτηση

$$S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(X_j^{(i)} - \bar{X}_{n_i})^2$$

είναι αμερόληπτη εκτιμήτρια της διασποράς σ_i^2 του i στρώματος. (Εδώ, $X_j^{(i)}$ είναι η j παρατήρηση του i δείγματος). Κατά συνέπεια, ισχύει το εξής θεώρημα.

Θεώρημα 3.1.2: Η στατιστική συνάρτηση

$$S_{\hat{\mu}_n}^2 = \sum_{i=1}^{k}\left(\frac{N_i}{N}\right)^2 \frac{S_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)$$

είναι αμερόληπτη εκτιμήτρια της $V(\hat{\mu}_n)$.

Παρατήρηση: Προφανώς, μια αμερόληπτη εκτιμήτρια της $V\left(\hat{Y}_{st}\right)$ είναι η $N^2 S^2_{\hat{\mu}_n}$.

Μπορεί να αποδειχθεί, κάτω από την υπόθεση ότι η $X_j^{(i)}$ είναι κανονική τυχαία μεταβλητή, ότι η κατανομή της στατιστικής συνάρτησης $(\hat{\mu}_n - \mu)/S_{\hat{\mu}_n}$ είναι κατά προσέγγιση η t με n_e βαθμούς ελευθερίας, όπου n_e είναι μια τιμή μεταξύ του $\min\{n_1\text{-}1,\ldots,n_k\text{-}1\}$ και του $\sum_{i=1}^{k}(n_i - 1)$. Συνήθως, η τιμή του n_e που χρησιμοποιείται είναι η

$$n_e = \left(\sum_{i=1}^{k} g_i S_i^2 \right)^2 \Bigg/ \sum_{i=1}^{k} \frac{g_i^2 S_i^4}{n_i - 1}, \qquad (3.1.6)$$

όπου $g_i=N_i(N_i-n_i)/n_i$, i=1, 2, …, k. Επομένως, ένα 100(1-α)% διάστημα εμπιστοσύνης για την παράμετρο μ έχει άκρα

$$\hat{\mu}_n \pm t_{n_e,1-\alpha/2}\, S_{\hat{\mu}_n} . \qquad (3.1.7)$$

Παράδειγμα: Έστω ένας πληθυσμός μεγέθους 1000, ο οποίος είναι κατά το 60% αγροτικός. Για να εκτιμηθεί το μέσο μηνιαίο εισόδημα (σε εκατοντάδες ευρώ) επελέγη ένα στρωματοποιημένο τυχαίο δείγμα μεγέθους 5 το οποίο οδήγησε στα αποτελέσματα που δίνονται στη συνέχεια. Να κατασκευασθεί ένα 95% διάστημα εμπιστοσύνης για το πραγματικό μέσο εισόδημα του πληθυσμού.

Αγροτικό Τμήμα	Μη αγροτικό Τμήμα
10	10
15	6
11	

Λύση: Έστω $X_j^{(i)}$ = η j παρατήρηση του i στρώματος

$\bar{X}_{n_i}$ = ο μέσος του i στρώματος

S_i^2 = η διασπορά του i στρώματος .

Τότε,

$$\bar{X}_{n_1} = \sum_{j=1}^{3} X_j^{(1)} \Big/ 3 = 12\,,$$

$$\bar{X}_{n_2} = \sum_{j=1}^{2} X_j^{(2)} \Big/ 2 = 8\,.$$

Επομένως,

$$\hat{\mu} = 12(0.6) + 8(0.4) = 10.4,$$

$$S_1^2 = \sum_{j=1}^{3} (X_j^{(1)} - \bar{X}_{n_1})^2 \Big/ 2 = (2^2 + 3^2 + 1^2)/2 = 7\,,$$

$$S_2^2 = \sum_{j=1}^{2} (X_j^{(2)} - \bar{X}_{n_2})^2 \Big/ 1 = (1^2 + 3^2)/1 = 10\,.$$

Κατά συνέπεια,

$$S_{\hat{\mu}}^2 = \frac{(0.6)(7) + (0.4)(10)}{5}\left(1 - \frac{5}{1000}\right) = 1.63\,,$$

οπότε

$$S_{\hat{\mu}} = 1.28.$$

Άρα τα άκρα του 95% διαστήματος εμπιστοσύνης είναι

$$10.8 \pm t_{n_e, 0.975}(1.28)\,,$$

όπου $n_e = (\sum_{i=1}^{2} g_i S_i^2)^2 \Big/ \sum_{i=1}^{2} \frac{g_i^2 \, S_i^4}{n_i - 1}$, $g_i = \frac{N_i(N_i - n_i)}{n_i}$, i=1, 2.

Εδώ,

$$g_1 = \frac{N_1(N_1 - n_1)}{n_1} = \frac{(N_1/N)[(N_1/N) - (n_1/N)]}{n_1/N_1^2} = \frac{0.6(0.6 - 0.003)}{0.00003} = 11940$$

και

$$g_2 = \frac{0.4(0.4 - 0.002)}{0.00002} = 7960.$$

Άρα,

$$(\sum_{i=1}^{2} g_i S_i^2)^2 = (163180)^2,$$

$$\sum_{i=1}^{2} \frac{g_i \, S_i^2}{n_i - 1} = 98289682$$

και, επομένως, $n_e = 2.708 \simeq 3$.

