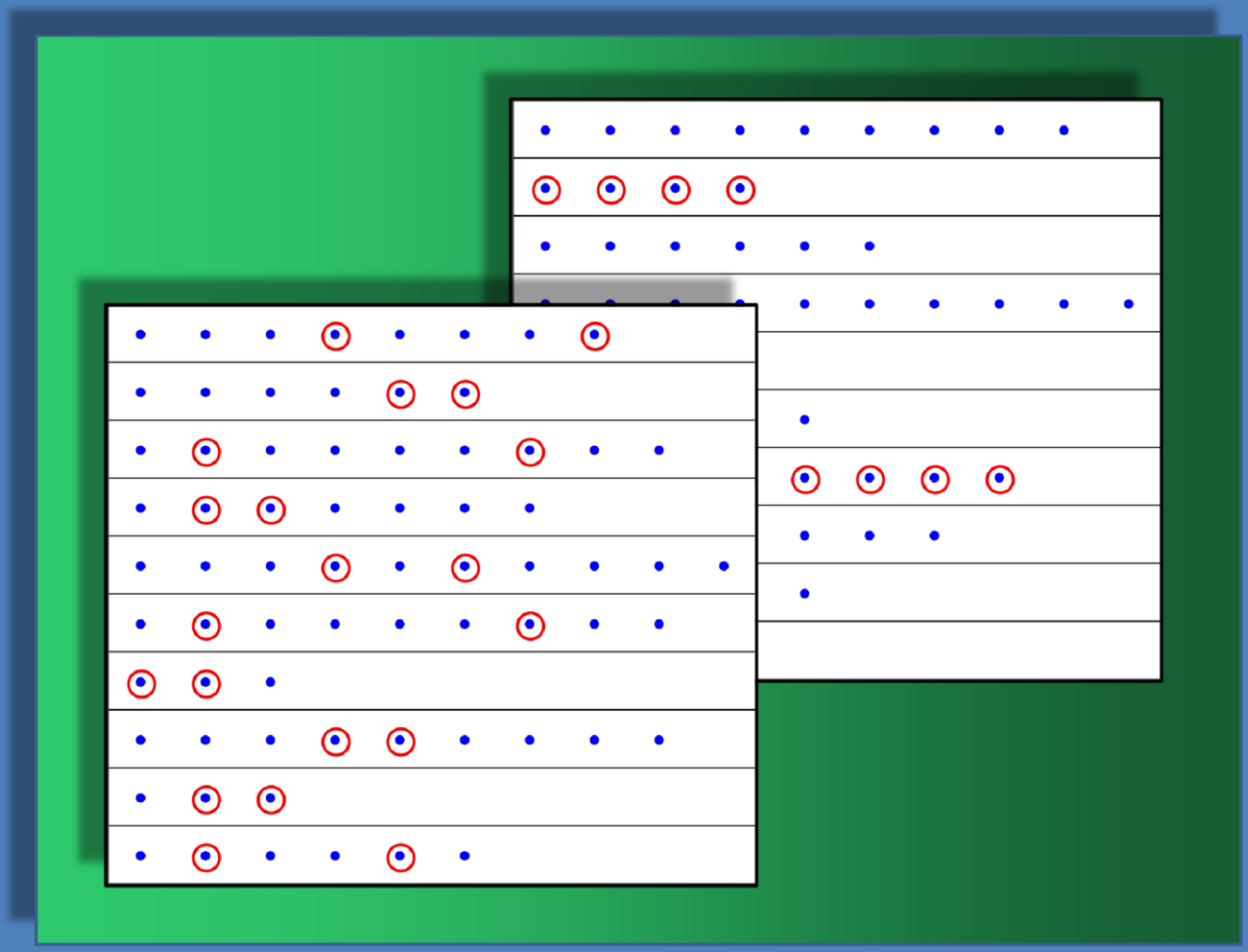
# Θεωρία Δειγματοληψίας

Ιουλία Παπαγεωργίου



Αθήνα 2015

ΙΟΥΛΙΑ ΠΑΠΑΓΕΩΡΓΙΟΥ
Επίκουρη Καθηγήτρια Οικονομικού Πανεπιστημίου Αθηνών

# *Θεωρία Δειγματοληψίας*

# Θεωρία Δειγματοληψίας


***Συγγραφή***

Ιουλία Παπαγεωργίου

***Κριτικός αναγνώστης***

Κωνσταντίνος Ξ. Καρακώστας

***Συντελεστές***

ΓΛΩΣΣΙΚΗ & ΓΡΑΦΙΣΤΙΚΗ ΕΠΙΜΕΛΕΙΑ: Συγγραφική Ομάδα
ΤΕΧΝΙΚΗ ΕΠΕΞΕΡΓΑΣΙΑ: Χρήστος Μουρίκης





ΣΥΝΔΕΣΜΟΣ ΕΛΛΗΝΙΚΩΝ ΑΚΑΔΗΜΑΪΚΩΝ ΒΙΒΛΙΟΘΗΚΩΝ
Εθνικό Μετσόβιο Πολυτεχνείο
Ηρώων Πολυτεχνείου 9, 15780 Ζωγράφου

www.kallipos.gr

ISBN: 978-960-603-275-2

# Πίνακας Περιεχομένων

# Πρόλογος

Σήμερα, άλλος λιγότερο και άλλος περισσότερο, όλοι είμαστε εξοικειωμένοι με την έννοια της δημοσκόπησης. Επίσης, σε αρκετούς από εμάς θα έχει τύχει να μας τηλεφωνήσουν για να συμμετάσχουμε σε κάποια έρευνα. Για να είναι αυτού του είδους οι έρευνες σωστές, και συνεπώς αξιόπιστες, θα πρέπει από την αρχική οργάνωση έως την εξαγωγή και διατύπωση των συμπερασμάτων, να βασίζονται στη Θεωρία της Δειγματοληψίας.

Η Θεωρία της Δειγματοληψίας είναι το κομμάτι εκείνο της στατιστικής που ασχολείται με τις λεγόμενες δειγματοληπτικές έρευνες. Ασχολείται με την οργάνωση και υλοποίηση των ερευνών αυτών, μας δίνει τις έννοιες του πληθυσμού, της απογραφής και του τυχαίου δείγματος, και μελετά τρόπους (π.χ. απλή τυχαία δειγματοληψία, στρωματοποιημένη κτλ.) για την επιλογή ενός τέτοιου δείγματος. Όλοι αυτοί οι τρόποι δειγματοληψίας βασίζονται στις πιθανότητες. Το γεγονός αυτό μας επιτρέπει, τα συμπεράσματα που διατυπώνουμε για το δείγμα, να μπορούμε εύκολα να τα γενικεύσουμε και για τον πληθυσμό από τον οποίο επιλέξαμε το τυχαίο δείγμα.

Ένας βασικός κορμός της θεωρίας της δειγματοληψίας παρουσιάζεται με απλό και σαφή τρόπο στις σελίδες του παρόντος ηλεκτρονικού συγγράμματος. Κάθε κεφάλαιο περιλαμβάνει αριθμητικά παραδείγματα, η ανάλυση των οποίων πραγματοποιείται μέσω της γλώσσας προγραμματισμού `R` (η γλώσσα αυτή είναι δωρεάν, και περισσότερα γι' αυτήν ο αναγνώστης μπορεί να δει στα δύο Παραρτήματα). Οι αποδείξεις είναι κυρίως απλές και μπορούν να παραλειφθούν στην πρώτη ανάγνωση. Οι πιο σύνθετες αποδείξεις δίνονται στο τέλος του κεφαλαίου. Η κατανόηση των αποδείξεων αυτών είναι ένα χρήσιμο εργαλείο για να μπορέσει ο αναγνώστης να κάνει τους δικούς του συνδυασμούς των δειγματοληπτικών σχημάτων, που δεν παρουσιάζονται στο συγκεκριμένο σύγγραμμα, και να είναι σε θέση να βρει τους εκτιμητές και τις ιδιότητες που αντιστοιχούν στον συγκεκριμένο συνδυασμό των προς εκτίμηση παραμέτρων του πληθυσμού. Είναι λάθος να χρησιμοποιούμε ένα σύνθετο δειγματοληπτικό σχέδιο και στο τέλος οι εκτιμητές μας και οι ιδιότητές τους να βασίζονται σε απλά δειγματοληπτικά σχήματα.

Στο σημείο αυτό, θα ήθελα να συγχαρώ την ομάδα του kallipos για την ιδέα που είχαν να δημιουργήσουν μια ηλεκτρονική βάση με συγγράμματα που καλύπτουν διάφορα επιστημονικά πεδία.

Τελειώνοντας, θα πρέπει να εκφράσω τις ευχαριστίες μου στη συγγραφέα, την κ. Ιουλία Παπαγεωργίου, για τη δυνατότητα που μου έδωσε, με την πρόσκλησή της, να είμαι ο κριτικός αναγνώστης του συγκεκριμένου συγγράμματος. Αυτό σημαίνει ότι για οτιδήποτε δεν είναι σωστό στο παρόν σύγγραμμα υπεύθυνος είναι ο κριτικός αναγνώστης.

Κωνσταντίνος Καρακώστας

# Ορολογία – Συντομεύσεις – Ακρωνύμια

| Ελληνικός όρος | | Κεφ | Αγγλικός όρος | |
|---|---|---|---|---|
| | | | | |
| | αθροιστική τετραγωνική ρίζα των συχνοτήτων | 4 | rootfreq | cumulative root frequency method |
| | ακρίβεια | 1 | | Accuracy |
| | αμεροληψία | 1 | | Unbiasedness |
| | αναλογική καταμέριση | 3 | | proportional allocation |
| ΑΝΑΔΙΑ | ΑΝΑλυση της ΔΙΑκύμανσης | 3 | ANOVA | ANalysis Of VAriance |
| | αναμενόμενη τιμή | 1 | | expected value |
| | ανεξάρτητος εκτιμητής λόγου | 7 | | separate ratio estimate |
| | ανεξάρτητος εκτιμητής παλινδρόμησης | 7 | | separate regression estimate |
| | αντικειμενικός πληθυσμός – ή - θεωρητικός πληθυσμός | 1 | | target population<br>theoretical population |
| α.τ.δ. | απλή τυχαία δειγματοληψία | 2 | srs | simple random sampling |
| α.τ. | απλό τυχαίο (δείγμα) | 2 | | simple random (sample) |
| | απογραφή | 1 | | Census |
| | αποτελεσματικός (εκτιμητής) | 1 | | efficient (estimator) |
| | αυτο-σταθμισμένο (δείγμα) | 3 | | self-weighting (sample) |
| | βέλτιστη καταμέριση | 3 | | optimal allocation |
| | βήμα | 5 | | step |
| | βοηθητική μεταβλητή | 4 | | auxiliary variable |
| | δείγμα | 1 | | sample |
| | δείγματα κρίσης | 1 | | judgmental samples |
| | δείγματα ευκολίας | 1 | | accessibility or convenience samples |
| | δείγματα με προκαθορισμένα ποσοστά | 1 | | quota sampling |
| | δείγματα μη-πιθανότητας | 1 | | non-probability samples |
| | δείγματα πιθανότητας | 1 | | probability samples |
| | δείγματα χιονοστιβάδας | 1 | | snowball sampling |

| Ελληνικός όρος | | Κεφ | Αγγλικός όρος | |
|---|---|---|---|---|
| | δειγματικές ποσότητες | 1 | | sample quantities |
| | δειγματική διασπορά | 1 | | sample variance |
| | δειγματική κατανομή | 1 | | sampling distribution |
| | δειγματική μέση τιμή | 1 | | sample mean value |
| | δειγματοληπτική έρευνα | 1 | | sampling survey |
| | δειγματοληπτικό σχέδιο | 1 | | sampling design |
| | δειγματοληπτικό σχέδιο σε πολλά στάδια | 6 | | multistage sampling |
| | δειγματοληψία κατά ομάδες ή συστάδες | 6 | | cluster sampling |
| | δειγματοληψία κατά ομάδες σε δύο στάδια | 6 | | two-stage cluster sampling |
| | δειγματοληψία κατά ομάδες σε ένα στάδιο | 6 | | one-stage cluster sampling |
| | δειγματοληψία με πιθανότητες ανάλογες του εκτιμώμενου μεγέθους | 6 | ppes | probabilities proportional to estimated size sampling |
| | δειγματοληψία με πιθανότητες ανάλογες του μεγέθους | 6 | pps | probabilities proportional to size |
| | δειγματοληψία με προκαθορισμένα ποσοστά | 4 | | quota sampling |
| | δειγματοληψία χωρίς επανατοποθέτηση | 2 | swr | sampling without replacement |
| | δευτερογενής δειγματοληπτική μονάδα | 6 | ssu | secondary sampling unit |
| | δημοσκόπηση | 1 | | opinion poll |
| | διακύμανση ή διασπορά | 2 | Var | variance |
| ΔΕ | διάστημα εμπιστοσύνης | 2 | CI | confidence interval |
| | διεξαγωγή της έρευνας | 1 | | survey operations |
| | διόρθωση πεπερασμένου πληθυσμού | 2 | fpc | finite population correction |
| | εκ των υστέρων στρωματοποιημένη δειγματοληψία | 4 | | post-stratified sampling design |
| | εκτιμητής | 1 | | Estimator |
| | εκτιμητής λόγου | 6 | | ratio estimator |

| Ελληνικός όρος | | Κεφ | Αγγλικός όρος | |
|---|---|---|---|---|
| | εκτιμητής παλινδρόμησης | 7 | | regression estimate |
| | ερωτηματολόγιο | 1 | | questionnaire |
| | Θεωρία Δειγματοληψίας | 1 | | Sampling Theory |
| | κεντρικά τοποθετημένη συστηματική δειγματοληψία | 5 | | centrally located systematic sampling |
| | μέγεθος δείγματος | 1 | | sample size |
| | μέγεθος του πληθυσμού | 1 | | population size |
| | μέση τιμή | 1 | | mean value |
| | μέσο τετραγωνικό σφάλμα | 1 | | mean square error |
| | μετρήσεις της έρευνας | 1 | | survey measurements |
| | πεπερασμένος πληθυσμός | 1 | | finite population |
| | πηλίκο δείγματος | 3 | f | sampling fraction |
| | πιθανότητες συμπερίληψης στο δείγμα πρώτης και δεύτερης τάξης | 6 | | first and second order inclusion probabilities |
| | πιλοτική έρευνα | 1 | | pilot study |
| | πληθυσμιακές ποσότητες - ή - πληθυσμιακά μεγέθη – ή - παράμετροι του πληθυσμού | 1 | | population quantities, population parameters |
| | πολλαπλή γραμμική παλινδρόμηση | 7 | | multiple linear regression |
| | ποσό μεροληψίας | 1 | | bias |
| | ποσοστό | 1 | | percentage |
| | πρωτογενής δειγματοληπτική μονάδα | 6 | psu | primary sampling unit |
| | στατιστική ανάλυση | 1 | | statistical analysis |
| | στατιστική συνάρτηση (σ.σ.) - ή - στατιστικό | 1 | | statistical function<br>statistic |
| | στατιστικό λάθος | 1 | | statistical error |
| | στρώματα | 3 | | strata |
| | στρωματοποιημένη | 3 | st | stratified |
| | στρωματοποιημένη με τυχαία προκαθορισμένα ποσοστά | 3 | | stratified with random quotas |
| | συμμεταβλητότητα ή συνδιακύμανση | 1 | | covariance |

| Ελληνικός όρος | | Κεφ | Αγγλικός όρος | |
|---|---|---|---|---|
| | συνδυασμένος ή από κοινού εκτιμητής λόγου | 7 | | combined ratio estimate |
| | συνδυασμένος ή από κοινού εκτιμητής παλινδρόμησης | 7 | | combined regression estimate |
| | σύνολο | 1 | | population total |
| | συντελεστής μεταβλητότητας | 2 | CV | coefficient of variation |
| | συντελεστής συσχέτισης intracluster | 5 | | intracluster correlation |
| | συστηματική δειγματοληψία | 5 | | systematic sampling |
| | συσχέτιση | 1 | | correlation |
| | σφάλματα μη-κάλυψης | 1 | | non-coverage error |
| τ.σ. | τυπικό σφάλμα | 2 | se | standard error |
| τ.μ. | τυχαία μεταβλητή | 2 | rv | random variable |
| | τυχαία συστηματική δειγματοληψία | 5 | | random systematic sampling |
| | τυχαίο στρωματοποιημένο | 3 | | random stratified |
| | υπό μελέτη πληθυσμός - ή - δειγματοληπτικό πλαίσιο | 1 | | sampling frame |
| | χαρακτηριστικό ενός πληθυσμού | 1 | | population characteristic |
| | | | | |

*Σημειώσεις*:

(i) Στη στήλη «Κεφ» αναγράφεται το Κεφάλαιο στο οποίο συναντάται για πρώτη φορά ο όρος.

(ii) Όταν δεν υπάρχει καθιερωμένη ελληνική συντομογραφία, συνήθως χρησιμοποιείται η αντίστοιχη αγγλική (π.χ. Var για τη διακύμανση).

# Κεφάλαιο 1 - ΕΙΣΑΓΩΓΗ

**Σύνοψη**

*Στο κεφάλαιο αυτό, περιγράφονται τα κυριότερα στάδια μιας δειγματοληπτικής έρευνας και δίνεται μια εισαγωγή της Θεωρίας Δειγματοληψίας. Οι δειγματοληπτικές έρευνες βασίζονται σε δεδομένα που καταγράφονται για ένα υποσύνολο του πληθυσμού, το δείγμα (sample), και έχουν ως στόχο την εξαγωγή συμπερασμάτων που αφορούν άγνωστες παραμέτρους του πληθυσμού. Τόσο η χρήση, όσο και η θεματολογία τους, είναι ευρύτατη. Οι μέθοδοι δειγματοληψίας χωρίζονται σε μεθόδους πιθανότητας (probability sampling) και μεθόδους μη-πιθανότητας (non-probability sampling). Για τα δείγματα που επιλέγονται σύμφωνα με μια μέθοδο πιθανότητας, η εξαγωγή συμπερασμάτων και η επέκτασή τους στον πληθυσμό είναι εφικτή κάνοντας χρήση της Στατιστικής συμπερασματολογίας. Για τα δεδομένα που συλλέγονται σύμφωνα με ένα δείγμα μη-πιθανότητας, είναι εφικτό να εφαρμοστεί περιγραφική στατιστική ώστε να παρουσιαστούν τα αποτελέσματα της έρευνας, αλλά δεν μπορεί να γίνει επέκταση των αποτελεσμάτων για τον πληθυσμό. Συνοπτικά, η Θεωρία Δειγματοληψίας (Sampling Theory) είναι το πεδίο της στατιστικής που περιλαμβάνει τη μεθοδολογία (α) για την επιλογή ενός δείγματος από έναν μεγαλύτερο πληθυσμό και (β) για την εξαγωγή αξιόπιστων συμπερασμάτων για τα ερωτήματα της έρευνας που θα έχουν ισχύ για τον πληθυσμό. Οι πιο γνωστές από τις μεθόδους δειγματοληψίας που βασίζονται στις πιθανότητες είναι η απλή τυχαία (simple random), η συστηματική (systematic), η στρωματοποιημένη (stratified) και η δειγματοληψία κατά ομάδες (cluster).*

## 1.1. Δειγματοληπτική Έρευνα

Με την ανάπτυξη των υπολογιστών, του διαδικτύου (internet) και των μέσων μαζικής ενημέρωσης και επικοινωνίας, η πρόσβαση σε δεδομένα είναι πολύ πιο εύκολη απ'ό,τι παλιότερα. Τα δεδομένα αυτά προέρχονται ή αφορούν ζητήματα γύρω από κάθε τομέα της προσωπικής, κοινωνικής ή επαγγελματικής ζωής μας. Ο μεγάλος όγκος των δεδομένων καθιστά αναγκαία την έκφρασή τους με ποσοτικά στοιχεία και γραφήματα, με σκοπό την καλύτερη οργάνωση, παρουσίαση και κατανόησή τους, και την εξαγωγή συμπερασμάτων. Η ανάγκη για την ύπαρξη ποσοτικών συμπερασμάτων για διάφορα θέματα της καθημερινότητας είναι το αντικείμενο των δειγματοληπτικών ερευνών.

Χαρακτηριστικό παράδειγμα δειγματοληπτικών ερευνών είναι οι *δημοσκοπήσεις (opinion polls)*, όπου επιλέγεται ένα δείγμα από ένα σύνολο ανθρώπων και όσοι το αποτελούν καλούνται να απαντήσουν εκφράζοντας την προσωπική τους άποψη ή προτίμηση γύρω από ένα θέμα, π.χ. εάν είναι ικανοποιημένοι από το σύστημα υγείας, σε τι βαθμό προτιμούν ένα καταναλωτικό προϊόν, ποιο πολιτικό κόμμα τούς εκφράζει περισσότερο κτλ. Οι δημοσκοπήσεις χρησιμοποιούνται με μεγάλη συχνότητα σε έρευνες μαρκετιγκ, διαφήμισης και έρευνες πολιτικού χαρακτήρα.

Οι δειγματοληπτικές έρευνες μπορεί να αφορούν και πληθυσμούς που δεν έχουν ως μέλη τους ανθρώπους. Π.χ. μια έρευνα που έχει ως στόχο να απαντήσει στο ερώτημα πόσα στρέμματα καλλιεργήσιμης γης μιας χώρας καλλιεργούνται με σιτάρι, ή ποιο ποσοστό των προϊόντων από μια γραμμή παραγωγής είναι ελαττωματικά.

Άλλα παραδείγματα δειγματοληπτικών ερευνών είναι οι έρευνες που διεξάγονται σε επίπεδο χωρών ή μικρότερων διοικητικών μονάδων, με σκοπό την εκτίμηση π.χ. του ποσοστού ανεργίας των κατοίκων και πώς αυτό μεταβάλλεται ανά μήνα, του ποσοστού των κατοίκων που ολοκληρώνουν την εκπαίδευση σε κάθε εκπαιδευτική βαθμίδα (βασική, υποχρεωτική, ανώτερη, ανώτατη), του μεγέθους του πληθωρισμού, του ποσοστού των νοικοκυριών της χώρας που δεν έχουν κεντρική θέρμανση κτλ.

Οι δειγματοληπτικές έρευνες που θα μας απασχολήσουν στη συνέχεια αφορούν έρευνες που έχουν ως στόχο τη μελέτη του πληθυσμού σε μια συγκεκριμένη χρονική στιγμή (χωρίς διάρκεια) και από τις οποίες ενδιαφερόμαστε να βγάλουμε συμπεράσματα για άγνωστες ποσότητες του πληθυσμού. Οι έρευνες αυτές λέγονται cross-sectional και ιδανικά αποτελούν ένα στιγμιότυπο (snapshot) του πληθυσμού τη στιγμή της έρευνας.

Αντίθετα, οι cohort δειγματοληπτικές έρευνες είναι έρευνες οι οποίες στοχεύουν στη μελέτη ενός χαρακτηριστικού του πληθυσμού στη διάρκεια του χρόνου, και κατά πόσο ή σε ποιο βαθμό επηρεάζεται το υπό μελέτη χαρακτηριστικό από άλλους παράγοντες.

Η ανάγκη για γρήγορη εξαγωγή αποτελεσμάτων και για σύγκριση αυτών σε τακτά χρονικά διαστήματα καθιστά τη χρήση των δειγματοληπτικών ερευνών επιβεβλημένη σε σχέση με την εναλλακτική επιλογή που είναι η απογραφή του πληθυσμού. Στην παράγραφο που ακολουθεί, δίνονται αναλυτικά οι ορισμοί του πληθυσμού, της απογραφής και του δείγματος (βλ. Groves et al, 2009).

## 1.2. Πληθυσμός – Δείγμα

Όλα τα παραδείγματα που δόθηκαν στην προηγούμενη παράγραφο έχουν ως κοινό γνώρισμα ότι αναφέρονται σε ένα σύνολο με συγκεκριμένο πλήθος μελών, π.χ. οι κάτοικοι μιας χώρας, το σύνολο των νοικοκυριών της χώρας κτλ. Το σύνολο των μελών που είναι πεπερασμένου, και όχι άπειρου, πλήθους ονομάζεται *πεπερασμένος πληθυσμός (finite population)*. Σε όλα τα κεφάλαια που ακολουθούν, υποθέτουμε ότι ο πληθυσμός είναι πεπερασμένος.

### Ορισμός 1.1

*Αντικειμενικός πληθυσμός (target population)* είναι το σύνολο των μελών για το οποίο ενδιαφερόμαστε να εξαγάγουμε κάποια συμπεράσματα.

### Παράδειγμα 1.1

Για τη δειγματοληπτική έρευνα που στοχεύει στον υπολογισμό του ποσοστού ανεργίας των κατοίκων μιας χώρας, ο αντικειμενικός πληθυσμός είναι όλοι οι κάτοικοι της χώρας (άνω των 18 ετών, οι οποίοι μπορούν να εργαστούν).

### Ορισμός 1.2

Η έρευνα που βασίζεται στη μελέτη όλων των στοιχείων ενός πληθυσμού, ονομάζεται *απογραφή (census)*.

Στην περίπτωση της απογραφής, το ερώτημα (ή τα ερωτήματα) της έρευνας τίθεται σε κάθε μέλος του πληθυσμού. Τα στοιχεία που θα συγκεντρωθούν με το τέλος της έρευνας χρησιμοποιούνται για τον υπολογισμό συγκεντρωτικών ποσοτήτων (π.χ. μέση τιμή) του πληθυσμού. Τα δεδομένα μιας απογραφής προσφέρουν τη δυνατότητα έγκυρων αποτελεσμάτων για τα ερωτήματα της έρευνας, δηλαδή αποτελεσμάτων που βασίζονται στην πλήρη πληροφορία για τον πληθυσμό και δεν εμπεριέχουν 'στατιστικό λάθος' (statistical error). Η έκφραση 'στατιστικό λάθος' ή 'αβεβαιότητα', που αναφέρεται συχνά σε δημοσιεύσεις αποτελεσμάτων δειγματοληπτικών ερευνών, θα διευκρινιστεί επακριβώς στη συνέχεια του κεφαλαίου. Παρόλο το πλεονέκτημα της απογραφής ως προς την ακρίβεια των αποτελεσμάτων, είναι προφανές ότι η διεξαγωγή μιας τέτοιας διαδικασίας είναι χρονοβόρα υπόθεση, με μεγάλο κόστος σε σχεδιασμό, οργάνωση και υλοποίηση. Επίσης, λόγω αυτών των χαρακτηριστικών, η απογραφή δεν είναι εφικτό να πραγματοποιείται σε πυκνά χρονικά διαστήματα, και κατά συνέπεια δεν είμαστε σε θέση να διαπιστώνουμε κατά πόσον οι ποσότητες του πληθυσμού που μας ενδιαφέρουν μεταβάλλονται στην πάροδο του χρόνο. Π.χ. εάν ενδιαφερόμαστε για το ποσοστό ανεργίας σε μια μεγάλη πόλη ή χώρα, ο υπολογισμός του βάσει απογραφής κάθε 2 μήνες για παράδειγμα, θα ήταν ανέφικτος. Επίσης, ένα άλλο πρακτικό πρόβλημα είναι ότι για την υλοποίηση μιας απογραφής είναι απαραίτητο να υπάρχει μια λίστα που περιλαμβάνει ένα προς ένα όλα τα μέλη του πληθυσμού.

Εναλλακτικά της απογραφής, η έρευνα βασίζεται σε ένα υποσύνολο του πληθυσμού που λέγεται *δείγμα (sample)* και κάνουμε λόγο για *δειγματοληπτική έρευνα (sampling survey)*. Κατά τη διεξαγωγή μιας δειγματοληπτικής έρευνας, τα μέλη του πληθυσμού που αντιστοιχούν στο δείγμα εντοπίζονται και καλούνται να απαντήσουν στο ερώτημα ή τα ερωτήματα της έρευνας. Τα στοιχεία που συλλέγονται κατά τη διεξαγωγή μιας δειγματοληπτικής έρευνας χρησιμοποιούνται και αυτά, όπως και της απογραφής, για τον υπολογισμό συγκεντρωτικών ποσοτήτων, όχι όμως του πληθυσμού, αλλά των αντίστοιχων ποσοτήτων στο δείγμα. Η διατύπωση οποιουδήποτε συμπεράσματος για ποσότητες του πληθυσμού, με βάση τις αντίστοιχες του δείγματος, απαιτεί την ανάπτυξη κατάλληλης μεθοδολογίας. Λόγω της επιλογής μέρους μόνο του πληθυσμού,

τα αποτελέσματα μιας δειγματοληπτικής έρευνας εμπεριέχουν μια ‘τυχαιότητα’ ή ‘αβεβαιότητα’, σε αντίθεση με την απογραφή.

Στα προφανή πλεονεκτήματα της δειγματοληπτικής έρευνας περιλαμβάνονται το μικρό κόστος, η σύντομη διάρκεια της συλλογής δεδομένων και η επαναληψιμότητα της διαδικασίας, δηλ. τα τρία στοιχεία που αποτελούν τα βασικά μειονεκτήματα της απογραφής. Γενικά, κατά τον σχεδιασμό και τη διεξαγωγή μιας δειγματοληπτικής έρευνας, γίνεται προσπάθεια να επιτευχθεί μείωση του κόστους και του χρόνου διεξαγωγής της έρευνας, με ταυτόχρονο έλεγχο της αβεβαιότητας που θα εμπεριέχεται στα αποτελέσματα. Η ισορροπία μεταξύ αυτών των δύο συνιστωσών εξαρτάται από πολλούς παράγοντες και θα είναι ένα από τα αντικείμενα που θα μας απασχολήσουν σε επόμενα κεφάλαια.

Τέλος, μπορούμε να προσθέσουμε ένα ακόμα πλεονέκτημα της δειγματοληπτικής έρευνας. Επειδή ο αριθμός των συμμετεχόντων στην έρευνα είναι μικρότερος σε σύγκριση με την απογραφή, ο όγκος των δεδομένων που έχουμε να χειριστούμε είναι μικρότερος, με αποτέλεσμα να ελαχιστοποιούνται σφάλματα τα οποία σχετίζονται με την καταγραφή, την κωδικοποίηση και την επεξεργασία των δεδομένων, και στα οποία είναι πιθανόν να υποπέσουμε. Ένα άλλο πρακτικό πλεονέκτημα της δειγματοληπτικής έρευνας έναντι της απογραφής, είναι ότι, σε ορισμένες περιπτώσεις, η καταγραφή μιας μέτρησης για μια μονάδα του πληθυσμού συνεπάγεται αλλοίωση ή καταστροφή της μονάδας. Παράδειγμα, η διάρκεια ζωής ενός ηλεκτρικού λαμπτήρα από μια γραμμή παραγωγής λαμπτήρων, ή ο έλεγχος της ικανοποιητικής ανάπτυξης των ριζών ενός φυτού που καλλιεργείται σε ένα φυτώριο. Στις περιπτώσεις αυτές, η δειγματοληψία αποτελεί μοναδική λύση και όχι απλώς εναλλακτική της απογραφής.

Απ’ την άλλη πλευρά, ενώ ο όγκος των δεδομένων για μια δειγματοληψία είναι μικρότερος σε σύγκριση με την απογραφή, η υιοθέτηση της δειγματοληπτικής έρευνας ως επιλογής είναι πιο απαιτητική σε ό,τι αφορά το μαθηματικό και στατιστικό υπόβαθρο που είναι απαραίτητο προκειμένου (i) να επιλεγούν οι μονάδες του πληθυσμού που θα αποτελούν το δείγμα και (ii) να εξαχθούν αξιόπιστα αποτελέσματα για τα ερωτήματα της έρευνας, που θα έχουν ισχύ για τον πληθυσμό. Τα δύο τελευταία αυτά χαρακτηριστικά της δειγματοληπτικής έρευνας αποτελούν και το αντικείμενο της *Θεωρίας Δειγματοληψίας (Sampling Theory)*.

Δίνουμε στη συνέχεια έναν ακόμα ορισμό για τον πληθυσμό εκτός του αντικειμενικού, ο οποίος συνδέεται με την επιλογή της δειγματοληπτικής έρευνας έναντι της απογραφής για την εξαγωγή των αποτελεσμάτων.

### Ορισμός 1.3

*Υπό μελέτη πληθυσμός* ή *δειγματοληπτικό πλαίσιο (sampling frame)* είναι το σύνολο των μελών στα οποία θα βασιστούμε προκειμένου να υλοποιήσουμε τη δειγματοληπτική έρευνα.

## Παράδειγμα 1.2

Για τη δειγματοληπτική έρευνα του Παραδείγματος 1.1, που στοχεύει στον υπολογισμό του ποσοστού ανεργίας των κατοίκων μιας χώρας, εάν η έρευνα διεξαχθεί επιλέγοντας τυχαία τηλεφωνικούς αριθμούς από τους τηλεφωνικούς καταλόγους της χώρας, τότε ο υπό μελέτη πληθυσμός είναι όλοι οι κάτοικοι της χώρας (άνω των 18 ετών, οι οποίοι μπορούν να εργαστούν) και οι οποίοι έχουν τηλέφωνο.

## Παράδειγμα 1.3

Εάν ενδιαφερόμαστε για τον υπολογισμό του ποσοστού των παιδιών μιας πόλης, ηλικίας 6 ετών, τα οποία δεν έχουν κάνει τα βασικά εμβόλια, τότε:

(i) Στην περίπτωση που η έρευνα γίνει εντοπίζοντας πρώτα όλα τα παιδιά ηλικίας 6 ετών της πόλης, τα οποία είναι εγγεγραμμένα στα δημοτολόγια της πόλης, και επιλέγοντας ένα δείγμα, τότε ο *υπό μελέτη πληθυσμός* είναι τα παιδιά 6 ετών της πόλης που είναι εγγεγραμμένα στα δημοτολόγια.

(ii) Εάν η έρευνα διεξαχθεί επιλέγοντας ένα υποσύνολο Δημοτικών Σχολείων της πόλης, και τα παιδιά της Πρώτης Τάξης των Σχολείων αυτών συμπεριληφθούν στο δείγμα, *ο υπό μελέτη* πληθυσμός είναι τα παιδιά 6 ετών της πόλης που φοιτούν σε ένα από τα Δημοτικά Σχολεία.

Ο *αντικειμενικός πληθυσμός* για το παράδειγμά μας είναι τα παιδιά της πόλης ηλικίας 6 ετών.

Γενικότερα, ενώ ο αντικειμενικός πληθυσμός είναι μονοσήμαντα ορισμένος, ο υπό μελέτη πληθυσμός ενδέχεται να έχει πολλούς ορισμούς, ακόμα κι αν αναφερόμαστε στην ίδια έρευνα. Αυτό συμβαίνει γιατί ο υπό μελέτη πληθυσμός συνδέεται με τον τρόπο υλοποίησης της έρευνας και κατά συνέπεια θα διαφέρει, δηλ. θα ορίζεται διαφορετικά, ανάλογα με τη μέθοδο επιλογής του δείγματος και με τη μέθοδο πρόσβασης στα μέλη του συνόλου για την επικοινωνία και την καταγραφή της μέτρησης.

Ο υπό μελέτη πληθυσμός είναι στη γενική περίπτωση ένα υποσύνολο του αντικειμενικού πληθυσμού. Στο Παράδειγμα 1.2, ο υπό μελέτη πληθυσμός δεν συμπεριλαμβάνει τους κατοίκους της χώρας που δεν έχουν τηλέφωνο. Στο Παράδειγμα 1.3(ii), ο υπό μελέτη πληθυσμός δεν συμπεριλαμβάνει τα παιδιά της πόλης που δεν φοιτούν σε κάποιο από τα σχολεία. Προφανώς, όσο πιο κοντά είναι ο υπό μελέτη πληθυσμός στον αντικειμενικό πληθυσμό, τόσο καλύτερο είναι το δειγματοληπτικό πλαίσιο, γιατί αποφεύγονται τα *σφάλματα μη-κάλυψης (non-coverage error)*.

Ο αντικειμενικός πληθυσμός πολλές φορές αναφέρεται και ως *θεωρητικός πληθυσμός (theoretical population)*, γιατί είναι το σύνολο των μελών στο οποίο θεωρητικά στοχεύουμε. Αντίθετα, ο υπό μελέτη πληθυσμός, όπως είδαμε, είναι ο πληθυσμός που χρησιμοποιούμε στην πράξη.

## 1.3. Δείγματα πιθανότητας και δείγματα μη-πιθανότητας

Ανάλογα με τον μηχανισμό επιλογής των μονάδων του πληθυσμού στο δείγμα, ο οποίος ονομάζεται και *δειγματοληπτικό σχέδιο (sampling design)*, τα δείγματα χωρίζονται αρχικά σε δύο μεγάλες κατηγορίες. Τα *δείγματα πιθανότητας (probability samples)* και τα *δείγματα μη-πιθανότητας (non-probability samples)*. Αντίστοιχα, η δειγματοληψία ονομάζεται δειγματοληψία πιθανότητας και δειγματοληψία μη-πιθανότητας.

### 1.3.1 Δείγματα Πιθανότητας

Ένα δείγμα λέγεται δείγμα πιθανότητας όταν η κάθε μονάδα του πληθυσμού έχει μια πιθανότητα, συγκεκριμένη και μη-μηδενική, να συμπεριληφθεί στο δείγμα. Η πιθανότητα αυτή είναι προκαθορισμένη πριν από την επιλογή του δείγματος.

Συνεπώς, σύμφωνα με τα δείγματα πιθανότητας, η μέθοδος δειγματοληψίας δεν αποκλείει κάποιες μονάδες του πληθυσμού από το ενδεχόμενο να είναι μέρη του δείγματος. Επιπλέον, η προκαθορισμένη, και κατά συνέπεια γνωστή, πιθανότητα επιλογής της κάθε μονάδας του πληθυσμού συνεπάγεται ή εγγυάται ότι στη διαδικασία επιλογής του δείγματος υπεισέρχεται ο παράγοντας της τυχαιότητας.

### Παράδειγμα 1.4

Έστω ότι ενδιαφερόμαστε για τη γνώμη (θετική ή αρνητική) 200 δημοτών για μια απόφαση του δημοτικού συμβουλίου της πόλης τους.

Εάν επιλέξουμε τυχαία 200 άτομα από το δημοτολόγιο της πόλης και επικοινωνήσουμε στη συνέχεια μαζί τους, π.χ. μέσω ταχυδρομείου, τότε το δείγμα είναι ένα δείγμα πιθανότητας. Το κάθε μέλος τού υπό μελέτη πληθυσμού, ο κάθε δημότης στην προκειμένη περίπτωση, έχει μια προκαθορισμένη πιθανότητα να ανήκει στο δείγμα. Ειδικότερα, για το παράδειγμά μας, οι πιθανότητες επιλογής είναι ίσες μεταξύ των μελών του πληθυσμού, και συγκεκριμένα ίσες με 1 προς το πλήθος των εγγεγραμμένων δημοτών.

### Παράδειγμα 1.5

Έστω ότι το θέμα της έρευνας αφορά την ίδρυση παιδικών σταθμών, και συνεπώς ενδιαφέρει περισσότερο τους γονείς που έχουν παιδιά σε προσχολική ηλικία. Θα μπορούσαμε να επιλέξουμε 150 άτομα από τη συγκεκριμένη αυτή κατηγορία και 50 άτομα από όλους τους υπόλοιπους. Το δείγμα αυτό παραμένει ένα δείγμα πιθανότητας, όπου οι πιθανότητες είναι άνισες μεταξύ των δύο κατηγοριών των δημοτών της πόλης, αλλά είναι γνωστές και προκαθορισμένες. Πιο συγκεκριμένα, οι πιθανότητες εδώ προσδιορίζονται από τα πλήθη, έστω $N_A$ και $N_B$, των δημοτών που ανήκουν στην πρώτη και τη δεύτερη κατηγορία, αντίστοιχα. Η

πιθανότητα που θα έχει ο κάθε δημότης με παιδί σε προσχολική ηλικία να επιλεγεί για το δείγμα θα είναι $1/N_A$ ενώ η πιθανότητα για τους δημότες χωρίς παιδιά αυτής της ηλικίας θα είναι $1/N_B$.

Και στα δύο παραδείγματα υπάρχει ο παράγοντας της τυχαιότητας.

Εάν, προκειμένου να διεξάγουμε την ίδια έρευνα με αυτή του Παραδείγματος 1.4, αποφασίζαμε να συμπληρώσουμε το δείγμα των 200 δημοτών ρωτώντας τους πρώτους 200 που θα περάσουν από ένα συγκεκριμένο σημείο ενός πολυσύχναστου δρόμου, το δείγμα δεν είναι δείγμα πιθανότητας. Αυτό συμβαίνει γιατί σύμφωνα με την παραπάνω διαδικασία επιλογής, ο κάθε δημότης δεν έχει μια θετική, μη-μηδενική πιθανότητα επιλογής. Οι δημότες που δεν θα περάσουν από το σημείο αυτό του δρόμου θα έχουν μηδενική πιθανότητα επιλογής.

Οι πιο γνωστές μέθοδοι δειγματοληψίας που ανήκουν στην κατηγορία των δειγματοληψιών πιθανότητας είναι οι:

- Απλή τυχαία,
- Στρωματοποιημένη,
- Συστηματική,
- Δείγματα με άνισες πιθανότητες - Δειγματοληψία ανάλογα του μεγέθους,
- Κατά ομάδες ή σε πολλά στάδια,
- Συνδυασμός των παραπάνω.

Στα κεφάλαια που ακολουθούν, γίνεται μελέτη της κάθε μιας από τις παραπάνω μεθόδους, αναλυτικά.

Το μεγάλο πλεονέκτημα των δειγμάτων πιθανότητας είναι ότι επειδή η διαδικασία επιλογής του δείγματος διέπεται από τους νόμους των πιθανοτήτων, είναι εφικτή η επέκταση των όποιων συμπερασμάτων από το δείγμα στον πληθυσμό, με τη βοήθεια της στατιστικής συμπερασματολογίας. Ορισμένα στοιχεία της περιοχής της στατιστικής συμπερασματολογίας, θα δώσουμε στην παράγραφο 1.5.

### 1.3.2 Δείγματα Μη-Πιθανότητας

Τα δείγματα μη-πιθανότητας είναι στον αντίποδα των δειγμάτων πιθανότητας. Η μέθοδος επιλογής των μονάδων του δείγματος δεν διέπεται από τους νόμους της πιθανότητας, αλλά βασίζεται σε κριτήρια όπως η ευκολία, η εύκολη πρόσβαση, η διαθεσιμότητα, ο σύντομος χρόνος συλλογής των δεδομένων κτλ. Τα κριτήρια αυτά δεν εξασφαλίζουν μια θετική και προκαθορισμένη πιθανότητα επιλογής στο δείγμα για το κάθε μέλος του πληθυσμού. Αντίθετα η επιλογή ή μη των μελών του πληθυσμού στο δείγμα γίνεται με βεβαιότητα.

Τα βασικότερα είδη δειγματοληψίας που ανήκουν στην κατηγορία των δειγμάτων μη-πιθανότητας είναι:

#### Δείγματα Ευκολίας (accessibility or convenience samples)

Είναι τα δείγματα όπου οι δειγματοληπτικές μονάδες επιλέγονται από τον πληθυσμό με κριτήριο την ευκολία, και όχι την τυχαιότητα ή την επιδίωξη της αντιπροσωπευτικότητας του πληθυσμού.

Ένα παράδειγμα δείγματος ευκολίας είναι όταν το δείγμα συλλέγεται μέσω μιας εφαρμογής στο διαδίκτυο όπου όποιος θέλει να συμμετέχει στην έρευνα και να απαντήσει στην ερώτηση μπορεί να το κάνει μόνος του. Επίσης, στο Παράδειγμα 1.4, το δείγμα που συλλέγεται ρωτώντας τους 200 πρώτους που θα περάσουν από το σταθερό σημείο όπου βρίσκεται ο συνεντεύκτης, είναι ένα δείγμα ευκολίας. Και στις δύο περιπτώσεις, υπάρχει έντονος ο κίνδυνος των μεροληπτικών αποτελεσμάτων. Στη μεν πρώτη περίπτωση, η συμμετοχή στην έρευνα γίνεται με πρωτοβουλία του μέλους του πληθυσμού, και όχι με τυχαιότητα, άρα η συμμετοχή ή όχι εξαρτάται από το αν το άτομο έχει θετική ή όχι άποψη για το θέμα της έρευνας. Επίσης, αποκλείονται όλοι όσοι, είτε δεν έχουν πρόσβαση στο διαδίκτυο, είτε δεν είδαν τη συγκεκριμένη εφαρμογή της έρευνας.

Στη δεύτερη περίπτωση, η μεροληψία είναι υπρκτή γιατί, για παράδειγμα, αποκλείονται τα μέλη του πληθυσμού που λόγω ηλικίας, ή επειδή εργάζονται, δεν διέρχονται από τον συγκεκριμένο δρόμο την ώρα της έρευνας.

Γενικότερα, στα δείγματα ευκολίας δεν είναι εφικτό να εξαχθούν αποτελέσματα τα οποία στη συνέχεια θα γενικευτούν για τον πληθυσμό, γιατί το δείγμα δεν είναι αντιπροσωπευτικό. Η ίδια αυτή παρατηρήση ισχύει για όλα τα δείγματα μη-πιθανότητας συνολικά.

### Δείγματα Κρίσης (Judgmental samples)

Στα δείγματα κρίσης ο ερευνητής επιλέγει τις μονάδες του πληθυσμού με βάση την προσωπική του κρίση, ή την εμπειρία του από προηγούμενες έρευνες με παρόμοιο θέμα στο ίδιο σύνολο πληθυσμού. Για παράδειγμα, σε δημοσκοπήσεις με στόχο πολιτικές έρευνες, αποτελέσματα εκλογών κτλ., ο ερευνητής μπορεί να επιλέξει το δείγμα του συμπεριλαμβάνοντας με βεβαιότητα περιοχές του πληθυσμού που έχουν ιδιαίτερα χαρακτηριστικά, π.χ. στις πιο πρόσφατες εκλογές, τα εκλογικά αποτελέσματα των περιοχών αυτών ήταν πολύ κοντά στα τελικά αποτελέσματα όλης της επικράτειας. Οι περιοχές αυτές αποκαλούνται «περιοχές βαρόμετρο» και επιλέγονται με βεβαιότητα στο δείγμα, γιατί βάσει της εμπειρίας από προηγούμενες εκλογές θεωρούνται αντιπροσωπευτικές.

### Δείγματα Χιονοστιβάδας (Snowball sampling)

Στα δείγματα χιονοστιβάδας, το δείγμα γίνεται προσβάσιμο στον ερευνητή μέσω ενός μικρού αρχικού συνόλου δείγματος που είναι διαθέσιμο σε εκείνον. Η κάθε μιά δειγματοληπτική μονάδα του αρχικού δείγματος προσφέρει τα στοιχεία, και άρα την πρόσβαση, σε ένα σύνολο από άλλα μέλη του πληθυσμού, τα οποία συμπεριλαμβάνονται στο δείγμα, και τα οποία με τη σειρά τους προσφέρουν πρόσβαση σε ένα άλλο σύνολο κοκ. Ο στατιστικός αναλυτής, συνεπώς, αποκτά το δείγμα μέσω των αρχικών εκπροσώπων, χωρίς προσπάθεια εντοπισμού και χωρίς ανάγκη να διαθέτει στοιχεία για τον πληθυσμό (λίστα μελών, πλήθος κτλ).

### Δείγματα με προκαθορισμένα ποσοστά (Quota sampling)

Σύμφωνα με αυτή τη μέθοδο δειγματοληψίας, ο στατιστικός αναλυτής συμπεριλαμβάνει στο δείγμα του μονάδες του πληθυσμού έτσι, ώστε το τελικό δείγμα να έχει εκπροσώπους από κάθε κατηγορία του πληθυσμού, και μάλιστα με αναλογία ίση με εκείνη που ισχύει για τον πληθυσμό. Οι κατηγορίες ορίζονται συνήθως με βάση ένα δημογραφικό κριτήριο, π.χ φύλο, ηλικιακές ομάδες κτλ.

Έστω, για παράδειγμα, ότι ενδιαφερόμαστε να συλλέξουμε ένα δείγμα 100 φοιτητών ενός τμήματος ΑΕΙ, και ότι θεωρούμε το φύλο των φοιτητών σημαντικό παράγοντα για την έρευνα. Αν είναι γνωστό ότι 40% των φοιτητών είναι γυναίκες και 60% είναι άντρες, και το δείγμα συλλεγεί επιλέγοντας τις πρώτες 40 φοιτήτριες του τμήματος που θα εντοπίσουμε σε μια επίσκεψή μας στο προαύλιο της Σχολής, και αντίστοιχα τους πρώτους 60 άντρες φοιτητές, τότε το δείγμα θα είναι ένα δείγμα με προκαθορισμένα ποσοστά.

Η δειγματοληψία με προκαθορισμένα ποσοστά έχει ομοιότητες με τη στρωματοποιημένη, και ειδικότερα την αναλογική στρωματοποιημένη, αλλά δεν αποτελεί δειγματοληψία πιθανότητας όπως η στρωματοποιημένη. Μια περιγραφή και σύγκριση της δειγματοληψίας με προκαθορισμένα ποσοστά με τη στρωματοποιημένη δίνεται στην παράγραφο 4.4.

Για περαιτέρω ανάλυση των δειγματοληπτικών σχεδίων μη-πιθανότητας, παραπέμπουμε μεταξύ άλλων στο κεφάλαιο 2 του βιβλίου Henry (1990).

Στα κεφάλαια που ακολουθούν, θα ασχοληθούμε αποκλειστικά με τα δείγματα πιθανότητας. Τα δείγματα μη-πιθανότητας δεν επιτρέπουν γενίκευση για τον πληθυσμό, ούτε την ανάπτυξη της στατιστικής μεθοδολογίας η οποία παρέχει ακριβή στοιχεία και ιδιότητες για τα αποτελέσματα του δείγματος, με τη βοήθεια των οποίων μπορούμε στη συνέχεια να εξαγάγουμε ασφαλέστερα συμπεράσματα για τον πληθυσμό. Στην παράγραφο που

ακολουθεί, θα δοθεί ο απαραίτητος μαθηματικός συμβολισμός. Στη συνέχεια, δίνουμε μια συνοπτική επισκόπηση των κυριότερων στοιχείων από τη Στατιστική συμπερασματολογία.

## 1.4. Συμβολισμός

Υποθέτουμε ότι ο πληθυσμός για τον οποίο ενδιαφερόμαστε να εξαγάγουμε συμπεράσματα με τη βοήθεια ενός δείγματος είναι πεπερασμένος, και έστω ότι $N$ είναι το πλήθος των μελών του πληθυσμού. Το πλήθος αυτό ονομάζεται και *μέγεθος του πληθυσμού (population size)*. Ανάλογα, συμβολίζουμε με $n$ και το ονομάζουμε *μέγεθος δείγματος (sample size)* το πλήθος των μονάδων του πληθυσμού που επιλέγονται για το δείγμα.

### Ορισμός 1.4

*Χαρακτηριστικό ενός πληθυσμού (population characteristic)* ονομάζεται το ερώτημα για τον πληθυσμό, στο οποίο μας ενδιαφέρει να δώσουμε μια απάντηση.

### 1.4.1 Πληθυσμιακά μεγέθη

Το χαρακτηριστικό του πληθυσμού συμβολίζεται συνήθως με ένα κεφαλαίο γράμμα, έστω $Y$. Το χαρακτηριστικό $Y$ λαμβάνει, για κάθε μέλος του πληθυσμού, μια συγκεκριμένη τιμή, την απάντηση στο ερώτημα της έρευνας. Για παράδειγμα, εάν το χαρακτηριστικό του πληθυσμού για το οποίο επιθυμούμε να εξαγάγουμε συμπεράσματα είναι το εισόδημα, υπάρχει μια τιμή που αντιστοιχεί στο εισόδημα για το κάθε μέλους του πληθυσμού. Το χαρακτηριστικό $Y$ μπορεί να θεωρηθεί ως μια *τυχαία μεταβλητή (τ.μ.) (random variable, (rv))* με δυνατές τιμές τις δυνατές τιμές του χαρακτηριστικού για τα μέλη του πληθυσμού.

Η τιμή του χαρακτηριστικού $Y$ για το $i$ μέλος του πληθυσμού συμβολίζεται με $Y_i$. Επίσης, το διάνυσμα με στοιχεία τα $Y_i$ για όλα τα δυνατά μέλη του πληθυσμού, $\boldsymbol{U} = \{Y_1, Y_2, \dots, Y_N\}$, ονομάζεται διάνυσμα του πληθυσμού (population vector) για το χαρακτηριστικό $Y$. Είναι φανερό ότι το διάνυσμα $\boldsymbol{U}$ είναι γνωστό μόνο μετά από μια απογραφή.

Αν $Y$ είναι το υπό μελέτη χαρακτηριστικό του πληθυσμού, τότε οι ποσότητες για τις οποίες πιο συχνά ενδιαφερόμαστε να εξαγάγουμε συμπεράσματα είναι συνήθως:

- η *μέση τιμή (mean value)* του χαρακτηριστικού για τον πληθυσμό. Όταν ο υπό μελέτη πληθυσμός είναι πεπερασμένος, η μέση τιμή ορίζεται ως το άθροισμα των μετρήσεων του $Y$ για όλα τα μέλη του πληθυσμού διαιρούμενο με το πλήθος τους. Εάν συμβολίσουμε τη μέση τιμή του $Y$ για τον πληθυσμό με $\bar{Y}$, τότε:

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

- το *σύνολο (population total)* των τιμών του χαρακτηριστικού για τον πληθυσμό, είναι το άθροισμα των τιμών του χαρακτηριστικού $Y_i$ για όλα τα μέλη του πληθυσμού και συμβολίζεται με $Y_T$. Θα είναι:

$$Y_T = \sum_{i=1}^{N} Y_i$$

Προκύπτει άμεσα από τους παραπάνω ορισμούς ότι $Y_T = N\bar{Y}$. Οι ποσότητες $\bar{Y}$ και $Y_T$ έχουν νόημα για ένα συνεχές χαρακτηριστικό $Y$ ή, ισοδύναμα, για μια συνεχή τυχαία μεταβλητή. Για παράδειγμα, εάν $Y$ είναι το εισόδημα των κατοίκων μιας χώρας, τότε ενδιαφερόμαστε να εκτιμήσουμε το μέσο εισόδημα των κατοίκων, δηλαδή το $\bar{Y}$ ή το συνολικό εισόδημα των κατοίκων $Y_T$. Ανάλογα, εάν $Y$ είναι η κατανάλωση νερού (σε λίτρα)

των νοικοκυριών μιας πόλης για έναν μήνα, τότε $\bar{Y}$ είναι η μέση κατανάλωση νερού σε ένα μήνα για τα νοικοκυριά της πόλης και $Y_T$ η συνολική κατανάλωση νερού (σε λίτρα) της πόλης για ένα μήνα.

- το *ποσοστό (percentage)* ενός χαρακτηριστικού στον πληθυσμό. Εάν το υπό μελέτη χαρακτηριστικό αντιστοιχεί σε μια κατηγορική με 2 ή περισσότερα επίπεδα, τότε το ενδιαφέρον της έρευνας επικεντρώνεται στην εξαγωγή συμπερασμάτων σχετικά με το ποσοστό των μελών του πληθυσμού που ανήκουν σε ένα από τα δυνατά επίπεδα του χαρακτηριστικού. Για παράδειγμα, εάν το χαρακτηριστικό $Y$ αντιστοιχεί στο τύπο καπνιστή και λαμβάνει 3 δυνατές τιμές: 'μη-καπνιστής', 'μέτρια καπνιστής', 'έντονα καπνιστής', τότε ως θέμα μιας δειγματοληπτικής έρευνας μπορεί να είναι η εκτίμηση του ποσοστού των μελών του πληθυσμού που ανήκουν στην κατηγορία 'έντονα καπνιστής'.

  Το ποσοστό συμβολίζεται με $P$ και ορίζεται ως:

  $$P = \frac{A}{N}$$

  όπου $A$ είναι το πλήθος των μελών του πληθυσμού που ανήκουν στην υπό μελέτη κατηγορία.

- Η ποσότητα $A$, δηλαδή το πλήθος των μελών του πληθυσμού που ανήκουν στην υπό μελέτη κατηγορία. Προφανώς ισχύει:

  $$A = NP$$

Μια επιπλέον πολύ σημαντική ποσότητα του πληθυσμού είναι η διασπορά των τιμών του πληθυσμού γύρω από τη μέση τιμή για το χαρακτηριστικό, η οποία συμβολίζεται με $S^2$ και ορίζεται ως:

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

Η *διασπορά (variance)* $S^2$ δίνει πληροφορία σχετικά με την ετερογένεια ή ομοιογένεια των μετρήσεων του χαρακτηριστικού στο σύνολο του πληθυσμού. Όσο μεγαλύτερη η διασπορά $S^2$ για έναν πληθυσμό, τόσο πιο ετερογενής είναι ο πληθυσμός. Όπως θα δούμε στη συνέχεια, ο ρόλος της διασποράς του πληθυσμού για το χαρακτηριστικό παίζει ουσιαστικό ρόλο σχεδόν σε κάθε στάδιο της έρευνας: στον σχεδιασμό και την επιλογή της δειγματοληπτικής μεθόδου, στον καθορισμό του μεγέθους του δείγματος, και βέβαια στην ανάλυση των δεδομένων και τον υπολογισμό των ιδιοτήτων των εκτιμητών.

Σύμφωνα με τον ορισμό των ποσοτήτων $\bar{Y}$, $Y_T$, $P$, $A$ και $S^2$ είναι φανερό ότι οι ποσότητες αυτές είναι γνωστές, μόνο αν είναι γνωστές οι τιμές του χαρακτηριστικού για όλα τα μέλη του πληθυσμού. Για να τονιστεί το γεγονός ότι οι παραπάνω ποσότητες αναφέρονται στο σύνολο του πληθυσμού ονομάζονται και *πληθυσμιακές ποσότητες* ή *πληθυσμιακά μεγέθη* ή *παράμετροι του πληθυσμού (population quantities, population parameters)*, πχ πληθυσμιακή μέση τιμή, πληθυσμιακό σύνολο κτλ. Γενικά, οι πληθυσμιακές ποσότητες είναι σταθερές ποσότητες αφού αναφέρονται σε απογραφικά στοιχεία και δεν εμπεριέχουν τυχαιότητα, είναι όμως συνήθως άγνωστες, και η εκτίμησή τους αποτελεί τον σκοπό της διενέργειας της έρευνας.

### 1.4.2 Δειγματικές ποσότητες

Ανάλογα με τον συμβολισμό και τους ορισμούς των πληθυσμιακών ποσοτήτων, μπορούν να οριστούν οι αντίστοιχες *δειγματικές ποσότητες (sample quantities)*. Έστω ότι το μέγεθος του δείγματος είναι $n$. Για τα δείγματα πιθανότητας, οι $n$ παρατηρήσεις του δείγματος είναι τυχαίες μεταβλητές, γιατί εμπεριέχουν τον παράγοντα της τυχαιότητας.

Στη βιβλιογραφία, έχει επικρατήσει να συμβολίζουμε το δείγμα ως $s$. Για το διάνυσμα των παρατηρήσεων του δείγματος, χρησιμοποιούνται δύο εναλλακτικοί συμβολισμοί:

(α) $s = \{y_1,\ y_2, \dots, y_n\}$

Δηλαδή, διατηρούμε το ίδιο γράμμα που έχουμε δώσει για το χαρακτηριστικό του πληθυσμού, αλλά με μικρούς χαρακτήρες ώστε να διαχωρίζονται οι δειγματικές από τις πληθυσμιακές ποσότητες, και:

(β) $s = \{X_1,\ X_2, \dots, X_n\}$

Στον συμβολισμό αυτό, χρησιμοποιούμε ένα άλλο κεφαλαίο γράμμα για τις μετρήσεις του δείγματος σε σχέση με τις μετρήσεις του πληθυσμού.

Στα επόμενα κεφάλαια, για να τονίσουμε το γεγονός ότι οι μετρήσεις του δείγματος είναι τυχαίες μεταβλητές, τις οποίες συνήθως συμβολίζουμε με κεφαλαία γράμματα, υιοθετούμε τον (β) συμβολισμό για το δείγμα.

Η τυχαία μεταβλητή $X_i$ $(i = 1,2, \dots, n)$ είναι η μεταβλητή που καταγράφει την $i-$οστή μέτρηση του δείγματος. Οι δυνατές τιμές της είναι το σύνολο $\{Y_1,\ Y_2, \dots, Y_N\}$, π.χ. $X_3 = Y_{20}$, σημαίνει ότι ως 3η μέτρηση του δείγματος επιλέχθηκε η 20ή μέτρηση του πληθυσμού. Προφανώς, σε κάθε επανάληψη του δείγματος, η τιμή του πληθυσμού που επιλέγεται για την πλήρωση της $i$ θέσης του δείγματος μπορεί να διαφέρει.

Αν $\{X_1,\ X_2, \dots, X_n\}$ είναι ένα δείγμα μεγέθους $n$, τότε ορίζεται:

- η *δειγματική μέση τιμή (sample mean value)* η οποία συμβολίζεται με $\bar{X}$ και υπολογίζεται ως:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- η *δειγματική διασπορά (sample variance)* των τιμών του δείγματος η οποία συμβολίζεται με $s^2$. Η δειγματική διασπορά ορίζεται ανάλογα με την πληθυσμιακή $S^2$. Αναλυτικά:

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

Η διασπορά $s^2$ υπολογίζεται αριθμητικά μετά τη συλλογή και καταγραφή των μετρήσεων του δείγματος και η πληροφορία που προσφέρει είναι ανάλογη εκείνης της $S^2$. Όσο μεγαλύτερη είναι η $s^2$, τόσο μεγαλύτερη ετερογένεια παρατηρείται στις μετρήσεις του δείγματος, και αντίστροφα.

Οι τυχαίες μεταβλητές του δείγματος $X_i$ έχουν χαρακτηριστικά τα οποία σχετίζονται τόσο με τον ίδιο τον πληθυσμό από τον οποίο έχει επιλεγεί το δείγμα, όσο και με τον τρόπο δειγματοληψίας. Η εξάρτηση της $X_i$ από τον τρόπο δειγματοληψίας θα γίνει περισσότερο κατανοητή στην παράγραφο που ακολουθεί.

## 1.5. Στοιχεία από τη Στατιστική Συμπερασματολογία

Στην παρούσα παράγραφο, δίνονται επιγραμματικά στοιχεία από τη στατιστική συμπερασματολογία, με έμφαση στη διαφοροποίηση των συνήθων εννοιών όταν οι παρατηρήσεις του δείγματος προέρχονται από έναν πεπερασμένο πληθυσμό.

Υποθέτουμε ένα δείγμα πιθανότητας $s = \{X_1,\ X_2, \dots, X_n\}$ το οποίο έχει επιλεγεί από έναν πεπερασμένο πληθυσμό μεγέθους $N$. Έστω επίσης $\theta$ μια πληθυσμιακή ποσότητα ή παράμετρος (π.χ. $\bar{Y}$, $Y_T$ κτλ.) η οποία είναι άγνωστη, και για την οποία ενδιαφερόμαστε να εξαγάγουμε συμπεράσματα μέσω του δείγματος.

### Ορισμός 1.5

*Εκτιμητής (estimator)* $\hat{\theta}$ του $\theta$ ονομάζεται μια συνάρτηση των τυχαίων μεταβλητών του δείγματος, που χρησιμοποιείται με σκοπό την εκτίμηση του $\theta$.

Σύμφωνα με τον ορισμό, θα είναι $\hat{\theta} = \hat{\theta}(s) = \hat{\theta}(X_1, X_2, \dots, X_n)$ και, κατά συνέπεια: (α) ο εκτιμητής είναι επίσης μια τυχαία μεταβλητή ως συνάρτηση τυχαίων μεταβλητών και (β) είναι εφικτό να υπολογίσουμε την αριθμητική τιμή του αμέσως μετά τη διεξαγωγή της έρευνας, όταν θα είναι γνωστές οι αριθμητικές τιμές των $X_i$ $(i = 1,2, \dots, n)$.

Ένας όρος γενικότερος του εκτιμητή είναι η *στατιστική συνάρτηση (σ.σ.)* ή *στατιστικό (statistical function* ή *statistic*). Στατιστικό ονομάζεται μια συνάρτηση των τ.μ. του δείγματος. Σύμφωνα με τον Ορισμό 1.5, ένας εκτιμητής είναι στατιστική συνάρτηση. Τα περισσότερα από τα συμπεράσματα στη θεωρία δειγματοληψίας βασίζονται στη *δειγματική κατανομή (sampling distribution)*. Ως δειγματική κατανομή μιας στατιστικής συνάρτησης ορίζεται η κατανομή των τιμών της στατιστικής συνάρτησης που προκύπτουν, εάν γίνει εξάντληση όλων των δυνατών δειγμάτων σύμφωνα με τον τρόπο που ακολουθείται για την επιλογή του δείγματος.

Η *αναμενόμενη τιμή (expected value)* μιας σ.σ. $t = t(X_1, X_2, \dots, X_n) = t(s)$ υπολογίζεται λαμβάνοντας υπόψη τις δυνατές τιμές της συνάρτησης $t = t(s)$, δηλαδή για όλα τα δυνατά δείγματα $s$ και τις αντίστοιχες πιθανότητες πραγματοποίησης του κάθε δείγματος, σύμφωνα με τον τρόπο δειγματοληψίας. Θα είναι συνεπώς:

$$\mathrm{E}(t) = \sum_{s} \pi(s)t(s) \qquad (1.1)$$

όπου ο δείκτης $s$ του αθροίσματος λαμβάνει όλες τις δυνατές τιμές του δείγματος, $t(s)$ είναι η τιμή της σ.σ $t$ υπολογισμένη για το δείγμα $s$, και $\pi(s)$ η πιθανότητα επιλογής του $s$.

Ο ορισμός του εκτιμητή είναι αρκετά γενικός, με αποτέλεσμα να επιτρέπει οποιαδήποτε συνάρτηση (οποιασδήποτε μορφής) να είναι θεωρητικά ένας εκτιμητής της ποσότητας που μας ενδιαφέρει. Υπάρχει συνεπώς ανάγκη για αξιολόγηση και σύγκριση των εκτιμητών μεταξύ τους. Η αξιολόγηση γίνεται με βάση μια σειρά κριτηρίων, ορισμένα εκ των οποίων αποτελούν ταυτόχρονα και ιδιότητες των εκτιμητών. Παραθέτουμε στη συνέχεια τα πιο σημαντικά από αυτά.

### Ορισμός 1.6

*Αμεροληψία (unbiasedness)* Ένας εκτιμητής $\hat{\theta}$ θα λέγεται αμερόληπτος εκτιμητής της παραμέτρου $\theta$ εάν η αναμενόμενη τιμή του ισούται με $\theta$, δηλ. $\mathrm{E}(\hat{\theta}) = \theta$.

Η αναμενόμενη τιμή του εκτιμητή, $\mathrm{E}(\hat{\theta})$, υπολογίζεται σύμφωνα με τη σχέση (1.1). Ας δούμε την εφαρμογή του ορισμού της αναμενόμενης τιμής και της ιδιότητας της αμεροληψίας ενός εκτιμητή μέσα από ένα παράδειγμα. Υποθέτουμε ένα μικρό θεωρητικό παράδειγμα πληθυσμού, όπου για χάρη κατανόησης οι τιμές του πληθυσμού είναι γνωστές.

## Παράδειγμα 1.6

Σε μια μικρή πόλη που έχει 5 δημοτικά σχολεία, επιθυμούμε να εκτιμήσουμε τον μέσο αριθμό μαθητών ανά σχολείο, λαμβάνοντας ένα δείγμα 2 σχολείων. Έστω ότι τα στοιχεία των συνολικών αριθμών των μαθητών στα σχολεία είναι γνωστά για τον πληθυσμό:

| α/α σχολείου | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Συνολικός αριθμός μαθητών | 59 | 28 | 90 | 44 | 36 |

Η παράμετρος του πληθυσμού που ενδιαφερόμαστε να εκτιμήσουμε είναι η μέση τιμή $\bar{Y}$ του αριθμού μαθητών στα σχολεία, έστω $\theta$, δηλ. $\bar{Y} = \theta$. Επειδή ο πληθυσμός είναι γνωστός, είναι $\theta = \frac{257}{5} = 51.4$.

Εάν η επιλογή του δείγματος γίνει επιλέγοντας τυχαία 2 σχολεία από τα 5 χωρίς επανατοποθέτηση, τότε τα δυνατά δείγματα που μπορούν να προκύψουν είναι όσοι οι συνδυασμοί 5 ανά 2, δηλαδή $\binom{5}{2} = \frac{5!}{2!3!} = 10$. Επειδή οι πιθανότητες επιλογής κατά την τυχαία επιλογή είναι ίσες, το καθένα από τα 10 δείγματα έχει ίση πιθανότητα επιλογής και κατά συνέπεια ίση με $\frac{1}{10}$. Θεωρούμε στη συνέχεια ως εκτιμητή $\hat{\theta}$ του ποσοστού $\theta$, τη δειγματική μέση τιμή, $\hat{\theta} = \bar{X}$. Ο πίνακας που ακολουθεί δίνει ένα προς ένα όλα τα δυνατά δείγματα μεγέθους 2, καθώς και την εκτίμηση $\hat{\theta}$ που λαμβάνεται από το κάθε δείγμα χωριστά.

| **Δείγμα** $s_i$ | **Πιθανότητα επιλογής** $\pi(s_i)$ | **Εκτιμητής** $\hat{\theta}(s_i)$ |
|---|---|---|
| **{1,2}** | $\frac{1}{10}$ | $\frac{59+28}{2} = 43.5$ |
| **{1,3}** | $\frac{1}{10}$ | $\frac{59+90}{2} = 74.5$ |
| **{1,4}** | $\frac{1}{10}$ | $\frac{59+44}{2} = 51.5$ |
| **{1,5}** | $\frac{1}{10}$ | $\frac{59+36}{2} = 47.5$ |
| **{2,3}** | $\frac{1}{10}$ | $\frac{28+90}{2} = 59$ |
| **{2,4}** | $\frac{1}{10}$ | $\frac{28+44}{2} = 36$ |
| **{2,5}** | $\frac{1}{10}$ | $\frac{28+36}{2} = 32$ |
| **{3,4}** | $\frac{1}{10}$ | $\frac{90+44}{2} = 67$ |
| **{3,5}** | $\frac{1}{10}$ | $\frac{90+36}{2} = 63$ |
| **{4,5}** | $\frac{1}{10}$ | $\frac{44+36}{2} = 40$ |

Από την εξάντληση των δειγμάτων μεγέθους 2, διαπιστώνουμε ότι η εκτίμηση του μέσου αριθμού μαθητών ανά σχολείο (το οποίο στην προκειμένη περίπτωση γνωρίζουμε ότι είναι 51.4) λαμβάνει 10 διαφορετικές τιμές, με εύρος από 32 έως 74.5. Η δειγματική κατανομή του εκτιμητή είναι οι 10 αυτές διαφορετικές τιμές, ενώ η πιθανότητα πραγματοποίησης της κάθε τιμής ισούται με την πιθανότητα επιλογής του αντίστοιχου δείγματος. Το στατιστικό σφάλμα που εμπεριέχεται στις δειγματοληπτικές έρευνες συνδέεται με το γεγονός ότι ο εκτιμητής είναι μια τυχαία μεταβλητή και λαμβάνει τόσες δυνατές τιμές, όσες και το πλήθος των δυνατών δειγμάτων. Αντίθετα, η πληθυσμιακή ποσότητα $\theta = 51.4$ είναι σταθερή.

Χρησιμοποιώντας τις δύο τελευταίες στήλες του πίνακα και εφαρμόζοντας τον ορισμό της αναμενόμενης τιμής (1.1), θα έχουμε:

$$\mathrm{E}(\hat{\theta}) = \sum_{i=1}^{10} \pi(s_i)\hat{\theta}(s_i) = \sum_{i=1}^{10} \frac{1}{10}\hat{\theta}(s_i)$$

$$= \frac{1}{10}(43.5 + 74.5 + 51.5 + 47.5 + 59 + 36 + 32 + 67 + 63 + 40) = 51.4$$

Άρα ο εκτιμητής $\hat{\theta}$ που υιοθετήσαμε είναι ένας αμερόληπτος εκτιμητής του μέσου αριθμού των μαθητών ανά σχολείο.

Διαπιστώνουμε ότι για την εύρεση της μέσης τιμής του εκτιμητή στο παράδειγμα, λάβαμε υπόψη μας (α) όλα τα δυνατά δείγματα μεγέθους που μπορούν να επιλεγούν και (β) τον τρόπο δειγματοληψίας, δηλ. τις πιθανότητες επιλογής του καθενός από τα δυνατά δείγματα. Συνεπώς, η αναμενόμενη τιμή του ίδιου εκτιμητή $\hat{\theta} = \bar{X}$, θα είναι ενδεχομένως διαφορετική, εάν υιοθετήσουμε ένα εναλλακτικό δειγματοληπτικό σχέδιο, π.χ. με άνισες πιθανότητες για τα 10 δυνατά δείγματα.

Το σύνολο των δυνατών δειγμάτων που προκύπτουν με μια μέθοδο δειγματοληψίας ονομάζεται δειγματοληπτικός χώρος (sampling space) και συμβολίζεται συνήθως με $\mathcal{S}$. Για το παράδειγμά μας, ο δειγματοληπτικός χώρος είναι:

$$\mathcal{S} = \{\{1,2\},\{1,3\},\{1,3\},\dots,\{4,5\}\}$$

Ο αριθμός των στοιχείων του $\mathcal{S}$ ταυτίζεται με το πλήθος των δυνατών δειγμάτων που προκύπτουν σύμφωνα με τον τρόπο δειγματοληψίας που ακολουθούμε στην έρευνα. Εάν για το Παράδειγμα 1.5 θεωρούμε ως τρόπο δειγματοληψίας: επιλογή με τυχαίο τρόπο ενός σχολείου μεταξύ αυτών με α/α 1,2 και 3 και ενός σχολείου μεταξύ των 4 και 5, ο δειγματικός χώρος θα είναι

$$\mathcal{S} = \{\{1,4\},\{1,5\},\{2,4\},\{2,5\},\{3,4\},\{4,5\}\}$$

που αποτελείται από $\binom{3}{1}\binom{2}{1} = 6$ δείγματα.

### Ορισμός 1.7

Το *ποσό μεροληψίας (bias)* ενός εκτιμητή $\hat{\theta}$, συμβολίζεται ως $\mathrm{bias}(\hat{\theta})$ και δίνεται από τη σχέση $\mathrm{bias}(\hat{\theta}) = \mathrm{E}(\hat{\theta}) - \theta$.

Εάν $\mathrm{bias}(\hat{\theta}) > 0$, αυτό σημαίνει ότι ο εκτιμητής παρουσιάζει θετική μεροληψία, δηλ. για το δοθέν δειγματοληπτικό σχέδιο αναμένουμε η δειγματική κατανομή του $\hat{\theta}$ να έχει μέση τιμή μεγαλύτερη του $\theta$ και, κατά συνέπεια, ο εκτιμητής να υπερ-εκτιμά την παράμετρο. Αντίστοιχα, αν $\mathrm{bias}(\hat{\theta}) < 0$, έχουμε αρνητική μεροληψία ή ισοδύναμα ο εκτιμητής υπο-εκτιμά την παράμετρο. Προφανώς, αν $\mathrm{bias}(\hat{\theta}) = 0$ ο εκτιμητής είναι αμερόληπτος.

Ένα άλλο κριτήριο σύγκρισης εκτιμητών είναι η *ακρίβεια (accuracy)*. Η ακρίβεια ως ιδιότητα ενός εκτιμητή δίνει ένα μέτρο της συγκέντρωσης, ή, αντίθετα, της απόκλισης που παρουσιάζουν μεταξύ τους οι δυνατές τιμές του εκτιμητή. Όσο πιο πυκνά είναι οι δυνατές τιμές του εκτιμητή, όσο δηλαδή μεγαλύτερη είναι η συγκέντρωση, τόσο μεγαλύτερη είναι η ακρίβεια του εκτιμητή.

Ένα μέτρο της ακρίβειας του εκτιμητή είναι το *μέσο τετραγωνικό σφάλμα, (mean square error)* που συμβολίζεται με MSE.

### Ορισμός 1.8

Το μέσο τετραγωνικό σφάλμα ενός εκτιμητή $\hat{\theta}$ ορίζεται ως η αναμενόμενη τιμή της τετραγωνικής απόκλισης του εκτιμητή από την προς εκτίμηση ποσότητα:

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{E}(\hat{\theta} - \theta)^2 \tag{1.2}$$

Η αναμενόμενη τιμή στον ορισμό του MSE υπολογίζεται όπως και στη σχέση (1.1), δηλαδή:

$$\mathrm{MSE}(\hat{\theta}) = \sum_{s \in S} \pi(s) \left(\hat{\theta}(s) - \theta\right)^2$$

*Ακριβής (accurate)* είναι ένας εκτιμητής με μικρό μέσο τετραγωνικό σφάλμα.

Το MSE σχετίζεται με το γνωστό μέτρο της διακύμανσης ενός εκτιμητή. Υπενθυμίζουμε ότι η *διακύμανση (variance)* ενός εκτιμητή ορίζεται από τη σχέση:

$$\mathrm{Var}(\hat{\theta}) = \mathrm{E}\left(\hat{\theta} - \mathrm{E}(\hat{\theta})\right)^2 \tag{1.3}$$

Από τις σχέσεις (1.2) και (1.3) εύκολα προκύπτει ότι εάν $\mathrm{E}(\hat{\theta}) = \theta$, δηλαδή εάν ο εκτιμητής $\hat{\theta}$ είναι αμερόληπτος, τότε $\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta})$.

Γενικότερα, αποδεικνύεται ότι η σχέση που συνδέει το MSE ενός εκτιμητή και τη διακύμανσή του είναι

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \mathrm{bias}^2(\hat{\theta}) \tag{1.4}$$

Μεταξύ όλων των αμερόληπτων εκτιμητών, ο εκτιμητής με την ελάχιστη διακύμανση λέγεται *αποτελεσματικός (efficient)*. Αν $\hat{\theta}_1$ και $\hat{\theta}_2$ είναι δύο αμερόληπτοι εκτιμητές με $\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2)$ τότε ο εκτιμητής $\hat{\theta}_1$ λέγεται σχετικά αποτελεσματικός.

### Ορισμός 1.9

*Τυπικό σφάλμα (standard error)* του εκτιμητή $\hat{\theta}$, συμβολικά $\mathrm{se}(\hat{\theta})$, ονομάζεται η θετική τετραγωνική ρίζα της διακύμανσης του εκτιμητή, δηλ.

$$\mathrm{se}(\hat{\theta}) = \sqrt{\mathrm{Var}(\hat{\theta})}$$

Για τον υπολογισμό της αναμενόμενης τιμής, του μέσου τετραγωνικού σφάλματος ή της διακύμανσης ενός εκτιμητή λαμβάνεται υπόψη ο δειγματικός χώρος και οι πιθανότητες επιλογής των δειγμάτων. Συνεπώς, όλες οι παραπάνω ποσότητες εξαρτώνται από το δειγματοληπτικό σχέδιο που υιοθετήθηκε στην έρευνα.

Τέλος, ως υπενθύμιση, ο Πίνακας 1.1 συγκεντρώνει ορισμένες από τις βασικές ιδιότητες της αναμενόμενης τιμής, της διακύμανσης και ποσοτήτων που ορίζονται με τη βοήθεια αυτών, όπως η *συμμεταβλητότητα (covariance)* και η *συσχέτιση (correlation)*.

Τα σύμβολα $X, Y, Z$ αντιπροσωπεύουν τυχαίες μεταβλητές, ενώ τα $a, b, c$ πραγματικούς αριθμούς.

| Ονομασία | Ιδιότητες |
|---|---|
| **Αναμενόμενη Τιμή** | 1. $\mathrm{E}(a) = a$<br>2. $\mathrm{E}(aX + bY) = a\mathrm{E}(X) + b\mathrm{E}(Y)$ |
| **Διακύμανση** | 1. $\mathrm{Var}(a) = 0$<br>2. $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$<br>3. Αν $X, Y$ είναι ασυσχέτιστες:<br>$\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ |

| | |
|---|---|
| **Συμμεταβλητότητα**<br>$\mathbf{Cov}(\boldsymbol{X},\boldsymbol{Y}) =$<br>$\mathrm{E}[(X-\bar{X})(Y-\bar{Y})] =$<br>$\mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y)$ | 1. Αν $\mathrm{Cov}(X,Y)=0$, τότε $X$, $Y$ ασυσχέτιστες<br>2. $\mathrm{Cov}(X,Y+Z)=\mathrm{Cov}(X,Y)+\mathrm{Cov}(X,Z)$<br>3. $\mathrm{Cov}(aX,Y)=a\,\mathrm{Cov}(X,Y)$<br>4. $\mathrm{Cov}(aX,bY)=ab\mathrm{Cov}(X,Y)$<br>5. $\mathrm{Var}(X\pm Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)\pm 2\mathrm{Cov}(X,Y)$ |
| **Συσχέτιση**<br>$\rho(X,Y)=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$ | $-1\le\rho(X,Y)\le 1$ |

***Πίνακας 1.1:*** *Βασικές ιδιότητες κυριότερων στατιστικών ποσοτήτων*

## 1.6. Κύρια στοιχεία μιας δειγματοληπτικής έρευνας

Στην παράγραφο 1.1 είδαμε ορισμένα παραδείγματα δειγματοληπτικών ερευνών. Έχοντας δώσει ενδιάμεσα το θεωρητικό υπόβαθρο και τη σχετική ορολογία, είμαστε σε θέση να δώσουμε σε πιο συγκεκριμένη και αναλυτική μορφή τα στοιχεία από τα οποία αποτελείται μια δειγματοληπτική έρευνα. Για μεγαλύτερη ανάπτυξη και εξειδίκευση ανάλογα με το θέμα της έρευνας, παραπέμπουμε στους Henry (1990), Blair, Czaja and Blair (2013), Fink (2013) και Floyd and Fowel (2001).

- *Το Δειγματοληπτικό σχέδιο (sampling design).*

  Είναι ο τρόπος επιλογής των μονάδων στο δείγμα. Το δειγματοληπτικό σχέδιο επιλέγεται μετά από συνεργασία (i) του στατιστικού επιστήμονα, (ii) εκείνου που αναθέτει την έρευνα και καθορίζει τα ερωτήματά της και (iii) τους ανθρώπους που θα διεξαγάγουν τη δειγματοληψία, των συνεντευκτών (interviewers) όπως λέγονται.

  Για την επιλογή του δειγματοληπτικού σχεδίου τονίζουμε τον σημαντικό, αν και όχι προφανή, ρόλο εκείνου που αναθέτει την έρευνα. Ο ρόλος του είναι σημαντικός, επειδή θέτει τα ερωτήματα και συνεπώς καθορίζει ποιες μεταβλητές πρέπει να καταγραφούν. Επίσης, προσδιορίζει τους χρηματικούς πόρους που είναι διαθέσιμοι για την έρευνα και θέτει τις απαιτήσεις ή προδιαγραφές της έρευνας ως προς την επιθυμητή ακρίβεια των εκτιμητών που θα παραχθούν. Για παράδειγμα, εάν το πρόβλημα είναι η εκτίμηση ενός ποσοστού, μπορεί να επιθυμεί η έρευνα να διεξαχθεί έτσι ώστε ο εκτιμητής του ποσοστού να έχει σφάλμα που δεν ξεπερνά το 2%. Όλα τα παραπάνω συνυπολογίζονται από τον στατιστικό αναλυτή, προκειμένου να επιλέξει το δειγματοληπτικό σχέδιο ή τον συνδυασμό των δειγματοληπτικών σχεδίων και τη μέθοδο εκτίμησης των πληθυσμιακών παραμέτρων.

  Ο ρόλος του συνεντεύκτη στη διαδικασία σχεδιασμού της έρευνας είναι συμπληρωματικός και βοηθητικός. Δίνει πληροφορίες για την εκτίμηση του χρόνου και της δυσκολίας της συλλογής των δεδομένων. Τα στοιχεία αυτά συνεκτιμώνται από τον στατιστικό αναλυτή.

- *Οι μετρήσεις της έρευνας (survey measurements).*

  Μετά την ανάθεση της έρευνας από κάποιο φορέα/άτομο, ένα ειδικευμένο προσωπικό επιλέγει τις μεταβλητές που θα συμπεριληφθούν και τη μονάδα μέτρησης της καθεμιάς. Το ίδιο αυτό προσωπικό, που συχνά απαρτίζεται από ψυχολόγους και κοινωνιολόγους, συντάσσει στη συνέχεια το *ερωτηματολόγιο (questionnaire)*.

  Η σύνταξη του ερωτηματολογίου είναι σημαντική και πολλές φορές πολύπλοκη. Ο τελικός στόχος ενός καλά σχεδιασμένου ερωτηματολογίου είναι να προσφέρει με το πέρας της έρευνας ένα αξιόπιστο σύνολο δεδομένων, το οποίο στη συνέχεια θα αναλύσει ο στατιστικός αναλυτής.

  Υπάρχουν αρκετοί κανόνες σύμφωνα με τους οποίους σχεδιάζεται ένα καλό ερωτηματολόγιο (βλ. Floyd and Fowler, 2001). Εν συντομία, η σύνταξη ενός ερωτηματολογίου γίνεται επιδιώκοντας (α) να έχει συγκεκριμένο στόχο (focus), (β) συντομία και (γ) απλότητα. Τα τρία αυτά στοιχεία

εξασφαλίζουν ελαχιστοποίηση των σφαλμάτων λόγω κακής διατύπωσης των ερωτήσεων που έχει ως αποτέλεσμα λανθασμένη κατανόηση και καταγραφή της μέτρησης, απαλοιφή της μεροληψίας που προέρχεται από κατευθυνόμενες ερωτήσεις, και μετρήσεις που είναι προϊόν είτε κούρασης είτε τυχαίας επιλογής απαντήσεων.

- *Διεξαγωγή της έρευνας (survey operations)*

  Κατά τη διεξαγωγή της έρευνας γίνεται προσπάθεια να μην υπάρξουν παρεκκλίσεις από τον αρχικό σχεδιασμό σε κανένα από τα ενδιάμεσα στάδια. Αυτό είναι σημαντικό, γιατί η αξιοπιστία των δεδομένων και των αποτελεσμάτων εξασφαλίζεται μόνο εφόσον η πραγματοποίηση της έρευνας ακολουθεί τις υποθέσεις της στατιστικής θεωρίας βάσει της οποίας ισχύουν οι ιδιότητες των εκτιμητών. Για παράδειγμα, εάν η καταγραφή της μέτρησης σε μια ερώτηση γίνεται συστηματικά με λανθασμένο τρόπο από τον συνεντεύκτη, ο εκτιμητής που θα υπολογιστεί για τη μεταβλητή που αντιστοιχεί στη συγκεκριμένη ερώτηση θα είναι μεροληπτικός, ακόμα και αν θεωρητικά, ως στατιστική συνάρτηση, ο εκτιμητής έχει την ιδιότητα της αμεροληψίας.

  Ο αριθμός των μελών της ομάδας διεξαγωγής της έρευνας εξαρτάται από τους σκοπούς, την έκταση και τον προϋπολογισμό της έρευνας.

  Τέλος, σημαντικό ρόλο κατά τη διεξαγωγή της έρευνας στην αποφυγή και πρόληψη προβλημάτων τα οποία ενδέχεται να οδηγήσουν σε απόκλιση από τον σχεδιασμό, παίζει η *πιλοτική έρευνα (pilot study)*.

