2. ΑΠΛΗ ΤΥΧΑΙΑ ΔΕΙΓΜΑΤΟΛΗΨΙΑ (Simple Random Sampling)

Η στοιχειωδέστερη μορφή δειγματοληψίας κατά πιθανότητα είναι η **απλή τυχαία δειγματοληψία**. Το σχήμα αυτό χρησιμοποιείται ευρύτατα, κυρίως λόγω της απλότητάς του από την άποψη της στατιστικής συμπερασματολογίας. Πέρα από την αυτοτελή χρήση του, το σχήμα αυτό χρησιμεύει και ως βάση για συνθετότερα δειγματοληπτικά σχήματα, όπως, για παράδειγμα, η *στρωματοποιημένη απλή τυχαία δειγματοληψία (stratified simple random sampling)* και η *δειγματοληψία κατά ομάδες (cluster sampling)*. Τα σχήματα αυτά μελετώνται στα κεφάλαια που ακολουθούν.

2.1 Απλό Τυχαίο Δείγμα

Βασική προϋπόθεση για τον σχηματισμό ενός δείγματος από ένα πληθυσμό είναι ο σαφής καθορισμός του πληθυσμού.

Ας υποθέσουμε ότι έχουμε ένα σαφώς καθορισμένο πληθυσμό μεγέθους 5, έστω

$$\{1, 2, 3, 4, 5\}$$

Τα δυνατά διακεκριμένα δείγματα μεγέθους 2 που μπορούμε να σχηματίσουμε από αυτόν τον πληθυσμό είναι τα εξής 10:

$$\{1,2\},\ \{1,3\},\ \{1,4\},\ \{1,5\},\ \{2,3\},$$
$$\{2,4\},\ \{2,5\},\ \{3,4\},\ \{3,5\},\ \{4,5\}.$$

Γενικότερα, αν ο πληθυσμός αποτελείται από N μονάδες και επιθυμούμε δείγμα μεγέθους n, το πλήθος των δυνατών διακεκριμένων δειγμάτων είναι:

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} = \frac{N(n-1)...(N-n+1)}{n},$$

όπου $n! = 1 \times 2 \times 3 \times \ldots \times n$.
(Βλέπε π.χ. Ε. Ξεκαλάκη και Ι. Πανάρετου: *Πιθανότητες και Στοιχεία Στοχαστικών Ανελίξεων*, Αθήνα 1993).

Η διαδικασία επιλογής ενός δείγματος από τα $\binom{N}{n}$ δυνατά δείγματα ονομάζεται **απλή τυχαία δειγματοληψία**, αν κάθε ένα από αυτά έχει πιθανότητα ίση με $1/\binom{N}{n}$ να επιλεγεί.

Στην πράξη δεν είναι πάντα εύκολος ο σχηματισμός όλων των δυνατών διακεκριμένων δειγμάτων, ιδίως όταν το μέγεθος N του πληθυσμού είναι πολύ μεγάλο. Στην θέση της διαδικασίας που περιγράφηκε παραπάνω ακολουθείται η εξής εναλλακτική διαδικασία:

Μια μονάδα του πληθυσμού επιλέγεται τυχαία, δηλαδή με τρόπο που εξασφαλίζει την ίδια πιθανότητα επιλογής σε κάθε μια από τις N μονάδες του πληθυσμού. Η μέθοδος που χρησιμοποιείται για τον σκοπό αυτό αποτελείται από τα εξής βήματα: (α) Αντιστοιχίζουμε σε κάθε μονάδα του πληθυσμού έναν αριθμό από το 1 μέχρι το N και (β) διαλέγουμε μια σειρά n τυχαίων αριθμών από το 1 μέχρι το N με την βοήθεια πινάκων τυχαίων αριθμών. Εύκολα μπορεί να διαπιστωθεί ότι κάθε ένα από τα δυνατά $\binom{N}{n}$ διακεκριμένα τυχαία δείγματα έχει πιθανότητα $1/\binom{N}{n}$ να επιλεγεί. Πράγματι, ας θεωρήσουμε ένα τέτοιο δείγμα, δηλαδή ένα τέτοιο σύνολο n διακεκριμένων μονάδων. Κατά την πρώτη δοκιμή, η πιθανότητα ότι κάποια από τις n συγκεκριμένες μονάδες του δείγματος θα επιλεγεί από τον δοθέντα πληθυσμό είναι $\frac{n}{N}$. Στην δεύτερη δοκιμή η πιθανότητα ότι κάποια από τις n-1 απομένουσες μονάδες του δείγματος θα επιλεγεί από τις N-1 απομένουσες μονάδες του

πληθυσμού είναι $\frac{n-1}{N-1}$ κ.ο.κ. Επομένως, η πιθανότητα με την οποία και οι n μονάδες του επιθυμούμενου δείγματος θα επιλεγούν είναι

$$\frac{n}{N} \times \frac{n-1}{N-1} \times \frac{n-2}{N-2} \times \ldots \times \frac{1}{N-n+1} = \frac{n!(N-n)!}{N!} = \frac{1}{\binom{N}{n}}.$$

Παρατήρηση 1: Στο παραπάνω δειγματοληπτικό σχέδιο, κάθε μονάδα απομακρύνεται από τον πληθυσμό μετά την επιλογή της στο τυχαίο δείγμα. (Δεν επανατοποθετείται). Για τον λόγο αυτό η δειγματοληψία αυτή ονομάζεται και **απλή τυχαία δειγματοληψία χωρίς επανάθεση**. Στα επόμενα, εκτός και αν υποδειχθεί διαφορετικά, με τον όρο απλή τυχαία δειγματοληψία εννοούμε απλή τυχαία δειγματοληψία χωρίς επανάθεση.

Παρατήρηση 2: Μπορεί κανείς εύκολα να δει ότι κατά την απλή τυχαία δειγματοληψία χωρίς επανάθεση, η πιθανότητα με την οποία η i μονάδα του πληθυσμού επιλέγεται στην j δοκιμή είναι $\frac{1}{N}$. Πράγματι, ας θεωρήσουμε την παρακάτω σχηματική παράσταση.

					i μονάδα ↓			
Δυνατοί τρόποι επιλογής μιας μονάδας	N-1	N-2	…	N-j+1	1	N-j	…	N-n+1
Δοκιμή	1	2	…	j-1	j	j+1	…	N
Διαθέσιμες μονάδες του πληθυσμού	N	N-1	…	N-j+2	N-j+1	N-j	…	N-n+1

Τότε

P(να επιλεγεί η i μονάδα την j δοκιμή) =

$$\frac{(N-1)(N-2)...(N-j+1).1.(N-j)...(N-n+1)}{N(N-1)(N-2)...(N-j+2)(N-j+1)...(N-n+1)} = \frac{1}{N}.$$

Παρατήρηση 3: Είναι δυνατόν μετά από κάθε δοκιμή η εκάστοτε επιλεγόμενη μονάδα να επανατοποθετείται στον πληθυσμό. Για παράδειγμα, αν N = 5 και n = 2, ένα τέτοιο δειγματοληπτικό σχήμα επιτρέπει τον σχηματισμό δειγμάτων της μορφής {1,1}, {2,2}, {3,3}, {4,4} και {5,5}. Στην περίπτωση αυτή, η δειγματοληψία ονομάζεται **απλή τυχαία δειγματοληψία με επανάθεση**.

Ο υπολογισμός της διασποράς και των εκτιμητριών της διασποράς είναι ευχερέστερος όταν η δειγματοληψία είναι με επανάθεση. Για τον λόγο αυτό, η δειγματοληψία με επανάθεση χρησιμοποιείται μερικές φορές στα πιο σύνθετα δειγματοληπτικά σχήματα, αν και εκ πρώτης όψεως δεν φαίνεται και τόσο πρακτικό να επιτρέπουμε τη δυνατότητα επιλογής της ίδιας μονάδας του πληθυσμού δύο ή περισσότερες φορές στο δείγμα.

2.2 Επιλογή Απλού Τυχαίου Δείγματος

Η επιλογή ενός απλού τυχαίου δείγματος γίνεται συνήθως με την βοήθεια πινάκων τυχαίων αριθμών. Αυτοί είναι πίνακες των ψηφίων 0,1,2,…,9 στους οποίους η πιθανότητα επιλογής σε οποιαδήποτε δοκιμή είναι η ίδια (1/10) για το κάθε ψηφίο. Ο πίνακας 2.2.1, που δίνεται στην συνέχεια, είναι απόσπασμα 1000 ψηφίων του πίνακα τυχαίων αριθμών των Snedecor και Cochran (1967).

Για την επιλογή ενός απλού τυχαίου δείγματος μεγέθους n από ένα πληθυσμό μεγέθους N, αντιστοιχίζουμε σε κάθε μια από τις μονάδες του πληθυσμού έναν αριθμό από το 1 μέχρι το N (διαφορετικό για κάθε μονάδα). Διαλέγουμε τυχαία τόσες στήλες όσα τα ψηφία του N και διαβάζουμε προς μια κατεύθυνση, π.χ. προς τα κάτω την συγκεκριμένη ομάδα στηλών επιλέγοντας τους αριθμούς που είναι ≤N. Για παράδειγμα, έστω N=198 και n=5. Έστω ότι διαλέγουμε τις στήλες 10-12.

Πίνακας 2.2.1
Τυχαίοι αριθμοί

	00-04	05-09	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49
00	54463	22662	65905	70639	79365	67382	29085	69831	47058	08186
01	15389	85205	18850	39226	42249	90669	96325	23248	60933	26927
02	85941	40756	82414	02015	13858	78030	16269	65978	01385	15345
03	61149	69440	11286	88218	58925	03638	52862	62733	33451	77455
04	05219	81619	10651	67079	92511	59888	84502	72095	83463	75577
05	41417	98326	87719	92294	46614	50948	64886	20002	97365	30976
06	28357	94070	20652	35774	16249	75019	21145	05217	47286	76305
07	17783	00015	10806	83091	91530	36466	39981	62481	49177	75779
08	40950	84820	29881	85966	62800	70326	84740	62660	77379	90279
09	82995	64157	66164	41180	10089	41757	78258	96488	88629	37231
10	96754	17676	55659	44105	47361	34833	86679	23930	53249	27083
11	34357	88040	53364	71726	45690	66334	60332	22554	90600	71113
12	06318	37403	49927	57715	50423	67372	63116	48888	21505	80182
13	62111	52820	07243	79931	89292	84767	85693	73947	22278	11551
14	47534	09243	67879	00544	23410	12740	02540	54440	32949	13491
15	98614	75993	84460	62846	59844	14922	48730	73443	48167	34770
16	24856	03648	44898	09351	98795	18644	39765	71058	90368	44104
17	96887	12479	80621	66223	86085	78285	02432	53342	42846	94771
18	90801	21472	42815	77408	37390	76766	52615	32141	30268	18106
19	55165	77312	83666	36028	28420	70219	81369	41943	47366	41067

Ξεκινώντας από την γραμμή 00 και διαβάζοντας π.χ. προς τα κάτω, οι πρώτοι 5 διακεκριμμένοι αριθμοί είναι 188, 112, 106, 108, 72. Το μειονέκτημα της μεθόδου αυτής είναι ότι οι τριψήφιοι αριθμοί 199 έως 999 δεν χρησιμοποιούνται (απορρίπτονται).

Μια άλλη μέθοδος που απαιτεί λιγότερες απορρίψεις αριθμών είναι η εξής: Αφαιρούμε 200 από τους τριψήφιους αριθμούς μεταξύ 201 και 400, 400 από τους τριψήφιους μεταξύ 401 και 600, 600 από τους τριψήφιους μεταξύ 601 και 800 και 800 από τους τριψήφιους μεταξύ 801 και 999. Φυσικά αφαιρούμε 000 από κάθε τριψήφιο μεταξύ 000 και 200. Στο δείγμα περιλαμβάνουμε τις παραπάνω διαφορές, ενώ απορρίπτουμε τους αριθμούς 000, 199, 200 και όλους τους αριθμούς που μετά την αφαίρεση είναι μεγαλύτεροι του 198. Για παράδειγμα, αν ξεκινήσουμε από την γραμμή 05 του πίνακα και διαβάζουμε προς τα κάτω τους τριψήφιους αριθμούς των στηλών 15-17, επιλέγουμε τους εξής διακεκριμένους αριθμούς:
122, 157, 30, 59, 11.

Η επιλογή των στοιχειωδών παρατηρήσεων του δείγματος μπορεί να επιτευχθεί, προφανώς με την εφαρμογή οποιουδήποτε κατάλληλου τυχαίου μηχανισμού επιλογής στις μονάδες του πληθυσμού. Η χρήση ενός πίνακα τυχαίων αριθμών αποτελεί μια τέτοια περίπτωση. Σήμερα, βέβαια, που η πρόσβαση στους υπολογιστές είναι ευκολότερη και η χρήση των στατιστικών πακέτων ευρύτατη, μπορεί να χρησιμοποιηθεί η γεννήτρια τυχαίων αριθμών ενός υπολογιστή, ή ενός στατιστικού πακέτου, για την παραγωγή των τυχαίων αριθμών που θα αποτελέσουν τους δείκτες των μονάδων του πληθυσμού που θα περιληφθούν στο δείγμα.

2.3 Ορισμοί και Συμβολισμοί

Στις δειγματοληπτικές έρευνες κυρίως μας ενδιαφέρει η μέτρηση (εκτίμηση) ορισμένων παραμέτρων ή χαρακτηριστικών ενός πληθυσμού.