Κατά συνέπεια, το 95% διάστημα εμπιστοσύνης για το πραγματικό μέσο εισόδημα του πληθυσμού ορίζεται από τα άκρα

$$10.8 \pm t_{3,0.975}(1.28),$$

όπου $t_{3,0.975}$ = 3.182. Δηλαδή, τελικά, τα άκρα του ζητούμενου διαστήματος είναι

$$10.8 \pm 4.073.$$

Άρα, με πιθανότητα ίση με 95%, το πραγματικό μέσο μηνιαίο εισόδημα είναι μεταξύ 672,70 και 1487,30 ευρώ.

3.2 Επιλογή των $n_1, n_2, \ldots, n_k$

Το πρόβλημα που συνδέεται άμεσα με την τεχνική της στρωματοποιημένης δειγματοληψίας είναι ο καταμερισμός του συνολικού δειγματικού μεγέθους n στα k διαθέσιμα στρώματα, δηλαδή ο καθορισμός των τιμών των μεγεθών $n_1, n_2, \ldots, n_k$ των k απλών τυχαίων δειγμάτων.

Αν το δειγματοληπτικό κόστος ανά μονάδα είναι το ίδιο σε όλα τα στρώματα και οι διασπορές των στρωμάτων δεν διαφέρουν σημαντικά, τα μεγέθη $n_1, n_2, \ldots, n_k$ συνηθίζεται να επιλέγονται έτσι ώστε

$$\frac{n_i}{N_i} = \frac{n}{N}, \quad i=1, 2, \ldots, k.$$

Ο σχεδιασμός αυτός είναι γνωστός ως **αναλογικός καταμερισμός του n (proportional allocation)** και η δειγματοληπτική τεχνική ονομάζεται **αναλογική στρωματοποιημένη τυχαία δειγματοληψία (proportional stratified random sampling)**. Στην περίπτωση αυτή,

$$n_i = n \frac{N_i}{N}, \quad i=1, 2, \ldots, k, \qquad (3.2.1)$$

δηλαδή, το μέγεθος του δείγματος από ένα στρώμα είναι ανάλογο του ποσοστού των μονάδων του πληθυσμού που το στρώμα εκπροσωπεί.

Υπάρχουν, όμως, περιπτώσεις όπου οι τιμές του πληθυσμού έχουν μεγαλύτερη διακύμανση σε μερικά στρώματα από ό,τι σε άλλα. Διαφέρουν δηλαδή σημαντικά οι διασπορές των στρωμάτων. Επομένως, για να αντιπροσωπευθούν επαρκώς τα στρώματα αυτά στο δείγμα, θα πρέπει ο λόγος $\frac{n_i}{N_i}$ να είναι ανάλογος της τυπικής απόκλισης σ_i του στρώματος. Αυτό σημαίνει ότι στρώματα με μεγαλύτερη διακύμανση από άλλα πρέπει να εκπροσωπούνται από μεγαλύτερο τμήμα του

δείγματος, για να αυξηθεί η ακρίβεια των εκτιμήσεων. Υποθέτοντας ότι το δειγματοληπτικό κόστος ανά μονάδα είναι το ίδιο για όλα τα στρώματα, αποδεικνύεται ότι η $V(\hat{\mu}_n)$ γίνεται ελάχιστη αν τα $n_1, n_2, \ldots, n_k$ επιλεγούν έτσι ώστε

$$n_i = n \frac{N_i \, \sigma_i}{\sum_{j=1}^{k} N_j \, \sigma_j}, \quad i=1, 2, \ldots, k. \qquad (3.2.2)$$

Ο σχεδιασμός αυτός είναι γνωστός ως **βέλτιστος καταμερισμός του n με σταθερό κόστος ανά δειγματοληπτική μονάδα (optimum allocation with constant cost per unit)** ή **καταμερισμός κατά Neyman (Neyman allocation)**.

Αν το δειγματοληπτικό κόστος ανά μονάδα διαφέρει από στρώμα σε στρώμα, τότε είναι φυσικό να προσπαθήσει ο ερευνητής να αυξήσει την ακρίβεια των εκτιμήσεών του επιλέγοντας τα n_i, αντιστρόφως ανάλογα των c_i, i=1, 2, …, k.