  Πιλοτική είναι μια έρευνα η οποία διεξάγεται πριν από την κύρια έρευνα. Ακολουθεί τον σχεδιασμό της κύριας έρευνας σε κάθε στάδιο (δειγματοληπτικό σχέδιο, ερωτηματολόγιο, καταγραφή, ανάλυση), αλλά είναι μικρότερης κλίμακας, δηλ. επιλέγεται ένα αρκετά μικρότερο δείγμα από εκείνο της κύριας. Η πιλοτική έρευνα είναι απαραίτητη κυρίως σε νέες έρευνες, και η σκοπιμότητά της είναι πολύπλευρη. Σε αρχική φάση, ελέγχεται το ερωτηματολόγιο και εντοπίζονται τυχόν παρερμηνείες ή ερωτήσεις που δεν είναι κατανοητές. Επίσης, γίνεται μια εκτίμηση του χρόνου της διεξαγωγής της έρευνας και ένας έλεγχος για την καλή εκπαίδευση και προετοιμασία των συνεντευκτών. Τέλος, και σημαντικότερο, λαμβάνονται κάποιες πρώτες εκτιμήσεις για τις άγνωστες παραμέτρους του πληθυσμού, οι οποίες θα χρησιμεύσουν στην υλοποίηση της κύριας έρευνας. Για παράδειγμα, όπως θα δούμε στο Κεφάλαιο 2 -, για τον προσδιορισμό του απαιτούμενου μεγέθους του δείγματος ώστε να πληρούνται ορισμένες προδιαγραφές, είναι απαραίτητο να διαθέτουμε στοιχεία για τον πληθυσμό, όπως η διακύμανση του υπό μελέτη χαρακτηριστικού. Εάν τα στοιχεία αυτά δεν παρέχονται από μια πρόσφατη απογραφή ή από μια προγενέστερη έρευνα, τότε η πιλοτική είναι αναγκαία.

- *Στατιστική Ανάλυση (Statistical Analysis)*

  Μετά τη διεξαγωγή της έρευνας και την καταγραφή των μετρήσεων, είναι διαθέσιμο ένα σύνολο δεδομένων. Η ανάλυση των δεδομένων θα γίνει από τον στατιστικό ερευνητή εφαρμόζοντας τα αποτελέσματα της Θεωρίας Δειγματοληψίας και λαμβάνοντας υπόψη το δειγματοληπτικό σχέδιο που υιοθετήθηκε κατά την έρευνα. Τα κύρια σημεία της ανάλυσης είναι ο υπολογισμός των εκτιμητών των αγνώστων ποσοτήτων του πληθυσμού βάσει του δείγματος, ο υπολογισμός των σφαλμάτων της εκτίμησης και η κατασκευή διαστημάτων εμπιστοσύνης.

  Σε σύνθετα δειγματοληπτικά σχέδια, όπου γίνεται συνδυασμός αρκετών επιμέρους δειγματοληπτικών σχεδίων, ο υπολογισμός των τυπικών σφαλμάτων των εκτιμητών γίνεται με τη βοήθεια προσεγγιστικών μεθόδων.

  Τέλος, τα αποτελέσματα της ανάλυσης καταγράφονται και παρουσιάζονται.

# Βιβλιογραφικές Αναφορές


Blair, J., Czaja, R. F. & Blair, E. (2013). *Designing Surveys: A Guide to Decisions and Procedures*. 3rd Edition. Sage Publications.

Fink, A. (2013). *How to Conduct Surveys: A Step-by-Step Guide*. 5th Edition. Sage Publications.

Floyd, J. & Fowler, Jr. (2001). *Survey Research Methods (Applied Social Research Methods)*. 3rd Edition. Sage Publications.

Groves, R.M. Floyd, J., Fower, Jr., Couper, M.P., Lepkowski, J.M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology*. 2nd Edition, Wiley-Blackwell.

Henry, G. T. (1990). *Practical Sampling, Sage Publications* (CA).

# Κεφάλαιο 2 - ΑΠΛΗ ΤΥΧΑΙΑ ΔΕΙΓΜΑΤΟΛΗΨΙΑ

**Σύνοψη**

*Με το κεφάλαιο αυτό ξεκινά η μελέτη των στοιχειωδών δειγματοληπτικών σχημάτων που ανήκουν στη γενικότερη κατηγορία των μεθόδων δειγματοληψίας στις οποίες η επιλογή του δείγματος βασίζεται στη θεωρία πιθανοτήτων. Στα δειγματοληπτικά αυτά σχήματα, οι στοιχειώδεις μονάδες του πληθυσμού αποτελούν ταυτόχρονα και τις δειγματοληπτικές μονάδες. Η χρήση των στοιχειωδών δειγματοληπτικών σχημάτων έχει μεγάλη ευρύτητα στις δειγματοληπτικές έρευνες, αλλά συνήθως τα σχήματα αυτά εμφανίζονται σε συνδυασμό ή ως επιμέρους βήματα ενός μεγαλύτερου και πιο σύνθετου δειγματοληπτικού σχεδίου. Η μελέτη της πρώτης δειγματοληπτικής μεθόδου, της απλής τυχαίας δειγματοληψίας είναι σημαντική, όχι ίσως από πρακτικής άποψης, δεδομένου ότι σπάνια η μέθοδος αυτή θα εφαρμοστεί αποκλειστικά σε μια δειγματοληπτική έρευνα, αλλά επειδή αποτελεί τη βάση πάνω στην οποία η Θεωρία Δειγματοληψίας (Sampling Theory) έχει χτιστεί. Η μελέτη της απλής τυχαίας δειγματοληψίας (α.τ.δ.) (αγγλικά: simple random sampling ή srs) επίσης θα βοηθήσει στην κατανόηση των εννοιών που περιγράφηκαν στο Κεφάλαιο 1 - και οι οποίες αφορούν τον σκοπό μιας δειγματοληπτικής έρευνας, τη διαδικασία επιλογής του δείγματος, την εκτίμηση, τις δειγματοληπτικές κατανομές και τις ιδιότητες των εκτιμητών. Επίσης, αρκετά από τα δειγματοληπτικά σχήματα που θα αναπτυχθούν σε επόμενα κεφάλαια κάνουν χρήση της απλής τυχαίας δειγματοληψίας.*

**Προαπαιτούμενη γνώση**

*Κεφάλαιο 1 -, Εκτιμητική, Κανονική κατανομή, t κατανομή, Διάστημα Εμπιστοσύνης.*

## 2.1. Ορισμός και περιγραφή της Απλής Τυχαίας Δειγματοληψίας

Η απλή τυχαία δειγματοληψία είναι μια μέθοδος δειγματοληψίας με πιθανότητες, κατά την οποία οι πιθανότητες επιλογής των μονάδων από τον πληθυσμό είναι ίσες μεταξύ τους.

Έστω $N$ το πλήθος των μονάδων σ'έναν πληθυσμό και έστω $n$ το μέγεθος του δείγματος που επιθυμούμε να επιλέξουμε. Συνολικά υπάρχουν $\binom{N}{n}$ δυνατά δείγματα μεγέθους $n$ που μπορούν να επιλεγούν από τον πληθυσμό (με την ιδιότητα να μην εμφανίζεται περισσότερες από μία φορά μια μονάδα του πληθυσμού στο ίδιο δείγμα. Βλ. πιο κάτω Παρατήρηση 2.1).

Σύμφωνα με τον συμβολισμό του Κεφάλαιο 1 -, το σύνολο των δυνατών δειγμάτων για την περίπτωση αυτή θα είναι:

$$S = \left\{ s_1, s_2, \dots, s_{\binom{N}{n}} \right\}$$

όπου $\binom{N}{n} = \frac{N!}{(N-n)!n!}$ και $n! = n(n-1)(n-2)\cdots 1$. Με τη βοήθεια του συνόλου $S$, δίνεται ο ορισμός της απλής τυχαίας δειγματοληψίας.

### Ορισμός Απλής Τυχαίας Δειγματοληψίας

Απλή τυχαία δειγματοληψία είναι η δειγματοληψία κατά την οποία όλα τα δυνατά δείγματα του πληθυσμού, μεγέθους $n$ $(n = 1,2,..,N)$, $s_1, s_2, \dots, s_{\binom{N}{n}}$, έχουν την ίδια πιθανότητα να επιλεγούν. Η πιθανότητα αυτή ισούται με $1/\binom{N}{n}$.

### Παρατήρηση 2.1

Στον παραπάνω ορισμό έγινε εφαρμογή της ιδιότητας ότι όλα τα στοιχεία του δείγματος είναι διαφορετικά μέλη του πληθυσμού ή, όπως αλλιώς λέγεται, **δειγματοληψία χωρίς επανατοποθέτηση (sampling without**

**replacement)**. Εάν η δειγματοληψία επιτρέπει επανατοποθέτηση (δηλ. οποιαδήποτε μονάδα του πληθυσμού να έχει τη δυνατότητα να εμφανισθεί περισσότερες από μία φορές στο δείγμα) ο ορισμός ισχύει ως έχει, με τη μόνη διαφορά ότι το πλήθος των δυνατών δειγμάτων θα είναι διευρυμένο και, συγκεκριμένα, θα είναι $N^n$. Κατά συνέπεια, η πιθανότητα επιλογής του κάθε δείγματος, για την απλή τυχαία δειγματοληψία, θα είναι $1/N^n$.

Στα παρακάτω θα γίνει ανάπτυξη και ανάλυση για τη δειγματοληψία χωρίς επανατοποθέτηση, μιας και αποτελεί στην πράξη την πιο συνηθισμένη εκδοχή της απλής τυχαίας δειγματοληψίας. Η απλή τυχαία δειγματοληψία αποτελεί αντικείμενο σχεδόν κάθε διδακτικού ή ερευνητικού βιβλίου με αντικείμενο τη θεωρία δειγματοληψίας ή τη διεξαγωγή μιας έρευνας. Μεταξύ αυτών, προτείνονται τα συγγράμματα των Cochran (1977, Κεφ. 2), Des Raj (1968, Κεφ. 3), Rao (2000, Κεφ. 2, 3 και 4), Barnett (2002, Κεφ. 2) και Levy & Lemeshow (1999, Κεφ. 3). Για τη μελέτη της α.τ.δ. με επανατοποθέτηση, παραπέμπουμε στους Thompson (2012, Κεφ. 2) και Des Raj (1968, Κεφ. 3).

## 2.2. Επιλογή ενός δείγματος σύμφωνα με την Απλή Τυχαία Δειγματοληψία

Για την εφαρμογή της απλής τυχαίας δειγματοληψίας, προκειμένου να επιλεγεί ένα δείγμα μεγέθους $n$, χρησιμοποιείται η παρακάτω Πρόταση. Ένα δείγμα που επιλέγεται με τη διαδικασία που περιγράφεται στην Πρόταση είναι απλό τυχαίο σύμφωνα με τον ορισμό, γιατί διατηρείται η ιδιότητα των ίσων πιθανοτήτων για όλα τα δυνατά δείγματα.

### Πρόταση 2.1

Ένα δείγμα είναι απλό τυχαίο εάν η επιλογή των μονάδων του από τον πληθυσμό γίνεται μία προς μία διαδοχικά, χωρίς επανατοποθέτηση, και έτσι ώστε κάθε φορά όλες οι διαθέσιμες μονάδες στον πληθυσμό να παραμένουν ισοπίθανες.

### Απόδειξη

Έστω $X_1, X_2, \dots, X_n$ μια ακολουθία $n$ μονάδων του πληθυσμού, η οποία προήλθε με τον παραπάνω τρόπο. Η πιθανότητα της πραγματοποίησης αυτής της ακολουθίας είναι:

$$\frac{1}{N}\frac{1}{N-1}\cdots\frac{1}{N-n+1}=\frac{(N-n)!}{N!}$$

Ο τρόπος αυτός υπολογισμού της πιθανότητας λαμβάνει υπόψη τη διάταξη. Για την απλή τυχαία δειγματοληψία χωρίς επανατοποθέτηση, η σειρά εμφάνισης των μονάδων του πληθυσμού στο δείγμα δεν λαμβάνεται υπόψη. Κατά συνέπεια, όλα τα δείγματα τα οποία μπορούν να προκύπτουν με αναδιάταξη των στοιχείων του αρχικού δείγματος είναι ισοδύναμα. Ο συνολικός αριθμός των δειγμάτων αυτών ισούται με $n!$. Συνολικά, η επιλογή της τυχαίας ακολουθίας $X_1, X_2, \dots, X_n$ σύμφωνα με το θεώρημα, έχει πιθανότητα εμφάνισης:

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

η οποία συμφωνεί με τον ορισμό της srs ■

Συνοπτικά, η υλοποίηση μιας απλής τυχαίας δειγματοληψίας γίνεται ως εξής:

- Αντιστοιχίζουμε κάθε μονάδα του πληθυσμού με έναν αριθμό από το 1 έως το $N$.

- Με μια διαδικασία παραγωγής τυχαίων αριθμών (υπολογιστή ή πίνακες τυχαίων αριθμών) επιλέγουμε ένα σύνολο από $n$ τυχαίους αριθμούς. Οι αριθμοί αυτοί θα πρέπει να είναι διακριτοί (όχι ίσοι) και μικρότεροι του $N$.
- Οι μονάδες του πληθυσμού οι οποίες αντιστοιχούν στους τυχαίους αριθμούς που έχουν παραχθεί σύμφωνα με την ανάθεση επιλέγονται για το δείγμα.

### Παράδειγμα 2.1

Σε μια τάξη 30 μαθητών δημοτικού πρόκειται να γίνει μια δειγματοληπτική έρευνα με θέμα την εκτίμηση του αριθμού των παιδιών που δεν έχουν κάνει τα βασικά εμβόλια.

Έστω ότι στο πλαίσιο της έρευνας σκοπεύουμε να επιλέξουμε ένα δείγμα πέντε παιδιών. Όλα τα δυνατά δείγματα τα οποία αποτελούνται από 5 παιδιά της τάξης είναι:

$$\binom{30}{5} = 142506$$

Για την επιλογή του δείγματος, αρχικά αντιστοιχίζεται ένας αριθμός από το 1 έως το 30 με κάθε μαθητή. Στη συνέχεια, επιλέγονται 5 τυχαίοι αριθμοί. Π.χ., με τη βοήθεια του στατιστικού πακέτου `R`, η εντολή η οποία παράγει τους αριθμούς αυτούς θα είναι:

```
> sample(30,5)
```

και το αποτέλεσμα:

```
[1] 23 8 16 4 29
```

δηλώνει ότι ο 4ος, 8ος, 16ος, 23ος και 29ος μαθητής είναι οι μαθητές που θα αποτελούν το δείγμα και θα συμμετέχουν στην έρευνα για την εκτίμηση του ποσοστού των παιδιών τα οποία δεν έχουν κάνει τα βασικά εμβόλια.

## 2.3. Εκτίμηση παραμέτρων του πληθυσμού κάτω από την Απλή Τυχαία Δειγματοληψία

Στην παράγραφο αυτή υποθέτουμε ότι το χαρακτηριστικό του πληθυσμού $Y$ το οποίο αποτελεί το αντικείμενο της δειγματοληπτικής έρευνας είναι ποσοτικό, δηλ. αντιστοιχεί σε μια τυχαία μεταβλητή που είναι συνεχής. Όπως αναφέρθηκε στο Κεφάλαιο 1 -, στην περίπτωση κατά την οποία το χαρακτηριστικό $Y$ είναι συνεχές, οι ποσότητες οι οποίες πιο συχνά ενδιαφέρουν για εκτίμηση είναι η μέση τιμή και το σύνολο του χαρακτηριστικού $Y$ για τον πληθυσμό. Υπενθυμίζουμε ότι $Y_i$ $(i = 1,2, \dots, N)$ συμβολίζει την τιμή ή απάντηση του $i$ μέλους του πληθυσμού για τη μεταβλητή $Y$. Η μέση τιμή και το σύνολο είναι γραμμικοί συνδυασμοί των $Y_i$, οι οποίοι δίνονται πιο αναλυτικά από τις εκφράσεις:

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

και

$$Y_T = \sum_{i=1}^{N} Y_i$$

αντίστοιχα. Στην επόμενη παράγραφο θεωρούμε την περίπτωση της εκτίμησης του μέσου $\bar{Y}$ του πληθυσμού από ένα δείγμα μεγέθους $n$, το οποίο έχει επιλεγεί σύμφωνα με την απλή τυχαία δειγματοληψία.

### 2.3.1 Εκτίμηση του μέσου για το χαρακτηριστικό του πληθυσμού

Για το πρόβλημα της εκτίμησης της μέσης τιμής για το χαρακτηριστικό $Y$ του πληθυσμού, ο δειγματικός μέσος του απλού τυχαίου δείγματος το οποίο έχει επιλεγεί είναι ο εκτιμητής κάτω από την α.τ.δ.

Οι ιδιότητες του εκτιμητή ο οποίος προκύπτει από ένα απλό τυχαίο δείγμα είναι ταυτόχρονα οι ιδιότητες του τρόπου δειγματοληψίας ή του δειγματοληπτικού σχεδίου, όπως επίσης λέγεται, που έχει εφαρμοστεί. Συνεπώς, η μελέτη του εκτιμητή και των ιδιοτήτων του είναι σημαντική για δύο λόγους. Ο πρώτος λόγος αφορά τον ίδιο τον σκοπό της διεξαγωγής της έρευνας. Το αποτέλεσμα της μελέτης θα προσφέρει τα θεωρητικά εργαλεία, η εφαρμογή των οποίων στο δείγμα μιας έρευνας θα επιτρέψει τον αριθμητικό υπολογισμό του εκτιμητή, καθώς και επιπλέον σημαντικών χαρακτηριστικών του, όπως η ακρίβεια ή **διακύμανση (variance)**, **το τυπικό σφάλμα (standard error)** και η κατασκευή **διαστημάτων εμπιστοσύνης (confidence intervals)**. Όλα τα παραπάνω στοιχεία μαζί συνθέτουν τα αποτελέσματα της έρευνας. Ο δεύτερος λόγος είναι ότι, παράλληλα, η μελέτη των ιδιοτήτων του εκτιμητή σε θεωρητικό επίπεδο προσφέρει το μέσο για τη σύγκριση του απλού τυχαίου δειγματοληπτικού σχεδίου με ένα άλλο δειγματοληπτικό σχέδιο γενικότερα.

### 2.3.2 Ιδιότητες του εκτιμητή του μέσου για το χαρακτηριστικό του πληθυσμού

Για τον εκτιμητή του μέσου $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ ενός χαρακτηριστικού $Y$ του πληθυσμού, χρησιμοποιείται ο δειγματικός μέσος $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ ενός δείγματος $s = \{X_1, X_2, \dots, X_n\}$, το οποίο έχει προέλθει με απλό τυχαίο τρόπο. Για τον εκτιμητή αυτόν μπορούν να αποδειχτούν οι παρακάτω προτάσεις.

**Πρόταση 2.2**

Για την απλή τυχαία δειγματοληψία, ο δειγματικός μέσος $\bar{X}$ είναι αμερόληπτος εκτιμητής του $\bar{Y}$, δηλαδή $\mathrm{E}(\bar{X}) = \bar{Y}$.

**Απόδειξη**

Θεωρούμε τις τυχαίες μεταβλητές $V_i, i = 1,2, \dots, N$, η καθεμιά από τις οποίες αντιστοιχεί στις μονάδες του πληθυσμού $Y_i$ και οι οποίες ορίζονται ως εξής:

$$V_i = \begin{cases} 1, \text{εάν η } Y_i \text{ επιλεγεί στο srs δείγμα} \\ 0, \text{εάν η } Y_i \text{ δεν επιλεγεί στο δείγμα} \end{cases}$$

Από τον τρόπο με τον οποίο έχουν οριστεί οι $V_i$ και σύμφωνα με τον ορισμό της απλής τυχαίας δεογματοληψίας, εύκολα προκύπτει ότι $V_i \sim \mathrm{Bernoulli}(p)$ με $p = \frac{n}{N}$.

Η πιθανότητα επιτυχίας $p$ ισούται με $\frac{n}{N}$ γιατί κάθε μονάδα του πληθυσμού μπορεί να καταλάβει μία από τις $n$ θέσεις του δείγματος. Κατά συνέπεια, για τις τυχαίες μεταβλητές (τ.μ.) $V_i$ ισχύει $\mathrm{E}(V_i) = p, i = 1,2, \dots, N$. Επιπλέον, με τη βοήθεια των τ.μ. $V_i$ η δειγματική μέση τιμή γράφεται ισοδύναμα ως:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{N} V_i Y_i \qquad (2.1)$$

δηλ. ένας γραμμικός συνδυασμός των τ.μ. $V_i$ με συντελεστές τις τιμές του χαρακτηριστικού $Y_i$.

Παίρνοντας την αναμενόμενη τιμή στην τελευταία σχέση, προκύπτει:

$$\mathrm{E}(\bar{X}) = \frac{1}{n}\sum_{i=1}^{N} Y_i \mathrm{E}(V_i) = \frac{1}{n}\sum_{i=1}^{N} Y_i \frac{n}{N} = \frac{1}{N}\sum_{i=1}^{N} Y_i = \bar{Y}$$

Άρα ο εκτιμητής $\bar{X}$ εκτιμά αμερόληπτα τον $\bar{Y}$■

### Πρόταση 2.3

Η διακύμανση του εκτιμητή $\bar{X}$ κάτω από την απλή τυχαία δειγματοληψία είναι:

$$\mathrm{Var}(\bar{X}) = \frac{1-f}{n} S^2$$

όπου $f = \frac{n}{N}$ και $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$ η διακύμανση των τιμών του πληθυσμού για το χαρακτηριστικό $Y$.

### Απόδειξη

Χρησιμοποιώντας τις τ.μ. $V_i$ που ορίστηκαν στην απόδειξη της Πρότασης 2.2 και την έκφραση (2.1) για τη διακύμανση του $\bar{X}$, προκύπτει:

$$\mathrm{Var}(\bar{X}) = \frac{1}{n^2}\left[\sum_{i=1}^{N} Y_i^2 \mathrm{Var}(V_i) + 2\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j>i}}^{N} \mathrm{Cov}(V_i, V_j)\right]. \qquad (2.2)$$

Από τον ορισμό των $V_i$ και της Bernoulli κατανομής τους, ισχύουν

(α) $\mathrm{Var}(V_i) = p(1-p)$ και

(β) $\mathrm{Cov}(V_i, V_j) = -\frac{n}{N(N-1)}\left(1 - \frac{n}{N}\right)$

αφού

$$\mathrm{Cov}(V_i, V_j) = \mathrm{E}(V_i V_j) - \mathrm{E}(V_i)\mathrm{E}(V_j).$$

Επίσης, για την $\mathrm{E}(V_i V_j)$ εάν ληφθεί υπόψην ότι ισούται με την πιθανότητα $\mathrm{P}(\{V_i = 1\} \cap \{V_j = 1\})$ μπορεί εύκολα να δειχθεί ότι:

$$\mathrm{E}(V_i V_j) = \frac{n(n-1)}{N(N-1)}$$

Αν αντικατασταθούν τα επιμέρους αποτελέσματα (α) και (β) στην (2.2) και εκτελεστούν οι αλγεβρικές πράξεις στα αθροίσματα, προκύπτει το αποτέλεσμα

## Παρατηρήσεις

**2.2** Η ποσότητα $f = \frac{n}{N}$ ονομάζεται πηλίκο δείγματος (sampling fraction). Από τον ορισμό του, είναι ένας αριθμός από 0 έως 1 και δίνει το μέγεθος του δείγματος σε σχετικούς όρους ως προς τον πληθυσμό. Π.χ. 5%, 10% κτλ.

**2.3** Η ποσότητα $1 - f$ που εμφανίζεται στην έκφραση της διακύμανσης του εκτιμητή λέγεται και διόρθωση του πεπερασμένου πληθυσμού (finite population correction, fpc). Η ονομασία οφείλεται στο γεγονός ότι η ποσότητα αυτή είναι η μόνη που διαφοροποιεί το αποτέλεσμα της $\mathrm{Var}(\bar{X})$ από το αντίστοιχο αποτέλεσμα για άπειρους πληθυσμούς.

**2.4** Εάν ο παράγοντας fpc δεν συμπεριληφθεί στον υπολογισμό της $\mathrm{Var}(\bar{X})$, αυτό θα έχει ως αποτέλεσμα την υπερεκτίμηση της διασποράς του εκτιμητή. Κατά συνέπεια, όσο μεγαλύτερες τιμές έχει το $f = {}^{n}/_{N}$ (γενικά $0 < f < 1$) σε μια δειγματοληπτική έρευνα, τόσο πιο επιβεβλημένη είναι η χρήση του fpc για την$\mathrm{Var}(\bar{X})$. Και αντίθετα, για μικρές τιμές του $f$ η επίδραση του fpc στην $\mathrm{Var}(\bar{X})$ θα είναι μικρή και συνεπώς μπορεί να παραλειφθεί.

**2.5** Από την έκφραση της $\mathrm{Var}(\bar{X})$ όπως αυτή αποδείχτηκε στην Πρόταση 2.3, προκύπτει ότι το δειγματοληπτικό σφάλμα, που προέρχεται από το γεγονός ότι εκτιμούμε το $\bar{Y}$ αντί να το υπολογίζουμε με απογραφή, είναι ανάλογο του $S^2$. Η ποσότητα $S^2$, από τον ορισμό της, είναι ένα μέτρο της μεταβλητότητας των τιμών του συνόλου (εδώ ο πληθυσμός) σε σχέση με τη μέση του τιμή. Άρα, η ακρίβεια που θα επιτευχθεί κατά την εκτίμηση μετά από μια απλή τυχαία δειγματοληψία καθορίζεται από την ομοιογένεια ή ετερογένεια του πληθυσμού ως προς το χαρακτηριστικό $Y$.

Άλλες ποσότητες που απορρέουν από τη διακύμανση του εκτιμητή και είναι χρήσιμες στην πράξη είναι το **τυπικό σφάλμα** του εκτιμητή (**standard error**) $\mathrm{se}(\bar{X}) = \sqrt{\mathrm{Var}(\bar{X})}$ και ο **συντελεστής μεταβλητότητας** του εκτιμητή (**coefficient of variation**) $\mathrm{CV}(\bar{X}) = {\sqrt{\mathrm{Var}(\bar{X})}}/{\mathrm{E}(\bar{X})} = {\mathrm{se}(\bar{X})}/{\bar{Y}}$.

Από την Πρόταση 2.3, γίνεται φανερό ότι για τον αριθμητικό υπολογισμό της διακύμανσης ή του τυπικού σφάλματος του $\bar{X}$ κάτω από την απλή τυχαία δειγματοληψία, απαιτείται προηγούμενη γνώση της διασποράς του πληθυσμού $S^2$. Για τον υπολογισμό του συντελεστή μεταβλητότητας, απαιτείται γνώση τόσο του $S^2$ όσο και της μέσης τιμής $\bar{Y}$ για τον πληθυσμό.

### 2.3.3 Εκτίμηση του δειγματοληπτικού σφάλματος

Η παρακάτω πρόταση προσφέρει ένα αποτέλεσμα με τη βοήθεια του οποίου είναι εφικτό να εκτιμηθεί η διακύμανση του εκτιμητή $\mathrm{Var}(\bar{X})$, όταν δεν υπάρχει προηγούμενη γνώση ή πληροφορία για τη διασπορά του πληθυσμού $S^2$.

**Πρόταση 2.4**

Για την απλή τυχαία δειγματοληψία, ισχύει:

$$\mathrm{E}(s^2) = S^2$$

όπου:

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

**Απόδειξη** (βλ. Des Raj, 1968, Κεφ. 1)

Γράφοντας:

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

και λαμβάνοντας αναμενόμενες τιμές, προκύπτει:

$$(n-1)\mathrm{E}(s^2) = \mathrm{E}\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = \mathrm{E}\left(\sum_{i=1}^{n}X_i^2 - \frac{(\sum_{i=1}^{n}X_i)^2}{n}\right)$$

$$= \mathrm{E}\left(\sum_{i=1}^{n} X_i^2\right) - \frac{1}{n}\mathrm{E}\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] = E\left(\sum_{i=1}^{n} X_i^2\right) - n\mathrm{E}(\bar{X}^2)$$

$$= \mathrm{E}\left(\sum_{i=1}^{n} X_i^2\right) - n\left(\mathrm{Var}(\bar{X}) + \left(\mathrm{E}(\bar{X})\right)^2\right)$$

$$= \mathrm{E}\left(\sum_{i=1}^{n} X_i^2\right) - (1-f)S^2 - n\bar{Y}^2$$

Για την αναμενόμενη τιμή, κάνοντας και πάλι χρήση των τ.μ. $V_i, (i = 1,2,..,N)$ που ορίστηκαν στην απόδειξη της Πρότασης 2.2, ισχύει:

$$\mathrm{E}\left(\sum_{i=1}^{n} X_i^2\right) = \mathrm{E}\left(\sum_{i=1}^{N} V_i Y_i^2\right) = \sum_{i=1}^{N} Y_i^2 \mathrm{E}\,(V_i) = \frac{n}{N}\sum_{i=1}^{N} Y_i^2.$$

Αντικαθιστώντας στην προηγούμενη σχέση την αναμενόμενη τιμή, θα είναι:

$$(n-1)\mathrm{E}(s^2) = \frac{n}{N}\sum_{i=1}^{N} Y_i^2 - (1-f)S^2 - n\bar{Y}^2 = \frac{n}{N}\left(\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2\right) - (1-f)S^2$$

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

$$= (n-1)S^2$$

που αποδεικνύει την πρόταση. ■

Με τη βοήθεια της Πρότασης 2.4 προκύπτουν τα ακόλουθα πορίσματα, τα οποία δίνουν τις εκτιμώμενες ποσότητες για τη διασπορά και το τυπικό σφάλμα του εκτιμητή $\bar{X}$ που έχει προέλθει από ένα απλό τυχαίο δείγμα.

### Πόρισμα 2.1

Για την απλή τυχαία δειγματοληψία, ισχύει ότι ένας αμερόληπτος εκτιμητής της διακύμανσης $\mathrm{Var}(\bar{X})$ είναι ο:

$$\hat{\mathrm{V}}\mathrm{ar}(\bar{X}) = \frac{1-f}{n}s^2.$$

### Πόρισμα 2.2

Ένας εκτιμητής του τυπικού σφάλματος του εκτιμητή $\bar{X}$ δίνεται από τη σχέση:

$$\hat{\mathrm{se}}(\bar{X}) = \sqrt{\frac{1-f}{n}s^2}$$

Αξίζει να σημειωθεί ότι ένας λόγος για τον οποίο έχει χρησιμοποιηθεί ο παρονομαστής $N-1$ αντί για $N$ στον ορισμό του $S^2$ (και αντίστοιχα n-1 αντί n στον ορισμό του s2) είναι η απλότητα του αποτελέσματος της Πρότασης 2.4.

## Παράδειγμα 2.2

Επιλέγονται τυχαία 9 από τους 25 συνολικά γιατρούς μιας πόλης, με σκοπό την εκτίμηση του μέσου αριθμού των επισκέψεων των γιατρών σε σπίτια ασθενών μέσα σε μία εβδομάδα. Οι γιατροί, που επιλέχθηκαν με απλή τυχαία δειγματοληψία, είναι αυτοί με αύξοντα αριθμό 13, 3, 17, 1, 14, 12, 7, 18, 4. Τα δεδομένα του Πίνακα 2.1 που ακολουθεί δίνουν τις απαντήσεις των 9 γιατρών στο ερώτημα της έρευνας:

| **α/α Γιατρών** | 1 | 3 | 4 | 7 | 12 | 13 | 14 | 17 | 18 |
|---|---|---|---|---|---|---|---|---|---|
| **Αριθμ. Επισκέψεων** | 5 | 1 | 4 | 12 | 5 | 6 | 4 | 7 | 0 |

***Πίνακας 2.1*** *Μετρήσεις γιατρών για επισκέψεις ασθενών*

Για τις μετρήσεις του δείγματος, ο δειγματικός μέσος είναι $\bar{X} = \frac{44}{9} = 4.89$ .

Άρα, ο μέσος αριθμός εβδομαδιαίων επισκέψεων σε σπίτια ασθενών για τους γιατρούς της πόλης εκτιμάται σε 4.89.

Για την τυπική απόκλιση της εκτίμησης, μπορεί να υπολογιστεί μια εκτίμηση αυτής, αφού η διακύμανση $S^2$ για τον πληθυσμό, δηλ. τους 25 γιατρούς, είναι άγνωστη. Αρχικά:

$$s^2 = \frac{1}{8}\sum_{i=1}^{9}(X_i - \bar{X})^2 = \frac{1}{8}\left(\sum_{i=1}^{9} X_i^2 - \frac{\left(\sum_{i=1}^{9} X_i\right)^2}{9}\right) = \frac{1}{8}\left(312 - \frac{44^2}{9}\right) = 12.11$$

Άρα, η διασπορά της εκτίμησης $\bar{X} = 4.89$ εκτιμάται ως:

$$\hat{\mathrm{V}}\mathrm{ar}(\bar{X}) = \frac{1-f}{n}s^2 = \frac{1-{}^{9}/_{25}}{9}\,12.11 = 0.86$$

Το τυπικό σφάλμα της εκτίμησης είναι:

$$\widehat{se}(\bar{X}) = \sqrt{0.86} = 0.93$$

Άρα, βάσει του δείγματος, συμπεραίνεται ότι οι γιατροί της πόλης πραγματοποιούν μέσα σε μία εβδομάδα κατά μέσον όρο 4.89 επισκέψεις σε σπίτια ασθενών. Το τυπικό σφάλμα της εκτίμησης είναι 0.93 επισκέψεις.

Για την έρευνα του παραδείγματος, το πηλίκο του δείγματος $f$ είναι 0.36, αρκετά μεγάλο για να μπορεί να παραλειφθεί. Πράγματι, εάν το $f$ αγνοούνταν, η διακύμανση του εκτιμητή θα ήταν:

$$\hat{\mathrm{V}}\mathrm{ar}(\bar{X}) = \frac{1}{n}s^2 = \frac{1}{9}\,12.11 = 1.34$$

αρκετά μεγαλύτερη (υπερεκτίμηση) του 0.86.

### 2.3.4 Εκτίμηση συνόλου για το χαρακτηριστικό του πληθυσμού

Το σύνολο του πληθυσμού ή το άθροισμα των τιμών ενός χαρακτηριστικού για τον πληθυσμό ορίζεται ως το άθροισμα $\sum_{i=1}^{N} Y_i$ και συμβολίζεται με $Y$ ή $Y_T$.

Η εκτίμηση του συνόλου αντί του μέσου ενός χαρακτηριστικού είναι αρκετά συχνά το ερώτημα του ενδιαφέροντος σε μια έρευνα. Για παράδειγμα, έχει ενδιαφέρον να εκτιμηθεί το συνολικό εισόδημα των κατοίκων μιας χώρας, ή η συνολική κατανάλωση σε νερό από τα νοικοκυριά μιας πόλης κτλ.

Από τον ορισμό του $Y_T$ ισχύει ότι $Y_T = N\bar{Y}$και η εκτίμησή του προκύπτει ως άμεση συνέπεια από την εκτίμηση του $\bar{Y}$. Με τη βοήθεια του εκτιμητή τού $\bar{Y}$ και των ιδιοτήτων του, προκύπτουν τα παρακάτω.

**Πρόταση 2.5**

Για την εκτίμηση του συνόλου του πληθυσμού κάτω από την απλή τυχαία δειγματοληψία ισχύουν τα εξής:

(i) Ο εκτιμητής του συνόλου του πληθυσμού είναι ο $X_T = N\bar{X}$

(ii) Ο εκτιμητής $X_T$ εκτιμά αμερόληπτα το $Y_T$, δηλ. $\mathrm{E}(X_T) = Y_T$.

(iii) Η διασπορά του εκτιμητή $X_T$ δίνεται από τη σχέση $\mathrm{Var}(X_T) = N^2 \frac{1-f}{n} S^2$.

(iv) Το τυπικό σφάλμα του $X_T$ είναι $\mathrm{se}(X_T) = N\sqrt{\frac{1-f}{n} S^2}$.

(v) Η εκτιμώμενη διακύμανση και το εκτιμώμενο τυπικό σφάλμα του $\mathrm{X_T}$ δίνονται από τις σχέσεις:

$\hat{\mathrm{V}}\mathrm{ar}(X_T) = N^2 \frac{1-f}{n} s^2$ και $\hat{\mathrm{s}}\mathrm{e}(X_T) = N\sqrt{\frac{1-f}{n} s^2}$ αντίστοιχα.

**Απόδειξη** (Προφανής)

## Παράδειγμα 2.3

Σε συνέχεια του Παραδείγματος 2.2, εάν το ερώτημα της έρευνας είναι πόσες επισκέψεις σε σπίτια ασθενών πραγματοποιούν οι γιατροί της πόλης, τότε ο αριθμός αυτός εκτιμάται από το δείγμα μέσω του $X_T$ και είναι

$X_T = N\bar{X} = 25 \times 4.89 = 122.22$ επισκέψεις.

Το τυπικό σφάλμα της εκτίμησης εκτιμάται ως $\hat{\mathrm{s}}\mathrm{e}(X_T) = N\hat{\mathrm{s}}\mathrm{e}(\bar{X}) = 25 \times 0.93 = 23.20$

Άρα, για τους συνολικά 25 γιατρούς της πόλης, εκτιμάται ότι συνολικά πραγματοποιούν 122.22 επισκέψεις ανά εβδομάδα σε σπίτια ασθενών, με εκτιμώμενο τυπικό σφάλμα 23.20 επισκέψεις.

### 2.3.5 Εκτίμηση ποσοστού για το χαρακτηριστικό του πληθυσμού

Μια ποσότητα, της οποίας πολύ συχνά η εκτίμηση αποτελεί το αντικείμενο δειγματοληπτικής έρευνας, είναι το ποσοστό $P$ των μονάδων του πληθυσμού οι οποίες ανήκουν σε μια κατηγορία, έστω $C$. Για παράδειγμα, το ποσοστό των ανέργων, το ποσοστό των κατοίκων με τουλάχιστον τριτοβάθμια εκπαίδευση, το ποσοστό των γυναικών που φοιτούν στο 1ο έτος μιας πανεπιστημιακής σχολής κτλ.

Αν $N$ είναι το μέγεθος του πληθυσμού και $A$ είναι ο αριθμός των μονάδων του πληθυσμού οι οποίες ανήκουν στην κατηγορία $C$, τότε το προς εκτίμηση ποσοστό $P$ του πληθυσμού ορίζεται ως:

$$P = \frac{A}{N}.$$

Αν θεωρήσουμε τις τυχαίες μεταβλητές $Y_i, i = 1,2,\dots,N$ που αντιστοιχούν σε κάθε μονάδα του πληθυσμού $i = 1,2,\dots,N$ και ορίζονται ως:

$$Y_i = \begin{cases} 1, & \text{αν το } i \text{ μέλος ανήκει στην } C \\ 0, & \text{αν το } i \text{ μέλος δεν ανήκει στην } C \end{cases} \qquad (2.3)$$

(βλ. για παράδειγμα Cochran (1977, Κεφ. 3) και Rao (2000, Κεφ. 4)), τότε το σύνολο του πληθυσμού $\sum_{i=1}^{N} Y_i$ για το χαρακτηριστικό $Y$ ισούται με $A$ και το ποσοστό εκείνων που ανήκουν στην κατηγορία $C$ είναι:

$$P = \frac{A}{N} = \frac{\sum_{i=1}^{N} Y_i}{N} = \bar{Y} \tag{2.4}$$

Συνεπώς, το πρόβλημα της εκτίμησης του ποσοστού $P$ ανάγεται σε ένα πρόβλημα εκτίμησης μέσης τιμής. Η σύνδεση των δύο προβλημάτων είναι η μεταβλητή $Y$, η οποία κωδικοποιεί σε 1 ή 0 («επιτυχία» ή «αποτυχία» αντίστοιχα) τις απαντήσεις της κάθε μονάδας του πληθυσμού στο ερώτημα «ανήκει» ή «δεν ανήκει» στην κατηγορία $C$.

Έχοντας γράψει το ποσοστό $P$ ως μια μέση τιμή, το πρόβλημα της εκτίμησης και των ιδιοτήτων του εκτιμητή από ένα απλό τυχαίο δείγμα μπορούν να εξαχθούν ως άμεση συνέπεια από τη μελέτη του προβλήματος της εκτίμησης μιας μέσης τιμής πληθυσμού στην παράγραφο 2.3.1.Το πρώτο αποτέλεσμα αφορά τον εκτιμητή και δίνεται από το πόρισμα που ακολουθεί.

### Πόρισμα 2.3

Για ένα απλό τυχαίο δείγμα μεγέθους $n$, ο εκτιμητής του ποσοστού $P$ είναι ο $p = \frac{a}{n}$, όπου $a$ είναι οι μονάδες του δείγματος που ανήκουν στην κατηγορία $C$.

### Απόδειξη

Πράγματι, λόγω της σχέσης (2.4), ο εκτιμητής του $P$ ταυτίζεται με τον εκτιμητή του $\bar{Y}$, δηλ. τον:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{a}{n}$$

Ο αριθμητής ισούται με $a$, γιατί από τις $n$ τιμές του δείγματος μόνο οι $a$ θα είναι 1 ενώ οι υπόλοιπες $n - a$ θα είναι μηδέν.

Αναλόγως, για την εκτίμηση του $A$ και επειδή $A = \sum Y_i = Y_T$, δηλ. το σύνολο του πληθυσμού για την $Y$, ο εκτιμητής του θα είναι:

$$\hat{A} = N\frac{a}{n}$$

Το $A$ δηλώνει το πλήθος των μονάδων του πληθυσμού που ανήκουν στην κατηγορία $C$ και αποτελεί αρκετά συχνά στην πράξη το ερώτημα της έρευνας. Για παράδειγμα, στην περίπτωση όπου $P$ είναι το ποσοστό ανεργίας σε μια χώρα, η εκτίμηση του $A$ θα προσφέρει πληροφορία ως προς το πόσοι είναι αριθμητικά οι άνεργοι στη χώρα.

Οι ιδιότητες των εκτιμητών $p$ και $\hat{A}$ μπορούν να εξαχθούν επίσης ως συνέπειες των αντιστοίχων ιδιοτήτων των $\bar{X}$και $X_T$. Η παρακάτω πρόταση προσθέτει ένα ενδιάμεσο και χρήσιμο αποτέλεσμα για τη διασπορά $S^2$, στην περίπτωση που οι μεταβλητές $Y_i$ ορίζονται όπως στην παρούσα παράγραφο (με τιμές 0 ή 1).

### Πρόταση 2.6

Εάν οι τ.μ. $Y_i$ $(i = 1,2, \dots, N)$ ορίζονται σύμφωνα με την (2.3), τότε η διασπορά του πληθυσμού $S^2$ δίνεται μέσω του ποσοστού $P$ από τη σχέση:

$$S^2 = \frac{NP(1-P)}{N-1}$$

### Απόδειξη

Από τον ορισμό της, η διασπορά του πληθυσμού ισούται με:

$$S^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} Y_i^2 - \frac{\left(\sum_{i=1}^{N} Y_i\right)^2}{N}\right)$$

Για την περίπτωση που οι $Y_i$ ορίζονται όπως στην (2.3) και παίρνουν τιμές μόνο 0 και 1, ισχύει ότι $\sum_{i=1}^{N} Y_i^2 = \sum_{i=1}^{N} Y_i = A$.

Συνεπώς,

$$S^2 = \frac{1}{N-1}\left(A - \frac{(A)^2}{N}\right) = \frac{1}{N-1}\left(NP - \frac{(NP)^2}{N}\right) = \frac{NP(1-P)}{N-1}$$

που ολοκληρώνει την απόδειξη ■

Τονίζεται και πάλι ότι το αποτέλεσμα της Πρότασης 2.6 ισχύει μόνον εφόσον η $Y$ είναι δίτιμη διακριτή μεταβλητή.

Με τη βοήθεια της Πρότασης 2.6 μπορούν εύκολα να αποδειχθούν τα παρακάτω πορίσματα.

### Πόρισμα 2.4

Ο εκτιμητής $p$ του άγνωστου ποσοστού $P$, ο οποίος υπολογίζεται από ένα απλό τυχαίο δείγμα μεγέθους $n$, έχει τις ιδιότητες:

- Είναι αμερόληπτος εκτιμητής του $P$, δηλ. $\mathrm{E}(p) = P$
- $\mathrm{Var}(p) = \frac{1-f}{n}\frac{NP(1-P)}{N-1}$
- $\mathrm{se}(p) = \sqrt{\frac{1-f}{n}\frac{NP(1-P)}{N-1}}$

Ανάλογο αποτέλεσμα με εκείνο της Πρότασης 2.6 για την $S^2$ μπορεί να αποδειχθεί και για τη δειγματική διασπορά $s^2$. Πιο συγκεκριμένα, αποδεικνύεται με ακριβώς ανάλογο τρόπο όπως στην απόδειξη της Πρότασης 2.6, ότι:

$$s^2 = \frac{np(1-p)}{n-1}$$

Με τη βοήθεια του αποτελέσματος αυτού και των Πορισμάτων 2.1 και 2.2 προκύπτει το παρακάτω

### Πόρισμα 2.5

Η εκτιμώμενη διασπορά και το εκτιμώμενο τυπικό σφάλμα του $p$ δίνονται από τις εκφράσεις:

$$\hat{\mathrm{V}}\mathrm{ar}(p) = (1-f)\frac{p(1-p)}{n-1}$$

και

$$\hat{\mathrm{s}}\mathrm{e}(p) = \sqrt{(1-f)\frac{p(1-p)}{n-1}}$$

αντίστοιχα.

## Παράδειγμα 2.4

Από ένα πελατολόγιο που αποτελείται από $N = 3042$ ονόματα και διευθύνσεις πελατών, εκλέγεται με τυχαίο τρόπο ένα δείγμα μεγέθους $n = 200$. Από τους 200 πελάτες, διαπιστώθηκε ότι οι 38 είχαν στο μεταξύ αλλάξει διεύθυνση.

Αν το ενδιαφέρον του ιδιοκτήτη του πελατολογίου είναι να γνωρίζει σε τι ποσοστό τα στοιχεία των πελατών του χρειάζονται ανανέωση γιατί έχουν τροποποιηθεί, τότε βάσει του δείγματος το ποσοστό αυτό είναι:

$$p = \frac{a}{n} = \frac{38}{200} = 0.19$$

Δηλαδή 19% των πελατών έχει αλλάξει διεύθυνση. Το τυπικό σφάλμα της εκτίμησης εκτιμάται με:

$$\hat{se}(p) = \sqrt{(1-f)\frac{p(1-p)}{n-1}} = \sqrt{\left(1-\frac{200}{3042}\right)\frac{0.19*(1-0.19)}{199}} = 0.0269$$

Για το συγκεκριμένο παράδειγμα, $f = 0.065$ (μικρό) και αν δεν συμπεριληφθεί στον παραπάνω υπολογισμό, θα δώσει $\hat{se}(p) = 0.0278$, δηλ. όχι μακρυά από το $0.0269$.

Όσον αφορά την εκτίμηση του συνολικού αριθμού των διευθύνσεων του πελατολογίου για τις οποίες εκτιμάται ότι έχουν αλλάξει και δεν ισχύουν πλέον, είναι:

$$\hat{A} = N\frac{a}{n} = 3042\frac{38}{200} = 577.98 \cong 578$$

με τυπικό σφάλμα εκτίμησης:

$$\hat{se}\left(\hat{A}\right) = N\hat{se}(p) = 3042 * 0.0269 = 81.7\ .$$

## Παρατήρηση 2.6

Το πρόβλημα της εκτίμησης ποσοστού, όπως αυτό έχει περιγραφεί στην παρούσα παράγραφο, είναι προφανές ότι υφίσταται σε περιπτώσεις κατά τις οποίες το υπό μελέτη χαρακτηριστικό $Y$ για τον πληθυσμό παίρνει δύο δυνατές τιμές, $C$ και το συμπλήρωμά του $\bar{C}$. Με άλλα λόγια, όταν το $Y$ είναι μια δίτιμη κατηγορική μεταβλητή. Το πρόβλημα παρόλα αυτά ισχύει και αντιμετωπίζεται όμοια και για την πιο γενική περίπτωση, όταν δηλ. η $Y$ είναι μια κατηγορική μεταβλητή με περισσότερα από δύο επίπεδα.

Γενικά, εάν $Y$ είναι ένα χαρακτηριστικό που αντιστοιχεί σε μια κατηγορική μεταβλητή με $k$ επίπεδα, έστω $C_1, C_2, \dots, C_k$, ορίζονται τα ποσοστά $P_1, P_2, \dots, P_k$ , με $P_1 + P_2 + \cdots + P_k = 1$, όπου $P_i$το ποσοστό των μελών του πληθυσμού που ανήκουν στην κατηγορία $C_i$ $(i = 1, 2, \dots, k)$. Για την εκτίμηση ενός από τα ποσοστά $P_i$, έστω του $P_r$ $(r = 1,2, \dots, k)$, μπορεί να γίνει χρήση της θεωρίας που αναπτύχθηκε, αρκεί να μετατραπεί η $Y$ σε δίτιμη, όπου $C = C_r$ και $\bar{C}$ η ένωση όλων των κατηγοριών εκτός της $C_r$.

Για παράδειγμα, μια έρευνα θα μπορούσε να στοχεύει στην εκτίμηση του ποσοστού των καπνιστών μιας πόλης. Στην περίπτωση αυτή, οι κάτοικοι χωρίζονται σε 2 κατηγορίες, καπνιστές και μη καπνιστές, και η μεταβλητή $Y$, το κάπνισμα, είναι δίτιμη. Μια έρευνα όμως με παρόμοιο θέμα μπορεί να έχει ως στόχο την εκτίμηση του ποσοστού των πολύ έντονα καπνιστών, εάν θεωρήσει κανείς 4 διαβαθμίσεις στον τύπο καπνίσματος Α: καθόλου (0 τσιγάρα ανά μέρα), Β: σπάνια (μέχρι 2 τσιγάρα την ημέρα), Γ: μέτριος καπνιστής (από 2 έως 15 τσιγάρα) και Δ: έντονα καπνιστής (πάνω από 15 τσιγάρα την ημέρα).

### 2.3.6 Αξιοπιστία εκτιμητών

Το τυπικό σφάλμα των εκτιμητών είναι ένα μέτρο της δειγματοληπτικής μεταβλητότητας ενός εκτιμητή σε σχέση με τους εκτιμητές που προκύπτουν από όλα τα δυνατά δείγματα. Κάτω από την υπόθεση ότι δεν

υπεισέρχονται άλλου είδους σφάλματα σε μια έρευνα, όπως για παράδειγμα σφάλματα μέτρησης, τότε η αξιοπιστία ενός εκτιμητή μετριέται από το μέγεθος του τυπικού του σφάλματος. Όσο μεγαλύτερο είναι το τυπικό σφάλμα ενός εκτιμητή, τόσο λιγότερο αξιόπιστος είναι.

Ένα άλλο μέτρο αξιοπιστίας για έναν εκτιμητή είναι ο συντελεστής μεταβλητότητας. Σε αναλογία με τον ορισμό που δόθηκε στην παράγραφο 2.3.2 για τον εκτιμητή της μέσης τιμής του πληθυσμού, ο συντελεστής μεταβλητότητας για τον εκτιμητή $\hat{\theta}$ που εκτιμά την παράμετρο $\theta$ του πληθυσμού (π.χ. $\bar{Y}, Y_T, P$) ορίζεται ως:

$$\mathrm{CV}(\hat{\theta}) = \frac{\mathrm{se}(\hat{\theta})}{\theta}$$

Ως μέτρο, το $\mathrm{CV}(\hat{\theta})$ μετρά τη δειγματοληπτική μεταβλητότητα ενός εκτιμητή σε σχέση με την ποσότητα που εκτιμά. Παρατηρείστε ότι ενώ το τυπικό σφάλμα ενός εκτιμητή είναι ένα μέγεθος το οποίο μετριέται σε μονάδες, αυτές του χαρακτηριστικού $Y$, ο συντελεστής μεταβλητότητας είναι αδιάστατο μέγεθος, δηλ. δεν έχει μονάδες.

Τέλος, προσθέτοντας μια υπόθεση για την κατανομή του εκτιμητή, ένα εναλλακτικό μέτρο αξιοπιστίας είναι το διάστημα εμπιστοσύνης για την παράμετρο $\theta$ το οποίο κατασκευάζεται με τη βοήθεια του εκτιμητή $\hat{\theta}$. Όσο πιο μικρό μήκος έχει ένα διάστημα, τόσο πιο αξιόπιστος είναι ο εκτιμητής. Η κατασκευή διαστημάτων εμπιστοσύνης για τους εκτιμητές των ποσοτήτων του πληθυσμού οι οποίοι έχουν αναφερθεί έως τώρα παρουσιάζεται στην επόμενη παράγραφο.

## **Παράδειγμα 2.5** (Δαμιανού, 2006, Κεφ. 2)

Από τα 300 σχολεία μιας περιοχής, επιλέχθηκε ένα τυχαίο δείγμα μεγέθους 50 σχολείων με σκοπό τη διερεύνηση του ποσοστού των παιδιών που απουσιάζουν από το σχολείο μια συγκεκριμένη ημέρα. Σε δέκα σχολεία βρέθηκε να απουσιάζουν 1, 4, 3, 1, 2, 5, 2, 7, 8, 4 μαθητές αντίστοιχα και στα άλλα σχολεία κανένας.

Έστω ότι αρχικά ενδιαφερόμαστε να απαντήσουμε στην ερώτηση, πόσα σχολεία παρουσιάζουν τουλάχιστον μία απουσία στην πόλη, ποιο το τυπικό σφάλμα της εκτίμησης και ποιος είναι ο συντελεστής μεταβλητότητας.

Ο αριθμός των σχολείων στα οποία απουσίαζε τουλάχιστον ένας μαθητής τη δεδομένη ημέρα εκτιμάται με:

$$\hat{A} = Np = 300\,\frac{a}{n} = 300\frac{10}{50} = 300 \times 0.20 = 60$$

Για την εκτίμηση αυτή, λαμβάνεται υπόψη μόνο εάν σ' ένα σχολείο παρατηρούνται απουσίες ή όχι, χωρίς περαιτέρω ανάλυση για τον συγκεκριμένο αριθμό απουσιών.

Το τυπικό σφάλμα της εκτίμησης είναι:

$$\widehat{\mathrm{se}}(\hat{A}) = N\sqrt{(1-f)\frac{p(1-p)}{n-1}} = 300\sqrt{\left(1-\frac{50}{300}\right)\frac{0.2\times 0.8}{49}} = 15.65$$

Ο συντελεστής μεταβλητότητας της εκτίμησης είναι:

$$\mathrm{CV}(\hat{A}) = \frac{\widehat{\mathrm{se}}(\hat{A})}{\hat{A}} = \frac{15.65}{60} = 0.26$$

Έστω ότι τώρα ενδιαφερόμαστε να εκτιμήσουμε τον συνολικό αριθμό απουσιών και να υπολογίσουμε το τυπικό σφάλμα και τον συντελεστή μεταβλητότητας για την εκτίμηση.

Για την εκτίμηση του συνολικού αριθμού των παιδιών που απουσιάζουν εκείνη τη μέρα, χρειάζεται να εκτιμηθεί αρχικά ο μέσος αριθμός παιδιών που απουσιάζουν ανά σχολείο. Αυτός είναι:

$$\bar{X} = \frac{\sum_{i=1}^{50} X_i}{n}$$

όπου $X_i$ οι απουσίες μαθητών που παρατηρήθηκαν στο $i$ σχολείο ($i = 1,2,\ldots,50$).

Βάσει των μετρήσεων που δόθηκαν, θα είναι:

$$\bar{X} = \frac{37}{50} = 0.74$$

Για τον συνολικό αριθμό των μαθητών που απουσιάζουν στο σύνολο των 300 σχολείων, θα είναι:

$$X_T = N\,\bar{X} = 300 \times 0.74 = 222$$

Άρα, 222 μαθητές εκτιμάται βάσει του δείγματος ότι απουσιάζουν την ημέρα αυτή συνολικά από όλα τα σχολεία.

Το τυπικό σφάλμα της εκτίμησης είναι:

$$\widehat{se}(X_T) = 300\sqrt{\frac{(1-f)}{n}s^2}$$

όπου, βάσει του δείγματος και πάλι,

$$s^2 = \frac{1}{49}\left(\sum_{i=1}^{50} X_i^2 - \frac{\left(\sum_{i=1}^{50} X_i\right)^2}{50}\right) = \frac{1}{49}\left(189 - \frac{37^2}{50}\right) = 3.29$$

Άρα συνολικά:

$$\widehat{se}(X_T) = 300\sqrt{\frac{(1-50/300)}{50}3.29} = 70.24$$

Συνεπώς, οι συνολικές απουσίες εκτιμώνται σε 222 και τυπικό σφάλμα εκτίμησης 70.24 απουσίες.

Ο συντελεστής μεταβλητότητας για την εκτίμηση αυτή είναι:

$$\mathrm{CV}(X_T) = \frac{\widehat{se}(X_T)}{X_T} = \frac{70.24}{222} = 0.32$$

## 2.4. Διαστήματα Εμπιστοσύνης

Στην παράγραφο αυτή υιοθετείται η προσέγγιση του διαστήματος εμπιστοσύνης για την εκτίμηση μιας παραμέτρου του πληθυσμού. Δοθέντος ενός δείγματος $(X_1, X_2, \dots, X_n)$, το οποίο έχει προέλθει με απλή τυχαία δειγματοληψία, το ερώτημα είναι πώς κατασκευάζεται το διάστημα εμπιστοσύνης για τις παραμέτρους $\bar{Y}, Y_T, P, A$ του πληθυσμού. Η αντιμετώπιση του προβλήματος θα είναι ενιαία.

Έστω $\theta = \bar{Y}$ η πληθυσμιακή μέση τιμή και $\hat{\theta} = \bar{X}$ ο δειγματικός μέσος ο οποίος αποτελεί τον εκτιμητή για την παράμετρο $\theta$ κάτω από την απλή τυχαία δειγματοληψία. Η δειγματική μέση τιμή $\bar{X}$, ως άθροισμα ανεξάρτητων τυχαίων μεταβλητών, και για μέγεθος δείγματος μεγάλο, ακολουθεί προσεγγιστικά την κανονική κατανομή. Το αποτέλεσμα αυτό ισχύει για άπειρους πληθυσμούς, με μόνη υπόθεση, εκτός από το μέγεθος δείγματος $n$, η κατανομή του πληθυσμού να έχει πεπερασμένη διασπορά.

Το ίδιο αποτέλεσμα, για να ισχύει στην περίπτωση ενός πεπερασμένου πληθυσμού, απαιτεί επιπλέον υποθέσεις. Οι υποθέσεις αυτές σχετίζονται περισσότερο με τη συμμετρία ή λοξότητα της κατανομής του πληθυσμού και παρέχουν το ικανό μέγεθος δείγματος $n$, για το οποίο η υπόθεση της κανονικότητας για τον δειγματικό μέσο ισχύει. Αν ο πληθυσμός είναι συμμετρικός, τότε οι ίδιοι κανόνες με τους άπειρους πληθυσμούς ισχύουν και στους πεπερασμένους. Αρκετά συχνά όμως, σε πεπερασμένους πληθυσμούς το ιστόγραμμα των τιμών $Y_i$ παρουσιάζει μεγάλη λοξότητα. Για παράδειγμα, εάν το χαρακτηριστικό $Y$ είναι εισόδημα, τότε περιμένουμε αρκετές τιμές $Y_i$ να είναι μικρές, και ορισμένες μόνο να είναι αρκετά μεγαλύτερες, με αποτέλεσμα να δημιουργείται έντονη δεξιά λοξότητα. Μέσα σ' αυτό το πλαίσιο, ένας κανόνας που πρακτικά προτείνεται για τον προσδιορισμού του ικανού μεγέθους $n$ είναι:

$$n \geq 25\, G_1^2$$

όπου $G_1 = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^3}{NS^3}$ (Cochran, 1977). Το $G_1$ μπορεί να εκτιμηθεί, εφόσον κρίνεται απαραίτητο, με στοιχεία από μια προηγούμενη έρευνα για τον ίδιο πληθυσμό και το χαρακτηριστικό $Y$, ή με τη βοήθεια μιας πιλοτικής έρευνας. Οι Sugden, Smith, & Jones (2000) επεκτείνουν το παραπάνω αποτέλεσμα προτείνοντας εναλλακτικά:

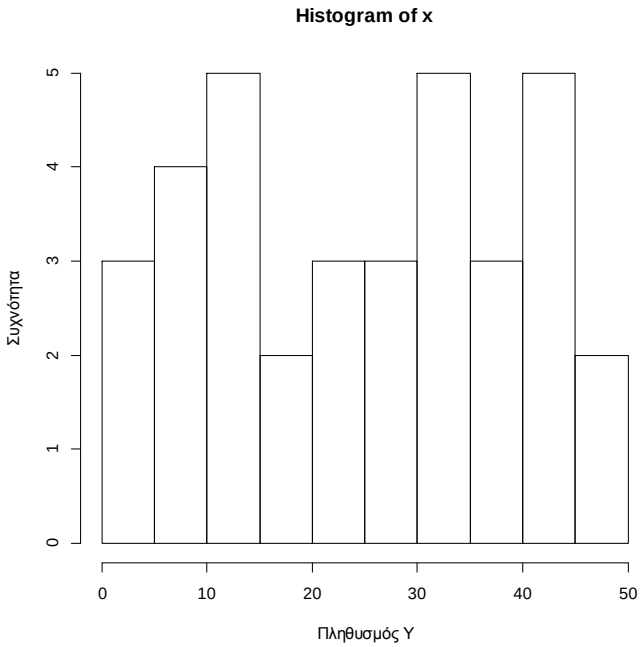
$$n \geq 28 + 25\, G_1^2$$

Στην πράξη, σε πληθυσμούς που δεν παρουσιάζουν έντονη λοξότητα, η υπόθεση της κανονικότητας επιτυγχάνεται αρκετά εύκολα και με σχετικά μικρό μέγεθος δείγματος, όπως φαίνεται στο παράδειγμα που ακολουθεί.

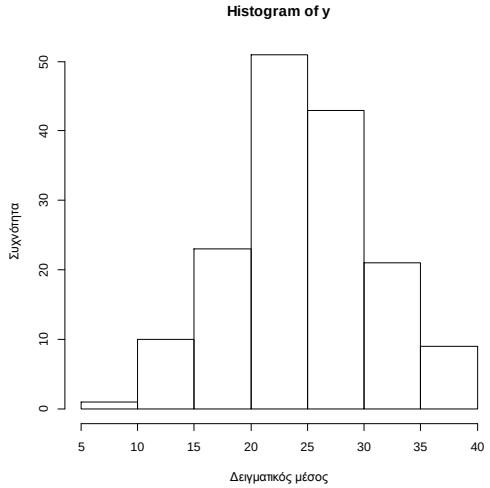
## Παράδειγμα 2.6

Το Σχήμα 2.1 είναι το ιστόγραμμα των 35 τιμών ενός πληθυσμού με εύρος τιμών από 0 έως 50. Έστω ότι η παράμετρος προς εκτίμηση είναι η μέση τιμή του πληθυσμού και για τον σκοπό αυτό λαμβάνεται ένα δείγμα μεγέθους $n = 5$ και έστω $\bar{X} = \sum_{i=1}^{5} X_i/5$ η δειγματική μέση τιμή που προκύπτει από το δείγμα των 5 μετρήσεων και η οποία χρησιμεύει ως ο εκτιμητής για την αληθινή μέση τιμή $\bar{Y}$. Ας προσπαθήσουμε να διερευνήσουμε τη δειγματική κατανομή του $\bar{X}$ γραφικά. Τα δυνατά δείγματα μεγέθους $n = 5$ που μπορούν να προκύψουν με απλή τυχαία δειγματοληψία είναι σε πλήθος $\binom{35}{5} = 324632$. Κάθε δείγμα δίνει μία τιμή για τον $\bar{X}$ και, αν εξαντλούσαμε πλήρως τα δείγματα, οι 324632 τιμές των μέσων οι οποίες θα προέκυπταν, θα έδιναν τη δειγματική κατανομή για τον εκτιμητή $\bar{X}$.

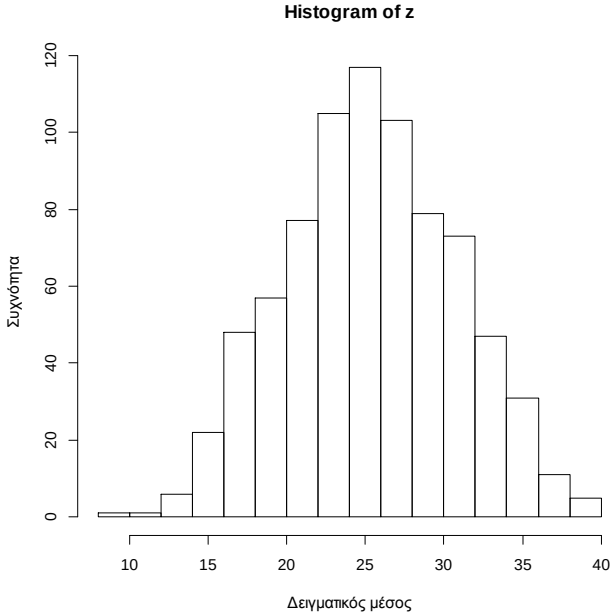
Έστω ότι επιλέγουμε 1000 τέτοια δείγματα από τα 324632 συνολικά, υπολογίζουμε τη δειγματική μέση τιμή για το καθένα και κάνουμε ένα ιστόγραμμα αυτών των μέσων τιμών. Το ιστόγραμμα αυτό παρουσιάζεται στο Σχήμα 2.2 και αποτελεί μια εκτίμηση της κατανομής του $\bar{X}$ βασισμένη σε 1000 πραγματοποιήσεις του. Ως προς το σχήμα του ιστογράμματος, παρατηρούμε ότι αυτό προσεγγίζει μια μονοκόρυφη και συμμετρική κατανομή. Επαναλαμβάνουμε το πείραμα για 5000 δείγματα και το Σχήμα 2.3 δίνει το νέο ιστόγραμμα των δειγματικών μέσων. Η ίδια παρατήρηση ως προς το σχήμα του ιστογράμματος ισχύει και είναι ακόμα πιο έντονη. Παρατηρείται συνεπώς για το παράδειγμα ότι, παρότι οι τιμές του πληθυσμού δεν πλησιάζουν την κανονικότητα, η κατανομή του δειγματικού μέσου για δείγμα μεγέθους μόλις 5 είναι πολύ κοντά στην κανονική.



***Σχήμα 2.1*** *Ιστόγραμμα των τιμών του πληθυσμού.*




***Σχήμα 2.2*** *Ιστόγραμμα των εκτιμητών της μέσης τιμής για 1000 επαναλήψεις απλού τυχαίου δείγματος μεγέθους* $n = 5$.




***Σχήμα 2.3*** *Ιστόγραμμα των εκτιμητών της μέσης τιμής για 5000 επαναλήψεις απλού τυχαίου δείγματος μεγέθους* $n = 5$.

Στη συνέχεια, βασιζόμενοι στην υπόθεση της κανονικότητας, θα δώσουμε ένα διάστημα εμπιστοσύνης (ΔΕ) για τη μέση τιμή του πληθυσμού $\bar{Y}$.

Το διάστημα αυτό, το οποίο καλύπτει την αληθινή τιμή $\bar{Y}$ με πιθανότητα $1 - a$ $(0 < a < 1)$, είναι το:

$$\left(\bar{X} - z_{a/2}S\sqrt{\frac{1-f}{n}}, \qquad \bar{X} + z_{a/2}S\sqrt{\frac{1-f}{n}}\right) \tag{2.5}$$

όπου $z_a$ είναι το άνω α-εκατοστιαίο σημείο για την τυπική κανονική, δηλ. $\mathrm{P}(Z \geq z_a) = a$ για $Z \sim N(0,1)$. Συνηθισμένες τιμές του $a$ είναι 5% και 1%. Για τις τιμές αυτές, $z_{0.975} = 1.96 \cong 2$, και $z_{0.005} = 2.56 \cong 2.5$.

Το ανωτέρω διάστημα είναι συμμετρικό λόγω της συμμετρικότητας της κανονικής κατανομής, έχει κέντρο τον εκτιμητή $\bar{X}$ και μήκος το διπλάσιο του γινομένου του $z_{a/2}$ εκατοστιαίου σημείου με το τυπικό σφάλμα του εκτιμητή. Το τυπικό σφάλμα του εκτιμητή συνεπώς καθορίζει το μήκος του διαστήματος εμπιστοσύνης και για το λόγο αυτό, τόσο το τυπικό σφάλμα, όσο και το διάστημα εμπιστοσύνης, είναι και τα δύο μέτρα αξιοπιστίας της εκτίμησης.

Για τον υπολογισμό του διαστήματος (2.5), χρειάζεται η διασπορά των μετρήσεων του πληθυσμού $S^2$ να είναι γνωστή. Αν αυτό δεν είναι εφικτό, τότε χρησιμοποιείται ο αμερόληπτος δειγματικός εκτιμητής $s^2$. Στην περίπτωση αυτή, αντί για την κανονική κατανομή $Z$ χρησιμοποιείται η $t_{n-1}$ (Student's t) κατανομή. Έτσι, το αντίστοιχο $(1-a)100\%$ διάστημα εμπιστοσύνης για $S^2$ άγνωστο είναι:

$$\left(\bar{X} - z_{a/2}S\sqrt{\frac{1-f}{n}}, \qquad \bar{X} + z_{a/2}S\sqrt{\frac{1-f}{n}}\right) \tag{2.6}$$

Σημειώνεται ότι αν $n \geq 30$ η $t_n$ κατανομή προσεγγίζεται ικανοποιητικά από την $Z$, δηλ. $t_{n-1,a/2} = z_{a/2}$ στο διάστημα εμπιστοσύνης το οποίο δίνεται από την (2.6).

## Παράδειγμα 2.7

Για τα δεδομένα του Παραδείγματος 2.2, έχει βρεθεί ότι ο εκτιμητής της μέσης τιμής του πληθυσμού είναι $\bar{X} = 4.89$, με εκτιμώμενη διακύμανση $\widehat{\mathrm{Var}}(\bar{X}) = 0.86$ ή εκτιμώμενο τυπικό σφάλμα $\widehat{\mathrm{se}}(\bar{X}) = 0.93$.

Ένα διάστημα εμπιστοσύνης με βαθμό 95% υπολογίζεται με τη βοήθεια της (2.6) γιατί το μέγεθος δείγματος είναι μικρό (n = 9). Θα είναι συνεπώς:

$$\left(4.89 - t_{8,0.025}\ 0.93,\ 4.89 + t_{8,0.025}\ 0.93\right)$$

$$= (4.89 - 2.31 \times 0.93,\ 4.89 + 2.31 \times 0.93)$$

$$= (2.75, 7.03).$$

Δηλαδή, εκτιμούμε ότι η αληθινή μέση τιμή περιέχεται στο διάστημα $(2.75, 7.03)$, με πιθανότητα σφάλματος 5%.

Ανάλογα με τη μέση τιμή του πληθυσμού, μπορεί να κατασκευαστεί και το διάστημα εμπιστοσύνης για το σύνολο $Y_T$, το ποσοστό $P$ και τον αριθμό μελών του πληθυσμού $A$. Για εκτιμώμενη (όχι ακριβή) διασπορά εκτιμητών και για μέγεθος δείγματος μεγάλο ($n \geq 30$), το z-διάστημα εμπιστοσύνης των ποσοτήτων αυτών θα είναι:

- ΔΕ για το σύνολο $Y_T$: $\left(N\bar{X} - z_{a/2}Ns\sqrt{\frac{1-f}{n}},\ N\bar{X} + z_{a/2}Ns\sqrt{\frac{1-f}{n}}\right)$
- ΔΕ για το σύνολο $P$: $\left(p - z_{a/2}\sqrt{\frac{(1-f)p(1-p)}{n-1}},\ p + z_{a/2}\sqrt{\frac{(1-f)p(1-p)}{n-1}}\right)$
- ΔΕ για το σύνολο $A$: $\left(N\,p - z_{a/2}\,N\sqrt{\frac{(1-f)p(1-p)}{n-1}},\ N\,p + z_{a/2}\,N\sqrt{\frac{(1-f)p(1-p)}{n-1}}\right)$

Για τις περιπτώσεις των εκτιμητών $P$ και $A$, επειδή η κατανομή των εκτιμητών είναι διακριτή και προσεγγίζεται με μια συνεχή, κάνουμε χρήση της διόρθωσης της συνεχείας προκειμένου να βελτιωθεί το αποτέλεσμα της προσέγγισης. Για παράδειγμα, το $(1-a)100\%$ ΔΕ για το $P$ θα δίνεται από τη σχέση:

$$\left(p - \left[z_{a/2}\sqrt{\frac{(1-f)p(1-p)}{n-1}} + \frac{1}{2n}\right], \qquad p + \left[z_{a/2}\sqrt{\frac{(1-f)p(1-p)}{n-1}} + \frac{1}{2n}\right]\right)$$

Σημειώνεται ότι, θεωρητικά, το διάστημα εμπιστοσύνης για τις παραμέτρους $P$ και $A$ μπορεί να υπολογιστεί και από την ακριβή κατανομή των εκτιμητών τους, αντί της προσέγγισης μέσω της κανονικής. Εάν η δειγματοληψία γίνεται χωρίς επανατοποθέτηση, η ακριβής κατανομή της τ.μ. $\alpha$, του αριθμού των μελών του δείγματος τα οποία ανήκουν στην κατηγορία $C$, είναι η υπεργεωμετρική $\mathrm{Hg}(N, n, P)$. Στη συνέχεια, από τις σχέσεις $p = \frac{a}{n}$ και $\hat{A} = N\frac{a}{n}$, υπολογίζονται οι ακριβείς κατανομές των εκτιμητών των $P$ και $A$ αντίστοιχα. Όταν η απλή τυχαία δειγματοληψία διεξάγεται με επανατοποθέτηση, η κατανομή του $\alpha$ είναι η διωνυμική $\mathrm{B}(n, P)$.

Στην πράξη, παρότι η αναγνώριση των κατανομών του $\alpha$ είναι εφικτή και στις δύο περιπτώσεις, το διάστημα εμπιστοσύνης για τα $P$ και $A$ κατασκευάζεται συνήθως από την κανονική κατανομή, κυρίως λόγω της δυσκολίας που παρουσιάζει η κατασκευή ΔΕ με χρήση διακριτών τυχαίων μεταβλητών.

## Παράδειγμα 2.8

Σε συνέχεια του Παραδείγματος 2.4, εναλλακτικά με την εκτίμηση και το τυπικό σφάλμα για το ποσοστό των πελατών με λάθος διεύθυνση στο πελατολόγιο, το 99% διάστημα εμπιστοσύνης για την ίδια ποσότητα είναι:

$$p \pm \left[ z_{a/2}\, \hat{se}(p) + \frac{1}{2n} \right]$$

Χρησιμοποιώντας τα αποτελέσματα που βρέθηκαν μέσω της έρευνας, και συγκεκριμένα $p = 0.19$, $\hat{se}(p) = 0.0269$ , $z_{a/2} = 2.58$ και $n = 200$, το 99% διάστημα εμπιστοσύνης για τους πελάτες οι οποίοι έχουν αλλάξει διεύθυνση είναι $(0.12, 0.26)$.

Αντίστοιχα, για τον συνολικό αριθμό των πελατών $A$ που έχουν αλλάξει διεύθυνση, η εκτίμησή του βρέθηκε ότι είναι $\hat{A} = 578$ και $\hat{se}(\hat{A}) = 81.7$.

Ένα 99% ΔΕ για τον ίδιο αριθμό θα είναι:

$$Np \pm \left[ z_{a/2}\, N\, \hat{se}(p) + \frac{1}{2n} \right]$$

που δίνει αποτέλεσμα (366.86, 789.10).

## Παράδειγμα 2.9

Για την εκτίμηση του μέσου χρόνου αναμονής των τηλεφωνικών κλήσεων των πελατών μιας εταιρείας στη γραμμή εξυπηρέτησης, επιλέγονται τυχαία 15 κλήσεις μέσα σε μία ημέρα και καταγράφεται η διάρκεια αναμονής (σε λεπτά). Τα δεδομένα δίνονται στον πίνακα που ακολουθεί. Υποθέτοντας ότι ο συνολικός αριθμός των κλήσεων που δέχεται το τηλεφωνικό κέντρο της εταιρείας είναι πολύ μεγάλος, η εταιρεία ενδιαφέρεται να γνωρίζει ένα 95% διάστημα εμπιστοσύνης του μέσου χρόνου αναμονής των πελατών της.

| **α/α κλήσης** | **Διάρκεια αναμονής (min)** | **α/α κλήσης** | **Διάρκεια αναμονής (min)** |
|---|---|---|---|
| 1 | 9.44 | 9 | 25.46 |
| 2 | 24.25 | 10 | 7.05 |
| 3 | 20.49 | 11 | 11.40 |
| 4 | 14.40 | 12 | 19.33 |
| 5 | 14.20 | 13 | 7.08 |
| 6 | 19.51 | 14 | 9.58 |

| α/α κλήσης | Διάρκεια αναμονής (min) | α/α κλήσης | Διάρκεια αναμονής (min) |
|---|---|---|---|
| 7 | 6.53 | 15 | 25.18 |
| 8 | 5.03 | **Σύνολο** | **218.93** |

***Πίνακας 2.2*** *Μετρήσεις χρόνου αναμονής (σε λεπτά) στο τηλεφωνικό κέντρο.*

Ο δειγματικός μέσος χρόνος αναμονής στο τηλεφωνικό κέντρο θα είναι:

$$\bar{X} = \frac{218.93}{15} = 14.59$$

Άρα, εκτιμάται ότι οι πελάτες περιμένουν κατά μέσον όρο 14.59 λεπτά της ώρας. Για τις μετρήσεις του Πίνακα 2.2 υπολογίζεται εύκολα ότι $\sum X_i^2 = 3932.056$ και στη συνέχεια:

$$s^2 = 52.62$$

Επειδή η διακύμανση του πληθυσμού $S^2$ είναι άγνωστη και το μέγεθος του δείγματος σχετικά μικρό ($n = 15$), το $(1-a)100\%$ διάστημα εμπιστοσύνης για τη μέση τιμή του πληθυσμού θα είναι αυτό που βασίζεται στην $t$ κατανομή, δηλ. το:

$$\left(\bar{X} - t_{n-1,a/2}s\sqrt{\frac{1-f}{n}}, \qquad \bar{X} + t_{n-1,a/2}s\sqrt{\frac{1-f}{n}}\right)$$

για $n = 15$ και $a = 0.01$.

Επιπλέον, επειδή δίδεται ότι το $N$ είναι αρκετά μεγάλο, μπορεί να θεωρηθεί ότι η διόρθωση πεπερασμένου πληθυσμού είναι αμελητέα και συνεπώς να παραλειφθεί. Οπότε, τελικά, το 99% ΔΕ για το μέσο χρόνο αναμονής είναι:

$$\left(14.59 - 2.98\sqrt{\frac{52.62}{15}}, \qquad 14.59 + 2.98\sqrt{\frac{52.62}{15}}\right) = (9.01, 20.17)$$

## Παράδειγμα 2.10 (ομαδοποιημένα δεδομένα)

Μια έρευνα διεξήχθη σε μεγάλη βιομηχανία με σκοπό να μελετήσει τον μέσο αριθμό ημερών κατά τις οποίες απουσιάζει ένας εργάτης με άδεια χωρίς προειδοποίηση. Ένα τυχαίο δείγμα από $n = 1000$ εργάτες λαμβάνεται από το σύνολο των $N = 36000$ εργαζομένων στο εργοστάσιο. Ο Πίνακας 2.3 δίνει τις χωρίς προειδοποίηση απουσίες που είχαν στην εργασία τους, για τους τελευταίους 6 μήνες, οι εργάτες του δείγματος.

| **Ημέρες άδειας χωρίς προειδοποίηση** | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Αριθμός εργατών** | 451 | 162 | 187 | 112 | 49 | 21 | 5 | 11 | 2 | 0 |

***Πίνακας 2.3*** *Απουσίες εργαζομένων τους τελευταίους 6 μήνες.*

Για τα δεδομένα του προβλήματος, επειδή οι μετρήσεις $X_i (i = 1,2, \dots, 1000)$ είναι ομαδοποιημένες σε 10 κατηγορίες/τιμές, η δειγματική μέση τιμή, που είναι ο εκτιμητής του μέσου αριθμού απουσιών, υπολογίζεται από τη σχέση:

$$\bar{X} = \frac{\sum_{i=0}^{9} X_i f_i}{\sum_{i=0}^{9} f_i}$$

όπου $f_i$ οι συχνότητες που καταγράφηκαν για το δείγμα στις 10 κατηγορίες και δίνονται στη δεύτερη γραμμή του Πίνακα 2.3.

Αντικαθιστώντας, θα είναι:

$$\bar{X} = \frac{1269}{1000} = 1.27$$

Συνεπώς, βάσει των μετρήσεων του δείγματος, ο μέσος αριθμός ημερών που οι εργαζόμενοι απουσιάζουν χωρίς προειδοποίηση, είναι 1.27.

Για το τυπικό σφάλμα της εκτίμησης:

$$s^2 = \frac{1}{\sum_{i=0}^{9} f_i - 1}\left(\sum_{i=0}^{9} X_i^2 f_i - \frac{\left(\sum_{i=0}^{9} X_i f_i\right)^2}{\sum_{i=0}^{9} f_i}\right) = 2.397$$

και:

$$\hat{se}(\bar{X}) = \sqrt{\frac{1-f}{1000}\, 2.397} = 0.0023$$

Ένα 95% διάστημα εμπιστοσύνης για τον ίδιο αριθμό θα είναι το:

$$(1.27 - 1.96 * 0.0023, 1.27 + 1.96 * 0.0023) = (1.16, 1.39)$$

δηλ. ο μέσος αριθμός ημερών που ένας εργάτης δεν εμφανίζεται στην εργασία του χωρίς ενημέρωση ή κανονική άδεια για ένα διάστημα 6 μηνών εκτιμάται ότι περιέχεται στο διάστημα $(1.16, 1.39)$ με πιθανότητα σφάλματος 5%.

## 2.5. Προσδιορισμός μεγέθους δείγματος για την Απλή Τυχαία Δειγματοληψία

Ένα από τα βασικότερα θέματα που απασχολούν τους ερευνητές κατά την οργάνωση και τον σχεδιασμό των δειγματοληπτικών ερευνών, είναι ο προσδιορισμός του μεγέθους του δείγματος. Το μέγεθος του δείγματος, $n$, παίζει καθοριστικό ρόλο στην αξιοπιστία των εκτιμητών που θα παραχθούν. Άλλες συνιστώσες της έρευνας που επίσης επηρεάζονται και συνυπολογίζονται για τον προσδιορισμό του μεγέθους του δείγματος, είναι το κόστος και ο χρόνος που θα απαιτηθεί για τη διεξαγωγή της έρευνας. Η αξιοπιστία των εκτιμητών αποτελεί τη σημαντικότερη συνιστώσα των δειγματοληπτικών ερευνών και ο συνήθης τρόπος προσδιορισμού του μεγέθους του δείγματος γίνεται με την υιοθέτηση ορισμένων απαιτήσεων ως προς την αξιοπιστία των εκτιμητών και την αναζήτηση του ελάχιστου αριθμού $n$, για τον οποίο οι απαιτήσεις αυτές ή οι προδιαγραφές, όπως λέγονται, πληρούνται.

Είναι προφανές ότι για δεδομένες προδιαγραφές επί των αποτελεσμάτων της έρευνας, το απαιτούμενο μέγεθος του δείγματος θα αλλάζει ανάλογα με τον τρόπο δειγματοληψίας ή, γενικότερα, το δειγματοληπτικό σχέδιο. Ένα δειγματοληπτικό σχέδιο είναι πιο αποδοτικό σε σχέση με ένα άλλο, εάν οι ίδιες προδιαγραφές πληρούνται με μικρότερο μέγεθος δείγματος σε σχέση με το εναλλακτικό σχέδιο. Συνεπώς, κατά τον σχεδιασμό μιας έρευνας, μετά την επιλογή του δειγματοληπτικού σχεδίου που θα υιοθετηθεί, πρέπει να επιλυθεί το πρόβλημα του προσδιορισμού του μεγέθους του δείγματος.

Στο παρόν κεφάλαιο, θα παρουσιαστεί το πρόβλημα του προσδιορισμού του ικανού μεγέθους του δείγματος για την απλή τυχαία δειγματοληψία. Ανάλογα αντιμετωπίζεται το πρόβλημα και για διαφορετικά δειγματοληπτικά σχέδια ή για συνδυασμό αυτών. Επίσης, για δεδομένο δειγματοληπτικό σχέδιο, η λύση του προβλήματος ως προς την εύρεση του $n$ θα διαφέρει, ανάλογα με την παράμετρο του πληθυσμού που αποτελεί τον στόχο της έρευνας. Για παράδειγμα, εάν αυτό που επιδιώκεται είναι να εκτιμηθεί η μέση τιμή του πληθυσμού για ένα χαρακτηριστικό $Y$, η αξιοπιστία, και συνεπώς και το μέγεθος δείγματος που θα απαιτείται, θα είναι διαφορετικά από το ανάλογο πρόβλημα της εκτίμησης του συνόλου του πληθυσμού.

Τα βασικά στοιχεία του προβλήματος, τα οποία ισχύουν για κάθε πρόβλημα εκτίμησης παραμέτρου και για κάθε δειγματοληπτικό σχέδιο, είναι τα εξής:

- Το πρώτο βήμα είναι να τεθούν οι προδιαγραφές ή οι απαιτήσεις τις οποίες επιδιώκεται να ικανοποιούν τα αποτελέσματα της έρευνας, όταν αυτή ολοκληρωθεί. Οι προδιαγραφές αυτές συνήθως σχετίζονται με την αξιοπιστία των εκτιμητών και εκφράζονται είτε σε απόλυτη μορφή είτε με ένα περιθώριο λάθους.
- Οι προδιαγραφές που έχουν οριστεί διατυπώνονται με μαθηματικό τρόπο, ώστε το πρόβλημα να μεταφραστεί σε μια μαθηματική εξίσωση ή ανίσωση που θα περιέχει τον άγνωστο $n$.
- Η εξίσωση ενδέχεται να περιέχει ποσότητες του πληθυσμού ή χαρακτηριστικά, τα οποία πιθανότατα θα είναι άγνωστα στο στάδιο αυτό και θα πρέπει είτε να εκτιμηθούν είτε να προσδιοριστούν με κάποιο τρόπο, πριν από τη διεξαγωγή της έρευνας.
- Από τη λύση της εξίσωσης (ή της ανίσωσης) ως προς το $n$ προκύπτει το προτεινόμενο μέγεθος του δείγματος για την έρευνα. Το μέγεθος αυτό πρέπει να συμφωνεί και να μην υπερβαίνει τα επιτρεπτά όρια που είναι διαθέσιμα για το συνολικό κόστος και τον χρόνο διεξαγωγής της έρευνας. Μια παράμετρος που επίσης πρέπει να συνυπολογιστεί είναι το ποσοστό μη-απόκρισης (non-response) που αναπόφευκτα συνοδεύει μια δειγματοληπτική έρευνα και, ανάλογα με το θέμα ή τον τρόπο διεξαγωγής, μπορεί να είναι αρκετά μεγάλο.

### 2.5.1 Εύρεση του ελάχιστα απαιτούμενου μεγέθους δείγματος

Έστω ότι ο στόχος της διεξαγωγής της έρευνας είναι η εκτίμηση του μέσου του πληθυσμού $\bar{Y}$, για το χαρακτηριστικό $Y$. Η πιο συνηθισμένη μορφή απαιτήσεων ή περιορισμών στα προσδοκώμενα αποτελέσματα είναι να θέτει ο ερευνητής ορισμένα όρια ανοχής στην εκτίμηση που θα επιτευχθεί ως προς την αληθινή τιμή στον πληθυσμό. Με άλλα λόγια, να προσδιορίζει μια περιοχή γύρω από την παράμετρο $\theta = \bar{Y}$ (συνήθως η περιοχή είναι συμμετρική ως προς το $\theta$) μέσα στην οποία, τιμές των εκτιμητών $\hat{\theta}$ που θα προκύψουν από την έρευνα θα είναι αποδεκτές, ενώ τιμές εκτός της περιοχής, όχι. Ή, αντίστροφα, το μέγεθος του δείγματος $n$ να είναι ικανό, ώστε να εγγυάται εκτιμητές εντός της περιοχής ανοχής. Το όριο ανοχής μπορεί να θεωρηθεί και ως το περιθώριο του λάθους που επιτρέπει ο ερευνητής για την εκτίμηση. Το όριο αυτό είναι πάντα σχετικό με το πρόβλημα και τις τιμές της παραμέτρου $\theta$. Συνήθως, καθορίζεται σε συνεργασία του στατιστικού ερευνητή και του ερευνητή από τον χώρο που αναφέρεται το αρχικό πρόβλημα (ιατρική, βιολογία, παιδαγωγικά κτλ).