Το ενδιαφέρον συγκεντρώνεται συνήθως στα εξής χαρακτηριστικά:

(α) Μέση τιμή του πληθυσμού (μ)
(β) Συνολικό μέγεθος κάποιου χαρακτηριστικού (π.χ. συνολικό εισόδημα των κατοίκων μιας περιοχής).
(γ) Λόγος δυο συνολικών μεγεθών ή δυο μέσων τιμών.
(δ) Ποσοστό μονάδων ενός πληθυσμού που ανήκουν σε μια ορισμένη κατηγορία.

Έστω πληθυσμός μεγέθους N. Οι τιμές των μονάδων του πληθυσμού, όσο αφορά κάποιο υπό μελέτη χαρακτηριστικό, θα συμβολίζονται με μικρά γράμματα π.χ. $y_1, y_2,\ldots,y_N$ ενώ οι αντίστοιχες παρατηρήσεις ενός δείγματος μεγέθους n με κεφαλαία γράμματα, π.χ. $X_1, X_2, \ldots, X_n$. (Τα κεφαλαία γράμματα αναφέρονται σε τυχαίες μεταβλητές, ενώ τα μικρά σε συγκεκριμένες σταθερές τιμές). Επίσης $\hat{\theta}$ παριστάνει μια εκτιμήτρια της παραμέτρου θ ενός πληθυσμού.

Ο πίνακας 2.3.1 συνοψίζει τον συμβολισμό που θα ακολουθηθεί στα επόμενα.

Δύο από τις επιθυμητές ιδιότητες μιας εκτιμήτριας $\hat{\theta}_n$ της παραμέτρου θ ενός πληθυσμού είναι η **συνέπεια** και η **αμεροληψία**.

Ορισμός 2.3.1: Μια εκτιμήτρια $\hat{\theta}_n$ της παραμέτρου θ ενός πληθυσμού λέγεται **συνεπής** αν η πιθανότητα να απέχει από την θ περισσότερο από οποιαδήποτε δοθείσα ποσότητα τείνει στο μηδέν, καθώς το δείγμα μεγαλώνει, δηλαδή αν, για οποιαδήποτε τιμή ε>0,

$\lim_{n\to N} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0$ (βλέπε Ι. Πανάρετου και Ε. Ξεκαλάκη: *Εισαγωγή στην Στατιστική Σκέψη, Τόμος ΙΙ (Εισαγωγή στις Πιθανότητες και στην Στατιστική Συμπερασματολογία)*, Αθήνα 2000).

Ορισμός 2.3.2: Μια εκτιμήτρια $\hat{\theta}_n$ λέγεται **αμερόληπτη** εκτιμήτρια της παραμέτρου θ ενός πληθυσμού τότε και μόνο τότε αν $E(\hat{\theta}_n) = \theta$.

(Δηλαδή τότε και μόνο τότε αν ο μέσος των τιμών της $\hat{\theta}_n$ για όλα τα $\binom{N}{n}$ δυνατά τυχαία δείγματα μεγέθους n από τον δοθέντα πληθυσμό, είναι ίσος με θ).

Πίνακας 2.3.1

Πληθυσμός		Δείγμα
μέση τιμή	$\mu = \frac{\sum_{i=1}^{N} y_i}{N}$	$\hat{\mu} = \bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$ (μέσος)
συνολικό μέγεθος	$y = \sum_{i=1}^{N} y_i$	$\hat{Y} = N\bar{X}_n = N\frac{\sum_{i=1}^{n} X_i}{n}$
λόγος δύο συνολικών μεγεθών	$R = \frac{\sum_{i=1}^{N} y_i^{(1)}}{\sum_{i=1}^{N} y_i^{(2)}} = \frac{\mu_1}{\mu_2}$	$\hat{R} = \frac{\sum_{i=1}^{n} X_i^{(1)}}{\sum_{i=1}^{n} X_i^{(2)}} = \frac{\bar{X}_n^{(1)}}{\bar{X}_n^{(2)}}$
αριθμός μονάδων σε μια κατηγορία	A	X (αριθμός μονάδων του δείγματος στην κατηγορία)
ποσοστό μονάδων σε μια κατηγορία	$p = \frac{A}{N}$	$\hat{p} = \frac{X}{n}$

Είναι προφανές, ότι ο μέσος $\bar{X}_n$ ενός δείγματος από κάποιο πληθυσμό με μέση τιμή μ είναι συνεπής εκτιμήτρια του μ όπως και η $\hat{Y} = N\bar{X}_n$ είναι συνεπής εκτιμήτρια του συνολικού μεγέθους y.

Το θεώρημα που ακολουθεί αποδεικνύει ότι ο μέσος $\bar{X}_n$ αποτελεί αμερόληπτη εκτιμήτρια της μέσης μ.

Θεώρημα 2.3.1: Ο μέσος $\bar{X}_n$ ενός απλού τυχαίου δείγματος μεγέθους n από ένα πληθυσμό μεγέθους N είναι αμερόληπτη εκτιμήτρια της μέσης τιμής μ του πληθυσμού.

Απόδειξη: Αρκεί να δειχθεί ότι $E(\bar{X}_n) = \mu$.
Ισχύει ότι

$$E(\bar{X}_n) = \frac{\sum \bar{X}_n}{\binom{N}{n}} = \frac{\sum (X_1 + X_2 + ... + X_n)}{n\binom{N}{n}}, \qquad (2.3.1)$$

όπου το άθροισμα στον αριθμητή του κλάσματος εκτείνεται σε όλα τα δυνατά $\binom{N}{n}$ δείγματα. Για τον υπολογισμό του αθροίσματος αυτού, αρκεί να υπολογισθεί σε πόσα δείγματα ανήκει μια τύχουσα τιμή y_i του πληθυσμού: Επειδή υπάρχουν N-1 άλλες μονάδες διαθέσιμες στον πληθυσμό (διαφορετικές της i) και n-1 διαθέσιμες θέσεις στο δείγμα, ο απαιτούμενος αριθμός των δειγμάτων που περιέχουν την τιμή y_i της i μονάδας του πληθυσμού είναι ίσος με τον αριθμό των τρόπων που οι n-1 θέσεις του δείγματος μπορούν να καλυφθούν από τις υπόλοιπες N-1 μονάδες του πληθυσμού. Είναι δηλαδή ίσος με $\binom{N-1}{n-1}$. Επομένως, η τιμή κάθε μονάδας του πληθυσμού υπολογίζεται $\binom{N-1}{n-1}$ φορές στον αριθμητή του κλάσματος και, κατά συνέπεια,

$$\sum(X_1 + X_2 + ... + X_n) = \binom{N-1}{n-1}(y_1 + y_2 + ... + y_N). \qquad (2.3.2)$$

Άρα

$$E(\bar{X}_n) = \frac{\binom{N-1}{n-1}}{n\binom{N}{n}}(y_1 + y_2 + ... + y_N) = \frac{y_1 + y_2 + ... + y_N}{N} = \mu.$$

Παρατήρηση: Η παραπάνω απόδειξη μπορεί να θεωρηθεί ότι στηρίζεται στην εξής απλή συλλογιστική: Επειδή κάθε μονάδα του πληθυσμού περιέχεται στον ίδιο αριθμό δειγμάτων, η $E(X_1 + \ldots + X_n)$ πρέπει να είναι υποπολλαπλάσιο του $y_1 + y_2 + \ldots + y_N$. Επειδή η πρώτη ποσότητα έχει n όρους και η δεύτερη N όρους, έπεται ότι ο πολλαπλασιαστής έχει την τιμή $\frac{n}{N}$.

Θεώρημα 2.3.2: Η διασπορά του μέσου $\bar{X}_n$ ενός απλού τυχαίου δείγματος μεγέθους n από ένα πληθυσμό μεγέθους N είναι

$$V(\bar{X}_n) = \frac{\sigma^2}{n} \frac{N-n}{N}.$$

Απόδειξη: Ισχύει ότι

$$V(\bar{X}_n) = E(\bar{X}_n - \mu)^2.$$

Αλλά,

$$n(\bar{X}_n - \mu) = (X_1 - \mu) + (X_2 - \mu) + \ldots + (\bar{X}_n - \mu)$$

και, επομένως,

$$n^2(\bar{X}_n - \mu)^2 = (X_1 - \mu)^2 + (X_2 - \mu)^2 + \ldots + (X_n - \mu)^2 + 2\{(X_1 - \mu)(X_2 - \mu) +$$

$$(X_1-\mu)(X_3-\mu)+\ldots+(X_{n-1}-\mu)(X_n-\mu)\Big\}.$$

Άρα

$$n^2E(\bar{X}_n-\mu)^2=E\left(\sum_{i=1}^{n}(X_i-\mu)^2\right)+2\ E\left(\sum_{i<j}^{n}(X_i-\mu)(X_j-\mu)\right). \qquad (2.3.3)$$

Επιχειρηματολογία όμοια με αυτή της προηγούμενης παρατήρησης μας οδηγεί στο συμπέρασμα ότι

$$E(\sum_{i=1}^{n}(X_i-\mu)^2)=\frac{n}{N}\sum_{i=1}^{N}(y_i-\mu)^2,$$

$$E(\sum_{1\le i<j\le n}(X_i-\mu)(X_j-\mu))=\frac{n(n-1)}{N(N-1)}\sum_{1\le i<j\le N}(y_i-\mu)(y_j-\mu).$$

(Η τελευταία σχέση στηρίζεται στο ότι το αριστερό μέλος της έχει $\binom{n}{2}$ όρους, ενώ το δεξί μέλος της έχει $\binom{N}{2}$ όρους).

Τότε η (2.3.3) γίνεται

$$n^2E\left(\bar{X}_n-\mu\right)^2=\frac{n}{N}\ \left\{\sum_{i=1}^{N}(y_i-\mu)^2+2\ \frac{n-1}{N-1}\sum_{i<j}^{N}(y_i-\mu)(y_j-\mu)\right\}.$$

Προσθέτοντας και αφαιρώντας $\frac{n-1}{N-1}\sum_{i=1}^{N}(y_i-\mu)^2$ στο δεξί μέλος της παραπάνω ισότητας, έχουμε

$$n^2 E\left(\bar{X}_n - \mu\right)^2 = \frac{n}{N}\left\{\left(1-\frac{n-1}{N-1}\right)\sum_{i=1}^{N}(y_i-\mu)^2 + \frac{n-1}{N-1}\left\{\sum_{i=1}^{N}(y_i-\mu)^2\right\}^2\right\}$$

$$= (\text{επειδή ο } \beta' \text{ προσθετέος είναι } 0)$$

$$= \frac{n(N-n)}{N(N-1)}\sum_{i=1}^{N}(y_i-\mu)^2 .$$

Επομένως,

$$E(\bar{X}_n - \mu)^2 = \frac{\sigma^2}{n}\,\frac{N-n}{n}.$$

Πόρισμα 1: Το τυπικό σφάλμα του μέσου $\bar{X}_n$ είναι

$$\sigma_{\bar{X}_n} = \sigma\sqrt{\frac{N-n}{n\,N}} .$$

Παρατήρηση 1: Ο λόγος $\frac{n}{N}$ αντιπροσωπεύει την αναλογία δείγματος – πληθυσμού και συνήθως συμβολίζεται με f. Η εισαγωγή του συμβολισμού αυτού στους παραπάνω τύπους διευκολύνει την απομνημόνευσή τους:

$$V(\bar{X}_n) = \frac{\sigma^2}{n}(1-f),\ \ \sigma_{\bar{X}_n} = \frac{\sigma}{\sqrt{n}}\sqrt{1-f} .$$

Παρατήρηση 2: Είναι προφανές ότι μια αμερόληπτη εκτιμήτρια της τιμής $y = \sum_{i=1}^{N} y_i$ είναι η στατιστική συνάρτηση $\hat{Y} = N\bar{X}_n$ με διασπορά $V(\hat{Y}) = \frac{N^2\sigma^2}{n}(1-f)$ και τυπικό σφάλμα $\sigma_{\hat{Y}} = \frac{N\sigma}{\sqrt{n}}\sqrt{1-f}$.

Παρατήρηση 3: Είναι γνωστό ότι η διασπορά του μέσου ενός απλού τυχαίου δείγματος μεγέθους n από ένα άπειρο πληθυσμό είναι $\frac{\sigma^2}{n}$. Η διαφορά που υπάρχει στην περίπτωση του πεπερασμένου πληθυσμού μεγέθους N είναι ο παράγοντας 1-f. Ο παράγοντας αυτός ονομάζεται **διόρθωση πεπερασμένου πληθυσμού** και ο παρονομαστής του είναι $N-1$, αν τα αποτελέσματα αναφέρονται στην έκφραση $\frac{1}{N}\sum_{i=1}^{N}(y_i-\mu)^2$ για την διασπορά σ^2 του πληθυσμού. Προφανώς, όταν ο λόγος $f = \frac{n}{N}$ είναι μικρός, η διόρθωση 1-f είναι κοντά στο 1 και το μέγεθος N του πληθυσμού δεν επηρεάζει την διασπορά της εκτιμήτριας $\bar{X}_n$ του μ. Για παράδειγμα, αν δυο πληθυσμοί μεγέθους $N_1 = 200000$ και $N_2 = 20000$ αντίστοιχα έχουν την ίδια διασπορά σ^2, ένα απλό τυχαίο δείγμα μεγέθους 500 από κάθε ένα από τους δυο πληθυσμούς θα οδηγήσει σε εκτίμηση της μέσης τιμής του κάθε ενός με το ίδιο περίπου τυπικό σφάλμα.

Στην πράξη, η διόρθωση 1-f αγνοείται όταν $f \le 0.05$ και πολλές φορές ακόμη και αν $f = 0.10$. Το αποτέλεσμα είναι να υπερεκτιμάται το τυπικό σφάλμα της εκτιμήτριας $\bar{X}_n$.