Έστω ότι το συνολικό κόστος c μιας δειγματοληψίας είναι συνάρτηση των c_i, i=1, 2, …, k, δηλαδή, έστω ότι

$$c = c_0 + \sum_{i=1}^{k} n_i \, c_i, \qquad c_0 > 0.$$

Τότε, αποδεικνύεται ότι αν το κόστος c έχει μια δοθείσα τιμή, οι τιμές των n_i που ελαχιστοποιούν την διασπορά $V(\hat{\mu}_n)$ δίνονται από τον τύπο

$$n_i = n \frac{N_i \, \sigma_i / \sqrt{c_i}}{\sum_{j=1}^{k} N_j \, \sigma_j / \sqrt{c_j}}, \quad i=1, 2, \ldots, k \qquad (3.2.4)$$

Οι τιμές (3.2.4) ελαχιστοποιούν επίσης το κόστος εάν η διασπορά $V(\hat{\mu}_n)$ έχει μια δοθείσα τιμή.

Ο παραπάνω καταμερισμός του n ονομάζεται **βέλτιστος καταμερισμός (optimum allocation).**

Η σχέση (3.2.4) υπονοεί ότι το μέγεθος του απλού τυχαίου δείγματος, που επιλέγεται από ένα στρώμα, πρέπει να είναι μεγαλύτερο από τα μεγέθη των δειγμάτων άλλων στρωμάτων, αν το μέγεθος του στρώματος είναι μεγαλύτερο, η διασπορά του στρώματος είναι μεγαλύτερη και το κόστος ανά μονάδα του στρώματος είναι χαμηλότερο.

Παρατήρηση 1: Η (3.2.4) οδηγεί στην (3.2.2) αν $c_i=c$, i=1, 2, …, k, δηλαδή, αν υπάρχει ομοιόμορφο κόστος. Επίσης, η (3.2.4) οδηγεί στην (3.2.1), αν, εκτός από ομοιόμορφο κόστος, υπάρχει και ομοιόμορφη διακύμανση στις τιμές των μονάδων των διαφόρων στρωμάτων, δηλαδή, αν $c_i=c$ και $\sigma_i^2 = \sigma_0^2$, i=1, 2, …, k.

Παρατήρηση 2: Στην περίπτωση (3.2.4) είναι δυνατόν το μέγεθος n του συνολικού δείγματος να μην είναι προκαθορισμένο. Αυτό μπορεί να καθορισθεί με την βοήθεια της συνάρτησης κόστους, αν το συνολικό κόστος c έχει μια δοθείσα τιμή. Πράγματι, αντικαθιστώντας τις τιμές των n_i, όπως αυτές δίνονται από την (3.2.4) στην (3.2.3) και λύνοντας ως προς n, έχουμε

$$n = \frac{(c-c_0)\sum_{i=1}^{k} N_i\,\sigma_i/\sqrt{c_i}}{\sum_{i=1}^{k} N_i\,\sigma_i/\sqrt{c_i}}.$$

Στην περίπτωση που η επιθυμητή τιμή της $V(\hat{\mu}_n)$ είναι V, αντικαθιστώντας τις τιμές των n_i από την (3.2.4) στην (3.1.3) και λύνοντας ως προς n , έχουμε

$$n = \frac{\sum_{i=1}^{k} \frac{N_i\ \sigma_i\ \sqrt{c_i}}{N} \sum_{i=1}^{k} \frac{N_i\ \sigma_i}{N\ \sqrt{c_i}}}{V + \frac{1}{N} \sum_{i=1}^{k} \frac{N_i\ \sigma_i^2}{N}}.$$

Παράδειγμα: Έστω ότι σε κάποιο επάγγελμα, οι μηνιαίοι βασικοί μισθοί ανδρών – γυναικών σε εκατοντάδες ευρώ είναι οι εξής:

Κατηγορία i		1	2	3	4	5	6	7
Αποδοχές	(z_i)	10	15	20	25	30	35	40
Άνδρες	(α_i)	0	0	100	200	400	200	100
Γυναίκες	(γ_i)	500	500	500	0	0	0	0

(α) Να υπολογισθεί ο μέσος βασικός μισθός του πληθυσμού αυτού
(β) Να εκτιμηθεί το σφάλμα που ένας ερευνητής θα κάνει αν προσπαθήσει να εκτιμήσει τον μέσο βασικό μισθό με βάση ένα δείγμα μεγέθους n=25 χρησιμοποιώντας

(i) απλή τυχαία δειγματοληψία,
(ii) αναλογική στρωματοποιημένη τυχαία δειγματοληψία,
(iii) βέλτιστη στρωματοποιημένη τυχαία δειγματοληψία.

Λύση:

(**α)** Υπάρχουν δυο στρώματα μεγέθους $N_1 = \sum_{i=1}^{7} \alpha_i = 1000$

και $N_2 = \Sigma\gamma_i = 1500$.
Αν μ_i είναι η μέση τιμή του i στρώματος, τότε

$$\mu = (N_1\mu_1 + N_2\mu_2)/N.$$

Αλλά,

$$\mu_1 = \sum_{i=1}^{7} \alpha_i z_i / N_1 = 30,$$

$$\mu_2 = \sum_{i=1}^{7} \gamma_i z_i / N_2 = 15 \text{ και } N = N_1 + N_2 = 2500.$$

Άρα, $\mu = 21$.