Αν $d$ συμβολίζει το όριο ανοχής και το διάστημα ανοχής είναι συμμετρικό, αυτό σημαίνει ότι η περιοχή ανοχής για την εκτίμηση της παραμέτρου $\hat{\theta}$, είναι $(\theta - d, \theta + d)$ ή, ισοδύναμα,

$$\left|\theta - \hat{\theta}\right| < d$$

Εάν η επιλογή του $n$ είναι τέτοια, ώστε η παραπάνω ανισότητα να ικανοποιείται ένα μεγάλο αριθμό φορών, συγκεκριμένα σε ποσοστό $(1 - a)100\%$ για διαδοχικές ανεξάρτητες επαναλήψεις της δειγματοληψίας, ή, ισοδύναμα, το ενδεχόμενο η εκτίμηση να είναι εκτός των ορίων ανοχής για την παράμετρο να έχει πιθανότητα $a$, τότε συνολικά η πιθανότητα:

$$\mathrm{P}\left(\left|\theta - \hat{\theta}\right| \geq d\right) = a \qquad (2.7)$$

καθορίζει την εξίσωση η οποία μεταφράζει τις απαιτήσεις του ερευνητή και η επίλυση της οποίας θα οδηγήσει στον προσδιορισμό του $n$.

Προκειμένου να λυθεί η ανίσωση, πρέπει αρχικά να υπολογιστεί η πιθανότητα στο αριστερό μέρος και στη συνέχεια να ταυτιστεί με $a$. Για $\hat{\theta} = \bar{X}$ κάτω από την απλή τυχαία δειγματοληψία και με την επιπλέον υπόθεση της κανονικότητας για τον δειγματικό μέσο, θα ισχύει

$$\mathrm{P}(|\bar{Y} - \bar{X}| \geq d) = \mathrm{P}\left(\frac{|\bar{Y} - \bar{X}|}{\sqrt{\mathrm{Var}(\bar{X})}} \geq \frac{d}{\sqrt{\mathrm{Var}(\bar{X})}}\right) = \mathrm{P}\left(|Z| \geq \frac{d}{\sqrt{\mathrm{Var}(\bar{X})}}\right)$$

όπου $Z$ η τυπική κανονική.

Άρα, συνολικά, η (2.7) γίνεται:

$$2\,\Phi\left(\frac{d}{\mathrm{se}(\bar{X})}\right) = a$$

όπου $\Phi$ η αθροιστική συνάρτηση της τυπικής κανονικής κατανομής. Με τη βοήθεια των $z$ (άνω) εκατοστιαίων σημείων της τυπικής κανονικής, η τελευταία σχέση συνεπάγεται:

$$\frac{d}{\mathrm{se}(\bar{X})} = z_{\alpha/2} \tag{2.8}$$

Τέλος, λαμβάνοντας υπόψη τον τρόπο δειγματοληψίας και την έκφραση του τυπικού σφάλματος του δειγματικού μέσου στον παρονομαστή, θα είναι:

$$\frac{d}{\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) S^2}} = z_{\alpha/2}$$

Λύνοντας ως προς το $n$, προκύπτει:

$$n = N\left\{1 + N\left(\frac{d}{z_{\alpha/2}\,S}\right)^2\right\}^{-1} \tag{2.9}$$

Η τιμή του $n$ η οποία δίνεται από τη (2.9) είναι η ελάχιστη ικανή, ώστε ο εκτιμητής που θα προκύψει με επιλογή απλού τυχαίου δείγματος για εκτίμηση της μέσης τιμής του πληθυσμού να πληροί τη (2.7). Προφανώς, οποιαδήποτε τιμή του $n$ μεγαλύτερη από αυτήν ικανοποιεί επίσης τους περιορισμούς.

Η (2.9) ισοδύναμα γράφεται:

$$n = \left(\frac{z_{\alpha/2}\,S}{d}\right)^2 \left\{1 + \frac{1}{N}\left(\frac{z_{\alpha/2}\,S}{d}\right)^2\right\}^{-1} \tag{2.10}$$

Από την τελευταία σχέση, προκύπτει ότι $n = \left(\frac{z_{\alpha/2}S}{d}\right)^2$ για $N$ αρκετά μεγάλο.

Σημειώνεται ότι η ίδια διαδικασία μπορεί να εφαρμοστεί και για παράμετρο $\theta = Y_T$ και ομοίως $\theta = P$ ή $A$, για διακριτές περιπτώσεις. Όλες οι περιπτώσεις ως προς την παράμετρο $\theta$ μπορούν να συμπεριληφθούν στην εξίσωση:

$$\frac{d}{\mathrm{se}(\hat{\theta})} = z_{\alpha/2} \tag{2.11}$$

όπου $\theta$είναι η παράμετρος

προς εκτίμηση, $d$ το μέγιστο επιτρεπτό όριο ανοχής για την εκτίμηση του $\theta$ από το $\hat{\theta}$ και $\text{se}(\hat{\theta})$ είναι το τυπικό σφάλμα του εκτιμητή, όπως αυτό υπολογίζεται μέσω της απλής τυχαίας δειγματοληψίας.

## Παρατήρηση 2.7

Αντί της απαίτησης η εκτίμηση να είναι εντός προκαθορισμένων επιτρεπτών ορίων από την αληθινή τιμή της παραμέτρου, άλλοι συνήθεις τρόποι προσδιορισμού του μεγέθους του δείγματος είναι (i) μια επιτρεπτή ανώτατη τιμή για το τυπικό σφάλμα της εκτίμησης ή (ii) μια επιτρεπτή ανώτατη τιμή για τον συντελεστή μεταβλητότητας. Για παράδειγμα, η περίπτωση (i) ισοδυναμεί μαθηματικά με την εξίσωση $\text{se}(\hat{\theta}) = s_0$, με $s_0$ τη δοθείσα ανώτατη επιτρεπτή τιμή. Αν η εξίσωση λυθεί ως προς $n$, προσδιορίζεται το ελάχιστο ικανό μέγεθος του απλού τυχαίου δείγματος το οποίο καλύπτει την απαίτηση ως προς το τυπικό σφάλμα του εκτιμητή.

### 2.5.2 Εκτίμηση άγνωστων ποσοτήτων του πληθυσμού πριν από τη διεξαγωγή της έρευνας

Από τη μελέτη της προηγούμενης παραγράφου, προκύπτει ότι, γενικά, ο υπολογισμός του μεγέθους του δείγματος $n$ επιτυγχάνεται με τη λύση της εξίσωσης που ενσωματώνει τις αρχικές συνθήκες. Προκειμένου η λύση να έχει μοναδικό άγνωστο το $n$ και η επίλυσή της να προσφέρει αριθμητικό και όχι συναρτησιακό αποτέλεσμα, ποσότητες όπως η διακύμανση του πληθυσμού $S^2$ στην εξίσωση π.χ. (2.9) θα πρέπει να είναι γνωστές κατά τη φάση του σχεδιασμού της έρευνας. Ανάλογα με την παράμετρο $\theta$ που είναι προς εκτίμηση, θα είναι διαφορετική η ποσότητα του πληθυσμού που είναι αναγκαίο να είναι γνωστή πριν από την έρευνα. Για παράδειγμα, αν $\theta = P$, η (2.11) δίνει:

$$\frac{d}{\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)\frac{N\,P\,(1-P)}{N-1}}} = z_{\alpha/2} \qquad (2.12)$$

από όπου συνάγεται ότι είναι απαραίτητη μια πληροφορία για την τιμή του ποσοστού $P$ στον πληθυσμό, προκειμένου να υπολογιστεί το $n$. Σημειώνεται ότι σ' αυτή την περίπτωση, η ποσότητα $P$ είναι η ίδια η ποσότητα του πληθυσμού, για την εκτίμηση της οποίας πρόκειται να διεξαχθεί η έρευνα.

Άγνωστες ποσότητες από τον πληθυσμό απαιτούνται και όταν το όριο ανοχής $d$ δεν προσφέρεται από τον ερευνητή σε απόλυτο μέγεθος, αλλά ως ποσοστό επί της αληθινής ποσότητας. Π.χ., αν μια έρευνα αφορά το μέσο ετήσιο εισόδημα ενός συνόλου πολιτών, το όριο ανοχής $d$, αντί να εκφράζεται σε χρηματικές μονάδες, έστω $d = 300$ χρηματικές μονάδες, μπορεί να ορίζεται ως το $10\%$ επί της αληθινής τιμής, δηλ. το σύνολο αποδεκτών τιμών για την εκτίμηση να είναι $(0.9\bar{Y}, 1.1\bar{Y})$. Σε αυτή την περίπτωση, χρειάζεται επιπλέον του $S^2$ να είναι διαθέσιμη πληροφορία και για το $\bar{Y}$.

Παρότι η έρευνα βρίσκεται ακόμα στο στάδιο του σχεδιασμού και δεν υπάρχουν διαθέσιμοι εκτιμητές από το δείγμα, π.χ. ο εκτιμητής $s^2$ για το $S^2$, είναι αρκετά συνηθισμένο στην πράξη να υπάρχουν προηγούμενες έρευνες πάνω στο ίδιο θέμα για τον ίδιο πληθυσμό ή να είναι διαθέσιμα στοιχεία από μια πρόσφατη απογραφή. Εάν δεν υπάρχει πρόσφατη απογραφή ή προηγούμενη έρευνα για τον ίδιο πληθυσμό, είναι αναγκαία η διεξαγωγή μιας μικρής κλίμακας έρευνας, πιλοτικής, η οποία μεταξύ άλλων πολύτιμων αποτελεσμάτων που θα προσφέρει για την κανονικής κλίμακας έρευνα που θα ακολουθήσει, θα δώσει και κάποιες αρχικές εκτιμήσεις για τις άγνωστες παραμέτρους του πληθυσμού.

Στις περιπτώσεις όπου η έρευνα αφορά ποσοστό $P$, ή, ισοδύναμα, το πλήθος των μελών $A$ που ανήκουν σε μια κατηγορία $C$, είναι αρκετά συχνό στην πράξη να υπάρχουν διαθέσιμες προηγούμενες έρευνες που να παρέχουν πληροφορία για το $P$ με μορφή διαστήματος. Π.χ. η έρευνα αφορά το ποσοστό ανεργίας σε μια χώρα και, από στοιχεία προηγούμενων ερευνών, εκτιμάται ότι το ποσοστό αυτό την παρούσα χρονική στιγμή κυμαίνεται από $0.15$ έως $0.21$. Στις περιπτώσεις αυτές, η τιμή του $P$ που θα αντικατασταθεί στην εξίσωση

(2.11) είναι εκείνη η τιμή του διαστήματος $(0.15,\ 0.21)$ που μεγιστοποιεί τη διασπορά του πληθυσμού $S^2 = \frac{NP(1-P)}{N-1}$. Για το παράδειγμα, θα είναι $P = 0.21$ .

## Παράδειγμα 2.11

Μια έρευνα διεξάγεται στην αρχή του ημερολογιακού έτους, με σκοπό την εκτίμηση του συνολικού αριθμού Χριστουγεννιάτικων καρτών που θα πουληθούν συνολικά στην περιοχή το τρέχον έτος (Barnett, 2002, Κεφ. 2). Στην περιοχή είναι καταγεγραμμένα συνολικά 243 καταστήματα. Η έρευνα θα διεξαχθεί με τυχαία επιλογή ενός δείγματος των καταστημάτων και θα ζητήσει στοιχεία για το πόσες κάρτες πούλησαν τη χρονιά που πέρασε. Το ερώτημα είναι πόσα καταστήματα πρέπει να επιλεγούν, έτσι ώστε η εκτίμηση να γίνει με απόκλιση το πολύ 10% επί της πραγματικής τιμής και με βεβαιότητα 95%;

Για την περιοχή, υπάρχουν τα συνολικά στοιχεία για τις πωλήσεις των καταστημάτων σε Χριστουγεννιάτικες κάρτες, όπως αυτά δηλώνονται κάθε Ιούλιο. Για τα τρία τελευταία χρόνια, τα στοιχεία για τις συνολικές πωλήσεις, καθώς και για την τυπική τους απόκλιση (σε 10000 μονάδες), είναι

| $Y_T$ | $S$ |
|---|---|
| 321.7 | 0.826 |
| 366.8 | 0.776 |
| 401.0 | 0.804 |

Για τον υπολογισμό του μεγέθους του δείγματος, γίνεται χρήση της εξίσωσης (2.11), λαμβάνοντας υπόψη ότι $\theta = Y_T$ και ότι ο τρόπος δειγματοληψίας είναι η απλή τυχαία, οπότε το τυπικό σφάλμα του εκτιμητή δίνεται από την Πρόταση 2.5 (iv). Θα είναι συνεπώς:

$$\frac{d}{N\sqrt{\frac{1-f}{n}S^2}} = z_{\alpha/2}$$

Αν η σχέση αυτή επιλυθεί στη συνέχεια ως προς $n$, προκύπτει:

$$n = N\left\{1 + \frac{1}{N}\left(\frac{d}{z_{\alpha/2}\,S}\right)^2\right\}^{-1}.$$

Για τον αριθμητικό υπολογισμό του $n$ από την τελευταία έκφραση, είναι $N = 243$, $z_{\alpha/2} = 1.96$ και $d = 0.1\,Y_T$ και τιμές για τα $Y_T$, $S$ λαμβάνονται από τα στοιχεία προηγουμένων ετών.

Εάν γίνει χρήση των διαθέσιμων στοιχείων από το πιο πρόσφατο παρελθόν έτος, τότε $Y_T = 401.0$ και $S = 0.804$ και, αντικαθιστώντας στη λύση της εξίσωσης για το $n$, θα είναι:

$$n = 243\left\{1 + \frac{1}{243}\left(\frac{0.10 \times 401.0}{1.96 \times 0.804}\right)^2\right\}^{-1} = 66.30$$

Δηλαδή, απαιτείται να συμπεριληφθούν τουλάχιστον 67 καταστήματα στην έρευνα, προκειμένου να πληρούνται οι προδιαγραφές.

Εάν, αντί για τα στοιχεία της αμέσως προηγούμενης χρονιάς, ληφθεί υπόψη η ανοδική τάση η οποία διακρίνεται στις πωλήσεις για τα 3 διαδοχικά έτη, τότε μπορεί να γίνει μια εκτίμηση για το $Y_T$ του τρέχοντος έτους, η οποία θα είναι ενδεχομένως πιο κοντά στην αληθινή τιμή $Y_T$ σε σύγκριση με προηγούμενα έτη. Για μια, έστω συντηρητική, αύξηση πωλήσεων, εάν τεθεί $Y_T = 420$ και $S = 0.8$, τότε η ίδια διαδικασία

υπολογισμού του $n$ δίνει $n = 61.48$, δηλ. απαιτούνται 62 καταστήματα για την έρευνα προκειμένου να πληρούνται οι προϋποθέσεις.

Ως γενική παρατήρηση, η μεθοδολογία που έχει αναπτυχθεί στην παρούσα παράγραφο δίνει μια κατώτατη τιμή για το μέγεθος του δείγματος που απαιτείται. Όσο μεγαλύτερη αξιοπιστία έχουν για τον ερευνητή οι εκτιμήσεις των ποσοτήτων από τον πληθυσμό που έχει χρησιμοποιήσει προκειμένου να υπολογίσει αυτή την κατώτατη τιμή, τόσο πιο κοντά σε αυτή την τιμή μπορεί να επιλέξει το $n$ κατά τη διεξαγωγή της έρευνας.

## 2.6. Απλή Τυχαία Δειγματοληψία στην πράξη

Η απλή τυχαία δειγματοληψία, παρόλο που αποτελεί τη βάση για την ανάπτυξη όλων των υπόλοιπων δειγματοληπτικών σχεδίων, δεν χρησιμοποιείται αρκετά συχνά στην πράξη. Η απλή τυχαία δειγματοληψία διαθέτει αρκετά πλεονεκτήματα λόγω του απλού τρόπου στον σχεδιασμό της και συνεπώς και στην εφαρμογή. Οι μονάδες στον πληθυσμό είναι όλες ισοπίθανες να συμπεριληφθούν στην έρευνα και αυτό είναι επίσης ένα πλεονέκτημα γιατί δίνεται ίδια βαρύτητα σε όλες τις μονάδες.

Ταυτόχρονα όμως, και παρά την απλότητά της, η απλή τυχαία δειγματοληψία παρουσιάζει και μειονεκτήματα. Σύμφωνα με τον ορισμό της απλής τυχαίας δειγματοληψίας, θα πρέπει πριν από τη διεξαγωγή της έρευνας να υπάρχει διαθέσιμη η λίστα με ένα προς ένα τα στοιχεία του πληθυσμού. Τα μέλη δηλ. του πληθυσμού να είναι εφικτό να αναγνωριστούν και να καταγραφούν ένα προς ένα. Πρακτικά, αυτό δεν είναι εφικτό για κάθε πληθυσμό, και τα μέλη του δεν είναι διαθέσιμα με πλήρη στοιχεία για τον ερευνητή. Παράλληλα, αν η διαδικασία αυτή ενσωματωθεί στη διεξαγωγή της έρευνας, το κόστος θα αυξηθεί αρκετά. Οι πρακτικοί λόγοι λοιπόν, είναι η μία κατηγορία των μειονεκτημάτων της απλής τυχαίας.

Στην ίδια κατηγορία, καθώς και στην αύξηση του κόστους της έρευνας, εντάσσεται επίσης το επόμενο από τα προβλήματα που παρουσιάζει η απλή τυχαία. Εάν ο πληθυσμός εκτείνεται σε μεγάλη γεωγραφική περιοχή, η επιλογή των μελών του δείγματος σύμφωνα με την απλή τυχαία μπορεί να οδηγήσει σε μέλη του πληθυσμού για το δείγμα που είναι διάσπαρτα σε όλη την έκταση του πληθυσμού. Σ΄ αυτές τις περιπτώσεις, ο εντοπισμός και η συλλογή της παρατήρησης για το δείγμα, ειδικά εάν γίνεται με επί τόπου επίσκεψη ή προσωπική συνέντευξη, απαιτεί την πλήρη γεωγραφική κάλυψη του πληθυσμού. Αυτό με τη σειρά του συνεπάγεται επιβάρυνση τόσο του κόστους, όσο και της διάρκειας της διεξαγωγής της έρευνας.

Το τελευταίο χαρακτηριστικό έχει μεγαλύτερες προεκτάσεις και επηρεάζει και την ακρίβεια των εκτιμητών που παράγονται. Αυτό συμβαίνει γιατί η καλή γεωγραφική κάλυψη του πληθυσμού από το δείγμα, δεν σημαίνει ταυτόχρονα και καλή κάλυψη όλου του εύρους τιμών του χαρακτηριστικού που μελετάται για τον πληθυσμό, δηλ. αντιπροσωπευτικότητα. Ενδέχεται μερικές ειδικές ομάδες που παρουσιάζουν μια συγκεκριμένη συμπεριφορά ως προς το χαρακτηριστικό να είναι τόσο μικρές, ώστε η α.τ.δ. να μην περιλάβει κανέναν, ή να περιλάβει πολύ λίγους εκπροσώπους τους στο δείγμα. Για την απλή τυχαία δειγματοληψία, οι διάφορες ομάδες του πληθυσμού θα εκπροσωπούνται στο δείγμα λαμβάνοντας υπόψη μόνο το μέγεθός τους, και κανένα επιπλέον χαρακτηριστικό, όπως η ομοιογένεια ή ετερογένεια που εμφανίζουν, η σπουδαιότητα που ενδεχομένως έχουν για μια συγκεκριμένη έρευνα κτλ.

Οι δειγματοληπτικές τεχνικές που θα αναπτυχθούν στα επόμενα κεφάλαια έχουν στόχο να εκμεταλλευτούν περισσότερα χαρακτηριστικά που πληθυσμού και να απαντήσουν στο πρόβλημα της εκτίμησης με μεγαλύτερη στατιστική ακρίβεια, διατηρώντας τα υπόλοιπα στοιχεία της έρευνας, όπως το κόστος, σταθερά.

# Βιβλιογραφικές Αναφορές


Δαμιανού, Χ. Χ. (2006). *Μεθοδολογία δειγματοληψίας. Τεχνικές και εφαρμογές*. Θεσσαλονίκη: Εκδόσεις Σοφία Α.Ε.

Barnett, V. (2002). *Sample survey: Principles and methods*. 3rd Edition, London: Arnold.

Cochran, W. G. (1977). *Sampling techniques* (3rd Edition). New York: John Wiley and Sons.

Des Raj (1968). *Sampling Theory*. New York: McGraw-Hill.

Levy, P.S. and Lemeshow, S. (1999). *Sampling of Populations. Methods and Applications* (3rd Edition). New York: John Wiley and Sons.

Rao, P.S.R.S. (2000). *Sampling methodologies with applications*. Boca Raton, Fla: Chapman and Hall/CRC.

Sugden, R. A., Smith, T. M. F. and Jones, R. P.(2000). Cochran's Rule for Simple Random Sampling. *Journal of the Royal Statistical Society B 62*, 787-793. doi: 10.1111/1467-9868.00264

Thompson, S. K. (2012). *Sampling* (3rd Edition). Hoboken, NJ: John Wiley and Sons.

# Κεφάλαιο 3 - ΣΤΡΩΜΑΤΟΠΟΙΗΜΕΝΗ ΔΕΙΓΜΑΤΟΛΗΨΙΑ

**Σύνοψη**

*Η στρωματοποιημένη δειγματοληψία αποτελεί μία από τις πιο διαδεδομένες μεθόδους δειγματοληψίας. Η χαρακτηριστική ιδιότητα και ταυτόχρονα το κυριότερο πλεονέκτημα της στρωματοποιημένης δειγματοληψίας είναι τα μειωμένα σε σχέση με άλλες μεθόδους δειγματοληψίας τυπικά σφάλματα των εκτιμητών. Στο στάδιο του σχεδιασμού της έρευνας, ο πληθυσμός χωρίζεται σε στρώματα. Τα στρώματα αποτελούν υποσύνολα του πληθυσμού τα οποία πληρούν συγκεκριμένες ιδιότητες. Στη συνέχεια, διεξάγεται μια ανεξάρτητη δειγματοληψία σε κάθε στρώμα και το τελικό δείγμα της έρευνας αποτελείται από τη συλλογή όλων των επιμέρους δειγμάτων ανά στρώμα. Με τον τρόπο αυτό, το συνολικό δείγμα που επιλέγεται από τον πληθυσμό περιέχει εκπροσώπους από κάθε διαφορετικό στρώμα του πληθυσμού. Εάν τα στρώματα έχουν οριστεί έτσι, ώστε να αντιστοιχούν σε διακριτές ομάδες του πληθυσμού, το δείγμα που επιλέγεται με τη βοήθεια της στρωματοποιημένης δειγματοληψίας είναι αντιπροσωπευτικό. Η στρωματοποιημένη δειγματοληψία εφαρμόζεται τόσο αποκλειστικά, όσο και σε συνδυασμό με άλλους τρόπους δειγματοληψίας, σ' ένα πιο σύνθετο δειγματοληπτικό σχέδιο.*

**Προαπαιτούμενη γνώση**

*Κεφάλαιο 1 -, Κεφάλαιο 2 -, Εκτιμητική.*

## 3.1. Εισαγωγή

Οι μέθοδοι δειγματοληψίας που ακολουθούν την εισαγωγή και τη μελέτη της απλής τυχαίας δειγματοληψίας είναι μέθοδοι που επιχειρούν να βελτιώσουν ορισμένα από τα μειονεκτήματα της απλής τυχαίας, είτε στην κατεύθυνση της ευκολίας και του κόστους ή του χρόνου εκτέλεσης της έρευνας, είτε στην κατεύθυνση της ακρίβειας των εκτιμητών που θα παραχθούν. Έχει ήδη σημειωθεί ότι η απλή τυχαία δειγματοληψία μειονεκτεί στο γεγονός ότι το δείγμα δεν επιτυγχάνει πάντα να επιλέξει εκπροσώπους από διακριτές ομάδες του πληθυσμού. Αυτό έχει ως αποτέλεσμα με τη σειρά του να μην επιτυγχάνεται μεγάλη ακρίβεια στην εκτίμηση κάτω από τον απλό τυχαίο τρόπο ως μέθοδο δειγματοληψίας.

Έστω, για παράδειγμα, ότι η έρευνα αφορά επιχειρήσεις που δραστηριοποιούνται σε μια περιοχή. Υπάρχουν επιχειρήσεις που απασχολούν έναν ή δύο εργαζομένους και υπάρχουν επιχειρήσεις που απασχολούν χιλιάδες εργαζομένους. Μία απλή τυχαία δειγματοληψία για τον εν λόγω πληθυσμό θα οδηγούσε σε μεγάλα τυπικά σφάλματα εκτιμητών, γιατί η ακρίβεια των εκτιμητών εξαρτάται από τη μεταβλητότητα του χαρακτηριστικού στον πληθυσμό. Η μεταβλητότητα για τον αριθμό εργαζομένων είναι αρκετά μεγάλη όταν ο πληθυσμός θεωρηθεί ως ένα ενιαίο σύνολο. Αν, αντίθετα, ήταν εφικτό να χωριστούν οι επιχειρήσεις σε κατηγορίες ανάλογα με τα άτομα που απασχολούν, οι ομάδες που θα προέκυπταν θα είχαν μικρότερη μεταβλητότητα η καθεμιά. Π.χ. μικρές επιχειρήσεις, επιχειρήσεις μεσαίου μεγέθους και μεγάλες επιχειρήσεις. Αν, στη συνέχεια, κάθε ομάδα του πληθυσμού καλύπτεται λαμβάνοντας ένα δείγμα χωριστά, τότε με μεγαλύτερη βεβαιότητα θα καλυφθεί όλο το εύρος των επιχειρήσεων στην περιοχή. Επιπλέον, η εκτίμηση που θα γίνει χωριστά σε κάθε ομάδα θα είναι πιο ακριβής, γιατί η μεταβλητότητα του κάθε υποπληθυσμού θα είναι μικρή. Με κατάλληλο συνδυασμό στη συνέχεια των τριών επιμέρους εκτιμήσεων, προκύπτει ένας εκτιμητής που αναφέρεται στον συνολικό πληθυσμό και είναι επίσης βελτιωμένος σε ακρίβεια, δεδομένου ότι και οι τρεις συνιστώσες εκτιμήσεις είναι ακριβείς.

Η μέθοδος δειγματοληψίας που δόθηκε στην περιγραφή του παραδείγματος λέγεται **στρωματοποιημένη (stratified)** και οι υποπληθυσμοί στους οποίους χωρίζεται ο πληθυσμός λέγονται **στρώματα (strata)**. Όπως είναι ήδη φανερό, όσο πιο ομοιογενή είναι τα στρώματα, τόσο πιο ακριβής θα είναι η κάθε επιμέρους δειγματοληψία, άρα και η συνολική τελική εκτίμηση.

Ο πληθυσμός ενδέχεται να είναι χωρισμένος σε στρώματα και η πληροφορία αυτή να είναι διαθέσιμη στον ερευνητή μέσω του δειγματοληπτικού πλαισίου. Για παράδειγμα, όταν το δειγματοληπτικό πλαίσιο για τους κατοίκους μιας πόλης αποτελείται από τους ταχυδρομικούς κώδικες των διευθύνσεων των κατοίκων. Σε άλλες περιπτώσεις, ο χωρισμός του πληθυσμού σε στρώματα γίνεται από τον ίδιο τον ερευνητή, με σκοπό την εφαρμογή στη συνέχεια της στρωματοποιημένης δειγματοληψίας, λόγω των πλεονεκτημάτων που διαθέτει.

Για παράδειγμα, στην έρευνα για την εκτίμηση του ποσοστού ανεργίας των κατοίκων μιας χώρας, οι κατάλογοι των κατοίκων της χώρας χωρίζονται από τον ερευνητή κατά ηλικιακές ομάδες.

Στην πράξη, συμβαίνει αρκετά συχνά να υπάρχει πληροφορία για τα στρώματα στον πληθυσμό, και αυτό συντελεί στην ευρεία χρήση της στρωματοποιημένης δειγματοληψίας.

Συνοπτικά, η **ιδέα** της στρωματοποιημένης δειγματοληψίας είναι να χωριστεί ο πληθυσμός σε όσο το δυνατό πιο διακριτά και ομοιογενή στο εσωτερικό τους στρώματα ή επίπεδα του υπό μελέτη χαρακτηριστικού. Στη συνέχεια, πραγματοποιείται **ανεξάρτητα** μία δειγματοληψία σε κάθε στρώμα. Το τελικό δείγμα είναι η συλλογή όλων των επιμέρους δειγμάτων από τα στρώματα. Με τον τρόπο αυτό, επιτυγχάνεται να περιέχονται αντιπρόσωποι όλων των δυνατών στρωμάτων του πληθυσμού στο τελικό δείγμα. Επιπλέον, ο βαθμός εκπροσώπησης του κάθε στρώματος στο δείγμα είναι στον έλεγχο του ερευνητή, αφού το μέγεθος του δείγματος ανά στρώμα καθορίζεται ανεξάρτητα.

Ως προς την εκτίμηση των παραμέτρων του πληθυσμού, αυτή προκύπτει με τον κατάλληλο συνδυασμό των επιμέρους εκτιμήσεων των παραμέτρων που θα προκύψουν από τις δειγματοληψίες ανά στρώμα.

Ο πληθυσμός μπορεί να χωριστεί σε στρώματα με πολλούς διαφορετικούς τρόπους. Ο χωρισμός όμως που θα δώσει πιο ακριβή αποτελέσματα είναι εκείνος που θα γίνει κάνοντας χρήση μιας μεταβλητής ή ενός κριτηρίου που είναι σχετικό με το χαρακτηριστικό που μελετάμε. Η ιδανική περίπτωση είναι τα χωρισμένα στρώματα να είναι όσο το δυνατό **πιο ομοιογενή.** Δηλαδή "παρόμοιες" (ως προς το χαρακτηριστικό που μελετάμε) μονάδες να τοποθετηθούν στο ίδιο στρώμα. Επίσης, ταυτόχρονα, το κάθε στρώμα **να διαφέρει** από τα υπόλοιπα. Π.χ. Έστω $Y$: ο συνολικός χρόνος ανά εβδομάδα που βλέπουν τηλεόραση οι κάτοικοι μιας περιοχής. Αν το χαρακτηριστικό $Y$ μεταβλητή είναι το αντικείμενο της έρευνας, κριτήρια χωρισμού των κατοίκων σε στρώματα θα μπορούσε να είναι: η ηλικιακή ομάδα στην οποία ανήκουν, το αν κάποιος εργάζεται ή όχι, το φύλο κ.ά.

Θέματα που αφορούν την επιλογή της βοηθητικής μεταβλητής και την κατασκευή των στρωμάτων θα απασχολήσουν το επόμενο κεφάλαιο. Στο παρόν κεφάλαιο, γίνεται μελέτη της εκτίμησης και των ιδιοτήτων των εκτιμητών που προκύπτουν από ένα στρωματοποιημένο δειγματοληπτικό σχέδιο. Ενδεικτική βιβλιογραφία, όπως και για την α.τ.δ., είναι των Cochran (1977, Κεφ. 5), Kish (1965, Κεφ. 3), Des Raj (1968, Κεφ. 4), Rao (2000, Κεφ. 4), Barnett (2002, Κεφ. 2), Levy & Lemeshow (1999, Κεφ. 5) και Thompson (2012, Κεφ. 11).

## 3.2. Εκτίμηση στη στρωματοποιημένη δειγματοληψία

### 3.2.1 Συμβολισμός

Πριν από την ανάπτυξη του προβλήματος της εκτίμησης στη στρωματοποιημένη δειγματοληψία είναι αναγκαίο να δοθεί ο συμβολισμός που θα χρησιμοποιηθεί για το κεφάλαιο αυτό. Ο Πίνακας 3.1 περιέχει τον απαραίτητο συμβολισμό και την επεξήγηση κάθε ποσότητας που εισάγεται. Ο συμβολισμός είναι σε συμφωνία με εκείνον που δόθηκε για την α.τ.δ., με την επιπλέον αναγκαία επέκτασή του και εφαρμογή του σε κάθε στρώμα του πληθυσμού. Έτσι, για παράδειγμα, εκτός από την πληθυσμιακή μέση τιμή $\bar{Y}$ του χαρακτηριστικού $Y$, για την περίπτωση του στρωματοποιημένου πληθυσμού έχει νόημα και ορίζεται ανάλογα η πληθυσμιακή μέση τιμή του $Y$, τόσο ανά στρώμα, όσο και συνολικά για ολόκληρο τον πληθυσμό.

| **Σύμβολο** | **Ερμηνεία** |
|---|---|
| $L$ | Αριθμός των στρωμάτων. |
| $N_1, N_2, \dots, N_L$ | Πληθυσμιακά μεγέθη των στρωμάτων. Αν $N$ είναι το μέγεθος του πληθυσμού, τότε: $N_1 + N_2 + \dots + N_L = N$. |

| | |
|---|---|
| $Y_{ij}$ | Η τιμή του χαρακτηριστικού $Y$ για το $j$ μέλος του πληθυσμού που ανήκει στο στρώμα $i$. |
| $W_h = \frac{N_h}{N}$ | Βάρος του στρώματος $h$ $(h = 1,2, \dots, L)$. Ισχύει: $\sum_{h=1}^{L} W_h = 1$. |
| $\bar{Y}_h = \frac{1}{N_h}\sum_{j=1}^{N_h} Y_{hj}$ | Μέση τιμή του χαρακτηριστικού $Y$ για το στρώμα $h$ $(h = 1,2, \dots, L)$. |
| $S_h^2 = \frac{1}{N_h - 1}\sum_{j=1}^{N_h}\left(Y_{hj} - \bar{Y}_h\right)^2$ | Πληθυσμιακή διασπορά για το στρώμα $h$ $(h = 1,2, \dots, L)$. |
| $X_{ij}$ | Η $j$ μονάδα του δείγματος που προέρχεται από το στρώμα $i$. |
| $n_1, n_2, \dots, n_L$ | Μεγέθη δείγματος ανά στρώμα. Το στρωματοποιημένο δείγμα είναι μεγέθους $n$, όπου $n$ τέτοιο ώστε: $\sum_{h=1}^{L} n_h = n$. |
| $f_h = \frac{n_h}{N_h}$ | Πηλίκο δείγματος για το στρώμα $h$ $(h = 1,2, \dots, L)$. |
| $\bar{X}_h = \frac{1}{N_h}\sum_{j=1}^{N_h} X_{hj}$ | Ο δειγματικός μέσος για το στρώμα $h$ $(h = 1,2, \dots, L)$. |
| $s_h^2 = \frac{1}{n_h - 1}\sum_{j=1}^{n_h}\left(X_{hj} - \bar{X}_h\right)^2$ | Δειγματική διασπορά μετρήσεων για το στρώμα $h$ $(h = 1,2, \dots, L)$. |

***Πίνακας 3.1*** *Συμβολισμός για τη στρωματοποιημένη δειγματοληψία*

## Παράδειγμα 3.1

Μια τάξη σχολείου αποτελείται από 3 τμήματα. Το σύνολο των μαθητών σε κάθε τμήμα είναι $N_1$ =20, $N_2$ =35 και $N_3$ =30 αντίστοιχα (αυτό σημαίνει ότι $N = 85$ και τα βάρη των τριών στρωμάτων είναι $W_1 = 0.23$, $W_2 = 0.42$ και $W_3 = 0.35$). Από κάθε τμήμα, επιλέχθηκε ένα τυχαίο δείγμα μαθητών. Έτσι, από το 1ο τμήμα επιλέξαμε $n_1 = 6$ μαθητές, από το 2ο $n_2 = 4$ και από το 3ο τμήμα $n_3 = 5$ μαθητές. Το συνολικό μέγεθος δείγματος είναι $n$ =15. Οι μαθητές του δείγματος έδωσαν ένα τεστ Μαθηματικών και τα αποτελέσματα του τεστ για κάθε μαθητή δίνονται στον Πίνακα 3.2.

| | **Παρατηρήσεις Δείγματος** | **Μέγεθος Δείγματος** | **Δειγματικοί μέσοι** |
|---|---|---|---|
| **Στρώμα 1 (Τμήμα 1)** | 87, 34, 29, 62, 75, 90 | $n_1 = 6$ | $\bar{X}_1$=62.83 |
| **Στρώμα 2 (Τμήμα 2)** | 24, 15, 85, 69 | $n_2 = 4$ | $\bar{X}_2$=48.25 |
| **Στρώμα 3 (Τμήμα 3)** | 99, 81, 97, 9, 61 | $n_3 = 5$ | $\bar{X}_3$=69.4 |

***Πίνακας 3.2*** *Παρατηρήσεις του στρωματοποιημένου δείγματος των μαθητών.*

### 3.2.2 Εκτίμηση του μέσου του πληθυσμού

Η εκτίμηση του μέσου του πληθυσμού $\bar{Y}$ με βάση ένα στρωματοποιημένο δείγμα προκύπτει εύκολα, εάν η άγνωστη παράμετρος $\bar{Y}$ εκφραστεί συναρτήσει των μέσων τιμών του $\bar{Y}_h$ μέσα σε κάθε υποπληθυσμό ή στρώμα.

**Πρόταση 3.1**

Για έναν στρωματοποιημένο πληθυσμό, η μέση τιμή $\bar{Y}$ δίνεται από τη σχέση:

$$\bar{Y} = \sum_{h=1}^{L} W_h \bar{Y}_h \tag{3.1}$$

όπου $\bar{Y}_h$ η μέση τιμή του στρώματος $h$ για το χαρακτηριστικό $Y$ και $W_h$ το βάρος του στρώματος $h$ για τον συνολικό πληθυσμό.

**Απόδειξη**

Σύμφωνα με τον ορισμό, η μέση τιμή του $Y$ για τον συνολικό πληθυσμό είναι:

$$\bar{Y} = \frac{1}{N} \sum_{h=1}^{L} \sum_{j=1}^{N_h} Y_{hj}$$

Παράλληλα, από τον ορισμό μέσων τιμών του $Y$ ανά στρώμα, $\bar{Y}_h$, προκύπτει ότι $\sum_{j=1}^{N_h} Y_{hj} = N_h \bar{Y}_h$. Αν αντικαταστήσουμε το αποτέλεσμα αυτό στην προηγούμενη σχέση, προκύπτει:

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

που συμπληρώνει την απόδειξη ∎

Από την ερμηνεία του αποτελέσματος της Πρότασης 3.1 και λαμβάνοντας υπόψη ότι $W_1 + \cdots + W_L = 1$, προκύπτει ότι η συνολική μέση τιμή του υπό μελέτη χαρακτηριστικού για τον πληθυσμό είναι ένας σταθμισμένος μέσος όρος των επιμέρους μέσων τιμών του χαρακτηριστικού ανά στρώμα. Η στάθμη του μέσου για το κάθε στρώμα ισούται με το βάρος τού στρώματος. Με άλλα λόγια, η βαρύτητα, ή η συνεισφορά της επιμέρους μέσης τιμής $\bar{Y}_h$ του στρώματος $h$ στη συνολική μέση τιμή του πληθυσμού, θα είναι τόσο μεγαλύτερη, όσο μεγαλύτερο αριθμητικά είναι το στρώμα σε σχέση με τον συνολικό πληθυσμό, και αντίστροφα.

**Ορισμός 3.1**

Σε αναλογία με την Πρόταση 3.1, ορίζεται ο δειγματικός μέσος του στρωματοποιημένου δείγματος που συμβολίζεται $\bar{X}_{st}$ και δίνεται από τη σχέση:

$$\bar{X}_{st} = \sum_{h=1}^{L} W_h \bar{X}_h \tag{3.2}$$

Ο στρωματοποιημένος μέσος $\bar{X}_{st}$ είναι ο εκτιμητής του $\bar{Y}$ όταν η δειγματοληψία έχει γίνει με στρωματοποιημένο σχήμα. Η σχέση (3.2) δίνει μαθηματικά τον τρόπο που συνδυάζονται οι εκτιμήσεις ανά στρώμα, ώστε να προκύψει μία συνολική εκτίμηση για τον μέσο του πληθυσμού. Εκτός από τους εκτιμητές ανά στρώμα $\bar{X}_h$, οι ποσότητες που παίζουν σημαντικό ρόλο στη σύνθεση του τελικού εκτιμητή είναι και τα

πληθυσμιακά βάρη των στρωμάτων, $W_h$. Ο ρόλος των $W_h$ στην εκτίμηση (3.2) είναι ταυτόσημος με εκείνον στην περίπτωση της σχέσης (3.1) που ισχύει για την αληθινή ποσότητα.

Αξίζει να σημειωθεί ότι ο μέσος του στρωματοποιημένου δείγματος **δεν** ισούται με τον γνωστό απλό δειγματικό μέσο του δείγματος $\bar{X}$. Η παρακάτω πρόταση εξετάζει αναλυτικά τη σχέση μεταξύ των δύο δειγματικών μέσων.

### Πρόταση 3.2

Η δειγματική στρωματοποιημένη μέση τιμή $\bar{X}_{st}$ ισούται με τη δειγματική απλή μέση τιμή ενός δείγματος $\bar{X}$, αν και μόνον αν $W_h = w_h$ για κάθε $h$ $(h = 1,2,\ldots,L)$, όπου $w_h$ τα δειγματικά βάρη των στρωμάτων, δηλ. $w_h = \frac{n_h}{n}$, $h = 1,2,\ldots,L$.

### Απόδειξη

Οι δύο μέσοι θα ταυτίζονται όταν $\bar{X}_{st}$=$\bar{X}$ ή, ισοδύναμα:

$$\sum_{h=1}^{L} W_h \bar{X}_h = \frac{1}{n}\sum_{h=1}^{L}\sum_{j=1}^{n_h} X_{hj}$$

ή:

$$\sum_{hj}^{L} \frac{N_h}{N}\bar{X}_h = \frac{1}{n}\sum_{h=1}^{L} n_h \bar{X}_h$$

λόγω του ορισμού της δειγματικής μέσης τιμής ανά στρώμα. Η τελευταία ισότητα θα ισχύει, αν και μόνον αν οι συντελεστές των δειγματικών μέσων ανά στρώμα ταυτίζονται, δηλ.

$$\frac{N_h}{N} = \frac{n_h}{n} \tag{3.3}$$

για κάθε δυνατή τιμή του $h$ ■

## Παρατήρηση 3.1

Κατά την εφαρμογή της στρωματοποιημένης δειγματοληψίας, η Πρόταση 3.2, εφόσον ισχύει, επιτρέπει στον ερευνητή να ενώσει όλες τις δειγματικές μονάδες, αγνοώντας την προέλευση του στρώματος για την καθεμιά και να υπολογίσει την εκτίμηση του μέσου του πληθυσμού με έναν δειγματικό απλό μέσο όρο. Λόγω της ευκολίας που έχει ο υπολογισμός του απλού δειγματικού μέσου όρου σε σχέση με εκείνον του $\bar{X}_{st}$ από την (3.2), επιδιώκεται αρκετά συχνά στην πράξη τα μεγέθη των δειγμάτων ανά στρώμα $n_h$, να επιλέγονται έτσι, ώστε να ικανοποιούν τη συνθήκη (3.3).

### Παρατήρηση 3.2

Οι δύο εκτιμητές, $\bar{X}_{st}$ και $\bar{X}$, αποτελούν και οι δύο μέσες τιμές των μετρήσεων του δείγματος, με τη διαφορά ότι στη μεν περίπτωση του $\bar{X}_{st}$ ο μέσος είναι σταθμισμένος και το βάρος των παρατηρήσεων του δείγματος διαφέρει από στρώμα σε στρώμα, ενώ στην περίπτωση του $\bar{X}$ τα βάρη είναι ίσα για κάθε παρατήρηση, δηλ. είναι απλός μέσος όρος. Αναλυτικά, στη μεν περίπτωση του $\bar{X}_{st}$ το βάρος της τυχαίας παρατήρησης $X_{hj}$ $(h = 1,2,..,L$ και $j = 1,2,..,n_h)$ του δείγματος είναι $\frac{N_h}{N n_h}$ και εξαρτάται από το $h$, ενώ για την περίπτωση του $\bar{X}$ βάση, βάσει του ορισμού του, τα βάρη για κάθε παρατήρηση είναι $\frac{1}{n}$.

### Παρατήρηση 3.3

Κάτω από τις υποθέσεις της Πρότασης 3.2, το στρωματοποιημένο δείγμα λέγεται και **αυτο-σταθμισμένο (self-weighting)**. Η ονομασία εξηγείται από το αποτέλεσμα της Παρατήρησης 3.2.

Για τον εκτιμητή $\bar{X}_{st}$, στη γενική περίπτωση αποδεικνύονται οι ιδιότητες που περιλαμβάνονται στις Προτάσεις που ακολουθούν.

### Πρόταση 3.3

Ο εκτιμητής $\bar{X}_{st}$ είναι αμερόληπτος εκτιμητής του $\bar{Y}$, δηλ. $\mathrm{E}(\bar{X}_{st}) = \bar{Y}$, αν και μόνον αν $\mathrm{E}(\bar{X}_h) = \bar{Y}_h$ για κάθε $h$ $(h = 1, .., L)$.

**Απόδειξη** (προφανής).

Σύμφωνα με την Πρόταση 3.3, ο εκτιμητής του μέσου για ολόκληρο τον πληθυσμό είναι αμερόληπτος, εφόσον η μέθοδος δειγματοληψίας που έχει υλοποιηθεί μέσα σε κάθε στρώμα εγγυάται αμεροληψία στους εκτιμητές που θα παραχθούν.

Το στρωματοποιημένο δειγματοληπτικό σχέδιο, σύμφωνα με τον ορισμό του, δεν επιβάλλει περιορισμό ως προς τον τρόπο δειγματοληψίας από στρώμα σε στρώμα ούτε στο μέγεθος δείγματος για το κάθε στρώμα. Η μόνη απαίτηση είναι η δειγματοληψία να διενεργείται ανεξάρτητα από στρώμα σε στρώμα. Ο διαφορετικός χειρισμός του κάθε στρώματος από τον ερευνητή αποτελεί ένα από τα βασικά πλεονεκτήματα της στρωματοποιημένης δειγματοληψίας. Ειδικές συνθήκες, πρόσβασης ή ομοιογένειας/ετερογένειας ανά στρώμα, μπορούν να αντιμετωπιστούν τοπικά κι όχι για το ευρύ σύνολο του πληθυσμού.

Μια ειδική περίπτωση στρωματοποιημένης δειγματοληψίας είναι εκείνη κατά την οποία εφαρμόζεται απλή τυχαία στο εσωτερικό του κάθε στρώματος. Το δειγματοληπτικό αυτό σχέδιο ονομάζεται **τυχαίο στρωματοποιημένο (random stratified)**. Σύμφωνα με την Πρόταση 3.3 και τα γνωστά αποτελέσματα για την απλή τυχαία, προκύπτει ότι ο εκτιμητής $\bar{X}_{st}$ είναι αμερόληπτος για το τυχαίο στρωματοποιημένο δειγματοληπτικό σχέδιο.

### Παράδειγμα 3.2

Για τα δεδομένα του Παραδείγματος 3.1 η μέση απόδοση των μαθητών στο τεστ Μαθηματικών, βάσει του στρωματοποιημένου δείγματος μεγέθους $n = 15$ που επιλέχθηκε, είναι:

$$\bar{X}_{st} = \sum_{h=1}^{L} W_h \bar{X}_h = \frac{20}{85} \times 62.83 + \frac{35}{85} \times 48.25 + \frac{30}{85} \times 69.4 = 59.15.$$

### Παράδειγμα 3.3

Μια έρευνα πραγματοποιείται με σκοπό την εκτίμηση του μέσου αριθμού ασθενών που επισκέπτονται τους γιατρούς μιας πόλης σε μια ημέρα. Ανάλογα με την εμπειρία τους, οι γιατροί της πόλης χωρίζονται σε 3 κατηγορίες. Κατηγορία 1: εμπειρία κάτω από 5 χρόνια, κατηγορία 2: εμπειρία από 5 μέχρι 10 χρόνια και κατηγορία 3: εμπειρία πάνω από 10 χρόνια. Η επιλογή του δείγματος γίνεται με τυχαία επιλογή 200 γιατρών ανά κατηγορία. Ο Πίνακας 3.3 δίνει συνοπτικά τα αποτελέσματα της έρευνας από το δείγμα. Στοιχεία ως προς τον συνολικό αριθμό γιατρών ανά κατηγορία έχουν ζητηθεί από τους ιατρικούς συλλόγους και δίνονται επίσης στον Πίνακα 3.3.

| | Αριθμός γιατρών | Δειγματικοί μέσοι |
|---|---|---|
| **Κατηγορία 1** | 500 | $\bar{X}_1$=9.5 |
| **Κατηγορία2** | 1000 | $\bar{X}_2$=15.4 |
| **Κατηγορία 3** | 2500 | $\bar{X}_3$=21.3 |

***Πίνακας 3.3*** *Παρατηρήσεις του στρωματοποιημένου δείγματος των γιατρών.*

Η δειγματοληψία που πραγματοποιήθηκε είναι η στρωματοποιημένη, γιατί έγινε μία δειγματοληψία σε κάθε κατηγορία, και ειδικότερα είναι τυχαία στρωματοποιημένη με μέγεθος δείγματος $n_1 = n_2 = n_3 = 200$ για κάθε στρώμα. Από τα στοιχεία του δείγματος, συμπεραίνεται ότι το συνολικό μέγεθος του πληθυσμού είναι $N = 4000$. Επίσης, σύμφωνα με τον συμβολισμό της παραγράφου 3.2.1, $N_1 = 500$, $N_2 = 1000$, $N_3 = 2500$ είναι τα μεγέθη των τριών στρωμάτων. Κατά συνέπεια, τα βάρη για τα στρώματα είναι:

$$W_1 = \frac{500}{4000} = 0.125, \quad W_2 = \frac{1000}{4000} = 0.25, \; W_1 = \frac{2500}{4000} = 0.625$$

απ' όπου διαπιστώνουμε ότι το τελευταίο στρώμα έχει μεγάλο βάρος σχετικά με τα δύο πρώτα. Η εκτίμηση του μέσου αριθμού επισκέψεων ασθενών στους γιατρούς της πόλης θα υπολογιστεί από τον στρωματοποιημένο δειγματικό μέσο $\bar{X}_{st}$, που δίνεται στην (3.2). Αναλυτικά,

$$\bar{X}_{st} = W_1\bar{X}_1 + W_2\bar{X}_2 + W_3\bar{X}_3 = 0.125 \times 9.5 + 0.25 \times 15.4 + 0.625 \times 21.3 = 18.35$$

Συνεπώς, βάσει του στρωματοποιημένου δείγματος, η εκτίμηση για τον μέσο αριθμό επισκέψεων ασθενών στα ιατρεία της πόλης είναι 18.35.

Το αποτέλεσμα αποτελεί σταθμισμένο μέσο όρο των επιμέρους εκτιμητών και, λόγω της μεγάλης βαρύτητας του στρώματος 3 και παράλληλα της μικρής βαρύτητας στο στρώμα 1, ο συνολικός εκτιμητής είναι πιο κοντά στον εκτιμητή που προέρχεται από το στρώμα 3.

### Πρόταση 3.4

Η διακύμανση του εκτιμητή $\bar{X}_{st}$ όταν η δειγματοληψία ανά στρώμα γίνεται ανεξάρτητα δίνεται από τη σχέση:

$$\mathrm{Var}(\bar{X}_{st}) = \sum_{h=1}^{L} W_h^2 \mathrm{Var}(\bar{X}_h) \qquad (3.4)$$

### **Απόδειξη** (προφανής)

Η απόδειξη της Πρότασης 3.4 είναι προφανής και το αποτέλεσμα για τη διασπορά είναι αρκετά γενικό, αλλά αρκετά χρήσιμο στην πράξη, ειδικά όταν το στρωματοποιημένο σχήμα είναι πολύπλοκο και δεν εφαρμόζεται ο ίδιος τρόπος δειγματοληψίας μέσα σε κάθε στρώμα. Επίσης, από την (3.4), παρόλο που η $\mathrm{Var}(\bar{X}_{st})$ είναι σε αρχική ακόμα μορφή, είναι φανερό ότι η συνολική διακύμανση του εκτιμητή προκύπτει ως άθροισμα των διακυμάνσεων των επιμέρους εκτιμητών μέσα στα στρώματα. Εάν ο χωρισμός έχει επιτύχει ομοιογένεια στο εσωτερικό των στρωμάτων, οι διακυμάνσεις των επιμέρους εκτιμήσεων είναι βελτιωμένες και, κατά συνέπεια, το ίδιο ισχύει και για τη διακύμανση του τελικού εκτιμητή. Στην ακραία περίπτωση όπου οι διακυμάνσεις $S_1^2, S_2^2, \ldots, S_L^2$ για τα στρώματα θα ήταν όλες ίσες με μηδέν, ο εκτιμητής που θα προέκυπτε θα είχε μηδενικό σφάλμα. Σημειώνεται ότι τα βάρη $W_h$ είναι αριθμοί μικρότεροι της μονάδας και επιπλέον εμφανίζονται στην (3.4) εκφρασμένοι στο τετράγωνο.

Ειδικές περιπτώσεις για το αποτέλεσμα της Πρότασης 3.4 προκύπτουν όταν λάβει κανείς ειδικές περιπτώσεις για τον τρόπο δειγματοληψίας ανά στρώμα και το μέγεθος του δείγματος ανά στρώμα.

### Πρόταση 3.5

Η διακύμανση του εκτιμητή $\bar{X}_{st}$ για την τυχαία στρωματοποιημένη δειγματοληψία δίνεται από τη σχέση:

$$\mathrm{Var}(\bar{X}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{1-f_h}{n_h} S_h^2 \qquad (3.5)$$

όπου $n_h$ και $f_h$ είναι το μέγεθος και το πηλίκο του δείγματος ανά στρώμα αντίστοιχα, και $S_h^2$ είναι η διασπορά των μετρήσεων του χαρακτηριστικού στο στρώμα $h$.

### Απόδειξη

Η απόδειξη της διακύμανσης (3.5) προκύπτει, εάν εφαρμοστεί η Πρόταση 2.3 για τη διακύμανση του εκτιμητή του μέσου κάτω από την απλή τυχαία, σε κάθε στρώμα $h$ ■

### Πόρισμα 3.1

Αν το πηλίκο του δείγματος για κάθε στρώμα μπορεί να θεωρηθεί αμελητέο, δηλ. $f_h \approx 0$ για κάθε $h$, τότε:

$$\mathrm{Var}(\bar{X}_{st}) = \sum_{h\ 1}^{L} W_h^2 \frac{S_h^2}{n_h}$$

### Πρόταση 3.6

Για την τυχαία στρωματοποιημένη δειγματοληψία με $W_h = w_h$ για κάθε $h$ $(h = 1,2,\ldots,L)$, ισχύει:

$$\mathrm{Var}(\bar{X}_{st}) = \frac{1-f}{n} \sum_{h=1}^{L} W_h\, S_h^2 \qquad (3.6)$$

### Απόδειξη

Εάν $W_h = w_h$ για κάθε $h$, ισοδύναμα ισχύει επίσης $\frac{n_h}{N_h} = \frac{n}{N}$ ή $n_h = nW_h$. Επίσης, από την ίδια αναλογία προκύπτει και $f_h = f$, όπου $f = \frac{n}{N}$, δηλ. ίδιο πηλίκο δείγματος για κάθε στρώμα $h$.

Με τη βοήθεια των αποτελεσμάτων αυτών, η σχέση (3.5) γίνεται:

$$\mathrm{Var}(\bar{X}_{st}) = \sum_{h\ 1}^{L} W_h^2 \frac{1-f}{n\,W_h} S_h^2 = \frac{1-f}{n} \sum_{h=1}^{L} W_h\, S_h^2$$

που συμπληρώνει την απόδειξη ■

## Παρατήρηση 3.4

Από την Πρόταση 3.6 προκύπτει το συμπέρασμα ότι ακόμα και για την περίπτωση όπου ο στρωματοποιημένος μέσος $\bar{X}_{st}$ ταυτίζεται αριθμητικά με τον εκτιμητή $\bar{X}$ που προκύπτει από ένα απλό τυχαίο δείγμα, οι διακυμάνσεις των δύο εκτιμητών διαφέρουν. Η γνώση ότι το δείγμα έχει προέλθει από μια ανεξάρτητη δειγματοληψία ανά στρώμα επιτρέπει τη χρήση του αποτελέσματος της Πρότασης 3.6 για τον $\bar{X}_{st}$, ενώ είναι $\mathrm{Var}(\bar{X}) = \frac{1-f}{n} S^2$ για τον $\bar{X}$. Σε επόμενες παραγράφους του παρόντος κεφαλαίου, θα γίνει η σύγκριση των δύο διακυμάνσεων.

**Πόρισμα 3.2**

Αν επιπλέον των υποθέσεων της Πρότασης 3.6 ισχύει ότι $S_h^2 = S_w^2$ για κάθε $h$, δηλ. η διασπορά του υπό μελέτη χαρακτηριστικού είναι σταθερή για όλα τα στρώματα, τότε:

$$\mathrm{Var}(\bar{X}_{st}) = \frac{1-f}{n_h} S_w^2$$

**Απόδειξη**

Προκύπτει άμεσα από την Πρόταση 3.6 και λαμβάνοντας υπόψη ότι $\sum_{h=1}^{L} W_h = 1$ ∎

**Πόρισμα 3.3**

Για το τυπικό σφάλμα του εκτιμητή, ισχύει ο γενικός ορισμός $\mathrm{se}(\bar{X}_{st}) = \sqrt{\mathrm{Var}(\bar{X}_{st})}$, όπου $\mathrm{Var}(\bar{X}_{st})$ η διακύμανση του στρωματοποιημένου μέσου $\bar{X}_{st}$.

Περισσότερες ειδικές περιπτώσεις για τον υπολογισμό της διακύμανσης του $\bar{X}_{st}$ θα προκύψουν όταν συμπεριληφθούν περαιτέρω ειδικές περιπτώσεις για τα μεγέθη του δείγματος ανά στρώμα.

### 3.2.3 Εκτίμηση της διακύμανσης εκτιμητή

Για την εκτίμηση της διακύμανσης του εκτιμητή $\bar{X}_{st}$ στη στρωματοποιημένη, ισχύουν ανάλογα αποτελέσματα με εκείνα της απλής τυχαίας. Το βασικό συστατικό της εκτίμησης είναι η εκτίμηση της πληθυσμιακής διασποράς $S_h^2$ του κάθε στρώματος από την αντίστοιχη δειγματική $s_h^2$ (βλ. Πίνακα 3.1).Αν η δειγματοληψία στο εσωτερικό κάθε στρώματος είναι απλή τυχαία, τότε η κάθε μία από τις $L$ εκτιμήσεις διασπορών είναι αμερόληπτη. Για την περίπτωση αυτή, η διακύμανση του εκτιμητή του μέσου του πληθυσμού, όπως δίνεται από την Πρόταση 3.5, εκτιμάται από την:

$$\hat{\mathrm{V}}\mathrm{ar}(\bar{X}_{st}) = \sum_{h\ 1}^{L} W_h^2 \frac{1-f_h}{n_h} s_h^2$$

Στην ειδική περίπτωση κατά την οποία οι διασπορές $S_h^2$ σε κάθε στρώμα είναι ίσες, και έστω ίσες με $S_w^2$, η εκτίμηση του $S_h^2$ μπορεί να προκύψει από τον από κοινού (pooled) εκτιμητή της διασποράς, δηλ.

$$\hat{S}_w^2 = \frac{1}{N-L} \sum_{h=1}^{L} \sum_{j=1}^{n_h} \left(X_{hj} - \bar{X}_h\right)^2.$$

Οι βαθμοί ελευθερίας του $\hat{S}_w^2$ εκτιμητή είναι $N - L$.

**Παράδειγμα 3.4**

Στα δεδομένα του Παραδείγματος 3.3, προστίθεται ότι οι δειγματικές διασπορές των μετρήσεων για τα τρία στρώματα είναι $s_1^2 = 73.6$, $s_2^2 = 340.10$ και $s_3^2 = 730.89$. Για την εκτίμηση του μέσου αριθμού επισκέψεων στο Παράδειγμα 3.3 με $\bar{X}_{st} = 18.35$, η εκτιμώμενη διακύμανση είναι:

$$\hat{\mathrm{V}}\mathrm{ar}(\bar{X}_{st}) = \sum_{h\ 1}^{L} W_h^2 \frac{1-f_h}{n_h} s_h^2 = 0.125^2 \frac{1-200/500}{200} 73.6 +$$

$$0.25^2 \frac{1-200/1000}{200} 340.10 + 0.625^2 \frac{1-200/2500}{200} 730.89 = 1.40$$

Άρα ο μέσος αριθμός επισκέψεων εκτιμάται σε 18.35 και το εκτιμώμενο τυπικό σφάλμα της εκτίμησης είναι $\hat{se}(\bar{X}_{st}) = \sqrt{1.40} = 1.183973 \cong 1.18$ επισκέψεις.

## 3.3. Διάστημα Εμπιστοσύνης

Η κατασκευή ενός διαστήματος εμπιστοσύνης για την αληθινή μέση τιμή του πληθυσμού $\bar{Y}$ βάσει ενός στρωματοποιημένου δείγματος γίνεται με χρήση της υπόθεσης της κανονικής κατανομής του εκτιμητή $\bar{X}_{st}$. Κατά συνέπεια, ένα προσεγγιστικό διάστημα εμπιστοσύνης για τη μέση τιμή, με βαθμό εμπιστοσύνης $(1-a)100\%$, έχει άκρα:

$$\left(\bar{X}_{st} - z_{a/2}\hat{se}(\bar{X}_{st}), \quad \bar{X}_{st} + z_{a/2}\hat{se}(\bar{X}_{st})\right)$$

όπου $z_a$ είναι το άνω $a$-εκατοστιαίο σημείο για την τυπική κανονική κατανομή, $Z$. Η προσέγγιση από την κανονική κατανομή και η χρήση των εκατοστιαίων σημείων της $Z$ θα είναι αποδεκτή και θα δώσει αξιόπιστα αποτελέσματα, μόνο αν η κάθε μία από τις $L$ εκιμήσεις των $S_h^2$ από τις δειγματικές $s_h^2$ $(h = 1, \dots, L)$ είναι ικανοποιητική. Αν η συνθήκη αυτή δεν πληρούται, τότε στο διάστημα εμπιστοσύνης πρέπει να αντικατασταθούν σημεία από την $t$-κατανομή, αντί της $Z$. Παρόλα αυτά, για τη στρωματοποιημένη δειγματοληψία, και επειδή ο εκτιμητής της διακύμανσης του $\bar{X}_{st}$ είναι πολύπλοκος και αποτελείται από πολλές επιμέρους εκτιμήσεις, δεν είναι ξεκάθαρο ποια είναι τα αποδεκτά μεγέθη δείγματος – ανά στρώμα και συνολικά - που εξασφαλίζουν ικανοποιητική προσέγγιση. Παράλληλα, για τη χρήση της $t$-κατανομής είναι επίσης αρκετά δύσκολη επιλογή των βαθμών ελευθερίας της κατανομής, θέμα για το οποίο δεν δίνονται σαφείς οδηγίες στη βιβλιογραφία.

### Παράδειγμα 3.5

Για τα δεδομένα του Παραδείγματος 3.2, το 95% διάστημα εμπιστοσύνης για την εκτίμηση του μέσου αριθμού επισκέψεων στα ιατρεία της πόλης είναι

$$\left(\bar{X}_{st} - z_{a/2}\hat{se}(\bar{X}_{st}),\ \bar{X}_{st} + z_{a/2}\hat{se}(\bar{X}_{st})\right)$$

$$= (18.35 - 1.96 \times 1.18,\ 18.35 + 1.96 \times 1.18) = (16.04,\ 20.66)$$

Συνεπώς, η άγνωστη παράμετρος του πληθυσμού, που αφορά τον μέσο αριθμό επισκέψεων ασθενών σε ιατρεία της πόλης, εκτιμάται ότι ανήκει στο διάστημα $(16.04,\ 20.66)$ με πιθανότητα σφάλματος 5%.

### Παράδειγμα 3.6

Ο χρόνος (σε ώρες) που απαιτείται για τη διεκπεραίωση των φακέλων μιας υπηρεσίας από τους υπαλλήλους της εξαρτάται σημαντικά από τη σοβαρότητα κάθε υπόθεσης. Στο πλαίσιο μιας έρευνας για την εκτίμηση των εργατοωρών που χρειάζεται η υπηρεσία, οι 1500 συνολικά υποθέσεις του περασμένου μήνα χωρίζονται βάσει της σοβαρότητάς τους σε δύο κατηγορίες, όπως παρουσιάζεται στον Πίνακα 3.4. Ένα self-weighting δείγμα, μεγέθους $n = 100$ φακέλων, επιλέγεται από τις δύο κατηγορίες και τα αποτελέσματα της δειγματοληψίας παρουσιάζονται επίσης στον Πίνακα 3.4.

| Σοβαρότητα Υπόθεσης | Αριθμός Υποθέσεων | Δειγματικός μέσος χρόνος (ώρες) | Τυπική απόκλιση των χρόνων (ώρες) |
|---|---|---|---|
| **Μικρή** | 975 | 1.8 | 0.7 |
| **Μεγάλη** | 525 | 7.5 | 2.4 |

***Πίνακας 3.4*** *Στοιχεία φακέλων και αποτελέσματα έρευνας για τον χρόνο επεξεργασίας.*

Το δείγμα είναι self-weighting, άρα τα μεγέθη του δείγματος σε κάθε στρώμα είναι:

$$n_1 = 100\,\frac{975}{1500} = 65 \quad \text{και} \quad n_2 = 100\,\frac{525}{1500} = 35.$$

Ο μέσος χρόνος που απαιτείται για τη διεκπεραίωση των φακέλων εκτιμάται ως:

$$\bar{X}_{st} = \frac{1}{n}\sum_{h=1}^{L}\sum_{j=1}^{N_h} X_{hj} = \frac{\sum_{j=1}^{65} X_{1j} + \sum_{j=1}^{35} X_{2j}}{100} = \frac{65 * 1.8 + 35 * 7.5}{100} = 3.795\ \text{ώρες}$$

Ο εκτιμητής υπολογίζεται ισοδύναμα από τον τύπο του απλού μέσου, γιατί το δείγμα έχει επιλεγεί με $w_h = W_h$ και $n_h = nW_h$ $(h = 1,2)$.

Για τη διακύμανση της παραπάνω εκτίμησης εφαρμόζεται η έκφραση (3.6), και ένας εκτιμητής της είναι ο:

$$\hat{\text{V}}\text{ar}(\bar{X}_{st}) = \frac{1-f}{n}\sum_{h=1}^{L} W_h\, S_h^2 = \frac{1 - 100/1500}{100}(0.65 \times 0.7^2 + 0.35 \times 2.4^2) = 0.0218$$

ή το τυπικό σφάλμα της εκτίμησης είναι $\hat{\text{se}}(\bar{X}_{st}) = 0.148$ ώρες.

## 3.4. Εκτίμηση συνόλου και ποσοστού

Το πρόβλημα της εκτίμησης του συνόλου του πληθυσμού με τη βοήθεια ενός στρωματοποιημένου δείγματος αντιμετωπίζεται με ανάλογο τρόπο, όπως και στην απλή τυχαία δειγματοληψία. Η εκτίμηση του συνόλου συνδέεται με την εκτίμηση του μέσου, βάσει της σχέσης $Y_T = N\bar{Y}$ που ισχύει μεταξύ των αληθινών ποσοτήτων $Y_T$ και $\bar{Y}$ αντίστοιχα. Η έκφραση αυτή εφαρμόζεται για την εκτίμηση του συνόλου σε κάθε επιμέρους δειγματοληψία ανά στρώμα, έστω $Y_{hT}$ .

Τα αποτελέσματα για την εκτίμηση, τη διασπορά και την εκτιμώμενη διασπορά του εκτιμητή λαμβάνονται στη συνέχεια ως άμεση συνέπεια από τα αντίστοιχα αποτελέσματα για την εκτίμηση του μέσου. Τα αποτελέσματα αυτά παρουσιάζονται συνοπτικά στον Πίνακα 3.5.

| | |
|---|---|
| Εκτιμητής συνόλου ανά στρώμα $h$ | $\hat{Y}_{hT} = N_h \bar{X}_h$ |
| Εκτιμητής συνόλου για τον πληθυσμό | $\hat{Y}_{T,st} = \sum_{h=1}^{L} N_h\, \bar{X}_h$ |
| Διακύμανση συνόλου πληθυσμού για την τυχαία στρωματοποιημένη | $\text{Var}(\hat{Y}_{T,st}) = \sum_{h=1}^{L} N_h^2 \frac{1-f_h}{n_h} S_h^2$ |
| Διακύμανση συνόλου πληθυσμού για self-weighting στρωματοποιημένο δείγμα | $\text{Var}(\hat{Y}_{T,st}) = \frac{1-f}{n} N \sum_{h=1}^{L} N_h\, S_h^2$ |

| | |
|---|---|
| Τυπικό σφάλμα εκτιμητή του συνόλου για την τυχαία στρωματοποιημένη | $\mathrm{se}(\hat{Y}_{T,st}) = \sqrt{\sum_{h=1}^{L} N_h^2 \frac{1-f_h}{n_h} S_h^2}$ |
| Εκτιμώμενο τυπικό σφάλμα εκτιμητή του συνόλου για την τυχαία στρωματοποιημένη | $\widehat{\mathrm{se}}(\hat{Y}_{T,st}) = \sqrt{\sum_{h=1}^{L} N_h^2 \frac{1-f_h}{n_h} s_h^2}$ |

***Πίνακας 3.5*** *Εκτιμητής συνόλου και οι ιδιότητες για τη στρωματοποιημένη.*

## Παράδειγμα 3.7

Για τα δεδομένα του Παραδείγματος 3.3, αν αντί του μέσου αριθμού επισκέψεων ασθενών στα ιατρεία της πόλης, ενδιαφερόμαστε για τον συνολικό αριθμό επισκέψεων που πραγματοποιούνται, τότε:

$$\hat{Y}_{T,st} = N_1\bar{X}_1 + N_2\bar{X}_2 + N_3\bar{X}_3 = 500 \times 9.5 + 1000 \times 15.4 + 2500 \times 21.3 = 73400$$

Άρα, βάσει της έρευνας, εκτιμάται ότι οι συνολικές επισκέψεις που πραγματοποιούνται είναι 73400. Το τυπικό σφάλμα της εκτίμησης εκτιμάται ως:

$$\mathrm{se}(\hat{Y}_{T,st}) = \sqrt{\sum_{h,}^{L} N_h^2 \frac{1-f_h}{n_h} S_h^2}$$

$$= \sqrt{500^2 \frac{1-2/5}{200} 73.6 + 1000^2 \frac{1-2/10}{200} 340.10 + 2500^2 \frac{1-2/25}{200} 730.89}$$

$$= 4735.89$$

Για το συγκεκριμένο παράδειγμα, κι επειδή έχει προηγηθεί η λύση του προβλήματος για εκτίμηση του μέσου στα Παραδείγματα 3.2 και 3.3, ένας πιο σύντομος, ισοδύναμος, τρόπος για τη στατιστική συμπερασματολογία του συνόλου είναι ο ακόλουθος:

$$\hat{Y}_{T,st} = N\bar{X}_{st} = 4000 \times 18.35 = 73400$$

Το 95% διάστημα εμπιστοσύνης για τον συνολικό αριθμό επισκέψεων είναι:

$$\left(\hat{Y}_{T,st} - z_{a/2}\widehat{\mathrm{se}}(\hat{Y}_{T,st}),\ \hat{Y}_{T,st} + z_{a/2}\widehat{\mathrm{se}}(\hat{Y}_{T,st})\right)$$

$$= (73400 - 1.96 \times 4735.89,\quad 73400 + 1.96 \times 4735.89) = (64117.66,\ 82682.34)$$

Άρα, η αληθινή τιμή του συνολικού αριθμού των επισκέψεων εκτιμάται ότι ανήκει στο διάστημα $(64117.66, 82682.34)$ με πιθανότητα σφάλματος 5%.

Στην ίδια βάση με εκείνη της εκτίμησης του ποσοστού για την απλή τυχαία δειγματοληψία γίνεται και η εκτίμηση του ποσοστού για τη στρωματοποιημένη. Έστω ότι το υπό μελέτη χαρακτηριστικό έχει διακριτή φύση και αντιστοιχεί σε μια κατηγορική μεταβλητή. Έστω ακόμα, ότι το ενδιαφέρον της έρευνας επικεντρώνεται στην εκτίμηση του ποσοστού $P$, σύμφωνα με το οποίο τα μέλη του πληθυσμού ανήκουν σε μια κατηγορία. Η εκτίμηση του ποσοστού αυτού ανά στρώμα, έστω $P_h$, γίνεται μέσω της εκτίμησης της μέσης

τιμής $\bar{X}_h$ ενός χαρακτηριστικού $Y$ που έχει κωδικοποιηθεί κατάλληλα, ώστε να παίρνει τις τιμές 0 και 1. Εκτιμώντας το ποσοστό ανά στρώμα, δίνουμε στη συνέχεια τον εκτιμητή για το συνολικό ποσοστό, ακολουθώντας ανάλογη διαδικασία, όπως και για το συνεχές χαρακτηριστικό, δηλ. την έκφραση με τον σταθμισμένο μέσο (3.2). Άρα, η εκτίμηση ενός ποσοστού, όπως και στην απλή τυχαία δειγματοληψία, παραπέμπει στην εκτίμηση μιας μέσης τιμής.

Ανάλογα αποτελέσματα ισχύουν και για την εκτίμηση του $A$, δηλ. του πλήθους των μελών του πληθυσμού που ανήκουν στην υπό μελέτη κατηγορία. Έστω $P_h$ η τιμή του ποσοστού $P$ στο στρώμα $h$ και $A_h$ το πλήθος των μελών του στρώματος $h$ που ανήκουν στην κατηγορία. Έστω ακόμα ότι $\hat{P}_{st}$ και $\hat{A}_{st}$ συμβολίζουν τις εκτιμήσεις των συνολικών (για ολόκληρο τον πληθυσμό) $P$ και $A$ αντίστοιχα, κάτω από ένα στρωματοποιημένο δειγματοληπτικό σχήμα. Για ένα στρωματοποιημένο δείγμα με $n_h$ μονάδες ανά στρώμα και $a_h$ καταγεγραμμένες μονάδες να ανήκουν στην υπό μελέτη κατηγορία, ο Πίνακας 3.6 παρουσιάζει τα αποτελέσματα για τις εκφράσεις του εκτιμητή, τη διασπορά, το τυπικό σφάλμα και την εκτιμώμενη διακύμανση για $P$ και $A$, ανά στρώμα και συνολικά.

| | |
|---|---|
| Εκτιμητής ποσοστού ανά στρώμα $h$ | $\hat{P}_h = \frac{a_h}{n_h}$ |
| Εκτιμητής ποσοστού | $\hat{P}_{st} = \sum_{h=1}^{L} W_h \hat{P}_h$ |
| Διακύμανση εκτιμητή του ποσοστού για την τυχαία στρωματοποιημένη | $\mathrm{Var}(\hat{P}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{1-f_h}{n_h} \frac{N_h P_h (1-P_h)}{N_h - 1}$ |
| Εκτιμώμενη διακύμανση εκτιμητή του ποσοστού για την απλή στρωματοποιημένη | $\hat{\mathrm{Var}}(\hat{P}_{st}) = \sum_{h=1}^{L} W_h^2 (1-f_h) \frac{\hat{P}_h (1-\hat{P}_h)}{n_h - 1}$ |
| Τυπικό σφάλμα εκτιμητή του ποσοστού για την τυχαία στρωματοποιημένη | $\mathrm{se}(\hat{P}_{st}) = \sqrt{\sum_{h=1}^{L} W_h^2 \frac{1-f_h}{n_h} \frac{N_h P_h (1-P_h)}{N_h - 1}}$ |
| Εκτιμητής συνόλου μελών $\boldsymbol{A}$ | $\hat{A}_{st} = \sum_{h=1}^{L} N_h \hat{P}_h$ |
| Διακύμανση εκτιμητή του $\boldsymbol{A}$ στη στρωματοποιημένη δειγματοληψία | $\mathrm{Var}(\hat{A}_{st}) = \sum_{h=1}^{L} N_h^2 \frac{1-f_h}{n_h} \frac{N_h P_h (1-P_h)}{N_h - 1}$ |
| Εκτιμώμενη διακύμανση εκτιμητή του $\boldsymbol{A}$ στη στρωματοποιημένη δειγματοληψία | $\hat{\mathrm{Var}}(\hat{A}_{st}) = \sum_{h=1}^{L} N_h^2 (1-f_h) \frac{\hat{P}_h (1-\hat{P}_h)}{n_h - 1}$ |

| | |
|---|---|
| Τυπικό σφάλμα εκτιμητή του ***A*** στη στρωματοποιημένη δειγματοληψία | $\text{se}(\hat{A}_{st}) = \sqrt{\sum_{h=1}^{L} N_h^2 \frac{1-f_h}{n_h} \frac{N_h P_h (1-P_h)}{N_h - 1}}$ |

***Πίνακας 3.6*** *Εκτιμητές ποσοστού και συνόλου για τη στρωματοποιημένη.*

## Παράδειγμα 3.8

Κατά τη διάρκεια της δειγματοληψίας για την έρευνα που παρουσιάζεται στο Παράδειγμα 3.3, καταγράφονται επιπλέον τα ποσοστά των ασθενών που δεν λαμβάνουν απόδειξη από τον γιατρό. Τα συνολικά στοιχεία της έρευνας δίνονται στον Πίνακα 3.7.

| | **Αριθμός γιατρών** | **Δειγματικοί μέσοι** $\bar{X}_h$ | **Ποσοστά** $\hat{P}_h$ |
|---|---|---|---|
| **Κατηγορία 1** | 500 | 9.5 | 0.15 |
| **Κατηγορία2** | 1000 | 15.4 | 0.25 |
| **Κατηγορία 3** | 2500 | 21.3 | 0.40 |

***Πίνακας 3.7*** *Παρατηρήσεις του στρωματοποιημένου δείγματος των γιατρών.*

Εάν είναι επιθυμητό με βάση το δείγμα να δοθεί μια εκτίμηση για το ποσοστό των αποδείξεων που δεν εκδίδονται από το σύνολο των γιατρών της πόλης, τότε η εκτίμηση αυτή υπολογίζεται από το στρωματοποιημένο δείγμα και είναι:

$$\hat{P}_{st} = \sum_{h=1}^{L} W_h \hat{P}_h = 0.125 \times 0.15 + 0.25 \times 0.25 + 0.625 \times .40 = 0.3312$$

Συνεπώς, βάσει του στρωματοποιημένου δείγματος, εκτιμάται ότι για την πόλη που διεξάγεται η έρευνα, το ποσοστό των αποδείξεων που δεν εκδίδονται είναι 33.12%.

Για ένα 95% διάστημα εμπιστοσύνης της ίδιας ποσότητας, χρειάζεται πρώτα να υπολογιστεί το τυπικό σφάλμα της εκτίμησης, $\widehat{\text{se}}(\hat{P}_{st})$. Αρχικά, $\widehat{\text{se}}(\hat{P}_{st}) = \sqrt{\widehat{\text{Var}}(\hat{P}_{st})}$ και από τον Πίνακα 3.6 υπολογίζουμε την $\widehat{\text{Var}}(\hat{P}_{st})$. Θα είναι:

$$\widehat{\text{se}}(\hat{P}_{st}) = \sqrt{\widehat{\text{Var}}(\hat{P}_{st})} = \sqrt{\sum_{h=1}^{L} W_h^2 (1-f_h) \frac{\hat{P}_h(1-\hat{P}_h)}{n_h - 1}}$$

$$= \sqrt{0.125^2 \frac{1-\frac{200}{500}}{200} \frac{0.15 \times 0.85}{199} + 0.25^2 \frac{1-\frac{200}{1000}}{200} \frac{0.25 \times 0.75}{199} + 0.625^2 \frac{1-\frac{200}{2500}}{200}}$$

$$= \sqrt{0.000484} = 0.022$$

Άρα το τυπικό σφάλμα της εκτίμησης του ποσοστού $\hat{P}_{st} = 0.33$ είναι 0.022.

Εάν, περαιτέρω, ενδιαφερόμαστε για τον αριθμό των αποδείξεων που δεν εκδίδονται, τότε:

$$\hat{A}_{st} = \sum_{h=1}^{L} \hat{Y}_{T,h} \hat{P}_h = \hat{Y}_{T,st} \hat{P}_{st}$$

όπου ο πληθυσμός, σ' αυτή την περίπτωση, είναι το σύνολο των επισκέψεων που πραγματοποιούνται στα ιατρεία της πόλης και το μέγεθος αυτού έχει εκτιμηθεί στο Παράδειγμα 3.7 ως $\hat{Y}_{T,st} = 73400$. Ανάλογα, $\hat{Y}_{T,h}$ είναι ο αριθμός των επισκέψεων ασθενών ανά κατηγορία γιατρών. Συνολικά:

$$\hat{A}_{st} = 73400 \times 0.33 = 24222$$

Συνεπώς, βάσει των δεδομένων από τη στρωματοποιημένη δειγματοληψία που έχει διεξαχθεί, εκτιμάται ότι δεν εκδίδονται 24222 αποδείξεις συνολικά από τους γιατρούς της πόλης.

## 3.5. Καταμερισμός δείγματος στη στρωματοποιημένη δειγματοληψία

Στην παρούσα παράγραφο, το αντικείμενο μελέτης είναι ένα ειδικότερο πρόβλημα της στρωματοποιημένης δειγματοληψίας που ανακύπτει κατά το σχεδιασμό της έρευνας και είναι καθοριστικό για την ακρίβεια των εκτιμητών που θα επιτευχθούν. Το πρόβλημα αφορά τον καταμερισμό του μεγέθους του δείγματος στο εσωτερικό του κάθε στρώματος. Κάτω από την υπόθεση, αρχικά, ότι το συνολικό μέγεθος του δείγματος είναι γνωστό, έστω $n$, το ερώτημα στο οποίο πρέπει να απαντήσει ο ερευνητής πριν από τη διεξαγωγή της έρευνας είναι πώς θα μοιράσει το συνολικό δείγμα $n$ στα $L$ στρώματα. Το πλήθος των δυνατών τρόπων καταμερισμού είναι μεγάλο και ο κάθε καταμερισμός δεν οδηγεί σε εκτιμητές με ισοδύναμες ως προς όλους τους υπόλοιπους συνδυασμούς ιδιότητες.

Μαθηματικά, το πρόβλημα είναι ισοδύναμο με τον προσδιορισμό των $n_1, n_2, \dots, n_L$ οι οποίοι είναι θετικοί ακέραιοι αριθμοί, τέτοιοι ώστε $n_1 + \cdots + n_L = n$. Οι διάφοροι τρόποι προσέγγισης του προβλήματος, κατά τη βιβλιογραφία, χωρίζονται σε δύο κατηγορίες. Στη μία κατηγορία, ο κανόνας χωρισμού είναι η ευκολία στην υλοποίηση της έρευνας και την εξαγωγή των αποτελεσμάτων, δηλ. πρακτικό πλεονέκτημα ενώ στη δεύτερη κατηγορία, ο κανόνας είναι η βελτιστοποίηση των ιδιοτήτων του εκτιμητή που θα προκύψει. Στη συνέχεια, αναπτύσσονται οι κυριότερες μεθοδολογίες καταμερισμού του δείγματος σε στρώματα. Για περισσότερες λεπτομέρειες ως προς τις μαθηματικές αποδείξεις, παραπέμπουμε στο βιβλίο του Cochran (1977, Κεφ. 5).

### 3.5.1 Αναλογικός Καταμερισμός (Proportional Allocation)

Η αναλογική μέθοδος καταμερισμού είναι μια μέθοδος που εμπίπτει στην πρώτη κατηγορία μεθοδολογιών. Το κριτήριο επιλογής των μεγεθών του δείγματος κάτω από τον καταμερισμό αυτό είναι τα πληθυσμιακά μεγέθη, ή, ισοδύναμα, τα βάρη των στρωμάτων στον πληθυσμό να είναι ίσα με τα αντίστοιχα στο δείγμα. Η ιδέα πίσω από τον αναλογικό καταμερισμό είναι η αναλογία του μεγέθους του δείγματος $n_h$ κάθε στρώματος ως προς το σύνολο $n$, να είναι ίση με την αναλογία του μεγέθους του στρώματος $N_h$ ως προς το $N$.

Ο προσδιορισμός των αγνώστων $n_h$, για τα διάφορα $h$ $(h = 1, \dots, L)$ προκύπτει έτσι από τη λύση των ισοτήτων των αναλογιών:

$$\frac{n_1}{n} = \frac{N_1}{N}$$

$$\frac{n_2}{n} = \frac{N_2}{N}$$

$$\dots$$

$$\frac{n_L}{n} = \frac{N_L}{N}$$

ή, ισοδύναμα:

$$n_h = n\frac{N_h}{N}, \qquad h = 1,2, \dots, L \tag{3.7}$$

Κατά τον αναλογικό καταμερισμό, τηρούνται οι αναλογίες και συνεπώς ένα μεγάλο πληθυσμιακά στρώμα θα εκπροσωπηθεί σε μεγάλη αναλογία στο δείγμα, ενώ αντίστοιχα ένα μικρό αριθμητικά στρώμα, ανεξαρτήτως του περιεχομένου του, θα έχει λίγους, σχετικά με τα υπόλοιπα στρώματα, εκπροσώπους.

Η κατανομή του δείγματος $n$ στα στρώματα σύμφωνα με τον τύπο (3.7) λέγεται **αναλογικός καταμερισμός (proportional allocation).**

Η έκφραση (3.7) δίνει επίσης:

$$\frac{n_h}{n} = \frac{N_h}{N}, \qquad h = 1,2, \dots, L$$

ή $w_h = W_h$, που είναι η συνθήκη για τον self-weighting εκτιμητή. Άρα, ο αναλογικός καταμερισμός συνεπάγεται την ισότητα των δύο μέσων $\bar{X}_{st}$ και $\bar{X}$ ή, πιο αναλυτικά:

$$\bar{X}_{st} = \frac{\sum_{h=1}^{L} \sum_{j=1}^{N_h} X_{hj}}{n}$$

και, κατά συνέπεια, η υιοθέτηση του καταμερισμού αυτού εγγυάται μία εύκολη έκφραση για τον εκτιμητή που θα προκύψει.

Για τη διακύμανση του εκτιμητή κάτω από τη συνθήκη (3.7), ισχύουν οι προϋποθέσεις της Πρότασης 3.6, και η διακύμανση για τον αναλογικό καταμερισμό δίνεται από τη σχέση (3.6),

$$\mathrm{Var}(\bar{X}_{st}) = \frac{1-f}{n} \sum_{h=1}^{L} W_h \, S_h^2$$

Συνολικά, τόσο το κριτήριο καταμερισμού, όσο και οι εκφράσεις του εκτιμητή και της διακύμανσής του, υπολογίζονται εύκολα κατά τον αναλογικό καταμερισμό, και γι'αυτό είναι αρκετά συχνή στην πράξη η εφαρμογή του, ειδικά εάν δεν είναι διαθέσιμα άλλα περαιτέρω στοιχεία για τα στρώματα παρά μόνο τα μεγέθη τους $N_h$.

### 3.5.2 Βέλτιστος Καταμερισμός (Optimal Allocation)

Ο βέλτιστος τρόπος καταμερισμού του συνολικού μεγέθους του δείγματος $n$ στα στρώματα προκύπτει, εάν τα μεγέθη $n_h$ ανά στρώμα θεωρηθούν άγνωστοι αριθμοί και υπολογιστούν υιοθετώντας ένα κριτήριο βελτιστοποίησης. Το σύνηθες κριτήριο βελτιστοποίησης είναι η ελαχιστοποίηση της διακύμανσης του εκτιμητή που θα προκύψει για ένα σταθερό συνολικό κόστος που είναι διαθέσιμο για την έρευνα και ακολουθώντας τη μέθοδο δειγματοληψίας που πρόκειται να υλοποιηθεί.

Εάν το συνολικό κόστος $C$ της έρευνας δίνεται από μια απλή μορφή, π.χ. γραμμική,

$$C = c_0 + \sum_{h=1}^{L} c_h n_h$$

όπου $c_0$ είναι ένα αρχικό κόστος της οργάνωσης της έρευνας και $c_h$ ο συντελεστής κόστους ανά μονάδα δειγματοληψίας για το στρώμα $h$ $(h = 1,2, \dots, L)$, ο υπολογισμός των αγνώστων $n_h$ μπορεί να γίνει αναλυτικά, και το αποτέλεσμα δίνει η Πρόταση 3.7 που ακολουθεί.