Παρατήρηση 4: Η συναγωγή συμπερασμάτων και ο υπολογισμός διασπορών και άλλων παραμέτρων ενός πληθυσμού γίνονται ευκολότερα αν ο πληθυσμός είναι άπειρος. Πολλές φορές, είναι δυνατή η

χρησιμοποίηση ″μεθόδων απείρου πληθυσμού″ ακόμη και σε περιπτώσεις πεπερασμένου πληθυσμού. Για παράδειγμα, η απόδειξη των δυο παραπάνω θεωρημάτων γίνεται απλούστερη με την χρήση της εξής θεωρίας.

Έστω $U_i = \begin{cases} 1 & \text{αν } y_i \text{ ανήκει στο δείγμα} \\ 0 & \text{διαφορετικά} \end{cases}$

Η τυχαία μεταβλητή U_i ακολουθεί την διωνυμική κατανομή με παράμετρο $p = \frac{n}{N}$. Δηλαδή

$$P(U_i = 1) = \frac{n}{N} = 1\text{-}P(U_i = 0),$$

$$P(U_i = 1,\ U_j = 1) = P(U_i = 1)P(U_j = 1 \mid U_i = 1) = \frac{n}{N}\frac{n-1}{N-1}.$$

Ισχύει προφανώς ότι

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{N} U_i y_i .$$

Τότε

$$E(\bar{X}_n) = \frac{1}{n}\sum_{i=1}^{N} E(U_i) y_i$$

και

$$V\left(\bar{X}_n\right) = \frac{1}{n^2}\sum_{i=1}^{N} y_i^2 V\left(U_i\right) + 2\sum_{i<j} y_i y_j \operatorname{Cov}(U_i, U_j) .$$

Αλλά

$$E(U_i) = 1 \cdot P(U_i = 1) + 0 \cdot P(U_i = 0) = \frac{n}{N}\ (= f),$$

$$V(U_i) = p\,q = \frac{n}{N}\left(1 - \frac{n}{N}\right)\ (= f\ (1\text{-}f)),$$

$$\operatorname{Cov}(U_i, U_j) = E(U_i U_j) - E(U_i)\cdot E(U_j)$$

$$= \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2}$$

$$= \frac{n}{N(N-1)}\left(1 - \frac{n}{N}\right) \left(= \frac{n}{N(N-1)} \ (1-f)\right)$$

(αφού $E(U_i\, U_j) = 1 \cdot P(U_i = 1,\ U_j = 1) + 0 \cdot P(U_i = 0,\ U_j = 1) +$

$+ 0 \cdot P(U_i = 1,\ U_j = 0) + 0 \cdot P(U_i = 0,\ U_j = 0) \ = \frac{n}{N} \cdot \frac{n-1}{N-1}$).

Άρα,

$$E(\bar{X}_n) = \frac{1}{n} \cdot \frac{n}{N} \sum_{i=1}^{N} y_i = \mu$$

και

$$V\left(\bar{X}_n\right) = \frac{1}{n^2}\left\{\frac{n}{N}\left(1 - \frac{n}{N}\right) \sum_{i=1}^{N} y_i^2 + 2\,\frac{n}{N(N-1)}\left(1 - \frac{n}{N}\right) \sum_{i<j} y_i y_j\right\}$$

$$= \frac{1-f}{n\ N}\left(\sum y_i^2 - \frac{2}{N-1} \sum_{i<j} y_i y_j\right)$$

(προσθαφαιρώντας το $\frac{1}{N-1}\sum y_i^2$)

$$= \frac{1-f}{n\ N}\left\{\frac{N}{N-1}\sum y_i^2 - \frac{1}{N-1}\left(\sum y_i\right)^2\right\}$$

$$= \frac{1-f}{n(N-1)}\left\{\sum y_i^2 - N\mu^2\right\}$$

$$= \frac{1-f}{n(N-1)} \sum_{i=1}^{N}\left(y_i - \mu\right)^2 = \frac{\sigma^2}{n}(1-f).$$

2.4 Εκτίμηση του Τυπικού Σφάλματος

Η χρησιμότητα του αποτελέσματος του θεωρήματος 2.3.2 για το τυπικό σφάλμα είναι μεγάλη για τους εξής λόγους:
(α) Δίνει την δυνατότητα μέτρησης του βαθμού ακρίβειας της εκτίμησης της μέσης τιμής του πληθυσμού και σύγκρισής του με τον βαθμό ακρίβειας που παρέχει οποιαδήποτε άλλη μέθοδος δειγματοληψίας και
(β) Δίνει την δυνατότητα εκτίμησης του μεγέθους του δείγματος που απαιτείται σε μια δειγματοληπτική έρευνα, ώστε να επιτευχθεί ο επιθυμητός βαθμός ακρίβειας.

Βέβαια, η γνώση του σ^2 είναι απαραίτητη. Στην πράξη όμως αυτό συμβαίνει σπάνια. Για τον λόγο αυτό, απαιτείται μια εκτίμηση του σ^2 από τα δεδομένα του δείγματος και ως τέτοια συνήθως θεωρείται η τιμή της εκτιμήτριας

$$\hat{\sigma}^2 = S^{*2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{n-1}$$

στην περίπτωση απλού τυχαίου δείγματος. Στην συνέχεια, για απλότητα, θα χρησιμοποιείται ο συμβολισμός S^2 για την εκτιμήτρια $\hat{\sigma}^2$ της διασποράς σ^2 που ορίζεται από την παραπάνω σχέση.

Θεώρημα 2.4.1: Στην απλή τυχαία δειγματοληψία χωρίς επανάθεση από πεπερασμένο πληθυσμό μεγέθους N, η στατιστική συνάρτηση

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{n-1}$$

είναι αμερόληπτη εκτιμήτρια της διασποράς σ^2 του πληθυσμού.

Απόδειξη:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[(X_i-\mu)-(\bar{X}_n-\mu)\right]^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}(X_i-\mu)^2 - n(\bar{X}_n-\mu)^2\right].$$

Τότε ο μέσος της διασποράς όλων των δυνατών δειγμάτων μεγέθους n σύμφωνα με την επιχειρηματολογία των θεωρημάτων 2.3.1 και 2.3.2. θα δίνεται από τον τύπο

$$\begin{aligned}
E(S^2) &= \frac{1}{n-1}\left[E(\sum_{i=1}^{N}(X_i-\mu)^2 - E(n(\bar{X}_n-\mu)^2)\right] \\
&= \frac{1}{n-1}\left[\frac{n}{N}\sum_{i=1}^{N}(y_i-\mu)^2 - \frac{N-n}{N}\sigma^2\right] \\
&= \frac{1}{n-1}\left[\frac{n(N-1)}{N}\sigma^2 - \frac{N-n}{N}\sigma^2\right] \\
&= \frac{\sigma^2}{(n-1)N}\left[n(N-1)-(N-n)\right] \\
&= \sigma^2.
\end{aligned}$$

Πόρισμα: Οι στατιστικές συναρτήσεις

$$\hat{\sigma}^2_{\bar{X}_n}\ S^2_{\bar{X}_n} = \frac{S^2}{n}\frac{N-n}{N} = \frac{S^2}{n}(1-f)$$

και

$$\hat{\sigma}^2_{\hat{Y}} \equiv S^2_{\hat{Y}} = \frac{N^2S^2}{n}\frac{N-n}{N} = \frac{N^2S^2}{n}(1-f)$$

είναι αμερόληπτες εκτιμήτριες των διασπορών του μέσου $\bar{X}_n$ και της τυχαίας μεταβλητής $\hat{Y} = N\bar{X}_n$ αντίστοιχα.
(Για την εκτίμηση των τυπικών σφαλμάτων θεωρούμε τις θετικές τετραγωνικές ρίζες των παραπάνω εκφράσεων).

Παράδειγμα: Για κάποιο αίτημα, μαζεύτηκαν υπογραφές που κάλυψαν 676 σελίδες. Κάθε σελίδα είχε χώρο για 42 υπογραφές, αλλά σε πολλές από τις σελίδες υπήρχε αριθμός υπογραφών διαφορετικός του 42. Ένα απλό τυχαίο δείγμα 50 σελίδων επελέγη και ο αριθμός των υπογραφών ανά σελίδα καταγράφηκε. Ο παρακάτω πίνακας συχνότητας συνοψίζει τα αποτελέσματα. (X_i=αριθμός υπογραφών, ν_i=αριθμός σελίδων με X_i υπογραφές).

X_i	42	41	36	32	29	27	23	19	16	15
ν_i	23	4	1	1	1	2	1	1	2	2
X_i	14	11	10	9	7	6	5	4	3	Σύνολο
ν_i	1	1	1	1	1	3	2	1	1	50

Να εκτιμηθεί ο συνολικός αριθμός των υπογραφών που μαζεύτηκαν στις 676 σελίδες και να υπολογισθεί το τυπικό σφάλμα της εκτίμησης αυτής.

Λύση: Έχουμε

$$n = \sum \nu_i = 50$$
$$\sum \nu_i X_i = 1471$$
$$\sum \nu_i X_i^2 = 54497.$$

Άρα $\bar{X}_n = 1471/50 = 29.52$ και επομένως η τιμή της εκτιμήτριας $\hat{Y}$ του συνολικού αριθμού υπογραφών y είναι

$$\hat{Y} = N\bar{X}_n = 19888.$$

Το τυπικό σφάλμα της $\hat{Y}$ είναι ίσο με

$$S_{\hat{Y}} = \frac{NS}{\sqrt{n}}\sqrt{1-f},$$

όπου

$$S = \sqrt{\frac{1}{n-1}\sum \nu_i \left(X_i - \bar{X}_n\right)^2} = \sqrt{\frac{1}{n-1}\left[\sum \nu_i X_i^2 - \frac{\left(\sum \nu_i X_i\right)^2}{n}\right]}$$

$$= \sqrt{229} = 15.13.$$

και $f = \frac{50}{676} = 0.0740,$ δηλαδή

$$S_{\hat{Y}} = \frac{(676)(15.13)}{\sqrt{50}}\sqrt{1-0.0740} = 1391.$$

(Το τυπικό σφάλμα του μέσου του δείγματος είναι

$$S_{\bar{X}_n} = \frac{S}{\sqrt{n}}\sqrt{1-f} = 2.05).$$

2.5 Διαστήματα Εμπιστοσύνης – Η Ισχύς της Κανονικής Προσέγγισης

Έχει αποδειχθεί (βλέπε π.χ. Ε. Ξεκαλάκη και Ι. Πανάρετου: *Πιθανότητες και Στοιχεία Στοχαστικών Ανελίξεων*, Αθήνα 1993) ότι η κανονική κατανομή είναι η οριακή μορφή της κατανομής του μέσου $\bar{X}_n$ ενός τυχαίου δείγματος μεγέθους n, το οποίο προέρχεται από έναν άπειρο πληθυσμό με πεπερασμένη διασπορά, όταν το n τείνει στο ∞. Δηλαδή, αν μ και σ^2 συμβολίζουν την μέση τιμή και την διασπορά του πληθυσμού αντίστοιχα, τότε, όταν το δείγμα είναι αρκετά μεγάλο ($n \to \infty$),

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ ή, ισοδύναμα, } \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Το πόσο μεγάλο πρέπει να είναι το μέγεθος n του δείγματος ώστε η προσέγγιση της πραγματικής κατανομής του $\bar{X}_n$ από την κανονική κατανομή να είναι ικανοποιητική δεν καθορίζεται από κάποιο γενικό κανόνα. Στις περισσότερες εφαρμογές, το n δεν συνηθίζεται να είναι μικρότερο του 25 ($n \geq 25$).

Το παραπάνω αποτέλεσμα είναι γνωστό ως **κεντρικό οριακό θεώρημα.** Η πρακτική αξία του θεωρήματος αυτού είναι μεγάλη εξ αιτίας των δυνατοτήτων που δίνει στον ερευνητή όσο αφορά την συναγωγή στατιστικών συμπερασμάτων. Πράγματι, το γεγονός ότι

$$P\left[\left|\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right| < z_{1-\alpha/2}\right] = 1-\alpha,$$

όπου $z_{1-\alpha/2}$ συμβολίζει το (1-α/2)-ποσοστιαίο σημείο της N(0,1) οδηγεί αμέσως στο συμπέρασμα ότι στο 100(1-α)% των περιπτώσεων η πραγματική τιμή του μ ανήκει στο διάστημα με άκρα

$$\bar{X}_n \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

(Στην περίπτωση που η διασπορά σ^2 του πληθυσμού είναι άγνωστη και χρησιμοποιείται η εκτίμηση της S^2, η κατανομή του λόγου $\frac{\bar{X}_n - \mu}{\sigma\sqrt{n}}$ είναι η t με n-1 βαθμούς ελευθερίας και τα άκρα του 100(1-α)% διαστήματος εμπιστοσύνης είναι ως γνωστό

$$\bar{X}_n \pm t_{n-1,1-\alpha/2} \frac{S}{\sqrt{n}}.$$

Εδώ $t_{n-1,1-\alpha/2}$ συμβολίζει το (1-α/2) – ποσοστιαίο σημείο της κατανομής t με n-1 βαθμούς ελευθερίας.