(β)

(i) Αγνοώντας την διαίρεση του πληθυσμού σε δυο στρώματα, έστω $\upsilon_i = \alpha_i + \gamma_i$. Τότε, η χρησιμοποιούμενη εκτιμήτρια $\bar{X}_n$ της μέσης τιμής μ του πληθυσμού έχει διασπορά

$$\sigma^2_{\bar{X}_n} = \frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right),$$

όπου

$$\sigma^2 = \{\sum_{i=1}^{7} \upsilon_i z_i^2 - (\sum_{i-1}^{7} \upsilon_i z_i)^2 \Big/ N\} \Big/ (N-1) = 76.03.$$

Δηλαδή,

$$\sigma^2_{\bar{X}_n} = \frac{76.03}{25}\left(1 - \frac{25}{2500}\right) = 3.01 \Rightarrow \sigma_{\bar{X}_n} = 1.735.$$

(ii) $n_1 = N_1(n/N) = 10$
$n_2 = N_2(n/N) = 15.$

Η χρησιμοποιούμενη εκτιμήτρια είναι η

$$\hat{\mu}_n = \sum_{i=1}^{2} N_i \bar{X}_{n_i} \Big/ N.$$

Άρα,

$$\sigma^2_{\hat{\mu}} = V(\hat{\mu}_n) = \sum_{i=1}^{2} \left(\frac{N_i}{N} \right)^2 \frac{\sigma_i^2}{n_i} \left(1 - \frac{n_i}{N_i} \right).$$

Αλλά,

$$\sigma_1^2 = \{ \sum_{i=1}^{7} \alpha_i z_i^2 - (\sum_{i=1}^{7} \alpha_i z_i)^2 / N_1 \} / (N_1\text{-}1) = 30.03.$$

$$\sigma_2^2 = \{ \sum_{i=1}^{7} \gamma_i z_i^2 - (\sum_{i=1}^{7} \gamma_i z_i)^2 / N_2 \} / (N_2\text{-}1) = 16.67.$$

$$\Rightarrow \sigma^2_{\hat{\mu}_n} = 0.872 \Rightarrow \sigma_{\hat{\mu}_n} = 0.934.$$

(iii) $n_1 = N_1\sigma_1 / (N_1\sigma_1 + N_2\sigma_2) = 11.8 \approx 12,$

$n_2 = N_2\sigma_2 / (N_1\sigma_1 + N_2\sigma_2) = 13.19 \approx 13.$

$$\Rightarrow \sigma^2_{\hat{\mu}_n} = 0.854,$$

$$\Rightarrow \sigma_{\hat{\mu}_n} = 0.924.$$

Παρατήρηση: Και στις δύο περιπτώσεις της στρωματοποιημένης δειγματοληψίας, η διασπορά της εκτίμησης της παραμέτρου μ είναι ίση με το 30% περίπου της διασποράς της εκτίμησης, στην οποία οδηγούμεθα με απλή τυχαία δειγματοληψία.

3.3 Σχετική Ακρίβεια Στρωματοποιημένης Τυχαίας Δειγματοληψίας και Απλής Τυχαίας Δειγματοληψίας

Η στρωματοποίηση έχει εν γένει ως αποτέλεσμα μικρότερη διασπορά για την εκτιμήτρια της μέσης τιμής του πληθυσμού. Δεν είναι, όμως, αληθές ότι οποιοδήποτε στρωματοποιημένο δείγμα δίνει μικρότερη διασπορά από ένα απλό τυχαίο δείγμα. Αν οι τιμές των

$n_1, n_2, \ldots, n_k$ απέχουν πολύ από αυτές του βέλτιστου καταμερισμού, τότε η εκτίμηση της μέσης τιμής μπορεί να έχει μεγαλύτερη διασπορά.

Έστω $V_A(\hat{\mu}_n)$ και $V_B(\hat{\mu}_n)$ η διασπορά της εκτιμήτριας $\hat{\mu}_n$ της μέσης τιμής μ στην περίπτωση αναλογικής στρωματοποιημένης δειγματοληψίας και βέλτιστης στρωματοποιημένης δειγματοληψίας, αντίστοιχα. Έστω $\overline{X}_n$ ο μέσος ενός απλού τυχαίου δείγματος μεγέθους n. Τότε, ισχύει το εξής θεώρημα.

Θεώρημα 3.3.1: Αν το μέγεθος n του δείγματος είναι δεδομένο και στις εκφράσεις των $V(\overline{X}_n), V_A(\hat{\mu}_n)$ και $V_B(\hat{\mu}_n)$ όροι ως προς $1/N_i$ είναι αμελητέοι (και επομένως όροι ως προς 1/N είναι αμελητέοι), ισχύει ότι

$$V_B(\hat{\mu}_n) \le V_A(\hat{\mu}_n) \le V(\overline{X}_n).$$

Απόδειξη: Ισχύει ότι

$$V\left(\overline{X}_n\right) = \left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n}, \tag{3.3.1}$$

$$V_A\left(\hat{\mu}_n\right) = \left(1 - \frac{n}{N}\right)\frac{\sum_{i=1}^{k}(N_i/N)\sigma_i^2}{n}$$

$$= \frac{\sum_{i=1}^{k}(N_i/N)\sigma_i^2}{n} - \frac{\sum_{i=1}^{k}(N_i/N)\sigma_i^2}{N}, \tag{3.3.2}$$

$$V_B(\hat{\mu}_n) = \frac{(\sum_{i=1}^{k}(N_i/N)\sigma_i)^2}{n} - \frac{\sum_{i=1}^{k}(N_i/N)\sigma_i^2}{N}. \tag{3.3.3}$$

Οι δυο τελευταίες σχέσεις προκύπτουν από την (3.1.5) σε συνδυασμό με την (3.2.1) και (3.2.2) αντίστοιχα.