**Πρόταση 3.7**

Για την τυχαία στρωματοποιημένη δειγματοληψία με γραμμική συνάρτηση κόστους, η διακύμανση του εκτιμητή του μέσου γίνεται ελάχιστη για δοθέν σταθερό κόστος, όταν τα $n_h$ είναι ανάλογα των ποσοτήτων $N_h S_h / \sqrt{c_h}$.

**Απόδειξη**

Για την τυχαία στρωματοποιημένη δειγματοληψία, η διακύμανση του εκτιμητή $\bar{X}_{st}$ είναι:

$$\mathrm{Var}(\bar{X}_{st}) = \sum_{h=1}^{L} \frac{1-f_h}{n_h} W_h^2 S_h^2 = \sum_{h\ 1}^{L} \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{N_h}$$

Το ζητούμενο πρόβλημα, συνεπώς, ισοδυναμεί με την ελαχιστοποίηση της τελευταίας έκφρασης ως προς $n_h$, δοθέντος ότι $\sum_{h=1}^{L} c_h n_h = C - c_0$.

Χρησιμοποιώντας τους πολλαπλασιαστές Lagrange για την ελαχιστοποίηση, ορίζεται αρχικά η συνάρτηση:

$$\varphi(n_1, n_2, .., n_L, \lambda) = \sum_{h\lambda}^{L} \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{N_h} + \lambda(\sum_{h\lambda}^{L} c_h n_h - C + c_0)\,.$$

Παραγωγίζοντας ως προς τους αγνώστους $n_h$ και $\lambda$ και εξισώνοντας τις πρώτες παραγώγους με μηδέν, προκύπτουν οι εξισώσεις:

$$-\frac{W_h^2 S_h^2}{n_h^2} + \lambda c_h = 0 \qquad (h = 1,2,\dots,L)$$

από τις οποίες ισοδύναμα προκύπτει ότι:

$$\lambda n_h^2 = \frac{W_h^2 S_h^2}{c_h}$$

ή:

$$n_h \sqrt{\lambda} = \sqrt{\frac{W_h^2 S_h^2}{c_h}}\ (h = 1,2,\dots,L) \tag{3.8}$$

Προσθέτοντας τις τελευταίες εξισώσεις για όλα τα $h$ δίνει, προκύπτει:

$$n\sqrt{\lambda} = \sum_{i=1}^{L} \sqrt{\frac{W_h^2 S_h^2}{c_h}} \tag{3.9}$$

και διαιρώντας την (3.8) με την (3.9) κατά μέλη, παίρνουμε:

$$\frac{n_h}{n} = \frac{\frac{W_h S_h}{\sqrt{c_h}}}{\sum_{h=1}^{L} \frac{W_h S_h}{\sqrt{c_h}}} \quad \text{ή} \quad \frac{n_h}{n} = \frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum_{h=1}^{L} \frac{N_h S_h}{\sqrt{c_h}}}$$

απ' όπου τελικά λαμβάνουμε:

$$n_h = n \frac{N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} N_h S_h / \sqrt{c_h}} \tag{3.10}$$

Η τελευταία σχέση ολοκληρώνει την απόδειξη της Πρότασης μιας, και ο όρος $\frac{n}{\sum_{h=1}^{L} \frac{N_h S_h}{\sqrt{c_h}}}$ είναι σταθερός και εκφράζει τη σταθερά αναλογίας ■

Από τη σχέση (3.8), είναι φανερό ότι για τον βέλτιστο καταμερισμό λαμβάνεται τόσο πιο μεγάλο μέγεθος δείγματος στο στρώμα $n_h$,

(i) όσο πιο μεγάλο είναι το στρώμα $h$

(ii) όσο πιο μεγάλη είναι η τυπική διασπορά $S_h$ του στρώματος και

(iii) όσο πιο μικρός είναι ο συντελεστής κόστους για το στρώμα $c_h$.

Αντικαθιστώντας τις τιμές των $n_h$ που δίνονται από την (3.10) στη συνάρτηση κόστους και λύνοντας ως προς το $n$, προκύπτει ότι:

$$n = \frac{(C - c_0)\sum_{h=1}^{L} N_h S_h/\sqrt{c_h}}{\sum_{h=1}^{L} N_h S_h \sqrt{c_h}}$$

και, αντικαθιστώντας στη συνέχεια την τιμή αυτή του $n$ στις (3.10), τα $n_h$ προσδιορίζονται ακριβώς. Συγκεκριμένα, θα είναι:

$$n_h = \frac{(C - c_0)\, N_h S_h/\sqrt{c_h}}{\sum_{h=1}^{L} N_h S_h \sqrt{c_h}} \tag{3.11}$$

Συνοψίζοντας, για την τυχαία στρωματοποιημένη με σταθερό κόστος $C$ γραμμικής μορφής, η βέλτιστη κατανομή του δείγματος ανά στρώμα δίνεται από την (3.11).

Ανάλογα μπορεί κάποιος να εργαστεί και για το συμμετρικό πρόβλημα ελαχιστοποίησης. Δηλαδή, σταθεροποιώντας τη διασπορά του εκτιμητή $\mathrm{Var}(\bar{X}_{st}) = V$, να ενδιαφέρεται για τον βέλτιστο καταμερισμό του δείγματος, ελαχιστοποιώντας τη συνάρτηση του κόστους. Αποδεικνύεται ότι στην περίπτωση που η συνάρτηση του κόστους είναι γραμμική, ισχύει το ίδιο αποτέλεσμα με την Πρόταση 3.7, δηλ. τα $n_h$ είναι ανάλογα των ποσοτήτων $N_h S_h/\sqrt{c_h}$. Το συνολικό μέγεθος δείγματος $n$ για προκαθορισμένη διασπορά $V$, είναι:

$$n = \frac{\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right)\left(\sum_{h=1}^{L} W_h S_h/\sqrt{c_h}\right)}{V + (1/N)\sum_{h=1}^{L} W_h S_h^2}$$

Ο βέλτιστος καταμερισμός επιτυγχάνει την ελάχιστη διακύμανση για σταθερό κόστος και το ελάχιστο κόστος για σταθερή διακύμανση. Στην πρώτη περίπτωση, η τιμή της **ελάχιστης διακύμανσης** προκύπτει αν στον τύπο της διακύμανσης $\mathrm{Var}(\bar{X}_{st})$ αντικατασταθούν οι τιμές των $n_h$ από την (3.10). Η ελάχιστη λοιπόν διακύμανση του εκτιμητή θα είναι:

$$\mathrm{Var}(\bar{X}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h/\sqrt{c_h}\right)\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right) - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2 \tag{3.12}$$

Ο βέλτιστος καταμερισμός προκύπτει με βελτιστοποίηση ιδιοτήτων του εκτιμητή (είτε του κόστους της έρευνας) και ανήκει στη δεύτερη γενική κατηγορία καταμερισμών, όπου η βελτιστοποίηση, παρά η ευκολία, είναι ο κανόνας που καθορίζει την επιλογή των $n_h$ ανά στρώμα.

### 3.5.3 Καταμερισμός Neyman (Neyman Allocation)

Μια ειδική περίπτωση για τον βέλτιστο καταμερισμό του $n$, έχουμε όταν ο συντελεστής του κόστους δεν διαφέρει ανά στρώμα, δηλ. $c_h = c$, $(h = 1,2,\ldots,L)$. Η κατανομή αυτή ονομάζεται Neyman και τα αποτελέσματα που ισχύουν για τον καταμερισμό και τη βέλτιστη τιμή της διακύμανσης που επιτυγχάνεται από τον εκτιμητή μπορούν να εξαχθούν από τα γενικότερα αποτελέσματα του βέλτιστου καταμερισμού.

**Πόρισμα 3.4**

Για τον καταμερισμό Neyman, τα μεγέθη του δείγματος ανά στρώμα δίνονται από τη σχέση:

$$n_h = n\frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} \tag{3.13}$$

**Πόρισμα 3.5**

Για τον καταμερισμό Neyman, η τιμή της ελάχιστης διακύμανσης του εκτιμητή $\bar{X}_{st}$ δίνεται από τη σχέση:

$$\mathrm{Var}(\bar{X}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2 \tag{3.14}$$

## Παρατήρηση 3.5

Εάν, επιπλέον της υπόθεσης της ισότητας των συντελεστών κόστους ανά στρώμα, ισχύει και η ισότητα μεταξύ των τυπικών αποκλίσεων $S_h$ ανά στρώμα, τότε η (3.13) δίνει $n_h = \frac{nN_h}{N}$, δηλ. ο καταμερισμός Neyman είναι ισοδύναμος με τον αναλογικό. Κατά συνέπεια, κάτω από τις υποθέσεις αυτές, ο αναλογικός καταμερισμός, που χρησιμοποιείται αρκετά συχνά στην πράξη λόγω ευκολίας, συμπίπτει με τον βέλτιστο και άρα επιτυγχάνει βέλτιστη αποτελεσματικότητα.

## 3.6. Καθορισμός του ελάχιστου απαιτούμενου μεγέθους δείγματος για τη στρωματοποιημένη δειγματοληψία

Έστω ότι το συνολικό μέγεθος δείγματος $n$ δεν είναι γνωστό, αλλά προσδιορίζεται βάσει κάποιων προδιαγραφών που θα πρέπει να ισχύουν στα αποτελέσματα της έρευνας. Ανάλογα με την περίπτωση της απλής τυχαίας δειγματοληψίας, που εξετάστηκε στο Κεφάλαιο 2 -, οι προδιαγραφές αυτές μπορεί να είναι μια ανώτατη επιτρεπτή τιμή για τη διακύμανση του εκτιμητή που θα παραχθεί ή η εκτίμηση να είναι εντός ενός επιτρεπτού ορίου σε σχέση με την άγνωστη παράμετρο.

Το πρόβλημα αντιμετωπίζεται ανάλογα όπως και στο Κεφάλαιο 2 -, κάνοντας χρήση των κατάλληλων εκφράσεων για τη διακύμανση του εκτιμητή κάτω από ένα στρωματοποιημένο σχήμα, αντί εκείνων από ένα απλό τυχαίο. Εάν, για παράδειγμα, ο περιορισμός είναι η εκτίμηση του μέσου $\bar{Y}$ από τον $\bar{X}_{st}$ να έχει μια μέγιστη επιτρεπτή απόσταση $d$ και πιθανότητα σφάλματος $a$, τότε, υποθέτοντας κανονική κατανομή για τον εκτιμητή $\bar{X}_{st}$, η έκφραση (2.10) στην περίπτωση της στρωματοποιημένης γίνεται:

$$\frac{d}{\sqrt{\mathrm{Var}(\bar{X}_{st})}} = z_{\alpha/2}$$

Στην τελευταία σχέση, η έκφραση της διακύμανσης $\mathrm{Var}(\bar{X}_{st})$ στον παρονομαστή αντικαθιστάται από την ειδική έκφραση της στρωματοποιημένης που πρόκειται να εφαρμοστεί και η εξίσωση λύνεται στη συνέχεια ως προς $n$.

Για τον καταμερισμό Neyman, αντικαθιστώντας την $\mathrm{Var}(\bar{X}_{st})$ από την (3.14) και ακολουθώντας την παραπάνω διαδικασία προκύπτει:

$$n = \frac{\left(\sum_{h=1}^{L} W_h S_h\right)^2}{\frac{1}{N}\sum_{h=1}^{L} W_h S_h^2 + \left(\frac{d}{z_{a/2}}\right)^2}$$

Μετά τον προσδιορισμό του συνολικού $n$, το μέγεθος αυτό κατανέμεται στα στρώματα βάσει του καταμερισμού Neyman.

## Παράδειγμα 3.9

Επιστρέφουμε στο Παράδειγμα 2.11, όπου έγινε ο προσδιορισμός του μεγέθους του δείγματος $n$ για εφαρμογή στην α.τ.δ. στο σύνολο των καταστημάτων της περιοχής, με σκοπό την εκτίμηση του συνόλου των Χριστουγεννιάτικων καρτών που θα πουληθούν. Έστω ότι τα καταστήματα χωρίζονται σε τρεις κατηγορίες με βάση τον ετήσιο τζίρο τους, και ο ερευνητής ενδιαφέρεται να προσδιορίσει το μέγεθος δείγματος που απαιτείται προκειμένου να υλοποιήσει μια τυχαία στρωματοποιημένη δειγματοληψία. Για λόγους σύγκρισης, έστω ότι επιθυμούμε ο εκτιμητής που θα προκύψει να έχει τις ίδιες ιδιότητες με εκείνες που τέθηκαν στο Παράδειγμα 2.11. Δηλαδή, η εκτίμηση του συνολικού αριθμού καρτών να μη διαφέρει περισσότερο από 10% από την αληθινή τιμή, με πιθανότητα σφάλματος 5%.

Τα στοιχεία που αφορούν τα στρώματα του πληθυσμού δίνονται στον Πίνακα 3.8.

| **Ετήσιος τζίρος (σε χιλ. Χρημ. Μονάδες)** | $N_h$ | $S_h^2$ |
|---|---|---|
| **≤ 50** | 146 | 0.16 |
| **Μεταξύ 50 και 100** | 62 | 0.58 |
| **≥ 100** | 35 | 0.31 |

***Πίνακας 3.8****Στρωματοποίηση καταστημάτων βάσει ετήσιου τζίρου.*

Εάν ο ερευνητής ενδιαφέρεται να υλοποιήσει τον αναλογικό καταμερισμό, τότε το απαιτούμενο μέγεθος δείγματος συνολικά για το δείγμα υπολογίζεται λύνοντας την εξίσωση:

$$\frac{d}{\sqrt{\mathrm{Var}_{prop}\,(N\bar{X}_{st})}} = z_{\alpha/2}$$

ως προς $n$, το οποίο περιέχεται στον εκτιμητή του τυπικού σφάλματος στον παρονομαστή. Αντικαθιστώντας τις προδιαγραφές που δίνονται για την έρευνα, δηλ. $d = 0.1 * 420$, $a = 0.05$ και τον τύπο (3.6) για τη διασπορά του εκτιμητή στον παρονομαστή, που είναι σε συμφωνία με το δειγματοληπτικό σχήμα που πρόκειται να εφαρμοστεί, προκύπτει

$$\frac{42}{N\sqrt{\left(\frac{1}{n}-\frac{1}{N}\right)\sum_{i=1}^{3} W_i S_i^2}} = 1.96$$

ή

$$\left(\frac{1}{n}-\frac{1}{420}\right)\left(\frac{146}{243}\times 0.16+\frac{62}{243}\times 0.58+\frac{35}{243}0.31\right)=\left(\frac{42}{243\times 1.96}\right)^2$$

απ' όπου με πράξεις προκύπτει $n = 32.21$. Συνεπώς, αν εφαρμοστεί η δεδομένη στρωματοποίηση για την έρευνα, θα χρειαστεί να επιλεγούν συνολικά 33 καταστήματα. Μοιράζοντας αναλογικά τα 33 ανά στρώμα, προκύπτει:

$$n_1 = n\frac{N_1}{N} = 33\frac{146}{243} = 19.82,\ n_2 = n\frac{N_2}{N} = 8.42\ \ και\ \ n_3 = n\frac{N_3}{N} = 4.75$$

ή $n_1 = 20,\ \ n_2 = 8\ \ και\ \ n_3 = 5.$

Αξίζει να σημειωθεί ότι για απλό τυχαίο δείγμα, το αντίστοιχο μέγεθος που απαιτούνταν για ίδιες προδιαγραφές ήταν 62. Η μείωση είναι αρκετά μεγάλη, γεγονός που δηλώνει ότι η στρωματοποίηση του Πίνακα 3.8 ήταν αρκετά ικανοποιητική, δηλ. η διακύμανση στο κάθε στρώμα που προκύπτει είναι αρκετά μικρότερη από εκείνη του συνολικού πληθυσμού.

## 3.7. Σύγκριση εκτίμησης μέσου από την απλή τυχαία και τη στρωματοποιημένη δειγματοληψία

Η σύγκριση θα αναπτυχθεί για την περίπτωση της τυχαίας στρωματοποιημένης δειγματοληψίας. Οι δύο εκτιμητές $\bar{X}$ και $\bar{X}_{st}$ που προκύπτουν από ένα απλό τυχαίο δείγμα και ένα στρωματοποιημένο είναι αμερόληπτοι, οπότε η σύγκριση ως προς την ακρίβεια μεταξύ των δύο εκτιμητών θα πραγματοποιηθεί με κριτήριο την ελάχιστη διακύμανση. Για τη σύγκριση, υποθέτουμε ίσο μέγεθος δείγματος και στις δύο περιπτώσεις.

Θα χρειαστούμε ένα πρώτο αποτέλεσμα, αυτό της διάσπασης (ανάλυσης) της συνολικής αληθινής διασποράς $S^2$ σε δύο επιμέρους αθροίσματα τετραγώνων, όπως δίνεται στην Πρόταση που ακολουθεί. Η ανάλυση διασποράς είναι ο τύπος που συνδέει τις δύο διαφορετικές καταστάσεις του πληθυσμού, δηλ. ενιαίος ή χωρισμένος σε στρώματα. Βάση του τύπου αυτού, είναι εφικτή η σύγκριση των εκτιμητών που προκύπτουν από τις δύο διαφορετικές δειγματοληψίες.

### Πρόταση 3.8

Ο τύπος ανάλυσης διακύμανσης (ΑΝΑΔΙΑ) για τη διακύμανση $S^2$ ενός πληθυσμού ο οποίος έχει χωριστεί σε $L$ στρώματα είναι:

$$(N-1)S^2 = \sum_{h=1}^{L}(N_h - 1)\,S_h^2 + \sum_{h=1}^{L} N_h(\bar{Y}_h - \bar{Y})^2 \qquad (3.15)$$

### Απόδειξη

Από τον ορισμό της διασποράς $S^2$ ισχύει:

$$(N-1)S^2 = \sum_{h=1}^{L}\sum_{j=1}^{N_h}\left(Y_{hj} - \bar{Y}\right)^2.$$

Προσθαφαιρώντας τη μέση τιμή ανά στρώμα $\bar{Y}_h$ σε κάθε όρο του αθροίσματος και αναπτύσσοντας τη γνωστή ταυτότητα, προκύπτει:

$$(N-1)S^2 = \sum_{h=1}^{L}\sum_{j=1}^{N_h}\left(Y_{hj} - \bar{Y}_h + \bar{Y}_h - \bar{Y}\right)^2$$

$$= \sum_{h=1}^{L}\sum_{j=1}^{N_h}\left(Y_{hj} - \bar{Y}_h\right)^2 + \sum_{h=1}^{L}\sum_{j=1}^{N_h}(\bar{Y}_h - \bar{Y})^2 + 2\sum_{h=1}^{L}\sum_{j=1}^{N_h}\left(Y_{hj} - \bar{Y}_h\right)(\bar{Y}_h - \bar{Y})$$

$$= \sum_{h=1}^{L}(N_h - 1)\,S_h^2 + \sum_{h=1}^{L} N_h(\bar{Y}_h - \bar{Y})^2 + 2\sum_{h=1}^{L}(\bar{Y}_h - \bar{Y})\sum_{j=1}^{N_h}\left(Y_{hj} - \bar{Y}_h\right)$$

όπου χρησιμοποιήσαμε τον ορισμό των $S_h^2$ και κάναμε πράξεις στα αθροίσματα. Περαιτέρω, το τελευταίο άθροισμα είναι μηδέν (προφανές από τον ορισμό της μέσης τιμής ανά στρώμα), άρα συνολικά:

$$(N-1)S^2 = \sum_{h-}^{L}(N_h - 1)\,S_h^2 + \sum_{h=1}^{L} N_h(\bar{Y}_h - \bar{Y})^2$$

που συμπληρώνει την απόδειξη ■

Σύμφωνα με το αποτέλεσμα, που είναι γνωστό και σε άλλες στατιστικές εφαρμογές, η συνολική διασπορά του πληθυσμού αναπτύσσεται σε δύο αθροίσματα. Το πρώτο άθροισμα δίνει την πληροφορία για τη διασπορά των μονάδων στο εσωτερικό κάθε στρώματος (ονομάζεται within διασπορά) και το δεύτερο άθροισμα δίνει την πληροφορία πόσο μακριά είναι οι μέσες τιμές των στρωμάτων μεταξύ τους (between). Οι μέσες τιμές $\bar{Y}_h$ συγκρίνονται όλες με τη σταθερή, ανεξάρτητη της στρωματοποίησης, μέση τιμή $\bar{Y}$.

Διαιρώντας τη σχέση (3.15) με $N$ και στα δύο μέλη και θεωρώντας ότι οι όροι $1/N_h$ είναι αμελητέοι, προκύπτει ο πιο απλός τύπος:

$$S^2 = \sum_{h=1}^{L} W_h\,S_h^2 + \sum_{h=1}^{L} W_h(\bar{Y}_h - \bar{Y})^2 \qquad (3.16)$$

Ένα μέτρο της σχετικής αποτελεσματικότητας ενός δειγματοληπτικού σχεδίου σε σύγκριση με την απλή τυχαία είναι η ***επίδραση του δειγματοληπτικού σχεδίου (design effect)***. Η επίδραση ενός δειγματοληπτικού σχεδίου p συμβολίζεται **deff** και ορίζεται ως το κλάσμα της διακύμανσης του εκτιμητή όταν η δειγματοληψία γίνει σύμφωνα με το δειγματοληπτικό σχέδιο p, προς τη διακύμανση του εκτιμητή εάν είχε πραγματοποιηθεί α.τ.δ. με ίσο μέγεθος δείγματος. Αναλυτικά:

$$\text{deff} = \frac{\text{Var}_p(\hat{\theta})}{\text{Var}_{srs}(\hat{\theta})}$$

Το δειγματοληπτικό σχέδιο p στον αριθμητή είναι το σχέδιο που συγκρίνουμε με την srs και είναι συνήθως ένα πιο πολύπλοκο σχέδιο. Η αριθμητική τιμή του deff δίνει ένα μέτρο της απόστασης της ακρίβειας που επιτυγχάνεται από τον εκτιμητή υιοθετώντας το δειγματοληπτικό σχέδιο $p$, ως προς την αντίστοιχη υιοθετώντας την α.τ.δ. Ανάλογα με το αριθμητικό αποτέλεσμα του μέτρου deff προκύπτει η ερμηνεία του. Για παράδειγμα εάν deff=1, τότε $\text{Var}_p(\hat{\theta}) = \text{Var}_{srs}(\hat{\theta})$ και κατά συνέπεια, το δειγματοληπτικό σχέδιο p είναι ισοδύναμο με την α.τ.δ. ως προς την ακρίβεια. Αν deff<1, τότε το δειγματοληπτικό σχέδιο p έχει θετική επίδραση πάνω στο α.τ.δ. γιατί η διακύμανση του εκτιμητή είναι μικρότερη, άρα επιτυγχάνει μεγαλύτερη ακρίβεια σε σχέση με την α.τ.δ. Όσο μικρότερο της μονάδας είναι το deff τόσο μεγαλύτερη είναι η θετική επίδραση του p. Π.χ. αν deff=0.85, το σχέδιο p έχει θετική επίδραση στο α.τ. και ειδικότερα η διακύμανση του εκτιμητή θα μειωθεί κατά 15% εάν υιοθετήσουμε το δειγματοληπτικό σχέδιο p έναντι της α.τ.δ. Ανάλογα, εάν deff>1 η επίδραση του δειγματοληπτικού σχεδίου p είναι αρνητική, δηλαδή η διακύμανση του εκτιμητή σύμφωνα με τη δειγματοληψία p αυξάνεται σε σχέση με την αντίστοιχη που θα είχαμε για ένα α.τ.δ. ίσου μεγέθους. Όσο μεγαλύτερη της μονάδας είναι η τιμή του deff, τόσο μεγαλύτερη είναι η αρνητική επίδραση του p στο α.τ.δ. Π.χ. εάν για ένα δειγματοληπτικό σχέδιο $p_1$ είναι deff=1.7 και για ένα δεύτερο σχέδιο $p_2$ είναι deff=2.5, τότε αρχικά συμπεραίνουμε ότι και τα δύο σχέδια $p_1$ και $p_2$ είναι λιγότερο αποτελεσματικά από το α.τ.δ. Επιπλέον, μεταξύ των δύο σχεδίων $p_1$ και $p_2$ καλύτερο είναι το $p_1$ γιατί η διακύμανση του εκτιμητή αυξάνεται κατά 70% σε σχέση με εκείνη της α.τ.δ., ενώ η διακύμανση του εκτιμητή κάτω από το $p_2$ αυξάνεται κατά 150%.

Επίσης, το deff χρησιμοποιείται όταν ο ερευνητής θέλει να επιβεβαιώσει εάν το δείγμα των μετρήσεων που διαθέτει έχει συλλεχθεί σύμφωνα με την α.τ.δ. Εάν η τιμή του deff είναι μακρυά από το 1 τότε ο τρόπος δειγματοληψίας που χρησιμοποιήθηκε για τη συλλογή του δείγματος δεν είναι η α.τ.δ.

## Παράδειγμα 3.10

Για τα δεδομένα του Παραδείγματος 3.6, η συνολική διασπορά των μετρήσεων στον πληθυσμό, δηλ. των χρόνων που απαιτούνται για τη διεκπεραίωση των φακέλων ανεξαρτήτως κατηγορίας στην οποία ανήκουν, υπολογίζεται από τον τύπο (3.16) και θα είναι:

$$S^2 = \sum_{h\ 1}^{2} W_h\, S_h^2 + \sum_{h=1}^{2} W_h(\bar{Y}_h - \bar{Y})^2$$

Επειδή τα στοιχεία $S_h^2, \bar{Y}_h$ και $\bar{Y}$ δεν είναι γνωστά για τον πληθυσμό, αλλά εκτιμώνται μέσω του δείγματος, ο εκτιμητής του $S^2$ θα δίνεται τελικά από τον τύπο:

$$\hat{S}^2 = W_1\hat{S}_1^2 + W_2\hat{S}_2^2 + W_1\left(\hat{\bar{Y}}_1 - \hat{\bar{Y}}\right)^2 + W_2\left(\hat{\bar{Y}}_2 - \hat{\bar{Y}}\right)^2$$

$$= 0.65 \times 0.7^2 + 0.35 \times 2.4^2 + 0.65(1.8 - 3.795)^2 + 0.35(7.5 - 3.795)^2$$

$$= 9.73.$$

Άρα, ο πληθυσμός των χρόνων επεξεργασίας των φακέλων, εάν θεωρηθεί ως ενιαίο σύνολο, έχει (εκτιμώμενη) διασπορά ίση με 9.73. Αξίζει να σημειωθεί ότι οι διασπορές των δύο στρωμάτων που προκύπτουν εάν γίνει κατηγοριοποίηση των φακέλων με βάση τον βαθμό δυσκολίας είναι $0.7^2 = 0.49$ και $2.4^2 = 5.76$ αντίστοιχα.

Η διακύμανση των δύο στρωμάτων είναι αρκετά μικρότερη από την αρχική, γεγονός που οδηγεί σ' έναν εκτιμητή με μικρότερο τυπικό σφάλμα. Πράγματι, όπως έχει δειχθεί στο Παράδειγμα 3.6, ο εκτιμητής που κατασκευάζεται με τη βοήθεια ενός στρωματοποιημένου δείγματος μεγέθους $n = 100$ έχει τυπικό σφάλμα 0.148 ώρες, ενώ ο εκτιμητής από ένα απλό τυχαίο δείγμα ίδιου μεγέθους από τον πληθυσμό θα έχει εκτιμώμενο τυπικό σφάλμα:

$$\hat{\mathrm{Var}}(\bar{X}) = \frac{1-f}{n}S^2 = \frac{1 - 100/1500}{100}9.73 = 0.091$$

ή $\hat{\mathrm{se}}(\bar{X}) = 0.30$ ώρες, αρκετά μεγαλύτερο (διπλάσιο) από το αντίστοιχο της στρωματοποιημένης.

Ένας τρόπος να συγκριθεί η ακρίβεια των δύο εκτιμητών είναι με τη βοήθεια της επίδρασης, deff, του στρωματοποιημένου δειγματοληπτικού σχεδίου πάνω στο απλό τυχαίο. Για τις τιμές του παραδείγματος, θα είναι:

$$\mathrm{deff} = \frac{\mathrm{Var}_{prop}\,(\bar{X}_{st})}{\mathrm{Var}\,(\bar{X})} = \frac{0.0218}{0.091} = 0.24$$

δηλ. η μεταβλητότητα του εκτιμητή ενός δείγματος που επιλέγεται με την απλή τυχαία βελτιώνεται κατά 76% εάν υιοθετηθεί αντί το απλό τυχαίο το αναλογικό στρωματοποιημένο σχήμα.

Στη συνέχεια, για τη σύγκριση της ακρίβειας των εκτιμητών της μέσης τιμής από ένα απλό τυχαίο δείγμα και από ένα στρωματοποιημένο ίσου μεγέθους δείγμα, αποδεικνύεται το ακόλουθο αρκετά σημαντικό και πολύ γενικό αποτέλεσμα.

## Πρόταση 3.9

Για τις διακυμάνσεις των εκτιμητών $\bar{X}$ και $\bar{X}_{st}$ που προκύπτουν από το απλό τυχαίο δείγμα και ένα στρωματοποιημένο αναλογικό και βέλτιστο, και αν οι όροι $1/N_h$, $1/N$ θεωρηθούν αμελητέοι, τότε ισχύει:

$$\mathrm{Var}_{opt}\,(\bar{X}_{st}) \le \mathrm{Var}_{prop}\,(\bar{X}_{st}) \le \mathrm{Var}(\bar{X})$$

όπου ο δείκτης $opt$ δηλώνει τον βέλτιστο καταμερισμό και ο δείκτης $prop$ τον αναλογικό.

**Απόδειξη**

Αν πολλαπλασιάσουμε και τα δύο μέλη της σχέσης (3.16) που προκύπτει από τον τύπο ΑΝΑΔΙΑ με τον όρο $\frac{1-f}{n}$, λαμβάνουμε:

$$\frac{1-f}{n}S^2 = \frac{1-f}{n}\sum_{h=1}^{L} W_h\, S_h^2 + \frac{1-f}{n}\sum_{h=1}^{L} W_h(\bar{Y}_h - \bar{Y})^2$$

ή ισοδύναμα:

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}_{prop}\,(\bar{X}_{st}) + \frac{1-f}{n}\sum_{h=1}^{L} W_h(\bar{Y}_h - \bar{Y})^2 \qquad (3.17)$$

Η τελευταία σχέση αποδεικνύει τη δεξιά ανισότητα της Πρότασης, δηλ. την:

$$\mathrm{Var}_{prop}\,(\bar{X}_{st}) \le \mathrm{Var}(\bar{X}),$$

δεδομένου ότι το τελευταίο άθροισμα είναι ένα άθροισμα τετραγώνων και συνεπώς είναι πάντα θετικό.

Για την απόδειξη της ανισότητας που συγκρίνει τα δύο στρωματοποιημένα σχήματα ως προς την ακρίβεια του εκτιμητή που θα προσφέρουν, θεωρούμε τη διαφορά των δύο διακυμάνσεων:

$$\mathrm{Var}_{prop}\,(\bar{X}_{st}) - \mathrm{Var}_{opt}\,(\bar{X}_{st})$$

και, αντικαθιστώντας τις αναλυτικές τους εκφράσεις, προκύπτει:

$$\mathrm{Var}_{prop}\,(\bar{X}_{st}) - \mathrm{Var}_{opt}\,(\bar{X}_{st}) = \frac{1}{n}\sum_{h=1}^{L} W_h(S_h - \bar{S})^2 \qquad (3.18)$$

όπου $\bar{S} = \sum_h W_h S_h$. Επειδή και πάλι το άθροισμα της διαφοράς είναι σταθερού προσήμου, και ειδικότερα πάντα θετικό, αποδεικνύεται η ισχύς της ζητούμενης ανισότητας ■

Η Πρόταση 3.9 είναι πολύ σημαντική, γιατί δίνει τη σύγκριση της απλής τυχαίας με τη στρωματοποιημένη δειγματοληψία, αλλά επίσης γιατί παρέχει χρήσιμες εκφράσεις, και πιο συγκεκριμένα τις (3.17) και (3.18), που δίνουν ένα μέτρο της διαφοράς των δύο πιο γνωστών τυχαίων στρωματοποιημένων σχεδίων και της απλής τυχαίας. Οι παρατηρήσεις που ακολουθούν συνοψίζουν τα αποτελέσματα που εξάγονται από την Πρόταση 3.9.

## Παρατηρήσεις

**3.6** Το βασικό αποτέλεσμα της Πρότασης είναι ότι, ανεξαρτήτως της δομής του πληθυσμού (της μεταβλητότητας $S^2$ για το χαρακτηριστικό), η στρωματοποιημένη δειγματοληψία βελτιώνει πάντα την απλή τυχαία. Το αποτέλεσμα, με άλλα λόγια, ισχύει πάντα και όχι ανά περίπτωση ή υπό συνθήκη. Η μόνη προϋπόθεση είναι τα μεγέθη $N_h$ να είναι αρκετά μεγάλα, ώστε τα $1/N_h$ να θεωρούνται αμελητέα.

**3.7** Μια στρωματοποίηση βελτιώνει πάντα την ακρίβεια της εκτίμησης και, δεδομένου ότι υπάρχουν πολλοί τρόποι χωρισμού του πληθυσμού σε στρώματα, αξίζει να αναζητηθεί ποιο κριτήριο χωρισμού και ποιος ορισμός στρωμάτων θα επιτύχει μεγαλύτερη αποτελεσματικότητα.

**3.8** Η έκφραση (3.17) συνεπάγεται ότι η ποσότητα $d_1 = \frac{1-f}{n}\sum_{h=1}^{L} W_h(\bar{Y}_h - \bar{Y})^2$ είναι η διαφορά των διακυμάνσεων κάτω από την απλή τυχαία και από την αναλογική στρωματοποιημένη. Η πρώτη διακύμανση δεν εξαρτάται από στρώματα, αλλά από τον ενιαίο πληθυσμό, και είναι σταθερή. Η διακύμανση $\mathrm{Var}_{prop}\,(\bar{X}_{st})$ θα είναι τόσο μικρότερη της $\mathrm{Var}(\bar{X})$ από την απλή τυχαία, όσο μεγαλύτερη

είναι η διαφορά $d_1$. Η διαφορά αυτή με τη σειρά της γίνεται μέγιστη, όταν οι τετραγωνικές αποκλίσεις $\bar{Y}_h - \bar{Y}$ γίνονται μέγιστες. Συνεπώς, όσο περισσότερο διαφέρουν οι μέσες τιμές των στρωμάτων μεταξύ τους, τόσο καλύτερη απόδοση αναμένουμε από τη στρωματοποιημένη δειγματοληψία.

**3.9** Η έκφραση (3.18) συγκρίνει την ακρίβεια των εκτιμητών κάτω από τον ίδιο ορισμό στρωμάτων και διαφορετικό καταμερισμό δείγματος. Ανάλογα με την περίπτωση αυτή, η ποσότητα $d_2 = \frac{1}{n}\sum_{h=1}^{L} W_h (S_h - \bar{S})^2$ είναι το μέτρο της διαφοράς των δύο διακυμάνσεων. Αν $d_2 = 0$, δηλ. τα $S_h$ είναι ίσα για κάθε στρώμα, οι δύο καταμερισμοί είναι ισοδύναμοι και, όσο μεγαλύτερη είναι η απόκλιση των $S_h$ από τη μέση τους τιμή, τόσο περισσότερο υπερτερεί ο βέλτιστος καταμερισμός σε σχέση με τον αναλογικό.

Γενικότερα θέματα, όπως, επιλογή βοηθητικού χαρακτηριστικού για την κατασκευή των στρωμάτων, πλήθος στρωμάτων και καθορισμός ορίων στρωμάτων, θα απασχολήσουν την ανάπτυξη του επόμενου κεφαλαίου. Το συμπέρασμα που μέχρι το σημείο αυτό είναι σημαντικό να καταγραφεί είναι ότι η στρωματοποιημένη δειγματοληψία είναι μια ευρέως διαδεδομένη δειγματοληπτική μέθοδος και, ενώ βελτιώνει πάντα την απλή τυχαία, η προσπάθεια του ερευνητή επικεντρώνεται στην όσο το δυνατό βέλτιστη εφαρμογή της στρωματοποίησης, ώστε το όφελος να είναι μέγιστο.

## Παράδειγμα 3.11

Μια έρευνα διεξάγεται με σκοπό να εκτιμηθεί το ποσοστό ανεργίας των κατοίκων $P$ μιας πόλης. Από προηγούμενες έρευνες, είναι γνωστό ότι το ποσοστό ανεργίας είναι μεγαλύτερο για τις μικρότερες ηλικίες. Για το λόγο αυτό, πριν από τη δειγματοληψία, ο πληθυσμός χωρίζεται σε τρία στρώματα, όπως παρουσιάζεται στον Πίνακα 3.9. Τα βάρη του πληθυσμού για τα ηλικιακά στρώματα είναι γνωστά και δίνονται στη $2^{η}$ στήλη του ίδιου πίνακα. Με βάση τα στοιχεία για τον πληθυσμό, επιλέγεται ένα αναλογικό στρωματοποιημένο δείγμα μεγέθους $n = 300$. Τα αποτελέσματα για τους εκτιμητές των ποσοστών δίνονται επίσης στον Πίνακα 3.9.

| **Ηλικία** | $W_h$ | $\hat{P}_h$ |
|---|---|---|
| **15 έως 24** | 0.25 | 0.47 |
| **24 έως 40** | 0.35 | 0.30 |
| **40 και άνω** | 0.40 | 0.20 |

***Πίνακας 3.9*** *Στρωματοποίηση πληθυσμού με βάση την ηλικία.*

Η διακύμανση του εκτιμητή του ποσοστού ανεργίας με βάση το στρωματοποιημένο δείγμα εκτιμάται από τον τύπο:

$$\hat{\mathrm{V}}\mathrm{ar}(\hat{P}_{st}) = \frac{1-f}{n}\sum_{h=1}^{L} W_h \frac{n_h \hat{P}_h (1-\hat{P}_h)}{n_h - 1}$$

Θεωρώντας το πηλίκο του δείγματος ($f$) αμελητέο και αντικαθιστώντας τα στοιχεία από τον πίνακα, προκύπτει:

$$\hat{\mathrm{V}}\mathrm{ar}(\hat{P}_{st}) = \frac{1}{300}\left(0.25\frac{75 \times 0.47 \times 0.53}{74} + 0.35\frac{105 \times 0.3 \times 0.7}{104} + 0.4\frac{120 \times 0.2 \times 0.8}{119}\right)$$

$$= 0.000673$$

ή, το εκτιμώμενο τυπικό σφάλμα της $\hat{\mathrm{s}}\mathrm{e}(\hat{P}_{st}) = 0.026$.

Αν η δειγματοληψία γινόταν με απλή τυχαία αντί στρωματοποιημένη για σταθερό μέγεθος δείγματος, $n = 300$, η διακύμανση του εκτιμητή, σύμφωνα με την έκφραση (3.17) θα ήταν:

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}_{prop}(\bar{X}_{st}) + \frac{1-f}{n}\sum_{h=1}^{L} W_h(\bar{Y}_h - \bar{Y})^2$$

$$= 0.000673 + \frac{1}{300}\sum_{h=1}^{L} W_h(P_h - P)^2$$

όπου $P$ ο συνολικός εκτιμητής του δείγματος, δηλ. για ολόκληρο τον πληθυσμό. Από τα στοιχεία της έρευνας, προκύπτει ότι $P = \sum_h W_h P_h = 0.3025$ και, κατά συνέπεια,

$$\mathrm{Var}(\bar{X}) = 0.000673 + \frac{1}{300}(W_1(P_1 - P)^2 + W_2(P_2 - P)^2 + W_3(P_3 - P)^2)$$

$$= 0.00071$$

Η επίδραση του στρωματοποιημένου πάνω στο απλό τυχαίο δείγμα ίδιου μεγέθους είναι:

$$\mathrm{deff} = \frac{\mathrm{Var}_{prop}(\hat{P}_{st})}{\mathrm{Var}(\hat{P})} = \frac{0.000673}{0.00071} = 0.95$$

δηλ. μείωση της διακύμανσης του εκτιμητή κατά 5%.

Η διακύμανση του εκτιμητή που θα προέκυπτε αν η στρωματοποιημένη δειγματοληψία γινόταν με τον βέλτιστο καταμερισμό, μπορεί να υπολογιστεί με τη βοήθεια του τύπου (3.18) και είναι:

$$\mathrm{Var}_{prop}(\hat{P}_{st}) - \mathrm{Var}_{opt}(\hat{P}_{st}) = \frac{1}{n}\sum_{h=1}^{L} W_h(S_h - \bar{S})^2$$

όπου $\bar{S} = \sum_h W_h S_h = 0.45$ και $S_h^2 = \frac{n_h P_h(1-P_h)}{n_h - 1}$

και τελικά:

$$\mathrm{Var}_{opt}(\hat{P}_{st}) = \mathrm{Var}_{prop}(\hat{P}_{st}) - \frac{1}{n}\sum_{h=1}^{L} W_h(S_h - \bar{S})^2 = 0.000667.$$

# Βιβλιογραφικές Αναφορές


Barnett, V. (2002). *Sample survey: Principles and methods* (3rd Edition). London: Arnold.

Cochran, W. G. (1977). *Sampling techniques* (3rd Edition). New York: John Wiley and Sons.

Des Raj (1968). *Sampling theory*. New York: McGraw-Hill.

Kish, L. (1965). *Survey sampling*. New York: John Wiley and Sons.

Levy P.S. and Lemeshow, S. (1999). *Sampling of populations. Methods and applications* (3rd Edition). New York: John Wiley and Sons.

Rao, P.S.R.S. (2000). *Sampling methodologies with applications*. Boca Raton, Fla: Chapman and Hall/CRC.

Thompson, S. K. (2012). S*ampling* (3rd Edition). Hoboken, NJ: John Wiley and Sons.

# Κεφάλαιο 4 - ΣΤΡΩΜΑΤΟΠΟΙΗΜΕΝΗ ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΠΕΡΑΙΤΕΡΩ ΘΕΜΑΤΑ

**Σύνοψη**

*Η στρωματοποιημένη δειγματοληψία εφαρμόζεται πολύ συχνά στην πράξη, λόγω της αποτελεσματικότητας των εκτιμητών που προκύπτουν. Η υλοποίηση της στρωματοποιημένης δειγματοληψίας έχει ως προϋπόθεση την ύπαρξη μιας* **βοηθητικής μεταβλητής (auxiliary variable)**, *έστω Z, που, ιδανικά, να σχετίζεται με την κύρια μεταβλητή Y της έρευνας και να υπάρχουν διαθέσιμα στοιχεία της για τον πληθυσμό. Η βοηθητική μεταβλητή χρησιμεύει στην κατασκευή των στρωμάτων, προκειμένου στη συνέχεια η επιλογή του δείγματος να γίνει σύμφωνα με τη στρωματοποιημένη δειγματοληψία. Προηγούμενες έρευνες ή απογραφές παρέχουν τα απαραίτητα στοιχεία για μία ή περισσότερες βοηθητικές μεταβλητές στον πληθυσμό και πρακτικά είναι αρκετά εύκολη η στρωματοποίηση του πληθυσμού. Η ευκολία στην υλοποίηση, σε συνδυασμό με τις καλές ιδιότητες της μεθόδου, καθιστούν τη στρωματοποιημένη μια αρκετά διαδεδομένη μέθοδο δειγματοληψίας. Για το λόγο αυτό, το παρόν κεφάλαιο θα ασχοληθεί με περαιτέρω θέματα που αφορούν πρακτικά προβλήματα και ανακύπτουν κατά τον σχεδιασμό και την υλοποίηση της στρωματοποιημένης (βλέπε επίσης: Kish (1965, Κεφ. 2 & 3), Cochran (1977, Κεφ. 5) και Levy & Lemeshow (1999, Κεφ. 6). Αναλυτικότερα, τα θέματα της περαιτέρω μελέτης επικεντρώνονται (α) στην επιλογή της βοηθητικής μεταβλητής που θα χρησιμοποιηθεί ως κριτήριο χωρισμού των στρωμάτων στον πληθυσμό, (β) την επιλογή του αριθμού των στρωμάτων και (γ) την κατασκευή των στρωμάτων, ειδικότερα στην περίπτωση που η βοηθητική μεταβλητή είναι συνεχής. Επίσης, γίνεται περιγραφή και μελέτη της* **εκ των υστέρων στρωματοποιημένης δειγματοληψίας (post-stratification sampling)** *και των ιδιοτήτων που τη χαρακτηρίζουν.*

**Προαπαιτούμενη γνώση**

*Κεφάλαιο 1 -, Κεφάλαιο 2 -Κεφάλαιο 3 -Ομαδοποίηση Παρατηρήσεων σε Πίνακα Συχνοτήτων, Εκατοστημόρια.*

## 4.1. Επιλογή βοηθητικής μεταβλητής και αριθμού στρωμάτων στη Στρωματοποιημένη Δειγματοληψία

Ένα από τα πρωταρχικά ερωτήματα που θα απασχολήσουν τον ερευνητή στο στάδιο της οργάνωσης της έρευνας, όταν αυτή περιλαμβάνει τη στρωματοποιημένη δειγματοληψία, είναι ποια βοηθητική μεταβλητή θα επιλέξει για τον χωρισμό του πληθυσμού σε στρώματα. Η ανάγκη της επιλογής μιας βοηθητικής μεταβλητής προκύπτει από το γεγονός ότι κατά κανόνα υπάρχουν αρκετές βοηθητικές μεταβλητές που θα μπορούσαν να σχετίζονται με την έρευνα και που ταυτόχρονα είναι διαθέσιμα τα στοιχεία τους για τον πληθυσμό. Η επιλογή δε είναι σημαντική, γιατί το όφελος στην αποτελεσματικότητα του εκτιμητή με χρήση της στρωματοποίησης έναντι της α.τ.δ. εξαρτάται από τη βοηθητική μεταβλητή και τον τρόπο που σύμφωνα με αυτή θα οριστούν τα στρώματα.

Ειδικότερα, όπως αναλυτικά έχει αποδειχθεί στο Κεφάλαιο 3, η ακρίβεια της στρωματοποιημένης δειγματοληψίας έναντι της α.τ.δ. είναι τόσο μεγαλύτερη, όσο μεγαλύτερη ομοιογένεια έχει επιτευχθεί στο εσωτερικό των στρωμάτων και ταυτόχρονα όσο πιο διαφορετικοί (ετερογενείς) είναι οι μέσοι των στρωμάτων μεταξύ τους.

Είναι προφανές ότι το χαρακτηριστικό αυτό, στην περίπτωση των συνεχών βοηθητικών μεταβλητών, επιτυγχάνεται όταν για τη στρωματοποίηση επιλεγεί εκείνη η μεταβλητή που έχει τη μεγαλύτερη δυνατή συσχέτιση με την κύρια μεταβλητή της έρευνας. Το αποτέλεσμα αυτό επιβεβαιώνεται και μαθηματικά και παρουσιάζεται στην Παράγραφο 4.1.1, παράλληλα με την παρουσίαση του προβλήματος της επιλογής του αριθμού των στρωμάτων.

Στην περίπτωση των κατηγορικών μεταβλητών, η επιλογή θα γίνει με κριτήριο την ποσότητα $\sum_{h=1}^{L} W_h(\bar{Y}_h - \bar{Y})^2$. Σύμφωνα με τη σχέση (3.17), η στρωματοποίηση που επιτυγχάνει μεγιστοποίηση της ποσότητας αυτής βελτιώνει σε μεγαλύτερο βαθμό τη στρωματοποιημένη και, κατά συνέπεια, είναι η επικρατέστερη. Προκειμένου να γίνει αυτός ο έλεγχος, είναι απαραίτητες οι εκτιμήσεις των $\bar{Y}_h$ και $\bar{Y}$ από προηγούμενες έρευνες. Για κατανόηση, ας υποθέσουμε ένα μικρό παράδειγμα ενός πληθυσμού με 10 παρατηρήσεις. Ο Πίνακας 4.1 δίνει τις τιμές της βασικής μεταβλητής $Y$ για τον πληθυσμό και τιμές για δύο διαθέσιμες βοηθητικές κατηγορικές μεταβλητές, $Z$ και $U$, για κάθε μονάδα του πληθυσμού. Η μέση τιμή του πληθυσμού

είναι $\bar{Y} = 6.23$. Αν οι τιμές του πληθυσμού χωριστούν σύμφωνα με τη μεταβλητή $Z$, θα είναι: $\bar{Y}_1 = 6.2$, $\bar{Y}_2 = 6.26$ και $\sum_{h=1}^{L} W_h(\bar{Y}_h - \bar{Y})^2 = 0.0009$. Για τη μεταβλητή $U$, οι αντίστοιχες ποσότητες είναι $\bar{Y}_1 = 7.56$, $\bar{Y}_2 = 6.95$, $\bar{Y}_3 = 3.53$ και $\sum_{h=1}^{L} W_h(\bar{Y}_h - \bar{Y})^2 = 3.17$. Η μεταβλητή $U$ θα είναι επομένως καταλληλότερη για χρήση ως βοηθητική μεταβλητή στο συγκεκριμένο παράδειγμα πληθυσμού $Y$. Σύμφωνα με τη μεταβλητή αυτή, οι μέσες τιμές του πληθυσμού ανά στρώματα είναι αρκετά διαφορετικές μεταξύ τους, δηλ. επιτυγχάνεται η επιθυμητή ετερογένεια μεταξύ των στρωμάτων.

| **Μεταβλητή Y** | 2.3 | 4.5 | 5.6 | 3.4 | 9.4 | 12.3 | 4.7 | 3.6 | 11.2 | 5.3 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Μεταβλητή Z** | 1 | 2 | 2 | 1 | 1 | 1 | 2 | 1 | 2 | 2 |
| **Μεταβλητή U** | 3 | 2 | 1 | 1 | 2 | 1 | 3 | 3 | 1 | 1 |

***Πίνακας 4.1*** *Βασική Μεταβλητή (Y) και δύο βοηθητικές μεταβλητές (Z), (U).*

Είναι δυνατόν να χρησιμοποιηθούν περισσότερες από μία βοηθητικές μεταβλητές για τη στρωματοποίηση. Για παράδειγμα, το φύλο και η ηλικιακή ομάδα θεωρούνται για τη μελέτη του ποσοστού ανεργίας αρκετά σημαντικοί παράγοντες. Στην περίπτωση αυτή, το πλήθος των στρωμάτων του πληθυσμού είναι το γινόμενο των στρωμάτων που προκύπτουν από την κάθε μία μεταβλητή χωριστά. Η υιοθέτηση δύο ή περισσότερων βοηθητικών μεταβλητών για χωρισμό των στρωμάτων έχει ως πλεονέκτημα ότι μετά το πέρας της έρευνας είναι δυνατό να εξαχθούν συμπεράσματα για ειδικές κατηγορίες πληθυσμού, π.χ. το ποσοστό ανεργίας για τις γυναίκες ηλικίας έως 25 ετών, στο παράδειγμα. Το μειονέκτημα της χρήσης δύο ή περισσότερων βοηθητικών μεταβλητών για τη στρωματοποιημένη είναι το μεγάλο πλήθος των στρωμάτων που είναι δυνατό να προκύψει και η επιβάρυνση της οργάνωσης και της διενέργειας της έρευνας που αυτό θα έχει ως συνέπεια.

### 4.1.1 Επιλογή βοηθητικής μεταβλητής και αριθμού στρωμάτων στη Στρωματοποιημένη Δειγματοληψία

Μετά την επιλογή της βοηθητικής μεταβλητής για τον σχεδιασμό των στρωμάτων, το επόμενο ερώτημα είναι ο αριθμός των στρωμάτων στα οποία πρέπει να χωριστεί ο πληθυσμός. Ο αριθμός αυτός, για τις κατηγορικές μεταβλητές είναι συνήθως προφανής γιατί συμπίπτει με το πλήθος των επιπέδων της μεταβλητής. Για παράδειγμα, αν η βοηθητική μεταβλητή είναι το φύλο, τότε τα δύο στρώματα "άνδρες" και "γυναίκες" είναι τα προφανή στρώματα. Το ίδιο αποτέλεσμα ισχύει γενικότερα για μια οποιαδήποτε μη διατάξιμη κατηγορική μεταβλητή.

Αν η βοηθητική μεταβλητή είναι κατηγορική διατάξιμη με πολλά ενδεχομένως επίπεδα, ή αν είναι συνεχής, τότε έχει νόημα να διερευνηθεί περαιτέρω σε πόσα στρώματα θα χωρίσουμε τον πληθυσμό βάσει της μεταβλητής αυτής. Για παράδειγμα, αν η βοηθητική μεταβλητή είναι ο βαθμός προτίμησης ενός υποψήφιου αγοραστή για ένα προϊόν, και τα επίπεδα της απάντησης είναι από 1 έως 10, τότε θα μπορούσαν να οριστούν 10 στρώματα, όπως σε κάθε άλλη κατηγορική βοηθητική μεταβλητή, αλλά και λιγότερα, αν υπάρξουν συνενώσεις κάποιων επιπέδων. Π.χ. 1$^{ο}$ στρώμα: προτίμηση 1 έως 3, 2$^{ο}$ στρώμα: προτίμηση 4-6, 3$^{ο}$ στρώμα: προτίμηση 7 ή 8 και 4$^{ο}$ στρώμα: προτίμηση 9 ή 10.

Επίσης για μια βοηθητική μεταβλητή που είναι συνεχής, π.χ. ηλικία κατοίκων, το ερώτημα είναι ακόμα πιο σημαντικό, και απαιτείται ομαδοποίηση των ηλικιών, έτσι ώστε να προκύψουν τα $L$, σύμφωνα με τη θεωρία, στρώματα. Υπενθυμίζεται ότι η αύξηση του αριθμού των στρωμάτων συνδέεται με το κόστος της έρευνας, οπότε η βέλτιστη επιλογή του $L$ είναι επιθυμητή.

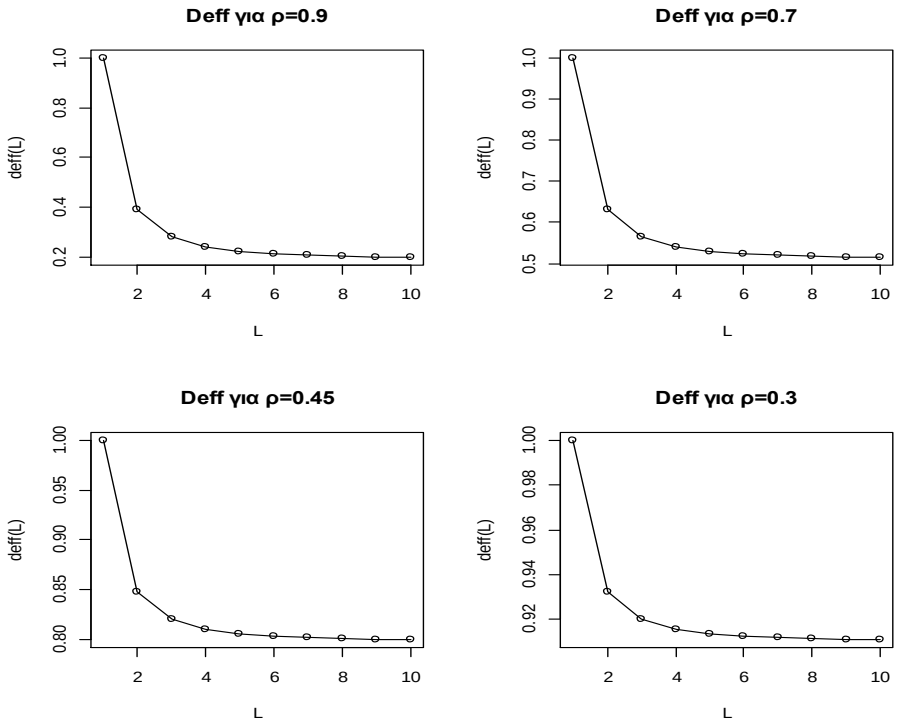
Αποδεικνύεται ότι, αν $Y$ είναι η κύρια μεταβλητή την οποία ενδιαφερόμαστε να εκτιμήσουμε, και $Z$ είναι η βοηθητική, με τη βοήθεια της οποίας θα οριστούν τα στρώματα στον πληθυσμό, τότε

$$\mathrm{Var}_{st}(\bar{Y}) \approx \frac{S_Y^2}{n}\left[\frac{\rho^2}{L^2} + (1-\rho^2)\right] \qquad (4.1)$$

όπου $\rho$ είναι ο συντελεστής συσχέτισης των μεταβλητών $Z$ και $Y$. Ο πρώτος παράγοντας στη σχέση (4.1) είναι η διακύμανση του εκτιμητή του μέσου του πληθυσμού κάτω από το απλό τυχαίο δειγματοληπτικό σχέδιο, όταν η διόρθωση του πεπερασμένου πληθυσμού θεωρηθεί αμελητέα (βλ. Παρατήρηση 2.4). Είναι συνεπώς φανερό από τη σχέση () ότι ο παράγοντας $\frac{\rho^2}{L^2} + (1 - \rho^2)$ είναι ο παράγοντας που αντιπροσωπεύει τη μείωση που θα υπάρξει στη διακύμανση του εκτιμητή από την α.τ.δ., όταν εφαρμοστεί η στρωματοποιημένη με τη βοήθεια της $Z$. Με άλλα λόγια, είναι η επίδραση του στρωματοποιημένου δειγματοληπτικού σχεδίου στην απλή τυχαία.

Στο Σχήμα 4.1 δίνονται ορισμένες γραφικές παραστάσεις της συνάρτησης $\text{deff}(L) = \frac{\rho^2}{L^2} + (1 - \rho^2)$, $L = 1,2,3, \ldots$ για διάφορες τιμές του $\rho$. Με τη βοήθεια των γραφικών παραστάσεων, διαπιστώνουμε γραφικά κάποια από τα βασικά χαρακτηριστικά της συνάρτησης $\text{deff}(L)$ και επιπλέον επιβεβαιώνουμε ορισμένα από τα θεωρητικά αποτελέσματα που ισχύουν για τη συνάρτηση της επίδρασης και είναι χρήσιμα κατά τον σχεδιασμό της έρευνας.

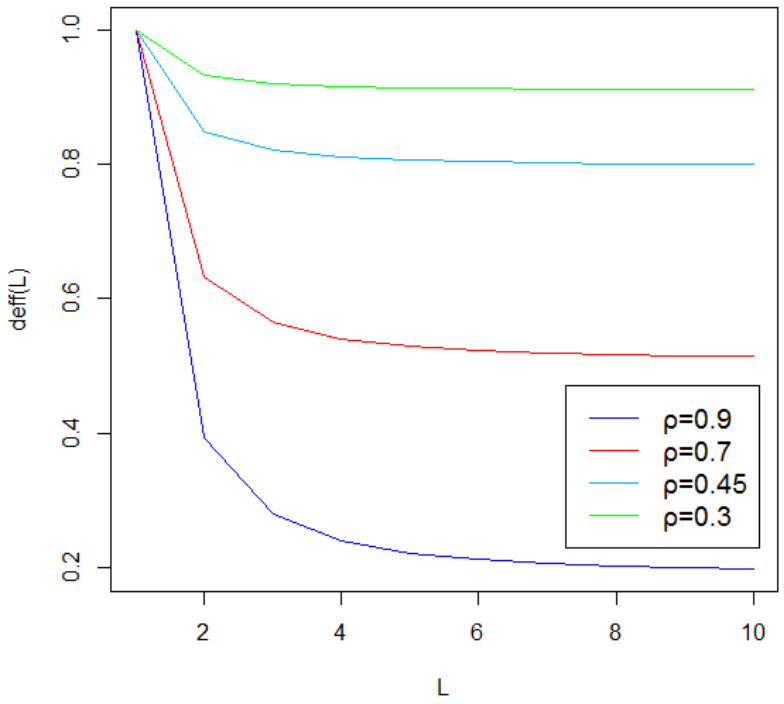
Αναλυτικά, ένα πρώτο συμπέρασμα είναι ότι η συνάρτηση είναι φθίνουσα ως προς το $L$ ($L = 1,2,3, \ldots$) και, συνεπώς, όσο αυξάνεται ο αριθμός των στρωμάτων $L$, τόσο πιο αποδοτική θα είναι η στρωματοποιημένη έναντι της απλής τυχαίας. Προφανώς, για $L = 1$ η δειγματοληψία ταυτίζεται με την απλή τυχαία και η ακρίβεια είναι η ίδια. Η φθίνουσα τάση της συνάρτησης σταθεροποιείται από έναν αριθμό $L$ και μετά, και η καμπύλη της επίδρασης του στρωματοποιημένου σχεδίου στο α.τ. δείγμα είναι σχεδόν επίπεδη. Ο αριθμός αυτός για το $L$ είναι μικρός, γεγονός που σημαίνει ότι η υιοθέτηση της στρωματοποιημένης δειγματοληψίας, αρκετά νωρίς, δηλ. ακόμη και για μικρό αριθμό στρωμάτων, προσφέρει το μεγαλύτερο μέρος της βελτίωσης της ακρίβειας του εκτιμητή. Όσο το $L$ μεγαλώνει περισσότερο, το πρόσθετο όφελος πλέον θα είναι αμελητέο. Ο αριθμός αυτός $L$ δεν διαφέρει σημαντικά για τις διάφορες τιμές του $\rho$, και ειδικότερα οι γραφικές παραστάσεις του Σχήματος 4.1 επιβεβαιώνουν το θεωρητικό συμπέρασμα ότι πρακτικά ο μέγιστος αριθμός των στρωμάτων για τις εφαρμογές δεν ξεπερνά το έξι.




***Σχήμα 4.1*** *Επίδραση στρωματοποιημένου σχεδίου ως προς την α.τ.δ.*

Το δεύτερο σημαντικό αποτέλεσμα που εξάγεται από το Σχήμα 4.1 είναι ότι η επίδραση του στρωματοποιημένου σχήματος στο απλό τυχαίο εξαρτάται από τον συντελεστή συσχέτισης $\rho$. Για παράδειγμα, όταν $\rho = 0.9$ το deff είναι $0.25$ για $L = 4$, δηλ. το στρωματοποιημένο σχήμα μειώνει τη διακύμανση $\text{Var}_{srs}(\bar{Y})$ κατά $75\%$. Για $\rho = 0.7$, το αντίστοιχο ποσοστό είναι $45\%$, ενώ για $\rho = 0.3$, το ίδιο

ποσοστό είναι μόλις 8%. Το ίδιο συμπέρασμα παρουσιάζεται πιο παραστατικά στο Σχήμα 4.2, όπου γίνεται η γραφική παράσταση και των 4 συναρτήσεων deff του Σχήματος 4.1 πάνω στο ίδιο σύστημα αξόνων. Η σχέση (4.1) επιβεβαιώνει θεωρητικά το γεγονός ότι αν υπάρχουν πολλές υποψήφιες συνεχείς βοηθητικές μεταβλητές για τη στρωματοποίηση, επικρατέστερη είναι εκείνη που συνδέεται πιο έντονα με τη βασική μεταβλητή.




***Σχήμα 4.2**: Επίδραση στρωματοποιημένης ως προς την α.τ.δ. για $\rho$ = 0.3, 0,45, 0.7 και 0.9.*

Κλείνοντας το ερώτημα για τον αριθμό των στρωμάτων στη στρωματοποιημένη, συμπεραίνουμε ότι ο αριθμός των στρωμάτων μπορεί να προσδιοριστεί με τη βοήθεια μιας εκτίμησης του $\rho$ και θέτοντας ένα προκαθορισμένο μέγεθος για το deff στο πλαίσιο των δυνατών τιμών του. Π.χ. για $\rho = 0.7$, η επίδραση deff για $L$ = 2 έως $L$ = 5 είναι

| ***L*** | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| **deff** | 0.63 | 0.56 | 0.54 | 0.53 |

Εάν το επιπλέον 2% όφελος που θα επιτύχει ο ερευνητής από την αύξηση των στρωμάτων κατά ένα, μεταβαίνοντας από $L = 3$ σε $L = 4$, είναι σημαντικό, θα επιλέξει $L = 4$. Εάν το deff $= 0.63$, δηλ. το όφελος 37% που επιτυγχάνει με την επιλογή $L = 2$ κρίνεται επαρκές, τότε τα δύο στρώματα είναι αρκετά για την εφαρμογή της στρωματοποιημένης.

## 4.2. Καθορισμός ορίων των στρωμάτων στη Στρωματοποιημένη Δειγματοληψία

Όταν τα ερωτήματα της βοηθητικής μεταβλητής που θα χρησιμοποιηθεί για τη στρωματοποίηση και του αριθμού των στρωμάτων έχουν απαντηθεί, το επόμενο πρόβλημα κατά τον σχεδιασμό της έρευνας είναι να καθοριστούν τα όρια της βοηθητικής μεταβλητής που θα ορίζουν ταυτόχρονα και τα όρια των στρωμάτων.

Σε ορισμένες εφαρμογές, τα όρια καθορίζονται από τη διαθεσιμότητα των στοιχείων για τον πληθυσμό και δεν υπάρχει η δυνατότητα καθορισμού των ορίων από τον ερευνητή. Για παράδειγμα, αν τα στρώματα

χωρίζονται βάσει του ταχυδρομικού κώδικα, τότε όλες οι διευθύνσεις που έχουν τον ίδιο ταχυδρομικό κώδικα ανήκουν στο ίδιο στρώμα. Δεν θα είναι εφικτό να χωρίσουμε τις διευθύνσεις που έχουν τον ίδιο ταχυδρομικό κώδικα.

Όταν η βοηθητική μεταβλητή είναι συνεχής και είναι διαθέσιμα τα στοιχεία στον ερευνητή, τότε είναι δυνατόν ο ίδιος να καθορίσει τα όρια του κάθε στρώματος. Για τον σκοπό αυτό, έχουν αναπτυχθεί διάφορες προτάσεις (βλ. Levy & Lemeshow, 1999, Κεφ. 6), τις σημαντικότερες των οποίων θα αναπτύξουμε στη συνέχεια. Αξίζει να σημειωθεί ότι η υλοποίηση αυτών των μεθοδολογιών είναι διαθέσιμη μέσα από στατιστικά πακέτα. Στο παρόν κεφάλαιο, θα γίνει η περιγραφή και η εφαρμογή ορισμένων από τις μεθόδους καθορισμού των ορίων των στρωμάτων με τη βοήθεια του στατιστικού πακέτου `stratification` της R.

### 4.2.1 Κατασκευή στρωμάτων στη Στρωματοποιημένη Δειγματοληψία

#### A. Μέθοδος ίσου εύρους (equal range method)

Η μέθοδος αυτή είναι η πιο απλή και, όπως δηλώνει το όνομά της, τα όρια των στρωμάτων ορίζονται έτσι ώστε τα στρώματα να έχουν ίσο εύρος τιμών για τη βοηθητική μεταβλητή. Εάν το συνολικό εύρος της βοηθητικής μεταβλητής $Z$ για τον πληθυσμό είναι $R$ και αν τα στρώματα που επιθυμούμε να χωρίσουμε τον πληθυσμό είναι $L$, τότε τα όρια της βοηθητικής μεταβλητής είναι:

$$\min_i W, \min_i W + \frac{R}{L}, \ \min_i W + 2\frac{R}{L}, \dots, \min_i W + (L-1)\frac{R}{L}$$

και κατά συνέπεια, τα όρια για τα αντίστοιχα στρώματα που ορίζονται θα είναι αναλυτικά:

$$\left[\min_i Z, \ \min_i Z + \frac{R}{L}\right),$$
$$\left[\min_i Z + \frac{R}{L}, \ \min_i Z + 2\frac{R}{L}\right),$$
$$\dots$$
$$\left[\min_i Z + (L-2)\frac{R}{L}, \min_i Z + (L-1)\frac{R}{L}\right].$$

Σύμφωνα με τον ορισμό των ορίων, το πρώτο στρώμα θα αποτελείται από τις μονάδες του πληθυσμού που η τιμή τους για τη βοηθητική μεταβλητή $Z$ κυμαίνεται στο διάστημα $\left[\min_i Z, \ \min_i Z + \frac{R}{L}\right)$, το δεύτερο στρώμα περιλαμβάνει τις μονάδες για τις οποίες η $Z$ ανήκει στο $\left[\min_i Z + \frac{R}{L}, \min_i Z + 2\frac{R}{L}\right)$ κ.ο.κ.

Η μέθοδος ίσου εύρους για τον καθορισμού των ορίων των στρωμάτων, έχει ως πλεονέκτημα την ευκολία στην εφαρμογή, αλλά είναι φανερό, από τον τρόπο που ορίζονται τα όρια της $W$, ότι τα στρώματα που θα προκύψουν έχουν ιδιότητες που εξαρτώνται από το σχήμα της κατανομής τής $Z$. Αναλυτικότερα, αν η $Z$ κατανέμεται ομοιόμορφα, τότε το κάθε στρώμα θα έχει ίδιο αριθμό παρατηρήσεων. Όταν η κατανομή της βοηθητικής μεταβλητής παρουσιάζει θετική ή αρνητική ασυμμετρία, τότε ο αριθμός των παρατηρήσεων ανά στρώμα μπορεί να διαφέρει αρκετά. Ορισμένα στρώματα θα έχουν λίγες μόνο παρατηρήσεις, ενώ αντίθετα ορισμένα μπορεί να είναι πολυπληθή. Το γεγονός ότι το εύρος παρατηρήσεων είναι ίδιο για όλα τα στρώματα συνεπάγεται ότι οι διακυμάνσεις μέσα σε κάθε στρώμα, σύμφωνα με την ίσου εύρους μεθοδολογία, είναι συγκρίσιμες. Αυτό με τη σειρά του, συνεπάγεται πως στην περίπτωση αυτή, ο βέλτιστος καταμερισμός, εάν επιπλέον το κόστος ανά στρώμα είναι ίσο, ταυτίζεται με τον αναλογικό καταμερισμό (βλ. Σχέση 3.13). Η περίπτωση συνεπώς ενός πληθυσμού με θετική ή αρνητική ασυμμετρία θα συνιστούσε μέγεθος δείγματος πάρα πολύ μεγάλο για τα στρώματα με τη μεγάλη συγκέντρωση παρατηρήσεων.

Για την αντιμετώπιση του παραπάνω φαινομένου, έχει προταθεί η επόμενη μέθοδος, που βασίζεται στα εκατοστημόρια.

### Β. Μέθοδος εκατοστημορίων (quantile method)

Σύμφωνα με τη μέθοδο των εκατοστημορίων, τα όρια της βοηθητικής μεταβλητής ορίζονται έτσι, ώστε μέσα στο καθένα από τα $L$ στρώματα που θα προκύψουν να ανήκει ίσος αριθμός παρατηρήσεων για τον πληθυσμό.

Αυτό επιτυγχάνεται με τη βοήθεια των εκατοστημορίων για τη βοηθητική μεταβλητή. Πιο συγκεκριμένα, αν $L = 4$ είναι ο επιθυμητός αριθμός των στρωμάτων, τα όρια της $Z$ καθορίζονται με τη βοήθεια του $25^{ου}$, $50^{ού}$ και $75^{ου}$ εκατοστημορίου, δηλ. το $1^{ο}$ τεταρτημόριο $R_1$, τη διάμεσο $M$ και το τρίτο τεταρτημόριο $R_3$. Τα 4 στρώματα που προκύπτουν από τα όρια αυτά είναι οι τιμές του πληθυσμού για τις οποίες η $Z$ παίρνει τιμές στα διαστήματα $[\min_i Z\ ,\ R_1)$, $[R_1, M)$, $[M, R_3)$ και $[R_3, \max_i Z]$.

Σύμφωνα με τη μεθοδολογία αυτή, τα στρώματα που προκύπτουν έχουν ίδιο αριθμό παρατηρήσεων, αλλά διαφορετική διακύμανση. Η ιδιότητα αυτή, σε ορισμένες περιπτώσεις, μπορεί να οδηγήσει σε μη βέλτιστα αποτελέσματα. Για παράδειγμα, εάν η κατανομή του βοηθητικού χαρακτηριστικού $Z$ παρουσιάζει θετική ασυμμετρία ή λοξότητα προς τα δεξιά, τα τελευταία στρώματα, εκείνα που αντιστοιχούν στην 'ουρά' της κατανομής, θα έχουν μεγαλύτερη διακύμανση. Κατά συνέπεια, ο βέλτιστος καταμερισμός του δείγματος σε μια τέτοια περίπτωση για τον πληθυσμό θα πρότεινε περισσότερες – δυσανάλογα περισσότερες – μονάδες του πληθυσμού για το δείγμα από τα στρώματα αυτά.

### Γ. Αθροιστική τετραγωνική ρίζα των συχνοτήτων (cumulative root frequency method)

Η μέθοδος της αθροιστικής τετραγωνικής ρίζας των συχνοτήτων (rootfreq) είναι μια προσπάθεια να απαλειφθούν τα ακραία φαινόμενα που παρατηρούνται ως προς τον αριθμό των μονάδων ανά στρώμα σύμφωνα με τις προηγούμενες δύο μεθόδους. Για την εφαρμογή της rootfreq μεθόδου κατασκευάζεται αρχικά ένας πίνακας συχνοτήτων για τις τιμές της $Z$ στις μονάδες του πληθυσμού. Υπολογίζεται στη συνέχεια η στήλη με την τετραγωνική ρίζα των συχνοτήτων για τις κλάσεις της $Z$ και τέλος μία στήλη με τις αθροιστικές τιμές της στήλης αυτής. Ο καθορισμός των ορίων στη $Z$, γίνεται με τη βοήθεια των αθροιστικών τετραγωνικών ριζών των συχνοτήτων. Συγκεκριμένα, το εύρος του κάθε υποσυνόλου τιμών για τη $Z$ που ορίζουν τα στρώματα υπολογίζεται από τη διαίρεση της μέγιστης αθροιστικής τετραγωνικής ρίζας συχνότητας με το $L$.

Το παράδειγμα που ακολουθεί δείχνει με πιο παραστατικό τρόπο την υλοποίηση της κάθε μεθόδου.

## 4.2.2 Εφαρμογή: SHS data

Για την εφαρμογή θα γίνει χρήση του πακέτου `stratification` της `R` (βλ. Baillargeon and Rivest, 2011). Το πακέτο αυτό δεν συμπεριλαμβάνεται στο βασικό κορμό εντολών της `R` και χρειάζεται να εγκατασταθεί από τον χρήστη. Το πακέτο, όπως και η `R`, είναι διαθέσιμα στον χρήστη με ελεύθερη πρόσβαση. (Σύντομη περιγραφή βασικών εντολών της `R` και οδηγίες για την εγκατάσταση ενός πακέτου δίνονται στο Παράρτημα Ι).

Εφόσον το πακέτο εγκατασταθεί, η εντολή

```
>library(stratification)
```

καθιστά διαθέσιμο το πακέτο στον τρέχοντα φάκελο εργασίας. Οι συναρτήσεις και τα αρχεία δεδομένων που περιέχονται στο πακέτο είναι άμεσα προσβάσιμα.

Θα χρησιμοποιήσουμε το αρχείο δεδομένων με τίτλο SHS, που περιέχεται στο πακέτο `stratification`, ώστε να εφαρμόσουμε τις ανωτέρω τεχνικές για την κατασκευή των στρωμάτων στη στρωματοποιημένη δειγματοληψία. Τα δεδομένα περιέχουν πληροφορίες από την έρευνα "2001 Survey of Household Spending (SHS)" και αποτελούνται από 16057 παρατηρήσεις και μετρήσεις σε 7 μεταβλητές.

Αναλυτικότερες πληροφορίες για την κάθε μεταβλητή είναι δυνατόν να εξαχθούν με την εντολή `help(SHS)`. Εκτελώντας την εντολή

```
>help(SHS)
```

παίρνουμε τις παρακάτω πληροφορίες για το αρχείο δεδομένων SHS.

A data frame with 16057 observations on the following 7 variables.

`CASEID` Identification number

`WEIGHT` Weight at household level

`PROVINCP` Province or territory code

`URBRUR` Urban rural code

`URBSIZEP` Size of area of residence code

`HHINCTOT` Household income before taxes

`M101` Household spending on recreation

Details

In this package, `HHINCTOT` is used as a stratification variable and `M101` as a survey variable.

Source

Income Statistics Division, Statistics Canada.

References

Information on Statistics Canada Survey of Household Spending:

http://www.statcan.ca/cgi-bin/imdb/p2SV.pl?Function=getSurvey&SDDS=3508&lang=en&db=IMDB&dbg=f&adm=8&dis=2

Για την εφαρμογή των μεθοδολογιών που αναπτύχθηκαν στην παράγραφο 4.2.1, θα γίνει χρήση των δεδομένων SHS με βασική μεταβλητή $(Y)$ την `M101`: έξοδα του νοικοκυριού για αναψυχή και βοηθητική μεταβλητή $(Z)$ για την κατασκευή των στρωμάτων την `HHINCTOT`: συνολικά έσοδα του νοικοκυριού (προ φόρων).

Έστω ότι σκοπεύουμε να κατασκευάσουμε $L = 5$ στρώματα και να επιλεγεί ένα δείγμα μεγέθους $n = 200$. Σκοπός της εφαρμογής είναι

(i) Να γίνει η υλοποίηση των μεθοδολογιών (α) ίσου εύρους, (β) εκατοστημορίων και (γ) αθροιστική τετραγωνική ρίζα των συχνοτήτων.

(ii) Να συγκριθούν οι τρεις παραπάνω μεθοδολογίες ως προς την ακρίβεια του εκτιμητή που θα παραχθεί.

1(α). Ξεκινώντας την εφαρμογή για τη μέθοδο ίσου εύρους, ορίζουμε ως z τη μεταβλητή του SHS που αντιστοιχεί στην `HHINCTOT`. Επιπλέον, κρατάμε μόνο τις θετικές τιμές αυτής της μεταβλητής. Στη συνέχεια, υπολογίζουμε το εύρος της βοηθητικής μεταβλητής και το διαιρούμε με $L$, ώστε να προσδιορίσουμε το πλάτος του κάθε στρώματος. Οι αντίστοιχες εντολές είναι:

```
> data(SHS)
> z<-SHS$HHINCTOT[SHS$HHINCTOT>0]
> range(z)
[1]   100 690000
> width=(range(z)[2]-range(z)[1])/5
> bounds<-seq(from=range(z)[1]+width,to=range(z)[2]-width, by= width)
```

```
> bounds
[1] 138080 276060 414040 552020
```

Η μεταβλητή `z` παίρνει τιμές από 100 έως 690000. Τα όρια για τα 5 στρώματα είναι: `[100, 138080, 276060, 414040, 552020, 690000]`. Στη μεταβλητή με όνομα `bounds` έχουν αποθηκευτεί τα ενδιάμεσα όρια των στρωμάτων, εκτός των δύο ακραίων που είναι η ελάχιστη και η μέγιστη τιμή που παρατηρείται. Η μεταβλητή `bounds` κατασκευάζεται με τη συγκεκριμένη μορφή, γιατί έτσι απαιτείται από τις υπόλοιπες εντολές του πακέτου στις οποίες θα αποτελεί ένα εσωτερικό όρισμα.

Η συνάρτηση `strata.bh` του πακέτου `stratification` κατασκευάζει τα στρώματα για τον πληθυσμό, εάν τα όρια είναι γνωστά. Τα βασικά ορίσματα που δέχεται είναι η βοηθητική μεταβλητή, τα όρια των στρωμάτων (χωρίς τα δύο ακραία: $\min_i Z$ και $\max_i Z$), τον αριθμό του δείγματος ή εναλλακτικά το επιθυμητό επίπεδο του συντελεστή μεταβλητότητας για τον εκτιμητή και τον αριθμό των στρωμάτων. Για τη σύνταξη της συνάρτησης και την πλήρη λίστα των ορισμάτων της, καθώς και του αποτελέσματος, βλέπε `>help(strata.bh)`. Στο παράδειγμά μας, η

```
>equal<-strata.bh(z, bounds, n=200, Ls=5)
```

κατασκευάζει τα 5 στρώματα που επιθυμούμε για τη $Z$ μεταβλητή και επιλέγει το δείγμα συνολικού μεγέθους $n = 200$, υλοποιώντας Neyman allocation (default επιλογή για καταμερισμό στη συνάρτηση). Όταν ζητηθεί το αποτέλεσμα, προκύπτει:

```
>equal
Given arguments:
x = z
n = 200, Ls = 5, takenone = 0, takeall = 0
allocation: q1 = 0.5, q2 = 0, q3 = 0.5
model = none

Strata information:
          |    typerhbh |      E(Y)     Var(Y)        Nh   nh   fh
stratum 1|take-some  1 138080| 47053.73  908539178 15422 191 0.01
stratum 2|take-some  1 276060| 168365.90 977791519   563   6 0.01
stratum 3|take-some  1 414040| 316000.00 1537333333   30   1 0.03
stratum 4|take-some  1 552020| 500000.00 1600000000    8   1 0.12
stratum 5|take-some  1 690001| 675000.00  225000000    2   1 0.50
Total                                              16025 200 0.01

Total sample size: 200
Anticipated population mean: 52123.73
Anticipated CV: 0.04094801
Note: CV=RRMSE (Relative Root Mean Squared Error) because takenone=0.
```

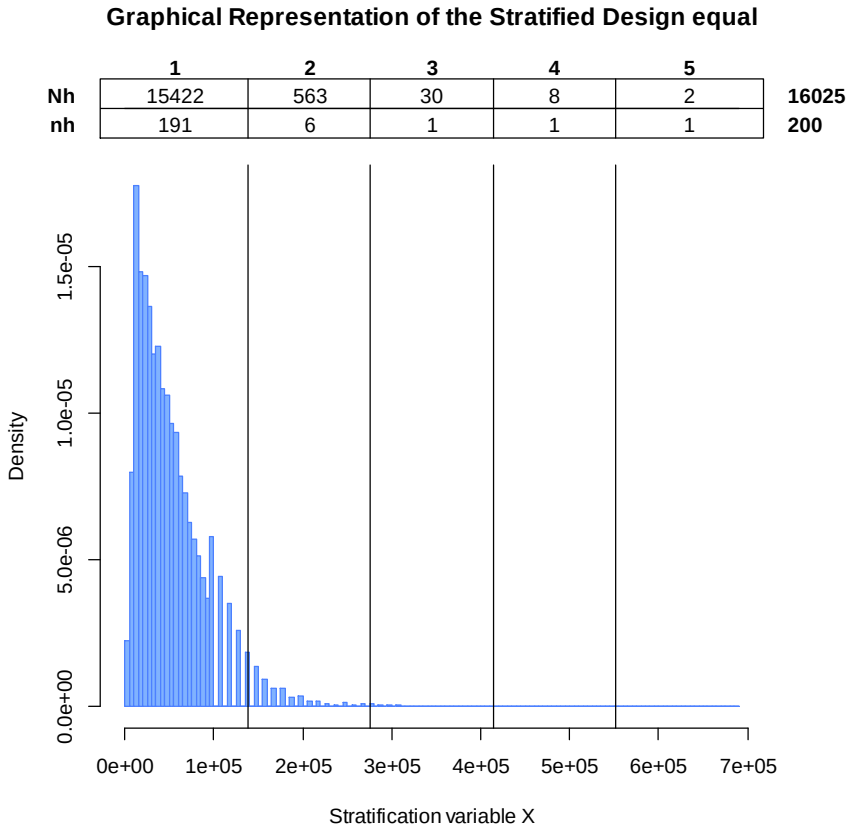
Οι τελευταίες τρεις στήλες (`Nh, nh, fh`) του πίνακα των αποτελεσμάτων περιλαμβάνουν το μέγεθος του πληθυσμού για το κάθε στρώμα που προκύπτει, το μέγεθος του δείγματος που επιλέγεται και το πηλίκο του δείγματος ανά στρώμα. Για το συγκεκριμένο αρχείο δεδομένων, ισχύει η παρατήρηση κατά την περιγραφή της μεθόδου ως προς την ανισοκατανομή του δείγματος στα στρώματα. Το πρώτο στρώμα περιέχει 15422 παρατηρήσεις από τις 16025 συνολικά, δηλ. ένα ποσοστό 96%. Αυτό έχει ως συνέπεια από τις 200 παρατηρήσεις του δείγματος οι 191 να προτείνεται να επιλεχθούν από το στρώμα 1 (το 95.5% του συνολικού

δείγματος), ενώ από τα τελευταία τρία στρώματα επιλέγεται μόλις 1 παρατήρηση από το καθένα. Αυτό συμβαίνει γιατί η κατανομή των μετρήσεων της $Z$ παρουσιάζει θετική λοξότητα.

Πράγματι, αν ζητήσουμε τη γραφική παράσταση της $W$ μέσω της εντολής `plot` του πακέτου stratification το αποτέλεσμα παρουσιάζεται στο Σχήμα 4.3. Για την εντολή, γράφουμε

```
>plot(equal)
```

Το γράφημα περιλαμβάνει τη γραφική παράσταση της κατανομής της $Z$ από την οποία διαπιστώνουμε τη λοξότητα προς τα δεξιά και, με μορφή καθέτων προς τον οριζόντιο άξονα, τα όρια των στρωμάτων. Για τη μέθοδο του ίσου εύρους, διαπιστώνουμε ότι οι κάθετοι άξονες που παριστούν τα όρια των στρωμάτων είναι σε ίσες αποστάσεις μεταξύ τους. Τέλος, στο πάνω μέρος του γραφήματος, υπάρχει ένας πίνακας που δίνει συνοπτικά τα αποτελέσματα της κατασκευής των στρωμάτων σύμφωνα με τη μέθοδο του ίσου εύρους.




***Σχήμα 4.3**: Γραφική αναπαράσταση στρωματοποίησης. SHSdata: Equal method.*

(β) Για τη μέθοδο που βασίζεται στα εκατοστημόρια, αφού πρώτα υπολογίσουμε τα όρια της βοηθητικής μεταβλητής με βάση τα ίσα εκατοστημόρια, και ακολουθώντας την ίδια διαδικασία όπως στο (α), καλούμε την ίδια εντολή για την κατασκευή των στρωμάτων. Επειδή $L = 5$, τα εκατοστημόρια που ορίζουν τα στρώματα είναι το $20^o$, $40^o$, $60^o$ και $80^o$. Συνεπώς, τα νέα όρια για την $W$ είναι

```
>bounds.p<-quantile(z, c(.2,.4,.6,.8))
>bounds.p
  20%   40%   60%   80%
20000 34000 52000 78000
```

Εισάγουμε τα νέα όρια στη συνάρτηση `strata.bh`, κρατώντας όλες τις υπόλοιπες επιλογές σταθερές με την (α) περίπτωση, ώστε να μπορεί να γίνει η σύγκριση. Αναλυτικά, είναι

```
>percentile<-strata.bh(z, bounds.p, n=200, Ls=5)
```

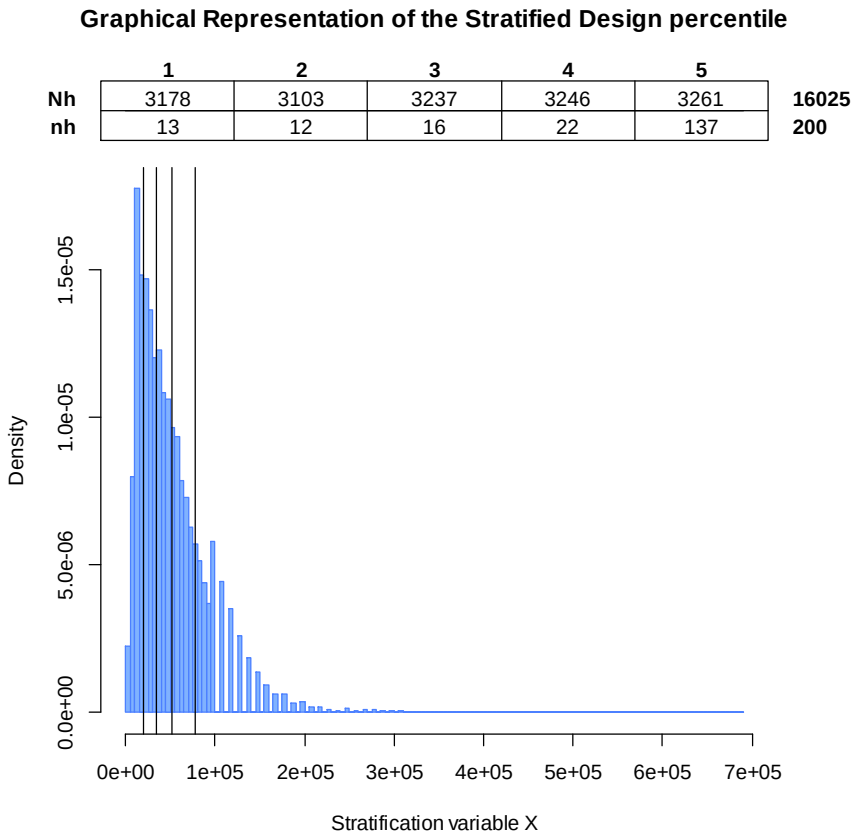
```
>percentile
Given arguments:
x = z
n = 200, Ls = 5, takenone = 0, takeall = 0
allocation: q1 = 0.5, q2 = 0, q3 = 0.5
model = none

Strata information:
   |type    rh        bh  | E(Y)     Var(Y)      Nh   nh  fh
stratum 1| take-some  1  20000| 12857.05  18070921  3178 13 0.00
stratum 2| take-some  1  34000| 26197.87  16097488  3103 12 0.00
stratum 3| take-some  1  52000| 42223.66  27046051  3237 16 0.00
stratum 4| take-some  1  78000| 63246.46  53912149  3246 22 0.01
stratum 5| take-some  1 690001|113816.31 2026673098 3261 137 0.04
Total                                            16025 200 0.01

Total sample size: 200
Anticipated population mean: 52123.73
Anticipated CV: 0.01779049
Note: CV=RRMSE (Relative Root Mean Squared Error) because takenone=0.
```

Η τρίτη από τα δεξιά στήλη (Nh) στον πίνακα των αποτελεσμάτων δίνει τα μεγέθη του πληθυσμού ανά στρώμα και διαπιστώνουμε ότι πράγματι τα μεγέθη αυτά είναι κατά προσέγγιση ίσα σύμφωνα με τη μέθοδο των εκατοστημορίων. Ως προς τα δειγματικά μεγέθη, το δείγμα δεν είναι πλέον συγκεντρωμένο στο πρώτο μόνο στρώμα, όπως συνέβαινε με τη μέθοδο του ίσου εύρους. Αντίθετα, σύμφωνα με τη μέθοδο των εκατοστημορίων, εμφανίζεται το ακριβώς αντίστροφο πρόβλημα στο αποτέλεσμα. Τα μεγαλύτερα στρώματα, δεξιά στον άξονα της βοηθητικής μεταβλητής, συγκεντρώνουν πάρα πολλές παρατηρήσεις για το δείγμα, ενώ ταυτόχρονα ο πληθυσμός έχει μικρό αριθμό παρατηρήσεων στα στρώματα αυτά. Ειδικά για το πέμπτο στρώμα, προτείνεται να επιλεγούν 137 παρατηρήσεις που αντιστοιχούν περίπου στο 70% του δείγματος. Το αποτέλεσμα αυτό οφείλεται στη μεγάλη διακύμανση των τιμών της βοηθητικής στα μεγάλα στρώματα, η οποία με τη σειρά της είναι συνέπεια της λοξότητας της κατανομής προς τα δεξιά.