Για την περίπτωση της δειγματοληψίας χωρίς επανάθεση από ένα πεπερασμένο πληθυσμό μεγέθους N, έχει επίσης αποδειχθεί η ισχύς της κανονικής προσέγγισης της κατανομής του μέσου $\bar{X}_n$ κάτω από ορισμένες συνθήκες. Και πάλι, το ερώτημα που αντιμετωπίζει ο ερευνητής είναι πόσο μεγάλο πρέπει να είναι το μέγεθος του δείγματος ώστε η κανονική προσέγγιση να είναι ικανοποιητική. Για πληθυσμούς με μακριά δεξιά ουρά, ένας πρόχειρος κανόνας είναι ο

$$n > 25G_1^2,$$

όπου η τιμή G_1 είναι γνωστή ως μέτρο ασυμμετρίας του Fisher και ορίζεται από την σχέση

$$G_1 = \frac{1}{N\sigma^3}\sum_{i=1}^{N}(y_i - \mu)^3$$

ή, ισοδύναμα,

$$G_1 = \frac{1}{N\sigma^3}\left\{\sum_{i=1}^{N} y_i^3 - 3\mu\sum_{i=1}^{N} y_i^2 + 2\mu^3\right\}$$

(βλέπε π.χ. Ι. Πανάρετου και Ε. Ξεκαλάκη: *Εισαγωγή στη Στατιστική Σκέψη, Τόμος Ι (Περιγραφική Στατιστική)*, Αθήνα 1993).

Ο κανόνας αυτός στηρίζεται στην υπόθεση ότι οποιεσδήποτε ροπές της κατανομής τάξης μεγαλύτερης του 3 παίζουν αμελητέο ρόλο στον καθορισμό του n. Έτσι, υπολογίζοντας (ή εκτιμώντας) το G_1 ενός συγκεκριμένου πληθυσμού, μπορούμε να έχουμε μια ιδέα όσο αφορά το απαιτούμενο δειγματικό μέγεθος.

Πίνακας 2.5.1

Κατανομή συχνότητας καλλιεργημένων εκτάσεων (σε εκτάρια) 556 αγροικιών

Κλάση	y_i (σε κωδικοποιημένη κλίμακα)	Συχνότητα ν_i	$\nu_i y_i$	$\nu_i y_i^2$	$\nu_i y_i^3$
0-29	-0.9	47	-42.3	38.1	-34.3
30-63	0	143	0	0	0
64-97	1	154	154	154	154
98-131	2	82	164	328	656
132-165	3	62	186	558	1674
166-199	4	33	132	528	2112
200-233	5	13	65	325	1625
234-267	6	6	36	216	1296
268-301	7	4	28	196	1372
302-335	8	6	48	384	3072
336-369	9	2	18	162	1458
370-403	10	0	0	0	0
404-437	11	2	22	242	2662
438-471	12	0	0	0	0
472-505	13	2	26	338	4394
Σύνολα		556	836.7	3469.1	20440.7

Παράδειγμα: Τα (κωδικοποιημένα) δεδομένα του πίνακα 2.5.1 αναφέρονται στην έκταση (σε εκτάρια) που χρησιμοποιήθηκε για καλλιέργεια από κάθε μια από 556 αγροικίες στην πολιτεία της Νέας Υόρκης. Τα δεδομένα αυτά προέρχονται από μια σειρά μελετών που έγιναν το 1951 και στηρίχθηκαν στην επανειλημμένη χρήση τυχαίων δειγμάτων μεγέθους 100 από τον πληθυσμό αυτών των αγροικιών. Οι μελέτες αυτές είχαν σκοπό την συμπερασματολογία για την κατανομή του $\bar{X}_n$ και ορισμένων άλλων χαρακτηριστικών που παρουσιάζουν ενδιαφέρον στις έρευνες διαχείρισης αγροικιών.

Ο υπολογισμός του G_1 δεν επηρεάζεται από το γεγονός ότι τα δεδομένα είναι κωδικοποιημένα εφ' όσον το G_1 είναι καθαρός αριθμός. Έχουμε

$$N = \sum \nu_i = 556,$$

$$\mu = \frac{\sum \nu_i y_i}{N} = \frac{836.7}{556} = 1.50486,$$

$$\frac{\sum \nu_i y_i^2}{N} = \frac{3469.1}{556} = 6.23939,$$

$$\frac{\sum \nu_i y_i^3}{N} = \frac{20440.7}{556} = 36.76385,$$

$$\sigma = \sqrt{\frac{\sum \nu_i y_i^2}{N} - \mu^2} = \sqrt{3.97476} = 1.99,$$

$$\frac{\sum \nu_i (y_i - \mu)^3}{N} = 36.76385 - 3\,(1.50486)\,(6.23939) + 2(1.50486)^2 =$$
$$= 15.411.$$

Άρα $G_1 = 1.9$ και, επομένως, $n \geq 25(1.9)^2 = 90.25 \simeq 91$.
Για δείγματα μεγέθους 100 βρέθηκε ότι η κατανομή του $\bar{X}_n$ δεν διαφέρει στατιστικά σημαντικά από την κανονική κατανομή.

Κάτω από τις παραπάνω προϋποθέσεις έχει λοιπόν αποδειχθεί ότι η κατανομή του μέσου $\bar{X}_n$ ενός απλού τυχαίου δείγματος μεγέθους n που έχει επιλεγεί χωρίς επανάθεση από ένα πεπερασμένο πληθυσμό μεγέθους N με μέση τιμή μ και διασπορά σ^2 είναι κατά προσέγγιση η

$$N(\mu, \frac{\sigma^2}{n}(1-\frac{n}{N})).$$

Επομένως, ισχύει ότι

$$\frac{\bar{X}_n - \mu}{\sigma\sqrt{(1-f)/n}} \sim N(0, 1)$$

και, άρα,

$$P\left[\left|\frac{\bar{X}_n - \mu}{\sigma\sqrt{(N-n)/(nN)}}\right| < z_{1-\alpha/2}\right] = 1-\alpha.$$

Η τελευταία σχέση οδηγεί στο εξής 100(1-α)% διάστημα εμπιστοσύνης για την μέση τιμή μ του πληθυσμού

$$\left(\bar{X}_n - z_{1-\alpha/2}\sqrt{\frac{1-f}{n}}, \qquad \bar{X}_n + z_{1-\alpha/2}\sigma\sqrt{\frac{1-f}{n}}\right).$$

Όταν η διασπορά δεν είναι γνωστή, χρησιμοποιείται μια εκτίμησή της, όπως αυτή δίνεται από την αμερόληπτη εκτιμήτρια

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Τότε, όπως και στην περίπτωση του άπειρου πληθυσμού,

$$\frac{\bar{X}_n - \mu}{S\sqrt{(1-f)/n}} \sim t_{n-1}$$

και, επομένως, τα άκρα του 100(1-α)% διαστήματος εμπιστοσύνης είναι τα

$$\bar{X}_n \pm t_{n-1,1-\alpha/2}\, S\sqrt{\frac{1-f}{n}}\,.$$

Προφανώς, επειδή $\hat{Y} = N\bar{X}_n$, ένα 100(1-α)% διάστημα εμπιστοσύνης για το συνολικό μέγεθος $y = \sum_{i=1}^{N} y_i$ έχει άκρα

$$N(\bar{X}_n \pm t_{n-1,1-\alpha/2}\, S\sqrt{\frac{1-f}{n}}\,).$$

Παράδειγμα: Στο παράδειγμα του δείγματος των υπογραφών, ένα 80% διάστημα εμπιστοσύνης για τον πραγματικό συνολικό αριθμό των υπογραφών που συνελέγησαν έχει άκρα

$$N(\bar{X}_n \pm t_{49,0.90}\, S\sqrt{\frac{1-f}{n}})\,.$$

Από τους πίνακες της κατανομής t, βρίσκουμε ότι $t_{49,\,0.90} = 1.299$. Άρα το 80% διάστημα εμπιστοσύνης είναι το (18107, 21669). (Η ακριβής καταμέτρηση των υπογραφών έδωσε ως αποτέλεσμα y =21045).

2.6 Εκτίμηση ενός Λόγου

Έστω πληθυσμός μεγέθους N και έστω x_i και y_i οι τιμές δύο διαφορετικών χαρακτηριστικών της i μονάδας του πληθυσμού, i = 1,2,…,N. (Για παράδειγμα, x_i μπορεί να συμβολίζει το ενοίκιο που πληρώνει η i οικογένεια και y_i το εισόδημά της, ή x_i μπορεί να αντιπροσωπεύει τις πωλήσεις της i επιχείρησης σε κάποιο μήνα και y_i τις πωλήσεις της σε κάποιο προηγούμενο μήνα). Τότε, $x = \sum_{i=1}^{N} x_i$ και

$y = \sum_{i=1}^{N} y_i$ αντιπροσωπεύουν τις συνολικές τιμές των δυο αυτών χαρακτηριστικών και πολλές φορές στην πράξη μας ενδιαφέρει να εκτιμήσουμε τον λόγο

$$R = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} = \frac{\sum_{i=1}^{N} y_i/N}{\sum_{i=1}^{N} x_i/N} = \frac{\mu_y}{\mu_x}.$$

Δηλαδή, μας ενδιαφέρει να εκτιμήσουμε τον λόγο των συνολικών μεγεθών x και y ή τον λόγο των μέσων τιμών μ_x και μ_y των δύο χαρακτηριστικών.

Η εκτιμήτρια που χρησιμοποιείται για τον σκοπό αυτό είναι η

$$\hat{R} = \frac{\bar{Y}_n}{\bar{X}_n} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i},$$

όπου $\{X_1, X_2, \ldots, X_n\}$, $\{Y_1, Y_2, \ldots, Y_n\}$, είναι οι τιμές των μονάδων ενός απλού τυχαίου δείγματος μεγέθους n για τα δύο υπό εξέταση χαρακτηριστικά και $\bar{X}_n$, $\bar{Y}_n$ οι μέσοι των τιμών αυτών αντίστοιχα.

Για την αξιολόγηση της ακριβείας της εκτιμήτριας αυτή και την εξαγωγή συμπερασμάτων, είναι αναγκαίος ο προσδιορισμός της κατανομής της και του τυπικού της σφάλματος. Για μικρές τιμές του n η κατανομή της $\hat{R}$ δεν είναι κανονική. Αντίθετα, είναι ασύμμετρη προς τα δεξιά (έχει μακριά δεξιά ουρά). Επί πλέον, η $\hat{R}$ δεν είναι αμερόληπτη εκτιμήτρια του R. Για μεγάλες όμως τιμές του n, η κατανομή της $\hat{R}$ τείνει στην κανονική κατανομή και ισχύει το εξής θεώρημα.

Θεώρημα 2.6.1: Αν $\{ X_1, X_2, \ldots, X_n \}$, $\{ Y_1, Y_2, \ldots, Y_n \}$, $\overline{X}_n$, $\overline{Y}_n$, $\{x_1, x_2, \ldots, x_N\}$, $\{y_1, y_2, \ldots, y_N\}$ και μ_x και μ_y ορίζονται όπως προηγουμένως, τότε για αρκετά μεγάλο n ($n \to N$) η στατιστική συνάρτηση

$$\hat{R} = \frac{\overline{Y}_n}{\overline{X}_n}$$

είναι αμερόληπτη εκτιμήτρια του λόγου $R = \mu_y / \mu_x$ και ισχύει ότι

$$V\left(\hat{R}\right) = \frac{1-f}{n\mu_x^2} \frac{\sum_{i=1}^{N} \left(y_i - Rx_i\right)^2}{N-1}. \qquad (2.6.1)$$

Απόδειξη: Ισχύει ότι

$$\hat{R} - R = \frac{\overline{Y}_n}{\overline{X}_n} - R = \frac{\overline{Y}_n - R\overline{X}_n}{\overline{X}_n}.$$

Αλλά ο μέσος ενός απλού τυχαίου δείγματος είναι συνεπής εκτιμήτρια της μέσης τιμής του πληθυσμού. Άρα, για μεγάλες τιμές του n, η τιμή του μέσου $\overline{X}_n$ δεν διαφέρει πολύ από την μ_x και, επομένως, ισχύει κατά προσέγγιση ότι

$$\hat{R} - R \simeq \frac{\overline{Y}_n - R\overline{X}_n}{\mu_x}.$$

Κατά συνέπεια,

$$E\left(\hat{R} - R\right) \simeq \frac{1}{\mu_x} E\left(\overline{Y}_n - R\overline{X}_n\right) = \frac{1}{\mu_x}\left(\mu_y - R\mu_x\right) = R - R = 0.$$

Άρα, με την προσέγγιση που θεωρήθηκε, αποδείχθηκε ότι η $\hat{R}$ είναι αμερόληπτη εκτιμήτρια του R. Επομένως,

$$V(\hat{R}) = E(\hat{R} - R)^2 \simeq \frac{1}{\mu_x^2} E(\bar{Y}_n - R\bar{X}_n)^2 .$$

Αλλά η τυχαία μεταβλητή $\bar{Y}_n - R\bar{X}_n$ είναι ο μέσος του απλού τυχαίου δείγματος $Y_1 - RX_1, Y_2 - RX_2, \ldots, Y_n - RX_n$ που προέρχεται από τον πληθυσμό $\{d_1, d_2, \ldots, d_N\}$, όπου $d_i = y_i - Rx_i$, i =1, 2, …, N με μέση τιμή $d = \mu_y - R\mu_x = 0$. Άρα,

$$E(\bar{Y}_n - R\bar{X}_n)^2 = d = 0$$

και, κατά συνέπεια,

$$\begin{aligned}
\mu_x^2 V(\hat{R}) &= V\left(\bar{Y}_n - R\bar{X}_n\right) \\
&= E\left(\bar{Y}_n - R\bar{X}_n\right)^2 \\
&= \frac{\sigma_d^2}{n}\left(1 - \frac{n}{N}\right) \\
&= \frac{\sum_{i=1}^{N} d_i^2}{N-1} \; \frac{1 - \frac{n}{N}}{n} \\
&= \frac{1 - \frac{n}{N}}{n} \sum_{i=1}^{N} \frac{(y_i - Rx_i)^2}{N-1}, \qquad \text{ο.ε.δ.}
\end{aligned}$$