Θα δειχθεί ότι $\sigma^2 \geq \sum_{i=1}^{k}(N_i/N)\sigma_i^2$.

Πράγματι,

$$\sigma^2(N-1) = \sum_{j=1}^{k}\sum_{i=1}^{N_j}(y_i^{(j)} - \mu)^2 = \text{ (προσθαφαιρώντας το } \mu_j)$$

$$= \sum_{j=1}^{k}\sum_{i=1}^{N_j}(y_i^{(j)} - \mu_j)^2 + \sum_{j=1}^{k}\sum_{i=1}^{N_j}(\mu_j - \mu)^2$$

$$= \sum_{j=1}^{k}(N_j - 1)\sigma_j^2 + \sum_{j=1}^{k}N_j(\mu_j - \mu)^2 .$$

Διαιρώντας και τα δυο μέλη με N και αγνοώντας τους όρους ως προς $1/N_i$ και ως προς $1/N$ (οι οποίοι είναι αμελητέοι), έχουμε

$$\sigma^2 = \sum_{j=1}^{k}(N_j/N)\sigma_j^2 + \sum_{j=1}^{k}(N_j/N)(\mu_j - \mu)^2 .$$

Επειδή ο δεύτερος προσθετέος είναι μη αρνητική ποσότητα, έπεται ότι

$$\sigma^2 \geq \sum_{j=1}^{k}(N_j/N)\sigma_j^2 .$$

Άρα, από τις σχέσεις (3.3.1) και (3.3.2), προκύπτει ότι

$$V(\bar{X}_n) \geq V_A(\hat{\mu}_n) .$$

Επιπλέον, είναι προφανές ότι

$$V_A(\hat{\mu}_n) - V_B(\hat{\mu}_n) = \frac{1}{n}\left\{\sum_{i=1}^{k}(N_i/N)\sigma_i^2 - \left[\sum_{i=1}^{k}(N_i/N)\sigma_i\right]^2\right\}$$

$$= \frac{1}{n}\sum_{i=1}^{k}\frac{N_i}{N}\left[\sigma_i - \left(\sum_{i=1}^{k}\frac{N_i}{N}\sigma_i\right)\right]^2 \geq 0.$$

Άρα, $V_A(\hat{\mu}_n) \geq V_B(\hat{\mu}_n)$ και, επομένως, αποδείχθηκε το θεώρημα.

3.4 Το μέγεθος του Δείγματος

Στην παράγραφο 3.2 αντιμετωπίσθηκε μια μορφή του προβλήματος καθορισμού του κατάλληλου μεγέθους n του συνολικού δείγματος στις περιπτώσεις προκαθορισμένου συνολικού κόστους και προκαθορισμένης διασποράς της εκτιμήτριας $\hat{\mu}_n$ (Παρατήρηση 2).

Γενικότερα, έστω ότι απαιτείται ένα στρωματοποιημένο τυχαίο δείγμα μεγέθους n του οποίου το $100W_i\%$ των μονάδων θα προέρχεται από το i στρώμα, i=1, 2, …, k. Έστω, δηλαδή, ότι το n καταμερίζεται σύμφωνα με την σχέση

$$n_i = W_i n, \quad i=1, 2, ..., k, \quad \sum_{i=1}^{k} W_i = 1. \tag{3.4.1}$$

Το πρόβλημα είναι ο προσδιορισμός της τιμής του n που εξασφαλίζει μια δοθείσα τιμή V για την διασπορά της εκτιμήτριας $\hat{\mu}_n$. Πρέπει δηλαδή το n να καθορισθεί έτσι ώστε

$$V(\hat{\mu}_n) = V \tag{3.4.2}$$

ή, ισοδύναμα, έτσι ώστε

$$\sum_{i=1}^{k}\left(\frac{N_i}{N}\right)^2\frac{\sigma_i^2}{n_i}\left(1-\frac{n_i}{N_i}\right)=V.$$

Η τελευταία σχέση σε συνδυασμό με την (3.4.1) γράφεται

$$\frac{1}{n}\sum_{i=1}^{k}\frac{(N_i/N)\sigma_i^2}{W_i}-\frac{1}{N}\sum_{i=1}^{k}(N_i/N)\sigma_i^2=V.$$