Το γράφημα με το αποτέλεσμα του καταμερισμού σύμφωνα με τη μέθοδο των ίσων εκατοστημορίων προκύπτει με την εντολή:

```
>plot(percentile)
```

και δίνεται στο Σχήμα 4.4. Στο γράφημα του καταμερισμού, οι κάθετοι άξονες που συμβολίζουν τα όρια των στρωμάτων δεν είναι πλέον ανά ίση απόσταση όπως με την πρώτη μέθοδο, αλλά είναι τοποθετημένοι αρκετά πιο πυκνά στα πρώτα στρώματα, λόγω της εμφάνισης μεγαλύτερης συχνότητας παρατηρήσεων για τα στρώματα αυτά.




***Σχήμα 4.4**: Γραφική αναπαράσταση στρωματοποίησης. SHS data: Percentile method.*

(γ) Για την εφαρμογή της μεθόδου της αθροιστικής τετραγωνικής ρίζας των συχνοτήτων, χρησιμοποιούμε την εντολή `strata.cumrootf` του πακέτου `stratification`. Κρατάμε τις ίδιες επιλογές ως προς το συνολικό μέγεθος δείγματος που επιδιώκουμε να επιλεγεί και τον αριθμό των στρωμάτων. Ένα από τα ορίσματα της εντολής μάς επιτρέπει να εισάγουμε τα όρια των ομάδων για την ομαδοποίηση των δεδομένων, προκειμένου να κατασκευαστεί ο πίνακας συχνοτήτων με τη βοήθεια του οποίου θα υπολογιστούν οι συχνότητες και η τετραγωνική ρίζα αυτών. Εάν δεν εισάγουμε τα όρια ή το μήκος των ομάδων, η επιλογή θα γίνει αυθαίρετα από το πακέτο.

Για το παράδειγμα μας, η εντολή είναι:

```
>cumrootf<-strata.cumrootf(z, n=200, Ls=5)
Warning message:
'nclass' value has been chosen arbitrarily
```

Το μήνυμα για τη μεταβλητή `nclass` προειδοποιεί ότι ένα όρισμα, που αφορά τον αριθμό των κλάσεων για την ομαδοποίηση των δεδομένων, δεν έχει προσδιοριστεί κατά τη σύνταξη της εντολής και αυτό θα υπολογιστεί αυθαίρετα από το πακέτο.

Το αποτέλεσμα που έχει αποθηκευθεί στο αντικείμενο (object της `R`) με όνομα `cumrootf` είναι αναλυτικά

```
>cumrootf
Given arguments:
x = w
nclass = 75, n = 200, Ls = 5
```

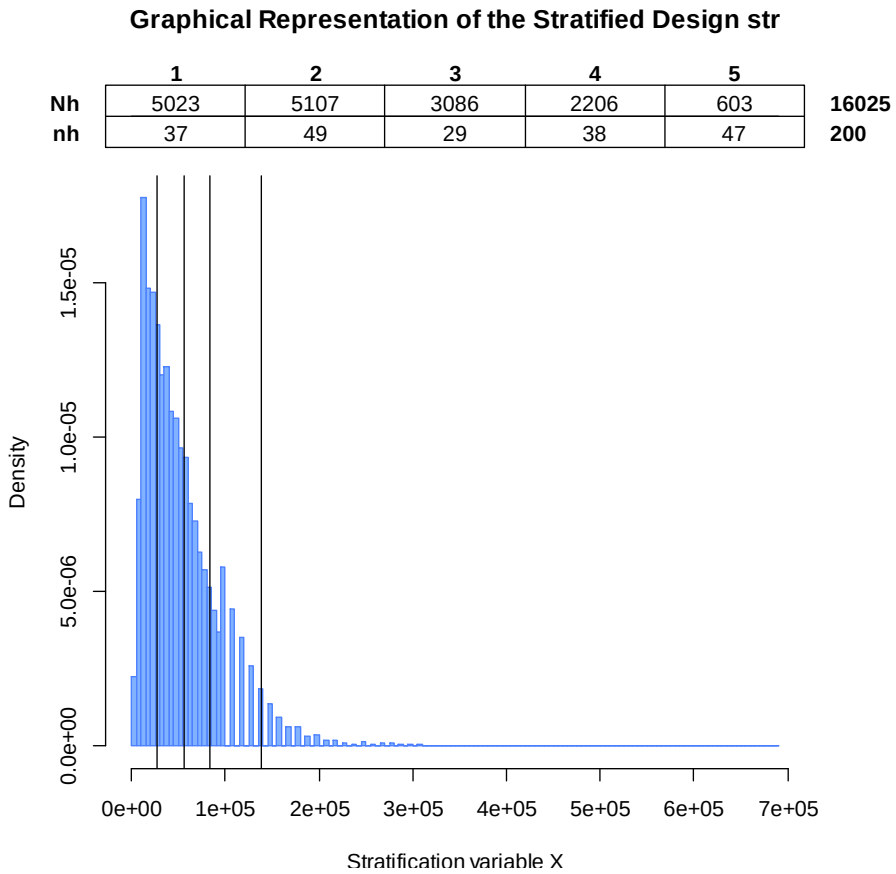
```
allocation: q1 = 0.5, q2 = 0, q3 = 0.5
model = none

Strata information:
          |    typerh  |     bh      E(Y)     Var(Y)    Nh   nh  fh
stratum 1| take-some 1|  27696  16714.45   38987293  5023 37 0.01
stratum 2| take-some 1|  55292  40654.59   65476150  5107 49 0.01
stratum 3| take-some 1|  82888  67628.32   59775983  3086 29 0.01
stratum 4| take-some 1| 138080 102167.72  207144126  2206 38 0.02
stratum 5| take-some 1| 690001 181791.04 4225979555   603 47 0.08
Total                                              16025 200 0.01

Total sample size: 200
Anticipated population mean: 52123.73
Anticipated CV: 0.01399077
```

Σύμφωνα με τον καταμερισμό που προτείνει η μέθοδος αυτή, το δείγμα είναι πιο ομοιόμορφα κατανεμημένο στα 5 στρώματα. Το πηλίκο του δείγματος είναι μεγαλύτερο στα μεγάλα στρώματα, λόγω της μεγαλύτερης συγκριτικά με τα υπόλοιπα στρώματα διασποράς, αλλά δεν εμφανίζονται ακραία φαινόμενα όπου το 90% ή το 70% των παρατηρήσεων του δείγματος να συγκεντρώνεται σε ένα μόνο στρώμα.

Το γράφημα του καταμερισμού του δείγματος για τη μέθοδο της αθροιστικής τετραγωνικής συχνότητας του δείγματος προκύπτει με την εντολή:

```
>plot(percentile)
```

και παρουσιάζεται στο Σχήμα 4.5.




***Σχήμα 4.5**: Γραφική αναπαράσταση στρωματοποίησης. SHS data: Αθροιστική τετραγωνική ρίζα συχνοτήτων.*

Για τη σύγκριση των τριών προτάσεων καταμερισμού του δείγματος, ο αριθμός των στρωμάτων και το συνολικό μέγεθος του δείγματος έχουν διατηρηθεί σταθερά. Η σύγκριση συνεπώς θα υλοποιηθεί ως προς την ακρίβεια του παραγόμενου εκτιμητή. Άρα, για $n = 200$ και εφαρμογή στρωματοποιημένης δειγματοληψίας σε πέντε στρώματα, συγκρίνουμε τον συντελεστή μεταβλητότητας του εκτιμητή που θα προκύψει ακολουθώντας τις 3 μεθοδολογίες για τον καθορισμό των ορίων των στρωμάτων. Υπενθυμίζουμε ότι η μέθοδος καθορισμού του μεγέθους δείγματος σε κάθε στρώμα έχει επιλεγεί να είναι η Neyman (default επιλογή) και στις τρεις περιπτώσεις.

Οι τελευταίες 3 γραμμές στο αποτέλεσμα της κάθε μεθοδολογίας μας δίνουν το μέγεθος του δείγματος, την αναμενόμενη τιμή του εκτιμητή του μέσου του πληθυσμού και την αναμενόμενη τιμή του συντελεστή μεταβλητότητας του εκτιμητή. Για τις τρεις μεθόδους (α), (β) και (γ) του (i) ερωτήματος, ο συντελεστής μεταβλητότητας, όπως προκύπτει από τα διαδοχικά αποτελέσματα, είναι 0.04094801, 0.01779049 και 0.01399077 αντίστοιχα. Άρα, ο εκτιμητής σύμφωνα με τη μέθοδο της αθροιστικής τετραγωνικής ρίζας των συχνοτήτων είναι πιο ακριβής και, συνεπώς, η μέθοδος αυτή για τον καθορισμό των στρωμάτων είναι επικρατέστερη των δύο άλλων.

## 4.3. Εκ των Υστέρων Στρωματοποιημένη Δειγματοληψία

**Η εκ των υστέρων στρωματοποιημένη (poststratification) δειγματοληψία** είναι το δειγματοληπτικό σχέδιο σύμφωνα με το οποίο η επιλογή του δείγματος γίνεται με απλό τυχαίο τρόπο, αλλά η εκτίμηση και η στατιστική συμπερασματολογία γενικότερα, γίνεται με τη βοήθεια της στρωματοποιημένης δειγματοληψίας. Η εκ των υστέρων στρωματοποιημένη δειγματοληψία ή, όπως λέγεται διαφορετικά, η στρωματοποίηση μετά τη συλλογή του δείγματος, εφαρμόζεται όταν το δείγμα έχει ήδη επιλεγεί και για λόγους είτε σχεδιασμού είτε πρακτικούς η στρωματοποίηση του πληθυσμού δεν είχε συμπεριληφθεί κατά τη διαδικασία επιλογής του δείγματος. Επειδή το όφελος για την αποτελεσματικότητα είναι μεγάλο κατά τη στρωματοποιημένη, γίνεται προσπάθεια να ανακτηθεί ένα τουλάχιστον μέρος από αυτό, κάνοντας τη στρωματοποίηση έστω μετά την

επιλογή του δείγματος. Πράγματι, το κύριο πλεονέκτημα της εκ των υστέρων στρωματοποιημένης δειγματοληψίας είναι ότι, παρόλο που εφαρμόζεται μετά τη δειγματοληψία, οι εκτιμητές που προκύπτουν έχουν τυπικά σφάλματα πολλές φορές λίγο μόνο μεγαλύτερα από εκείνα που θα είχαμε εάν εφαρμόζονταν η τυχαία στρωματοποιημένη. Για το λόγο αυτό, η εκ των υστέρων στρωματοποιημένη εφαρμόζεται αρκετά συχνά.

Αναλυτικότερα, έστω ότι δεν είναι διαθέσιμη η πληροφορία για μια βοηθητική μεταβλητή που θα επέτρεπε τον εκ των προτέρων χωρισμό των μελών του πληθυσμού σε στρώματα ή είναι πρακτικά δύσκολη ή χρονοβόρα η εφαρμογή της, ακόμα και αν υπάρχει διαθέσιμη η βοηθητική μεταβλητή. Για παράδειγμα, έστω ότι η έρευνα που πρόκειται να διεξαχθεί είναι για τον πληθυσμό των ασθενών ενός Νοσοκομείου το τελευταίο έτος και αφορά την εκτίμηση του ποσοστού των πρόωρων τοκετών. Σύμφωνα με έρευνες, είναι γνωστό ότι ο παράγοντας κάπνισμα για τη μητέρα συνδέεται με πρόωρο τοκετό. Συνεπώς, η στρωματοποίηση των γυναικών της Μαιευτικής κλινικής του Νοσοκομείου σε καπνίστριες και μη καπνίστριες θα οδηγούσε σε μια πιο αποτελεσματική εκτίμηση του ποσοστού των πρόωρων τοκετών. Για να γίνει όμως αυτή η στρωματοποίηση, πρέπει η καρτέλα της κάθε ασθενούς που έχει εισαχθεί στη συγκεκριμένη κλινική του Νοσοκομείου τον τελευταίο χρόνο να εξεταστεί ως προς το αν είναι καπνίστρια ή όχι, και αναλόγως να δημιουργηθούν τα στρώματα. Η διαδικασία αυτή μπορεί να είναι επίπονη και χρονοβόρα.

Σύμφωνα με την εκ των υστέρων στρωματοποιημένη δειγματοληψία, το δείγμα επιλέγεται από ολόκληρο τον πληθυσμό ως ενιαίο σύνολο και, κατόπιν της συλλογής του δείγματος, γίνεται ο διαχωρισμός των δειγματοληπτικών μονάδων σε στρώματα βάσει του κριτηρίου που εκτιμάται ότι δίνει ομοιογενή στρώματα. Στο παράδειγμα με το ποσοστό των πρόωρων τοκετών, το δείγμα επιλέγεται αρχικά από το σύνολο των ασθενών που έχουν εισαχθεί τον τελευταίο χρόνο στο Νοσοκομείο και στο τέλος, δηλ. μόνο για τις ασθενείς που ανήκουν στο δείγμα, εξετάζεται εάν είναι καπνίστριες ή όχι και αναλόγως χωρίζονται σε 2 στρώματα.

Τα δείγματα που προκύπτουν μετά τον διαχωρισμό του συνολικού αρχικού δείγματος παίζουν στη συνέχεια τον ρόλο των επιμέρους δειγμάτων ανά στρώμα που επιλέγονται σε μια κανονική στρωματοποιημένη δειγματοληψία. Με βάση τα δείγματα ανά στρώμα που προέκυψαν, εφαρμόζονται οι τεχνικές εκτίμησης της άγνωστης παραμέτρου του πληθυσμού, σύμφωνα με τη στρωματοποιημένη δειγματοληψία. Για το σκοπό αυτό, απαιτείται μόνο να είναι γνωστό το μέγεθος του κάθε στρώματος για τον πληθυσμό. Για το παράδειγμα του ποσοστού των πρόωρων τοκετών, είναι απαραίτητο να γνωρίζουμε για το σύνολο των ασθενών, ποια είναι η αναλογία των καπνιστών σε σχέση με τις μη καπνίστριες. Μια τέτοια πληροφορία είναι πιο εύκολο να εξαχθεί για το Νοσοκομείο χωρίς να απαιτεί τη μία προς μία εξέταση των καρτελών των ασθενών. Στην πράξη, η πληροφορία αυτή είναι συνήθως γνωστή εκ των προτέρων. Π.χ. το νοσοκομείο γνωρίζει εκ των προτέρων τα ποσοστά των καπνιστών ασθενών του, σε σχέση με τα ποσοστά των μη-καπνιστών.

### 4.3.1 Εκτίμηση μέσης τιμής πληθυσμού για την εκ των υστέρων στρωματοποιημένη δειγματοληψία

Το πρόβλημα της εκτίμησης και των ιδιοτήτων του εκτιμητή παρουσιάζεται, ακολουθώντας τους Cochran (1977, Κεφ. 5) και Levy & Lemeshow (1999, Κεφ. 6), για την περίπτωση του συνεχούς χαρακτηριστικού $Y$ για τον πληθυσμό και στην αναζήτηση της εκτίμησης της πραγματικής μέσης τιμής $\bar{Y}$ από ένα δείγμα μεγέθους $n$. Ανάλογα αντιμετωπίζονται τα προβλήματα της εκτίμησης συνόλου, και του ποσοστού $P$ ή του συνολικού αριθμού $A$, για την περίπτωση δίτιμου χαρακτηριστικού. Ένα απλό τυχαίο δείγμα μεγέθους $n$ επιλέγεται από τον πληθυσμό και, βάσει του κριτηρίου που χρησιμοποιείται για την εκ των υστέρων στρωματοποίηση, οι $n$ δειγματοληπτικές μονάδες χωρίζονται στα $L$ στρώματα.

Συμβολίζουμε με:

$$m_1, m_2, \dots, m_L$$

το πλήθος των μονάδων του δείγματος που ανήκουν στο καθένα από τα $L$ στρώματα. Προφανώς ισχύει $m_1 + m_2 + \cdots + m_L = n$. Ο εκτιμητής της μέσης τιμής για την εκ των υστέρων στρωματοποιημένη δειγματοληψία, $\bar{X}_{pst}$, δίνεται από τον ίδιο τύπο, όπως και στην τυχαία στρωματοποιημένη. Αναλυτικά,

$$\bar{X}_{pst} = \sum_{i=1}^{L} W_h \bar{X}_h$$

όπου $W_h$ είναι τα βάρη των στρωμάτων στον πληθυσμό.

Στην εκ των υστέρων στρωματοποιημένη, ο ερευνητής απλώς διαπιστώνει τα μεγέθη του δείγματος $m_h$ ανά στρώμα $h$ και δεν υπάρχει κάποιος έλεγχος ή εκ των προτέρων περιορισμός ως προς τις τιμές τους. Στην πραγματικότητα τα μεγέθη $m_h, h = 1,2,..L$ είναι τυχαίες μεταβλητές και η διαφορά αυτή σε σχέση με τα $n_h, h = 1,2,..L$ που είναι σταθεροί προκαθορισμένοι από τον ερευνητή αριθμοί για τη στρωματοποιημένη δειγματοληψία, έχει ως συνέπεια η αποτελεσματικότητα της εκ των υστέρων στρωματοποιημένης να υπολείπεται εκείνη της τυχαίας στρωματοποιημένης.

Πιο συγκεκριμένα, μπορεί να αποδειχθεί ότι η διακύμανση του εκτιμητή της μέσης τιμής του πληθυσμού που προκύπτει με τη στρωματοποίηση μετά τη δειγματοληψία είναι:

$$\mathrm{Var}(\bar{X}) = \frac{1-f}{n} \sum_{h=1}^{L} W_h S_h^2 + \frac{1}{n^2} \sum_{h=1}^{L} (1 - W_h)\, S_h^2 \qquad (4.2)$$

Η τελευταία σχέση δηλώνει ότι αν η στρωματοποίηση γίνει μετά τη δειγματοληψία, η διακύμανση του εκτιμητή επιβαρύνεται κατά τον όρο $\frac{1}{n^2}\sum_{h=1}^{L}(1 - W_h)S_h^2$ σε σχέση με τη διακύμανση του εκτιμητή εάν εφαρμοζόταν τυχαία στρωματοποιημένη με αναλογικό καταμερισμό του μεγέθους του δείγματος ανά στρώμα (βλ. σχέση (3.6)).

Ανάλογα με τις διακυμάνσεις ανά στρώμα και το μέγεθος του δείγματος, η ποσότητα αυτή ενδέχεται να είναι πολύ μικρή και το όφελος από την εκ των υστέρων στρωματοποιημένη να προσεγγίζει το όφελος της τυχαίας αναλογικής στρωματοποιημένης. Παρόλο που τα δύο δειγματοληπτικά σχήματα έχουν συγκρίσιμα τυπικά σφάλματα, η εκ των υστέρων στρωματοποιημένη δεν απαιτεί τον σχεδιασμό και τον χωρισμό των πληθυσμιακών μονάδων σε στρώματα. Συνεπώς, υπερέχει σε ταχύτητα, κόστος και απαιτούμενα στοιχεία για τον πληθυσμό, άρα από πρακτικής άποψης και υλοποίησης της δειγματοληψίας υπερτερεί έναντι της κανονικής στρωματοποιημένης.

## Παράδειγμα 4.1

Μια έρευνα διεξάγεται σε μια πόλη με σκοπό να εκτιμηθεί το χρηματικό ποσό που κατά μέσο όρο ξοδεύει μία οικογένεια ανά εβδομάδα για ιατρικούς λόγους. Επιλέγονται τυχαία 33 οικογένειες και τα στοιχεία της έρευνας δίνονται στον Πίνακα 4.2. Η δεύτερη στήλη του πίνακα δίνει τον αριθμό των μελών της οικογένειας, μια πληροφορία που έχουμε μόνο μετά την έρευνα και η τρίτη στήλη δίνει τα έξοδα σε χρηματικές μονάδες (χ.μ) που δήλωσε ότι έχει ξοδέψει η οικογένεια την περασμένη εβδομάδα. Για την πόλη, γνωρίζουμε τα ποσοστά των οικογενειών που έχουν 2, 3, 4, 5 και 6 και άνω παιδιά επί του συνόλου των οικογενειών. Τα ποσοστά αυτά είναι 0.2, 0.3, 0.35, 0.1 και 0.05 αντίστοιχα. Πόσο θα μπορούσε να βελτιώσει την εκτίμηση της μέσης τιμής των εξόδων η πληροφορία, έστω εκ των υστέρων, για τον αριθμό των μελών της οικογένειας;

| Οικογένεια | Αριθμός μελών (family) | Ιατρικά Έξοδα (χ.μ) (expenses) | Οικογένεια | Αριθμός μελών (family) | Ιατρικά Έξοδα(χ.μ) (expenses) |
|---|---|---|---|---|---|
| 1 | 2 | 28.6 | 18 | 4 | 72.0 |
| 2 | 3 | 41.6 | 19 | 2 | 21.2 |
| 3 | 3 | 45.4 | 20 | 4 | 55.4 |
| 4 | 5 | 61.0 | 21 | 2 | 51.9 |
| 5 | 4 | 82.4 | 22 | 5 | 46.6 |
| 6 | 7 | 86.4 | 23 | 3 | 79.6 |
| 7 | 2 | 48.4 | 24 | 4 | 33.6 |
| 8 | 4 | 60.0 | 25 | 7 | 75.6 |
| 9 | 2 | 48.4 | 26 | 3 | 69.6 |
| 10 | 5 | 88.8 | 27 | 3 | 57.4 |
| 11 | 3 | 26.8 | 28 | 6 | 106.0 |
| 12 | 6 | 79.6 | 29 | 2 | 39.1 |
| 13 | 4 | 58.8 | 30 | 2 | 43.2 |
| 14 | 4 | 54.2 | 31 | 6 | 76.4 |
| 15 | 2 | 44.2 | 32 | 4 | 40.2 |
| 16 | 5 | 75.4 | 33 | 2 | 41.4 |
| 17 | 3 | 45.2 | **Σύνολα** | **123** | **1884.4** |

***Πίνακας 4.2*** *Δείγμα οικογενειών και ιατρικά έξοδα.*

Α. Χωρίς χρήση στρωματοποίησης, ο εκτιμητής του μέσου ποσού που ξοδεύει μια οικογένεια της πόλης για ιατρικούς λόγους (ανά εβδομάδα) είναι ο απλός δειγματικός μέσος (βλ. Παράγραφο 2.3.2):

$$\bar{X} = \frac{1}{33} 1884.4 = 57.10\ \chi.\mu$$

και το τυπικό της σφάλμα, αν υποθέσουμε τη διόρθωση πεπερασμένου πληθυσμού αμελητέα, είναι (με βάση το Πόρισμα 2.2):

$$\text{se}(\bar{X}) = \sqrt{\frac{1}{33} s^2} = \sqrt{\frac{405.24}{33}} = 3.51\ \chi.\mu$$

Β. Έχοντας την πληροφορία για τον αριθμό των μελών της οικογένειας μετά τη στρωματοποίηση, μπορούμε να χωρίσουμε σε στρώματα τις 33 οικογένειες με τη βοήθεια αυτής της μεταβλητής και να εκτιμηθεί η ίδια ποσότητα εκ νέου, με εφαρμογή τής εκ των υστέρων στρωματοποίησης.

Το πρώτο βήμα είναι να χωριστεί το υπάρχον δείγμα σε στρώματα. Έστω ότι αποφασίζουμε να ορίσουμε τρία στρώματα, ως εξής: Οικογένειες με αριθμό μελών οικογένειας ≤ 2, αριθμό μελών 3-4 και αριθμό μελών≥5. Αν family και expenses είναι οι δύο μεταβλητές (object vectors για την R) όπου έχουν αποθηκευτεί οι τιμές των στοιχείων της έρευνας όπως παρουσιάζονται στον Πίνακα 4.2, οι εντολές στην `R` για τον υπολογισμό των δειγμάτων και των μέσων ανά στρώμα είναι:

```
>sample<-matrix(c(family, expenses), ncol=2)
```
# δημιουργία πίνακα με 2 στήλες: τις μεταβλητές `family` και `expenses`
```
>strata1<-sample[sample[,1]<=2,2]
```
# το πρώτο στρώμα κρατά τα στοιχεία της δεύτερης στήλης του πίνακα για
# εκείνες τις γραμμές που συναντά τιμή στη μεταβλητή `family` μικρότερη
# ή ίση του 2.
```
>strata2<-sample[sample[,1]>2 & sample[,1]<5,2]
>strata3<-sample[sample[,1]>=5,2]
```

Τα τρία δείγματα που προκύπτουν ανά στρώμα είναι

```
> strata1
[1] 28.6 48.4 48.4 44.2 21.2 51.9 39.1 43.2 41.4
> strata2
[1] 41.6 45.4 82.4 60.0 26.8 58.8 54.2 45.2 72.0 55.4 79.6 33.6
69.6 57.4 40.2
> strata3
[1] 61.0 86.4 88.8 79.6 75.4 46.6 75.6 106.0 76.4
```

Οι μέσοι ανά στρώμα εύκολα υπολογίζονται ως:

```
> means<-c(mean(strata1), mean(strata2), mean(strata3))
> means
[1] 40.71111 54.81333 77.31111
```

Από το αποτέλεσμα αυτό, ήδη διαπιστώνεται, όπως ήταν αναμενόμενο, ότι υπάρχει διαφορά στις μέσες τιμές της βασικής μεταβλητής για τα τρία στρώματα. Άρα, αναμένουμε μια μεγάλη βελτίωση στην εκτίμηση της μέσης τιμής των εξόδων κάτω από το στρωματοποιημένο σχήμα δειγματοληψίας, έστω αυτό της εκ των υστέρων στρωματοποιημένης. Πράγματι, λαμβάνοντας υπόψη τα βάρη των στρωμάτων από την πληροφορία που διαθέτουμε για την αναλογία των οικογενειών στην πόλη, η εκτίμηση αρχικά είναι:

```
>meanExpenses<-sum(weights* means)
>meanExpenses
[1] 55.36756
```

και το τυπικό σφάλμα της εκτίμησης, με τη βοήθεια της έκφρασης (4.1), υπολογίζεται ως:

```
>vars<-c(var(strata1), var(strata2), var(strata3))
>sqrt((1/33)*sum(weights*vars)+(1/33^2)*sum((1-weights)*vars))
[1] 2.744094
```

Διαπιστώνουμε ότι το τυπικό σφάλμα της εκτίμησης είναι 2.74 από 3.51 που ήταν όταν δεν έγινε χρήση της στήλης με τον αριθμό των μελών της οικογένειας. Το τυπικό σφάλμα της εκτίμησης είναι αρκετά βελτιωμένο,

χωρίς να αλλάξει το δείγμα των μετρήσεων, παρά μόνον ο τρόπος υπολογισμού της εκτίμησης. Η πληροφορία που χρειάστηκε επιπλέον είναι εύκολο να αποκτηθεί κατά τη διάρκεια της συλλογής των δεδομένων και δεν επιβαρύνει την έρευνα με κόστος και χρόνο. Αξίζει επίσης να σημειωθεί ότι η εκτίμηση της μέσης τιμής των εξόδων για ιατρικούς λόγους δεν αλλάζει σημαντικά για τους δύο τρόπους εκτίμησης. Αυτό είναι αναμενόμενο, γιατί και οι δύο τρόποι δειγματοληψίας οδηγούν σε αμερόληπτο εκτιμητή. Το όφελος είναι στο σκέλος της ακρίβειας του εκτιμητή.

Αν υπολογίσουμε την επίδραση του τυπικού σφάλματος για τις δύο εκτιμήσεις, θα είναι:

$$\text{deff} = \frac{\text{Var}_{pst}(\bar{X})}{\text{Var}_{srs}(\bar{X})} = 0.61$$

δηλ. ο τρόπος εκτίμησης της μέσης τιμής των εξόδων με την εκ των υστέρων στρωματοποιημένη δειγματοληψία βελτιώνει τη διακύμανση του εκτιμητή από την α.τ.δ. κατά 39%. Τέλος, εάν αντί της εκ των υστέρων στρωματοποιημένης εφαρμόζονταν η τυχαία στρωματοποιημένη με αναλογικό καταμερισμό, το τυπικό σφάλμα του εκτιμητή, όπως προκύπτει από το πρώτο άθροισμα στο δεύτερο μέρος της ισότητας (4.2), είναι

```
>sqrt((1/33)*sum(weights*vars))
[1] 2.674118
```

αντί 2.74 για την εκ των υστέρων στρωματοποιημένη. Άρα, για τα δεδομένα του παραδείγματος, και παρόλο που το μέγεθος δείγματος είναι μικρό, το τυπικό σφάλμα του εκτιμητή με την εκ των υστέρων στρωματοποίηση δεν διαφέρει σημαντικά από το αντίστοιχο τυπικό σφάλμα για αναλογική τυχαία στρωματοποιημένη.

## 4.4. Δειγματοληψία με προκαθορισμένα ποσοστά.

Μια μέθοδος δειγματοληψίας αρκετά διαδεδομένη σε έρευνες αγοράς και δημοσκοπήσεις είναι η δειγματοληψία με προκαθορισμένα ποσοστά (**quota sampling**). Η μέθοδος αυτή έχει ομοιότητες με τη στρωματοποιημένη δειγματοληψία, αλλά η βασική διαφορά είναι ότι δεν αποτελεί δειγματοληψία πιθανότητας (βλ. για παράδειγμα Barnett 2002, Κεφ. 4).

Αναλυτικότερα, σύμφωνα με τη δειγματοληψία με προκαθορισμένα ποσοστά, ο ερευνητής επιλέγει τις δειγματοληπτικές μονάδες έτσι, ώστε στο τελικό δείγμα να υπάρχουν εκπρόσωποι από κάθε στρώμα του πληθυσμού και μάλιστα με αναλογία ίση με εκείνη για τον πληθυσμό. Τα στρώματα ορίζονται συνήθως με βάση κάποιο δημογραφικό κριτήριο, π.χ. φύλο, ηλικία, φυλή κτλ.

Εάν ο χωρισμός του πληθυσμού σε στρώματα έχει γίνει πριν από τη διεξαγωγή της έρευνας και η επιλογή των $n_h$ δειγματοληπτικών μονάδων ανά στρώμα γίνει με τυχαίο τρόπο, τότε η μέθοδος ταυτίζεται με την **τυχαία στρωματοποιημένη**.

Στην πράξη, πολλές φορές ο χωρισμός του πληθυσμού σε στρώματα ενδέχεται να έχει μεγάλο κόστος. Ένας ισοδύναμος τρόπος επιλογής του δείγματος με προκαθορισμένα ποσοστά, που δεν απαιτεί τον χωρισμό των μονάδων του πληθυσμού σε στρώματα εκ των προτέρων, είναι όταν οι δειγματοληπτικές μονάδες επιλέγονται τυχαία από τον ενιαίο πληθυσμό μέχρις ότου συμπληρωθεί ο αριθμός $n_h$ για κάθε στρώμα του πληθυσμού. Με τον τρόπο αυτό, η ανάθεση της κάθε δειγματοληπτικής μονάδας σε ένα στρώμα γίνεται μετά την επιλογή της στο δείγμα. Η δειγματοληψία λέγεται **στρωματοποιημένη με τυχαία προκαθορισμένα ποσοστά (stratified with random quotas)**. Το μειονέκτημα της στρωματοποιημένης με τυχαία προκαθορισμένα ποσοστά είναι ότι ενδέχεται να αποβεί χρονοβόρα -και κατά συνέπεια με μεγάλο κόστος- γιατί προς τα τελευταία στάδια της έρευνας, όσοι επιλέγονται από τον πληθυσμό είναι πολύ πιθανό να ανήκουν σε στρώματα των οποίων τα $n_h$ έχουν ήδη καλυφθεί.

Για τη μείωση του κόστους της διεξαγωγής της έρευνας, ο ερευνητής αρκετές φορές στην πράξη δεν επιλέγει τις δειγματοληπτικές μονάδες από τον πληθυσμό τυχαία, αλλά με βάση την προσωπική του κρίση (**υποκειμενική δειγματοληψία**), έτσι ώστε να καλυφθούν τα προκαθορισμένα ποσοστά για το κάθε στρώμα όσο το δυνατό πιο σύντομα. Στην περίπτωση αυτή, η μέθοδος δειγματοληψίας λέγεται **στρωματοποιημένη με προκαθορισμένα ποσοστά (stratified with quota)** και δεν ανήκει στην κατηγορία των μεθόδων δειγματοληψίας με πιθανότητες. Κατά συνέπεια, δεν μπορεί να γίνει χρήση της θεωρίας που έχει αναπτυχθεί

για τη στρωματοποιημένη δειγματοληψία στο Κεφάλαιο 3 - για τον υπολογισμό των δειγματοληπτικών σφαλμάτων των εκτιμητών.

## Βιβλιογραφικές Αναφορές


Baillargeon, S. and Rivest L.-P. (2011). The construction of stratified designs in R with the package stratification. *Survey Methodology*, 37(1), 53-65.

Barnett, V. (2002). *Sample survey: Principles and methods* (3rd Edition). London: Arnold.

Cochran, W. G. (1977). *Sampling techniques* (3rd Edition). New York: John Wiley and Sons.

Kish, L. (1965). *Survey sampling*. New York: John Wiley and Sons.

Levy, P. S. and Lemeshow, S. (1999). *Sampling of populations. Methods and applications* (3rd Edition). New York: John Wiley and Sons.

# Κεφάλαιο 5 - ΣΥΣΤΗΜΑΤΙΚΗ ΔΕΙΓΜΑΤΟΛΗΨΙΑ

**Σύνοψη**

*Στο κεφάλαιο αυτό, εισάγεται ένα ακόμα δειγματοληπτικό σχέδιο, η* ***συστηματική δειγματοληψία (systematic sampling)****.*

*Το βασικό πλεονέκτημα του σχεδίου αυτού είναι η απλότητα και η ευκολία στην υλοποίησή του από τον ερευνητή. Ένα άλλο πλεονέκτημα του συστηματικού δειγματοληπτικού σχεδίου είναι ότι το δείγμα που επιλέγεται με τη μέθοδο αυτή θα είναι ομοιόμορφα κατανεμημένο ως προς τον πληθυσμό (αφού, όπως θα δούμε στη συνέχεια, οι δειγματικές μονάδες στη συστηματική δειγματοληψία ισαπέχουν). Ακόμη και στην α.τ.δ. ενδέχεται οι μονάδες ενός δείγματος να παρουσιάζουν ομαδοποιήσεις. Αυτό είναι αδύνατον να συμβεί στη συστηματική δειγματοληψία.*

*Ένα μειονέκτημα της συστηματικής δειγματοληψίας είναι ότι μπορεί να αλληλεπιδρά με πιθανή λανθάνουσα περιοδικότητα μεταξύ των μονάδων του πληθυσμού (Βλ. Παράγραφο 5.6).*

*Ως προς την ακρίβεια των εκτιμητών που προκύπτουν, τα αποτελέσματα ποικίλλουν, ανάλογα με τη δομή του δειγματοληπτικού πλαισίου ως προς το χαρακτηριστικό της έρευνας. Κάτω από ορισμένες περιπτώσεις πληθυσμού, η συστηματική μπορεί να είναι πιο αποτελεσματική από την απλή τυχαία, ενώ σε άλλες περιπτώσεις η συστηματική είναι λιγότερο αποδοτική. Επίσης, το αποτέλεσμα της σύγκρισης ως προς την αποτελεσματικότητα σε σχέση με την α.τ.δ. μπορεί να αλλάξει ακόμα κι όταν αναφερόμαστε στον ίδιο πληθυσμό, αν χρησιμοποιούμε διαφορετικό δειγματοληπτικό πλαίσιο ή διάταξη των μελών του πληθυσμού. Τα ακριβή χαρακτηριστικά στον τρόπο υλοποίησης, τα πρακτικά πλεονεκτήματα της μεθόδου κατά την εφαρμογή, οι ιδιότητες των εκτιμητών και οι τρόποι βελτίωσης της συστηματικής δειγματοληψίας δίνονται στις επόμενες παραγράφους, όπου περιλαμβάνεται η αναλυτική μελέτη του δειγματοληπτικού αυτού σχεδίου το οποίο, είτε αποκλειστικά, είτε σε συνδυασμό με κάποιο άλλο, είναι αυτό που εφαρμόζεται πιο συχνά στην πράξη.*

**Προαπαιτούμενη γνώση**

*Απλή τυχαία Δειγματοληψία, Στρωματοποιημένη Δειγματοληψία, Εκτιμητική, Διαστήματα Εμπιστοσύνης.*

## 5.1. Περιγραφή – Ορισμός και βασικά χαρακτηριστικά

Όπως δηλώνεται και από την ονομασία της, η συστηματική δειγματοληψία (systematic sampling) είναι η δειγματοληψία κατά την οποία οι μονάδες του δείγματος επιλέγονται από τον πληθυσμό με συστηματικό τρόπο, δηλαδή ανά ίσα, διαδοχικά διαστήματα. Ως μήκος κάθε διαστήματος ορίζεται η απόσταση στη σειρά κατάταξης ή καταγραφής που έχουν δύο μονάδες του πληθυσμού σύμφωνα με το δειγματοληπτικό πλαίσιο που χρησιμοποιούμε.

Για παράδειγμα, έστω ότι ο πληθυσμός είναι τα $N = 30$ παιδιά μιας τάξης ενός Δημοτικού σχολείου και έστω ακόμα ότι επιθυμούμε να επιλέξουμε ένα δείγμα μεγέθους $n = 6$ των μαθητών της τάξης αυτής. Εάν συνταχθεί μια λίστα των μαθητών με βάση την αλφαβητική σειρά, τότε το δειγματοληπτικό πλαίσιο βάσει του οποίου θα διεξαχθεί η δειγματοληψία είναι η ανωτέρω λίστα. Αντιστοιχίζουμε έναν αύξοντα αριθμό, από 1 έως 30, σε κάθε μαθητή, ανάλογα με τη θέση που κατέχει στη λίστα, και ο πληθυσμός μπορεί να παρασταθεί ως το διάνυσμα:

$$\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30\}.$$

Το δείγμα των μαθητών που αποτελείται για παράδειγμα από τους μαθητές με α/α:

$$\{2, 7, 12, 17, 22, 27\}$$

είναι ένα συστηματικό δείγμα, μιας και οι έξι μαθητές που επιλέχθηκαν έχουν όλοι ίση απόσταση μεταξύ τους, συγκεκριμένα πέντε μονάδες, σύμφωνα με τη σειρά κατάταξης στο πλαίσιο που χρησιμοποιήσαμε.

Αντίστοιχα, το δείγμα $\{4, 9, 14, 19, 24, 29\}$ είναι επίσης ένα συστηματικό δείγμα. Συνεπώς, για το παράδειγμα αυτό, ένα συστηματικό δείγμα μεγέθους 6 μπορεί να προκύψει επιλέγοντας τις δειγματοληπτικές μονάδες με απόσταση μεταξύ τους πέντε, που αποκαλείται επίσης και **βήμα (step)** της συστηματικής. Η απόσταση αυτή είναι άμεσα συνδεδεμένη με τον τρόπο τοποθέτησης των μαθητών σε λίστα, π.χ. εάν αντί της αλφαβητικής σειράς οι μαθητές μπουν σε σειρά με βάση το ύψος τους, και μάλιστα από το χαμηλότερο ως το υψηλότερο, τότε το συστηματικό δείγμα που θα προκύψει θα περιλαμβάνει μαθητές που απέχουν μεταξύ τους πέντε μονάδες στη σειρά κατάταξης κατά ύψος, αντί στην αλφαβητική σειρά κατάταξης. Τέλος, η αριθμητική τιμή του βήματος, πέντε για το παράδειγμά μας, συνδέεται με το επιθυμητό τελικό μέγεθος του δείγματος. Για ένα μεγαλύτερο δείγμα το βήμα θα ήταν μικρότερο και αντίστροφα. Για παράδειγμα αν $n = 10$ τότε το βήμα είναι 3 και ένα συστηματικό δείγμα θα ήταν το $\{2, 5, 8, 11, 14, 17, 20, 23, 26, 29\}$. Το βήμα σε κάθε περίπτωση καθορίζεται έτσι ώστε:

*«το τελικό δείγμα να καλύπτει ομοιόμορφα, ανά ίση απόσταση, όλο το εύρος του πληθυσμού.»*

Αυτή είναι και η ιδιότητα που χαρακτηρίζει ένα συστηματικό δείγμα.

Η ίδια αυτή ιδιότητα της συστηματικής δειγματοληψίας, η οποία φαινομενικά δείχνει να είναι μια θετική ιδιότητα και συντελεί στην επιλογή ενός αντιπροσωπευτικού δείγματος, είναι και ο λόγος που ορισμένες φορές η συστηματική δειγματοληψία οδηγεί σε ένα κακό δείγμα και κατά συνέπεια σε κακή ή λανθασμένη εκτίμηση. Αυτό συμβαίνει όταν οι μονάδες του πληθυσμού εμφανίζουν κάποια περιοδικότητα. Για παράδειγμα, έστω ότι έχει παρατηρηθεί πως οι εισπράξεις ενός καταστήματος είναι πιο αυξημένες στην αρχή του μήνα και πιο μειωμένες προς το τέλος του μήνα. Αν θέλουμε να κάνουμε μια συστηματική δειγματοληψία επιλέγοντας κάποιες ημέρες μέσα στη διάρκεια ενός έτους με σκοπό να εκτιμήσουμε τις εισπράξεις του καταστήματος, τότε αν οι παρατηρήσεις γίνουν: 5 Ιανουαρίου, 5 Φεβρουαρίου, 5 Μαρτίου κοκ, δηλ. κάθε αρχή του μήνα, η εκτίμησή μας θα υπερεκτιμά τις αληθινές εισπράξεις του καταστήματος. Τον ρόλο που παίζει του παίζει το κριτήριο κατάταξης των μονάδων του πληθυσμού και η ενδεχόμενη περιοδικότητα θα τα μελετήσουμε αναλυτικότερα στην Παράγραφο 5.6.

Συνοψίζοντας, μπορούμε να δώσουμε την περιγραφή υλοποίησης της συστηματικής δειγματοληψίας στη γενική της περίπτωση. Η διαδικασία αυτή περιλαμβάνει τις φάσεις:

Φάση 1: Κατασκευή ή υιοθέτηση του **δειγματοληπτικού πλαισίου** του πληθυσμού. Έστω $\{Y_1, Y_2, \dots, Y_N\}$ το διάνυσμα του πληθυσμού κάτω από το δειγματοληπτικό πλαίσιο.

Φάση 2: Υπολογισμός του **βήματος** της συστηματικής δειγματοληψίας. Το βήμα συνήθως συμβολίζεται με $k$ και υπολογίζεται από το κλάσμα $k = \frac{N}{n}$.

Φάση 3: Επιλέγεται ένας τυχαίος αριθμός, έστω $j$ $(1 \leq j \leq k)$. Ο αριθμός αυτός, καθώς και η μονάδα του πληθυσμού που αντιστοιχεί στον αριθμό αυτό, λέγεται και **αφετηρία** της συστηματικής.

Φάση 4: Το συστηματικό δείγμα είναι το διάνυσμα $s$:

$$s = \{Y_j, Y_{j+k}, Y_{j+2k}, \dots, Y_{j+(n-1)k}\} \tag{5.1}$$

## Παρατηρήσεις

**5.1** Το βήμα $k$ όπως ορίστηκε στο Στάδιο 2 της διαδικασίας υλοποίησης της συστηματικής δειγματοληψίας, θα είναι ακέραιος αριθμός στην περίπτωση που το $N$ διαιρείται ακριβώς από το $n$. Σε κάθε άλλη περίπτωση το κλάσμα $\frac{N}{n}$ θα έχει ως αποτέλεσμα ένα δεκαδικό αριθμό. Σε αυτές τις περιπτώσεις, ως βήμα $k$ λαμβάνεται ο πλησιέστερος ακέραιος αριθμός, με τις αναπόφευκτες επιπτώσεις

στο μέγεθος του δείγματος. Π.χ. αν $N = 20$ και $n = 3$, τότε $\frac{N}{n} = 6.67$, οπότε $k = 7$, και οι μονάδες θα επιλέγονται ανά απόσταση 7. Αν $N = 20$ και $n = 9$, τότε $\frac{N}{n} = 2.22$ και στην περίπτωση αυτή υιοθετούμε $k = 2$.

**5.2** Για τη γενική περίπτωση που ο $\frac{N}{n}$ δεν είναι ακέραιος, το πραγματικό μέγεθος του δείγματος είναι, ανάλογα με τη στρογγύλευση, μεγαλύτερο ή μικρότερο του $n$. Η γενική μορφή του συστηματικού δείγματος (5.1), στην περίπτωση αυτή, γράφεται ως:

$$s = \{Y_j, Y_{j+k}, Y_{j+2k}, Y_{j+3k}, \dots\}$$

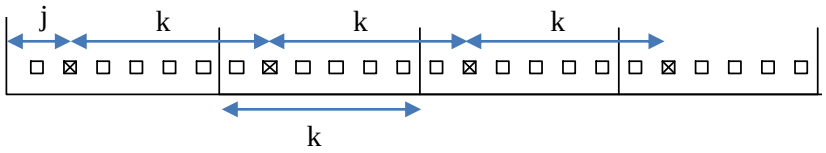
**5.3** Σύμφωνα με τη συστηματική δειγματοληψία, η μόνη επιλογή μονάδας που γίνεται με τυχαίο τρόπο είναι η πρώτη. Όλες οι υπόλοιπες μονάδες από τη 2η ως και τη n-οστή είναι πλήρως καθορισμένες και εξαρτώνται από την 1η μονάδα. Συγκεκριμένα, είναι οι μονάδες που απέχουν από την αρχική μονάδα όλα τα διαδοχικά πολλαπλάσια του βήματος, μέχρι την εξάντληση του πληθυσμού. Αυτή η διαδικασία επιλογής προγραμματίζεται εύκολα, κάτι που αποτελεί ένα από τα βασικότερα πλεονεκτήματα της συστηματικής δειγματοληψίας.

**5.4** Το σύνολο των δυνατών δειγμάτων, $\mathcal{S}$, κάτω από τη συστηματική δειγματοληψία περιλαμβάνει $k$ σε πλήθος δυνατά δείγματα (αριθμός αρκετά μικρότερος από το $\binom{N}{n}$ που αντιστοιχεί στην περίπτωση της α.τ.δ. χωρίς επανατοποθέτηση). Τα δυνατά δείγματα είναι ακριβώς τόσα, όσες και οι δυνατές αφετηρίες, δηλ. $1,2, \dots, k$.

**5.5** Τα $k$ δυνατά δείγματα του $S$ ονομάζονται και **συστηματικά δείγματα.**

**5.6** Τα δυνατά δείγματα στη συστηματική δειγματοληψία δεν έχουν κανένα κοινό στοιχείο μεταξύ τους.

Σχηματικά, ένα συστηματικό δείγμα μπορεί να παρασταθεί όπως στο Σχήμα 5.1. Έστω ότι οι μονάδες του πληθυσμού τοποθετούνται σε σειρά και παριστώνται σε μια ευθεία λαμβάνοντας θέσεις 1 έως $N$. Το διάστημα αυτό χωρίζεται στη συνέχεια σε $n$ ίσα υπο-διαστήματα μήκους $k$ το καθένα. Για το πρώτο διάστημα που περιέχει τις μονάδες του πληθυσμού από το 1 έως το $k$, επιλέγουμε έναν τυχαίο αριθμό, έστω $j$. Οι υπόλοιπες μονάδες του δείγματος είναι οι μονάδες που καταλαμβάνουν ακριβώς την ίδια θέση που καταλαμβάνει το $j$ στο 1ο υπο-διάστημα για όλα τα υπόλοιπα υποδιαστήματα μέχρι και την εξάντληση του πληθυσμού.




***Σχήμα 5.1*** *Γραφική αναπαράσταση συστηματικής δειγματοληψίας*

Αν χρησιμοποιήσουμε το πακέτο '`animation`' της `R`, μπορούμε να έχουμε μια οπτική αναπαράσταση της συστηματικής δειγματοληψίας μέσω των εντολών:

```
> library(animation)
> sample.system()
```

Για σύγκριση, στην α.τ.δ. η αντίστοιχη εντολή είναι:

```
> sample.simple(nrow=10, ncol=10, size=15, p.col=c("blue", "red"),
p.cex = c(1,3))
```

**Ορισμός**

Συστηματική δειγματοληψία είναι το δειγματοληπτικό εκείνο σχέδιο, στο οποίο το σύνολο των δυνατών δειγμάτων $\mathcal{S}$ αποτελείται από τα δείγματα $s = \{Y_j, Y_{j+k}, Y_{j+2k}, Y_{j+3k}, \dots\}$ για $k = N/n$ και $j \in \{1,2,\dots,k\}$, τα οποία είναι ισοπίθανα μεταξύ τους.

## Παρατήρηση 5.7

Εάν αντί της τυχαίας επιλογής ενός αριθμού $j$ $(1 \leq j \leq k)$ για την αφετηρία ενός συστηματικού δείγματος επιλεγεί το κεντρικό σημείο του 1ου διαστήματος μήκους $k$, η δειγματοληψία που προκύπτει ονομάζεται ειδικότερα **κεντρικά τοποθετημένη συστηματική δειγματοληψία (centrally located systematic sampling)**. Στην περίπτωση της τυχαίας επιλογής, η δειγματοληψία ονομάζεται **τυχαία συστηματική δειγματοληψία (random systematic sampling)**.

Σύμφωνα με τον τρόπο περιγραφής και τα βήματα υλοποίησης της δειγματοληψίας όπως δόθηκαν παραπάνω, η ύπαρξη του δειγματοληπτικού πλαισίου είναι απαραίτητη προϋπόθεση. Στην πράξη, η προϋπόθεση αυτή δεν είναι πάντα απαραίτητη. Αποτελεί δε το γεγονός αυτό ένα ακόμη πλεονεκτήματα της συστηματικής δειγματοληψίας έναντι άλλων μεθόδων. Για παράδειγμα, έστω ότι για λόγους ποιοτικού ελέγχου και υπολογισμού των εξόδων, μια κλινική θέλει να πραγματοποιεί μια διαρκή έρευνα στους ιατρικούς φακέλους των ασθενών της. Επειδή το συνολικό πλήθος των φακέλων δεν είναι γνωστό, αφού η έρευνα είναι διαρκής και οι φάκελοι ανανεώνονται, δεν είναι εφικτό να δημιουργηθεί μια λίστα και να εφαρμοστεί για παράδειγμα η απλή τυχαία δειγματοληψία. Παρόλα αυτά, μια προσέγγιση του συνολικού πλήθους είναι πιθανότατα εφικτή για το χρονικό εύρος της έρευνας και αυτό είναι αρκετό για τον υπολογισμό του $k$. Έχοντας υπολογίσει το $k$, μια συστηματική δειγματοληψία είναι εφικτή, μη έχοντας την πλήρη καταγραφή της λίστας του πληθυσμού, επιλέγοντας 1 φάκελο για την έρευνα ανά $k$ φακέλους ασθενών.

Στη συνέχεια του κεφαλαίου, ποσότητες με δείκτη sy θα αναφέρονται στη συστηματική δειγματοληψία. Προτεινόμενη βιβλιογραφία στο αντικείμενο της συστηματικής δειγματοληψίας είναι μεταξύ άλλων: Cochran (1977, κεφ. 8), Lohr (2010, κεφ. 5), Levy and Lemeshow (1999, κεφ. 4), Barnett (2002, παρ. 5.4) και Rao, (2000, κεφ. 7).

## 5.2. Εκτίμηση παραμέτρων του πληθυσμού και ιδιότητες εκτιμητών κάτω από τη συστηματική δειγματοληψία

Θεωρούμε την περίπτωση της εκτίμησης του πληθυσμιακού μέσου $\bar{Y}$ ενός πληθυσμού μεγέθους $N$ με τη βοήθεια ενός δείγματος μεγέθους $n$, που έχει επιλεγεί από τον πληθυσμό σύμφωνα με τη συστηματική δειγματοληψία.

Έστω $(X_1, X_2, \dots, X_n)$ το συστηματικό δείγμα που έχει επιλεγεί από τον πληθυσμό. Λόγω της συστηματικής επιλογής, θα ισχύει (όταν $N = kn$):

$$(X_1, X_2, \dots, X_n) = (Y_j, Y_{j+k}, \dots, Y_{j+(n-1)k})$$

για ένα τυχαίο $j$ μεταξύ του 1 και του $k$. Ο εκτιμητής του $\bar{Y}$ για τη συστηματική δειγματοληψία, τον οποίο θα συμβολίσουμε με $\bar{X}_{sy}$, είναι ο δειγματικός μέσος $\bar{X}$, δηλ.

$$\bar{X}_{sy} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Για τον εκτιμητή $\bar{X}_{sy}$ αποδεικνύεται η παρακάτω πρόταση, που αναφέρεται στην ιδιότητα της αμεροληψίας του εκτιμητή.

**Πρόταση 5.1**

Ο εκτιμητής $\bar{X}_{sy}$ για τη συστηματική δειγματοληψία είναι αμερόληπτος εκτιμητής του πληθυσμιακού μέσου $\bar{Y}$, όταν το μέγεθος του πληθυσμού $N$ διαιρείται ακριβώς με το μέγεθος του δείγματος $n$.

**Απόδειξη**

Ο $\bar{X}_{sy}$ θα είναι αμερόληπτος εκτιμητής του $\bar{Y}$, εάν και μόνον εάν $\mathrm{E}(\bar{X}_{sy}) = \bar{Y}$. Σύμφωνα με τον ορισμό, η αναμενόμενη τιμή $\mathrm{E}(\bar{X}_{sy})$ υπολογίζεται από τη σχέση:

$$\mathrm{E}(\bar{X}_{sy}) = \sum_{s \in S} \bar{X}_{sy}^{(s)} \pi(s)$$

όπου $s$ είναι ένα από τα δυνατά συστηματικά δείγματα του $S$, $\bar{X}_{sy}^{(s)}$ η τιμή του εκτιμητή $\bar{X}_{sy}$ για το δείγμα $s$ και $\pi(s)$ η πιθανότητα επιλογής του $s$.

Από τον ορισμό της συστηματικής δειγματοληψίας, προκύπτει ότι $\pi(s) = \frac{1}{k}$, και στην περίπτωση που το $N$ διαιρείται ακριβώς με το $n$, δηλ. $N = nk$, με $k$ ακέραιο, τότε όλα τα δυνατά συστηματικά δείγματα περιέχουν ακριβώς ίδιο αριθμό δειγματοληπτικών μονάδων, ίσο με $n$. Αναλυτικότερα, τα $k$ συστηματικά δείγματα είναι:

$s_1 = (Y_1, Y_{1+k}, \dots, Y_{1+(n-1)k})$ για τον τυχαίο αριθμό $j = 1$ με δειγματικό μέσο $\bar{X}_{sy}^{(s_1)} = \frac{1}{n}\sum_{i=1}^{n} Y_{1+(i-1)k}$

$s_2 = (Y_2, Y_{2+k}, \dots, Y_{2+(n-1)k})$ για τον τυχαίο αριθμό $j = 2$ με δειγματικό μέσο $\bar{X}_{sy}^{(s_2)} = \frac{1}{n}\sum_{i=1}^{n} Y_{2+(i-1)k}$

…

$s_k = (Y_k, Y_{2k}, \dots, Y_{nk})$ για τον τυχαίο αριθμό $j = k$ με δειγματικό μέσο $\bar{X}_{sy}^{(s_k)} = \frac{1}{n}\sum_{i=1}^{n} Y_{ik}$

Αντικαθιστώντας τις πιθανότητες $\pi(s)$ και τις εκφράσεις των μέσων των συστηματικών δειγμάτων στον ορισμό της αναμενόμενης τιμής, θα έχουμε:

$$\mathrm{E}(\bar{X}_{sy}) = \sum_{s \in S} \bar{X}_{sy}^{(s)} \pi(s) = \sum_{j=1}^{k} \frac{1}{k} \bar{X}_{sy}^{(s_j)}$$

$$= \sum_{j=1}^{k} \frac{n}{N} \frac{1}{n} \sum_{i=1}^{n} Y_{j+(i-1)k} = \frac{1}{N} \sum_{j=1}^{k} \sum_{i=1}^{n} Y_{j+(i-1)k} = \bar{Y}$$

Η τελευταία ισότητα ισχύει γιατί το τελικό διπλό άθροισμα ως προς $j$ και $i$ εξαντλεί όλα τα στοιχεία του πληθυσμού, χωρίς επαναλήψεις, δηλ. καταμετρά μία φορά ακριβώς την κάθε παρατήρηση και άρα ισούται με το σύνολο του πληθυσμού ■

Στην περίπτωση που $N \neq nk$, δηλ. το κλάσμα $\frac{N}{n}$ δεν είναι ακέραιος αριθμός, ο εκτιμητής $\bar{X}_{sy}$ δεν είναι αμερόληπτος, αλλά το ποσό μεροληψίας για μεγάλα $N$, $n$ είναι αμελητέο. Για μικρές τιμές των $N$, $n$ και στην περίπτωση που το δειγματοληπτικό πλαίσιο είναι διαθέσιμο, έχουν προταθεί δύο δυνατές τροποποιήσεις κατά την εφαρμογή της συστηματικής δειγματοληψίας, που εξασφαλίζουν την αμεροληψία του $\bar{X}_{sy}$.

**1η μέθοδος: Συμπλήρωση του πληθυσμού**

Σύμφωνα με τη μέθοδο αυτή, εάν $N < nk$ όπου $k$ ο μικρότερος ακέραιος έτσι ώστε το γινόμενο $nk$ να υπερβαίνει το $N$, τότε το διάνυσμα του πληθυσμού $(Y_1, Y_2, \dots, Y_N)$ συμπληρώνεται στο τέλος με τόσες μονάδες, όσες είναι απαραίτητες έτσι ώστε $N = nk$. Οι μονάδες που προστίθενται είναι οι αρχικές μονάδες του πληθυσμού, δηλ. ξεκινώντας από $Y_1, Y_2$, κοκ. Με τον τρόπο αυτό, το μέγεθος του πληθυσμού γίνεται

τελικά πολλαπλάσιο του $n$ και ο εκτιμητής $\bar{X}_{sy}$, σύμφωνα με την Πρόταση 5.1, είναι αμερόληπτος. Για παράδειγμα, εάν $N = 10$ και $n = 4$, τότε για $k = 3$ θεωρούμε τον πληθυσμό που συμπληρώνεται μέχρι και την $nk(= 12η)$ θέση δανειζόμενοι τις δύο πρώτες τιμές του πληθυσμού. Δηλ. ο πληθυσμός θα αποτελείται από τις $N^{'} = 16$ τιμές $(Y_1, Y_2, \dots, Y_9, Y_{10}, Y_1, Y_2)$. Τα τρία συστηματικά δείγματα που μπορούν να προκύψουν είναι:

$$(Y_1, Y_4, Y_7, Y_{10}) \text{ για } i = 1$$

$$(Y_2, Y_5, Y_8, Y_1) \text{ για } i = 2$$

$$(Y_3, Y_6, Y_9, Y_2) \text{ για } i = 3$$

και έχουν όλα ίσο μέγεθος δείγματος $n = 4$.

**2η μέθοδος: Άνισες πιθανότητες επιλογής της αφετηρίας του συστηματικού**

Στη μέθοδο αυτή, αντί ο αριθμός $j$ $(1 \le j \le k)$ να επιλέγεται με τυχαίο τρόπο, δηλαδή επιλογή ενός αριθμού από το 1 εως $k$ με πιθανότητα $1/k$, η επιλογή γίνεται με άνισες πιθανότητες, οι οποίες είναι σε συμφωνία με το μέγεθος του συστηματικού δείγματος που προκύπτει για το κάθε $j$. Για το ίδιο μικρό παράδειγμα που χρησιμοποιήσαμε στην 1η μέθοδο, τα τρία δυνατά δείγματα είναι:

$$(Y_1, Y_4, Y_7, Y_{10}) \text{ για } i = 1$$

$$(Y_2, Y_5, Y_8) \text{ για } i = 2$$

$$(Y_3, Y_6, Y_9) \text{ για } i = 3$$

Τα πρώτο δείγμα έχει 4 στοιχεία, ενώ το δεύτερο και το τρίτο έχουν 3. Οι πιθανότητες επιλογής συνεπώς των $\{1,2,3\}$ ως αφετηρίας, θα είναι $\left\{\frac{4}{10}, \frac{3}{10}, \frac{3}{10}\right\}$. Με τον τρόπο αυτό, εξασφαλίζεται επίσης η αμεροληψία του $\bar{X}_{sy}$. Πράγματι, σύμφωνα με τον ορισμό της αναμενόμενης τιμής ενός εκτιμητή που δώσαμε στο Κεφάλαιο 1 -, η αναμενόμενη τιμή του $\bar{X}_{sy}$ θα είναι:

$$\mathrm{E}(\bar{X}_{sy}) = \sum_{s \in \mathcal{S}} \pi(s)\left(\bar{X}_{sy}\right)_s$$

όπου $\mathcal{S}$ το σύνολο όλων των δυνατών δειγμάτων σύμφωνα με τον τρόπο δειγματοληψίας, $\pi(s)$ η πιθανότητα επιλογής του καθενός από τα δυνατά δείγματα $s$ του συνόλου $\mathcal{S}$ και $\left(\bar{X}_{sy}\right)_s$ η τιμή του εκτιμητή για το δείγμα $s$.

Για το παράδειγμά μας:

$$\mathrm{E}(\bar{X}_{sy}) = \frac{4}{10}\left[\frac{1}{4}(Y_1 + Y_4 + Y_7 + Y_{10})\right] + \frac{3}{10}\left[\frac{1}{3}(Y_2 + Y_5 + Y_8)\right] + \frac{3}{10}\left[\frac{1}{3}(Y_3 + Y_6 + Y_9)\right] = \bar{Y}$$

Για τον υπολογισμό της διακύμανσης και του τυπικού σφάλματος του εκτιμητή, αποδεικνύεται η παρακάτω πρόταση.

## Πρόταση 5.2

Η διακύμανση του εκτιμητή $\bar{X}_{sy}$ του πληθυσμιακού μέσου κάτω από τη συστηματική δειγματοληψία δίνεται από τη σχέση:

$$\mathrm{Var}\left(\bar{X}_{sy}\right) = \frac{N-1}{N}S^2 - \frac{k(n-1)}{N}S_{wsy}^2 \qquad (5.2)$$

όπου $S^2_{wsy} = \frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_i)^2$, και $\bar{Y}_i$ ο μέσος του $i$ συστηματικού δείγματος.

### Απόδειξη

Σύμφωνα με τον συμβολισμό της Πρότασης 5.1, οι $k$ δυνατές πραγματοποιήσεις του εκτιμητή $\bar{X}_{sy}$, είναι οι μέσες τιμές $\bar{X}_{sy}^{(s_1)}$, $\bar{X}_{sy}^{(s_2)}$, ..., $\bar{X}_{sy}^{(s_k)}$ ή, ισοδύναμα, τα $\bar{Y}_1$, $\bar{Y}_2$, ..., $\bar{Y}_k$ όπως ορίστηκαν παραπάνω.

Η διακύμανση του εκτιμητή $\bar{X}_{sy}$ ως προς την ποσότητα $\bar{Y}$ που εκτιμά είναι η αναμενόμενη τιμή $\mathrm{E}\left(\bar{X}_{sy} - \bar{Y}\right)^2$, η οποία, σύμφωνα με τον ορισμό στο Κεφάλαιο 1 -, υπολογίζεται αποκλειστικά με τη βοήθεια του συνόλου $S$ των δυνατών δειγμάτων και των πιθανοτήτων $\pi(s), s \in S$. Σύμφωνα με τη συστηματική δειγματοληψία και υποθέτοντας ότι $N = nk$, η ανωτέρω αναμενόμενη τιμή ισούται με:

$$\mathrm{Var}\left(\bar{X}_{sy}\right) = \frac{1}{k} \sum_{i=1}^{k} \left(\bar{X}_{sy}^{(s_i)} - \bar{Y}\right)^2 \tag{5.3}$$

Παράλληλα, για το διάνυσμα του πληθυσμού ισχύει η επόμενη σχέση, ανάλογη της οποίας έχουμε δει και για τη στρωματοποιημένη δειγματοληψία και η οποία είναι γνωστή και ως ΑΝΑλυση της ΔΙΑκύμανσης, σε συντομογραφία ΑΝΑΔΙΑ:

$$(N-1)S^2 = \sum_{i=1}^{k} \sum_{j=1}^{n} \left(Y_{ij} - \bar{Y}_j\right)^2 + n \sum_{i=1}^{k} (\bar{Y}_i - \bar{Y})^2$$

Διαιρώντας την τελευταία σχέση με $N$ και λαμβάνοντας υπόψη τον συμβολισμό της εκφώνησης της Πρότασης, προκύπτει ισοδύναμα:

$$\frac{(N-1)}{N} S^2 = \frac{1}{N} k(n-1) S^2_{wsy} + \frac{n}{N} \sum_{i=1}^{k} (\bar{Y}_i - \bar{Y})^2$$

η οποία, με τη βοήθεια της (5.3), δίνει ισοδύναμα:

$$\frac{(N-1)}{N} S^2 = \frac{k(n-1)}{N} S^2_{wsy} + \mathrm{Var}\left(\bar{X}_{sy}\right)$$

που αποδεικνύει τη ζητούμενη έκφραση ■

### Πόρισμα 5.1

Το τυπικό σφάλμα του εκτιμητή $\bar{X}_{sy}$ στη συστηματική δειγματοληψία, δίνεται από τη σχέση:

$$\mathrm{se}\left(\bar{X}_{sy}\right) = \sqrt{\frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S^2_{wsy}}$$

Ακολουθεί με συγκεκριμένα στοιχεία το παράδειγμα κατανόησης με την τάξη των 30 μαθητών, που δόθηκε στην αρχή της παραγράφου για την εισαγωγή της συστηματικής δειγματοληψίας.

## Παράδειγμα 5.1

Από το σύνολο των μαθητών μιας τάξης 30 ατόμων συνολικά, ενδιαφερόμαστε να επιλέξουμε ένα συστηματικό δείγμα μεγέθους 6, με σκοπό την εκτίμηση του μέσου όρου επίδοσης των μαθητών της τάξης

αυτής στο μάθημα των Μαθηματικών. Η σειρά κατάταξης των μαθητών είναι αλφαβητική και οι βαθμοί όλης της τάξης δίνονται στον Πίνακα 5.1 που ακολουθεί. Οι βαθμολογίες είναι με κλίμακα 0-100.

| α/α | Βαθμολογία στα Μαθηματικά | α/α | Βαθμολογία στα Μαθηματικά | α/α | Βαθμολογία στα Μαθηματικά |
|---|---|---|---|---|---|
| 1 | 20 | 11 | 34 | 21 | 18 |
| 2 | 5 | 12 | 7 | 22 | 31 |
| 3 | 42 | 13 | 54 | 23 | 70 |
| 4 | 56 | 14 | 67 | 24 | 80 |
| 5 | 76 | 15 | 86 | 25 | 25 |
| 6 | 20 | 16 | 16 | 26 | 73 |
| 7 | 27 | 17 | 70 | 27 | 83 |
| 8 | 65 | 18 | 9 | 28 | 25 |
| 9 | 97 | 19 | 40 | 29 | 13 |
| 10 | 44 | 20 | 79 | 30 | 16 |

***Πίνακας 5.1*** *Βαθμολογίες των μαθητών στα Μαθηματικά*

Για δείγμα μεγέθους $n = 6$, η επιλογή του δείγματος με τη βοήθεια της R γίνεται ως εξής:

Έστω x το διάνυσμα των 30 τιμών του πληθυσμού:

```
> x
 [1] 20  5 42 56 76 20 27 65 97 44 34  7 54 67 86 16 70  9 40
[20] 79 18 31 70 80 25 73 83 25 13 16
```

Στη συνέχεια υπολογίζεται το $k$:

```
> n<-6
> k<-length(x)/n
> k
[1] 5
```

και επιλέγεται η αφετηρία με τη βοήθεια της εντολής:

```
j<-sample(1:k,1)
> j
[1] 4
```

Άρα επιλέγεται το 4ο συστηματικό δείγμα, το οποίο συγκεκριμένα αποτελείται από τους μαθητές με αύξοντες αριθμούς:

```
> seq(from=j, to=length(x), by=k)
[1]  4  9 14 19 24 29
```

Από αυτό το δείγμα των μαθητών, το αντίστοιχο δείγμα των βαθμολογιών προκύπτει να είναι:

```
> s<-x[seq(from=j, to=length(x), by=k)]
> s
[1] 56 97 67 40 80 13
```

Αν ζητήσουμε  τον υπολογισμό της μέσης τιμής του πληθυσμού και της βαθμολογίας:

```
> mean(x)
[1] 44.93333
> mean(s)
[1] 58.83333
```

προκύπτει ότι η μέση βαθμολογία της τάξης των μαθητών στα μαθηματικά είναι 44.93 και η εκτίμηση αυτής της βαθμολογίας βάσει του συγκεκριμένου δείγματος είναι 58.83.

Οι μέσοι όλων των δυνατών δειγμάτων μπορούν να υπολογιστούν με εξάντληση όλων των δυνατών τυχαίων αριθμών ως αφετηρίας του συστηματικού δείγματος.

```
> means=rep(0,k)
> for (j in 1:k) means[j]<-mean(x[seq(from=j, to=length(x), by=k)])
> means
[1] 30.16667 37.16667 44.16667 58.83333 54.33333
```

Ο εκτιμητής είναι αμερόληπτος, γιατί πράγματι η μέση τιμή των δυνατών εκτιμήσεων ισούται με την αληθινή μέση τιμή:

```
> mean(means)
[1] 44.93333
```

και η διακύμανση του εκτιμητή του δειγματικού μέσου του συστηματικού δείγματος όπως δίνεται από την (5.3) είναι:

```
> sum((means-mean(means))^2)/k
[1] 112.1067
```

ή, το τυπικό σφάλμα της εκτίμησης είναι 10.59.

*Σημείωση*: Όπως σε κάθε πρόβλημα προγραμματισμού υπάρχουν πολλοί τρόποι υλοποίησης, έτσι και στην περίπτωση της δημιουργίας του συστηματικού δείγματος στην R, αντί της εντολής:

```
> seq(from=j, to=length(x), by=k)
```

που χρησιμοποιήθηκε στο Παράδειγμα 5.1, μπορεί εναλλακτικά να χρησιμοποιηθεί ο τελεστής modulus (x mod y) που στην R παριστάνεται με %%. Οι αντίστοιχες εντολές με τον τρόπο αυτό είναι:

```
> shift <- (1:30) - j
> positions <- (1:30)[(shift %% k) == 0]
```

■

Η ποσότητα $S^{2}_{wsy}$ που ορίστηκε στην Πρόταση 5.2 γράφεται ισοδύναμα:

$$S^2_{wsy} = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{(n-1)}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_i)^2\right) = \frac{1}{k}\sum_{i=1}^{k} S_i^2$$

όπου $S_i^2$, σύμφωνα με τον ορισμό, είναι η δειγματική διασπορά του $i$ συστηματικού δείγματος. Άρα, το $S_{wsy}^2$ είναι ο μέσος όρος των διακυμάνσεων $S_i^2$ για όλα τα δυνατά $i$ −συστηματικά δείγματα, $i = 1,2, \dots, k$. Για τον λόγο αυτόν, η ποσότητα $S_{wsy}^2$ ονομάζεται διακύμανση στο εσωτερικό ή διακύμανση εντός των συστηματικών δειγμάτων (within systematic variation ή intracluster variation) και αντιπροσωπεύει τη μέση δειγματική διακύμανση των συστηματικών δειγμάτων. Μια διαφορετική ερμηνεία της ίδιας ποσότητας είναι ότι παρέχει την πληροφορία για το πόσο συσχετισμένες είναι οι μονάδες στο εσωτερικό των συστηματικών δειγμάτων. Όσο μεγαλύτερη είναι η τιμή της $S_{wsy}^2$, τόσο μεγαλύτερη είναι η ετερογένεια των μονάδων στο εσωτερικό των συστηματικών και αντίστροφα.

Όπως είναι φανερό, πρακτικά η $S_{wsy}^2$ δεν είναι εφικτό να υπολογιστεί στην ακριβή της τιμή, αφού προϋποθέτει την εξάντληση όλων των δυνατών συστηματικών δειγμάτων, και αυτό με τη σειρά του συνεπάγεται την πλήρη εξάντληση του πληθυσμού, δηλ. απογραφή. Μια μεθοδολογία που επιτρέπει την εκτίμηση της ποσότητας και κατά συνέπεια και την εκτίμηση της διακύμανσης του εκτιμητή του μέσου κάτω από το συστηματικό σχήμα ονομάζεται επαναλαμβανόμενη συστηματική δειγματοληψία (repeated systematic sampling) και παρουσιάζεται στην επόμενη παράγραφο.

## 5.3. Εκτίμηση της διακύμανσης του εκτιμητή για τη συστηματική δειγματοληψία

Από τη μέχρι τώρα μελέτη των μεθόδων δειγματοληψίας και της εκτίμησης παραμέτρων του πληθυσμού, είναι σαφές ότι πρωταρχικής σημασίας επιδίωξη είναι η εκτίμηση της διακύμανσης ή του τυπικού σφάλματος του εκτιμητή. Οι εκφράσεις των αληθινών αντίστοιχων ποσοτήτων έχουν ενδιαφέρον θεωρητικό και χρησιμεύουν στη μελέτη και σύγκριση των δειγματοληπτικών σχεδίων, αλλά στην πράξη, αμέσως μετά την υλοποίηση μιας δειγματοληπτικής έρευνας ο ερευνητής ενδιαφέρεται να δώσει αριθμητικά τη διακύμανση του εκτιμητή. Αυτό, όπως έχουμε δει στα μέχρι τώρα κεφάλαια, είναι εφικτό μέσα από τον υπολογισμό μιας εκτίμησης της θεωρητικής διακύμανσης του εκτιμητή με βάση το δείγμα.

Για την περίπτωση της συστηματικής, η θεωρητική διακύμανση του $\bar{X}_{sy}$ δίνεται από την (5.2) και προκειμένου να εκτιμηθεί από το δείγμα, χρειάζεται να εκτιμηθούν οι 2 διασπορές που εμφανίζονται, δηλ. η συνολική διασπορά των μονάδων του πληθυσμού $S^2$ και η μέση διασπορά των συστηματικών στο εσωτερικό τους $S_{wsy}^2$. Η $S^2$ μπορεί να εκτιμηθεί από την αντίστοιχη δειγματική ποσότητα $s^2$, αλλά για την $S_{wsy}^2$ η εκτίμηση δεν είναι εφικτή, αφού είναι ένας μέσος όρος $k$ τιμών $(S_1^2, S_2^2, \dots, S_k^2)$ και η υλοποίηση της συστηματικής θα προσφέρει την αριθμητική τιμή μόνο μίας εξ αυτών.

Η παρακάτω μέθοδος, που προτείνεται ως τροποποίηση της συστηματικής δειγματοληψίας, έχει ως στόχο την εκτίμηση της $S_{wsy}^2$.

### 5.3.1 Επαναλαμβανόμενη συστηματική δειγματοληψία

Η ιδέα πίσω από την επαναλαμβανόμενη συστηματική δειγματοληψία, είναι η παραγωγή πολλαπλών, αντί ενός μοναδικού συστηματικού δείγματος, με σκοπό να είναι εφικτός ο υπολογισμός περισσότερων διακυμάνσεων $S_i^2$ μετά το πέρας της έρευνας.

Αναλυτικά, η μέθοδος αποτελείται από τις εξής φάσεις (βλ. για παράδειγμα Levy & Lemeshow 1999, παράγρ. 4.7):

Φάση 1. Έστω $k$ το βήμα της συστηματικής. Επιλέγουμε έναν αριθμό $m$ (σχετικό με το μέγεθος του $k$), ο οποίος δηλώνει τον αριθμό των επαναλήψεων της συστηματικής δειγματοληψίας.

Φάση 2. Με βήμα $k' = mk$, εφαρμόζουμε $m$ ανεξάρτητες συστηματικές δειγματοληψίες κατά τον συνήθη τρόπο.

Φάση 3. Το δείγμα που αποτελείται από τη συλλογή των μετρήσεων που θα προκύψουν από τις $m$ διαδοχικές συστηματικές δειγματοληψίες είναι το τελικό δείγμα.

Το $m$ επιλέγεται έτσι, ώστε να δίνει ένα ικανό μέγεθος δείγματος για να εκτιμηθεί ο μέσος όρος των $k$ διακυμάνσεων του $S^2_{wsy}$ και, ταυτόχρονα, να είναι διαιρέτης του $n$, ειδικά για μικρό μέγεθος δείγματος.

Το συνολικό τελικό δείγμα παραμένει ίσου μεγέθους με το αρχικό. Η διαφορά είναι κατά την εκτέλεση, και ότι το δείγμα εδώ λαμβάνεται σταδιακά, με $m$ επαναλήψεις του συστηματικού τρόπου δειγματοληψίας επί του πλήρους εύρους του πληθυσμού, αλλά χρησιμοποιώντας μεγαλύτερη απόσταση. Στην περίπτωση της επαναλαμβανόμενης συστηματικής δειγματοληψίας, ο εκτιμητής του μέσου υπολογίζεται από τον μέσο των μέσων των επιμέρους συστηματικών. Δηλ.

$$\bar{X}_{sy,rep} = \bar{\bar{X}}$$

## Παράδειγμα 5.2

Έστω ότι επιθυμούμε να επιλέξουμε ένα συστηματικό δείγμα μεγέθους 20, από ένα σύνολο 200 πρωτοετών φοιτητών ενός Τμήματος Πανεπιστημίου, με σκοπό την εκτίμηση του μέσου χρόνου που αφιερώνουν για μελέτη μέσα στην εβδομάδα, αν εξαιρεθεί ο χρόνος που αφιερώνουν για τις παρακολουθήσεις των διαλέξεων.

Το βήμα $k$ της συστηματικής είναι $k = \frac{200}{20} = 10$. Άρα, σύμφωνα με τον ορισμό της συστηματικής δειγματοληψίας, θα έπρεπε να επιλεγεί ένας τυχαίος αριθμός μεταξύ του 1 και του 10 και στη συνέχεια να επιλεγούν όλοι οι φοιτητές με απόσταση 10 από τον πρώτο. Π.χ. το δείγμα των φοιτητών που αντιστοιχούν στους α/α: 4 14 24 34 44 54 64 74 84 94 104 114 124 134 144 154 164 174 184 και 194 θα ήταν ένα τυχαίο συστηματικό δείγμα.

Για την επαναλαμβανόμενη συστηματική και επειδή $n = 4 * 5$ επιλέγουμε έστω $m = 5$ συστηματικά δείγματα με μέγεθος 4 το καθένα, έτσι ώστε να λαμβάνεται το ίδιο συνολικό μέγεθος δείγματος. Για την επιλογή αυτή του $m$, το βήμα της συστηματικής για το κάθε δείγμα θα είναι $k' = 5 * 10 = 50$. Στη συνέχεια επιλέγονται ανεξάρτητα $m = 5$ τυχαίοι αριθμοί από το 1 έως το 50. Οι αριθμοί αυτοί καθορίζουν τη θέση της αφετηρίας για το κάθε συστηματικό δείγμα. Οι υπόλοιπες θέσεις θα είναι με απόσταση $k' = 50$. Έστω ότι οι αφετηρίες για το παράδειγμα είναι:

```
> sample(1:50,5)
[1] 23  8 42 36 24
```

Τα πέντε συστηματικά θα περιλαμβάνουν αναλυτικά τους φοιτητές με α/α που υπολογίζονται κατά τα γνωστά:

```
> seq(from=23, to=200, by=50)
[1]  23  73 123 173
> seq(from=8, to=200, by=50)
[1]   8  58 108 158
> seq(from=42, to=200, by=50)
[1]  42  92 142 192
> seq(from=36, to=200, by=50)
[1]  36  86 136 186
> seq(from=24, to=200, by=50)
[1]  24  74 124 174
```

και το τελικό συνολικό δείγμα είναι η συλλογή των πέντε επιμέρους.

Η επαναλαμβανόμενη συστηματική δειγματοληψία έχει ως πλεονεκτήματα: (i) την καλύτερη κάλυψη του πληθυσμού, (ii) την αποφυγή της περιοδικότητας που ενδεχομένως εμφανίζει ο πληθυσμός ως προς το χαρακτηριστικό και (iii) την εκτίμηση της διακύμανσης $S^2_{wsy}$, και κατά συνέπεια της διακύμανσης του εκτιμητή $\bar{X}_{sy}$, βάσει ενός δείγματος συνολικού μεγέθους $n$.

Η εκτίμηση του $S^{2}_{wsy}$ γίνεται με τη βοήθεια των $m$ συστηματικών δειγμάτων που έχουν επιλεγεί σταδιακά. Συγκεκριμένα, ο εκτιμητής είναι

$$\hat{S}^2_{wsy} = \frac{1}{m}\sum_{i=1}^{m} S_i^2$$

Για την εκτίμηση της $\mathrm{Var}(\bar{X}_{sy})$ συνολικά, επειδή η επιλογή των $m$ δειγμάτων έγινε με απλή τυχαία δειγματοληψία, η διακύμανση του εκτιμητή του μέσου προκύπτει από το αντίστοιχο αποτέλεσμα της α.τ.δ. και είναι:

$$\mathrm{Var}(\bar{X}_{sy}) = \left(1-\frac{m}{k'}\right)\frac{1}{m}\left[\frac{1}{m-1}\sum_{i=1}^{m}(\bar{X}^{(s_i)}-\bar{\bar{X}})^2\right]$$

όπου $\bar{X}^{(s_i)}$ $(i = 1,2,\dots,m)$, είναι η μέση δειγματική τιμή του $i$ συστηματικού, από τα $m$ που θα επιλεχθούν συνολικά. Με τη βοήθεια της σχέσης αυτής, δίνουμε το παρακάτω πόρισμα.

### Πόρισμα 5.2

Οι εκτιμήσεις της διακύμανσης και του τυπικού σφάλματος του $\bar{X}_{sy}$ με τη βοήθεια της επαναλαμβανόμενης συστηματικής δειγματοληψίας, δίνονται από τις σχέσεις:

$$\mathrm{Var}(\bar{X}_{sy}) = \left(1-\frac{m}{k'}\right)\frac{1}{m}\left[\frac{1}{m-1}\sum_{i=1}^{m}(\bar{X}^{(s_i)}-\bar{\bar{X}})^2\right] \tag{5.4}$$

και:

$$\mathrm{se}(\bar{X}_{sy}) = \sqrt{\left(1-\frac{m}{k'}\right)\frac{1}{m}\left[\frac{1}{m-1}\sum_{i=1}^{m}(\bar{X}^{(s_i)}-\bar{\bar{X}})^2\right]} \tag{5.5}$$

αντίστοιχα.

## Παράδειγμα 5.3

Έστω ότι τα δεδομένα για τον χρόνο (σε ώρες) που αφιερώνουν οι φοιτητές για μελέτη του Παραδείγματος 5.2 δίνονται στον Πίνακα 5.2.

| α/α φοιτητών | Χρόνος για μελέτη | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1 έως 20** | 2 | 2 | 6 | 0 | 3 | 1 | 10 | 1 | 4 | 5 | 0 | 1 | 6 | 0 | 1 | 1 | 1 | 0 | 2 | 3 |
| **21 έως 40** | 4 | 6 | 4 | 2 | 4 | 1 | 2 | 2 | 4 | 6 | 5 | 1 | 0 | 0 | 1 | 4 | 6 | 3 | 3 | 2 |
| **41 έως 60** | 0 | 3 | 0 | 1 | 4 | 6 | 4 | 1 | 4 | 0 | 2 | 0 | 0 | 2 | 1 | 10 | 5 | 3 | 2 | 9 |
| **61 έως 80** | 1 | 0 | 3 | 1 | 8 | 6 | 3 | 2 | 4 | 9 | 1 | 5 | 4 | 1 | 2 | 6 | 6 | 2 | 4 | 4 |
| **81 έως 100** | 2 | 0 | 9 | 2 | 2 | 4 | 6 | 1 | 5 | 2 | 3 | 7 | 5 | 1 | 6 | 0 | 3 | 6 | 1 | 4 |
| **101 έως 120** | 3 | 6 | 1 | 2 | 4 | 3 | 1 | 0 | 1 | 1 | 0 | 4 | 4 | 3 | 2 | 0 | 11 | 1 | 9 | 3 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **121 έως 140** | 2 | 8 | 0 | 3 | 5 | 3 | 6 | 3 | 2 | 6 | 3 | 1 | 5 | 1 | 0 | 2 | 1 | 0 | 9 | 0 |
| **141 έως 160** | 5 | 0 | 8 | 1 | 9 | 1 | 7 | 3 | 5 | 0 | 1 | 0 | 2 | 1 | 6 | 2 | 1 | 6 | 4 | 4 |
| **161 έως 180** | 9 | 3 | 8 | 3 | 2 | 6 | 1 | 2 | 4 | 1 | 0 | 7 | 0 | 0 | 6 | 2 | 0 | 1 | 8 | 1 |
| **181 έως 200** | 1 | 8 | 0 | 1 | 3 | 3 | 2 | 2 | 3 | 3 | 0 | 5 | 4 | 0 | 2 | 3 | 3 | 3 | 1 | 6 |

***Πίνακας 5.2*** *Στοιχεία για τον εβδομαδιαίο χρόνο μελέτης (σε ώρες) των φοιτητών.*

Αν x είναι το διάνυσμα με τα δεδομένα, τότε τα πέντε συστηματικά δείγματα του Παραδείγματος 5.2 με αφετηρίες 23, 8, 42, 36 και 24 είναι:

```
> m<-5
> m1<-mean(x[seq(from=23, to=200, by=50)])
> m2<-mean(x[seq(from=8, to=200, by=50)])
> m3<-mean(x[seq(from=42, to=200, by=50)])
> m4<-mean(x[seq(from=36, to=200, by=50)])
> m5<-mean(x[seq(from=24, to=200, by=50)])
> mu<-mean(c(m1,m2,m3,m4,m5))
> mu
[1] 3.1
> syvar<-(1-m/10)*(1/m)*var(c(m1,m2,m3,m4,m5))
> syvar
[1] 0.133125
> syse<-sqrt(syvar)
> syse
[1] 0.364863
```

Συνεπώς, ο μέσος αριθμός ωρών που αφιερώνουν οι φοιτητές εβδομαδιαίως για διάβασμα είναι 3.1 ώρες σύμφωνα με το δείγμα των 20 φοιτητών, με τυπική απόκλιση της εκτίμησης 0.36 ώρες.