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

$$S_R^2 = \frac{1-\frac{n}{N}}{n\bar{X}_n^2} \frac{\sum_{i=1}^{n}(Y_i - \hat{R}X_i)^2}{n-1}$$

$$= \frac{1-\frac{n}{N}}{n\bar{X}_n^2} \frac{\sum_{i=1}^{n} Y_i^2 - 2\hat{R}\sum_{i=1}^{n} X_i Y_i + \hat{R}^2 \sum_{i=1}^{n} X_i^2}{n-1}.$$

Πίνακας 2.6.1

Μέγεθος, ετήσιο εισόδημα και ετήσια δαπάνη για φαγητό 33 οικογενειών (σε δεκάδες ευρώ)

Οικογένεια	$X^{(1)}$	$X^{(2)}$	Y	Οικογένεια	$X^{(1)}$	$X^{(2)}$	Y
1	2	62	14.3	18	4	83	36.0
2	3	62	20.8	19	2	85	20.6
3	3	87	22.7	20	4	73	27.7
4	5	65	30.5	21	2	66	25.9
5	4	58	41.2	22	5	58	23.3
6	7	92	28.2	23	3	77	39.8
7	2	88	24.2	24	4	69	16.8
8	4	79	30.0	25	7	65	37.8
9	2	83	24.2	26	3	77	34.8
10	5	62	44.4	27	3	69	28.7
11	3	63	13.4	28	6	95	63.0
12	6	62	19.8	29	2	77	19.5
13	4	60	29.4	30	2	69	21.6
14	4	75	27.1	31	6	69	18.2
15	2	90	22.2	32	4	67	20.1
16	5	75	37.7	33	2	63	20.7
17	3	69	22.6				
				Σύνολο	123	2394	907.2

Παράδειγμα: Από απλό τυχαίο δείγμα 33 οικογενειών που επελέγη από κάποιο πληθυσμό, συνελέγησαν στοιχεία για το μέγεθος ($X^{(1)}$) των οικογενειών, το ετήσιο εισόδημά τους ($X^{(2)}$) και την ετήσια δαπάνη τους για φαγητό (Y). Τα αποτελέσματα συνοψίζονται στον πίνακα 2.6.1.

Να εκτιμηθεί (i) η μέση ετήσια δαπάνη για φαγητό ανά οικογένεια, (ii) η μέση ετήσια δαπάνη για φαγητό ανά άτομο και (iii) το ποσοστό του εισοδήματος που δαπανάται για φαγητό. Να υπολογισθεί το τυπικό σφάλμα των εκτιμήσεων αυτών.

Λύση: Αγνοώντας την διόρθωση πεπερασμένου πληθυσμού, έχουμε

(i) $$\bar{Y}_{33} = \frac{907.2}{33} = 27.49$$

με τυπικό σφάλμα

$$S_{\bar{Y}_{33}} = \frac{1}{\sqrt{33}} \sqrt{\frac{\sum_{i=1}^{33} Y_i^2 - 33\bar{Y}_{33}^2}{32}} = \sqrt{\frac{28224 - (33)(27.49)^2}{(32)\cdot(33)}} = 1.76.$$

(ii) $$\hat{R}_1 = \frac{\sum Y_i}{\sum X_i^{(1)}} = \frac{907.2}{123} = 7.38$$

με τυπικό σφάλμα

$$S_{\hat{R}_1} = \sqrt{\frac{28224 - (14.7512)(3595.5) + (54.3996)(533)}{(33)(32)(3.7273)^2}} = 0.534.$$

(iii) $$\hat{R}_2 = 100\frac{\sum Y_i}{\sum X_i^{(2)}} = \frac{100(907.2)}{2394}\% = 37.9\%$$

με τυπικό σφάλμα

$$S_{\hat{R}_2} = 0.0238 .$$

Παρατήρηση: Προβλήματα, τα οποία απαιτούν την εκτίμηση του λόγου δύο μεταβλητών, συναντώνται πολύ συχνά στην πράξη. Η συνηθέστερη περίπτωση είναι όταν η δειγματοληπτική μονάδα αποτελείται από ένα σύνολο στοιχειωδέστερων μονάδων και το ενδιαφέρον του ερευνητή εστιάζεται στην εκτίμηση της μέσης τιμής του πληθυσμού ανά στοιχειώδη μονάδα. Τυπικό παράδειγμα αποτελεί η περίπτωση (ii) του προηγούμενου παραδείγματος. Λόγοι δύο μεταβλητών εμφανίζονται επίσης σε προβλήματα στα οποία ενδιαφέρει η μελέτη ενός ποσοτικού χαρακτηριστικού ως ποσοστού ενός άλλου ποσοτικού χαρακτηριστικού (περίπτωση (iii) του προηγούμενου παραδείγματος). Τέλος, προβλήματα σύγκρισης δύο χαρακτηριστικών ενός πληθυσμού ανάγονται επίσης σε προβλήματα εκτίμησης του λόγου δύο κατάλληλων μεταβλητών.

2.7 Εκτίμηση Μέσων Τιμών και Συνολικών Μεγεθών Υποπληθυσμών

Ας υποθέσουμε ότι ένα απλό τυχαίο δείγμα μεγέθους n επιλέγεται από κάποιο πεπερασμένο πληθυσμό μεγέθους N. Έστω N_0 ο αριθμός των μονάδων του πληθυσμού που ανήκει σε κάποιο υποσύνολο του πληθυσμού και n_0 ο αριθμός των μονάδων του δείγματος που ανήκουν σ' αυτό το υποσύνολο. Τότε, αποδεικνύεται η εξής πρόταση:

Πρόταση 2.7.1: Το υποσύνολο n_0 μονάδων του αρχικού τυχαίου δείγματος αποτελεί ένα απλό τυχαίο δείγμα από τον υποπληθυσμό των N_0 μονάδων και ισχύει ότι $E\left(\frac{n_0}{N_0}\right) = \frac{n}{N}$.

Απόδειξη: Για το πρώτο μέρος της πρότασης, αρκεί να δειχθεί ότι όλα τα δυνατά δείγματα μεγέθους n_0 είναι ισοπίθανα.

Έστω ένα απλό τυχαίο δείγμα μεγέθους n από τον αρχικό πληθυσμό. Υπάρχουν $\binom{N}{n}$ δυνατά δείγματα μεγέθους n. Σε κάθε ένα από αυτά, n_0 μονάδες ανήκουν στον υπ' όψη υποπληθυσμό και $n\text{-}n_0$ δεν ανήκουν σ' αυτόν με $\binom{N_0}{n_0}\binom{N\text{-}N_0}{n\text{-}n_0}$ τρόπους. Επομένως, η πιθανότητα με την οποία μια οποιαδήποτε επιλογή n_0 μονάδων από τις n ανήκει στον υπό εξέταση υποπληθυσμό είναι ίση με

$$\binom{N_0}{n_0}\binom{N-N_0}{n-n_0}\Big/\binom{N}{n}.$$

Άρα, όλα τα (υπο)δείγματα μεγέθους n_0 είναι ισοπίθανα. Για την απόδειξη του δεύτερου συμπεράσματος, έστω

$$X_i=\begin{cases}1 & \text{αν η i μονάδα του δείγματος}\\ & \text{ανήκει στον υποπληθυσμό}\\ 0 & \text{διαφορετικά}\end{cases},$$

και

$$x_i=\begin{cases}1 & \text{αν η i μονάδα του πληθυσμού}\\ & \text{ανήκει στον υποπληθυσμό}\\ 0 & \text{διαφορετικά}\end{cases}.$$

Τότε,

$$\sum_{i=1}^{n}X_i=n_0 \quad \text{και} \quad \sum_{i=1}^{N}x_i=N_0.$$

Επομένως,

$$E\left(\frac{n_0}{N_0}\right)=\frac{1}{N_0}E\left(\sum_{i=1}^{n}X_i\right)=\frac{1}{N_0}\ \frac{n}{N}\ \sum_{i=1}^{N}x_i=\frac{n}{N},\qquad \text{ο.ε.δ.}$$

Αυτό που ουσιαστικά αποδείχθηκε είναι ότι κάθε απλό τυχαίο δείγμα από ένα πληθυσμό περιέχει ένα απλό τυχαίο δείγμα από οποιονδήποτε υποπληθυσμό.

Η σημασία αυτού του συμπεράσματος, όσο αφορά τις εφαρμογές, είναι μεγάλη. Παρέχει την δυνατότητα στον ερευνητή να χρησιμοποιήσει πληροφορίες από το δείγμα για την συναγωγή συμπερασμάτων όχι μόνο για ολόκληρο τον πληθυσμό αλλά και για οποιονδήποτε υποπληθυσμό. Για παράδειγμα, με βάση ένα απλό τυχαίο δείγμα από κάποιο πληθυσμό εργαζόμενων ατόμων είναι δυνατή η μελέτη της κατανομής του εισοδήματος σε ολόκληρο τον πληθυσμό αλλά και παράλληλα σε υποπληθυσμούς, όπως οι άνδρες ηλικίας 30-40 ετών ή οι γυναίκες με δύο παιδιά κ.λ.π. Οι υποπληθυσμοί στους οποίους μπορεί να διαιρεθεί ένας πληθυσμός ονομάζονται συνήθως **περιοχές μελέτης**.

Έστω ο πληθυσμός $\{y_1, y_2, \ldots, y_N\}$ και έστω ότι κάθε μονάδα του πληθυσμού αυτού ανήκει σε ένα από k υποπληθυσμούς. Τότε, $N=N_1+N_2+\ldots+N_k$, όπου N_j αντιπροσωπεύει το μέγεθος του j υποπληθυσμού. Κατά συνέπεια, αν $X_1, X_2, \ldots, X_n$ είναι ένα απλό τυχαίο δείγμα μεγέθους n ισχύει ότι

$$n=n_1+n_2+\ldots+n_k\,,$$

όπου n_j είναι το μέγεθος του τμήματος του δείγματος του οποίου οι μονάδες ανήκουν στην j περιοχή. Έστω ότι $y_i^{(j)}$ αντιπροσωπεύει την i μονάδα της j περιοχής του πληθυσμού και $X_i^{(j)}$ την i μονάδα του τυχαίου δείγματος που αντιστοιχεί στην j περιοχή. Τότε, η μέση τιμή και η διασπορά του j υποπληθυσμού δίνονται προφανώς από τους τύπους

$$\mu_j=\frac{1}{N_j}\sum_{i=1}^{N_j}y_i^{(j)}$$

και

$$\sigma_j^2 = \frac{1}{N_j - 1} \sum_{i=1}^{N_j} (y_i^{(j)} - \mu_j)^2 ,$$

αντίστοιχα, ενώ, σύμφωνα με την θεωρία που αναπτύχθηκε στα προηγούμενα, οι αμερόληπτες εκτιμήτριές τους είναι οι στατιστικές συναρτήσεις

$$\bar{X}_{n_j} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_i^{(j)}$$

και

$$S_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (X_i^{(j)} - \bar{X}_{n_j})^2 ,$$

αντίστοιχα. Προφανώς η εκτιμήτρια του τυπικού σφάλματος της εκτιμήτριας $\bar{X}_{n_j}$ είναι ίση με

$$S_{\bar{X}_{n_j}} = \frac{S_j}{\sqrt{n_j}} \sqrt{1 - \frac{n_j}{N_j}}$$

Παρατήρηση: Αν η τιμή του μεγέθους του j υποπληθυσμού, N_j, δεν είναι γνωστή, μπορεί να χρησιμοποιηθεί στην θέση του λόγου n_j / N_j ο λόγος n/N για τον υπολογισμό της διόρθωσης πεπερασμένου πληθυσμού. Αυτό είναι απόρροια του γεγονότος ότι στην απλή τυχαία δειγματοληψία η τιμή n_j / N_j είναι αμερόληπτη εκτίμηση της τιμής n/N (πρόταση 2.7.1). Στην περίπτωση αυτή

$$S_{\bar{X}_{n_j}} = S_j \sqrt{\frac{1 - n/N}{n_j}} .$$

Για την εκτίμηση του συνολικού μεγέθους

$$y^{(j)} = \sum_{i=1}^{N_j} y_i^{(j)},$$

χρησιμοποιείται η στατιστική συνάρτηση

$$\hat{Y}^{(j)} = N_j S\bar{X}_{n_j}.$$

Αυτή, σύμφωνα με την θεωρία των προηγουμένων παραγράφων, είναι αμερόληπτη εκτιμήτρια του y_j με τυπικό σφάλμα εκτιμώμενο από την στατιστική συνάρτηση

$$S_{\hat{Y}^{(j)}} = N_j \; S\bar{X}_{n_j}$$

Η παραπάνω εκτιμήτρια του $y^{(j)}$ προϋποθέτει γνώση της τιμή του N_j. Στην πράξη όμως πολύ σπάνια συμβαίνει να είναι γνωστή αυτή η τιμή. Στις περιπτώσεις αυτές ως εκτιμήτρια του μεγέθους $y^{(j)}$ χρησιμοποιείται η συνάρτηση

$$\hat{Y}^{(j)} = \frac{N}{n} \sum_{i=1}^{n_j} X_i^{(j)} \qquad (2.7.1)$$

Η λογική πίσω από την επιλογή αυτής της στατιστικής συνάρτησης ως εκτιμήτριας του $y^{(j)}$ γίνεται ευκολότερα αντιληπτή με την βοήθεια της επιχειρηματολογίας που ακολουθεί. Έστω

$$y_i' = \begin{cases} y_i^{(j)} & \text{αν η i μονάδα του πληθυσμού} \\ & \text{ανήκει στην j περιοχή} \\ 0 & \text{διαφορετικά} \end{cases}, \qquad i = 1, 2, \ldots, N$$

και

$$X_i' = \begin{cases} X_i^{(j)} & \text{αν η i μονάδα του δείγματος} \\ & \text{ανήκει στην j περιοχή} \\ 0 & \text{διαφορετικά} \end{cases}, \qquad i = 1, 2, \ldots, n.$$

Τότε $\{ y_1', \; y_2', \; \ldots, \; y_N' \}$ είναι ένα σύνολο τιμών από τις οποίες N_j είναι ίσες με τις τιμές των N_j μονάδων της j περιοχής και $N - N_j$ είναι ίσες με 0.