Λύνοντας ως προς n, έχουμε

$$n=\frac{n_0}{1+\left[\sum_{i=1}^{k}(N_i/N)\sigma_i^2\right]\Big/(NV)}, \qquad (3.4.3)$$

όπου

$$n_0=\frac{1}{V}\sum_{i=1}^{k}\frac{(N_i/N)^2\sigma_i^2}{W_i}. \qquad (3.4.4)$$

Συνήθως, το μέγεθος του απαιτούμενου συνολικού δείγματος καθορίζεται με βάση τον βαθμό εμπιστοσύνης, με την οποία η εκτίμηση $\hat{\mu}_n$ της μέσης τιμής μ δεν απέχει από την μ περισσότερο από μια δοθείσα τιμή e. Δηλαδή, το πρόβλημα, όπως και στην απλή τυχαία δειγματοληψία, μπορεί να διατυπωθεί ως εξής:

"Ποιό είναι το μέγεθος n του στρωματοποιημένου τυχαίου δείγματος που απαιτείται ώστε

$$P\left(\left|\hat{\mu}_n-\mu\right|\le e\right)=1-\alpha;" \qquad (3.4.5)$$

Η (3.4.5) είναι ισοδύναμη με την

$$P\left(\left|\frac{\hat{\mu}_n-\mu}{\sqrt{V(\hat{\mu}_n)}}\right|\le\frac{e}{\sqrt{V(\hat{\mu}_n)}}\right)=1-\alpha\Leftrightarrow$$

$$\frac{e}{\sqrt{V(\hat{\mu}_n)}}=z_{1-\alpha/2}\,. \tag{3.4.6}$$

Αλλά,

$$V(\hat{\mu}_n)=\sum_{i=1}^{k}\left(\frac{N_i}{N}\right)^2\frac{\sigma_i^2}{n_i}\left(1-\frac{n_i}{N_i}\right)=\ (\text{από την (3.4.1)})$$

$$=\frac{1}{n}\sum_{i=1}^{k}\frac{(N_i/N)\sigma_i^2}{W_i}\ -\ \frac{1}{N}\sum_{i=1}^{k}(N_i/N)\,\sigma_i^2\,. \tag{3.4.7}$$

Αντικαθιστώντας στην (3.4.6) την τιμή της $V(\hat{\mu}_n)$ από την (3.4.7) και λύνοντας ως προς n, έχουμε

$$n=\frac{n_0}{1+\left[z_{1-\alpha/2}^2\sum_{i=1}^{k}(N_i/N)\sigma_i^2\right]/(Ne^2)}, \tag{3.4.8}$$

όπου

$$n_0=\frac{z_{1-\alpha/2}^2}{e^2}\sum_{i=1}^{k}\frac{(N_i/N)^2\sigma_i^2}{W_i}\,. \tag{3.4.9}$$

Στην πράξη, η τιμή του n καθορίζεται και στις δυο περιπτώσεις ως εξής:

$$n=\begin{cases}\ (3.4.4)\ \dot{\eta}\ (3.4.9), & \alpha\nu\quad n_0/N\le 0.05\\ \ (3.4.3)\ \dot{\eta}\ (3.4.8), & \alpha\nu\quad n_0/N>0.05\ .\end{cases}$$

Παράδειγμα: Στο παράδειγμα της παραγράφου 3.2, ποιο είναι το μέγεθος n του δείγματος που απαιτείται, ώστε, με πιθανότητα 95%, η εκτίμηση της μέσης τιμής του πληθυσμού να μην διαφέρει από την πραγματική τιμή περισσότερο από 1

(i) στην περίπτωση αναλογικού καταμερισμού και

(ii) στην περίπτωση βέλτιστου καταμερισμού;

Λύση: e=1, 1-α=0.95. Άρα, $z_{0.975}=1.96$ και, επομένως, από τις σχέσεις (3.4.9) και (3.4.8), έχουμε

$$n_0 = (1.96)^2\left[\frac{(0.4)^2(30.02)}{W_1}+\frac{(0.6)^2(16.67)}{W_2}\right]$$
$$=\frac{18.45}{W_1}+\frac{23.05}{W_2}$$

και

$$n=\frac{n_0}{1+\frac{(1.96)^2}{2500}\left[(0.4)(30.02)+(0.6)(16.67)\right]}=$$
$$=\frac{n_0}{1.03382}.$$

Κατά συνέπεια,

(i) $$W_i=\frac{N_i}{N}=\begin{cases}0.4\,, & i=1\\ 0.6\,, & i=2\end{cases} \quad\Rightarrow\quad n_0=84.54,$$

οπότε (επειδή $n_0/N=0.03<0.05$), $n=n_0\simeq 85$.

(ii) $$W_i=\frac{N_i\,\sigma_i}{N_1\,\sigma_1+N_2\,\sigma_2}=\begin{cases}0.47\,, & i=1\\ 0.53\,, & i=2\end{cases} \quad\Rightarrow\quad n_0=82.75$$

και, άρα, $n \simeq 83$.