### 5.3.2 Συντελεστής συσχέτισης intracluster (Intracluster correlation)

Ένας εναλλακτικός τρόπος για την εκτίμηση της διακύμανσης του εκτιμητή του πληθυσμιακού μέσου στη συστηματική δίνεται μέσω του συντελεστή συσχέτισης intracluster (intracluster ή intraclass correlation). Συμβολίζεται με $\rho$ ή $\rho_w$ (από το 'within') και ορίζεται ως ο συντελεστής συσχέτισης ανά δύο, των στοιχείων που ανήκουν στο ίδιο συστηματικό δείγμα. Αν συμβολίσουμε με $X_{ij}$ τη μονάδα του δείγματος που βρίσκεται στο $i$ συστηματικό, (δηλ. του δείγματος που προκύπτει με επιλογή του τυχαίου αριθμού $i$), και ανήκει στη $j$ θέση, τότε ο συντελεστής $\rho$ ορίζεται ως:

$$\rho_w = \frac{2}{(n-1)(N-1)S^2}\sum_{i=1}^{k}\left[\sum_{j=1}^{n}\sum_{\substack{u=1 \\ u>j}}^{n}(X_{ij}-\bar{Y})(X_{iu}-\bar{Y})\right]$$

όπου $\bar{Y}$ η πληθυσμιακή μέση τιμή και $S^2$ η πληθυσμιακή διασπορά (βλ. Cochran 1977, παράγρ. 8.3).

Ο συντελεστής $\rho_w$ σχετίζεται άμεσα με το $S^2_{wsy}$ και παρέχει την ίδια πληροφορία, αλλά με διαφορετική προσέγγιση. Τόσο το $\rho_w$, όσο και το $S^2_{wsy}$ δηλώνουν τη σχέση μεταξύ των μονάδων που ανήκουν στο ίδιο συστηματικό δείγμα. Όσο μεγαλύτερος είναι ο συντελεστής συσχέτισης $\rho_w$, τόσο μικρότερο αναμένεται να είναι το $S^2_{wsy}$, και αντίστροφα.

Μια εναλλακτική έκφραση της (5.2) για την ακριβή διακύμανση του $\bar{X}_{sy}$ μέσω της εσωτερικής συσχέτισης δίνεται από την παρακάτω Πρόταση.

### Πρόταση 5.3

Η διακύμανση $\mathrm{Var}(\bar{X}_{sy})$ δίνεται από τον τύπο:

$$\mathrm{Var}\left(\bar{X}_{sy}\right) = \frac{S^2(N-1)}{nN}[1 + (n-1)\rho_w] \tag{5.6}$$

(για την απόδειξη, βλ. Cochran 1977, παράγρ. 8.3).

Αν, επομένως, εκτιμήσουμε τον συντελεστή συσχέτισης $\rho_w$, προκύπτει μέσω της (5.6) ένας εναλλακτικός τρόπος εκτίμησης της διακύμανσης του εκτιμητή $\mathrm{Var}\left(\bar{X}_{sy}\right)$ για τη συστηματική δειγματοληψία. Η εκτίμηση του $\rho_w$ μπορεί να προκύψει με επιλογή ικανού αριθμού συστηματικών δειγμάτων (ανάλογο του $m$ στην επαναλαμβανόμενη συστηματική δειγματοληψία) και τον υπολογισμό του συντελεστή συσχέτισης ανά δύο, των στοιχείων που ανήκουν στο ίδιο συστηματικό δείγμα. Υπενθυμίζεται ότι το $\rho_w$ δεν αναφέρεται στις διαδοχικές μονάδες του πληθυσμού, αλλά στα στοιχεία του ίδιου συστηματικού δείγματος, δηλ. μονάδες του πληθυσμού με απόσταση $k$ μεταξύ τους.

Από την έκφραση (5.6) προκύπτουν ακόμα μια σειρά από ενδιαφέροντα συμπεράσματα που είναι χρήσιμα από πρακτικής άποψης, αλλά και από την πλευρά της στατιστικής συμπερασματολογίας και, ειδικότερα, της σύγκρισης της συστηματικής με άλλα δειγματοληπτικά σχέδια, ως προς την αποτελεσματικότητα.

### Παρατήρηση 5.8

Αν η τιμή του $\rho_w$ είναι μηδέν, δηλ. οι μονάδες που ανήκουν στο ίδιο συστηματικό είναι ασυσχέτιστες, τότε $\mathrm{Var}\left(\bar{X}_{sy}\right) = \frac{S^2(N-1)}{nN} \cong \frac{S^2}{n}$ που είναι η διακύμανση του εκτιμητή του μέσου κάτω από την α.τ.δ. (για αμελητέο $f$). Συνεπώς, αν $\rho_w = 0$, τότε:

(α) Ο υπολογισμός της διακύμανσης $\mathrm{Var}\left(\bar{X}_{sy}\right)$ και του εκτιμητή της μπορεί να γίνει με τη βοήθεια των αντίστοιχων αποτελεσμάτων για την α.τ.δ.

(β) Η συστηματική δειγματοληψία δίνει εκτίμηση ίσης αποτελεσματικότητας με την α.τ.δ.

(γ) Στις περιπτώσεις αυτές μπορεί να εφαρμοστεί στην πράξη η συστηματική δειγματοληψία αντί της απλής τυχαίας γιατί είναι πιο εύκολη και γιατί δεν απαιτεί μια λίστα των μελών του πληθυσμού όπως η α.τ.δ. Τα τυπικά σφάλματα θα υπολογίζονται μέσω των τύπων από την α.τ.δ.

Αρκετά από τα διαθέσιμα στατιστικά πακέτα που αφορούν την επιλογή του δείγματος και την εκτίμηση των παραμέτρων του πληθυσμού χρησιμοποιούν τις εκφράσεις από την α.τ.δ. για τον υπολογισμό της διακύμανσης του εκτιμητή της μέσης τιμής του πληθυσμού, υποθέτοντας ότι $\rho_w = 0$. Ένα τέτοιο στατιστικό πακέτο είναι το `TeachingSampling` της `R`.

### Παρατήρηση 5.9

Αν το $\rho_w$ είναι θετικό, έστω και ελάχιστα, τότε η διακύμανση του $\bar{X}_{sy}$ είναι μεγαλύτερη από την αντίστοιχη διακύμανση κάτω από την α.τ.δ. ίσου μεγέθους, και συνεπώς η συστηματική είναι λιγότερο αποτελεσματική

ακόμα και από την α.τ.δ. Συγκεκριμένα, η επίδραση του συστηματικού δειγματοληπτικού σχεδίου έναντι της α.τ.δ. με τη βοήθεια της (5.6) υπολογίζεται από την έκφραση:

$$\text{deff} = 1 + (n-1)\rho_w$$

κατά την οποία το $\rho_w$ πολλαπλασιάζεται με το $n-1$ και κατά συνέπεια ακόμα και μικρές θετικές τιμές για το $\rho_w$ οδηγούν σε deff αρκετά μεγαλύτερο του ένα.

Από τις Παρατηρήσεις 5.2 και 5.3 είναι ήδη φανερό ότι, σε αντίθεση με τη στρωματοποιημένη δειγματοληψία, η συστηματική δεν είναι ένας τρόπος δειγματοληψίας που εγγυάται μικρότερα τυπικά σφάλματα, δηλαδή πιο ακριβείς εκτιμήσεις. Το αποτέλεσμα εδώ ποικίλλει και εξαρτάται από τον βαθμό εξάρτησης των μονάδων που ανήκουν στο ίδιο συστηματικό δείγμα, δηλ. την κατανομή των μονάδων στον πληθυσμό ή, διαφορετικά, την κατασκευή (structure) ή δομή του πληθυσμού. Το θέμα αυτό θα μας απασχολήσει περισσότερο στην Παράγραφο 5.6.

## 5.4. Εκτίμηση συνόλου και ποσοστού. Διαστήματα εμπιστοσύνης. Ελάχιστο απαιτούμενο μέγεθος δείγματος

### 5.4.1 Εκτίμηση συνόλου και ποσοστού

Για την εκτίμηση του συνόλου $Y_T$ ισχύει η ίδια σύνδεση με το πρόβλημα της εκτίμησης της μέσης τιμής $\bar{Y}$ που εφαρμόστηκε και για τη μελέτη της α.τ.δ. και της στρωματοποιημένης. Επειδή:

$$Y_T = N\,\bar{Y}$$

θα είναι:

$$\hat{Y}_{T,sy} = N\,\bar{X}_{sy}$$

Οι ιδιότητες του $\hat{Y}_{T,sy}$ εξάγονται ως άμεσο συμπέρασμα από τις ιδιότητες του $\bar{X}_{sy}$.

Ομοίως για την εκτίμηση του ποσοστού, κάνοντας χρήση της σύνδεσης που αναπτύχθηκε στην παράγραφο 2.3.5, όπου το ποσοστό μπορεί να θεωρηθεί ως ο μέσος όρος ενός ειδικά κωδικοποιημένου πληθυσμού, η εκτίμησή του κάτω από το συστηματικό σχήμα ακολουθεί τα ίδια βήματα και έχει τις ίδιες ιδιότητες με τις ιδιότητες του $\bar{X}_{sy}$. Υπενθυμίζουμε ότι κάτω από την κωδικοποίηση σε 0 και 1 των τιμών του πληθυσμού, ισχύει $S^2 = \frac{NP(1-P)}{N-1}$, όπου $P$ το άγνωστο, προς εκτίμηση, ποσοστό του πληθυσμού.

Συνοπτικά οι χρήσιμες εκφράσεις για την εκτίμηση του ποσοστού δίνονται στον Πίνακα 5.3.

| | |
|---|---|
| **Εκτιμητής ποσοστού** | $$\hat{P}_{sy} = p_{sy}$$ όπου $p_{sy}$ το ποσοστό, υπολογισμένο στο συστηματικό που επιλέχθηκε. |
| **Διακύμανση εκτιμητή ποσοστού** | $$\mathrm{Var}(\hat{P}_{sy}) = P(1-P) - \frac{k(n-1)}{N} S^2_{wsy}$$ όπου $$S^2_{wsy} = \frac{1}{k}\sum_{i=1}^{k} \frac{np_i(1-p_i)}{n-1}$$ και $p_i$ το ποσοστό όπως υπολογίζεται από το $i$-οστό συστηματικό δείγμα. |
| **Εκτιμώμενη διακύμανση του εκτιμητή του ποσοστού** | $$\mathrm{Var}(\hat{P}_{sy}) = \left(1-\frac{1}{k}\right)\frac{1}{m}\left[\frac{1}{m-1}\sum_{i=1}^{m}(p^{(s_i)} - \bar{\bar{p}})^2\right]$$ όπου $p^{(s_i)}$ είναι το ποσοστό από το $i$, $(i = 1, \dots, m)$ επαναληπτικό συστηματικό και $\bar{\bar{p}}$ ο μέσος των $p^{(s_i)}$ για τα διάφορα $i$. |
| **Τυπικό σφάλμα Ποσοστού** | $$\mathrm{se}(\hat{P}_{sy}) = \sqrt{\mathrm{Var}(\hat{P}_{sy})}$$ |
| **Εκτιμώμενο Τυπικό σφάλμα Ποσοστού** | $$\hat{\mathrm{se}}(\hat{P}_{sy}) = \sqrt{\hat{\mathrm{Var}}(\hat{P}_{sy})}$$ |

***Πίνακας 5.3*** *Εκτίμηση Ποσοστού για τη συστηματική δειγματοληψία*

### 5.4.2 Διαστήματα εμπιστοσύνης

Για το διάστημα εμπιστοσύνης, η γενική θεωρία έχει αναπτυχθεί στην παράγραφο 2.3 και ο σχετικός τύπος που περιλαμβάνει καθεμία από τις παραπάνω περιπτώσεις πληθυσμιακής παραμέτρου είναι:

$$\left(\hat{\theta} - t_{n-1,a/2}\hat{\mathrm{se}}(\hat{\theta}),\ \ \hat{\theta} + t_{n-1,a/2}\hat{\mathrm{se}}(\hat{\theta})\right)$$

για μικρά $n$, και:

$$\left(\hat{\theta} - z_{a/2}\,\hat{\mathrm{se}}(\hat{\theta}),\ \ \hat{\theta} + z_{a/2}\,\hat{\mathrm{se}}(\hat{\theta})\right)$$

για $n > 30$, όπου $\hat{\theta}$ ο εκτιμητής της παραμέτρου $\theta$ ($\theta = \bar{Y}, Y_T$ ή $P$) και $\hat{\mathrm{se}}(\hat{\theta})$ το εκτιμώμενο τυπικό σφάλμα της εκτίμησης. Για ένα δείγμα που έχει προέλθει με συστηματική δειγματοληψία, το ΔΕ υπολογίζεται αναπτύσσοντας τόσο τις τιμές του εκτιμητή, όσο και του τυπικού σφάλματος, σύμφωνα με τους τύπους της συστηματικής.

### Παράδειγμα 5.10

Για τα δεδομένα του Παραδείγματος 5.3, ένα 95% διάστημα εμπιστοσύνης για τον μέσο χρόνο που αφιερώνουν οι φοιτητές μέσα στην εβδομάδα για διάβασμα υπολογίζεται μέσω του δείγματος από την επαναληπτική συστηματική.

Με τη βοήθεια των μέχρι τώρα αποτελεσμάτων για τη μέση τιμή και το τυπικό σφάλμα, το ΔΕ είναι:

```
> c(mu- qt(.975, 4)*syse, mu+qt(.975, 4)*syse)
[1] 2.086978 4.113022
```

Δηλ. ο μέσος χρόνος που αφιερώνουν οι φοιτητές για διάβασμα μέσα στην εβδομάδα εκτιμάται ότι ανήκει στο διάστημα (2.1, 4.1) ώρες με βαθμό εμπιστοσύνης 95%. Σημειώνεται ότι οι βαθμοί ελευθερίας στο παραπάνω ΔΕ είναι 4, γιατί η συστηματική δειγματοληψία είναι επαναληπτική και η εκτίμηση του τυπικού σφάλματος βασίζεται σε 5 δείγματα από τα 50 συνολικά συστηματικά.

## 5.4.3 Ελάχιστο απαιτούμενο μέγεθος δείγματος

Για τον υπολογισμό του ελάχιστου απαιτούμενου μεγέθους δείγματος, εφαρμόζουμε τον τύπο (2.11), δηλ.

$$\frac{d}{\text{se}(\hat{\theta})} = z_{\alpha/2} \tag{5.7}$$

όπου $\hat{\theta}$ ο εκτιμητής της πληθυσμιακής παραμέτρου $\theta$ σύμφωνα με τη συστηματική δειγματοληψία και $\text{se}(\hat{\theta})$ το τυπικό σφάλμα, όπως αυτό υπολογίζεται από μια προκαταρκτική έρευνα. Το μέγεθος της δειγματοληψίας είναι η λύση της εξίσωσης ως προς $n$. Εάν μπορούμε να υποθέσουμε ότι δεν υπάρχει κάποια διάταξη ή περιοδικότητα για τον πληθυσμό, τότε το $\text{se}(\hat{\theta})$ υπολογίζεται ειδικότερα από τον τύπο της α.τ.δ. Στην περίπτωση αυτή, η επίλυση της (5.7) είναι ένα πανομοιότυπο πρόβλημα με εκείνο του Κεφάλαιο 2 -.

Αν δεν μπορούμε να κάνουμε την υπόθεση της τυχαιότητας στη διάταξη, ή υποπτευόμαστε περιοδικότητα στις μονάδες του πληθυσμού, τότε για τον υπολογισμό του $\text{se}(\hat{\theta})$ στον παρονομαστή της (5.7) κάνουμε χρήση της έκφρασης (5.4) που έχουμε βρει για την επαναληπτική συστηματική δειγματοληψία:

$$\text{se}(\hat{\theta}) = \sqrt{\left(1-\frac{m}{k'}\right)\frac{1}{m}\left[\frac{1}{m-1}\sum_{i=1}^{m}(\bar{X}^{(s_i)}-\bar{\bar{X}})^2\right]}$$

Ο άγνωστος στην περίπτωση αυτή είναι ο αριθμός $m$, δηλ. πόσα συστηματικά δείγματα θα χρειαστεί να επιλέξουμε στην επαναληπτική συστηματική, ώστε να πληρούνται οι προδιαγραφές της έρευνας. Ένας αριθμός από συστηματικά δείγματα, έστω $m'$ θα είναι διαθέσιμος από την προκαταρκτική έρευνα. Με βάση τα $m'$ δείγματα εκτιμούμε την ποσότητα στην κλειστή παρένθεση. Αντικαθιστούμε στη συνέχεια στην (5.7) και λύνουμε ως προς το $m$. Στο ακόλουθο Παράδειγμα, εφαρμόζουμε την παραπάνω διαδικασία στα δεδομένα του Παραδείγματος 5.3.

### Παράδειγμα 5.5

Για τα δεδομένα του Παραδείγματος 5.3 και θεωρώντας τα επαναληπτικά συστηματικά δείγματα του Παραδείγματος 5.3 ως προκαταρκτική έρευνα, να υπολογιστεί το πλήθος των συστηματικών δειγμάτων με βήμα $k' = 50$ που είναι απαραίτητο να επιλεγούν, έτσι ώστε η εκτίμηση του μέσου χρόνου μελέτης των φοιτητών να απέχει από την αληθινή τιμή το πολύ 0.25% αυτής, με πιθανότητα σφάλματος 5%.

Η εξίσωση (5.7) για τα δεδομένα του Παραδείγματος είναι

$$\frac{0.25 * \hat{\bar{Y}}}{\text{se}(\hat{\bar{Y}})} = 1.96$$

και, σύμφωνα με την προκαταρκτική έρευνα των πέντε επαναληπτικών συστηματικών του Παραδείγματος 5.3, ένας εκτιμητής της ποσότητας:

$$\left[\frac{1}{m-1}\sum_{i=1}^{m}(\bar{X}^{(s_i)} - \bar{\bar{X}})^2\right]$$

είναι:

```
> var(c(m1,m2,m3,m4,m5))
[1] 1.33125
```

Άρα

$$\text{se}(\hat{\theta}) = \sqrt{\left(1-\frac{m}{k'}\right)\frac{1}{m}\left[\frac{1}{m-1}\sum_{i=1}^{m}(\bar{X}^{(s_i)} - \bar{\bar{X}})^2\right]} = \sqrt{\left(1-\frac{m}{k'}\right)\frac{1}{m}1.33125}$$

Για $k' = 50$ και $\hat{\bar{Y}} = 3.1$, από την προκαταρκτική έρευνα του Παραδείγματος 5.3, αντικαθιστώντας όλα τα παραπάνω στην:

$$\frac{0.25 * \hat{\bar{Y}}}{\text{se}(\hat{\bar{Y}})} = 1.96$$

ο μόνος άγνωστος που παραμένει είναι το μέγεθος των συστηματικών δειγμάτων $m$. Η λύση της εξίσωσης θα δώσει και τον αριθμό $m$:

$$\frac{0.25 * 3.1}{\sqrt{\left(\frac{1}{m}-\frac{1}{50}\right)1.33125}} = 1.96 \quad \text{ή} \quad \left(\frac{1}{m}-\frac{1}{50}\right)1.33125 = \left(\frac{0.25 * 3.1}{1.96}\right)^2$$

ή

$$m = \left[\frac{1}{50} + \left(\frac{0.25 * 3.1}{1.96}\right)^2 \frac{1}{1.33125}\right]^{-1}$$

```
> (1/50+((0.25*3.1)/1.96)^2*(1/1.33125))^(-1)
[1] 7.275679
```

Άρα, χρειάζεται δείγμα μεγέθους τουλάχιστον 8 συστηματικών δειγμάτων με βήμα 50 το καθένα, ή διαφορετικά δείγμα συνολικού μεγέθους 32, έτσι ώστε η δειγματοληψία να πραγματοποιηθεί με συστηματικό τρόπο και να πληρούνται οι προδιαγραφές της έρευνας.

## 5.5. Σύγκριση συστηματικής με την απλή τυχαία δειγματοληψία

Στην παράγραφο 5.3 έγινε ήδη η διαπίστωση ότι η συστηματική δειγματοληψία δεν είναι πάντα καλύτερη ως προς την ακρίβεια, σε σχέση με την απλή τυχαία. Στην παρούσα παράγραφο θα μελετήσουμε πιο αναλυτικά

τη σύγκριση αυτή και θα δώσουμε τις συνθήκες κάτω από τις οποίες η μία δειγματοληπτική μέθοδος υπερισχύει έναντι της άλλης.

Η παρακάτω Πρόταση περιέχει το σχετικό αποτέλεσμα.

### Πρόταση 5.4

Η συστηματική δειγματοληψία είναι πιο αποτελεσματική σε σχέση με την απλή τυχαία ως προς την εκτίμηση του πληθυσμιακού μέσου, αν και μόνον αν:

$$S^{\,2}_{wsy} > S^2 \tag{5.8}$$

### Απόδειξη

Συγκρίνουμε τις διακυμάνσεις του εκτιμητή του πληθυσμιακού μέσου κάτω από τη συστηματική και την α.τ.δ. ίσου μεγέθους δείγματος, λαμβάνοντας τη διαφορά τους. Από τη γνωστή σχέση που ισχύει στην α.τ.δ. και την (5.2) αντίστοιχα για τη συστηματική, θα είναι:

$$\mathrm{Var}(\bar{X}) - \mathrm{Var}\left(\bar{X}_{sy}\right) = \frac{(1-\frac{n}{N})}{n}S^2 - \frac{N-1}{N}S^2 + \frac{k(n-1)}{N}S^2_{wsy} =$$

$$= \left(\frac{1-\frac{n}{N}}{n} - \frac{N-1}{N}\right)S^2 + \frac{k(n-1)}{N}S^2_{wsy}.$$

Αν θέσουμε $N = kn$ και κάνουμε πράξεις στους συντελεστές των δύο όρων του αθροίσματος, προκύπτει ισοδύναμα:

$$\mathrm{Var}(\bar{X}) - \mathrm{Var}\left(\bar{X}_{sy}\right) = \left(\frac{n-1}{n}\right)(S^2_{wsy} - S^2) = \frac{k(n-1)}{N}\left(S^2_{wsy} - S^2\right)$$

Συνεπώς, η διαφορά $\mathrm{Var}(\bar{X}) - \mathrm{Var}\left(\bar{X}_{sy}\right)$ είναι θετική, και άρα η συστηματική δειγματοληψία είναι πιο αποτελεσματική από την α.τ.δ., αν και μόνον αν $S^2_{wsy} - S^2 > 0$ ■

Το αποτέλεσμα της Πρότασης 5.4 δηλώνει ότι η συστηματική δειγματοληψία θα έχει παράλληλα, εκτός από τα πλεονεκτήματα στην ευκολία υλοποίησης, και την ιδιότητα της αποτελεσματικότητας, στην περίπτωση που η $S^2_{wsy}$ είναι αρκετά μεγάλη, συγκεκριμένα τέτοια ώστε να ξεπερνά τη συνολική πληθυσμιακή μεταβλητότητα $S^2$. Αυτό με τη σειρά του σημαίνει ότι οι μονάδες που ανήκουν στο εσωτερικό του καθενός από τα συστηματικά δείγματα παρουσιάζουν όσο το δυνατό μεγαλύτερη ετερογένεια και επίσης ότι αναμένουμε οι μέσοι όλων των δυνατών συστηματικών να είναι κοντά.

Στην αντίθετη περίπτωση, όταν δηλ. τα συστηματικά δείγματα είναι ομοιογενή, τότε η συστηματική δειγματοληψία θα δώσει μεγάλα τυπικά σφάλματα για τους εκτιμητές. Το αποτέλεσμα αυτό θα είναι απόρροια του γεγονότος ότι το συστηματικό δείγμα που θα επιλεγεί δεν θα είναι αντιπροσωπευτικό του πληθυσμού, αφού η ομοιογένεια συνεπάγεται ότι παρόμοιες παρατηρήσεις θα έχουν τοποθετηθεί μέσα στο ίδιο δείγμα. Στην περίπτωση αυτή, οι μέσοι όλων των δυνατών συστηματικών αναμένονται να διαφέρουν μεταξύ τους.

Αντίστοιχα αποτελέσματα και ερμηνεία είναι εφικτά, αν η διακύμανση του εκτιμητή για τη συστηματική δειγματοληψία εκφραστεί μέσω του συντελεστή συσχέτισης $\rho_w$ με τη βοήθεια της σχέσης (5.6). Μεγάλη ομοιογένεια στο εσωτερικό των συστηματικών, δηλ. μεγάλο $\rho_w$, θα έχει ως αποτέλεσμα η συστηματική

δειγματοληψία να είναι λιγότερο αποδοτική από την α.τ.δ. Αντίθετα, εάν το $\rho_w$ έχει τιμή έστω και ελάχιστα αρνητική, τότε ο εκτιμητής $\bar{X}_{sy}$ είναι πιο αποτελεσματικός από τον $\bar{X}$.

## 5.6. Συστηματική δειγματοληψία και δομή του πληθυσμού

Η μελέτη των ιδιοτήτων του εκτιμητή $\bar{X}_{sy}$ έδειξε ότι ένα ιδιαίτερο χαρακτηριστικό της συστηματικής δειγματοληψίας που παίζει πολύ σημαντικό ρόλο στην ακρίβεια των εκτιμητών που παράγονται είναι η δομή στο εσωτερικό των δυνατών συστηματικών δειγμάτων. Εάν ο ερευνητής σκοπεύει να κάνει χρήση της συστηματικής δειγματοληψίας, τότε επιθυμεί και επιδιώκει – αν τα διαθέσιμα βοηθητικά στοιχεία του το επιτρέπουν – να επιτύχει μεγάλη ετερογένεια στο εσωτερικό των συστηματικών δειγμάτων. Το παρακάτω παράδειγμα επεξηγεί με πιο παραστατικό τρόπο το σχόλιο αυτό.

### Παράδειγμα 5.6

Χρησιμοποιούμε τα δεδομένα SHS που περιέχονται στο πακέτο `stratification` της `R`, (βλ. Εφαρμογή 4.2.2 του 4ου Κεφαλαίου). Υπενθυμίζουμε ότι τα δεδομένα περιέχουν πληροφορίες από την έρευνα «2001 Survey of Household Spending (SHS)» και αποτελούνται από 16057 παρατηρήσεις και μετρήσεις σε 7 μεταβλητές.

Έστω ότι επιδιώκεται να εκτιμηθεί η μέση τιμή της μεταβλητής `M101 (Household spending on recreation`, το χρηματικό ποσό που ξοδεύει ένα νοικοκυριό για διασκέδαση) βάσει ενός συστηματικού δείγματος που επιλέγεται από τον πληθυσμό για $f = 0.02$. Να υπολογιστεί το τυπικό σφάλμα της εκτίμησης και να συγκριθεί με το αντίστοιχο της α.τ.δ. ίσου μεγέθους δείγματος, για τις εξής περιπτώσεις:

(iii) Χωρίς χρήση βοηθητικής μεταβλητής,

(iv) Κάνοντας χρήση της βοηθητικής μεταβλητής `HHINCTOT (Household income before taxes`, συνολικό εισόδημα του νοικοκυριού προ φόρων).

Εισάγουμε τα δεδομένα και χρησιμοποιούμε ως φίλτρο για τη βασική μεταβλητή το νοικοκυριό να έχει μη-μηδενικό συνολικό εισόδημα. Έστω `y` η βασική μεταβλητή και `w` η βοηθητική.

```
> data(SHS)
> y <- SHS$M101[SHS$HHINCTOT>0]
> w <- SHS$HHINCTOT[SHS$HHINCTOT>0]
```

Υπολογίζουμε στη συνέχεια τα απαραίτητα στοιχεία για την εφαρμογή της συστηματικής, δηλ. $n$ και $k$. Βάσει του $f$ που δίνεται. θα είναι

```
> f<-0.02
> n<-f*N
> N<-length(y)
> N
[1] 16025
> n<-f*N
> n
[1] 320.5
> k<-round(N/n)
> k
[1] 50
```

Για την περίπτωση που η συστηματική δειγματοληψία εφαρμόζεται στη βασική μεταβλητή y χωρίς χρήση βοηθητικής, υπολογίζουμε την ακριβή διασπορά του $\bar{X}_{sy}$ από την (5.2), εξαντλώντας όλα τα συστηματικά δείγματα και υπολογίζοντας ακριβώς την $S^2_{wsy}$. Οι εντολές για το σκοπό αυτό είναι:

```
> vars_w<-rep(0,50)
> for (i in 1:50) vars_w[i]<-var(y[seq(from=i, to=N, by=k)])
> Swsy<-mean(vars_w)
> Swsy
[1] 18166431
```

Με βάση τα ανωτέρω, η διακύμανση (5.2) και στη συνέχεια το τυπικό σφάλμα είναι:

```
> var(y)*(N-1)/N- k*(n-1)* Swsy /N
[1] 74162.21
> sqrt(74162.21)
[1] 272.3274
```

Άρα, για μια συστηματική δειγματοληψία μεγέθους 320, ο εκτιμητής του μέσου ποσού που διαθέτουν τα νοικοκυριά για διασκέδαση έχει διακύμανση 74162.21 ή τυπικό σφάλμα 272.33.

Η αντίστοιχη διακύμανση του α.τ.δ. ίδιου μεγέθους είναι:

```
>  var(y)*(1-f)/n
[1] 55604.82
> sqrt( var(y)*(1-f)/n)
[1] 235.8067
```

δηλ. η α.τ.δ. είναι, στην περίπτωση αυτή, πιο αποτελεσματική από τη συστηματική.

Με χρήση της βοηθητικής μεταβλητής w που είναι το συνολικό εισόδημα του νοικοκυριού πριν από τους φόρους, επαναλαμβάνουμε την ίδια διαδικασία με το ερώτημα (i), διατάσσοντας όμως τις τιμές της y πριν από την εφαρμογή της συστηματικής, σύμφωνα με το συνολικό εισόδημα w. Η βοηθητική μεταβλητή εδώ χρησιμοποιείται μόνο για την αναδιάταξη των στοιχείων της βασικής μεταβλητής. Διατάσσουμε:

```
> y_ordered<-y[order(w)]
```

και επαναλαμβάνοντας τα προηγούμενα βήματα:

```
> vars_w<-rep(0,50)
> for (i in 1:50) vars_w[i]<-var(y_ordered[seq(from=i, to=N, by=k)])
> Swsy<-mean(vars_w)
> Swsy
[1] 18188610
>  var(y)*(N-1)/N- k*(n-1)*Swsy/N
[1] 52052.45
> sqrt(var(y)*(N-1)/N- k*(n-1)*Swsy/N)
[1] 228.1501
```

απ΄όπου συμπεραίνεται ότι η αποτελεσματικότητα του $\bar{X}_{sy}$ έχει βελτιωθεί μετά τη διάταξη των τιμών της βασικής μεταβλητής σύμφωνα με τη βοηθητική. Ο εκτιμητής $\bar{X}_{sy}$ είναι πλέον πιο αποτελεσματικός σε σχέση με τον $\bar{X}$ από την α.τ.δ.

Η διαφοροποίηση στο αποτέλεσμα επήλθε γιατί στη μεν πρώτη περίπτωση δεν ικανοποιείται η συνθήκη (5.8):

```
> var(y)
[1] 18185046
> 18166431>var(y)
[1] FALSE
```

ενώ στη δεύτερη περίπτωση:

```
> var(y_ordered)
[1] 18185046
> 18188610>var(y_ordered)
[1] TRUE
```

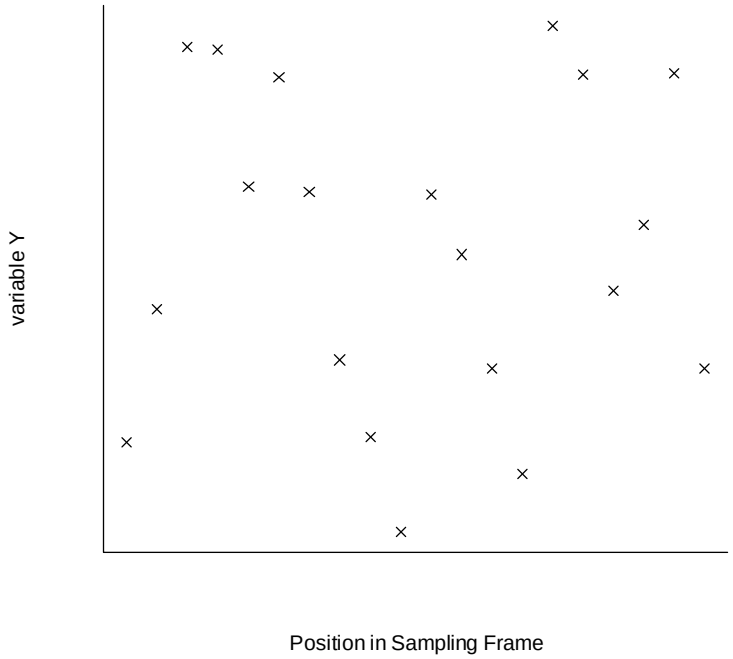
ικανοποιείται. Είναι προφανές ότι `var(y)=var(y_ordered)` ■

Το συμπέρασμα του παραδείγματος είναι ότι μια αναδιάταξη των μονάδων του πληθυσμού, χωρίς καμία άλλη αλλαγή από κει κι έπειτα ως προς τον τρόπο δειγματοληψίας και την έκφραση του εκτιμητή, ήταν ικανή να φέρει αλλαγή στις ιδιότητες του εκτιμητή και συγκεκριμένα να τον καταστήσει πιο αποτελεσματικό σε σχέση με εκείνον της α.τ.δ. ίσου μεγέθους. Αυτό που άλλαξε με την αναδιάταξη των μονάδων του πληθυσμού, ήταν η κατασκευή, ή η δομή, του πληθυσμού.

Γενικότερα, μπορούμε να δώσουμε τις παρακάτω τρεις περιπτώσεις για τη δομή ενός πληθυσμού και τα συμπεράσματα που ισχύουν κάθε φορά για τη συστηματική δειγματοληψία (βλ. Cochran 1977, παράγρ. 8.9 ή Lohr 2010, παράγρ. 5.5).

### 5.6.1 Τυχαίοι Πληθυσμοί

Εάν η λίστα των μελών του πληθυσμού είναι διατεταγμένη βάσει ενός κριτηρίου που δεν σχετίζεται με το χαρακτηριστικό $Y$ της έρευνας, τότε ο πληθυσμός θεωρείται τυχαίος. Για παράδειγμα, εάν $Y$ είναι το ατομικό εισόδημα ή το ποσοστό ανεργίας για τους κατοίκους μιας πόλης, και οι κάτοικοι είναι διαθέσιμοι σε μια λίστα με αλφαβητική σειρά, τότε ο πληθυσμός μπορεί να θεωρηθεί τυχαίος και στις δύο περιπτώσεις, αφού δεν υπάρχει λόγος να πιστεύουμε ότι το πρώτο γράμμα του επωνύμου ενός κατοίκου υποδηλώνει μεγαλύτερο ή μικρότερο εισόδημα και ομοίως πιθανότητα να είναι άνεργος ή όχι. Το ίδιο θα ίσχυε, αν η διάταξη ήταν με βάση τα τελευταία 5 ψηφία του τηλεφωνικού αριθμού. Ένας τυχαίος πληθυσμός θα έχει γραφική παράσταση όπως στο Σχήμα 5.2.



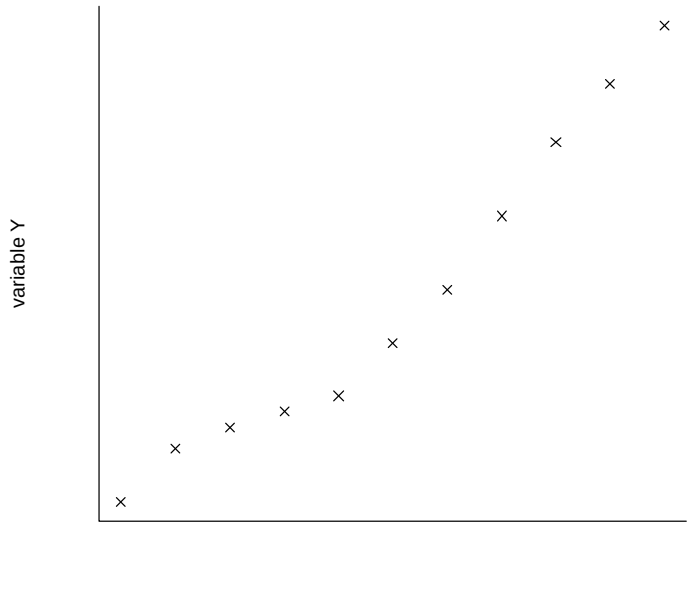

***Σχήμα 5.2*** *Γραφική αναπαράσταση τυχαίου πληθυσμού.*

Όταν ο πληθυσμός είναι τυχαίος, τότε η συστηματική είναι αρκετά πιθανό να συμπεριφέρεται όπως η α.τ.δ. Αυτό συμβαίνει γιατί αναμένουμε να ισχύει $\rho_w = 0$ και κατά συνέπεια $\mathrm{Var}(\bar{X}_{sy}) = \mathrm{Var}(\bar{X})$ (από τη σχέση (5.6)). Συνεπώς, σε τυχαίους πληθυσμούς, ακόμα κι αν εφαρμοστεί συστηματική δειγματοληψία αντί για την α.τ.δ., εφαρμόζουμε τους τύπους από την απλή τυχαία για τους υπολογισμούς των εκτιμητών, των τυπικών σφαλμάτων της εκτίμησης, του διαστήματος εμπιστοσύνης κτλ.

Η ακρίβεια των εκτιμητών $\bar{X}_{sy}$ και $\bar{X}$ είναι ίση, αλλά το πλεονέκτημα του ερευνητή είναι ότι για την υλοποίηση της συστηματικής δειγματοληψίας θα απαιτούνται λιγότερα στοιχεία ως προς το δειγματοληπτικό πλαίσιο και την κατασκευή του.

### 5.6.2 Πληθυσμοί με αύξουσα ή φθίνουσα διάταξη

Όταν οι τιμές του πληθυσμού $Y_i$ $(i = 1,2,\dots,N)$ είναι διατεταγμένες κατά αύξουσα ή φθίνουσα σειρά μεγέθους, τότε αναμένεται η συστηματική δειγματοληψία να είναι πιο αποτελεσματική από την απλή τυχαία. Είναι για παράδειγμα αρκετά συχνό, τα οικονομικά στοιχεία να καταγράφονται από το υψηλότερο μέχρι το χαμηλότερο ποσό. Οι πληθυσμοί αυτοί λέγονται πληθυσμοί με θετική αυτοσυσχέτιση (positive autocorrelation), με κύριο γνώρισμα ότι στοιχεία του πληθυσμού σε κοντινές θέσεις τείνουν να έχουν πιο κοντινές τιμές ως προς το χαρακτηριστικό $Y$ από ότι στοιχεία που είναι σε πιο μακρινές θέσεις μεταξύ τους. Το Σχήμα 5.3 δείχνει ένα παράδειγμα πληθυσμού με διάταξη.




***Σχήμα 5.3*** *Γραφική αναπαράσταση πληθυσμού με αύξουσα διάταξη.*

Για τους πληθυσμούς αυτούς, επειδή συστηματική – ανά ίση απόσταση – επιλογή μονάδων για το δειγματοληπτικό πλαίσιο (οριζόντιος άξονας) σημαίνει ταυτόχρονα και συστηματική κάλυψη των τιμών του χαρακτηριστικού $Y$ (κάθετος άξονας), το δείγμα που προκύπτει είναι αντιπροσωπευτικό. Για το λόγο αυτό, η συστηματική δειγματοληψία υπερτερεί της α.τ.δ. κατά την οποία, λόγω του μεγάλου αριθμού δυνατών δειγμάτων, ορισμένα από αυτά δεν θα είναι αντιπροσωπευτικά. Άρα ισχύει:

$$\mathrm{Var}(\bar{X}_{sy}) < Var(\bar{X})$$

για τις περιπτώσεις ενός διατεταγμένου πληθυσμού. Πρακτικά, όπως έχουμε δει και στο Παράδειγμα 5.6, αυτό σημαίνει ότι ο ερευνητής, πριν από τη διεξαγωγή της έρευνας, και εφόσον έχει διαθέσιμα στοιχεία, αναδιατάσσει τις μονάδες του πληθυσμού βάσει μιας βοηθητικής μεταβλητής που σχετίζεται πολύ έντονα με την $Y$, ώστε να προσεγγίσει τη μορφή του αυτοσυσχετισμένου πληθυσμού. Όσο μεγαλύτερη η συσχέτιση, τόσο καταλληλότερη είναι η βοηθητική μεταβλητή και τόσο πιο βελτιωμένα θα είναι τα αποτελέσματα ως προς την ακρίβεια.

Η διακύμανση και το τυπικό σφάλμα του εκτιμητή $\bar{X}_{sy}$ θα είναι μικρότερα σε σχέση με τα αντίστοιχα από την α.τ.δ. Παρόλα αυτά, η εκτίμησή τους απαιτεί εκτίμηση του $\rho_w$, ή ισοδύναμα του $S^2_{wsy}$. Συχνά στην πράξη, όταν αυτό δεν είναι εφικτό, χρησιμοποιούνται οι τύποι από την α.τ.δ. Στην περίπτωση αυτή, ο ερευνητής έχει υπόψη του ότι υπερεκτιμά τα τυπικά σφάλματα των εκτιμητών ή ότι προσφέρει ένα διάστημα εμπιστοσύνης με μεγαλύτερο μήκος από το αληθινό. Το γεγονός αυτό, ανάλογα με το αντικείμενο της έρευνας, μπορεί να έχει σημαντικές επιπτώσεις. Για παράδειγμα, όταν το διάστημα εμπιστοσύνης αντιπροσωπεύει τα όρια ελέγχου σε μια γραμμή παραγωγής, ενδέχεται η γραμμή παραγωγής να βγει εκτός ελέγχου, π.χ. ένα προϊόν με μικρότερο βάρος κατά τη συσκευασία, αλλά τα όρια ελέγχου που δώσαμε να μην ανιχνεύουν την αλλαγή.

## Παράδειγμα 5.7

Οι ετήσιες εισπράξεις των 24 καταστημάτων μιας περιοχής, σε χιλιάδες χρηματικές μονάδες, είναι: 54, 16, 250, 8, 145, 62, 64, 55, 87, 23, 60, 120, 18, 29, 320, 160, 102, 12, 28, 280, 130, 45, 74, 340. Σημειώνεται ότι η διάταξη των καταστημάτων έχει γίνει με βάση την αλφαβητική σειρά των ιδιοκτητών των καταστημάτων.

Έστω ότι ενδιαφερόμαστε να επιλέξουμε ένα συστηματικό δείγμα $n = 6$ καταστημάτων με σκοπό την εκτίμηση του μέσου ετήσιου ποσού εισπράξεων των καταστημάτων της περιοχής.

Ο Πίνακας 5.4 δίνει τα 4 δυνατά συστηματικά δείγματα που θα μπορούσαν να επιλεγούν από τον πληθυσμό, τη δειγματική μέση τιμή τους και τη δειγματική διακύμανση.

| **Αφετηρία** | **Δείγμα** | **Μέση τιμή** | **Διακύμανση** |
|---|---|---|---|
| $i = 1$ | 54 145 87 18 102 130 | 89.33333 | 2251.067 |
| $i = 2$ | 16 62 23 29 12 45 | 31.16667 | 362.1667 |
| $i = 3$ | 250 64 60 320 28 74 | 132.6667 | 14650.67 |
| $i = 4$ | 8 55 120 160 280 340 | 160.5 | 16505.5 |

***Πίνακας 5.4*** *Συστηματικά δείγματα και οι ιδιότητές τους.*

Παρατηρούμε ότι το δεύτερο δείγμα τυχαίνει να περιλαμβάνει καταστήματα με μικρές εισπράξεις και προσφέρει μια μικρή μέση τιμή εκτίμησης για τον πληθυσμό, ενώ το αντίθετο συμβαίνει για το 4ο συστηματικό δείγμα. Η μέση τιμή του πληθυσμού είναι 103.42, άρα επιλέγοντας το 2ο ή το 4ο συστηματικό δείγμα θα υποεκτιμούσαμε ή θα υπερεκτιμούσαμε την πληθυσμιακή μέση τιμή.

Οι διακυμάνσεις στο εσωτερικό του κάθε δείγματος διαφέρουν αρκετά. Οι δύο ακραίες τιμές λαμβάνονται για τους ίδιους λόγους στα δείγματα με αφετηρία 2 και 4.

Η διακύμανση $S^2_{wsy}$ προκύπτει ως ο μέσος όρος των 4 διασπορών της τελευταίας στήλης του Πίνακα 5.4 και είναι $S^2_{wsy} = 8442.351$. Επίσης, από τα πλήρη στοιχεία του πληθυσμού, η διακύμανση είναι $S^2 = 9827.906$ και, με τη βοήθεια της (5.2), προκύπτει ότι $\mathrm{Var}(\bar{X}_{sy}) = 2383.117$, ενώ ένα α.τ.δ. ίσου μεγέθους θα είχε $\mathrm{Var}(\bar{X}) = \frac{\left(1-\frac{6}{24}\right)S^2}{6} = 1228.488$, δηλ. αρκετά μικρότερη.

Διατάσσοντας τα στοιχεία του πληθυσμού από τη μικρότερη προς τη μεγαλύτερη μέτρηση, επαναλαμβάνουμε τους ίδιους υπολογισμούς με το (i).

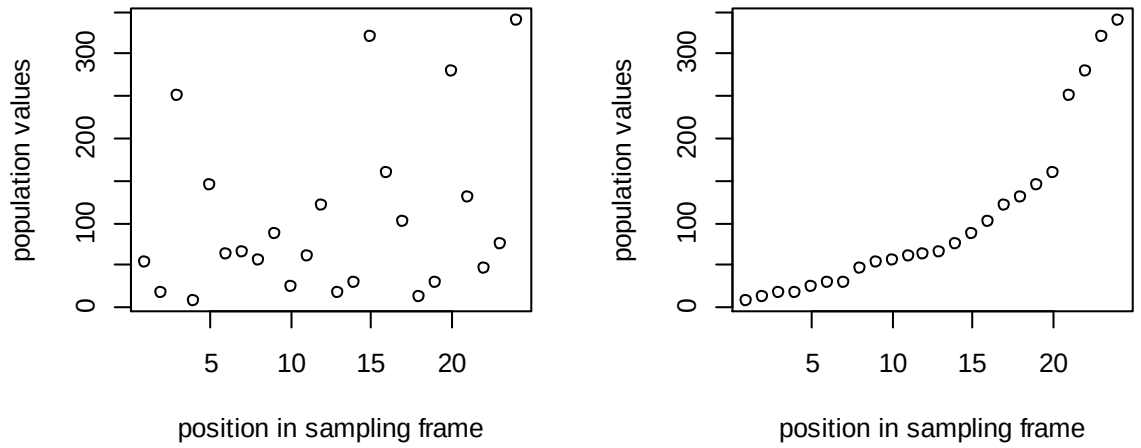
Τα αντίστοιχα με τον Πίνακα 5.4 στοιχεία των δειγμάτων δίνονται στον Πίνακα 5.5, και παρατηρούμε ότι όλα τα δυνατά δείγματα περιέχουν ομοιόμορφα εκπροσώπους από όλα τα διαφορετικά στρώματα καταστημάτων ως προς τις εισπράξεις. Οι μέσες τιμές τους δεν παρουσιάζουν μεγάλες αποκλίσεις μεταξύ τους και οι διακυμάνσεις είναι όλες αρκετά μεγάλες, γεγονός που επιβεβαιώνει την ετερογένεια των δειγμάτων.

| **Αφετηρία** | **Δείγμα** | **Μέση τιμή** | **Διακύμανση** |
|---|---|---|---|
| $i = 1$ | 8 23 54 64 120 250 | 86.5 | 7922.3 |
| $i = 2$ | 12 28 55 74 130 280 | 96.5 | 9771.1 |
| $i = 3$ | 16 29 60 87 145 320 | 109.5 | 12749.9 |
| $i = 4$ | 18 45 62 102 160 340 | 121.1667 | 13941.77 |

***Πίνακας 5.5*** *Συστηματικά δείγματα για τον διατεταγμένο πληθυσμό.*

Η διακύμανση του πληθυσμού προφανώς παραμένει $S^2 = 9827.906$, αλλά η διακύμανση $S^2_{wsy}$ για τον διατεταγμένο πληθυσμό είναι $S^2_{wsy} = \left(\frac{1}{4}\right)(7922.3 + 9771.1 + 12749.9 + 13941.77) = 11096.27$, μεγαλύτερη από την $S^2$ και η διακύμανση του συστηματικού δείγματος μέσω της (5.2) είναι $\mathrm{Var}(\bar{X}_{sy}) = 171.5181$.

Συνεπώς, μετά τη διάταξη του πληθυσμού κατά αύξουσα τάξη, η διακύμανση του εκτιμητή του συστηματικού δείγματος μειώθηκε κατά 13 φορές περίπου. Το Σχήμα 5.4 δείχνει τη γραφική παράσταση των 24 τιμών του πληθυσμού ως προς τα δύο δειγματοληπτικά πλαίσια: Με αλφαβητική σειρά των ιδιοκτητών των καταστημάτων και με αύξουσα σειρά ως προς τις εισπράξεις.




***Σχήμα 5.4*** *Η γραφική παράσταση των τιμών του πληθυσμού πριν και μετά τη διάταξη.*

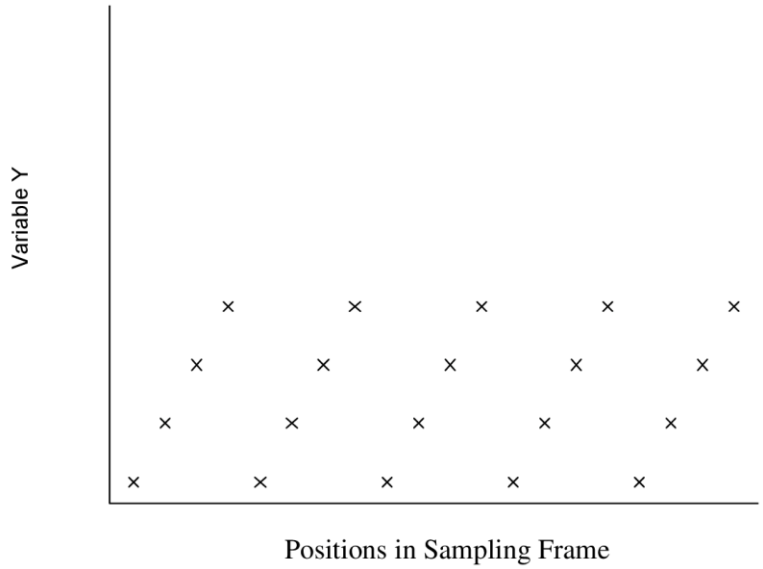
Στο συγκεκριμένο παράδειγμα, η διάταξη έγινε εύκολα, γιατί ήταν γνωστές όλες οι τιμές του πληθυσμού. Στην πράξη, ένας τρόπος να πετύχουμε την αύξουσα τάξη είναι να κατατάξουμε τα μέλη του πληθυσμού με βάση μια μεταβλητή που σχετίζεται με τις εισπράξεις τους, π.χ. εισπράξεις προηγούμενης χρονιάς.

### 5.6.3 Πληθυσμοί με περιοδικότητα

Πληθυσμοί για τους οποίους η διάταξη των μελών τους παρουσιάζει περιοδικότητα ως προς το χαρακτηριστικό $Y$, επισημαίνονται κατά τη μελέτη της συστηματικής δειγματοληψίας, γιατί η εφαρμογή της μεθόδου σε αυτούς είναι προβληματική και μπορεί να οδηγήσει σε κακή εκτίμηση. Εάν ο πληθυσμός παρουσιάζει περιοδική ή κυκλική τάση και το βήμα της συστηματικής ταυτιστεί ή είναι ακέραιο πολλαπλάσιο της περιόδου του πληθυσμού, τότε το συστηματικό δείγμα που θα επιλεγεί θα αποτελείται από $n$ παρόμοιες μεταξύ τους τιμές και συνεπώς δεν θα είναι αντιπροσωπευτικό του πληθυσμού. Για παράδειγμα, το Σχήμα 5.5 δίνει τη γραφική παράσταση των 20 τιμών του πληθυσμού:

1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4

Η περίοδος για τον πληθυσμό αυτόν είναι 4 και, αν το βήμα της περιόδου επιλεγεί να είναι επίσης 4, τότε για αφετηρία π.χ. $i = 3$, το συστηματικό δείγμα θα είναι $s = (3,3,3,3)$ με $\bar{X}_{sy} = 3$ και διακύμανση μηδέν. Η διακύμανση όμως του εκτιμητή $\bar{X}_{sy}$ είναι $\mathrm{Var}(\bar{X}_{sy}) = \frac{2}{3}$, ακριβώς όση και η συνολική μεταβλητότητα του πληθυσμού, δηλ. το συστηματικό δειγματοληπτικό σχέδιο για το παράδειγμα αυτό έχει τυπικό σφάλμα όσο ένα απλό τυχαίο δείγμα μεγέθους $n = 1$ από τον πληθυσμό.




***Σχήμα 5.5*** *Γραφική αναπαράσταση πληθυσμού με περιοδικότητα.*

Γενικά, η συστηματική μέθοδος δειγματοληψίας χρησιμοποιείται αρκετά συχνά στην πράξη, και ιδιαίτερα όταν:

(i) Ο ερευνητής ενδιαφέρεται για την επιλογή ενός αντιπροσωπευτικού δείγματος από τον πληθυσμό, και δεν έχει μια λίστα ή στοιχεία/πόρους για να κατασκευάσει το δειγματοληπτικό πλαίσιο.

(ii) Όταν υπάρχει η εκτίμηση ότι ο πληθυσμός είναι θετικά αυτοσυσχετισμένος.

(iii) Στο τελικό στάδιο ενός σύνθετου δειγματοληπτικού σχεδίου, όπου τα πλεονεκτήματα στην ευκολία της υλοποίησης της συστηματικής δειγματοληψίας μπορούν να αξιοποιηθούν. Άλλες μέθοδοι δειγματοληψίας που βελτιώνουν τα τυπικά σφάλματα εφαρμόζονται στα αρχικά στάδια μιας σύνθετης δειγματοληψίας και εξασφαλίζουν την ακρίβεια των εκτιμητών.

## Βιβλιογραφικές Αναφορές

Barnett, V. (2002). *Sample Survey: Principles and methods* (3rd Edition). London: Arnold.

Cochran, W. G. (1977). *Sampling techniques* (3rd Edition). New York: John Wiley and Sons.

Levy P.S. and Lemeshow, S. (1999). *Sampling of populations. Methods and applications* (3rd Edition). New York: John Wiley and Sons.

Lohr S. L. (2010). *Sampling: Design and analysis* (2nd Edition). Boston, Mass:Brooks/Cole, Cengage Learning.

Rao, P.S.R.S. (2000). *Sampling methodologies with applications*. Boca Raton, Fla: Chapman and Hall/CRC.

# Κεφάλαιο 6 - ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΚΑΤΑ ΟΜΑΔΕΣ

**Σύνοψη**

*Μια άλλη μέθοδος δειγματοληψίας είναι η κατά ομάδες δειγματοληψία (cluster sampling). Σύμφωνα με τη μέθοδο αυτή, η επιλογή του δείγματος γίνεται σε επίπεδο ομάδων (η έννοια της ομάδας θα γίνει αντιληπτή στη συνέχεια), και όχι στοιχειωδών μελών του πληθυσμού. Οι λόγοι που ο ερευνητής θα επιλέξει τη μέθοδο αυτή είναι παρόμοιοι με εκείνους κατά τη συστηματική δειγματοληψία. Πιο συγκεκριμένα, είτε δεν είναι διαθέσιμη μία λίστα των μελών του πληθυσμού, είτε είναι χρονοβόρα ή με μεγάλο κόστος η κατασκευή της. Η ευκολία συνεπώς, και όχι η ακρίβεια, είναι το κύριο χαρακτηριστικό της μεθόδου. Σύμφωνα με τη δειγματοληψία κατά ομάδες, ο ερευνητής έχει πρόσβαση στον πληθυσμό μέσω των ομάδων των μελών του. Επιλέγεται συνεπώς αρχικά ένα δείγμα ομάδων, με ίσες ή άνισες πιθανότητες, και στη συνέχεια, είτε όλα τα στοιχεία των ήδη επιλεγμένων ομάδων, είτε ένα υποσύνολό τους, συμπεριλαμβάνεται στο τελικό δείγμα. Γενικά, από την επιλογή των αρχικών ομάδων μέχρι και την επιλογή των στοιχειωδών μονάδων του πληθυσμού, μπορεί να μεσολαβούν αρκετά στάδια. Το πλήθος των σταδίων, ή των φάσεων της δειγματοληπτικής έρευνας όπως επίσης αποκαλούνται, εξαρτάται από τη γεωγραφική έκταση του πληθυσμού της έρευνας (εάν καλύπτει μία αρκετά μεγάλη περιοχή ή όχι) και το δειγματοληπτικό πλαίσιο σύμφωνα με το οποίο ορίζονται οι ομάδες και η διαμέρισή τους σε μικρότερες ομάδες. Η δειγματοληψία κατά ομάδες χρησιμοποιείται πάρα πολύ συχνά στην πράξη, και ιδιαίτερα σε μεγάλης κλίμακας έρευνες, λόγω του χαμηλού κόστους.*

## 6.1. Εισαγωγή

Όλα τα δειγματοληπτικά σχέδια που έχουν αναπτυχθεί έως τώρα υλοποιούνται βάσει ενός δειγματοληπτικού πλαισίου που επιτρέπει την άμεση πρόσβαση του ερευνητή στην πληθυσμιακή μονάδα. Για παράδειγμα, εάν η έρευνα αφορά την εκτίμηση του ποσοστού των παιδιών μιας πόλης που δεν έχουν κάνει τα βασικά τους εμβόλια, τότε, προκειμένου να εφαρμόσουμε απλή τυχαία, στρωματοποιημένη ή συστηματική δειγματοληψία, υποθέτουμε ότι υπάρχει μια λίστα των παιδιών ή ένα δειγματοληπτικό πλαίσιο που μας επιτρέπει την πρόσβαση στα μέλη του πληθυσμού, δηλ. τα παιδιά της πόλης. Τα παιδιά επιλέγονται κατευθείαν μέσω του δειγματοληπτικού πλαισίου και στη συνέχεια ερχόμαστε σε επικοινωνία μαζί τους, προκειμένου να γίνει η παρατήρηση ή η συλλογή των δεδομένων. Αντίστοιχα, για μια έρευνα με σκοπό την εκτίμηση των ποσοστών προτίμησης των πολιτικών κομμάτων στις προσεχείς εκλογές, τα μέλη του πληθυσμού είναι οι ψηφοφόροι της χώρας. Προκειμένου συνεπώς να εφαρμοστεί μία από τις παραπάνω μεθόδους δειγματοληψίας, θα πρέπει ομοίως να διαθέτουμε ένα δειγματοληπτικό πλαίσιο στο επίπεδο των ψηφοφόρων.

Στην πράξη, ειδικά σε έρευνες μεγάλης έκτασης ή έρευνες σε ανθρώπινους πληθυσμούς, π.χ. τους κατοίκους μιας πόλης ή χώρας, τους ψηφοφόρους, τα παιδιά μιας πόλης ή τους ηλικιωμένους κτλ., είναι αρκετά δύσκολο – και μερικές φορές αδύνατο – να ορίσουμε και να υιοθετήσουμε ένα δειγματοληπτικό πλαίσιο που να αποτελείται από τα στοιχειώδη μέλη του πληθυσμού, και συνεπώς να είναι άμεση η πρόσβαση σε αυτά. Είναι, για παράδειγμα, δύσκολο να έχουμε ένα δειγματοληπτικό πλαίσιο που να παρέχει πληροφορία για τα παιδιά της πόλης, π.χ. ηλικίας 6 ετών, τη χρονική περίοδο που διεξάγεται η έρευνα με σκοπό την εκτίμηση του ποσοστού των μη εμβολιασθέντων παιδιών με τα βασικά εμβόλια. Ανάλογα, στο παράδειγμα με την έρευνα για την εκτίμηση των ποσοστών των πολιτικών κομμάτων στις προσεχείς εκλογές, είναι δύσκολο να αποκτήσουμε πρόσβαση σε μια λίστα όλων των ψηφοφόρων της πόλης. Οι λόγοι που καθιστούν δύσκολη την κατασκευή ενός δειγματοληπτικού σχεδίου με αυτή την ιδιότητα είναι στην πλειονότητα των περιπτώσεων λόγοι κόστους και χρόνου. Για παράδειγμα, η διαδικασία αυτή μπορεί να απαιτεί άντληση στοιχείων από κάποια δημόσια αρχή, αίτηση για άδεια χρήσης στοιχείων από μια βάση δεδομένων η οποία μπορεί να είναι χρονοβόρα, ή, ακόμα περισσότερο, να μην είναι εφικτή η πρόσβαση σε στοιχεία, λόγω του προσωπικού απορρήτου.

Επιπλέον, ακόμα κι αν μια τέτοια λίστα υπήρχε από προηγούμενη έρευνα, η χρήση της ενέχει κινδύνους, και συγκεκριμένα, εισαγωγή σφαλμάτων μη κάλυψης στην έρευνα, επειδή η λίστα των μελών του πληθυσμού αλλάζει με την πάροδο του χρόνου, είτε λόγω μετακίνησης πληθυσμών, είτε με ανανεώσεις από γεννήσεις/θανάτους, είτε λόγω ωρίμανσης του πληθυσμού. Π.χ. για τους ψηφοφόρους μιας χώρας, η λίστα, εάν ήταν διαθέσιμη, θα έπρεπε να ενημερώνεται κάθε χρόνο με τις νέες εγγραφές των κατοίκων που

συμπληρώνουν το 18ο έτος ηλικίας και αποκτούν δικαίωμα ψήφου και με τις διαγραφές εκείνων που έχουν στερηθεί τα πολιτικά τους δικαιώματα (π.χ. λόγω καταδίκης) ή που έχουν εν τω μεταξύ πεθάνει κτλ.

Η δημιουργία, συνεπώς, και η επικαιροποίηση ενός δειγματοληπτικού πλαισίου με άμεση πρόσβαση στα μέλη του πληθυσμού είναι μια δαπανηρή διαδικασία και, για τον λόγο αυτόν, πολύ συχνά η έρευνα διεξάγεται με βάση ένα δειγματοληπτικό πλαίσιο που προσεγγίζει τα μέλη του πληθυσμού με έμμεσο τρόπο. Ο έμμεσος τρόπος είναι η πρόσβαση στα μέλη του πληθυσμού μέσω μεγαλύτερων ομάδων στις οποίες ανήκει το κάθε μέλος. Για παράδειγμα, τα Δημοτικά Σχολεία στα οποία φοιτούν τα παιδιά της πόλης είναι ομάδες που περιέχουν τα παιδιά, τα μέλη του πληθυσμού. Τα εκλογικά τμήματα στα οποία ανήκει ο κάθε ψηφοφόρος, η πόλη, ή το τετράγωνο στο οποίο κατοικεί ένας κάτοικος είναι επίσης ομάδες ψηφοφόρων. Η μεγαλύτερη ομάδα μέσω της οποίας συμπεριλαμβάνεται το κάθε μέλος του πληθυσμού στο δειγματοληπτικό πλαίσιο λέγεται **πρωτογενής δειγματοληπτική μονάδα (primary sampling unit, psu)** και η απλή, στοιχειώδης μονάδα του πληθυσμού λέγεται **στοιχείο (element)**. Τα μέλη της psu λέγονται **δευτερογενείς δειγματοληπτικές μονάδες (secondary sampling units, ssu)** και συνήθως είναι οι στοιχειώδεις μονάδες του πληθυσμού.

Προφανώς, για κάθε πρόβλημα υπάρχουν ενδεχομένως πολλοί διαφορετικοί τρόποι να οριστούν οι ομάδες και κατά συνέπεια το δειγματοληπτικό πλαίσιο, όπως στο παράδειγμα των ψηφοφόρων, όπου οι ομάδες μπορεί να είναι τα εκλογικά τμήματα, τα οικοδομικά τετράγωνα των πόλεων ή οι ταχυδρομικοί κωδικοί.

### 6.1.1 Ορισμός και ορολογία

#### Ορισμός

**Δειγματοληψία κατά ομάδες ή συστάδες (cluster sampling)** ονομάζεται η μέθοδος δειγματοληψίας που εφαρμόζεται όταν το δειγματοληπτικό πλαίσιο που υιοθετείται για τον πληθυσμό αποτελείται από μια λίστα πρωτογενών μονάδων αντί για μια λίστα των μεμονωμένων στοιχείων του πληθυσμού.

Για το υπόλοιπο του κεφαλαίου οι όροι psu, cluster, ομάδα ή συστάδα θα χρησιμοποιούνται με το ίδιο ακριβώς νόημα. Επίσης, ποσότητες με δείκτη cl θα αναφέρονται στη μέθοδο δειγματοληψίας κατά ομάδες.

Κατά τη μέθοδο δειγματοληψίας κατά ομάδες, η επιλογή του δείγματος γίνεται (α) στο επίπεδο των psu και (β) στο εσωτερικό των ομάδων, στο επίπεδο των μεμονωμένων στοιχείων του πληθυσμού. Ειδικότερα, για το δεύτερο βήμα, υπάρχουν δύο βασικές περιπτώσεις:

(i) Όλο το περιεχόμενο των psu που επιλέχθηκαν στο πρώτο στάδιο επιλέγεται με τη σειρά του για να ανήκει στο τελικό δείγμα. Στην περίπτωση αυτή, όταν μια ομάδα επιλεγεί, αυτομάτως όλα τα στοιχεία που ανήκουν σε αυτή την ομάδα συμμετέχουν στην έρευνα. Δεν λαμβάνει χώρα καμία περαιτέρω δειγματοληψία, παρά μόνον εκείνη στο επίπεδο των ομάδων, και η μέθοδος δειγματοληψίας λέγεται **δειγματοληψία κατά ομάδες σε ένα στάδιο (one-stage cluster sampling)**.

(ii) Από την κάθε ομάδα (psu) που επιλέχθηκε στο πρώτο στάδιο επιλέγεται περαιτέρω ένα υποσύνολό του από στοιχεία (ssu) για να συμμετέχει στο δείγμα της έρευνας. Η μέθοδος στην περίπτωση αυτή ονομάζεται **δειγματοληψία κατά ομάδες σε δύο στάδια (two-stage cluster sampling)**.

Για το παράδειγμα με το ποσοστό των παιδιών που δεν έχουν κάνει τα βασικά εμβόλια, έστω ότι το δειγματοληπτικό πλαίσιο είναι τα σχολεία της περιοχής που μας ενδιαφέρει να εξετάσουμε. Αν επιλεγούν π.χ. $n = 10$ σχολεία, και στη συνέχεια όλοι οι μαθητές των 10 αυτών σχολείων συμπεριληφθούν στο δείγμα και εξεταστούν ως προς τα εμβόλια, θα πρόκειται για ένα δειγματοληπτικό σχέδιο κατά ομάδες σε ένα στάδιο. Εάν αποφασιστεί ότι από το σύνολο των μαθητών των 10 σχολείων μόνο ένα ποσοστό π.χ. το 20% αυτών θα επιλεγεί και θα εξεταστεί ως προς τα εμβόλια, τότε το δειγματοληπτικό σχέδιο είναι μια δειγματοληψία κατά ομάδες σε δύο στάδια.

Είναι φανερό από τον ορισμό και τον τρόπο υλοποίησης της κατά ομάδες δειγματοληψίας, ότι πρόκειται για ένα **ιεραρχικό** δειγματοληπτικό σχέδιο. Η δειγματοληψία ξεκινά από το υψηλότερο επίπεδο μονάδων (psu) και συνεχίζεται προς το χαμηλότερο (ssu).

Μια γενίκευση είναι εφικτή, όταν τα επίπεδα διαβάθμισης της πρωτογενούς μονάδας μέχρι το στοιχείο του πληθυσμού είναι περισσότερα από 2. Π.χ. για την ομάδα Σχολείο στο παράδειγμα με τα παιδιά της πόλης, μπορεί να υπάρξει πιο λεπτή διαμέριση και αντί για τα 2 επίπεδα: Σχολείο/Μαθητής να υπάρξουν τέσσερα

επίπεδα: Σχολείο/Τάξη/Τμήμα/Μαθητής. Σε αυτή την περίπτωση, μπορεί θεωρητικά να διεξαχθεί μια δειγματοληψία σε τέσσερα στάδια (τόσα όσα είναι τα επίπεδα διαβάθμισης). Δειγματοληπτικά σχέδια τα οποία κάνουν χρήση ενός δειγματοληπτικού πλαισίου με λίστα ομάδων με περισσότερα από 2 επίπεδα διαβάθμισης λέγονται **δειγματοληπτικά σχέδια σε πολλά στάδια (multistage sampling).**

### 6.1.2 Πλεονεκτήματα και μειονεκτήματα της δειγματοληψίας κατά ομάδες

Κλείνοντας την εισαγωγή και την περιγραφή της δειγματοληψίας κατά ομάδες, ανακεφαλαιώνουμε με τα πλεονεκτήματα και τα μειονεκτήματα της μεθόδου.

Το πρώτο πλεονέκτημα της μεθόδου είναι οι πρακτικοί λόγοι. Υπάρχει διαθέσιμο ένα δειγματοληπτικό πλαίσιο ως προς τις ομάδες και όχι ως προς τα στοιχεία του πληθυσμού.

Από τον ορισμό και τον τρόπο επιλογής του δείγματος, συμπεραίνεται ότι στην πράξη η άντληση στοιχείων για τον πληθυσμό σύμφωνα με τη μέθοδο κατά ομάδες θα είναι απαραίτητη όχι για όλες τις ομάδες, αλλά μόνο για όσες έχουν επιλεγεί στο δείγμα. Π.χ. για το παράδειγμα με τους ψηφοφόρους και τα ποσοστά των πολιτικών κομμάτων, αν επιλέξουμε ως ομάδες τα εκλογικά τμήματα, θα χρειαστεί να ζητήσουμε στοιχεία μόνο για τα εκλογικά τμήματα που επιλέχθηκαν κατά το πρώτο στάδιο και όχι για το σύνολο των εκλογικών τμημάτων του πληθυσμού. Άρα μεγαλύτερη ταχύτητα κατά τη συλλογή στοιχείων.

Από τον ορισμό επίσης της μεθόδου, προκύπτει ότι τα στοιχεία του πληθυσμού που θα συμπεριληφθούν στο δείγμα είναι συγκεντρωμένα σε ένα αριθμό ομάδων που επιλέχθηκε στο πρώτο στάδιο. Επειδή συνήθως οι ομάδες ορίζονται με βάση τη γεωγραφική εγγύτητα, π.χ. ίδιο οικοδομικό τετράγωνο ή ίδιο σχολείο ή ίδιο ταχυδρομικό κώδικα κτλ., η διεξαγωγή της δειγματοληψίας κατά ομάδες απαιτεί τόσες επί τόπου επισκέψεις, όσες είναι και οι ομάδες. Στο παράδειγμα με τα παιδιά της πόλης, θα απαιτούνταν δέκα επισκέψεις στα δέκα σχολεία του δείγματος, και από το κάθε σχολείο θα συλλέγονταν στη συνέχεια στοιχεία για ένα μεγάλο αριθμό παιδιών. Αντίθετα, εάν για τη συλλογή στοιχείων από ίδιο αριθμό παιδιών εφαρμόζαμε την α.τ.δ. θα έπρεπε να γίνει χωριστά ο εντοπισμός και η επίσκεψη για το κάθε παιδί του δείγματος. Οι μαθητές στη δεύτερη περίπτωση μπορεί να είναι διάσπαρτοι σε όλα τα σχολεία της πόλης, γεγονός που απαιτεί περισσότερο χρόνο, ταξίδια και γενικά κόστος για την έρευνα.

Συνολικά, όλοι οι παραπάνω λόγοι, λόγοι πρακτικοί και ευκολίας, αλλά και λόγοι οικονομίας και ταχύτητας, κάνουν τη δειγματοληψία κατά ομάδες να πλεονεκτεί και να εφαρμόζεται αρκετά συχνά στην πράξη.

Ταυτόχρονα, οι ίδιοι λόγοι που κάνουν την κατά ομάδες δειγματοληψία να πλεονεκτεί κατά την εφαρμογή της έρευνας, έχουν ως αποτέλεσμα το μειονέκτημα της μεθόδου ως προς την ακρίβεια των εκτιμήσεων. Για παράδειγμα, θα ήταν δαπανηρό και χρονοβόρο να κάνουμε 500 διαφορετικές επισκέψεις σε κατοικίες μιας πόλης, για τη συλλογή με α.τ.δ. δείγματος μεγέθους $n = 500$, με σκοπό την εκτίμηση του μέσου εισοδήματος των κατοίκων μιας πόλης. Θα ήταν πιο εύκολο και σύντομο, να συλλέξουμε στοιχεία για τον ίδιο αριθμό κατοίκων εξετάζοντας μόνο 10 οικοδομικά τετράγωνα (υποθέτοντας ότι σε κάθε τετράγωνο κατοικούν 50 ενήλικες). Παράλληλα όμως, το α.τ.δ. των $n = 500$ κατοίκων θα εξασφαλίζει πιθανότατα καλύτερη κάλυψη της πόλης σε σύγκριση με τα 10 οικοδομικά τετράγωνα της δειγματοληψίας σε συστάδες, με αποτέλεσμα ένα πιο αντιπροσωπευτικό δείγμα και κατά συνέπεια μια εκτίμηση με μεγαλύτερη στατιστική ακρίβεια.

### 6.1.3 Παραδείγματα δειγματοληψίας κατά ομάδες

Στον Πίνακα 6.1, δίνονται ορισμένα πρακτικά παραδείγματα πληθυσμών χωρισμένων σε ομάδες, τα στοιχεία του πληθυσμού και τα πιθανά ερωτήματα της έρευνας που θα μπορούσε να εφαρμοστεί κάνοντας την αντίστοιχη θεώρηση των ομάδων.

| Ομάδα/Συστάδα | Στοιχείο | Πρόβλημα |
|---|---|---|
| **Οικοδομικά Τετράγωνα** | Νοικοκυριό | Εκτίμηση<br>του μέσου ποσού δαπανών για διατροφή ανά οικογένεια,<br>του ποσοστού οικογενειών που δεν πηγαίνουν διακοπές,<br>της μέσης κατανάλωσης νερού,<br>του ποσοστού οικογενειών χωρίς ιατροφαρμακευτική περίθαλψη. |
| **Οικοδομικά Τετράγωνα** | Κάτοικος | Εκτίμηση<br>του ποσοστού ανεργίας,<br>του ποσοστού κατοίκων που πάσχουν από μια ασθένεια,<br>του ποσοστού καπνιστών,<br>του μέσου εισοδήματος. |
| **Σχολεία** | Μαθητής | Εκτίμηση<br>της μέσης βαθμολογίας στα μαθήματα,<br>του ποσοστού μαθητών με καταγωγή εκτός από Ελληνική,<br>του ποσοστού παιδιών με χωρισμένους γονείς. |
| **Νοσοκομεία** | Ασθενής | Εκτίμηση<br>του συνολικού αριθμού ασθενών που επισκέπτονται τα Επείγοντα Περιστατικά των Νοσοκομείων της χώρας,<br>του μέσου χρόνου νοσηλείας των νοσηλευομένων. |
| **Νοσοκομεία** | Νοσηλευτής | Εκτίμηση<br>του μέσου χρόνου απασχόλησης με υπερωρία των νοσηλευτών,<br>του ποσοστού των ανδρών νοσηλευτών<br>της μέσης ηλικίας των νοσηλευτών. |
| **Δημόσια Νοσοκομεία** | Γιατρός | Εκτίμηση<br>του μέσου χρόνου συνολικής εβδομαδιαίας απασχόλησης των γιατρών στα δημόσια Νοσοκομεία της Χώρας,<br>του μέσου χρόνου προϋπηρεσίας των γιατρών που εργάζονται σε δημόσια νοσοκομεία. |
| **Γραμμή σελίδας βιβλίου** | Λέξη | Εκτίμηση<br>του ποσοστού ορθογραφικών λαθών ανά σελίδα του βιβλίου,<br>του συνολικού αριθμού λέξεων στη σελίδα. |
| **Πακέτο συσκευασίας προϊόντων** | Προϊόν | Εκτίμηση<br>του ποσοστού ελαττωματικών προϊόντων.<br>του μέσου αριθμού προϊόντων ανά πακέτο. |

***Πίνακας 6.1*** *Παραδείγματα πληθυσμών και ερευνητικά ερωτήματα.*

### 6.1.4 Δειγματοληψία κατά ομάδες και Στρωματοποιημένη Δειγματοληψία

Η Δειγματοληψία κατά ομάδες, σύμφωνα με τον ορισμό και την περιγραφή της, παρουσιάζει ομοιότητες με τη στρωματοποιημένη δειγματοληψία, γιατί τόσο η κατά ομάδες, όσο και η στρωματοποιημένη, κάνουν χρήση ομάδων των στοιχείων του πληθυσμού. Ωστόσο, οι δύο μέθοδοι δειγματοληψίας διαφέρουν σημαντικά και ως προς τον ορισμό και τον τρόπο υλοποίησης, αλλά και ως προς τις ιδιότητες των εκτιμητών που προκύπτουν. Ο Πίνακας 6.2 δίνει τις ομοιότητες και τις διαφορές των δύο αυτών μεθόδων δειγματοληψίας.

| **Στρωματοποιημένη Δειγματοληψία** | **Δειγματοληψία κατά ομάδες** |
|---|---|
| Κάθε στοιχείο του πληθυσμού ανήκει σε ένα ακριβώς στρώμα. | Κάθε στοιχείο του πληθυσμού ανήκει σε μία ακριβώς ομάδα ή cluster. |
| Ο πληθυσμός χωρίζεται σε $L$ στρώματα και το κάθε στρώμα έχει ένα αριθμό στοιχείων. | Ο πληθυσμός αποτελείται από ομάδες και η κάθε συστάδα έχει έναν αριθμό στοιχείων |
| Καλύπτονται όλα τα στρώματα με ανεξάρτητη δειγματοληψία στο εσωτερικό του καθενός. | Δεν καλύπτονται όλες οι ομάδες: επιλέγεται ένας αριθμός ομάδων. |
| Η δειγματοληψία γίνεται στο επίπεδο στοιχείων του πληθυσμού. | Η δειγματοληψία γίνεται στο επίπεδο των ομάδων. |
| Το πλήθος των στρωμάτων είναι μικρό και ο αριθμός των στοιχείων σε κάθε στρώμα συνήθως μεγάλος. | Το πλήθος των ομάδων είναι συνήθως μεγάλο και ο αριθμός των στοιχείων σε κάθε ομάδα μικρός. |
| Ο χωρισμός γίνεται από τον ερευνητή, και τα στρώματα που προκύπτουν δεν έχουν απαραίτητα γεωγραφική ή άλλου είδους εγγύτητα. Π.χ. κριτήριο χωρισμού σε στρώματα: φύλο, ηλικιακή ομάδα, επίπεδο εκπαίδευσης, εισόδημα. | Ο χωρισμός είναι δεδομένος γιατί είναι διαθέσιμος. Συνήθως τα μέλη της κάθε συστάδας έχουν και φυσική εγγύτητα. Π.χ. μένουν στο ίδιο τετράγωνο, φοιτούν στο ίδιο σχολείο, εργάζονται στο ίδιο νοσοκομείο. |
| Τα στρώματα κατασκευάζονται έτσι ώστε να επιτυγχάνεται ομοιογένεια στο εσωτερικό του κάθε στρώματος και διαφορετική μέση τιμή από στρώμα σε στρώμα. | Επιθυμούμε οι ομάδες να είναι ετερογενείς. Επειδή ορισμένες μόνο ομάδες θα επιλεγούν για το δείγμα, η δειγματοληψία κατά ομάδες θα είναι πιο ακριβής δειγματοληπτική μέθοδος εάν οι ομάδες είναι αντιπροσωπευτικές του πληθυσμού. |
| Μικρότερα τυπικά σφάλματα για τους εκτιμητές συγκρινόμενα με αυτά της α.τ.δ. | Μια μέθοδος που μπορεί να οδηγήσει σε μεγαλύτερα τυπικά σφάλματα συγκρινόμενη με την α.τ.δ. |
| Ο σκοπός της στρωματοποίησης του πληθυσμού είναι η βελτίωση των τυπικών σφαλμάτων των εκτιμητών. | Ο σκοπός της υιοθέτησης των ομάδων είναι η ευκολία στην πρόσβαση και η μείωση του κόστους και του χρόνου διεξαγωγής της έρευνας. |

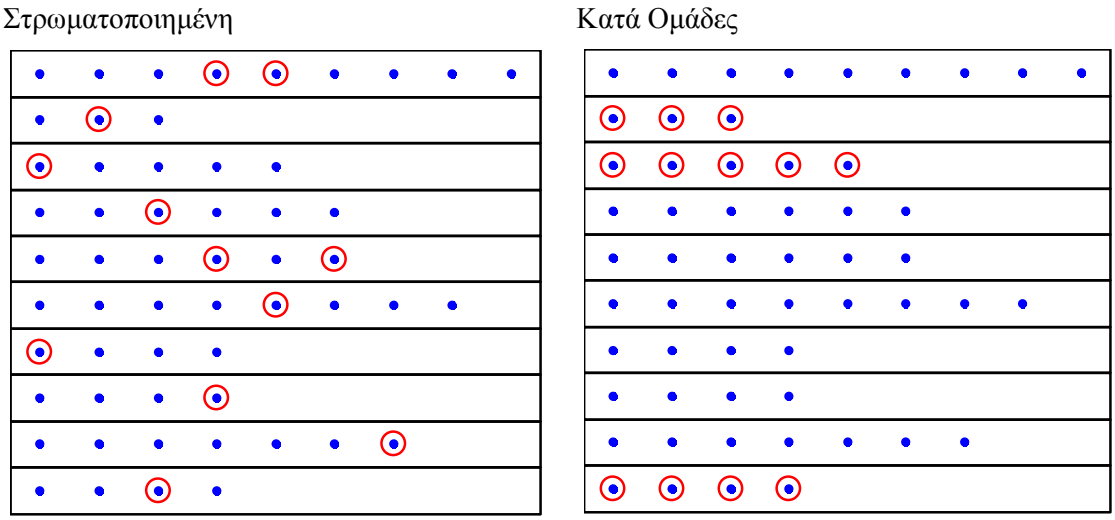
***Πίνακας 6.2*** *Ομοιότητες και διαφορές της στρωματοποιημένης και της δειγματοληψίας κατά ομάδες.*

Ένας τρόπος να παραστήσουμε γραφικά τον τρόπο επιλογής του δείγματος μέσω της στρωματοποιημένης και αντίστοιχα μέσω της κατά ομάδων, δίνεται με το Σχήμα 6.1. Για τον ίδιο πληθυσμό που είναι χωρισμένος σε στρώματα/ομάδες και παριστάνονται με τις γραμμές του πίνακα, στη μεν περίπτωση της στρωματοποιημένης λαμβάνεται ένα δείγμα από το κάθε στρώμα, ενώ στην κατά ομάδες μέθοδο λαμβάνονται μόνο ορισμένες ομάδες (λαμβάνονται εξ ολοκλήρου για τη μέθοδο κατά ομάδες σε ένα στάδιο).

Τα δύο γραφήματα, για τη στρωματοποιημένη και την κατά ομάδες δειγματοληψία αντίστοιχα, έγιναν κάνοντας χρήση του πακέτου 'animation' της R. Για περισσότερα παραδείγματα και κατανόηση της νέας μεθόδου δειγματοληψίας, δίνεται η εντολή που παράγει γραφήματα ενός δείγματος κατά ομάδες.

```
> library(animation)
> sample.cluster(pop = ceiling(10 * runif(10, 0.2, 1)), size = 3,
p.col = c('blue','red'), p.cex = c(1, 3))
```

Αλλάζοντας τις τιμές του ορίσματος `p.cex`, π.χ. `p.cex = c(1, 6)`, αλλάζει ο αριθμός των psu που επιλέγονται. Επίσης, δίνοντας διαφορετικές τιμές στη συνάρτηση `runif`, αλλάζει το πλήθος και το μέγεθος των psu.




***Σχήμα 6.1*** *Γραφική αναπαράσταση στρωματοποιημένου και κατά ομάδες δείγματος.*

## 6.2. Συμβολισμός

Στην παράγραφο αυτή δίνουμε τον συμβολισμό που θα χρησιμοποιήσουμε για τη μελέτη του προβλήματος εκτίμησης ποσοτήτων του πληθυσμού με τη βοήθεια ενός δείγματος κατά ομάδες. Στον Πίνακα 6.3, δίνεται ο απαραίτητος συμβολισμός για τη δειγματοληψία κατά ομάδες και η εξήγηση της κάθε ποσότητας που εισάγεται. Όπως και στη στρωματοποιημένη και στη συστηματική δειγματοληψία, γίνεται χρήση διπλού δείκτη για το κάθε στοιχείο (πληθυσμού και δείγματος) ώστε να δηλώνει τον α/α του στοιχείου και ως προς την ομάδα (psu) και ως προς το στοιχείο της ομάδας (ssu). Ο συμβολισμός του Πίνακα 6.3 καλύπτει τη δειγματοληψία κατά ομάδες, σε ένα και σε δύο στάδια.

| **Σύμβολο** | **Ερμηνεία** |
|---|---|
| $M$ | Αριθμός των Ομάδων |
| $K_i$ | Μέγεθος της $i$ ομάδας ή διαφορετικά πλήθος στοιχείων (ssu) που περιέχει η $i$ psu.<br>Αν $N$ είναι το μέγεθος του πληθυσμού τότε:<br>$$K_1 + K_2 + \ ... + K_M = N$$ |

| Σύμβολο | Ερμηνεία |
|---|---|
| $Y_{ij}$ | Η τιμή του χαρακτηριστικού $Y$ για το μέλος του πληθυσμού που ανήκει στην ομάδα $i$ και έχει αύξοντα αριθμό $j$. |
| $U_i = \{Y_{i1}, Y_{i2}, \dots, Y_{iK_i}\}$ | Τα στοιχεία του πληθυσμού για την ομάδα $i$. |
| $\bar{U}_i = \frac{1}{K_i}\sum_{j=1}^{K_i} Y_{ij}$ | Ο μέσος της $i$ ομάδας για το χαρακτηριστικό $Y$. |
| $t_i = \sum_{j=1}^{K_i} Y_{ij}$ | Το άθροισμα της ομάδας $i$ για το χαρακτηριστικό. |
| $Y_T = \sum_{i=1}^{M} t_i$ | Το άθροισμα του πληθυσμού για το χαρακτηριστικό. |
| $\bar{Y} = \frac{1}{N} Y_T$ | Ο μέσος του πληθυσμού για το χαρακτηριστικό. |
| $S_i^2 = \frac{1}{K_i - 1}\sum_{j=1}^{K_i}\left(Y_{ij} - \bar{U}_i\right)^2$ | Η διακύμανση στο εσωτερικό της ομάδας $i$. |
| $S^2 = \frac{1}{N-1}\sum_{i=1}^{M}\sum_{j=1}^{K_i}\left(Y_{ij} - \bar{Y}\right)^2$ | Η διακύμανση του πληθυσμού. |
| $m$ | Πλήθος ομάδων που επιλέγονται από τις $M$ συνολικά. |
| $k_i$ | Το μέγεθος του δείγματος των ssu που επιλέγονται από την ομάδα $i$. Ισχύει: $k_i \leq K_i$. |
| $X_{ij}$ | Το δειγματικό στοιχείο του πληθυσμού που προέρχεται από την ομάδα $i$ του δείγματος ομάδων και έχει α/α $j$. |
| $\bar{X}_i = \frac{1}{k_i}\sum_{j=1}^{k_i} X_{ij}$ | Ο δειγματικός μέσος της ομάδας $i$. |
| $s_i^2 = \frac{1}{k_i - 1}\sum_{j=1}^{k_i}\left(X_{ij} - \bar{X}_i\right)^2$ | Η δειγματική διακύμανση της ομάδας $i$. |

***Πίνακας 6.3*** *Συμβολισμός για την κατά ομάδες δειγματοληψία.*

## 6.3. Δειγματοληψία κατά ομάδες σε ένα στάδιο με ίσες πιθανότητες

Θεωρούμε την εκτίμηση των παραμέτρων του πληθυσμού: πληθυσμιακό μέσο $\bar{Y}$, σύνολο του πληθυσμού $Y_T$, το ποσοστό $P$ και αριθμό των μελών $A$ του πληθυσμού τα οποία ανήκουν σε μια κατηγορία ή κατέχουν μια

ιδιότητα. Η σύνδεση μεταξύ των προβλημάτων εκτίμησης των πληθυσμιακών παραμέτρων όπως έχει αναπτυχθεί στα προηγούμενα κεφάλαια ισχύει και για την περίπτωση της δειγματοληψίας κατά ομάδες. Για τον σκοπό αυτό, η ανάπτυξη του προβλήματος της εκτίμησης μιας άγνωστης ποσότητας του πληθυσμού με τη βοήθεια ενός δείγματος που προέκυψε με τη μέθοδο των ομάδων, θα γίνει μόνο για την περίπτωση του πληθυσμιακού μέσου $\bar{Y}$.