Επίσης X'_1, X'_2,..., X'_n είναι μια ακολουθία τυχαίων μεταβλητών, από τις οποίες n_j ταυτίζονται με τις τιμές $X_1^{(j)},...,X_{n_j}^{(j)}$ και n - n_j είναι ίσες με 0. Άρα

$$\text{(i)} \qquad y^{(j)} = \sum_{i=1}^{N_j} y_i^{(j)} = \sum_{i=1}^{N} y'_i$$

$$\text{(ii)} \qquad \bar{X}'_n = \frac{1}{n} \sum_{i=1}^{n} X'_i = \frac{1}{n} \sum_{i=1}^{n_j} X_i^{(j)}$$

Επομένως, μια εκτιμήτρια του συνολικού μεγέθους $y^{(j)}$ είναι η στατιστική συνάρτηση

$$\hat{Y}^{(j)} = NX'_n = \frac{N}{n} \sum_{i=1}^{n_j} X_i^{(j)}$$

Θεώρημα 2.7.1: Η στατιστική συνάρτηση $\hat{Y}^{(j)}$ είναι αμερόληπτη εκτιμήτρια του συνολικού μεγέθους $y^{(j)}$ με τυπικό σφάλμα εκτιμώμενο από την στατιστική συνάρτηση

$$S_{\hat{Y}^{(j)}} = NS' \sqrt{\frac{1 - n/N}{n}}, \qquad (2.7.2)$$

όπου

$$S'^2 = \frac{1}{n-1} \left[\sum_{i=1}^{n_j} (X_i^{(j)})^2 - \frac{\left(\sum_{i=1}^{n_j} X_i^{(j)} \right)^2}{n} \right] \qquad (2.7.3)$$

Απόδειξη: **Αμεροληψία**: Ισχύει ότι

$$E(\hat{Y}^{(j)}) = N\,E\,(\bar{X}'_n) = \frac{N}{n} E\,(\sum_{i=1}^{n} X') = \frac{N}{n}\,\frac{n}{N}\,\sum_{i=1}^{N} y'_i = \sum_{i=1}^{N_j} y_i^{(j)} = y^{(j)}$$

Υπολογισμός τυπικού σφάλματος: Είναι προφανές ότι η διασπορά των τιμών $y'_1,\ldots, y'_N$ είναι ίση με

$$\sigma'^{\,2} = \frac{1}{N-1}\left[\sum_{i=1}^{N} (y'_i)^2 - \frac{\left[\sum_{i=1}^{N} y'_i\right]^2}{N}\right]$$

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

και, επομένως, μια αμερόληπτη εκτιμήτρια του σ'^2 δίνεται από την (2.7.3). Κατά συνέπεια,

$$V(\hat{Y}^{(j)}) = N^2 V(\bar{X}'_n) = N^2\,\frac{\sigma'^2}{n}\left(1 - \frac{n}{N}\right).$$

Η τελευταία σχέση οδηγεί στην (2.7.2) και ολοκληρώνει την απόδειξη του θεωρήματος.

Παράδειγμα 2.7.1: Από μία λίστα 2422 διαφορετικών τύπων δαπανών των νοικοκυριών μιας περιοχής επελέγη ένα απλό τυχαίο δείγμα 180 δαπανών του νοικοκυριού. Από τους 180 τύπους δαπανών 28

θεωρήθηκαν μη καθοριστικής σημασίας. Το συνολικό ύψος των 152 ″καθοριστικών″ δαπανών ήταν ίσο με 343.5 ευρώ και το άθροισμα των τετραγώνων ίσο με 1491.38 (ευρώ)^2. Να εκτιμηθεί το συνολικό ύψος των δαπανών ενός νοικοκυριού και να υπολογισθεί το τυπικό σφάλμα της εκτίμησης.

Λύση: Προφανώς, αν n_1 είναι ο αριθμός των ″καθοριστικών″ δαπανών στο δείγμα ισχύει ότι n=180, n_1=152,

$$\sum_{i=1}^{152} X_i^{(1)} = 343.5 \quad \text{και} \quad \sum_{i=1}^{152} (X_i^{(1)})^2 = 1491.38$$

Άρα

$$\hat{Y}^{(1)} = \frac{N}{n} \sum_{i=1}^{n_1} X_i^{(1)} = \frac{2422}{180} \sum_{i=1}^{152} X_i^{(1)} = \frac{2422}{180} \ (343.5) = 4622$$

με τυπικό σφάλμα

$$S_{\hat{Y}^{(1)}} = 2422 \ S' \sqrt{\frac{1-180/2422}{180}},$$

όπου, από την σχέση (2.7.3),

$$S'^2 = \frac{1}{179}\left(1491.38 - \frac{(343.5)^2}{180}\right) = 4.67 .$$

Δηλαδή τελικά

$$S_{\hat{Y}^{(1)}} = 2422\sqrt{\frac{4.67}{180}\left(1-\frac{180}{2422}\right)} = 375 .$$

ΑΣΚΗΣΕΙΣ

1. Έστω ο πληθυσμός μεγέθους N=6 που αποτελείται από τις τιμές 8,3,1,11,4 και 7. Να υπολογισθεί ο μέσος όλων των δυνατών δειγμάτων

μεγέθους n=2 και να επαληθευθεί ότι αποτελεί μια αμερόληπτη εκτιμήτρια της μέσης τιμής μ του πληθυσμού με διασπορά σ^2 (1-f)/n.

2. Για τον ίδιο πληθυσμό, να υπολογισθεί η διασπορά όλων των δυνατών δειγμάτων μεγέθους 3 και να επαληθευθεί ότι $E(S^2)=\sigma^2$.

3. Ένα απλό τυχαίο δείγμα 30 νοικοκυριών επελέγη από ένα τμήμα 14848 νοικοκυριών κάποιας πόλης. Οι αριθμοί των προσώπων ανά νοικοκυριό στο δείγμα ήταν οι εξής:

5 6 3 3 2 3 3 3 4 4 3 2 7 4 3
5 4 4 3 3 4 3 3 1 2 4 3 4 2 4.

α) Να εκτιμηθεί ο συνολικός αριθμός των κατοίκων της περιοχής και να υπολογισθεί το τυπικό σφάλμα της εκτίμησης.
β) Να υπολογισθεί η πιθανότητα με την οποία η εκτίμηση διαφέρει το πολύ κατά 10% από τον πραγματικό αριθμό.

4. Σκοπός μιας μελέτης είναι η δυνατότητα χρησιμοποίησης ενός τυχαίου δείγματος για την ταχύτερη απογραφή του αποθέματος μιας αποθήκης. Η καταμέτρηση της αξίας των ειδών στα 36 ράφια της αποθήκης έδωσε τα εξής αποτελέσματα (σε ευρώ).

29 38 42 44 45 47 51 53 53 54 56 56 56 58 58 59 60 60
60 60 61 61 61 62 64 65 65 67 67 68 69 71 74 77 82 85.

Η εκτίμηση της συνολικής αξίας του αποθέματος δεν πρέπει να διαφέρει από την πραγματική αξία περισσότερο από 200 ευρώ εκτός από μια στις 20 περιπτώσεις. Θα συμφωνούσατε ότι ένα απλό τυχαίο δείγμα μεγέθους 12 πληροί τις προϋποθέσεις αυτές; (Δίνεται ότι $\Sigma y_i=2138$ και $\Sigma y_i^2=131,682$).

5. Από μια λίστα 468 σχολών διετούς φοίτησης επελέγη ένα απλό τυχαίο δείγμα 100 σχολών. Το δείγμα περιείχε 54 δημόσιες και 46 ιδιωτικές σχολές. Ο πίνακας που ακολουθεί συνοψίζει τα αποτελέσματα της δειγματοληψίας όσο αφορά τους αριθμούς των φοιτητών (Y_i) και καθηγητών (X_i).

Σχολές	n	ΣY_i	ΣY_i^2	ΣX_i	ΣX_i^2	$\Sigma X_i Y_i$
Δημόσιες	54	31281	29881219	2024	111090	1729349
Ιδιωτικές	46	13707	6366785	1075	33119	431041

α) Για κάθε τύπο σχολής, να εκτιμηθεί ο λόγος του αριθμού των φοιτητών προς τον αριθμό των καθηγητών.
β) Να υπολογισθούν τα τυπικά σφάλματα των εκτιμήσεων.
γ) Για τις δημόσιες σχολές να κατασκευασθεί ένα 90% διάστημα εμπιστοσύνης για την αναλογία φοιτητών / καθηγητών σε ολόκληρο τον πληθυσμό.

6. Στην προηγούμενη άσκηση να ελεγχθεί η υπόθεση ότι οι λόγοι αριθμού φοιτητών προς αριθμό καθηγητών διαφέρουν σημαντικά στους δυο τύπους σχολών.

7. Να εκτιμηθεί ο συνολικός αριθμός των καθηγητών των δημοσίων σχολών της άσκησης 5
α) αν είναι γνωστό ότι ο συνολικός αριθμός των δημοσίων σχολών είναι 251 και
β) αν το στοιχείο αυτό δεν είναι γνωστό.
Και στις δύο περιπτώσεις να υπολογισθεί το τυπικό σφάλμα της εκτίμησης.

8. Δυο οδοντίατροι Α και Β κάνουν μια έρευνα για την κατάσταση των δοντιών 200 παιδιών ενός χωριού. Ο Α επιλέγει ένα απλό τυχαίο δείγμα 20 παιδιών και μετρά τον αριθμό των χαλασμένων δοντιών για κάθε παιδί με τα εξής αποτελέσματα:

αριθμός χαλασμένων δοντιών ανά παιδί	0	1	2	3	4	5	6	7	8	9	10
αριθμός παιδιών	8	4	2	2	1	1	0	0	0	1	1

Ο Β χρησιμοποιώντας την ίδια οδοντιατρική τεχνική εξετάζει και τα 200 παιδιά και καταγράφει 60 παιδιά τα οποία δεν έχουν χαλασμένα δόντια.

Να εκτιμηθεί ο συνολικός αριθμός των χαλασμένων δοντιών όλων των παιδιών του χωριού.
α) χρησιμοποιώντας τα αποτελέσματα του Α και
β) χρησιμοποιώντας τα αποτελέσματα και των δυο οδοντιάτρων.

9. Ένα απλό τυχαίο δείγμα μεγέθους 3 επιλέγεται με επανάθεση από ένα πληθυσμό μεγέθους N. Να δειχθεί ότι οι πιθανότητες P_i το δείγμα να περιέχει i διαφορετικές μονάδες του πληθυσμού, i=1,2,3, δίνονται από τους τύπους

$$P_1 = 1/N^2,\ P_2 = 3(N-1)/N^2,\ P_3 = (N-1)(N-2)/N^2.$$

2.8 Ποιοτικά Χαρακτηριστικά

Πολλές φορές μας ενδιαφέρει να εκτιμήσουμε τον συνολικό αριθμό ή το ποσοστό των μονάδων ενός πληθυσμού οι οποίες εμπίπτουν σε κάποια κατηγορία. Για παράδειγμα, πολλές δειγματοληπτικές έρευνες έχουν σαν σκοπό την εκτίμηση του ποσοστού ανεργίας, του ποσοστού θανατηφόρων τροχαίων ατυχημάτων ή το ποσοστό των ψηφοφόρων μιας πόλης ή χώρας υπέρ κάποιου κόμματος ή νομοθεσίας. Στα πλαίσια προβλημάτων αυτής της μορφής, οι μονάδες ενός πληθυσμού μεγέθους N ταξινομούνται σε δυο κατηγορίες: Α αν έχουν το χαρακτηριστικό που μας ενδιαφέρει και Α′ διαφορετικά. Ο ακριβής αριθμός των μονάδων του πληθυσμού που ανήκουν στην κατηγορία Α, έστω N_A, δεν είναι γνωστός, όπως επίσης άγνωστο είναι και το ποσοστό $p = \frac{N_A}{N}$ των μονάδων αυτών. Η εκτίμηση αυτών των δυο μεγεθών είναι το αντικείμενο αυτής της ενότητας.

2.8.1 Εκτίμηση Ποσοστών

Έστω πληθυσμός μεγέθους N και έστω p το ποσοστό των μονάδων που ανήκουν σε κάποια κατηγορία Α. Ας υποθέσουμε ότι

επιθυμούμε να εκτιμήσουμε την παράμετρο p με βάση τις πληροφορίες (παρατηρήσεις) ενός απλού τυχαίου δείγματος μεγέθους n. Η στατιστική συνάρτηση που φαίνεται να είναι η πιο κατάλληλη για τον σκοπό αυτό είναι η

$$\hat{p} = \frac{X}{n},$$

όπου X παριστάνει τον αριθμό των μονάδων του δείγματος που ανήκουν στην κατηγορία A. Για τον υπολογισμό του τυπικού σφάλματος της εκτιμήτριας $\hat{p}$ είναι δυνατόν να εφαρμοσθεί η θεωρία που αναπτύχθηκε στις προηγούμενες παραγράφους για τον μέσο $\bar{X}_n$ ενός απλού τυχαίου δείγματος, με την εισαγωγή των εξής βοηθητικών μεγεθών.