3.5 Εκτίμηση Ποσοστών

Έστω $N_A^{(i)}$ ο αριθμός των μονάδων του i στρώματος του πληθυσμού που ανήκουν σε μια κατηγορία A και $X^{(i)}$ ο αριθμός των αντίστοιχων μονάδων του i απλού τυχαίου δείγματος. Τότε, το ποσοστό p των μονάδων του πληθυσμού που ανήκουν στην κατηγορία A είναι ίσο με

$$p = \sum_{i=1}^{k} N_A^{(i)} \Big/ N$$
$$= \sum_{i=1}^{k} (N_i / N) p_i ,$$

όπου

$$p_i = \frac{N_A^{(i)}}{N_i} = \text{ ποσοστό των μονάδων του i στρώματος που ανήκουν στην κατηγορία A.}$$

Θεώρημα 3.5.1: Η στατιστική συνάρτηση

$$\hat{p} = \sum_{i=1}^{k} (N_i / N) \hat{p}_i , \qquad (3.5.1)$$

όπου

$$\hat{p}_i = \frac{X^{(i)}}{n_i}, \qquad (3.5.2)$$

είναι αμερόληπτη εκτιμήτρια της παραμέτρου p και ισχύει ότι

$$V(\hat{p}) = \sum_{i=1}^{k} \left(\frac{N_i}{N} \right)^2 \frac{p_i(1-p_i)}{n_i} \frac{N_i - n_i}{N_i - 1}. \qquad (3.5.3)$$

Απόδειξη: Η αμεροληψία είναι προφανής. Για την απόδειξη της (3.5.3), αρκεί να παρατηρηθεί ότι

$$V(\hat{p}) = \sum_{i=1}^{k} \left(\frac{N_i}{N} \right)^2 V(\hat{p}_i),$$

όπου

$$V(\hat{p}_i) = \frac{p_i(1-p_i)}{n_i} \frac{N_i - n_i}{N_i - 1}. \qquad (3.5.4)$$

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

$$S^2_{\hat{p}} = \sum_{i=1}^{k} \left(\frac{N_i}{N} \right)^2 \frac{\hat{p}_i(1-\hat{p}_i)}{n_i - 1} \frac{N_i - n_i}{N_i}. \qquad (3.5.5)$$

Για τον υπολογισμό των n_i και τον καθορισμό του συνολικού δειγματικού μεγέθους n, ισχύουν οι τύποι της προηγούμενης παραγράφου, όπου ο παράγοντας σ_i^2 αντικαθίσταται από την $V(\hat{p}_i)$, όπως αυτή δίνεται από την (3.5.4).

Παρατήρηση: Αν σε μια δειγματοληπτική έρευνα ενδιαφερόμαστε να κάνουμε συγκρίσεις μεταξύ διαφορετικών στρωμάτων, οι κανόνες για τον καταμερισμό του n είναι διαφορετικοί. Αν, για παράδειγμα, ο πληθυσμός αποτελείται από δυο μόνο στρώματα και θέλουμε να συγκρίνουμε τις μέσες τιμές μ_1, μ_2 των στρωμάτων αυτών, τότε μια λογική επιλογή θα ήταν να διαλέξουμε τα n_1 και n_2 έτσι ώστε να

ελαχιστοποιείται η $V\left(\overline{X}_{n_1} - \overline{X}_{n_2}\right)$. Τότε, με την υπόθεση μιας γραμμικής συνάρτησης κόστους $c=c_0+c_1n_1+c_2n_2$, οι τιμές των n_1 και n_2, που ελαχιστοποιούν την $V\left(\overline{X}_{n_1} - \overline{X}_{n_2}\right)$ για σταθερό c, είναι

$$n_i = \frac{n\sigma_i/\sqrt{c_i}}{\sigma_1/\sqrt{c_1} + \sigma_2/\sqrt{c_2}} \quad , \quad i = 1,2 .$$

(Για τον καθορισμό των n_i αγνοήθηκε η διόρθωση πεπερασμένου πληθυσμού).

Όταν ο αριθμός k των στρωμάτων είναι μεγαλύτερος του 2, ο βέλτιστος καταμερισμός εξαρτάται από την ακρίβεια που απαιτεί ο ερευνητής για τις διάφορες συγκρίσεις.

ΑΣΚΗΣΕΙΣ

1. Ένας πληθυσμός μεγέθους N=6 αποτελείται από τα στρώματα {0,1,2} και {4,6,11}. Ένα δείγμα μεγέθους n=4 πρόκειται να επιλεγεί.

(α) Να δειχθεί ότι ο βέλτιστος καταμερισμός του n οδηγεί σε υποδείγματα μεγέθους $n_1=1$, $n_2=3$.

(β) Να εκτιμηθεί η μέση τιμή του πληθυσμού για κάθε δυνατό στρωματοποιημένο δείγμα μεγέθους 4 και να εκτιμηθεί το τυπικό σφάλμα των εκτιμητριών.