Θεωρούμε, συνεπώς, το πρόβλημα της εκτίμησης του μέσου $\bar{Y}$ ενός πληθυσμού μεγέθους $N$, με τη βοήθεια ενός δείγματος που έχει επιλεγεί με μια δειγματοληψία κατά ομάδες. Ειδικότερα, υποθέτουμε την κατά ομάδες σε ένα σταδιο δειγματοληψία, κατά την οποία, είτε όλες οι μονάδες μιας συστάδας περιέχονται στο δείγμα, είτε καμία. Σύμφωνα με τον συμβολισμό του Πίνακα 6.3, θα είναι $k_i = K_i$ και $\bar{X}_i = \bar{U}_i$, δηλ. η δειγματική μέση τιμή για την ομάδα $i$ ταυτίζεται με την πληθυσμιακή, επειδή, στο εσωτερικό της $i$ ομάδας, πραγματοποιείται απογραφή και όχι δειγματοληψία.

Για την εκτίμηση των πληθυσμιακών ποσοτήτων κάτω από τη δειγματοληψία κατά ομάδες, διακρίνουμε τις περιπτώσεις (i) οι ομάδες να είναι ίσου μεγέθους και (ii) οι ομάδες να είναι άνισου μεγέθους, καθώς και περιπτώσεις ως προς τον τρόπο επιλογής των ομάδων. Θεωρώντας την πιο απλή περίπτωση αρχικά ως προς τον τρόπο επιλογής των ομάδων, και ειδικότερα υποθέτοντας ίσες πιθανότητες κατά την επιλογή, εξετάζουμε τις περιπτώσεις (i) και (ii). Οι ίσες πιθανότητες επιλογής των psu ισοδυναμούν με υιοθέτηση της α.τ.δ.

### 6.3.1 Ομάδες ίσου μεγέθους

Η υπόθεση ότι όλες οι ομάδες έχουν το ίδιο πλήθος στοιχείων εξετάζεται γιατί είναι μια περίπτωση πληθυσμού χωρισμένου σε ομάδες που διαπιστώνουμε σε ορισμένες περιπτώσεις πληθυσμών στην πράξη. Π.χ. τάξεις μαθητών είναι παρόμοιου μεγέθους, οικοδομικά τετράγωνα μιας πόλης με ίση πυκνότητα δόμησης, ράφια βιβλίων σε μια βιβλιοθήκη, πακέτα συσκευασιών προϊόντων κ.ο.κ. Κάτω από την υπόθεση αυτή, ο γενικός συμβολισμός που δώσαμε στον Πίνακα 6.3 απλοποιείται περισσότερο. Τα μεγέθη $K_i$ είναι ίσα μεταξύ τους, και έστω ακόμα ότι:

$$K_i = K \ (i = 1,2, \dots, M)$$

απ΄ όπου προκύπτει $N = MK$. Αντίστοιχα, για το μέγεθος του δείγματος, θα ισχύει: $n = mK$.

Κάτω από αυτές τις υποθέσεις, και λαμβάνοντας ένα απλό τυχαίο δείγμα $m$ ομάδων του πληθυσμού από τις $M$ συνολικά, ένας εκτιμητής της πληθυσμιακής μέσης τιμής θεωρείται ο απλός μέσος του δείγματος. Αναλυτικότερα:

$$\bar{X}_{cl} = \hat{\bar{Y}} = \frac{1}{n}\sum_{i=1}^{m} t_i \qquad (6.1)$$

Αρχικά, διαπιστώνουμε ότι, λόγω της ισότητας του μεγέθους των ομάδων, ο εκτιμητής αυτός ταυτίζεται με:

$$\bar{X}_{cl} = \frac{1}{m}\sum_{i=1}^{m} \bar{U}_i$$

δηλ. $\bar{X}_{cl}$ είναι ένας απλός μέσος όρος των μέσων $m$ ομάδων ενός α.τ. δείγματος από τις $M$ συνολικά ομάδες.

Λαμβάνοντας υπόψη την τελευταία μορφή του $\bar{X}_{cl}$ και θεωρώντας στη συνέχεια το διάνυσμα $\{\bar{U}_1, \bar{U}_2, \dots, \bar{U}_M\}$ ως τον πληθυσμό από τον οποίο επιλέγεται ένα α.τ. δείγμα μεγέθους $m$, μπορούμε να συμπεράνουμε τις ιδιότητες του εκτιμητή $\bar{X}_{cl}$ κάνοντας χρήση των αποτελεσμάτων από την α.τ.δ. Πιο συγκεκριμένα, μέσω των Προτάσεων 2.2, 2.3 και του Πορίσματος 2.1, συνάγεται το ακόλουθο αποτέλεσμα.

**Πρόταση 6.1**

Για τον εκτιμητή $\bar{X}_{cl}$ που δίνεται από την (6.1), ισχύει:

(i) Είναι αμερόληπτος εκτιμητής του $\bar{Y}$.

(ii) Η διακύμανση του εκτιμητή $\bar{X}_{cl}$, δίνεται από τη σχέση:

$$\mathrm{Var}(\bar{X}_{cl}) = \frac{1}{m}(1-f)\sum_{i=1}^{M}\frac{(\bar{U}_i - \bar{Y})^2}{M-1} \tag{6.2}$$

όπου $f = m/M$, το πηλίκο δείγματος για την α.τ.δ. στο επίπεδο των ομάδων.

(iii) Η εκτιμώμενη διακύμανση του $\bar{X}_{cl}$ είναι

$$\hat{\mathrm{Var}}(\bar{X}_{cl}) = \frac{1}{m}(1-f)\sum_{i=1}^{m}\frac{(\bar{U}_i - \bar{X}_{cl})^2}{m-1} \tag{6.3}$$

Η ποσότητα $\sum_{i=1}^{M}\frac{(\bar{U}_i-\bar{Y})^2}{M-1}$ είναι η διακύμανση των πληθυσμιακών μέσων όλων συνολικά των ομάδων, ενώ η αντίστοιχη δειγματική ποσότητα $\sum_{i=1}^{m}\frac{(\bar{U}_i-\bar{X}_{cl})^2}{m-1}$ είναι η διακύμανση των μέσων εκείνων των ομάδων που επιλέχθηκαν στο δείγμα και συνεπώς είναι εφικτό να υπολογιστεί αμέσως μετά τη διεξαγωγή της έρευνας και τη συλλογή των δεδομένων. Και οι δύο ποσότητες είναι αντίστοιχες των γνωστών $S^2$ και $s^2$ που εμφανίζονται στις εκφράσεις της διακύμανσης και της εκτιμώμενης διακύμανσης του $\bar{X}$ για την α.τ.δ..

Αξίζει να σημειωθεί ότι η διακύμανση του εκτιμητή $\bar{X}_{cl}$ όπως δίνεται μέσω της (6.2), εξαρτάται αποκλειστικά από τη διακύμανση μεταξύ των μέσων των ομάδων, και όχι από τη διακύμανση στο εσωτερικό των ομάδων. Αυτό οφείλεται στο γεγονός ότι στο εσωτερικό των ομάδων πραγματοποιείται απογραφή, και όχι δειγματοληψία, συνεπώς δεν υπάρχει στατιστικό σφάλμα.

Κάτω από κατάλληλες προϋποθέσεις, ανάλογες με εκείνες που συζητήθηκαν στην α.τ.δ., υπολογίζεται το διάστημα εμπιστοσύνης της μέσης τιμής του πληθυσμού, κάνοντας χρήση της κανονικής κατανομής. Αναλυτικότερα, για $m$ μεγάλο αριθμό, το 99% διάστημα εμπιστοσύνης για τη μέση τιμή του πληθυσμού είναι:

$$\left(\bar{X}_{cl} - z_{\alpha/2}\sqrt{\hat{\mathrm{Var}}(\bar{X}_{cl})},\ \ \bar{X}_{cl} + z_{\alpha/2}\sqrt{\hat{\mathrm{Var}}(\bar{X}_{cl})}\right)$$

όπου η εκτιμώμενη διακύμανση του εκτιμητή $\hat{\mathrm{Var}}(\bar{X}_{cl})$ δίνεται από την (6.3), και $z_{\alpha/2}$ είναι το $\alpha/2$-άνω εκατοστιαίο σημείο της τυπικής κανονικής κατανομής.

Για $m$ μικρό αριθμό,το διάστημα εμπιστοσύνης υπολογίζεται με τη βοήθεια των εκατοστιαίων σημείων της $t$ κατανομής. Συγκεκριμένα, θα είναι:

$$\left(\bar{X}_{cl} - t_{m-1,\frac{\alpha}{2}}\sqrt{\hat{\mathrm{Var}}(\bar{X}_{cl})},\ \ \bar{X}_{cl} + t_{m-1,\frac{\alpha}{2}}\sqrt{\hat{\mathrm{Var}}(\bar{X}_{cl})}\right).$$

## Παράδειγμα 6.1

Μια έρευνα, με σκοπό την εκτίμηση του μέσου εισοδήματος των ενηλίκων κατοίκων, διεξάγεται σε μια πόλη η οποία είναι χωρισμένη σε 450 οικοδομικά τετράγωνα. Από προηγούμενες έρευνες, είναι γνωστό ότι ο μέσος αριθμός ενηλίκων κατοίκων ανά τετράγωνο είναι 120, χωρίς μεγάλη απόκλιση μεταξύ των τετραγώνων. Για την έρευνα, επιλέχθηκαν 25 τετράγωνα και ρωτήθηκαν όλοι οι ενήλικες που κατοικούν στα τετράγωνα αυτά ως προς το εισόδημά τους.

Τα δεδομένα της έρευνας μετά τη διεξαγωγή της δίνονται στον Πίνακα 6.4. Με βάση τα στοιχεία της έρευνας, να εκτιμηθεί το μέσο εισόδημα των κατοίκων της πόλης και να δοθεί ένα 95% διάστημα εμπιστοσύνης της εκτίμησης.

| α/α Οικοδομικού Τετραγώνου | Συνολικό Εισόδημα (σε χιλ. χ.μ) | Μέσο Εισόδημα (σε χιλ. χ.μ) | α/α Οικοδομικού Τετραγώνου | Συνολικό Εισόδημα (σε χιλ. χ.μ) | Μέσο Εισόδημα (σε χιλ. χ.μ) |
|---|---|---|---|---|---|
| 1 | 68.8598 | 0.5738 | 14 | 85.4589 | 0.7122 |
| 2 | 119.8689 | 0.9989 | 15 | 117.3347 | 0.9778 |
| 3 | 108.5510 | 0.9046 | 16 | 106.7089 | 0.8892 |
| 4 | 73.3973 | 0.6116 | 17 | 84.3756 | 0.7031 |
| 5 | 82.8918 | 0.6908 | 18 | 114.9435 | 0.9579 |
| 6 | 88.7509 | 0.7396 | 19 | 103.1808 | 0.8598 |
| 7 | 84.5846 | 0.7049 | 20 | 95.8406 | 0.7987 |
| 8 | 62.9579 | 0.5246 | 21 | 121.5084 | 1.0126 |
| 9 | 48.2354 | 0.4020 | 22 | 107.7965 | 0.8983 |
| 10 | 88.2420 | 0.7353 | 23 | 155.7251 | 1.2977 |
| 11 | 60.4749 | 0.5040 | 24 | 88.7936 | 0.7399 |
| 12 | 90.5054 | 0.7542 | 25 | 91.0228 | 0.7585 |
| 13 | 138.0609 | 1.1505 | **Άθροισμα** | **2388.1** | **19.9006** |

***Πίνακας 6.4*** *Δεδομένα για το εισόδημα των κατοίκων των 25 οικοδομικών τετραγώνων.*

Η εντολή για την εισαγωγή των δεδομένων στην R είναι

```
> tinc<-c(68.8598, 119.8689, 108.5510, 73.3973, 82.8918, 88.7509,
84.5846, 62.9579, 48.2354, 88.2420, 60.4749, 90.5054, 138.0609,
85.4589, 117.3347, 106.7089, 84.3756, 114.9435, 103.1808, 95.8406,
121.5084, 107.7965, 155.7251, 88.7936, 91.0228)
```

από την οποία προκύπτει η τρίτη στήλη του Πίνακα 6.4 με το μέσο εισόδημα ανά τετράγωνο ως:

```
> minc<-tinc/120
```

Σύμφωνα με τον συμβολισμό του κεφαλαίου, για τα δεδομένα του προβλήματος ισχύει: $M = 450$ ο αριθμός των ομάδων, $m = 25$ οι ομάδες που επιλέγονται με α.τ.δ. και $K = 120$ διότι υποθέτουμε ότι οι ομάδες είναι ίσου μεγέθους. Επίσης, προκύπτει ότι $N = 450 * 120 = 54000$, ενώ τα δεδομένα της δεύτερης στήλης του Πίνακα 6.4 είναι τα αθροίσματα $t_i$ $(i = 1,2, \dots ,25)$ και τα δεδομένα της τρίτης στήλης είναι οι μέσοι όροι $\overline{U}_i$ $(i = 1,2, \dots ,25)$.

Στη συνέχεια, εφαρμόζοντας την (6.1):

```
> est<-sum(tinc)/(120*25)
> est
[1] 0.7960234
```

λαμβάνουμε ότι η εκτίμηση του μέσου εισοδήματος των κατοίκων για την πόλη είναι 796.02 χρηματικές μονάδες.

Ισοδύναμα, η (6.1) υπολογίζεται ως η μέση τιμή των μέσων των ομάδων, δηλ.

```
> est<-mean(minc)
> est
[1] 0.7960234
```

Για τη διακύμανση του εκτιμητή, θα υπολογίσουμε τον εκτιμητή του από τη σχέση (6.3), η οποία απαιτεί να γνωρίζουμε τα συγκεντρωτικά στοιχεία των κατοίκων μόνο για τα 25 δειγματοληπτικά τετράγωνα. Για τον ακριβή υπολογισμό της διακύμανσης μέσω της (6.2), χρειάζεται να έχουμε τα αντίστοιχα στοιχεία ως προς το εισόδημα, για όλα τα 450 οικοδομικά τετράγωνα της πόλης. Οι υπολογισμοί για την (6.3) δίνουν:

```
> evar<-(1-25/450)*(1/25)*var(minc)
> evar
[1] 0.001582671
```

άρα η εκτιμώμενη διακύμανση της εκτίμησης ότι το μέσο εισόδημα είναι 0.796 χιλιάδες χρηματικές μονάδες είναι 0.001583. Το τυπικό σφάλμα της εκτίμησης της διασποράς είναι η τετραγωνική ρίζα του αριθμού αυτού, δηλ.

```
> sqrt(evar)
[1] 0.0397828
```

οπότε, το τυπικό σφάλμα της εκτίμησης είναι 0.03978 χιλιάδες χ.μ. ή 39.78 χ.μ.

Το 95% διάστημα εμπιστοσύνης για την εκτίμηση του μέσου εισοδήματος υπολογίζεται μέσω της προσέγγισης της κανονικής κατανομής και κάνοντας χρήση των $t$ εκατοστιαίων σημείων αντί για $z$, επειδή γίνεται εκτίμηση του τυπικού σφάλματος και το α.τ.δ. είναι μεγέθους 25. Με τη βοήθεια των κατάλληλων συναρτήσεων και εντολών της R, λαμβάνουμε

```
> c(est-qt(.975, 24)*sqrt(evar),est+qt(.975, 24)*sqrt(evar))
[1] 0.7139157 0.8781311
```

Άρα ένα 95% διάστημα εμπιστοσύνης για το μέσο εισόδημα είναι $(0.7139, 0.8781)$ σε χιλιάδες χρημ. μονάδες ή ισοδύναμα το μέσο εισόδημα των ενηλίκων κατοίκων της πόλης ανήκει στο διάστημα (713.9, 878.1) χρηματικές μονάδες με πιθανοτητα σφάλματος 5%.

Για την εφαρμογή της μεθόδου δειγματοληψίας κατά ομάδες, υπάρχουν εξειδικευμένα πακέτα στην R, μεταξύ άλλων τα `TeachingSampling`, `Sampling` και `survey`. Θα εξερευνήσουμε ορισμένα από αυτά στη συνέχεια του κεφαλαίου, με πιο απαιτητικά παραδείγματα.

### 6.3.2 Σύγκριση δειγματοληψίας κατά ομάδες σε ένα στάδιο και ίσο μέγεθος ομάδων, με την α.τ.δ.

Όπως και με κάθε άλλη μέθοδο δειγματοληψίας που έχουμε μελετήσει, ενδιαφέρον παρουσιάζει το ερώτημα πώς συγκρίνεται η μέθοδος δειγματοληψίας κατά ομάδες με την α.τ.δ. ως προς την αποτελεσματικότητα. Η παρακάτω πρόταση δίνει το σχετικό αποτέλεσμα.

**Πρόταση 6.2**

Ικανή και αναγκαία συνθήκη ώστε η δειγματοληψία κατά ομάδες σε ένα στάδιο με ίσο μέγεθος ομάδων να είναι πιο αποτελεσματική από την απλή τυχαία είναι:

$$\bar{S}^2 \geq S^2 \tag{6.4}$$

όπου $\bar{S}^2$ είναι η μέση διακύμανση των ομάδων των πληθυσμών.

**Απόδειξη**

Για την απόδειξη ακολουθούμε βήματα ανάλογα με την απόδειξη της Πρότασης 5.4, όπου αρχικά θεωρούμε τη διαφορά των δύο διακυμάνσεων προς σύγκριση και στη συνέχεια κάνουμε χρήση της γνωστής σχέσης ΑΝΑΔΙΑ των στοιχείων του πληθυσμού ώστε οι δύο εκφράσεις να γίνουν συγκρίσιμες.

Πιο αναλυτικά, η διαφορά των δύο διακυμάνσεων είναι

$$\mathrm{Var}(\bar{X}_{srs}) - \mathrm{Var}(\bar{X}_{cl}) = \frac{(1-\frac{n}{N})}{n} S^2 - \frac{1}{m}(1-f)\sum_{i=1}^{M}\frac{(\bar{U}_i - \bar{Y})^2}{M-1}$$

και η σχέση ΑΝΑΔΙΑ για τα στοιχεία του πληθυσμού, ανάλογη με την (3.15) που δείξαμε για την περίπτωση του στρωματοποιημένου πληθυσμού, δίνει

$$(N-1)S^2 = \sum_{i=1}^{M}(K-1)\,S_i^2 + \sum_{i=1}^{M} K(\bar{U}_i - \bar{Y})^2$$

από την οποία, λύνοντας ως προς τον τελευταίο όρο $\sum_{i=1}^{M}(\bar{U}_i - \bar{Y})^2$ και αντικαθιστώντας στη διαφορά των διακυμάνσεων που είναι προς σύγκριση, προκύπτει:

$$\mathrm{Var}(\bar{X}_{srs}) - \mathrm{Var}(\bar{X}_{cl}) = \frac{(1-f)M(K-1)}{mK(M-1)}(\bar{S}^2 - S^2) \qquad (6.5)$$

όπου,

$$\bar{S}^2 = \frac{1}{M}\sum_{i=1}^{M} S_i^2$$

η μέση διακύμανση των πρωτογενών μονάδων. Η σχέση (6.5) αποδεικνύει το ζητούμενο αποτέλεσμα γιατί το κλάσμα των σταθερών όρων με το οποίο πολλαπλασιάζεται η διαφορά $\bar{S}^2 - S^2$ είναι πάντα θετικό ■

Ερμηνεύοντας την Πρόταση 6.2 συμπεραίνουμε ότι: κατάλληλα δειγματοληπτικά πλαίσια για την εφαρμογή της δειγματοληψίας με συστάδες θα είναι εκείνα, για τα οποία οι ομάδες παρουσιάζουν μεγάλη διακύμανση στο εσωτερικό τους ως προς το χαρακτηριστικό $Y$ που μελετούμε. Η διακύμανση πρέπει να είναι μεγάλη για την κάθε ομάδα, έτσι ώστε ο μέσος όρος αυτών να είναι μεγαλύτερος από τη συνολική διακύμανση του πληθυσμού. Διαφορετικά, μπορούμε να πούμε ότι πρακτικά, η υιοθέτηση της δειγματοληψίας κατά ομάδες, (που πολύ συχνά είναι δεδομένη και δεν υπάρχει εναλλακτικό δειγματοληπτικό πλαίσιο για τον ερευνητή) θα δώσει αποτελεσματικούς εκτιμητές, εάν οι ομάδες είναι ετερογενείς.

Άρα, αντίθετα με τη στρωματοποιημένη δειγματοληψία όπου επιδιώκουμε ομοιογένεια στο εσωτερικό για μέγιστη αποτελεσματικότητα ως προς την α.τ.δ., στην κατά ομάδες δειγματοληψία οι συστάδες επιθυμούμε να είναι ετερογενείς. Η διαφορά με τη στρωματοποιημένη είναι αναμενόμενη, δοθέντος ότι ο τρόπος δειγματοληψίας είναι διαφορετικός. Κατά τη στρωματοποιημένη λαμβάνεται ένα δείγμα από κάθε στρώμα και, όσο πιο ομοιογενές είναι το στρώμα, τόσο πιο ακριβής θα είναι και ο εκτιμητής που αντιστοιχεί στο στρώμα αυτό. Αντίθετα, στη δειγματοληψία με συστάδες, δε λαμβάνεται δείγμα εντός της συστάδας, λαμβάνεται όλη η συστάδα. Συστάδα με ομοιογένεια συνεπάγεται ένα δείγμα με πολλές παρόμοιες παρατηρήσεις. Ενώ μια συστάδα με ετερογένεια, έχει μεγαλύτερη πιθανότητα να προσφέρει ένα αντιπροσωπευτικό δείγμα του πληθυσμού και άρα μια πιο ακριβή εκτίμηση.

## Παρατήρηση 6.1

Η συνθήκη (6.4) ταυτίζεται με την αντίστοιχη συνθήκη (5.8) που αποδείξαμε για τη συστηματική δειγματοληψία. Το αποτέλεσμα αυτό δεν είναι τυχαίο, αντίθετα προκύπτει ως άμεσο συμπέρασμα εξαιτίας του γενικότερου αποτελέσματος ότι η συστηματική δειγματοληψία μπορεί να θεωρηθεί ως ειδική περίπτωση της δειγματοληψίας κατά ομάδες. Πράγματι, αν θεωρήσουμε ως δειγματοληπτικό πλαίσιο το σύνολο των $k$ δυνατών συστηματικών δειγμάτων και επιλέξουμε με α.τ.δ. ένα εξ αυτών, τότε η μέθοδος ταυτίζεται με τη δειγματοληψία κατά ομάδες με ίσο μέγεθος ομάδων κατά την οποία $M = k$ και $m = 1$. Συνεπώς, όλα τα μαθηματικά αποτελέσματα που αποδείχθηκαν για τη συστηματική δειγματοληψία θα μπορούσαν να εξαχθούν ως ειδική περίπτωση της δειγματοληψίας κατά ομάδες σε ένα στάδιο με ίσο μέγεθος ομάδων.

Υιοθετώντας **τον συντελεστή συσχέτισης intracluster** που έχει εισαχθεί στην παράγραφο 5.3.2, μπορούμε να δώσουμε μια ισοδύναμη έκφραση για τις σχέσεις (6.2) και (6.5) αντίστοιχα.

Ο συντελεστής συσχέτισης intracluster $\rho_w$ δίνει ένα μέτρο συσχέτισης των στοιχείων του πληθυσμού που ανήκουν στην ίδια συστάδα. Υπενθυμίζουμε ότι ο συντελεστής $\rho_w$ σχετίζεται με τη διασπορά $S^2_{wsy}$ στη συστηματική και παρέχει την ίδια πληροφορία με αυτή (τη διασπορά), αλλά με διαφορετική προσέγγιση, ως προς την ομοιογένεια ή ετερογένεια του εσωτερικού των συστηματικών δειγμάτων. Με τον συμβολισμό του παρόντος κεφαλαίου, η μέση διασπορά $\bar{S}^2$ αντιστοιχεί στην $S^2_{wsy}$ και παρόμοια συμπεράσματα με εκείνα που δόθηκαν στη συστηματική ισχύουν και για τη δειγματοληψία κατά ομάδες. Όσο μεγαλύτερο είναι το $\rho_w$, τόσο μικρότερη τιμή θα λαμβάνει το $\bar{S}^2$ και θα δηλώνει ομάδες με ομοιογένεια στο εσωτερικό τους, και αντίστροφα. Αποδεικνύεται ότι η ακριβής σχέση που συνδέει τις ποσότητες $\rho_w$ και $\bar{S}^2$ για τις ειδικές υποθέσεις της δειγματοληψίας σε συστάδες σε ένα στάδιο με ίσο μέγεθος είναι

$$\rho_w = 1 - \left(\frac{MK}{MK-1}\right)\frac{\bar{S}^2}{S^2}.$$

Η απόδειξη βασίζεται στην εφαρμογή της ανάλυσης διακύμανσης κατά έναν παράγοντα στις μονάδες του πληθυσμού, θεωρώντας ως επίπεδα του παράγοντα τις ομάδες. Για την αναλυτική απόδειξη, παραπέμπουμε στο βιβλίο Cochran (1977), παρ. 9.4.

Ανάλογα με την Πρόταση 5.3 αποδεικνύεται το σχετικό αποτέλεσμα για την κατά ομάδες δειγματοληψία.

### Πρόταση 6.3

Η διακύμανση $\mathrm{Var}(\bar{X}_{cl})$ δίνεται από τον τύπο:

$$\mathrm{Var}(\bar{X}_{cl}) = \frac{(1-f)(MK-1)}{mK^2(M-1)} S^2[1+(K-1)\rho_w] \qquad (6.6)$$

(ομοίως για την απόδειξη, βλ. Cochran (1977), παρ. 9.4).

Η συνθήκη (6.4), για την οποία η κατά ομάδες είναι πιο αποτελεσματική από μια α.τ.δ. ίσου μεγέθους δείγματος, εκφράζεται μέσω του $\rho_w$ κάνοντας χρήση της σχέσης (6.6). Αν θεωρήσουμε τη διαφορά των διακυμάνσεων $\mathrm{Var}(\bar{X}_{cl})$ και $\mathrm{Var}(\bar{X}_{srs})$, αντικαταστήσουμε την (6.6) για τη διακύμανση του $\bar{X}_{cl}$ και κάνουμε πράξεις, καταλήγουμε στην ισοδύναμη με την (6.4) συνθήκη:

$$\rho_w < -\frac{1}{MK-1} \qquad (6.7)$$

Η ερμηνεία και τα συμπεράσματα για την αποτελεσματικότητα του εκτιμητή $\bar{X}_{cl}$ ως προς τις τιμές του $\rho_w$ είναι ακριβώς ανάλογα με εκείνα για τη συστηματική δειγματοληψία. Επιγραμματικά:

(i) Αν ο συντελεστής $\rho_w$ είναι θετικός, δηλ. υπάρχει ομοιογένεια στο εσωτερικό της συστάδας, τότε ο εκτιμητής $\bar{X}_{cl}$ έχει μεγαλύτερο στατιστικό σφάλμα από τον $\bar{X}_{srs}$, λόγω της (6.6). (Σημειώνεται ότι για μεγάλες τιμές του πλήθους $M$ των psu, ισχύει $(MK-1)/K(M-1) \cong 1$).

(ii) Όταν $\rho_w = 0$ τότε ο εκτιμητής $\bar{X}_{cl}$ ταυτίζεται με τον $\bar{X}_{srs}$.

(iii) Εάν $\rho_w < 0$ για το εσωτερικό των ομάδων, τότε ο εκτιμητής $\bar{X}_{cl}$ έχει μικρότερο στατιστικό σφάλμα από τον $\bar{X}_{srs}$.

(iv) Εάν το $M$ είναι μεγάλος αριθμός, τότε η επίδραση της δειγματοληψίας κατά ομάδες ως προς την α.τ. δίνεται από τη σχέση:

$$\mathrm{deff} = 1 + (K-1)\rho_w$$

## Παρατήρηση 6.2

Επειδή στους λόγους εφαρμογής της μεθόδου κατά ομάδες δεν συμπεριλαμβάνεται η αποτελεσματικότητα, αλλά η ευκολία, στην πράξη η μέθοδος χρησιμοποιείται πάρα πολύ συχνά, ανεξάρτητα από το εάν οι συνθήκες (6.4) ή (6.7) πληρούνται. Ωστόσο, είναι σημαντικό η συμπερασματολογία να βασίζεται στη μέθοδο που εφαρμόστηκε κατά τη συλλογή των δεδομένων (συστηματική ή cluster), γιατί εάν συμβαίνει το $\rho_w$ για τον πληθυσμό να είναι μακρυά από το μηδέν, οποιαδήποτε συμπεράσματα εξαχθούν κάνοντας χρήση των τύπων της α.τ.δ. θα είναι λανθασμένα. Το παρακάτω παράδειγμα με το πρόγραμμα intervals από τους Alf & Lohr (2007) βοηθά στην κατανόηση της παρατήρησης αυτής με έναν πιο παραστατικό τρόπο.

## Παράδειγμα 6.2

Κάνοντας χρήση του προγράμματος intervals γραμμένο στην R, διαθέσιμο στη σελίδα http://www.cengagebrain.com/cgi-wadsworth/course_products_wp.pl?fid=M20b&product_isbn_issn=9780495105275&token,

προσομοιώνουμε τις τιμές του πληθυσμού, για τον οποίο υποθέτουμε ότι είναι χωρισμένος σε ομάδες μεγέθους $K = 5$. Συνολικά, οι ομάδες του πληθυσμού είναι $M = 5000$, εκ των οποίων, στο πλαίσιο του πειράματος επιλέγονται $m = 10$. Η επιλογή του δείγματος επαναλαμβάνεται 100 φορές, με σκοπό την κατασκευή 100 διαστημάτων εμπιστοσύνης (ΔΕ) για την πληθυσμιακή μέση τιμή. Το κάθε ΔΕ υπολογίζεται με δύο τρόπους: (i) κάνοντας χρήση της α.τ.δ. για τον υπολογισμό του τυπικού σφάλματος και (ii) κάνοντας χρήση του δειγματοληπτικού σχεδίου που εφαρμόστηκε. Τέλος, το πείραμα επαναλαμβάνεται για 4 διαφορετικές περιπτώσεις της παραμέτρου $\rho_w$ του πληθυσμού.

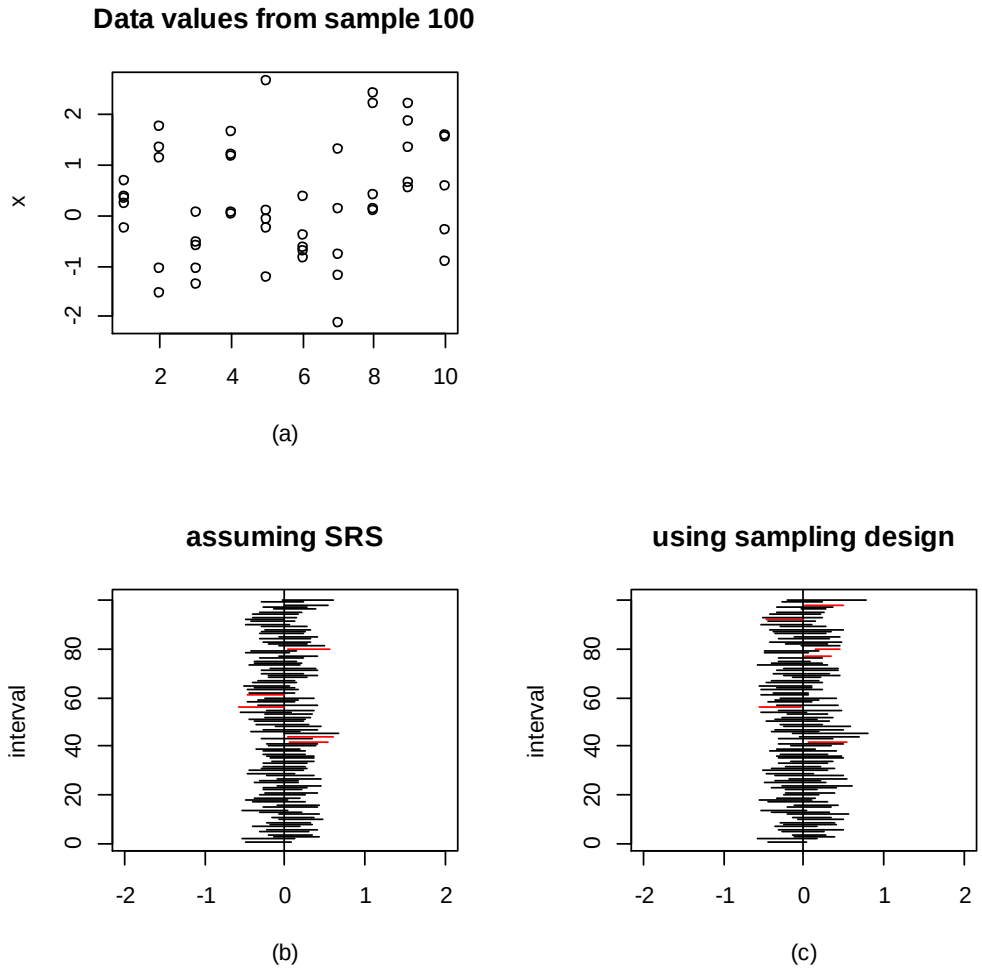
(Α) $\rho_w = 0$

(Β) $\rho_w = 0.25$

(Γ) $\rho_w = 0.5$ και

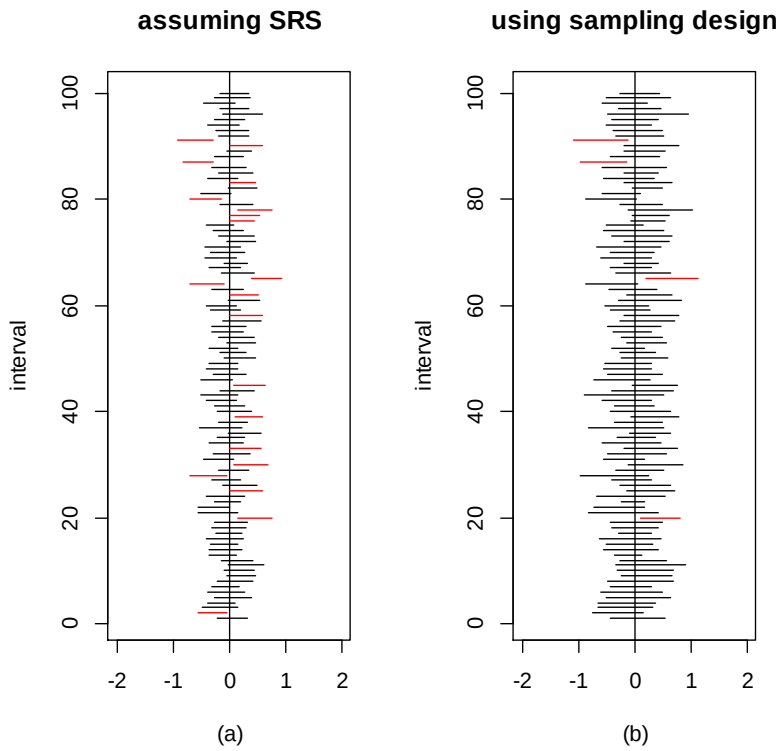
(Δ) $\rho_w = 0.8$

Για την περίπτωση (Α), τα στοιχεία του πληθυσμού δεν παρουσιάζουν συσχέτιση στο εσωτερικό των ομάδων. Στο Σχήμα 6.2 δίνονται τρία γραφήματα με τα αποτελέσματα του πειράματος. Στο γράφημα (a) δίνεται η γραφική παράσταση των στοιχείων ενός από τα 100 δείγματα που έχουν παραχθεί (του τελευταίου), στο (b) τα 100 ΔΕ που κατασκευάστηκαν από τα δείγματα υπολογισμένα σύμφωνα με τους τύπους από την α.τ.δ. και στο (c) με παρόμοιο τρόπο η γραφική παράσταση των ΔΕ, αλλά υπολογισμένα με την cluster δειγματοληψία. Τα διαστήματα με κόκκινο χρώμα είναι εκείνα που δεν περιέχουν την αληθινή τιμή της παραμέτρου του πληθυσμού. Όπως βλέπουμε για την περίπτωση αυτή του πληθυσμού, δεν υπάρχει διαφορά στον αριθμό των ΔΕ με κόκκινο χρώμα μεταξύ των γραφημάτων (b) και (c).



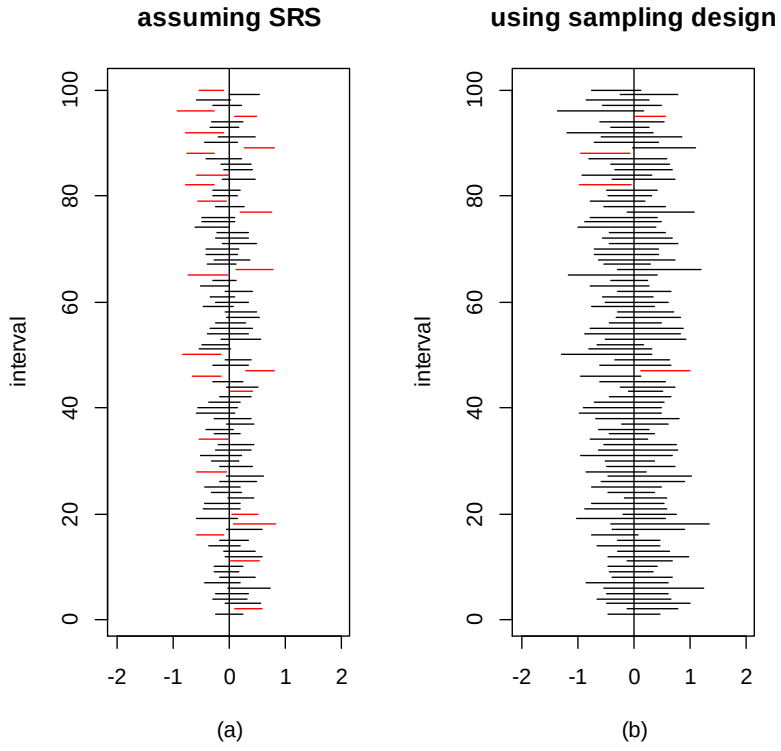

***Σχήμα 6.2*** *Γραφική παράσταση (a) ενός δείγματος από πληθυσμό με* $\rho_w = 0$, *(b) ΔΕ υπολογισμένα με α.τ.δ. και (c) ΔΕ υπολογισμένα σύμφωνα με την κατά ομάδες.*

(Β) όταν $\rho_w = 0.25$, το γράφημα των ΔΕ εμπιστοσύνης κάτω από την α.τ.δ. και τη δειγματοληψία κατά ομάδες δίνεται στο Σχήμα 6.3 (a) και (b) αντίστοιχα. Διαπιστώνουμε ότι τα διαστήματα εμπιστοσύνης σύμφωνα με την α.τ.δ. είναι μικρότερου μήκους σε σχέση με εκείνα βάσει της συμπερασματολογίας cluster και τα ΔΕ που είναι εσφαλμένα (με κόκκινο χρώμα) είναι περισσότερα.

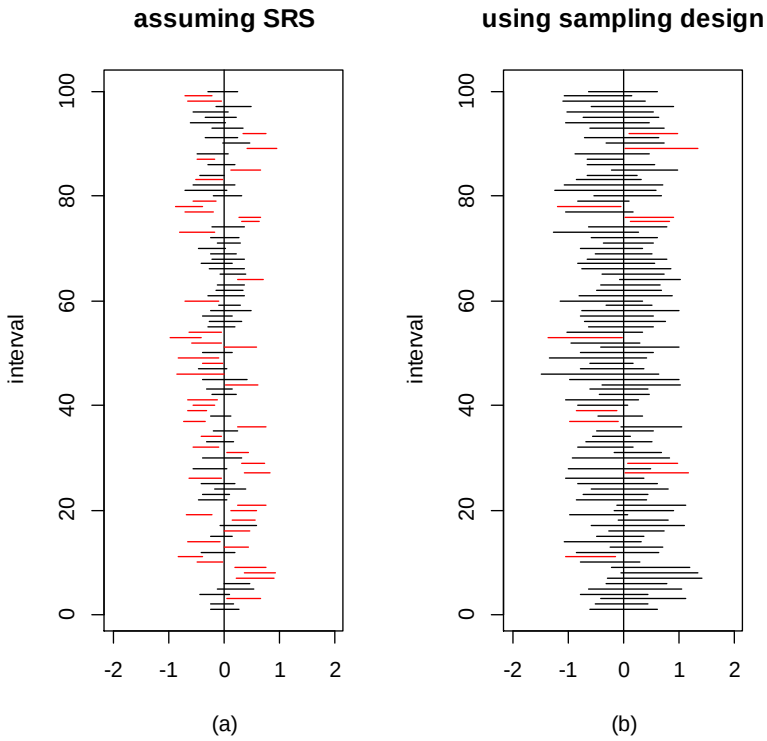



***Σχήμα 6.3*** *Γραφική παράσταση ΔΕ από δείγμα σε πληθυσμό με* $\rho_w = 0.25$, *(a) υπολογισμένα με α.τ.δ. και (b) υπολογισμένα σύμφωνα με την κατά ομάδες μέθοδο.*

Για τις περιπτώσεις πληθυσμού (Γ) και (Δ), οι γραφικές παραστάσεις των ΔΕ δίνονται στα Σχήματα 6.4 και 6.5 αντίστοιχα. Το φαινόμενο που παρατηρήθηκε για $\rho_w = 0.25$ εξακολουθεί να ισχύει και γίνεται πιο έντονο όσο το $\rho_w$ μεγαλώνει.

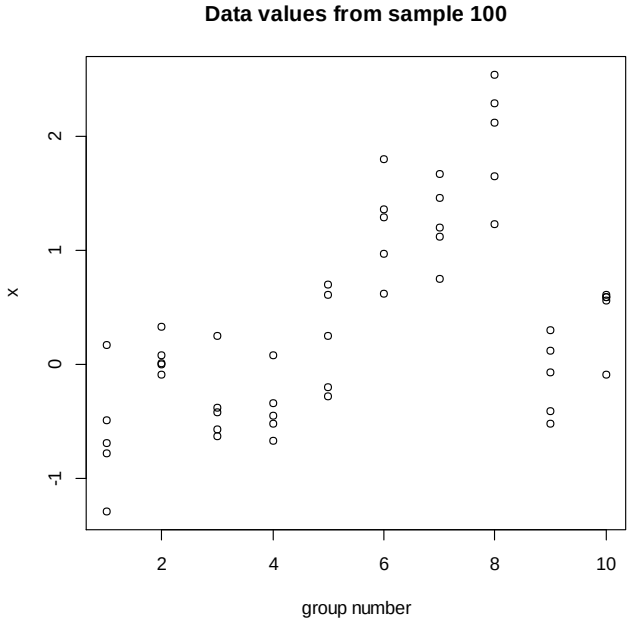



***Σχήμα 6.4*** *Γραφική παράσταση ΔΕ από δείγμα σε πληθυσμό με* $\rho_w = 0.5$, *(a) υπολογισμένα με α.τ.δ. και (b) υπολογισμένα σύμφωνα με την κατά ομάδες μέθοδο.*




***Σχήμα 6.5*** *Γραφική παράσταση ΔΕ από δείγμα σε πληθυσμό με* $\rho_w = 0.8$*, (a) υπολογισμένα με α.τ.δ. και (b) υπολογισμένα σύμφωνα με την κατά ομάδες μέθοδο.*

Η αντίστοιχη με το Σχήμα 6.2(a) γραφική παράσταση των στοιχείων ενός από τα δείγματα που επιλέχθηκαν από τον πληθυσμό για την περίπτωση $\rho_w = 0.8$ δίνεται στο Σχήμα 6.6. Διαπιστώνουμε ότι το εύρος των τιμών της κάθε ομάδας είναι πολύ μικρό, με αποτέλεσμα η κάθε ομάδα να προσφέρει εκπροσώπους μόνο από ένα τμήμα του πληθυσμού, και όχι από όλο συνολικά το εύρος του. Αντίθετα, οι ομάδες στο Σχήμα 6.2(a), όπου $\rho_w = 0$, εμφανίζουν μεγαλύτερο εύρος, που είναι συγκρίσιμο με εκείνο ολόκληρου του πληθυσμού.




***Σχήμα 6.6*** *Γραφική παράσταση ενός από τα δείγματα που επιλέχθηκαν από τον πληθυσμό με* $\rho_w = 0.8$*.*

### 6.3.3 Ομάδες άνισου μεγέθους

Η περίπτωση δειγματοληπτικού πλαισίου με ίσο μέγεθος ομάδων εμφανίζεται στην πράξη, αλλά με αρκετά μεγαλύτερη συχνότητα εμφανίζεται η περίπτωση όπου τα μεγέθη των ομάδων είναι άνισα. Π.χ. αν θεωρήσουμε ως ομάδες τα σχολεία, και ο πληθυσμός εκτείνεται σε μια έκταση με διαφορετική κατά τόπους πυκνότητα, ορισμένα σχολεία θα έχουν πάρα πολλούς μαθητές και ορισμένα ελάχιστους. Στην παράγραφο αυτή, γενικεύεται η υπόθεση για το μέγεθος των ομάδων και εξετάζεται η περίπτωση της εκτίμησης του πληθυσμιακού μέσου του πληθυσμού όταν το δείγμα και πάλι συλλέγεται με τη μέθοδο δειγματοληψίας κατά ομάδες σε ένα στάδιο με ίσες πιθανότητες.

Σύμφωνα με τον γενικότερο συμβολισμό, θα είναι $K_i$ το μέγεθος της ομάδας $i$ και $K_1 + K_2 + \cdots + K_M = N$. Επίσης, η δειγματοληψία παραμένει σε ένα στάδιο και, κατά συνέπεια, $k_i = K_i$ για κάθε ομάδα $i = 1,2, \dots, m$ που επιλέγεται στο δείγμα.

Κάτω από τις παραπάνω υποθέσεις, στη βιβλιογραφία έχουν επικρατήσει δύο εκτιμητές για την πληθυσμιακή μέση τιμή (α) ο αμερόληπτος εκτιμητής και (β) ο εκτιμητής λόγου ή πηλίκου, τα ονόματα των οποίων συνδέονται με τις ιδιότητές τους.

**(α) Αμερόληπτος Εκτιμητής**

Ο αμερόληπτος εκτιμητής της πληθυσμιακής μέσης τιμής $\bar{Y}$ δίνεται από τη σχέση:

$$\bar{X}_{cl,u} = \frac{1}{N}\frac{M}{m}\sum_{i=1}^{m} t_i \tag{6.8}$$

ή ισοδύναμα:

$$\bar{X}_{cl,u} = \frac{1}{N}\frac{M}{m}\sum_{i=1}^{m} K_i \bar{U}_i$$

όπου το $\frac{1}{m}\sum_{i=1}^{m} t_i$ εκτιμά το μέσο σύνολο των παρατηρήσεων ανά ομάδα. Αν, στη συνέχεια, κάνουμε χρήση της α.τ.δ. για την επιλογή των ομάδων, ακολουθώντας ανάλογη θεώρηση όπως και στην εξαγωγή του τύπου (6.2), προκύπτει:

**Πρόταση 6.4**

Ο εκτιμητής $\bar{X}_{cl,u}$ είναι αμερόληπτος για το $\bar{Y}$ και η διακύμανσή του:

$$\mathrm{Var}\left(\bar{X}_{cl,u}\right) = \frac{M^2}{N^2}\frac{1}{m}(1-f)\sum_{i=1}^{M}\frac{(K_i\bar{U}_i - \bar{t})^2}{M-1} = \frac{M^2}{N^2}\frac{1}{m}(1-f)\sum_{i=1}^{M}\frac{(t_i - \bar{t})^2}{M-1}$$

όπου $\bar{t}$ το μέσο $t_i$ για τις ομάδες, δηλ. $\bar{t} = \frac{1}{M}\sum_{i=1}^{M} t_i$.

Και με ανάλογο σκεπτικό, η εκτιμώμενη διακύμανση θα είναι:

$$\mathrm{Var}\left(\bar{X}_{cl,u}\right) = \frac{M^2}{N^2}\frac{1}{m}(1-f)\sum_{i=1}^{m}\frac{\left(K_i\bar{U}_i - \hat{\bar{t}}\right)^2}{m-1} = \frac{M^2}{N^2}\frac{1}{m}(1-f)\sum_{i=1}^{m}\frac{\left(t_i - \hat{\bar{t}}\right)^2}{m-1} \tag{6.9}$$

με $\hat{\bar{t}} = \frac{1}{m}\sum_{i=1}^{m} t_i$ το εκτιμώμενο βάσει του δείγματος μέσο άθροισμα των στοιχείων ανά ομάδα $\bar{t}$.

**(β) Εκτιμητής Λόγου (Ratio estimator)**

Ο εκτιμητής λόγου δίνεται ως ο λόγος του αθροίσματος των παρατηρήσεων για την $Y$ μεταβλητή στο δείγμα προς το άθροισμα των στοιχείων που έχουν συμπεριληφθεί συνολικά στο δείγμα. Αναλυτικότερα:

$$\bar{X}_{cl,r} = \hat{\bar{Y}} = \frac{\sum_{i=1}^{m} t_i}{\sum_{i=1}^{m} K_i} \tag{6.10}$$

Για τον εκτιμητή $\bar{X}_{cl,r}$ μπορεί να αποδειχθεί με τη βοήθεια της γενικότερης θεωρίας για τους εκτιμητές λόγου (θα δειχθεί στο Κεφάλαιο 7 -) η παρακάτω πρόταση.

### Πρόταση 6.5

Ο εκτιμητής $\bar{X}_{cl,r}$ δεν είναι αμερόληπτος και η διακύμανσή του δίνεται προσεγγιστικά από τον τύπο:

$$\mathrm{Var}\left(\bar{X}_{cl,r}\right) = \frac{1}{m\bar{K}^2}(1-f)\sum_{i=1}^{M}\frac{(K_i\bar{U}_i - K_i\bar{Y})^2}{M-1} \tag{6.11}$$

όπου $\bar{K} = \frac{1}{M}\sum_{i=1}^{M} K_i$ το μέσο μέγεθος των psu.

Μέσω της (6.11) για τη διακύμανση του εκτιμητή λόγου, συμπεραίνουμε ότι μια προσεγγιστική έκφραση για την εκτιμώμενη διακύμανση του $\bar{X}_{cl,r}$ δίνεται από τη σχέση:

$$\hat{\mathrm{Var}}\left(\bar{X}_{cl,r}\right) = \frac{1}{m\hat{\bar{K}}^2}(1-f)\sum_{i=1}^{m}\frac{\left(K_i\bar{U}_i - K_i\bar{X}_{cl,r}\right)^2}{m-1} \tag{6.12}$$

όπου $\hat{\bar{K}} = \frac{1}{m}\sum_{i=1}^{m} K_i$.

Τα ιδιαίτερα χαρακτηριστικά των εκτιμητών $\bar{X}_{cl,u}$ και $\bar{X}_{cl,r}$ συνοψίζονται στα παρακάτω σημεία.

(i) Από την έκφραση της διακύμανσης του $\bar{X}_{cl,u}$, είναι φανερό ότι η διακύμανση, και συνεπώς και το τυπικό σφάλμα του εκτιμητή, είναι μεγάλη, εάν η διακύμανση των αθροισμάτων $t_i$ $(i = 1,2,\dots,M)$ είναι μεγάλη μέσα στον πληθυσμό. Στην περίπτωση που εξετάζουμε, όπου τα μεγέθη $K_i$ είναι διαφορετικά ανά ομάδα, το ενδεχόμενο και τα αθροίσματά τους να διαφέρουν αρκετά είναι πολύ πιθανό. Π.χ. υποθέτουμε ως ομάδες τα ΑΕΙ μιας χώρας και έστω ότι το θέμα της έρευνας είναι η εκτίμηση της μέσης ηλικίας των εν ενεργεία μελών ΔΕΠ. Το άθροισμα των ηλικιών των μελών ΔΕΠ ανά ΑΕΙ, $t_i$, θα διαφέρει μεταξύ των διαφόρων ΑΕΙ εάν και το πλήθος των μελών ΔΕΠ διαφέρει για τα ΑΕΙ της χώρας, γιατί το πλήθος των μελών ΔΕΠ ταυτίζεται με το πλήθος των προσθετέων για την τιμή $t_i$. Άρα, μεγάλα ή παλιότερα ΑΕΙ με μεγάλο αριθμό μελών ΔΕΠ θα δώσουν μία μεγάλη τιμή για το άθροισμα $t_i$, ενώ μικρότερα ΑΕΙ με λιγότερα μέλη ΔΕΠ θα έχουν μία μικρότερη τιμή για το άθροισμα $t_i$.

(ii) Για τον υπολογισμό του $\bar{X}_{cl,u}$ χρειάζεται να γνωρίζουμε τον αριθμό $N$ ή, διαφορετικά, το άθροισμα $\sum_{i=1}^{M} K_i$. Σε αρκετές ρεαλιστικές περιπτώσεις δειγματοληψίας, δεν γνωρίζουμε το μέγεθος $K_i$ των $M$ ομάδων πριν από τη δειγματοληψία και, κατά συνέπεια, θα γίνουν γνωστά στο πέρας της έρευνας μόνο εκείνα τα $K_i$ που αντιστοιχούν σε ομάδες που επιλέχθηκαν στο δείγμα.

(iii) Η μεροληψία του εκτιμητή $\bar{X}_{cl,r}$ είναι της τάξης $1/m$. Για $m$ μεγάλο, ο εκτιμητής έχει αμελητέο ποσό μεροληψίας.

(iv) Ο εκτιμητής $\bar{X}_{cl,r}$ έχει μικρότερη διακύμανση, όταν το άθροισμα $t_i$ του χαρακτηριστικού που μελετάμε για την ομάδα $i$ συνδέεται με το μέγεθος της ομάδας $K_i$ (π.χ. ΑΕΙ και συνολική ηλικία των μελών ΔΕΠ).

(v) Ο εκτιμητής $\bar{X}_{cl,r}$ δεν απαιτεί γνώση του $N = \sum_{i=1}^{M} K_i$ για τον υπολογισμό του.

(vi) Οι δύο εκτιμητές δίνουν πολύ κοντινές τιμές εάν τα μεγέθη $K_i$ δεν διαφέρουν πολύ μεταξύ τους. Ειδικά για $K_i = K$, $i = 1,2,\dots,M$ οι δύο εκτιμητές ταυτίζονται.

### 6.3.4 Δειγματοληψία κατά ομάδες σε ένα στάδιο και στρωματοποιημένη

Όπως αναφέραμε στο 5ο Κεφάλαιο, η στρωματοποιημένη δειγματοληψία εφαρμόζεται αρκετά συχνά στην πράξη για βελτίωση των σφαλμάτων των εκτιμητών. Ως μέθοδος δειγματοληψίας συνδυάζεται επίσης πολύ συχνά με τη μέθοδο κατά ομάδες. Πιο συγκεκριμένα, κατά τον σχεδιασμό της έρευνας, οι ομάδες χωρίζονται σε στρώματα με κριτήριο γεωγραφικές περιοχές ή μέγεθος ομάδας κτλ. Στη συνέχεια, λαμβάνεται ένα α.τ.δ. ομάδων μέσα σε κάθε στρώμα. Η εκτίμηση της ποσότητας του πληθυσμού που μας ενδιαφέρει γίνεται αρχικά ανά στρώμα, κάνοντας χρήση της μεθόδου κατά ομάδες, ώστε να είναι σε συμφωνία με τον τρόπο που έχει συλλεγεί το δείγμα. Στη συνέχεια, συνθέτουμε τους εκτιμητές ανά στρώμα, με σκοπό τον υπολογισμό του συνολικού εκτιμητή για τον πληθυσμό, χρησιμοποιώντας τη στρωματοποιημένη δειγματοληψία και ειδικότερα τα βάρη των στρωμάτων.

Για την εκτίμηση και τις ιδιότητες των εκτιμητών, η επέκταση είναι άμεση και ισχύει για κάθε περίπτωση εκτιμητή που μελετήσαμε κάτω από το πλαίσιο της μεθόδου κατά ομάδες. Αν $\theta$ είναι η άγνωστη παράμετρος του πληθυσμού που ενδιαφερόμαστε να εκτιμήσουμε ($\theta = \bar{Y}$ ή $Y_T$ ή $P$ κτλ), ο εκτιμητής $\hat{\theta}_{st,cl}$ βάσει ενός στρωματοποιημένου δείγματος κατά ομάδες θα είναι ο:

$$\hat{\theta}_{st,cl} = \sum_h W_h \hat{\theta}_{h,cl}$$

όπου $\hat{\theta}_{h,cl}$ ο εκτιμητής του $\theta$ σύμφωνα με τη δειγματοληψία κατά ομάδες για το στρώμα $h$ και $W_h = N_h/N$ το πληθυσμιακό βάρος του ίδιου στρώματος. Επίσης, άμεσα προκύπτει ότι:

$$\mathrm{se}\left(\hat{\theta}_{st,cl}\right) = \sum_h W_h \mathrm{se}\left(\hat{\theta}_{h,cl}\right), \qquad \widehat{\mathrm{se}}\left(\hat{\theta}_{st,cl}\right) = \sum_h W_h \widehat{\mathrm{se}}\left(\hat{\theta}_{h,cl}\right).$$

### Παράδειγμα 6.3

Σε μια βιβλιοθήκη με 145 ράφια γεμάτα βιβλία, διεξάγεται μια δειγματοληψία, επιλέγοντας τυχαία 20 ράφια, με σκοπό την εκτίμηση του ποσοστού των βιβλίων που έχουν εκδοθεί μετά το 2000. Εξετάζονται όλα τα βιβλία των 20 ραφιών που επιλέχθηκαν και τα αριθμητικά αποτελέσματα δίνονται στον Πίνακα 6.5. Τα συνολικά ερωτήματα της έρευνας είναι:

(α) Να υπολογιστούν και να συγκριθούν ως προς την ακρίβεια ο αμερόληπτος και ο εκτιμητής λόγου για το ζητούμενο ποσοστό όταν:

(i) Η βιβλιοθήκη γνωρίζει ότι ο συνολικός αριθμός βιβλίων είναι 5200.

(ii) Δεν υπάρχει πληροφορία για τον συνολικό αριθμό των βιβλίων στη βιβλιοθήκη.

(β) Κάνοντας χρήση των δεδομένων της παρούσας έρευνας, να υπολογιστεί ο αριθμός ραφιών που χρειάζεται να επιλέξουμε σε μια μελλοντική επανάληψη της έρευνας, έτσι ώστε η εκτίμηση του ζητούμενου ποσοστού να μη διαφέρει περισσότερο από 1.5% από την αληθινή τιμή, με πιθανότητα σφάλματος 5%.

| **α/α Ραφιού** | **Αριθμός βιβλίων ($K_i$)** | **Βιβλία με έκδοση μετά το 2000 ($\alpha_i$)** | **α/α Ραφιού** | **Αριθμός βιβλίων($K_i$)** | **Βιβλία με έκδοση μετά το 2000 ($\alpha_i$)** |
|---|---|---|---|---|---|
| 1 | 39 | 8 | 11 | 44 | 9 |
| 2 | 35 | 6 | 12 | 28 | 4 |
| 3 | 40 | 8 | 13 | 41 | 10 |
| 4 | 49 | 12 | 14 | 22 | 3 |
| 5 | 36 | 6 | 15 | 46 | 8 |

| α/α Ραφιού | Αριθμός βιβλίων ($K_i$) | Βιβλία με έκδοση μετά το 2000 ($\alpha_i$) | α/α Ραφιού | Αριθμός βιβλίων($K_i$) | Βιβλία με έκδοση μετά το 2000 ($\alpha_i$) |
|---|---|---|---|---|---|
| 6 | 41 | 9 | 16 | 24 | 5 |
| 7 | 24 | 5 | 17 | 23 | 4 |
| 8 | 43 | 7 | 18 | 28 | 7 |
| 9 | 56 | 12 | 19 | 11 | 2 |
| 10 | 27 | 6 | 20 | 24 | 5 |

***Πίνακας 6.5*** *Δεδομένα για τον αριθμό και την ημερομηνία έκδοσης των βιβλίων στα 20 ράφια.*

(α) Η μέθοδος δειγματοληψίας που εφαρμόστηκε είναι κατά ομάδες σε ένα στάδιο με άνισο μέγεθος ομάδων. Οι εντολές για την εισαγωγή των δεδομένων στην R είναι

```
> K<-c(39, 35, 40, 49, 36, 41, 24, 43, 56, 27, 44, 28, 41, 22, 46,
24, 23, 28, 11, 24)
> a<-c(8, 6, 8, 12, 6, 9, 5, 7, 12, 6, 9, 4, 10, 3, 8, 5, 4, 7, 2, 5)
```

(i) Δίνεται ότι $N = 5200$. Για την εκτίμηση του ζητούμενου ποσοστού υπολογίζεται αρχικά ο αμερόληπτος εκτιμητής (6.8). Τα αθροίσματα $t_i$ για την περίπτωση του ποσοστού είναι τα αθροίσματα $\alpha_i$, τα οποία δίνουν τον συνολικό αριθμό στοιχείων του πληθυσμού ανά ομάδα με την ιδιότητα που μελετούμε. Οι εντολές για τις απαραίτητες πράξεις εφαρμογής της (6.8) είναι:

```
> est_unbiased<-145*sum(a)/(5200*20)
> est_unbiased
[1] 0.1896154
```

από όπου προκύπτει ότι η εκτίμηση του ποσοστού των βιβλίων που έχουν εκδοθεί μετά το 2000 είναι 0.1896 ή 18.96%.

Ο εκτιμητής λόγου για την ίδια ποσότητα, εφαρμόζοντας τον τύπο (6.10), θα είναι:

```
> est_ratio<-sum(a)/sum(K)
> est_ratio
[1] 0.1997063
```

οπότε το ζητούμενο ποσοστό με τη βοήθεια του εκτιμητή του λόγου εκτιμάται με 0.1997 ή 19.97% .

Οι εκτιμώμενες διακυμάνσεις των δύο εκτιμητών υπολογίζονται από τις (6.9) και (6.12) αντίστοιχα. Οι αριθμητικές τιμές για τα δεδομένα του παραδείγματος προκύπτουν, για μεν τον αμερόληπτο εκτιμητή:

```
> N<-5200
> M<-145
> m<-20
> f<-20/145
> var_est_u<-(M/N)^2*(1/m)*var(a)
> var_est_u
[1] 0.0002930142
> sqrt(var_est_u)
[1] 0.01711766
```

για δε τον εκτιμητή λόγου:

```
> meanK<-sum(K)/20
> meanK
```

```
[1] 34.05
>
> var_est_r<-(1/m)*(1/meanK)^2*(1-20/145)*sum((a-K*est_ratio)^2)/19
> var_est_r
[1] 4.553756e-05
> sqrt(var_est_r)
[1] 0.006748152
```

Το τυπικό σφάλμα του εκτιμητή του λόγου για το ποσοστό $\hat{P}_{cl,r} = 0.006748152$ είναι αρκετά μικρότερο από εκείνο του αμερόληπτου εκτιμητή $\hat{P}_{cl,u} = 0.01711766$.

(ii) Εάν το πλήθος συνολικά των βιβλίων για τη βιβλιοθήκη είναι άγνωστο, οι υπολογισμοί για τον εκτιμητή $\hat{P}_{cl,r}$ παραμένουν ανεπηρέαστοι, γιατί ο εκτιμητής δεν εξαρτάται από το $N$, ενώ ο εκτιμητής $\hat{P}_{cl,u}$ δεν είναι εφικτό να υπολογιστεί.

(β) Αξιοποιώντας τα δεδομένα της τωρινής έρευνας και λαμβάνοντας υπόψη τους περιορισμούς που πρέπει να πληροί ο εκτιμητής του ποσοστού, το μέγεθος του δείγματος των ομάδων $m$ προκύπτει από τη λύση της εξίσωσης (2.11):

$$\frac{d}{\mathrm{se}(\hat{\theta})} = z_{\alpha/2}$$

όπου $d = 0.015$, $z_{\alpha/2} = 1.96$ και

$$\mathrm{se}(\hat{\theta}) = \sqrt{\mathrm{Var}(\hat{P}_{cl,r})} = \sqrt{\frac{1}{m\bar{K}^2}(1-f)\sum_{i=1}^{M}\frac{(K_iP_i - K_iP)^2}{M-1}}$$

Εφαρμόζουμε τον εκτιμητή λόγου ως πιο αποτελεσματικό, όπως προέκυψε από το (α). Στην τελευταία έκφραση, το άθροισμα $A = \sum_{i=1}^{M}\frac{(K_iP_i - K_iP)^2}{M-1}$ αντικαθιστάται από τον εκτιμητή του βάσει της τρέχουσας έρευνας. Στη συνέχεια, η εξίσωση λύνεται ως προς $m$. Για τις πράξεις, οι κατάλληλες εντολές της R είναι

```
> A=sum((a-K*est_ratio)^2)/19
> A
[1] 1.224875
> d<-.015
> m<-(1/M+(d/1.96)^2*meanK^2/A)^(-1)
> m
[1] 16.0423
```

Άρα, για μια μελλοντική έρευνα θα χρειαστεί να επιλεγούν με α.τ. τρόπο 17 τουλάχιστον ράφια της βιβλιοθήκης, ακολουθώντας το ίδιο δειγματοληπτικό σχήμα, έτσι ώστε να πληρούνται οι προδιαγραφές που δόθηκαν.

## 6.4. Δειγματοληψία με άνισες πιθανότητες

Στη μέχρι τώρα ανάπτυξη της μεθόδου δειγματοληψίας κατά ομάδες, έχουμε υποθέσει ότι η επιλογή των συστάδων κατά το πρώτο στάδιο, όπως και η επιλογή των στοιχείων στο δεύτερο στάδιο, γίνεται με απλή τυχαία δειγματοληψία και κατά συνέπεια, σύμφωνα με τον ορισμό της α.τ.δ., με ίσες πιθανότητες. Η υιοθέτηση της α.τ.δ. για την επιλογή πρωτογενών μονάδων ή στοιχείων διευκολύνει τη διαδικασία επιλογής του δείγματος, αλλά ταυτόχρονα παρουσιάζει μειονεκτήματα, τα οποία κυρίως συνδέονται με την ακρίβεια των εκτιμητών. Για την κατανόηση της πηγής των προβλημάτων που δημιουργούνται με την υπόθεση των

ίσων πιθανοτήτων, χρησιμοποιούμε ένα παράδειγμα με μικρό μέγεθος ομάδων, μη ρεαλιστικό, μόνο για λόγους κατανόησης.

## Παράδειγμα 6.4

Έστω ότι σε μια περιοχή δραστηριοποιούνται τρία πολυκαταστήματα και ενδιαφερόμαστε, κάνοντας μια δειγματοληψία επιλέγοντας δύο εξ αυτών, να εκτιμήσουμε τη συνολική αξία των πωλήσεων που πραγματοποίησαν τα καταστήματα της περιοχής στη διάρκεια του περασμένου μήνα. Έστω ότι τα στοιχεία των πωλήσεων των καταστημάτων είναι γνωστά και δίνονται στον Πίνακα 6.6.

| **Πολυκατάστημα** | **Έκταση (επιφάνεια σε τ.μ.) του καταστήματος** | **Πωλήσεις σε χιλιάδες χρηματικές μονάδες** |
|---|---|---|
| 1 | 120 | 16 |
| 2 | 300 | 28 |
| 3 | 550 | 64 |
| **Σύνολα** | **970** | **108** |

***Πίνακας 6.6*** *Επιφάνεια και συνολικές εισπράξεις από πωλήσεις των καταστημάτων της περιοχής*

Τα δυνατά δείγματα μιας τυχαίας δειγματοληψίας μεγέθους 2 χωρίς επανατοποθέτηση, καθώς και οι εκτιμητές των συνολικών πωλήσεων των καταστημάτων της περιοχής βάσει του δείγματος, θα είναι:

| **Δείγμα** | **Δειγματικό Σύνολο** | **Εκτίμηση Συνόλου πληθυσμού** |
|---|---|---|
| 1, 2 | 44 | 66 |
| 1, 3 | 80 | 120 |
| 2, 3 | 92 | 138 |

Οι εκτιμήσεις του συνόλου του πληθυσμού που δίνονται στην τρίτη στήλη προκύπτουν από τη γνωστή έκφραση $\hat{Y}_T = N\hat{\bar{Y}}$ που δίνει τον εκτιμητή του συνόλου του πληθυσμού για την α.τ.δ. Για παράδειγμα, το δείγμα (1,2) έχει δειγματικό σύνολο 44, δηλ.δειγματικό μέσο 22, άρα ο εκτιμητής του συνόλου του πληθυσμού με βάση αυτό το δείγμα είναι $\hat{Y}_T = N\hat{\bar{Y}} = 3 * 22 = 66$.

Έχοντας τα αποτελέσματα των εκτιμητών για όλα τα δυνατά δείγματα υπολογίζουμε σύμφωνα με τον ορισμό την αναμενόμενη τιμή, τη διακύμανση και το τυπικό σφάλμα της εκτίμησης $\hat{Y}_T$. Συγκεκριμένα:

$$\mathrm{E}(\hat{Y}_T) = \frac{1}{3}(66 + 120 + 138) = 108,$$

$$\mathrm{Var}(\hat{Y}_T) = \frac{1}{3}[(66 - 108)^2 + (120 - 108)^2 + (138 - 108)^2] = 936 \text{ και}$$

$$\mathrm{se}(\hat{Y}_T) = 30.59 \text{ και } \mathrm{CV}(\hat{Y}_T) = 0.28.$$

Ο εκτιμητής $\hat{Y}_T$ είναι αμερόληπτος, λόγω της α.τ.δ., αλλά το τυπικό σφάλμα του εκτιμητή, $\mathrm{se}(\hat{Y}_T)$, καθώς και ο συντελεστής μεταβλητότητας, $\mathrm{CV}(\hat{Y}_T)$, έχουν πολύ μεγάλες τιμές για το δείγμα, το οποίο αποτελείται παρόλα αυτά από τα 2/3 του πληθυσμού.

Ο λόγος που παρατηρείται το μεγάλο τυπικό σφάλμα στην εκτίμηση είναι η μεγάλη διακύμανση μεταξύ των τιμών των πωλήσεων των τριών πολυκαταστημάτων. Το τρίτο πολυκατάστημα είναι αρκετά μεγαλύτερο σε

έκταση και σε πωλήσεις, με αποτέλεσμα το πρώτο δείγμα (1,2) που δεν το περιέχει να δίνει έναν εκτιμητή αρκετά μικρότερο από την αληθινή τιμή.

Γενικότερα, σε έρευνες όπου γνωρίζουμε ή υποπτευόμαστε ότι οι τιμές για τις μονάδες του πληθυσμού έχουν αρκετά μεγάλη διαφορά μεταξύ τους, η επιλογή του δείγματος με ίσες πιθανότητες, αν και εύκολη, δεν συνιστάται, γιατί οδηγεί σε εκτιμητές με μεγάλα σφάλματα. Ένα τυπικό παράδειγμα της περίπτωσης πληθυσμών που περιγράψαμε είναι όταν το δειγματοληπτικό πλαίσιο δίνεται σε μορφή ομάδων και τα μεγέθη των ομάδων είναι αρκετά διαφορετικά μεταξύ τους. Αναμένουμε σ΄αυτές τις περιπτώσεις μια συστάδα με μεγάλο μέγεθος να δώσει ένα σύνολο για το υπό μελέτη χαρακτηριστικό πολύ μεγαλύτερο από μια συστάδα με μικρό πλήθος στοιχείων. Για παράδειγμα, έστω ότι οι συστάδες είναι τα Πανεπιστήμια μιας χώρας και καταγράφουμε τον αριθμό μελών ΔΕΠ που απασχολούν. Ορισμένα Πανεπιστήμια θα είναι πιο παλιά και με περισσότερα Τμήματα, άρα και περισσότερο προσωπικό, ενώ άλλα Πανεπιστήμια, επαρχιακά ή με πρόσφατη χρονολογία δημιουργίας, θα έχουν λιγότερα μέλη ΔΕΠ. Ανάλογα, έστω ότι ομάδες είναι τα Νοσοκομεία μιας περιοχής και ενδιαφερόμαστε για ένα χαρακτηριστικό σχετικό με τους ασθενείς π.χ. τον μέσο χρόνο νοσηλείας τους ή το ποσοστό εκείνων που έχουν εισαχθεί στη μονάδα εντατικής θεραπείας (ΜΕΘ) κτλ. Είναι αναμενόμενο τα μεγάλα Νοσοκομεία με πολλές κλίνες ή αίθουσες ΜΕΘ να προσφέρουν μεγαλύτερο άθροισμα για το χαρακτηριστικό που ερευνούμε. Εάν, σε έρευνες παρόμοιες με αυτές που περιγράψαμε, εφαρμόσουμε τους εκτιμητές $\bar{X}_{cl,u}$ και $\bar{X}_{cl,r}$ που μελετήσαμε στην παράγραφο 6.3.3, θα έχουμε αρκετά μεγάλα σφάλματα στις εκτιμήσεις.

Όπως είναι φανερό από τα παραδείγματα, η μεγάλη διακύμανση μεταξύ των συνόλων των ομάδων $t_i$ στην κατά ομάδες σε ένα στάδιο δειγματοληψία παρατηρείται όταν το χαρακτηριστικό που μελετούμε συνδέεται θετικά με το μέγεθος της ομάδας $K_i$ και τα μεγέθη αυτά διαφέρουν από ομάδα σε ομάδα. Το γεγονός όμως αυτό συμβαίνει στην πλειονότητα των περιπτώσεων, γι΄αυτό και συνιστάται ένα εναλλακτικό δειγματοληπτικό σχέδιο από εκείνο που έχει παρουσιαστεί στην παράγραφο 6.3, δηλ. επιλογή ομάδων με ίσες πιθανότητες. Στο σχέδιο αυτό, οι πιθανότητες επιλογής των ομάδων είναι άνισες μεταξύ τους και είναι ανάλογες με το μέγεθος μιας συστάδας. Όσο μεγαλύτερη, είναι μια συστάδα τόσο μεγαλύτερη θα είναι η πιθανότητα επιλογής της στο δείγμα. Η επιλογή των ομάδων στο δείγμα με τον τρόπο αυτό ονομάζεται, **δειγματοληψία με πιθανότητες ανάλογες του μεγέθους (probabilities proportional to size)** και έχει διεθνώς επικρατήσει να συμβολίζεται **pps.**

Για το απλό παράδειγμα με τα πολυκαταστήματα, αν αντί για ίσες πιθανότητες επιλογής κάνουμε χρήση της μεταβλητής που δίνει την επιφάνεια του κάθε πολυκαταστήματος και προσαρμόσουμε ανάλογα τις πιθανότητες (π.χ. για το πρώτο κατάστημα η πιθανότητα επιλογής θα είναι 120/970 δηλ. όσο το ποσοστό που αναλογεί σε έκταση για το κατάστημα αυτό σε σχέση με το σύνολο), τότε οι νέες πιθανότητες θα είναι:

| **Πολυκατάστημα** | **Πιθανότητα επιλογής $p_i$** |
|---|---|
| 1 | 0.12 |
| 2 | 0.31 |
| 3 | 0.57 |

Σύμφωνα με τις πιθανότητες αυτές η πιθανότητα 0.12 του πολυκαταστήματος 1 είναι αρκετά μικρότερη από την πιθανότητα 0.33 που ήταν η πιθανότητα επιλογής του σύμφωνα με τις ίσες πιθανότητες. Το ίδιο ισχύει, αλλά προς την αντίθετη κατεύθυνση, για το πολυκατάστημα 3

Η μεθοδολογία που έχει αναπτυχθεί για τον υπολογισμό του εκτιμητή και των ιδιοτήτων του όταν το δείγμα επιλέγεται με άνισες πιθανότητες είναι γενικότερη της μεθόδου δειγματοληψίας κατά ομάδες. Όπως και στο παράδειγμα, αρκεί να είναι διαθέσιμη μια βοηθητική μεταβλητή η οποία να μας δίνει πληροφορία για το μέγεθος της μονάδας του πληθυσμού, και με τη βοήθειά της να προσδιορίσουμε τις (άνισες) πιθανότητες επιλογής για κάθε μέλος του πληθυσμού. Ωστόσο, όπως αναφέρθηκε, η δειγματοληψία κατά ομάδες είναι ένα δειγματοληπτικό σχέδιο για το οποίο η μεθοδολογία αυτή βρίσκει μεγάλη εφαρμογή.

Η κεντρική ιδέα της pps μεθοδολογίας είναι να γίνει χρήση της πληροφορίας που διαθέτουμε για το μέγεθος της κάθε μονάδας, και η πληροφορία αυτή να αξιοποιηθεί, τόσο κατά την επιλογή του δείγματος, όσο και κατά την εξαγωγή των εκτιμητών, με τελικό στόχο τους αποτελεσματικότερους εκτιμητές.

Θα δούμε την εφαρμογή ενός από τους τύπους εκτιμητών που έχουν προταθεί από τη βιβλιογραφία για επιλογή με άνισες πιθανότητες στο μικρό παράδειγμα με τα πολυκαταστήματα. Υπολογίζουμε αρχικά τις πιθανότητες που έχει να επιλεγεί το κάθε ζευγάρι καταστημάτων (1,2), (1,3) και (2,3). Υπενθυμίζουμε ότι η επιλογή των πολυκαταστημάτων γίνεται χωρίς επανατοποθέτηση. Κάνοντας χρήση κανόνων συνδυαστικής, εύκολα επαληθεύεται ότι οι πιθανότητες αυτές είναι $0.09619 (= p_1p_2/(p_2+p_3) + p_2p_1/(p_1+p_3))$, $0.23679\ (= p_1p_3/(p_2+p_3) + p_3p_1/(p_1+p_2))$ και $0.66701\ (= p_2p_3/(p_1+p_3) + p_3p_2/(p_1+p_2))$, αντίστοιχα.

| **Δείγμα (i,j)** | **Πιθανότητα επιλογής δείγματος $\pi_{ij}$** |
|---|---|
| (1,2) | 0.09619 |
| (1,3) | 0.23679 |
| (2,3) | 0.66701 |

Για να βρούμε την πιθανότητα το 1ο πολυκατάστημα να ανήκει σε ένα δείγμα μεγέθους 2 αρκεί να αθροίσουμε στον παραπάνω πίνακα τις πιθανότητες των δειγμάτων που περιέχουν το πολυκατάστημα 1. Έτσι, η πιθανότητα αυτή θα ισούται με 0.09619+0.23679=0.33298, ενώ για το 2ο πολυκατάστημα θα είναι 0.09619+0.66701=0.76320 και για το 3ο θα είναι 0.23679+0.66701=0.90380. Οι πιθανότητες αυτές εμφανίζονται στον επόμενο πίνακα.

| **Πολυκατάστημα $i$** | **Πιθανότητα συμπερίληψης στο δείγμα $\pi_i$** |
|---|---|
| 1 | 0.33298 |
| 2 | 0.76320 |
| 3 | 0.90380 |

Υιοθετούμε τον εκτιμητή:

$$\tilde{Y} = \sum_{i=1}^{n} \frac{t_i}{\pi_i}$$

όπου $t_i$ είναι οι συνολικές πωλήσεις του καταστήματος $i$. Η βασική διαφορά του εκτιμητή $\tilde{Y}$ σε σχέση με τους $\bar{X}_{cl,u}$ και $\bar{X}_{cl,r}$ που μελετήσαμε είναι ότι η κάθε τιμή του δείγματος $t_i$ έχει συντελεστή, ή βάρος, $1/\pi_i$, το οποίο εξαρτάται από τη μονάδα $i$ και είναι αντιστρόφως ανάλογο της πιθανότητας που είχε η συγκεκριμένη μονάδα να ανήκει στο δείγμα. Μεγάλη πιθανότητα, όπως π.χ. για το 3ο κατάστημα, σημαίνει μικρό βάρος για τη μέτρηση αυτή, εφόσον συμπεριληφθεί στο δείγμα.

Οι εκτιμήσεις των συνολικών πωλήσεων των καταστημάτων της περιοχής μπορούν να υπολογιστούν εκ νέου με τη βοήθεια του $\tilde{Y}$ και για $n = 2$ είναι:

| Δείγμα | Εκτίμηση Συνόλου πληθυσμού |
|---|---|
| 1,2 | 84.7386 |
| 1,3 | 118.8631 |
| 2,3 | 107.4998 |

Κάνοντας χρήση του ορισμού και των πιθανοτήτων $\pi_{ij}$ του κάθε δείγματος, η αναμενόμενη τιμή του νέου εκτιμητή είναι:

$$\mathrm{E}(\tilde{Y}) = 0.09619 * 84.7387 + 0.23679 * \ 118.8631 + 0.66701 * 107.4998 = 108$$

είναι δηλαδή επίσης αμερόληπτος, όπως κι εκείνος που υιοθετήσαμε για ίσες πιθανότητες. Η διακύμανση όμως του εκτιμητή $\tilde{Y}$ είναι

$$\mathrm{Var}(\tilde{Y}) = 0.09619 * (84.7387 - 108)^2$$

$$+0.23679 * (118.8631 - 108)^2$$

$$+ 0.66701 * (107.4998 - 108)^2 = \ 80.15701$$

(ενώ η αντίστοιχη διακύμανση ήταν 936 για τον εκτιμητή με ίσες πιθανότητες) και το τυπικό σφάλμα του $\tilde{Y}$, είναι $\mathrm{se}(\tilde{Y}) = 8.95$ αντί για 30.59. Προφανώς, το τυπικό σφάλμα του $\tilde{Y}$ είναι αρκετά μειωμένο, χωρίς να χάνει την ιδιότητα της αμεροληψίας.

Ο εκτιμητής $\tilde{Y}$ που χρησιμοποιήσαμε για το παράδειγμα ανήκει σε μια μεγαλύτερη κλάση εκτιμητών με όνομα Horvitz-Thompson (HT). Ο εκτιμητής HT είναι αμερόληπτος και η διακύμανση του εξαρτάται από τις πιθανότητες $\pi_i$ και $\pi_{ij}$, οι οποίες ονομάζονται **πιθανότητες συμπερίληψης στο δείγμα πρώτης και δεύτερης τάξης** αντίστοιχα (**first and second order inclusion probabilities**). Οι εκφράσεις για τη διακύμανση και το τυπικό σφάλμα του εκτιμητή είναι περισσότερο πολύπλοκες όταν η επιλογή του δείγματος γίνεται χωρίς επανατοποθέτηση (όπως στο παράδειγμα), γιατί γίνεται πιο δύσκολος ο υπολογισμός των $\pi_{ij}$. Ένα παράδειγμα εκτιμητή HT που έχουμε ήδη χρησιμοποιήσει είναι ο δειγματικός μέσος για την εκτίμηση της πληθυσμιακής μέσης τιμής κατά την α.τ.δ. Εκεί, αρκεί να θέσουμε $\pi_i = n/N$ στον ορισμό του HT και λαμβάνεται η γνωστή έκφραση για τον εκτιμητή που υιοθετήσαμε στο Κεφάλαιο 2 -. Για λεπτομερέστερη μελέτη του εκτιμητή Horvitz-Thompson βλ. Cochran (1977), Κεφ. 9A, Levy & Lemeshow (1999), Κεφ. 11 και Thompson (2012), Κεφ. 12.

Στη συνέχεια του κεφαλαίου, θα παρουσιάσουμε δύο από τους εκτιμητές που έχουν προταθεί στη βιβλιογραφία για δειγματοληψία με άνισες πιθανότητες και ειδικότερα pps, στο πλαίσιο της δειγματοληψίας κατά ομάδες σε ένα στάδιο. Οι δύο εκτιμητές που θα παρουσιαστούν είναι οι Horvitz-Thompson, ο οποίος αρχικά παρουσιάστηκε το 1952, και ο εκτιμητής Hansen-Hurwitz (1943).

## 6.5. Δειγματοληψία κατά ομάδες σε ένα στάδιο με άνισες πιθανότητες

### 6.5.1 Εκτιμητής Horvitz-Thompson

Ακολουθώντας τον ίδιο συμβολισμό που υιοθετήσαμε στην αρχή του κεφαλαίου (Πίνακας 6.3) υποθέτουμε ότι ο πληθυσμός αποτελείται από $M$ συστάδες άνισου μεγέθους $K_i$, $(i = 1,2, \dots, M)$ με $K_1 + K_2 + \cdots + K_M = N$. Υποθέτουμε ακόμα ότι η δειγματοληψία γίνεται στο επίπεδο των συστάδων και επιλέγονται $m$ από τις $M$ συστάδες. Για τον εκτιμητή Horvitz-Thompson, μπορούμε να υποθέσουμε δειγματοληψία με επανα-

τοποθέτηση ή χωρίς, αλλά για την παρουσίαση του προβλήματος και χάριν απλότητας στις εκφράσεις που θα προκύψουν, έστω ότι η δειγματοληψία είναι με επανατοποθέτηση. Αυτό σημαίνει ότι ένα psu μπορεί να επιλεγεί περισσότερες από μία φορές και το περιεχόμενό του ως προς τις ssu να προσμετρηθεί τον ίδιο αριθμό φορών στο δείγμα.

Ο εκτιμητής Horvitz-Thompson (HT) του πληθυσμιακού συνόλου στην περίπτωση της δειγματοληψίας κατά ομάδες σε ένα στάδιο παίρνει τη μορφή:

$$\hat{Y}_{HT} = \sum_{i=1}^{u} \frac{t_i}{\pi_i} \tag{6.13}$$

όπου $u$ είναι το πλήθος των ομάδων που είναι διακριτές από τις $m$ που επιλέγονται, $\pi_i$ η πιθανότητα να επιλεγεί η $i$ ομάδα στο δείγμα και $t_i$ το άθροισμα όλου του περιεχομένου της $i$ ομάδας (απογραφή εντός της ομάδας) ως προς το χαρακτηριστικό $Y$.