Έστω

$$y_i = \begin{cases} 1 & \text{αν η i μονάδα του πληθυσμού } \in A \\ 0 & \text{διαφορετικά} \end{cases}$$

i = 1,2,…, N και

$$X_i = \begin{cases} 1 & \text{αν η i μονάδα του δείγματος } \in A \\ 0 & \text{διαφορετικά} \end{cases}$$

i = 1,2,…, n. Τότε, προφανώς

$$N_A = \sum_{i=1}^{N} y_i \ , \quad X = \sum_{i=1}^{n} X_i$$

$$p = \frac{1}{N} \sum_{i=1}^{N} y_i = \mu \qquad \text{και} \qquad \hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}_n \, .$$

Άρα το πρόβλημα εκτίμησης του p και του N_A είναι ισοδύναμο με το πρόβλημα εκτίμησης της μέσης τιμής και του συνολικού ύψους των

τιμών ενός πληθυσμού, αντίστοιχα, του οποίου οι τιμές των μονάδων είναι 0 και 1.
Προφανώς,

$$\sum_{i=1}^{N} y_i^2 = \sum_{i=1}^{N} y_i = N_A = N\,p$$

και

$$\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i = X = n\,\hat{p}\,.$$

Άρα, η διασπορά του πληθυσμού είναι ίση με

$$\begin{aligned}
\sigma^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(y_i - p)^2 \\
&= \frac{1}{N-1}\left(\sum_{i=1}^{N} y^2 - Np^2\right) \\
&= \frac{Np - Np^2}{N-1} \\
&= \frac{N}{N-1}\,p\,(1-p) \qquad (2.8.1)
\end{aligned}$$

και μια εκτιμήτριά της παρέχεται από την στατιστική συνάρτηση

$$S^2 = \frac{n}{n-1}\,\hat{p}\,(1-\hat{p})\,.$$

Θεώρημα 2.8.1: Η στατιστική συνάρτηση $\hat{p}=X/n$ είναι αμερόληπτη εκτιμήτρια της παραμέτρου p και ισχύει ότι

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

Απόδειξη: Αρκεί να παρατηρηθεί ότι

$$E(\hat{p}) = E(\bar{X}_n) = \mu = p$$

και

$$V(\hat{p}) = V(\bar{X}_n)$$

$$= \frac{\sigma^2}{n} \frac{N-n}{N}$$

$$= \frac{p(1-p)}{n} \frac{N-n}{N-1}, \text{ (από την (2.8.1))} \quad \text{ο.ε.δ.}$$

Πόρισμα 1: Η στατιστική συνάρτηση $\hat{Y}_A = N\hat{p}$ είναι μια αμερόληπτη εκτιμήτρια του N_A με διασπορά

$$V(\hat{Y}_A) = \frac{N^2 p(1-p)}{n} \frac{N-n}{N-1}.$$

Θεώρημα 2.8.2: Η στατιστική συνάρτηση

$$S^2_{\hat{p}} = \frac{\hat{p}(1-\hat{p})}{n-1} \frac{N-n}{N}$$

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

Απόδειξη:

$$S^2_{\hat{p}} = \hat{\sigma}^2_{\bar{X}_n} = \frac{\hat{\sigma}^2}{n} \frac{N-n}{N}$$

$$= \frac{S^2}{n} \frac{N-n}{N}$$
$$= \frac{n\,\hat{p}\,(1-\hat{p})}{(n-1)\,n} \frac{N-n}{N}$$
$$= \frac{\hat{p}\,(1-\hat{p})}{n-1} \frac{N-n}{N}.$$

Πόρισμα 2: Η στατιστική συνάρτηση

$$S^2_{\hat{Y}_A} = \frac{N(N-n)}{n-1}\hat{p}\,(1-\hat{p})$$

είναι μια αμερόληπτη εκτιμήτρια της $V(\hat{Y}_A)$.

Παράδειγμα: Από τους 3042 φοιτητές ενός πανεπιστημίου επελέγη ένα απλό τυχαίο δείγμα 200 φοιτητών. Αν 38 από τους φοιτητές αυτούς ήταν υπέρ της εξωτερικής πολιτικής της χώρας τους, να εκτιμηθεί ο συνολικός αριθμός των φοιτητών του πανεπιστημίου που υποστηρίζουν την εξωτερική πολιτική της χώρας τους. Να υπολογισθεί το τυπικό σφάλμα της εκτίμησης.

Λύση:

$$N = 3042,\ n = 200,\ X = 38.$$

Άρα, $\hat{p} = 0.19$

και, επομένως,

$$\hat{Y}_A = N\hat{p} = 578,$$
$$S^2_{\hat{Y}_A} = (3042)\,(2842)\,(0.19)\,(0.81)\,/199 = 6686$$
$$\Rightarrow S^2_{\hat{Y}_A} = \sqrt{6686} = 81.8.$$

2.8.2 Κατανομή της Εκτιμήτριας $\hat{p}$ - Διαστήματα Εμπιστοσύνης για την Παράμετρο p

Η ακριβής κατανομή της στατιστικής συνάρτησης $\hat{p}$ μπορεί να προσδιορισθεί μέσω της κατανομής της τυχαίας μεταβλητής $X = \sum_{i=1}^{n} X_i$, όπου X_i, i =1,2,…,n ορίζονται όπως προηγουμένως.

Επειδή $P(X_i =1) = 1- P(X_i =0) = p = N_A/N$, έπεται ότι η κατανομή της X είναι

(α) διωνυμική, στην περίπτωση δειγματοληψίας με επανάθεση, οπότε

$$P(X = r) = \binom{n}{r} \left(\frac{N_A}{N}\right)^r \left(1 - \frac{N_A}{N}\right)^{n-r}, \qquad r = 0, 1, \ldots, n$$

ή

(β) υπεργεωμετρική, στην περίπτωση δειγματοληψίας χωρίς επανάθεση, οπότε

$$P(X = r) = \binom{N_A}{r} \binom{N - N_A}{n - r} \Big/ \binom{N}{n}, \qquad r = 0, 1, \ldots, n$$

Και στις δύο περιπτώσεις, η συνάρτηση κατανομής της στατιστικής συνάρτησης $\hat{p}$ προσδιορίζεται από την σχέση

$$\begin{aligned} F_{\hat{p}}(z) &= P(\hat{p} \le z) \\ &= P\left(\frac{X}{n} \le z\right) \\ &= P(X \le nz) \\ &= \sum_{x=0}^{nz} P(X = x). \end{aligned}$$

Σημείωση: Το αποτέλεσμα της (α) ισχύει και στην περίπτωση δειγματοληψίας με ή χωρίς επανάθεση από άπειρο πληθυσμό. Μια καλή

προσέγγιση της κατανομής της X (και συνεπώς της $\hat{p}$) παρέχεται από την κανονική κατανομή όταν $np \geq 5$ και $n(1-p) \geq 5$. Συγκεκριμένα, ισχύει στην περίπτωση αυτή ότι

$$\frac{X - np}{\sigma_x} \underset{\text{appr.}}{\sim} N(0,1)$$

ή, ισοδύναμα,

$$\frac{\hat{p} - p}{\sigma_{\hat{p}}} \underset{\text{appr.}}{\sim} N(0,1).$$

Κατά συνέπεια, ένα 100(1-α)% διάστημα εμπιστοσύνης για την παράμετρο p έχει άκρα

$$\hat{p} \pm z_{1-\alpha/2}\sigma_{\hat{p}} \qquad (2.8.2)$$

Στην θέση της τυπικής απόκλισης $\sigma_{\hat{p}}$ μπορεί να χρησιμοποιηθεί η εκτιμήτριά της, $S_{\hat{p}}$. Ο λόγος είναι ότι η $S_{\hat{p}}^2$ είναι συνεπής εκτιμήτρια του $\sigma_{\hat{p}}^2$ και, επομένως, για μεγάλο n, η $S_{\hat{p}}^2$ δεν διαφέρει πολύ από την $\sigma_{\hat{p}}^2$. Συνήθως, στην ακτίνα του διαστήματος που ορίζεται από τα άκρα (2.8.2) εφαρμόζεται και η λεγόμενη **διόρθωση συνεχείας**, η οποία είναι ίση με 1/(2n). Άρα τελικά, τα 100(1-α)% όρια εμπιστοσύνης για την παράμετρο p δίνονται από τους εξής τύπους:

α) Δειγματοληψία με επανάθεση ή δειγματοληψία από άπειρο πληθυσμό (με ή χωρίς επανάθεση)

$$\hat{p} \pm \left(z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} + \frac{1}{2n} \right) \qquad (2.8.3)$$

β) Δειγματοληψία χωρίς επανάθεση από πεπερασμένο πληθυσμό

$$\hat{p} \pm \left(z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}\left(1-\frac{n}{N}\right)} + \frac{1}{2n} \right) \qquad (2.8.4)$$

Σημείωση: Προφανώς, τα άκρα του 100(1-α)% διαστήματος εμπιστοσύνης του N_A είναι ίσα με N×(άκρα του 100(1-α)% διαστήματος εμπιστοσύνης του p).

Παράδειγμα: Από τα 500 ελαστικά αυτοκινήτων, που κατασκευάσθηκαν από μια εταιρεία σε μια συγκεκριμένη χρονική περίοδο, επελέγη ένα απλό τυχαίο δείγμα μεγέθους 100 που έδειξε ότι 37 από αυτά δεν συμφωνούσαν με τις προδιαγραφές. Να κατασκευασθεί ένα 95% διάστημα εμπιστοσύνης για τον συνολικό αριθμό των ελαστικών που δεν συμφωνούν με τις προδιαγραφές.

Λύση:

$$n = 100,\ N = 500,\ \hat{p} = 0.37 \text{ και } z_{0.975} = 1.96.$$

Άρα, από την (2.8.4), το ζητούμενο 95% διάστημα εμπιστοσύνης έχει άκρα

$$N\left[\hat{p} \pm \left(z_{0.975} \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}\left(1-\frac{n}{N}\right)} + \frac{1}{2n} \right) \right].$$

Δηλαδή,

$$500\left[0.37 \pm \left(1.96 \sqrt{\frac{(0.37)(0.63)}{99}\left(1-\frac{100}{500}\right)} + \frac{1}{200} \right) \right].$$

Επομένως, το 95% διάστημα εμπιστοσύνης για το συνολικό αριθμό των ελαστικών που δε συμφωνούν με τις προδιαγραφές είναι το διάστημα (140, 230).

Παρατήρηση: Είναι προφανές ότι η παραπάνω θεωρία εφαρμόζεται και σε κάθε περιοχή μελέτης του αρχικού πληθυσμού. Ας υποθέσουμε ότι έχουμε την εξής διάταξη:

	Περιοχή μελέτης					
	1		2		…	k	
Κατηγορία	A	A΄	A	A΄	…	A	A΄
αριθμός μονάδων δείγματος στην κατηγορία	X_1	n_1-X_1	X_2	n_2-X_2	…	X_k	n_k-X_k

Τότε, προφανώς, ισχύει ότι

$$\hat{p}_i = \frac{X_i}{n_i} \qquad S_{\hat{p}_i} = \sqrt{\frac{\hat{p}_i\,(1-\hat{p}_i)}{n_i-1}\left(1-\frac{n_i}{N_i}\right)}$$

$$\hat{Y}_A^{(i)} = N_i \frac{X_i}{n_i} \qquad S_{\hat{Y}_A^{(i)}} = N_i\, S_{\hat{p}_i}.$$

Στην περίπτωση που η τιμή του N_i είναι άγνωστη, ως εκτιμήτρια του $N_A^{(i)}$ (του αριθμού των ατόμων της i περιοχής μελέτης του πληθυσμού που ανήκουν στην κατηγορία A) χρησιμοποιείται η συνάρτηση

$$\hat{Y}_A^{(i)} = \frac{N}{n}\, X_i,$$

της οποίας το τυπικό σφάλμα είναι ίσο με

$$S_{\hat{Y}_A} = N\sqrt{\frac{\hat{p}_i'\,(1-\hat{p}_i')}{n-1}\left(1-\frac{n}{N}\right)},$$

όπου εδώ $\hat{p}_i' = \frac{X_i}{n}$.

2.9 Το Μέγεθος του Δείγματος

Το πρώτο ερώτημα με το οποίο βρίσκεται αντιμέτωπος ένας ερευνητής, όταν πρόκειται να κάνει μια δειγματοληπτική έρευνα, είναι πόσο μεγάλο πρέπει να είναι το δείγμα που θα χρησιμοποιήσει. Αν ορισμένοι παράγοντες, όπως το κόστος της δειγματοληψίας μιας μονάδας δεν λαμβάνονται υπ' όψη, δεν υπάρχει βασική δυσκολία όσον αφορά τον

προσδιορισμό του μεγέθους n του απαιτούμενου δείγματος. Η τιμή του n θα πρέπει προφανώς να εξαρτάται από τον βαθμό ακρίβειας με την οποία ο ερευνητής επιθυμεί να εκτιμήσει την οποιαδήποτε παράμετρο του πληθυσμού. Πιο συγκεκριμένα, θα πρέπει να εξαρτάται από το μέγιστο ανεκτό σφάλμα της εκτίμησης και την πιθανότητα με την οποία αυτό είναι επιτρεπτό.