2. Ο παρακάτω πίνακας δείχνει τους αριθμούς των κατοίκων 64 Αμερικανικών πόλεων το 1930 που ήταν 5ες έως 68ες κατά σειρά μεγέθους στις ΗΠΑ το 1920. Οι πόλεις είναι χωρισμένες σε δυο στρώματα 16 και 48 πόλεων. Ο συνολικός αριθμός των κατοίκων των 64 πόλεων κατά το έτος 1930 πρόκειται να εκτιμηθεί με ένα τυχαίο δείγμα μεγέθους 24. Να υπολογισθεί το τυπικό σφάλμα της εκτίμησης στις εξής περιπτώσεις:

(1) Απλή τυχαία δειγματοληψία,

(2) στρωματοποιημένη αναλογική δειγματοληψία και
(3) στρωματοποιημένη δειγματοληψία με 12 μονάδες από κάθε στρώμα.

Στρώμα	Πληθυσμός							
1	900	822	781	805	670	1238	573	634
	578	487	442	451	459	464	400	366
2	364	317	328	302	288	291	253	291
	308	272	284	255	270	214	195	260
	209	183	163	253	232	260	201	147
	292	164	143	169	139	170	150	143
	113	115	123	154	140	119	130	127
	100	107	114	111	163	116	122	134

3. Για να εκτιμηθεί το συνολικό απόθεμα ελαστικών που έχουν οι πελάτες μιας εταιρείας, η εταιρεία ελαστικών κάνει μια στρωματοποιημένη δειγματοληψία από τον πληθυσμό των πελατών της. Η στρωματοποίηση γίνεται σύμφωνα με το απόθεμα που οι πελάτες είχαν την προηγούμενη χρονιά. Ένα δείγμα μεγέθους n=1000 έδωσε τα αποτελέσματα που δίνονται στην συνέχεια.

Στρώμα	Πληθυσμός			Δείγμα (αναλογικός καταμερισμός)		
	Αριθμός πελατών N_i	Ni/N		n_i	Μέσο Απόθεμα $\overline{X}_{n_i}$	$\hat{\sigma}_i^2$
1	4000	0.20		200	105	1600
2	10000	0.50		500	180	2500
3	5400	0.27		270	270	2500
4	600	0.03		30	390	5600

(α) Να εκτιμηθεί το μέσο απόθεμα και, επομένως, το συνολικό απόθεμα.
(β) Να εκτιμηθούν τα n_i, αν γίνει βέλτιστος καταμερισμός του n.

(γ) Να εκτιμηθεί η διασπορά της εκτιμήτριας του μέσου αποθέματος στις περιπτώσεις (i) αναλογικού καταμερισμού, (ii) βέλτιστου καταμερισμού και (iii) απλού τυχαίου δείγματος.
(δ) Να κατασκευασθούν 95% διαστήματα εμπιστοσύνης για τις τρεις περιπτώσεις της προηγούμενης ερώτησης.

4. Ένας ερευνητής προτείνει να χρησιμοποιήσει στρωματοποίηση ενός πληθυσμού σε δυο στρώματα και για αυτό προτείνει την επιλογή ενός στρωματοποιημένου τυχαίου δείγματος. Ο ερευνητής περιμένει ότι το κόστος του θα είναι της μορφής $c_0 + \sum_{i=1}^{2} c_i n_i$. Οι εκ των προτέρων εκτιμήσεις του για τα δυο στρώματα είναι οι εξής:

Στρώμα i	N_i/N	σ_i	c_i
1	0.4	10	4
2	0.6	20	9

(α) Να υπολογισθούν οι τιμές των n_i/n που ελαχιστοποιούν το συνολικό κόστος για μια δοθείσα τιμή της $V(\hat{\mu}_n)$.

(β) Αγνοώντας την διόρθωση πεπερασμένου πληθυσμού, να υπολογισθεί πόσο μεγάλο πρέπει να είναι το μέγεθος του στρωματοποιημένου δείγματος, ώστε $V(\hat{\mu}_n) = 1$.

5. Σε ένα εργοστάσιο, 62% των εργαζομένων είναι ειδικευμένοι ή ανειδίκευτοι εργάτες, 31% ειδικευμένες ή ανειδίκευτες εργάτριες και 7% επόπτες. Η διοίκηση του εργοστασίου έχει εξασφαλίσει για τους εργαζομένους την δυνατότητα χρήσης ενός γειτονικού αθλητικού κέντρου και επιθυμεί να εκτιμήσει το ποσοστό των εργαζομένων, που κάνουν χρήση του κέντρου αυτού με βάση ένα δείγμα 400 εργαζομένων. Πρόχειρες εκτιμήσεις δείχνουν ότι τα ποσοστά είναι 40-45%, 20-25% και 1-7.5% για τις τρεις παραπάνω κατηγορίες.

(α) Πώς θα καταμερισθεί το δείγμα;

(β) Αν οι πραγματικές τιμές των ποσοστών των εργαζομένων που κάνουν χρήση του κέντρου είναι 48%, 21% και 4% αντίστοιχα, ποιο θα είναι το τυπικό σφάλμα της εκτιμήτριας του p;