2.9.1 Δειγματοληψία για την Εκτίμηση της Μέσης Τιμής ενός Πληθυσμού

Για την περίπτωση εκτίμησης της μέσης τιμής μ ενός πληθυσμού με διασπορά σ^2, το ερώτημα σχετικά με το μέγεθος του δείγματος μπορεί, λαμβάνοντας υπ' όψη τα παραπάνω, να διατυπωθεί ως εξής:

"Πόσο μεγάλο πρέπει να είναι το μέγεθος n του δείγματος, ώστε, με πιθανότητα 1-α, το σφάλμα που κάνει ο ερευνητής εκτιμώντας την άγνωστη μέση τιμή μ με τον μέσο $\bar{X}_n$ του δείγματος να μην υπερβαίνει την τιμή e;"

Το ερώτημα αυτό ισοδυναμεί με το ερώτημα:

"Ποια είναι η τιμή του n για την οποία ισχύει $P(\left| \bar{X}_n - \mu \right| \leq e) = 1 - \alpha$;"

Είναι γνωστό ότι

$$P(\left| \bar{X}_n - \mu \right| \leq e) = 1 - \alpha \Leftrightarrow$$

$$P\left(\left| \frac{\bar{X}_n - \mu}{\sigma_{\bar{X}_n}} \right| \leq \frac{e}{\sigma_{\bar{X}_n}} \right) = 1 - \alpha . \qquad (2.9.1)$$

Η παραπάνω σχέση μπορεί να οδηγήσει στον προσδιορισμό της τιμής του n, αν, για το συγκεκριμένο πρόβλημα, μπορεί να υποτεθεί ότι ο πληθυσμός είναι κανονικός ή κατά προσέγγιση κανονικός. Στην περίπτωση αυτή,

$$\frac{\bar{X}_n - \mu}{\sigma_{\bar{X}_n}} \sim N(0,1).$$

Τότε, η (2.9.1) ισχύει τότε και μόνο τότε αν

$$\frac{e}{\sigma_{\bar{X}_n}} = z_{1 - \alpha/2}$$

ή, ισοδύναμα, τότε και μόνο τότε αν

$$\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}} = \frac{e}{z_{1-\alpha/2}}.$$

Λύνοντας την τελευταία εξίσωση ως προς n έχουμε

$$n = \frac{N\,\sigma^2\;z_{1-\alpha/2}^2}{N\,e^2 + \sigma^2\;z_{1-\alpha/2}^2}. \qquad (2.9.2)$$

Από τον τύπο αυτό, είναι προφανές ότι αν $N\to\infty$ (περίπτωση απείρου πληθυσμού), τότε

$$n = \left(\frac{\sigma\, z_{1 - \alpha/2}}{e}\right)^2 .$$

Αν, δηλαδή, $\frac{n}{N} \to 0$, η τιμή του δειγματικού μεγέθους n είναι η λύση της εξίσωσης

$$\frac{\sigma}{\sqrt{n}} = \frac{e}{z_{1-\alpha/2}}.$$

Στην πράξη, ως πρώτη προσέγγιση του n, λαμβάνεται η τιμή

$$n_0 = \left(\frac{\sigma \, z_{1-\alpha/2}}{e} \right)^2 \qquad (2.9.3)$$

και, ως τελική τιμή του n, θεωρείται η τιμή

$$n = \begin{cases} n_0 & , \text{αν } \dfrac{n_0}{N} \le 0.05 \\ \dfrac{n_0}{1 + \dfrac{n_0}{N}} & , \text{αν } \dfrac{n_0}{N} > 0.05 \end{cases} \qquad (2.9.4)$$

Παρατήρηση: Αν η διασπορά σ^2 του πληθυσμού είναι άγνωστη, τότε χρησιμοποιείται μια εκτίμησή της βασισμένη σε κάποιο προκαταρκτικό δείγμα μεγέθους ≥ 30.

Παράδειγμα: Μια εταιρεία θέλει να εκτιμήσει τον μέσο μηνιαίο μισθό των 1000 υπαλλήλων της προκειμένου να κάνει συγκρίσεις με τους μέσους μηνιαίους μισθούς άλλων εταιρειών. Αν από παλιά εμπειρία γνωρίζει ότι οι μισθοί των υπαλλήλων της είναι κανονικά κατανεμημένοι με τυπική απόκλιση 100 ευρώ, πόσο μεγάλο πρέπει να είναι το δείγμα ώστε με πιθανότητα 95% η εκτίμηση να μην απέχει από την πραγματική τιμή περισσότερο από 20 ευρώ;

Λύση: $1 - \alpha = 0.95$, $N = 1000$, $\sigma = 15$. Άρα
$z_{1-\alpha/2} = z_{0.975} = 1.96$ και, επομένως, από την (2.9.3), έχουμε

$$n_0 = \left(\frac{100 \cdot (1.96)}{20} \right)^2 = 96.04 \simeq 96 .$$

Επειδή, όμως

$$\frac{n_0}{N} = \frac{96}{1000} = 0.096 > 0.05 ,$$

έπεται, από την (2.9.4), ότι

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{96}{1 + 0.096} = 87.59 \simeq 88 .$$

2.10 Δειγματοληψία για την Εκτίμηση ενός Ποσοστού

Ακολουθώντας την συλλογιστική της προηγούμενης παραγράφου, η τιμή του n θα πρέπει να καθορισθεί έτσι ώστε

$$P\left(\left| \hat{p} - p \right| \le e \right) = 1 - \alpha \Leftrightarrow P\left(\left| \frac{\hat{p} - p}{\sigma_{\hat{p}}} \right| < \frac{e}{\sigma_{\hat{p}}} \right) = 1 - \alpha$$

$$\Leftrightarrow \frac{e}{\sigma_{\hat{p}}} = z_{1-\alpha/2} .$$

Αλλά,

$$\sigma_{\hat{p}} = \frac{p(1-p)}{n} \; \frac{N-n}{N-1} .$$

Άρα, τελικά,

$$n = \frac{N \, z_{1-\alpha/2}^2 \; p(1-p)}{N \, e^2 + z_{1-\alpha/2}^2 \; p(1-p) - e^2} . \qquad (2.10.1)$$

Προφανώς, η (2.10.1) δεν μπορεί να δώσει την λύση, αφού η τιμή του n εξαρτάται από την τιμή της άγνωστης παραμέτρου p. Στην πράξη, συνήθως χρησιμοποιείται μια εκτίμηση της παραμέτρου p που βασίζεται σε παλαιότερη εμπειρία. Αν δεν υπάρχουν ενδείξεις όσον αφορά το ποια περίπου είναι η τιμή του p, τότε χρησιμοποιείται η συντηρητική τιμή p=1/2, η οποία μεγιστοποιεί το γινόμενο p(1-p). (Πράγματι, το γινόμενο $p(1-p) = p-p^2$ γίνεται μέγιστο $\Leftrightarrow (p-p^2)' = 0 \Leftrightarrow 1-2p=0 \Leftrightarrow p=1/2$).
Στην πράξη, και πάλι χρησιμοποιείται ως πρώτη προσέγγιση της τιμής του n η τιμή

$$n_0 = \frac{z^2_{1-\alpha/2}\, p(1-p)}{e^2},$$

οπότε, τελικά,

$$n = \begin{cases} n_0 & , \text{αν } \dfrac{n_0}{N} \le 0.05 \\ \dfrac{n_0}{1+(n_0-1)/N} & , \text{αν } \dfrac{n_0}{N} > 0.05 \end{cases} \qquad (2.10.2)$$

Παράδειγμα: Ένας ανθρωπολόγος θέλει να εκτιμήσει το ποσοστό των κατοίκων ενός απομακρυσμένου νησιού, οι οποίοι ανήκουν σε μια ορισμένη ομάδα αίματος. Το νησί έχει συνολικά 3200 κατοίκους. Πόσο μεγάλο πρέπει να είναι το δείγμα που θα χρησιμοποιήσει, αν επιθυμεί να έχει πιθανότητα 90% να εκτιμήσει το πραγματικό ποσοστό με σφάλμα το πολύ ίσο με 5% ;

Λύση: e =0.05, 1-α =0.90, p(1-p) =1/4. Άρα $z_{1-\alpha/2} = z_{0.95} = 1.645$ και κατά συνέπεια

$$n_0 = \frac{(1.645)^2}{4\ (0.05)^2} = 270.6 \simeq 271.$$

Επειδή

$$\frac{n_0}{N} = \frac{271}{3200} = 0.085 > 0.05,$$

έχουμε από την (2.10.2) ότι

$$n = \frac{271}{1+\ \dfrac{270}{3200}} = \ 249.9 \simeq 250.$$

Σημείωση: Αν στο παράδειγμα αυτό ο ερευνητής είχε λόγους να πιστεύει ότι το ποσοστό p είχε μια τιμή μεταξύ 20% και 35%, τότε θα μπορούσε να είναι λιγότερο συντηρητικός όσον αφορά την τιμή του γινομένου p(1-p) διαλέγοντας ως εκτίμηση του p την $\hat{p}=0.35$. Στην περίπτωση αυτή το απαιτούμενο δειγματικό μέγεθος θα ελαττωνόταν. Πράγματι, θα είχε σε πρώτη προσέγγιση την τιμή

$$n_0 = \frac{(1.645)^2\,(0.35)(0.65)}{(0.05)^2} = 246.25 \simeq 247$$

και τελικά (επειδή $\frac{n_0}{N} = \frac{247}{3200} = 0.077 > 0.05$)

$$n = \frac{247}{1 + \frac{246}{3200}} = 229.37 \simeq 230.$$

ΑΣΚΗΣΕΙΣ

1. Σε ένα απλό τυχαίο δείγμα 200 ατόμων από ένα πληθυσμό 2000 ατόμων, 120 άτομα ήταν υπέρ ενός νομοσχεδίου, 57 κατά και 23 δεν εξέφρασαν γνώμη. Να εκτιμηθεί ο συνολικός αριθμός των ατόμων του πληθυσμού που είναι υπέρ του νομοσχεδίου με ένα 95% διάστημα εμπιστοσύνης.

2. Αποτελούν τα αποτελέσματα της προηγούμενης άσκησης ένδειξη ότι η πλειοψηφία των ατόμων είναι υπέρ του νομοσχεδίου; ($\alpha=0.05$)

3. Ένα απλό τυχαίο δείγμα 290 νοικοκυριών επελέγη από μια περιοχή κάποιας πόλης που περιείχε 14828 νοικοκυριά. Κάθε οικογένεια

ερωτήθηκε αν ενοικίαζε ή όχι το σπίτι στο οποίο έμενε και αν είχε αποκλειστική χρήση λουτρού ή όχι. Τα αποτελέσματα ήταν τα εξής:

	Ιδιοκτήτες	Ενοικιαστές
Αποκλειστικό λουτρό	141	109
κοινό λουτρό	6	34

α) Για την περίπτωση των ενοικιαστών, να εκτιμηθεί το ποσοστό του πληθυσμού με αποκλειστικό λουτρό και να υπολογισθεί το τυπικό σφάλμα της εκτίμησης.

β) Να εκτιμηθεί ο συνολικός αριθμός ενοικιαστών του πληθυσμού με κοινό λουτρό και να υπολογισθεί το τυπικό σφάλμα της εκτίμησης.

4. Αν στην προηγούμενη άσκηση ήταν γνωστό ότι 7526 οικογένειες του πληθυσμού ενοικιάζουν σπίτια, να δοθεί μια νέα εκτίμηση του αριθμού των ενοικιαστών χωρίς αποκλειστικό λουτρό και να εκτιμηθεί το τυπικό σφάλμα της.

5. Σε μια περιοχή 4000 νοικοκυριών πρόκειται να γίνει μια έρευνα για την εκτίμηση του ποσοστού των νοικοκυριών που διατηρούν δυο αυτοκίνητα. Το πραγματικό ποσοστό υπολογίζεται μεταξύ 5 και 10 τοις εκατό. Αν το τυπικό σφάλμα της εκτίμησης δεν πρέπει να υπερβαίνει το 1%, με πιθανότητα 90%, πόσο μεγάλο πρέπει να είναι το δείγμα;

6. Μια δειγματοληπτική έρευνα πρόκειται να γίνει για να εκτιμηθεί το ποσοστό των υπαλλήλων μιας εταιρείας οι οποίοι παραπονούνται ότι ο μισθός τους δεν ανταποκρίνεται στην προσφορά εργασίας τους. Πόσο μεγάλο πρέπει να είναι το δείγμα που θα επιλεγεί από τους 1480 υπαλλήλους της εταιρείας, αν η εκτίμηση δεν πρέπει να διαφέρει από το πραγματικό ποσοστό περισσότερο από 0.02 με πιθανότητα 90%;

7. Πόσο μεγάλο πρέπει να είναι το δείγμα που απαιτείται για την κατασκευή ενός 95% διαστήματος εμπιστοσύνης για την μέση

περιεκτικότητα σε νικοτίνη μιας μάρκας τσιγάρων, αν η περιεκτικότητα σε νικοτίνη έχει την κανονική κατανομή με σ=8.5 mg και το μήκος του διαστήματος πρέπει σύμφωνα με την διαδικασία ελέγχου να είναι ίσο με 6 mg;

8. Πόσα άτομα ενός πληθυσμού μεγέθους 5000 πρέπει να ερωτηθούν σε μια δειγματοληπτική έρευνα ώστε, με 99% πιθανότητα, η εκτίμηση του μέσου ετήσιου εισοδήματος να μην υπερβαίνει το πραγματικό μέσο εισόδημα περισσότερο από 2000 ευρώ αν η τυπική απόκλιση είναι ίση με 5000 ευρώ;