Από τη σχέση (6.13) προκύπτει ότι ο εκτιμητής Horvitz-Thompson του πληθυσμιακού μέσου δίνεται στη συνέχεια ως:

$$\hat{\bar{Y}}_{HT} = \frac{1}{N} \sum_{i=1}^{u} \frac{t_i}{\pi_i}$$

Ο εκτιμητής HT έχει την ιδιότητα της αμεροληψίας και αποδεικνύεται (για την απόδειξη, βλ. Cochran (1977), Κεφ. 9Α, σελ. 260) ότι το τυπικό του σφάλμα είναι:

$$\mathrm{se}\left(\hat{Y}_{HT}\right) = \sqrt{\sum_{i=1}^{M} \frac{1-\pi_i}{\pi_i} t_i^2 + \sum_{i=1}^{M} \sum_{j \neq i} \left( \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) t_i \, t_j}$$

Για την εκτίμηση του τυπικού σφάλματος του εκτιμητή HT, έχουν προταθεί δύο εκτιμητές, ο πρώτος από τους Horvitz-Thompson (1952):

$$\widehat{\mathrm{se}}_1\left(\hat{Y}_{HT}\right) = \sqrt{\sum_{i=1}^{u} \frac{1-\pi_i}{\pi_i^2} t_i^2 + \sum_{i=1}^{u} \sum_{j \neq i} \left( \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}} \right) t_i \, t_j} \tag{6.14}$$

και ο δεύτερος από τους Yates & Grundy (1953) και Sen (1953)

$$\widehat{\mathrm{se}}_2\left(\hat{Y}_{HT}\right) = \sqrt{\sum_{i=1}^{u} \sum_{j>i} \left( \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \left( \frac{t_i}{\pi_i} - \frac{t_j}{\pi_j} \right)^2} \tag{6.15}$$

Ο υπολογισμός της εκτίμησης του τυπικού σφάλματος του εκτιμητή HT από τα στατιστικά πακέτα γίνεται συνήθως από τον δεύτερο εκτιμητή $\widehat{\mathrm{se}}_2\left(\hat{Y}_{HT}\right)$, γιατί παρουσιάζει λιγότερα προβλήματα από τον $\widehat{\mathrm{se}}_1\left(\hat{Y}_{HT}\right)$. Τα προβλήματα που παρουσιάζουν και οι δύο εκτιμητές οφείλονται στο γεγονός ότι για ορισμένα δειγματοληπτικά σχέδια, οι όροι $\pi_i \pi_j - \pi_{ij}$ μπορεί να παρουσιάζουν αρκετά μεγάλη μεταβλητότητα και να γίνονται και αρνητικοί, με αποτέλεσμα ο αριθμητικός υπολογισμός της εκτίμησης του τυπικού σφάλματος για ορισμένα δείγματα να δώσει αρνητική τιμή.

Τα πλεονεκτήματα του HT είναι ότι ανήκει στην κλάση των γραμμικών εκτιμητών, δηλαδή έχει μία απλή έκφραση, είναι αμερόληπτος και επιπλέον ορίζεται αρκετά γενικά, με τη βοήθεια μόνο των πιθανοτήτων συμπερίληψης στο δείγμα πρώτης τάξης $\pi_i$. Επιπλέον, καλύπτει τις περιπτώσεις δειγματοληψίας με επανατοποθέτηση και χωρίς επανατοποθέτηση. Κατά συνέπεια, μπορεί να εφαρμοστεί σε οποιοδήποτε δειγματοληπτικό σχέδιο, όσο πολύπλοκο κι αν είναι, αρκεί να είμαστε σε θέση να υπολογίσουμε τις

πιθανότητες $\pi_i$. Στην περίπτωση της δειγματοληψίας κατά ομάδες, οι πιθανότητες $\pi_i$ υπολογίζονται βάσει των μεγεθών των ομάδων $K_i$, ενώ για τη γενικότερη περίπτωση μιας δειγματοληψίας με άνισες πιθανότητες, τα $\pi_i$ υπολογίζονται με τη βοήθεια μιας βοηθητικής μεταβλητής.

Στα βασικά επίσης πλεονεκτήματα του HT συμπεριλαμβάνεται η ιδιότητα της αμεροληψίας και ότι υπάρχει ο γενικός τύπος υπολογισμού των εκτιμώμενων τυπικών σφαλμάτων, ο οποίος εξαρτάται επιπλέον από τις πιθανότητες συμπερίληψης δεύτερης τάξης $\pi_{ij}$. Για κάθε ζεύγος παρατηρήσεων του πληθυσμού $(Y_i, Y_j)$, η πιθανότητα $\pi_{ij}$ ορίζεται ως η πιθανότητα οι μονάδες $Y_i$, $Y_j$ να ανήκουν ταυτόχρονα σε ένα δείγμα (από κοινού συμπερίληψη στο δείγμα) σύμφωνα με το δειγματοληπτικό σχέδιο που εφαρμόζεται για την έρευνα. Στην πράξη, και ειδικότερα (i) για σύνθετα δειγματοληπτικά σχέδια και (ii) για επιλογή δείγματος χωρίς επανατοποθέτηση, ο υπολογισμός των πιθανοτήτων $\pi_{ij}$ είναι δύσκολος. Αυτός είναι ένας επιπλέον λόγος που ο υπολογισμός των σφαλμάτων του εκτιμητή HT μπορεί να γίνει προβληματικός. Παρόλα αυτά, υπάρχουν διαθέσιμες προσεγγίσεις του υπολογισμού των $\pi_{ij}$ στη βιβλιογραφία, με κυριότερες εκείνες των Brewer (2002), Overton (1990) και Stehman & Overton (1987). Οι προσεγγίσεις αυτές είναι μεταξύ των πιο γνωστών προσεγγίσεων που υλοποιούνται από τα στατιστικά πακέτα για ανάλυση δεδομένων από σύνθετες δειγματοληπτικές έρευνες.

Ένας εκτιμητής που ανήκει επίσης στην κλάση των γραμμικών εκτιμητών, αλλά ο υπολογισμός του τυπικού του σφάλματος στην πράξη είναι συνήθως πιο εύκολος, είναι ο εκτιμητής Hansen-Hurwitz (1943), ο οποίος όμως εφαρμόζεται μόνο για τη δειγματοληψία με επανατοποθέτηση.

### 6.5.2 Εκτιμητής Hansen-Hurwitz

Σύμφωνα με τον ορισμό του εκτιμητή Hansen-Hurwitz (HH), στην περίπτωση της δειγματοληψίας σε ένα στάδιο με επανατοποθέτηση, ο εκτιμητής του πληθυσμιακού συνόλου δίνεται από τη σχέση:

$$\hat{Y}_{HH} = \frac{1}{m}\sum_{i=1}^{m}\frac{t_i}{\pi_i'} \tag{6.16}$$

όπου $\pi_i'$ είναι η πιθανότητα της $i$ ομάδας να επιλεγεί σε κάθεμιά από τις $m$ ανεξάρτητες επιλογές μιας μονάδας του πληθυσμού (δειγματοληψία με επανατοποθέτηση).

Αν θέσουμε $u_i = t_i/\pi_i'$, ο εκτιμητής (6.16) γράφεται ισοδύναμα:

$$\hat{Y}_{HH} = \frac{1}{m}\sum_{i=1}^{m} u_i = \bar{u} \tag{6.17}$$

είναι δηλαδή ο δειγματικός μέσος όρος των τιμών $u_i$. Από την τελευταία αυτή έκφραση, προκύπτει εύκολα η σχέση που δίνει τη διακύμανση, και κατά συνέπεια το τυπικό σφάλμα, του εκτιμητή HH, καθώς και οι αντίστοιχοι εκτιμητές τους. Αναλυτικά:

$$\mathrm{se}\left(\hat{Y}_{HH}\right) = \sqrt{\frac{1}{m}\sum_{i=1}^{M}(u_i - Y_T)^2\pi_i'} \tag{6.18}$$

και ένας αμερόληπτος εκτιμητής του σφάλματος $\mathrm{se}\left(\hat{Y}_{HH}\right)$, είναι ο:

$$\hat{\mathrm{se}}\left(\hat{Y}_{HH}\right) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\frac{(u_i - \bar{u})^2}{m-1}} \tag{6.19}$$

Εάν η δειγματοληψία είναι pps, δηλαδή $\pi_i^{'} = \frac{K_i}{\sum_{i=1}^{M} K_i}$ και $\sum_{i=1}^{M} K_i = N$, τότε ο εκτιμητής (6.16) γίνεται:

$$\hat{Y}_{HH} = \frac{N}{m}\sum_{i=1}^{m}\frac{t_i}{K_i} = N\frac{1}{m}\sum_{i=1}^{m}\bar{U}_i$$

και ο εκτιμητής HH του πληθυσμιακού μέσου παίρνει αντίστοιχα τη μορφή:

$$\hat{\bar{Y}}_{HH} = \frac{1}{m}\sum_{i=1}^{m}\bar{U}_i$$

## Παρατήρηση 6.3

Στον ορισμό του εκτιμητή HH (6.16) και σε όσες εκφράσεις εκτιμητών προκύπτουν από αυτόν ισοδύναμα, οι όροι του αθροίσματος δεν είναι απαραίτητα διακριτές τιμές, διότι λόγω της επανατοποθέτησης ενδέχεται να υπάρξει πολλαπλότητα στον αριθμό επιλογής μιας ομάδας στο δείγμα. Ο εκτιμητής HH λαμβάνει υπόψη την κάθε τιμή ομάδας αλλά και την πολλαπλότητά της (δηλ. το πόσες φορές εμφανίζεται στο δείγμα), σε αντίθεση με τον εκτιμητή HT που λαμβάνει υπόψη τις διακριτές ομάδες που εμφανίζονται στο δείγμα.

## Παρατήρηση 6.4

Όπως έχουμε δει, σύμφωνα με τη μεθοδολογία pps ο υπολογισμός των πιθανοτήτων $\pi_i$ για την επιλογή των ομάδων του πληθυσμού γίνεται βάσει των μεγεθών των ομάδων $K_i$. Στην περίπτωση που τα μεγέθη $K_i$ δεν είναι γνωστά για όλες τις ομάδες του πληθυσμού, τότε γίνεται μια εκτίμηση αυτών με τη βοήθεια μιας μεταβλητής που συνδέεται με το μέγεθος $K_i$ των ομάδων και είναι διαθέσιμη από το δειγματοληπτικό πλαίσιο. Για παράδειγμα, αν ομάδες είναι τα σχολεία μιας περιοχής, και $K_i$ ο αριθμός των μαθητών του σχολείου $i$, τότε μπορεί ο αριθμός των μαθητών να μην είναι διαθέσιμος για όλα τα σχολεία της περιοχής, αλλά να είναι διαθέσιμο ένα άλλο στοιχείο που είναι σχετικό π.χ. το ποσό της κρατικής χρηματικής επιχορήγησης Αν υποθέσουμε ότι οι δύο αυτές μεταβλητές συνδέονται, τότε οι πιθανότητες $\pi_i$ μπορούν να υπολογιστούν, για την ακρίβεια να εκτιμηθούν, χρησιμοποιώντας τη μεταβλητή του μεγέθους της επιχορήγησης ανά σχολείο, αντί για τη μεταβλητή του πλήθους των μαθητών $K_i$. Στις περιπτώσεις αυτές, η μέθοδος δειγματοληψίας λέγεται **δειγματοληψία με πιθανότητες ανάλογες του εκτιμώμενου μεγέθους** (**probabilities proportional to estimated size sampling**) και συμβολίζεται **ppes**.

## Παράδειγμα 6.5

Τα δεδομένα `election` είναι διαθέσιμα στο πακέτο '`survey`' της `R`. Αφορούν δεδομένα από τις προεδρικές εκλογές στην Αμερική το 2004 και είναι καταχωρισμένα ανά πολιτεία και επαρχία (US 2004 presidential election data at state or county level). Τα δεδομένα αποτελούνται από 4600 παρατηρήσεις οι οποίες αντιστοιχούν στις επαρχίες, και 8 μεταβλητές για την κάθε παρατήρηση. Μια επεξήγηση της κάθε μεταβλητής δίνεται μέσω της web σελίδας του πακέτου:

http://127.0.0.1:16059/library/survey/html/election.html

και παρατίθεται στη συνέχεια.

`election` is a data frame with 4600 observations on the following 8 variables.

| | |
|---|---|
| `County` | A factor specifying the state or country |
| `TotPrecincts` | Number of precincts in the state or county |
| `PrecinctsReporting` | Number of precincts supplying data |
| `Bush` | Votes for George W. Bush |
| `Kerry` | Votes for John Kerry |
| `Nader` | Votes for Ralph Nader |

| | |
|---|---|
| `Votes` | Total votes for those three candidates |
| `P` | Sampling probability, proportional to votes |

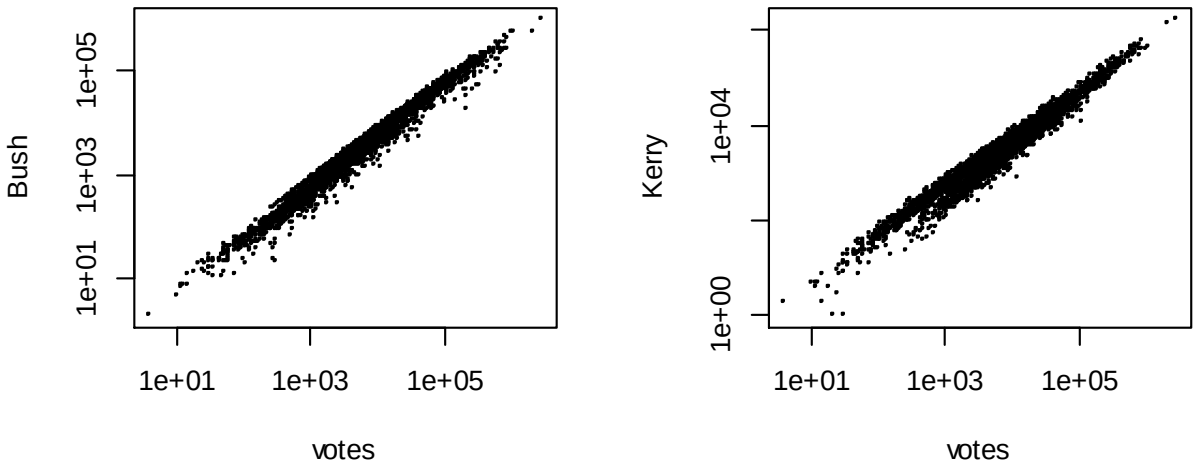
Οι 4600 παρατηρήσεις των δεδομένων παρουσιάζουν πολύ μεγάλη μεταβλητότητα ως προς τον συνολικό αριθμό των κατοίκων που ψήφισαν (στήλη: `votes`). Πράγματι, ζητώντας κάποια απλά στατιστικά στοιχεία για τη μεταβλητή `votes`, έχουμε:

```
> summary(election[,'votes'])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      4    2109    6536   25260   16640 2625000
```

Δηλαδή υπάρχει επαρχία στην οποία ψήφισαν συνολικά 4 κάτοικοι (αντιστοιχεί στο T3 ND Twp) και επαρχία που ψήφισαν 2625000 (αντιστοιχεί στο Los Angeles).

Οι ψήφοι που αποκόμισε ο καθένας από τους 3 υποψήφιους για την προεδρία στην κάθε επαρχία συνδέονται πολύ έντονα, όπως είναι αναμενόμενο, με τον συνολικό αριθμό ψήφων. Το Σχήμα 6.7 δίνει τη γραφική παράσταση (σε λογαριθμική κλίμακα) των ψήφων των 2 επικρατέστερων υποψηφίων Bush και Kerry ανά επαρχία, ως προς τον συνολικό αριθμό ψήφων στην επαρχία. Οι εντολές στην `R` με τις οποίες προκύπτουν οι γραφικές παραστάσεις είναι

```
> plot(election$votes,election$Bush, xlab="votes", ylab="Bush",log="xy")
> plot(election$votes,election$Kerry, xlab="votes", ylab="Kerry",log="xy")
```




***Σχήμα 6.7*** *Γραφική παράσταση των ψήφων των Bush και Kerry ανά επαρχία ως προς τις συνολικές ψήφους της επαρχίας (σε λογαριθμική κλίμακα).*

Θα χρησιμοποιήσουμε τα δεδομένα `election` με σκοπό να εφαρμόσουμε δειγματοληψία κατά ομάδες, με ομάδες τις επαρχίες, σε ένα στάδιο και με πιθανότητες ανάλογες του μεγέθους, όπου ως μέγεθος της επαρχίας θα είναι οι συνολικοί ψήφοι που καταγράφηκαν για την επαρχία και δίνονται στη στήλη με όνομα Votes. Το μέγεθος της δειγματοληψίας είναι 40 και η δειγματοληψία θα γίνει (i) χωρίς επανατοποθέτηση και (ii) με επανατοποθέτηση. Σκοπός μας είναι να εκτιμήσουμε τον συνολικό αριθμό ψήφων των 3 υποψηφίων για όλη τη χώρα καθώς και τα τυπικά σφάλματα των εκτιμήσεων. Το μέγεθος δείγματος έχει επιλεγεί ως 40 για σύγκριση με τα αποτελέσματα ίδιου μεγέθους δείγματος που βρίσκονται στη σελίδα https://cran.r-project.org/web/packages/survey/survey.pdf.

Για την απάντηση των ερωτημάτων του παραδείγματος χρησιμοποιούμε το πακέτο `TeachingSampling` της `R`. Εναλλακτικά, η ίδια ανάλυση μπορεί να γίνει σταδιακά με τη βοήθεια δύο άλλων πακέτων της `R`: (α) το πακέτο sampling για την επιλογή του δείγματος, και (β) το πακέτο survey για την ανάλυση των δεδομένων του δείγματος που προκύπτουν.

Τα διαδοχικά βήματα στο περιβάλλον της `R` είναι:

Αρχικά, καλούμε το πακέτο από το οποίο θα πάρουμε τα δεδομένα της εφαρμογής και, στη συνέχεια, φέρνουμε τα δεδομένα στο περιβάλλον εργασίας:

```
> library(survey)
> data(election)
```

Οι πέντε πρώτες γραμμές των δεδομένων είναι

```
> election[1:5,]
   County TotPrecincts PrecinctsReporting   Bush Kerry Nader  votes           p
1  Alaska          439                439 151876 86064  3890 241830 0.083246769
2 Autauga           22                 22  15212  4774    74  20060 0.006905389
3 Baldwin           50                 50  52910 15579   371  68860 0.023704141
4 Barbour           24                 24   5893  4826    26  10745 0.003698824
5    Bibb           16                 16   5471  2089    12   7572 0.002606561
```

(i) Για το πρώτο ερώτημα (**δειγματοληψία χωρίς επανατοποθέτηση**), επιθυμούμε να εφαρμόσουμε pps στο επίπεδο των ομάδων και να επιλέξουμε ένα δείγμα μεγέθους 40 χωρίς επανατοποθέτηση. Επιλέγουμε το δείγμα με χρήση της κατάλληλης συνάρτησης του πακέτου `TeachingSampling`, με όνομα `S.piPS`. Τα βασικά ορίσματα της συνάρτησης είναι το μέγεθος δείγματος που επιθυμούμε και ένα διάνυσμα που περιέχει τις πιθανότητες $\pi_i$ για κάθε ομάδα του πληθυσμού (άρα διάσταση όση ο πληθυσμός) ή, εναλλακτικά, μια βοηθητική μεταβλητή διαθέσιμη για όλον τον πληθυσμό βάσει της οποίας θα υπολογίσει τις $\pi_i$ υποθέτοντας pps. Για το παράδειγμά μας, η μεταβλητή αυτή είναι η στήλη `votes`.

```
> library(TeachingSampling)
> sample<-S.piPS(40, election$votes)
```

Το αποτέλεσμα της συνάρτησης `S.piPS` θα είναι ένας πίνακας δύο στηλών με γραμμές όσες και το μέγεθος του δείγματος, δηλ. 40 για το παράδειγμα. Στην πρώτη στήλη, είναι οι θέσεις των επιλεγμένων στο δείγμα ομάδων και στη δεύτερη, οι πιθανότητες επιλογής για την καθεμία.

Να σημειώσουμε ότι οι πιθανότητες $\pi_i$ μπορούν να υπολογιστούν και ανεξάρτητα, με την εντολή

```
> p<-40*election$votes/sum(election$votes)
```

και είναι η τελευταία στήλη στο αρχείο των δεδομένων μας.

Για τον υπολογισμό των εκτιμητών των συνολικών ψήφων που αποκόμισαν οι υποψήφιοι βάσει του δείγματος, κάνουμε χρήση της συνάρτησης `E.piPS` του ίδιου πακέτου. Τα δύο βασικά ορίσματα της `E.piPS` είναι το δείγμα (μία ή περισσότερες μεταβλητές για τις οποίες επιθυμούμε να εκτιμήσουμε το σύνολο στον πληθυσμό) και οι πιθανότητες $\pi_i$ που αντιστοιχούν σε κάθε παρατήρηση του δείγματος. Η `E.piPS` επιστρέφει την αριθμητική τιμή του εκτιμητή Horvitz Thompson, καθώς και μια εκτίμηση του τυπικού του σφάλματος και του συντελεστή μεταβλητότητας. Για την εφαρμογή:

```
> E.piPS(election[sample,c('Bush','Kerry', 'Nader')], Pik=sample[,2])
                          N         Bush        Kerry        Nader
Estimation     5034.27811 6.261107e+07 5.318155e+07 4.064882e+05
Standard Error 1423.55676 2.141374e+06 2.116914e+06 7.794600e+04
CVE              28.27728 3.420120e+00 3.980543e+00 1.917547e+01
DEFF                  Inf 2.251434e-04 8.518556e-05 5.705791e-03
```

Άρα, με βάση το δείγμα κατά ομάδες σε ένα στάδιο, μεγέθους 40 επαρχιών από τις 4600 συνολικά του πληθυσμού, εκτιμούμε ότι ο Bush συγκεντρώνει συνολικά 62611070 ψήφους, ο Kerry 53181550 και ο Nader 406488. Τα τυπικά σφάλματα είναι στη δεύτερη γραμμή των αποτελεσμάτων. Η τρίτη και τέταρτη γραμμή δίνουν το εκτιμώμενο CV της εκτίμησης και την επίδραση του δειγματοληπτικού σχεδίου σε σχέση με την α.τ.δ. αντίστοιχα.

(ii) **Δειγματοληψία pps με επανατοποθέτηση**. Το δείγμα λαμβάνεται με την εντολή

```
> wrsample<-S.PPS(40,election$votes)
```

κάνοντας χρήση της κατάλληλης συνάρτησης `S.PPS` του ίδιου πακέτου για επιλογή του δείγματος με επανατοποθέτηση. Η σύνταξη και το αποτέλεσμα της `S.PPS` είναι όπως εκείνο της `S.piPS` συνάρτησης για επιλογή χωρίς επανατοποθέτηση.

Η αντίστοιχη συνάρτηση για την εξαγωγή των εκτιμητών και των ιδιοτήτων τους είναι η `E.PPS`, η οποία ομοίως συντάσσεται όπως και η `E.piPS`. Για την εφαρμογή, θα είναι:

```
>E.PPS(election[wrsample,c('Bush','Kerry','Nader')], pk=wrsample[,2])
                        N         Bush        Kerry        Nader
Estimation     3453.7451 5.905086e+07 5.676093e+07 3.873155e+05
Standard Error 1462.6057 2.697460e+06 2.677947e+06 8.773848e+04
CVE              42.3484 4.568028e+00 4.717941e+00 2.265297e+01
DEFF                 Inf 4.945596e-04 1.375623e-04 7.783956e-03
```

Άρα, με βάση το δεύτερο δείγμα, εκτιμούμε ότι ο Bush συγκεντρώνει συνολικά 59050860 ψήφους, ο Kerry 56760930 και ο Nader 387315. Τα τυπικά σφάλματα των εκτιμητών είναι μεγαλύτερα από εκείνα του αντίστοιχου δείγματος με επανατοποθέτηση.

Τέλος, ως ένα επιπλέον ερώτημα στο Παράδειγμα, έστω ότι οι επαρχίες ήταν χωρισμένες σε στρώματα και επιθυμούσαμε να επαναλάβουμε την ίδια εκτίμηση του ερωτήματος (i) επιλέγοντας τις 40 επαρχίες όχι από τον ενιαίο πληθυσμό, αλλά από τα στρώματα που αποτελούν τον πληθυσμό. Μέσα σε κάθε στρώμα, η επιλογή των επαρχιών θα γίνει και πάλι με pps. Για τα δεδομένα του παραδείγματος δεν υπάρχουν στρώματα. Θα δημιουργήσουμε μια στήλη με στρώματα, μόνο για τους σκοπούς της εφαρμογής.

Έστω ότι οι πρώτες 1000 παρατηρήσεις είναι το 1ο στρώμα, οι επόμενες 1000 είναι το 2ο και όλες οι υπόλοιπες το 3ο στρώμα. Προκειμένου να κάνουμε χρήση της στρωματοποίησης, δημιουργούμε μία νέα στήλη, η οποία θα χρησιμεύσει ως δείκτης του στρώματος. Συνεπώς, στις 8 συνολικά στήλες των δεδομένων, προσθέτουμε μία τελευταία, την ένατη, η οποία θα έχει στοιχεία 1, 2 και 3 κατά μήκος των γραμμών. Ειδικότερα: 1 για τις γραμμές 1 έως 1000, 2 για τις γραμμές 1001 έως 2000 και 3 για τις υπόλοιπες γραμμές δηλ. 2001 έως 4600. Τέλος, ονομάζουμε τη στήλη αυτή `strata` γιατί αυτή δηλώνει το στρώμα στο οποίο ανήκει η κάθε παρατήρηση. Οι εντολές για τη δημιουργία και ονομασία της στήλης είναι:

```
> election[1:1000,9]<-1
> election[1001:2000,9]<-2
> election[2001:4600,9]<-3
> dimnames(election)[[2]][9]<-'strata'
```

Η συνάρτηση `S.STpiPS` είναι η ανάλογη της `S.piPS` για την επιλογή δείγματος pps σε στρωματοποιημένο πληθυσμό. Η σύνταξη είναι εύκολη: δίνει κανείς πρώτα τη στήλη που δηλώνει σε ποιο στρώμα ανήκει η κάθε παρατήρηση του πληθυσμού, τις πιθανότητες συμπερίληψης πρώτης τάξης ή τη βοηθητική μεταβλητή με τη βοήθεια της οποίας θα υπολογιστούν, και ένα διάνυσμα με τα μεγέθη του δείγματος που επιθυμούμε ανά στρώμα. Έστω ότι για το συνολικό δείγμα μεγέθους 40, επιθυμούμε 10 παρατηρήσεις να επιλεγούν από το 1ο στρώμα, 10 από το 2ο και 20 από το 3ο. Άρα, το διάνυσμα των μεγεθών του δείγματος ανά στρώμα θα είναι το `c(10,10,20)`.

```
> st.sample<-S.STpiPS(election$strata,election$votes,c(10,10,20))
```

Η εκτίμηση του αριθμού των ψήφων των υποψηφίων ανά στρώμα και συνολικά για τον πληθυσμό γίνεται μέσω της συνάρτησης `E.STpiPS`, ανάλογης της `E.piPS` για τον μη-στρωματοποιημένο πληθυσμό. Η

σύνταξη είναι παρόμοια, με ένα επιπλέον όρισμα που δίνει την πληροφορία του στρώματος από το οποίο προέρχεται η κάθε παρατήρηση του δείγματος.

```
>E.STpiPS(election[st.sample,c('Bush','Kerry','Nader')],
pik=st.sample[,2], election[st.sample,9])
, , Bush
                          1            2            3   Population
Estimation     1.839233e+07 6.493610e+06 3.515954e+07 6.004548e+07
Standard Error 2.220841e+06 3.636038e+05 2.244177e+06 2.197413e+03
CVE            1.207482e+01 5.599410e+00 6.382840e+00 7.806844e-05
DEFF           1.726668e-02 2.915425e-03 1.236179e-02 5.574997e-13
, , Kerry
                          1            2            3   Population
Estimation     2.071375e+07 4.249511e+06 3.078103e+07 5.574430e+07
Standard Error 2.205509e+06 3.666277e+05 2.216730e+06 2.188348e+03
CVE            1.064756e+01 8.627526e+00 7.201610e+00 8.391851e-05
DEFF           4.422341e-03 9.326844e-03 3.320159e-03 1.598574e-13
, , Nader
                          1            2            3   Population
Estimation     9.652006e+04 3.471623e+04 2.780948e+05 4.093311e+05
Standard Error 3.055273e+04 8.507722e+03 6.944690e+04 3.294045e+02
CVE            3.165428e+01 2.450647e+01 2.497239e+01 4.433943e-03
DEFF           6.517611e-01 1.116180e-02 1.467612e-02 1.442602e-09
```

Η τελευταία στήλη για τον κάθε υποψήφιο αναφέρεται στις εκτιμήσεις για τον συνολικό πληθυσμό και, όπως παρατηρούμε, τα τυπικά σφάλματα αυτά είναι αρκετά μικρότερα από τα αντίστοιχα των ερωτημάτων (i) και (ii) λόγω της στρωματοποιημένης δειγματοληψίας.

## 6.6. Δειγματοληψία κατά ομάδες, σε δύο στάδια με ίσες πιθανότητες

Σύμφωνα με τη μέθοδο δειγματοληψίας κατά ομάδες σε δύο στάδια, πραγματοποιείται μία δειγματοληψία σε καθένα από τα δύο επίπεδα της δειγματοληπτικής μονάδας. Συγκεκριμένα, επιλέγεται αρχικά ένα δείγμα από το σύνολο των ομάδων του πληθυσμού και στη συνέχεια επιλέγεται ένα δείγμα στοιχείων (ssu) από τις ήδη επιλεγμένες ομάδες. Οι λόγοι που οδηγούν σε δειγματοληψία κατά ομάδες σε δύο στάδια αντί για ένα είναι:

(i) Το μεγάλο μέγεθος ορισμένων ομάδων. Αν η δειγματοληψία είναι σε ένα στάδιο, τότε η επιλογή μιας ομάδας με μεγάλο μέγεθος προκαλεί μεγάλη αύξηση στο μέγεθος του δείγματος σε επίπεδο στοιχείων. Εάν λόγω κόστους της έρευνας ή χρόνου διεξαγωγής, υπάρχει μία ανώτατη επιτρεπτή τιμή ως προς το μέγεθος του τελικού δείγματος, υπάρχει ο κίνδυνος η επιλογή μιας μεγάλης ομάδας να συμπληρώσει το μέγεθος αυτό και να αποτρέψει την περαιτέρω επιλογή άλλων ομάδων.

(ii) Η μεγάλη ομοιογένεια των στοιχείων που απαρτίζουν την ομάδα. Σε αυτές τις περιπτώσεις, η απογραφή αυξάνει το κόστος και το μέγεθος της έρευνας, χωρίς να προσφέρει αποτελεσματικούς εκτιμητές.

(iii) Η επιθυμία μεγαλύτερης κάλυψης του πληθυσμού και

(iv) Το κόστος: Μερικές φορές κοστίζει εξίσου ακριβά η δειγματοληψία σε επίπεδο ssu με τη δειγματοληψία σε επίπεδο psu. Στις περιπτώσεις αυτές, είναι προτιμότερο το μέγεθος δείγματος (και άρα και το κόστος της έρευνας) να κατανεμηθεί σε περισσότερα psu του πληθυσμού, με σκοπό την αύξηση της αποτελεσματικότητας των εκτιμητών.

Για τη μελέτη των εκτιμητών και των ιδιοτήτων τους, θεωρούμε, όπως και στη δειγματοληψία σε ένα στάδιο, δύο υποπεριπτώσεις ως προς το μέγεθος των ομάδων.

### 6.6.1 Ομάδες ίσου μεγέθους

Για τη δειγματοληψία κατά ομάδες σε δύο στάδια, με $M$ ομάδες ίσου μεγέθους $K$ στοιχείων, θεωρούμε ότι η διαδικασία δειγματοληψίας αποτελείται από τα βήματα:

(i) Επιλογή με α.τ.δ. $m$ από τις $M$ ομάδες του πληθυσμού.

(ii) Επιλογή με α.τ.δ. $k$ από τα $K$ στοιχεία της $i$ $(i = 1,2, \dots, m)$ επιλεγμένης ομάδας.

Έστω $f_1 = m/M$ και $f_2 = k/K$, το πηλίκο δείγματος για το πρώτο και το δεύτερο στάδιο δειγματοληψίας, αντίστοιχα.

Ας εξετάσουμε στη συνέχεια την εκτίμηση της πληθυσμιακής μέσης τιμής, όταν το δείγμα έχει επιλεγεί σύμφωνα με το δειγματοληπτικό σχέδιο όπως έχει περιγραφεί παραπάνω. Το μέγεθος του δείγματος που θα συλλεγεί τελικά, σε επίπεδο ssu, είναι $n = mk$. Διατηρώντας τον συμβολισμό του Πίνακα 6.3, ένας εκτιμητής για την άγνωστη μέση τιμή $\bar{Y}$ δίνεται ανάλογα με την έκφραση (6.1) για την περίπτωση του ενός σταδίου με ίσα μεγέθη. Θα είναι:

$$\bar{X}_{cl,2} = \hat{\bar{Y}} = \frac{1}{mk}\sum_{i=1}^{m}\sum_{j=1}^{k} X_{ij} \tag{6.20}$$

ή ισοδύναμα:

$$\bar{X}_{cl,2} = \hat{\bar{Y}} = \frac{1}{m}\sum_{i=1}^{m} \bar{X}_i$$

Εύκολα επιβεβαιώνεται, με χρήση της υπόθεσης της α.τ.δ., ότι ο εκτιμητής αυτός, όπως και στην περίπτωση του (6.1), είναι αμερόληπτος. Η διακύμανση του εκτιμητή $\bar{X}_{cl,2}$ προκύπτει από τη διακύμανση του εκτιμητή (6.1) που δίνεται από τη σχέση (6.2), με έναν επιπλέον όρο, ο οποίος εξηγεί την επιπλέον μεταβλητότητα που υπεισέρχεται στην έρευνα λόγω της εκτίμησης αντί για απογραφή των μέσων τιμών των ομάδων $\bar{U}_i$ $(i = 1,2, \dots, M)$. Αναλυτικότερα, η διακύμανση δίνεται από τον τύπο:

$$\mathrm{Var}\left(\bar{X}_{cl,2}\right) = \frac{1}{m}(1 - f_1)\sum_{i=1}^{M}\frac{(\bar{U}_i - \bar{Y})^2}{M-1} + \frac{1}{mk}(1 - f_2)\sum_{i=1}^{M}\sum_{j=1}^{K}\frac{\left(Y_{ij} - \bar{U}_i\right)^2}{M(K-1)} \tag{6.21}$$

Αν συμβολίσουμε (σε αναλογία με την $S_{wsy}^2$ στη συστηματική) με:

$$S_w^2 = \sum_{i=1}^{M}\sum_{j=1}^{K}\frac{\left(Y_{ij} - \bar{U}_i\right)^2}{M(K-1)}$$

τη μέση διακύμανση στο εσωτερικό των ομάδων (average within-cluster variance) και με:

$$S_b^2 = \sum_{i=1}^{M}\frac{(\bar{U}_i - \bar{Y})^2}{M-1}$$

τη διακύμανση μεταξύ των μέσων των ομάδων για τον πληθυσμό (between clusters variance) τότε η διακύμανση του $\bar{X}_{cl,2}$ γράφεται ισοδύναμα:

$$\mathrm{Var}\left(\bar{X}_{cl,2}\right) = \frac{1}{m}(1-f_1)S_b^2 + \frac{1}{mk}(1-f_2)S_w^2$$

Επίσης:

$$\mathrm{Var}\left(\bar{X}_{cl,2}\right) = Var(\bar{X}_{cl}) + \frac{1}{mk}(1-f_2)S_w^2$$

όπου $\mathrm{Var}(\bar{X}_{cl})$ είναι η διακύμανση του εκτιμητή όταν έχει πραγματοποιηθεί δειγματοληψία σε ένα στάδιο. Διαπιστώνουμε συνεπώς, ότι η διακύμανση του $\bar{X}_{cl,2}$ εξαρτάται από την 'μεταξύ των ομάδων' μεταβλητότητα (= $\mathrm{Var}(\bar{X}_{cl})$) η οποία οφείλεται στη δειγματοληψία στο πρώτο στάδιο, και από τη μεταβλητότητα 'στο εσωτερικό των ομάδων' η οποία οφείλεται στη δειγματοληψία έναντι της απογραφής, κατά το δεύτερο στάδιο.

Στην πράξη, για τις περισσότερες έρευνες σε δύο στάδια, ο δεύτερος προσθετέος στη διακύμανση του $\bar{X}_{cl,2}$ είναι αρκετά μικρότερος σε σχέση με τον πρώτο, σε βαθμό που μπορεί να παραλειφθεί. Το φαινόμενο αυτό οφείλεται στο γεγονός ότι στην πράξη συναντάται συνήθως μεγαλύτερη διαφορά στους μέσους μεταξύ των ομάδων, παρά στις παρατηρήσεις που ανήκουν στην ίδια ομάδα.

Μια σχέση που συνδέει τις μεταβλητότητες $S_b^2$ και $S_w^2$ προκύπτει από τον τύπο ΑΝΑΔΙΑ για τον πληθυσμό ο οποίος είναι χωρισμένος σε ομάδες:

$$\frac{N-1}{N}S^2 = \frac{M-1}{M}S_b^2 + \frac{K-1}{K}S_w^2$$

Εκτιμώντας τις μεταβλητότητες $S_b^2$ και $S_w^2$ με τις αντίστοιχες δειγματικές:

$$s_b^2 = \sum_{i=1}^{m} \frac{\left(\bar{X}_i - \hat{\bar{Y}}\right)^2}{m-1}$$

και:

$$s_w^2 = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{\left(X_{ij} - \bar{X}_i\right)^2}{m(k-1)}$$

η εκτιμώμενη διακύμανση του $\bar{X}_{cl,2}$ αποδεικνύεται ότι δίνεται από τη σχέση:

$$\hat{\mathrm{Var}}\left(\bar{X}_{cl,2}\right) = \frac{1}{m}(1-f_1)s_b^2 + \frac{1}{Mk}(1-f_2)s_w^2 \qquad (6.22)$$

Το τυπικό σφάλμα και το εκτιμώμενο τυπικό σφάλμα του εκτιμητή $\bar{X}_{cl,2}$ προκύπτουν κατά τα γνωστά ως η τετραγωνική ρίζα των σχέσεων (6.21) και (6.22) αντίστοιχα.

### 6.6.2 Ομάδες άνισου μεγέθους

Κάτω από την υπόθεση του άνισου μεγέθους των συστάδων, η δειγματοληψία σε δύο στάδια αποτελείται από τα εξής βήματα:

(i) Επιλογή με α.τ.δ. $m$ από τις $M$ ομάδες του πληθυσμού.

(ii) Επιλογή με α.τ.δ. $k_i$ από τα $K_i$ στοιχεία της $i$ $(i = 1,2,\dots,m)$ επιλεγμένης ομάδας

Έστω $f_1 = m/M$ και $f_{2i} = k_i/K_i$ $(i = 1,2,\dots,m)$ το πηλίκο δείγματος για το πρώτο και το δεύτερο στάδιο δειγματοληψίας αντίστοιχα.

Ο εκτιμητής που υιοθετείται ως ο εκτιμητής του μέσου του πληθυσμού είναι ο:

$$\bar{X}'_{cl,2} = \hat{\bar{Y}} = \frac{M}{Nm}\sum_{i=1}^{m} K_i \frac{1}{k_i}\sum_{j=1}^{k_i} X_{ij} = \frac{M}{Nm}\sum_{i=1}^{m} K_i \hat{\bar{X}}_i \tag{6.23}$$

όπου $N = K_1 + K_2 + \cdots + K_M$.

Ο εκτιμητής αποδεικνύεται ότι είναι αμερόληπτος, και ένας αμερόληπτος εκτιμητής της διακύμανσής του δίνεται από τη σχέση:

$$\mathrm{Var}\left(\bar{X}'_{cl,2}\right) = \frac{1}{m\bar{K}^2}(1-f_1)\sum_{i=1}^{m}\frac{\left(K_i\hat{\bar{X}}_i - \hat{Y}_T/M\right)^2}{m-1} + \frac{f_1}{m^2\bar{K}^2}\sum_{i=1}^{m}(1-f_{2i})K_i^2 s_i^2/k_i \tag{6.24}$$

Ανάλογα με την περίπτωση της δειγματοληψίας σε ένα στάδιο, ορίζεται και ο εκτιμητής λόγου για την πληθυσμιακή μέση τιμή, που δεν έχει την ιδιότητα της αμεροληψίας.

## Παράδειγμα 6.6

Τα δεδομένα είναι από το βιβλίο Levy & Lemeshow (1999, σελ. 309). Έστω ότι σε μια περιοχή με δέκα Νοσοκομεία συνολικά, επιλέγονται 3 εξ αυτών με ίσες πιθανότητες, και από το κάθε Νοσοκομείο του δείγματος, επιλέγεται στη συνέχεια ένα υποσύνολο ασθενών του, με σκοπό να γίνει η καταγραφή και η εκτίμηση στη συνέχεια του συνολικού αριθμού ασθενών που δεν κατάφεραν να βγουν από το Νοσοκομείο εν ζωή. Τα δεδομένα είναι στη σελίδα http://www.ats.ucla.edu/stat/examples/sop/ και αποτελούνται από 708 γραμμές που αντιστοιχούν σε ασθενείς και 9 στήλες. Ενδεικτικά, τα δεδομένα έχουν τη μορφή:

```
> pt210
    HOSPNO  ID LIFETHRT DXDEAD       W1       W2       W  M X_TOTAL_
1        1   1        1      1 3.333332 9.995338 33.31778 10     4288
2        1   2        1      1 3.333332 9.995338 33.31778 10     4288
3        1   3        1      1 3.333332 9.995338 33.31778 10     4288
4        1   4        1      1 3.333332 9.995338 33.31778 10     4288
.        .   .        .      . .        .        .        .      .
443      4 443        1      0 3.333332  9.968750 33.22916 10     638
444      4 444        1      0 3.333332  9.968750 33.22916 10     638
445      4 445        1      0 3.333332  9.968750 33.22916 10     638
446      4 446        1      0 3.333332  9.968750 33.22916 10     638
.        .   .        .      . .        .        .        .      .
503     10 503        1      0 3.333332 10.018600 33.39532 10     2154
504     10 504        1      0 3.333332 10.018600 33.39532 10     2154
505     10 505        1      0 3.333332 10.018600 33.39532 10     2154
506     10 506        1      0 3.333332 10.018600 33.39532 10     2154
```

Η μεταβλητή `HOSPNO` είναι ο α/α του Νοσοκομείου που έχει επιλεγεί στο δείγμα (1, 4 και 10 για τα δεδομένα γιατί επιλέχθηκαν οι τρεις ομάδες με α/α 1, 4, 10), η μεταβλητή `LIFETHRT` είναι η δίτιμη μεταβλητή που δηλώνει εάν ο ασθενής κατά την εισαγωγή του στο Νοσοκομείο ήταν σε κρίσιμη κατάσταση υγείας ώστε να κινδυνεύει η ζωή του (1 = ναι, 0 = όχι), `DXDEAD` η δίτιμη μεταβλητή που δηλώνει εάν ο ασθενής επιβιώνει μετά την εισαγωγή του στο Νοσοκομείο (1 = όχι, 0 = ναι), `M` είναι ο αριθμός των Νοσοκομείων συνολικά, και `X_TOTAL_` η μεταβλητή που δίνει το σύνολο των ασθενών για τα Νοσοκομεία του δείγματος.

Έστω ότι έχουμε φέρει τα δεδομένα στο περιβάλλον της `R` και ότι είναι αποθηκευμένα με το όνομα `pt210`. Η ανάλυση των δεδομένων και ο υπολογισμός των εκτιμητών θα γίνει με το πακέτο `TeachingSampling`.

Η κατάλληλη συνάρτηση για τη δειγματοληψία κατά ομάδες σε δύο στάδια είναι η `E.2SI`, η οποία έχει μια σειρά από ορίσματα:

`NI:` αριθμός ομάδων στον πληθυσμό

`nI:` αριθμός ομάδων στο δείγμα

`Ni:` αριθμός στοιχείων (ssu) σε κάθε ομάδα του δείγματος

`ni:` αριθμός στοιχείων (ssu) που επιλέγονται (από τις Ni)

`y:` μεταβλητή για την οποία ενδιαφερόμαστε να εκτιμήσουμε το σύνολο

`PSU:` μεταβλητή που δηλώνει σε ποια ομάδα ανήκει η κάθε παρατήρηση (πρέπει να είναι μεταβλητή factor).

Για τα δεδομένα, θα είναι

```
> dim(pt210[pt210$HOSPNO==1,])
[1] 429  9
> dim(pt210[pt210$HOSPNO==4,])
[1] 64 9
> dim(pt210[pt210$HOSPNO==10,])
[1] 215  9
```

δηλ. 429 είναι οι ασθενείς του δείγματος που επιλέχθηκαν από το Νοσοκομείο με α/α 1 με συνολικό αριθμό ασθενών 4288 (τελευταία στήλη), 64 από το Νοσοκομείο 4 και 215 από το Νοσοκομείο 10. Τα υπόλοιπα στοιχεία είναι όλα άμεσα διαθέσιμα, οπότε:

```
> E.2SI(10, 3, c(4288,638,2154), c(429,64,215), pt210[,c(3,4)],
as.factor(pt210[,1]))
                            N    LIFETHRT      DXDEAD
Estimation     23600.00000 2932.320491 499.3791798
Standard Error  8857.60765  773.082694 116.0722562
CVE               37.53224   26.364195  23.2433111
DEFF                   Inf    7.185678   0.8502028
```

Άρα, βάσει του δείγματος, εκτιμάται ότι ο συνολικός αριθμός των ασθενών οι οποίοι δεν επιβιώνουν είναι 499.3791798 και ο συνολικός αριθμός εκείνων που ήταν σε κρίσιμη κατάσταση κατά την εισαγωγή τους είναι 2932.320491. Τα τυπικά σφάλματα των εκτιμητών είναι 116.0722562 και 773.082694 αντίστοιχα.

Για την εκτίμηση του ποσοστού των ασθενών που δεν επιβιώνουν μεταξύ εκείνων που κατά την εισαγωγή τους ήταν σε κρίσιμη κατάσταση:

```
> p<-499.3791798/2932.320491
> p
[1] 0.1703017
```

Άρα, το ποσοστό αυτό εκτιμάται σε 17%.

# Βιβλιογραφικές Αναφορές


Alf, C. & Lohr, S. (2007). Sampling Assumptions in Introductory Statistics Classes. *American Statistician, 6*, 71-77.

Brewer, K.R.W. (2002). *Combined survey sampling inference (Weighing Basu's elephants) [Chapter 9]*. London: Hodder Education

Cochran, W. G. (1977). *Sampling techniques* (3rd Edition). New York: John Wiley and Sons.

Levy, P.S. and Lemeshow, S. (1999). *Sampling of populations. Methods and applications* (3rd Edition). New York: John Wiley and Sons.

Hansen, M. M. and Hurwitz, W. N. (1943). On the Theory of Sampling from Finite Populations. *Annals of Mathematical Statistics*. *14:* 333-362.

Horvitz, D. G. and Thompson, M. E. (1952). A Generalization of Sampling without Replacement from a Finite Universe. *Journal of the American Statistical Association. 47* (260): 663-685.

Overton, W. S., D. White, and D. L. Stevens Jr. (1990). *Design report for EMAP, Environmental Monitoring and Assessment Program*. EPA 600/3-91/053. Corvallis, OR: U.S. Environmental. Protection Agency, Environmental Research Laboratory.

Stehman, S. V. and Overton, W. S. (1987). Estimating the Variance of the Horvitz-Thompson Estimator in Variable Probability, Systematic Samples. *Proceeding of the Section on Survey Research Methods. American Statistical Association Annual Meeting*. Pp. 743-748.

Sen, A. R. (1953). On the Estimate of the Variance in Sampling with Varying Probabilities. *Journal of the Indian Society of Agricultural Statistics. 5*: 119-127.

Thompson, S. K. (2012). *Sampling* (3rd Edition). Hoboken, NJ: John Wiley and Sons.

Yates, F. and Grundy, P. M. (1953). Selection without Replacement from Within. Strata with Probability Proportional to Size. *Journal of the Royal Statistical Society*. *15* (1): 253-261.

# Κεφάλαιο 7 - ΕΚΤΙΜΗΤΕΣ ΛΟΓΟΥ ΚΑΙ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

**Σύνοψη**

*Οι εκτιμητές λόγου και παλινδρόμησης (ratio and regression estimators) μιας πληθυσμιακής παραμέτρου ($\bar{Y}$, $\bar{Y}_T$, $P$) έχουν ως κύριο γνώρισμα ότι βασίζονται σε μια βοηθητική μεταβλητή. Η ιδέα και στις δύο περιπτώσεις εκτίμησης είναι να γίνει χρήση διαθέσιμων δεδομένων μίας ή περισσότερων μεταβλητών που σχετίζονται με την κύρια μεταβλητή της έρευνας, με σκοπό να βελτιωθεί η εκτίμηση των παραμέτρων του πληθυσμού για την κύρια μεταβλητή. Ο εκτιμητής λόγου ορίζεται με τη βοήθεια του λόγου των μέσων τιμών ή των συνόλων δύο χαρακτηριστικών, ενώ ο εκτιμητής παλινδρόμησης βασίζεται στην προσαρμογή ενός γραμμικού μοντέλου παλινδρόμησης με ανεξάρτητη μεταβλητή την κύρια μεταβλητή της έρευνας και εξαρτημένη/ες τις βοηθητικές. Ο εκτιμητής λόγου δίνει πιο ακριβή αποτελέσματα όταν ο συντελεστής συσχέτισης μεταξύ της κύριας και της βοηθητικής μεταβλητής είναι μεγάλος. Ο εκτιμητής παλινδρόμησης είναι πάντα πιο αποτελεσματικός από τον εκτιμητή που υπολογίζεται χωρίς χρήση βοηθητικής μεταβλητής η δε αποτελεσματικότητά του είναι τόσο μεγαλύτερη, όσο μεγαλύτερος είναι ο συντελεστής συσχέτισης μεταξύ των μεταβλητών.*

## 7.1. Εισαγωγή

Στο κεφάλαιο αυτό, παρουσιάζονται δύο τρόποι εκτίμησης οι οποίοι κάνουν χρήση μιας βοηθητικής μεταβλητής και **σκοπεύουν στη βελτίωση της ακρίβειας των εκτιμητών** που επιτυγχάνονται κατά τη δειγματοληπτική έρευνα. Είναι ο εκτιμητής λόγου ή πηλίκου και ο εκτιμητής παλινδρόμησης.

Στο Κεφάλαιο 4 -(παράγρ. 4.1 και 4.2), είδαμε τη χρήση μιας βοηθητικής μεταβλητής κατά το στάδιο του σχεδιασμού της έρευνας και, ειδικότερα, στον καθορισμό του αριθμού και των ορίων των στρωμάτων σε μια στρωματοποιημένη δειγματοληψία. Επίσης, στο Κεφάλαιο 5 - (παράγρ. 5.6), είδαμε πώς μπορεί να χρησιμοποιηθεί μια βοηθητική μεταβλητή κατά την εφαρμογή της συστηματικής δειγματοληψίας, ώστε να διαταχθούν οι μονάδες του πληθυσμού και να επιτευχθεί μια αύξουσα ή φθίνουσα διάταξη με σκοπό τη βελτίωση της ακρίβειας των εκτιμητών της συστηματικής. Στις περιπτώσεις που περιγράψαμε, η χρήση της βοηθητικής μεταβλητής γίνεται μόνο κατά τη διάρκεια του σχεδιασμού της έρευνας και βοηθά στον επακριβή καθορισμό του δειγματοληπτικού σχεδίου. Η κεντρική ιδέα του εκτιμητή λόγου και του εκτιμητή παλινδρόμησης είναι να γίνει χρήση μιας βοηθητικής μεταβλητής που σχετίζεται με την κύρια μεταβλητή της έρευνας στην ίδια την έκφραση του εκτιμητή και όχι μόνο στον σχεδιασμό της έρευνας. Η καθεμιά εκτίμηση αξιοποιεί με διαφορετικό τρόπο τη βοηθητική μεταβλητή, αλλά και στις δύο περιπτώσεις ο ορισμός του εκτιμητή εμπλέκει δύο χαρακτηριστικά. Το κυρίως χαρακτηριστικό ή την κύρια ερώτηση ή μεταβλητή της έρευνας και τη βοηθητική.

Ο ορισμός των εκτιμητών είναι αρκετά γενικός και μπορεί να εφαρμοστεί σε συνδυασμό με κάθε μέθοδο δειγματοληψίας που έχουμε μελετήσει. Ως προς τη διεξαγωγή της έρευνας, αυτό που αλλάζει και που είναι απαραίτητο συστατικό του εκτιμητή λόγου και του εκτιμητή παλινδρόμησης είναι ότι για κάθε μέλος του δείγματος καταμετρώνται ταυτόχρονα οι απαντήσεις σε **δύο** χαρακτηριστικά, το κύριο και το βοηθητικό. Το όφελος στην αποτελεσματικότητα που επιτυγχάνουν οι δύο μέθοδοι εκτίμησης βασίζεται κατά κύριο λόγο στη σχέση μεταξύ των 2 μεταβλητών. Τα ζεύγη των παρατηρήσεων του δείγματος για τις δύο μεταβλητές μάς δίνουν την πληροφορία για τη σχέση αυτή μεταξύ των δύο μεταβλητών.

Για την εφαρμογή του εκτιμητή λόγου και του εκτιμητή παλινδρόμησης, εκτός από τις μετρήσεις του δείγματος για τα 2 χαρακτηριστικά, απαιτείται από τον ερευνητή **προηγούμενη γνώση** (π.χ. από απογραφή) για την πληθυσμιακή μέση τιμή ή το σύνολο του βοηθητικού χαρακτηριστικού. Πρακτικά, αυτό δεν αποτελεί μειονέκτημα των μεθόδων, γιατί ο ερευνητής έχει συνήθως πρόσβαση σε στοιχεία από απογραφές, και μάλιστα για περισσότερες από μία βοηθητικές μεταβλητές που σχετίζονται με το κυρίως χαρακτηριστικό της έρευνας. Τα δύο κριτήρια επιλογής της βοηθητικής μεταβλητής είναι α) να σχετίζεται όσο το δυνατό περισσότερο με την κύρια μεταβλητή και β) να υπάρχουν πρόσφατα αξιόπιστα συγκεντρωτικά στοιχεία της για τον πληθυσμό.

Το κεφάλαιο χωρίζεται σε δύο βασικά μέρη, τα οποία περιλαμβάνουν τη μελέτη του **εκτιμητή λόγου** και τη μελέτη του **εκτιμητή παλινδρόμησης**. Για την κάθε εκτίμηση, δίνεται ο ορισμός και οι ιδιότητες του εκτιμητή, η σύγκρισή του με τον εκτιμητή που υπολογίζεται από ένα απλό τυχαίο δείγμα χωρίς τη χρήση

βοηθητικής μεταβλητής, η εφαρμογή του εκτιμητή στη στρωματοποιημένη δειγματοληψία, καθώς και η επέκτασή του στην περίπτωση περισσότερων του ενός βοηθητικών χαρακτηριστικών. Τέλος, οι δύο εκτιμητές συγκρίνονται μεταξύ τους.

## 7.2. Εκτίμηση λόγου δύο χαρακτηριστικών

Ο εκτιμητής λόγου για την πληθυσμιακή παράμετρο ενός χαρακτηριστικού, π.χ. μέση τιμή ή σύνολο, ορίζεται με τη βοήθεια του εκτιμητή του λόγου δύο χαρακτηριστικών. Πριν δώσουμε τον τρόπο βελτίωσης της εκτίμησης μιας παραμέτρου ενός χαρακτηριστικού με τη βοήθεια ενός δεύτερου, θα μελετήσουμε το πρόβλημα της εκτίμησης του λόγου δύο χαρακτηριστικών έστω $Y_1$ και $Y_2$ τα οποία ισοδύναμα ή ταυτόχρονα ενδιαφέρουν τον ερευνητή. Ειδικότερα, έστω ότι το ερώτημα που ενδιαφέρει είναι να εκτιμήσουμε την αναλογία ή τον λόγο που παρουσιάζουν οι δύο μεταβλητές $Y_1$ και $Y_2$ για το σύνολο των μελών του πληθυσμού.

Για παράδειγμα, αν $Y_1$ είναι το μηνιαίο συνολικό εισόδημα των νοικοκυριών μιας χώρας και $Y_2$ είναι το ποσό που διαθέτουν τα νοικοκυριά για την κάλυψη των βασικών ειδών διατροφής τους, ένα ερώτημα που θα ενδιέφερε τον ερευνητή είναι να εκτιμήσει τον λόγο των δύο αυτών μεταβλητών. Με τον τρόπο αυτό, θα μπορέσει να βγάλει συμπεράσματα για το τι μέρος του εισοδήματος ενός νοικοκυριού απαιτείται, προκειμένου να καλυφθούν οι βασικές ανάγκες διατροφής του. Κατά συνέπεια, μπορούν επίσης να εξαχθούν συμπεράσματα και για το βιοτικό επίπεδο ή για την αγοραστική ικανότητα που έχουν οι κάτοικοι της χώρας κτλ. Ένα άλλο παράδειγμα, όπου ο λόγος δύο χαρακτηριστικών μπορεί να αποτελεί το πρωταρχικό ερώτημα της έρευνας, είναι η συνολική έκταση των καλλιεργήσιμων κτημάτων μιας χώρας και η έκταση εκείνων που καλλιεργούν ένα συγκεκριμένο είδος σιτηρού (π.χ. βρώμη). Επίσης, σε ένα ιχθυοτροφείο θα είχε πιθανά ενδιαφέρον να εκτιμηθεί ο αριθμός των ψαριών που έχουν βάρος μεγαλύτερο π.χ. από 0.5 Kg ως προς τον συνολικό αριθμό των ψαριών του ιχθυοτροφείου. Αυτός ο λόγος, με τη σειρά του, θα δώσει πληροφορία για τη διατροφή των ψαριών και για την ετοιμότητα κάλυψης των προμηθευτών.

Ένας άλλος λόγος που αρκετά συχνά ενδιαφέρει να εκτιμήσουμε το πηλίκο δύο χαρακτηριστικών και όχι κατά απόλυτο μέγεθος το καθένα από αυτά, είναι όταν τα αποτελέσματα της έρευνας προορίζονται να συγκριθούν σε επόμενο επίπεδο με αντίστοιχα αποτελέσματα διαφορετικού πληθυσμού. Για παράδειγμα, μπορεί να πρέπει ο λόγος του μηνιαίου εισοδήματος των νοικοκυριών προς τα έξοδά τους για την κάλυψη των βασικών ειδών διατροφής να εκτιμηθεί χωριστά για κάθε χώρα της Ευρωπαϊκής Ένωσης και στη συνέχεια να γίνουν συγκρίσεις μεταξύ των χωρών.

### 7.2.1 Συμβολισμός και Ορισμός εκτιμητή λόγου δύο χαρακτηριστικών

Έστω ότι $Y_1$ και $Y_2$ είναι δύο χαρακτηριστικά του πληθυσμού, των οποίων ενδιαφερόμαστε να εκτιμήσουμε τον λόγο. Αν $\{Y_{11}, Y_{12}, \dots, Y_{1N}\}$ είναι οι τιμές του χαρακτηριστικού $Y_1$ για τα $N$ μέλη του πληθυσμού και, ανάλογα, $\{Y_{21}, Y_{22}, \dots, Y_{2N}\}$ είναι οι τιμές του χαρακτηριστικού $Y_2$ για τα μέλη του ίδιου πληθυσμού, τότε ο λόγος $R$ των δύο χαρακτηριστικών ορίζεται ως:

$$R = \frac{\bar{Y}_1}{\bar{Y}_2} \tag{7.1}$$

Προφανώς, αν $Y_{1T}$ και $Y_{2T}$ είναι τα σύνολα των δύο χαρακτηριστικών, τότε ισοδύναμα θα ισχύει:

$$R = \frac{Y_{1T}}{Y_{2T}}$$

γιατί τόσο οι υπολογισμοί των μέσων, όσο και των συνόλων των δύο χαρακτηριστικών, γίνονται για το ίδιο σύνολο των μελών του πληθυσμού.

Συνεπώς, $R$ είναι η αληθινή τιμή της παραμέτρου του λόγου των δύο χαρακτηριστικών για τον πληθυσμό ή ο **πληθυσμιακός λόγος** όπως θα λέμε στη συνέχεια. Το πρόβλημα της εκτίμησης είναι να βρούμε έναν

εκτιμητή του $R$, έστω $\hat{R}$, επιλέγοντας ένα δείγμα μεγέθους $n$ από τον πληθυσμό. Υποθέτουμε αρχικά ότι η επιλογή του δείγματος γίνεται σύμφωνα με την απλή τυχαία δειγματοληψία.

Ο προφανέστερος εκτιμητής $\hat{R}$ του πληθυσμιακού λόγου $R$ είναι αυτός που υπολογίζεται σε αναλογία με τον ορισμό (7.1) του $R$, αλλά βασίζεται στις δειγματικές, αντί για τις πληθυσμιακές, ποσότητες των χαρακτηριστικών $Y_1$ και $Y_2$. Πιο συγκεκριμένα:

$$\hat{R} = \frac{\hat{\bar{Y}}_1}{\hat{\bar{Y}}_2} = \frac{\hat{Y}_{1T}}{\hat{Y}_{2T}} \tag{7.2}$$

όπου $\hat{\bar{Y}}_1$ είναι ο εκτιμητής του $\bar{Y}_1$ βάσει του α.τ.δ. μεγέθους $n$, δηλαδή ο δειγματικός μέσος των μετρήσεων για το πρώτο ερώτημα της έρευνας και, ομοίως, $\hat{\bar{Y}}_2$ ο δειγματικός μέσος των μετρήσεων για το δεύτερο ερώτημα.

Αν, σύμφωνα με τον συμβολισμό του Κεφάλαιο 1 -, $\{X_{11}, X_{12}, \dots, X_{1n}\}$ και $\{X_{21}, X_{22}, \dots, X_{2n}\}$ είναι οι απαντήσεις των $n$ μελών του δείγματος για τις ερωτήσεις $Y_1$ και $Y_2$ αντίστοιχα, τότε ισχύει ισοδύναμα:

$$\hat{R} = \frac{\bar{X}_1}{\bar{X}_2} = \frac{X_{1T}}{X_{2T}}$$

όπου:

$$\bar{X}_1 = \frac{1}{n}\sum_{i=1}^{n} X_{1i}, \quad \bar{X}_2 = \frac{1}{n}\sum_{i=1}^{n} X_{2i}, \qquad X_{1T} = \sum_{i=1}^{n} X_{1i} \quad \text{και} \quad X_{2T} = \sum_{i=1}^{n} X_{2i}.$$

Προφανώς, $\bar{X}_1 = \hat{\bar{Y}}_1,\ \bar{X}_2 = \hat{\bar{Y}}_2,\ X_{1T} = \hat{Y}_1,\ X_{2T} = \hat{Y}_2$.

Για να τονίσουμε ότι και οι δύο μετρήσεις λαμβάνονται στα ίδια $n$ μέλη του δείγματος, γράφουμε το δείγμα ως ζεύγη παρατηρήσεων:

$$\{(X_{11}, X_{21}), (X_{12}, X_{22}), \dots, (X_{1n}, X_{2n})\}.$$

Ένας εκτιμητής του λόγου των δύο χαρακτηριστικών, διαφορετικός από τον $\hat{R}$, είναι εκείνος που ορίζεται ως ο δειγματικός μέσος των λόγων των τιμών των δύο χαρακτηριστικών για τα μέλη του δείγματος. Αν συμβολίσουμε $\hat{R}'$ τον εκτιμητή αυτόν, θα είναι αναλυτικότερα:

$$\hat{R}' = \frac{1}{n}\sum_{i=1}^{n} \frac{X_{1i}}{X_{2i}} \tag{7.3}$$

Ο εκτιμητής $\hat{R}'$ είναι μεροληπτικός και δεν χρησιμοποιείται συχνά στην πράξη, παρά την εύχρηστη μορφή του, γιατί το μέσο τυπικό σφάλμα του είναι μεγαλύτερο από το αντίστοιχο άλλων εκτιμητών και ειδικά από του εκτιμητή $\hat{R}$. Για πλήρη μελέτη του εκτιμητή $\hat{R}'$, βλ. Barnett, 2002, Κεφ. 3, παράγρ. 3.1.

### 7.2.2 Ιδιότητες του εκτιμητή λόγου δύο χαρακτηριστικών

Θα μελετήσουμε τις ιδιότητες του εκτιμητή $\hat{R}$ όπως αυτός ορίζεται μέσω της (7.2), και κάτω από την υπόθεση ότι πραγματοποιήθηκε α.τ.δ., τα αποτελέσματα της οποίας είναι τα ζεύγη των παρατηρήσεων $\{(X_{11}, X_{21}), (X_{12}, X_{22}), \dots, (X_{1n}, X_{2n})\}$.

**Πρόταση 7.1**

Ο εκτιμητής $\hat{R}$ δεν είναι αμερόληπτος εκτιμητής του $R$. Αναλυτικότερα, ισχύει ότι το ποσό μεροληψίας του $\hat{R}$ δίνεται από τη σχέση $\text{Bias}(\hat{R}) = \frac{1}{\bar{Y}_2}\text{Cov}(\hat{R}, \hat{\bar{Y}}_2)$.

**Απόδειξη**

Υπολογίζουμε το ποσό μεροληψίας για τον εκτιμητή $\hat{R}$. Σύμφωνα με τον ορισμό:

$$\text{Bias}(\hat{R}) = \text{E}(\hat{R}) - R = \text{E}(\hat{R}) - \frac{\bar{Y}_1}{\bar{Y}_2} = \text{E}(\hat{R}) - \frac{\bar{Y}_1}{\bar{Y}_2} = \frac{1}{\bar{Y}_2}\left[\text{E}(\bar{Y}_2\hat{R} - \bar{X}_1) + \text{E}(\bar{X}_1 - \bar{Y}_1)\right]$$

$$= \frac{1}{\bar{Y}_2}\text{E}(\bar{Y}_2\hat{R} - \bar{X}_1) = \frac{1}{\bar{Y}_2}\text{E}(\hat{R}\bar{Y}_2 - \hat{R}\bar{X}_2) = \frac{1}{\bar{Y}_2}\text{E}\left[\hat{R}(\bar{Y}_2 - \text{E}(\bar{X}_2))\right]$$

$$= \frac{1}{\bar{Y}_2}\text{Cov}(\hat{R}, \bar{X}_2) = \frac{1}{\bar{Y}_2}\text{Cov}\left(\hat{R}, \hat{\bar{Y}}_2\right)$$

Για τη μετάβαση από τη σχέση $\frac{1}{\bar{Y}_2}\text{E}\left[\hat{R}(\bar{Y}_2 - \bar{X}_2)\right]$ στη σχέση $-\frac{1}{\bar{Y}_2}\text{E}\left[(R - \hat{R})(\bar{Y}_2 - \bar{X}_2)\right]$, παρατηρήστε ότι $\text{E}[R(\bar{Y}_2 - \bar{X}_2)]$ ισούται με μηδέν (0), επειδή $E(\bar{X}_2) = \bar{Y}_2$.

Άρα, το ποσό μεροληψίας του εκτιμητή $\hat{R}$ δεν είναι μηδέν και κατά συνέπεια ο εκτιμητής δεν είναι αμερόληπτος. Ειδικότερα, το ποσό μεροληψίας του $\hat{R}$ είναι ανάλογο της συνδιακύμανσης μεταξύ του εκτιμητή του λόγου των δύο χαρακτηριστικών και του εκτιμητή του πληθυσμιακού μέσου για τη βοηθητική μεταβλητή. ■

Η συνδιακύμανση $\text{Cov}\left(\hat{R}, \hat{\bar{Y}}_2\right)$ γίνεται μικρή, όταν η γραμμή παλινδρόμησης μεταξύ των μεταβλητών $Y_1$ και $Y_2$ διέρχεται από την αρχή των αξόνων.

**Πόρισμα 7.1**

Για τον εκτιμητή $\hat{R}$ αποδεικνύεται ότι $\frac{|\text{Bias}(\hat{R})|}{\sqrt{\text{Var}(\hat{R})}} \leq \text{CV}\left(\hat{\bar{Y}}_2\right)$

**Απόδειξη** (Hartley and Ross, 1954)

Από την Πρόταση 7.1 και τον ορισμό του CV θα έχουμε:

$$\frac{|\text{Bias}(\hat{R})|}{\sqrt{\text{Var}(\hat{R})}} = \frac{\left|\frac{1}{\bar{Y}_2}\rho\left(\hat{R}, \hat{\bar{Y}}_2\right)\right|\sqrt{\text{Var}(\hat{R})\text{Var}\left(\hat{\bar{Y}}_2\right)}}{\sqrt{\text{Var}(\hat{R})}} \leq \frac{\sqrt{\text{Var}\left(\hat{\bar{Y}}_2\right)}}{|\bar{Y}_2|} = \text{CV}\left(\hat{\bar{Y}}_2\right)$$

■

Σύμφωνα με το Πόρισμα 7.1, το ποσό μεροληψίας του $\hat{R}$ ως προς το τυπικό του σφάλμα είναι μικρότερο από τον συντελεστή μεταβλητότητας του εκτιμητή της βοηθητικής μεταβλητής. Η συνδιακύμανση $\text{CV}\left(\hat{\bar{Y}}_2\right)$ είναι άγνωστη, αλλά είναι εφικτό να δώσουμε ένα άνω φράγμα για την ποσότητα αυτή, κάνοντας χρήση γνωστών αποτελεσμάτων από την srs. Πιο αναλυτικά, για την απλή τυχαία δειγματοληψία, γνωρίζουμε ότι:

$$\text{se}\left(\hat{\bar{Y}}_2\right) = \frac{\sqrt{1-f}\sqrt{S_2^2}}{\sqrt{n}}$$

όπου $S_2^2$ η πληθυσμιακή διακύμανση των τιμών της $Y_2$ μεταβλητής και $n$ το μέγεθος του δείγματος. Επειδή $1 - f \leq 1$, έπεται ότι:

$$\text{se}\left(\hat{\bar{Y}}_2\right) \leq \frac{\sqrt{S_2^2}}{\sqrt{n}}$$

απ' όπου λαμβάνεται ότι:

$$\mathrm{CV}\left(\hat{\bar{Y}}_2\right) \le \frac{\sqrt{S_2^2}}{\sqrt{n}|\bar{Y}_2|}$$

Άρα, συνολικά, το ποσό μεροληψίας του $\hat{R}$ ως προς το τυπικό του σφάλμα γίνεται μικρότερο αριθμητικά, όσο το μέγεθος του δείγματος αυξάνει. Μία εκτίμηση του άνω φράγματος της συνδιακύμανσης $\mathrm{CV}\left(\hat{\bar{Y}}_2\right)$ υπολογίζεται αντικαθιστώντας, στη θέση των ποσοτήτων $S_2^2$ και $\bar{Y}_2$ στο δεξί μέλος της ανισότητας, τους εκτιμητές τους, όπως αυτοί υπολογίζονται με τη βοήθεια του δείγματος.

Η Πρόταση 7.1 δίνει το ακριβές αποτέλεσμα για το ποσό μεροληψίας του εκτιμητή $\hat{R}$. Παρόλα αυτά, το αποτέλεσμα αυτό δεν είναι εύχρηστο για υπολογισμούς στην πράξη αμέσως μετά τη συλλογή ενός δείγματος, γιατί δεν είναι γνωστή η συνδιακύμανση $\mathrm{Cov}\left(\hat{R}, \hat{\bar{Y}}_2\right)$. Για το λόγο αυτό, έχουν προταθεί στη βιβλιογραφία μια σειρά από ασυμπτωτικά ή προσεγγιστικά αποτελέσματα για τον υπολογισμό του ποσού μεροληψίας. Το ίδιο ισχύει και για τη διασπορά του εκτιμητή $\hat{R}$. Θα δούμε τα κυριότερα από τα αποτελέσματα αυτά.

Για τις αποδείξεις, θα χρειαστεί να ορίσουμε δύο έννοιες, οι οποίες είναι γωστές από την κλασική θεωρία στατιστικής και που υιοθετούνται και στη θεωρία δειγματοληψίας για πεπερασμένους πληθυσμούς. Ο πρώτος ορισμός αφορά τη συμμεταβλητότητα ή συνδιακύμανση δύο μεταβλητών. Η συνδιακύμανση (covariance), στην περίπτωση του πεπερασμένου πληθυσμού, ορίζεται με ανάλογο τρόπο όπως και στην περίπτωση της μεταβλητότητας $S^2$ μίας μεταβλητής. Αναλυτικά, αν $X$ και $Y$ δύο μεταβλητές και $X_i$, $Y_i$οι μετρήσεις των μεταβλητών για ένα πεπερασμένο πληθυσμό μεγέθους $N$, τότε ορίζεται ως πληθυσμιακή συμμεταβλητότητα η ποσότητα:

$$\mathrm{Cov}(X,Y) = S_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$$

Ορίζουμε επίσης τον συντελεστή συσχέτισης (correlation coefficient) $\rho$ για πεπερασμένους πληθυσμούς σε αναλογία με τον γνωστό ορισμό, δηλ.

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{S_{XY}}{\sqrt{S_X^2 S_Y^2}} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{(N-1)S_X S_Y}$$

### Πρόταση 7.2

Το ποσό μεροληψίας του $\hat{R}$ δίνεται προσεγγιστικά από τη σχέση:

$$\mathrm{Bias}\left(\hat{R}\right) \cong \frac{1-f}{n\bar{Y}_2^2}\left(RS_2^2 - \rho S_1 S_2\right) \qquad (7.4)$$

όπου $S_1^2$ και $S_2^2$ είναι οι πληθυσμιακές διακυμάνσεις των μεταβλητών $Y_1$ και $Y_2$ αντίστοιχα, $f = n/N$ το πηλίκο του δείγματος και $\rho$ ο συντελεστής συσχέτισης των δύο μεταβλητών $Y_1$ και $Y_2$.

### Απόδειξη

Η απόδειξη δίνεται στο τέλος του κεφαλαίου.

Το προσεγγιστικό αποτέλεσμα της Πρότασης 7.2 δίνει με τη σειρά του το παρακάτω ασυμπτωτικό αποτέλεσμα για τον εκτιμητή.

**Πόρισμα 7.2**

Ο εκτιμητής $\widehat{R}$ είναι ασυμπτωτικά αμερόληπτος εκτιμητής του $R$.

**Απόδειξη**

Η απόδειξη είναι άμεση από τη σχέση (7.4) εφόσον λάβουμε το όριο στο δεύτερο μέρος της ανισότητας για $n$ να τείνει στο άπειρο ■

Συνεπώς, για μεγάλα μεγέθη δείγματος, μπορούμε να θεωρούμε ότι ο εκτιμητής του λόγου δύο χαρακτηριστικών όπως ορίζεται από την (7.2) είναι αμερόληπτος.

Το ίδιο αποτέλεσμα με αυτό της Πρότασης 7.2 μπορεί να προκύψει εάν γράψουμε:

$$\hat{R} - R = \frac{\bar{X}_1 - \bar{X}_2 R}{\bar{X}_2}$$

και αντικαταστήσουμε στον παρονομαστή τον πληθυσμιακό μέσο $\bar{Y}_2$ αντί του εκτιμητή του $\bar{X}_2$. Για μεγάλα μεγέθη δείγματος η εκτίμηση αυτή είναι συνεπής για τον πληθυσμιακό μέσο κάτω από ένα α.τ. δείγμα. Άρα:

$$\frac{\bar{X}_1 - \bar{X}_2 R}{\bar{X}_2} \approx \frac{\bar{X}_1 - \bar{X}_2 R}{\bar{Y}_2} \tag{7.5}$$

και, λαμβάνοντας στη συνέχεια αναμενόμενες τιμές και στα δύο μέλη της ισότητας, προκύπτει ότι:

$$\mathrm{E}\left(\hat{R} - R\right) \approx \frac{1}{\bar{Y}_2}\mathrm{E}(\bar{X}_1 - \bar{X}_2 R) = \frac{1}{\bar{Y}_2}(\bar{Y}_1 - \bar{Y}_2 R) = 0$$

Συνεπώς, $\mathrm{E}\left(\hat{R}\right) \approx R$ δηλ. ο εκτιμητής $\hat{R}$ είναι ασυμπτωτικά αμερόληπτος. Το αντίστοιχο ασυμπτωτικό αποτέλεσμα για τη διακύμανση του εκτιμητή $\widehat{R}$ δίνεται από την ακόλουθη Πρόταση.

**Πρόταση 7.3**

Η διακύμανση του εκτιμητή του $\hat{R}$ δίνεται ασυμπτωτικά από τη σχέση:

$$\mathrm{Var}\left(\hat{R}\right) = \frac{1 - f}{n\bar{Y}_2^2}\sum_{i=1}^{N}\frac{(Y_{1i} - RY_{2i})^2}{N - 1} \tag{7.6}$$

**Απόδειξη**

Η απόδειξη μπορεί να γίνει είτε με χρήση αναπτύγματος σε σειρά Taylor της κατάλληλης συνάρτησης, δουλεύοντας ανάλογα όπως και στην απόδειξη της Πρότασης 7.2, είτε κάνοντας τη θεώρηση ότι η (7.5) ισχύει. Με τον δεύτερο τρόπο, θα έχουμε:

$$\mathrm{Var}\left(\hat{R}\right) = \mathrm{E}\left(\hat{R} - R\right)^2 = \mathrm{E}\left(\frac{\bar{X}_1 - \bar{X}_2 R}{\bar{Y}_2}\right)^2 = \frac{1}{\bar{Y}_2^2}\mathrm{E}(\bar{X}_1 - \bar{X}_2 R)^2$$

Αν θέσουμε $Z_i = Y_{1i} - Y_{2i}R$ για $i = 1,2,\dots,N$, τότε οι μονάδες $Z_i$ αντιπροσωπεύουν έναν πληθυσμό με πληθυσμιακή μέση τιμή μηδέν (γιατί $\bar{Z} = \bar{Y}_1 - \bar{Y}_2 R = 0$) και η ποσότητα $\bar{X}_1 - \bar{X}_2 R$ είναι ο δειγματικός μέσος ενός α.τ. δείγματος που προέρχεται από τον πληθυσμό των $Z_i$ μετρήσεων. Ως αμερόληπτος εκτιμητής της πληθυσμιακής μέσης τιμής, ο εκτιμητής $\bar{X}_1 - \bar{X}_2 R$ θα έχει αναμενόμενη τιμή μηδέν και συνεπώς:

$$\mathrm{E}(\bar{X}_1 - \bar{X}_2 R)^2 = \mathrm{Var}(\bar{X}_1 - \bar{X}_2 R)$$

Η τελευταία διακύμανση, μέσω της α.τ.δ., είναι $\mathrm{Var}(\bar{X}_1 - \bar{X}_2 R) = \frac{1-f}{n} S_Z^2$, όπου $S_Z^2$ η πληθυσμιακή διασπορά των μετρήσεων $Z_i$, $i = 1,2,\ldots,N$, δηλ.

$$S_Z^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Z_i - \bar{Z})^2 = \frac{1}{N-1}\sum_{i=1}^{N} {Z_i}^2 = \sum_{i=1}^{N}\frac{(Y_{1i} - RY_{2i})^2}{N-1}$$

το οποίο ολοκληρώνει την απόδειξη ■

Το ανωτέρω αποτέλεσμα μπορεί να γραφεί ισοδύναμα με μια σειρά από χρήσιμες μορφές, οι οποίες δίνονται μέσα από τα Πορίσματα που ακολουθούν.

### Πόρισμα 7.3

Η διακύμανση του εκτιμητή του λόγου δύο χαρακτηριστικών, $Y_1$ και $Y_2$, γράφεται ισοδύναμα:

$$\mathrm{Var}(\hat{R}) = \frac{1-f}{n\bar{Y}_2^2}(S_1^2 - 2RS_{12} + R^2 S_2^2) \qquad (7.7)$$

όπου $S_1^2$ και $S_2^2$ οι διακυμάνσεις των μεταβλητών $Y_1$ και $Y_2$ αντίστοιχα και $S_{12}$ η συνδιακύμανση των μεταβλητών.

### Απόδειξη

Η απόδειξη είναι άμεση, αν γράψουμε:

$$\sum_{i=1}^{N}\frac{(Y_{1i} - RY_{2i})^2}{N-1} = \sum_{i=1}^{N}\frac{(Y_{1i} - \bar{Y}_1 - RY_{2i} + R\bar{Y}_2)^2}{N-1}$$

αναπτύξουμε την ταυτότητα στους όρους του αθροίσματος και λάβουμε υπόψη τους ορισμούς των $S_1^2$, $S_2^2$ και $S_{12}$ ■

Με τη βοήθεια του ορισμού του συντελεστή συσχέτισης $\rho$ για πεπερασμένους πληθυσμούς:

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{S_{XY}}{\sqrt{S_X^2 S_Y^2}} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{(N-1)S_X S_Y}$$

προκύπτει η παρακάτω ισοδύναμη μορφή γραφής της διακύμανσης του εκτιμητή του λόγου δύο χαρακτηριστικών.

### Πόρισμα 7.4

Η διακύμανση του εκτιμητή του λόγου δύο χαρακτηριστικών $Y_1$ και $Y_2$, γράφεται ισοδύναμα:

$$\mathrm{Var}(\hat{R}) = \frac{1-f}{n\bar{Y}_2^2}(S_1^2 - 2R\rho S_1 S_2 + R^2 S_2^2)$$

όπου $S_1^2$ και $S_2^2$ οι διακυμάνσεις των μεταβλητών $Y_1$ και $Y_2$ αντίστοιχα, και $\rho$ ο συντελεστής συσχέτισης των δύο μεταβλητών.

**Πόρισμα 7.5**

Ακολουθώντας τα Πορίσματα 7.3 και 7.4 που δίνουν τις εκφράσεις της προσεγγιστικής διακύμανσης για τον εκτιμητή του λόγου δύο χαρακτηριστικών, η προσεγγιστική εκτιμώμενη διασπορά του λόγου των χαρακτηριστικών $Y_1$ και $Y_2$ δίνεται προσεγγιστικά από τη σχέση:

$$\hat{\mathrm{V}}\mathrm{ar}(\hat{R}) = \frac{1-f}{n\hat{\bar{Y}}_2^2}\left(s_1^2 - 2\hat{R}\hat{S}_{12} + \hat{R}^2 s_2^2\right) \tag{7.8}$$

ή:

$$\hat{\mathrm{V}}\mathrm{ar}(\hat{R}) = \frac{1-f}{n\hat{\bar{Y}}_2^2}\left(s_1^2 - 2\hat{R}\hat{\rho}s_1 s_2 + \hat{R}^2 s_2^2\right) \tag{7.9}$$

όπου $s_1^2$ και $s_2^2$ είναι οι δειγματικές διακυμάνσεις των μεταβλητών $Y_1$ και $Y_2$ αντίστοιχα, $\hat{R}$ ο εκτιμητής του λόγου των δύο χαρακτηριστικών και $\hat{S}_{12}$ ο εκτιμητής της συνδιακύμανσης των μεταβλητών από τις μετρήσεις του δείγματος.

Συγκεκριμένα:

$$\hat{S}_{12} = s_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)$$

για το δείγμα $\{(X_{11}, X_{21}), (X_{12}, X_{22}), \ldots, (X_{1n}, X_{2n})\}$.

**Παράδειγμα 7.1** (Από το βιβλίο της S.L. Lohr, (2010): *Sampling: Design and Analysis*)

Τα δεδομένα του Πίνακα 7.1 αποτελούν στοιχεία μετρήσεων στο πλαίσιο της έρευνας Peart (1994), στο νησί Santa Cruz, της πολιτείας Καλιφόρνια, στις ΗΠΑ. Ο σκοπός της έρευνας ήταν η μελέτη της επίδρασης της δραστηριότητας των άγριων χοίρων και της ξηρασίας στη βλάστηση του νησιού. Η μελέτη στηρίχτηκε σε δείγματα 10 δέντρων βελανιδιάς (τα οποία επιλέχτηκαν τυχαία), την καταγραφή των νέων δενδρυλλίων που υπήρχαν κάτω από το κάθε δέντρο την άνοιξη του 1992 και την καταγραφή όσων εξ αυτών είχαν επιβιώσει δύο χρόνια μετά, την άνοιξη του 1994. Στα δενδρύλλια της πρώτης επίσκεψης προστέθηκε κατάλληλη σήμανση ώστε να αναγνωριστούν εύκολα κατά τη δεύτερη επίσκεψη.

| α/α δέντρου | Αριθμός νέων δενδρυλλίων (Μάρτιος 1992) | Αριθμός δενδρυλλίων που επιβίωσαν (Φεβρουάριος 1994) |
|---|---|---|
| 1 | 1 | 0 |
| 2 | 0 | 0 |
| 3 | 8 | 1 |
| 4 | 2 | 2 |
| 5 | 76 | 10 |
| 6 | 60 | 15 |
| 7 | 25 | 3 |
| 8 | 2 | 2 |
| 9 | 1 | 1 |

| α/α δέντρου | Αριθμός νέων δενδρυλλίων (Μάρτιος 1992) | Αριθμός δενδρυλλίων που επιβίωσαν (Φεβρουάριος 1994) |
|---|---|---|
| 10 | 31 | 27 |
| **Σύνολα** | **206** | **61** |

***Πίνακας 7.1*** *Δεδομένα από την εργασία Peart (1994).*

Το ζητούμενο είναι να εκτιμηθεί το ποσοστό των δενδρυλλίων που επιβιώνουν από την ξηρασία και τη δραστηριότητα των άγριων χοίρων στο, νησί και να δοθεί το τυπικό σφάλμα της εκτίμησης. Να κατασκευαστεί επίσης ένα 95% διάστημα εμπιστοσύνης για το ίδιο ποσοστό.

Τα δεδομένα αποτελούν εφαρμογή του προβλήματος εκτίμησης λόγου δύο χαρακτηριστικών, γιατί οι παρατηρήσεις αφορούν μετρήσεις δύο χαρακτηριστικών πάνω στις ίδιες πληθυσμιακές μονάδες. Οι πληθυσμιακές μονάδες για το παράδειγμα είναι τα δέκα δέντρα, το δείγμα είναι α.τ. και το μέγεθος του δείγματος δέκα.

Βάση των θεωρητικών αποτελεσμάτων για τον εκτιμητή του λόγου, υπολογίζουμε την τιμή του εκτιμητή και το εκτιμώμενο τυπικό του σφάλμα, με τη βοήθεια των (7.2) και (7.8) αντίστοιχα. Οι εντολές για την εισαγωγή των δεδομένων στην R και τους υπολογισμούς είναι

```
> x1<-c(0,0,1,2,10,15,3,2,1,27)
> x1
 [1]  0  0  1  2 10 15  3  2  1 27
> x2<-c(1,0,8,2,76,60,25,2,1,31)
> x2
 [1]  1  0  8  2 76 60 25  2  1 31

> sum(x1);sum(x2)
 [1] 61
 [1] 206
> estR<-mean(x1)/mean(x2)
> estR
[1] 0.2961165
> s12<-sum((x1-mean(x1))*(x2-mean(x2)))/9
> s12
[1] 148.0444
> varR<-(1/10/mean(x2)^2)*(var(x1)-2*estR*s12+estR^2*var(x2))
> seR<-sqrt(varR)
> seR
[1] 0.1152622
```

Άρα, βάσει του δείγματος, εκτιμούμε ότι ένα 29.61% των νέων δενδρυλλίων που αναπτύσσονται κάτω από τα δέντρα του νησιού καταφέρνουν να επιβιώσουν και το τυπικό σφάλμα της εκτίμησης είναι 0.115.

Ισοδύναμα, το αποτέλεσμα του τυπικού σφάλματος προκύπτει υπολογίζοντας τον εκτιμητή της έκφρασης (7.6):

```
> varR<-(1/10/mean(x2)^2)*sum((x1-estR*x2)^2)/9
> seR<-sqrt(varR)
> seR
[1] 0.1152622
```

Για το διάστημα εμπιστοσύνης, το δείγμα είναι μεγέθους 10, άρα θα γίνει χρήση των t-εκατοστιαίων σημείων. Θα είναι συνεπώς

```
> c(estR-qt(.975,9)*seR, estR+qt(.975,9)*seR)
[1] 0.03537532 0.55685769
```

Άρα, με βαθμό εμπιστοσύνης 5%, το ποσοστό των νέων δενδρυλλίων που θα επιβιώσουν από τις δραστηριότητες των χοίρων και την ξηρασία του νησιού εκτιμάται ότι παίρνει τιμές στο διάστημα (0.0353, 0.5568).

Δύο παρατηρήσεις πάνω στην ανάλυση του παραδείγματος είναι:

(i) το τυπικό σφάλμα της εκτίμησης, και κατά συνέπεια και το μήκος του διαστήματος εμπιστοσύνης, είναι αρκετά μεγάλο, γιατί το μέγεθος του δείγματος είναι μικρό και

(ii) το πρόβλημα δεν αποτελεί εκτίμηση ποσοστού όπως αυτό έχει παρουσιαστεί στο Κεφάλαιο Κεφάλαιο 2 -, παράγραφο 2.3.5, και η εκτίμηση του τυπικού σφάλματος δεν μπορεί να υπολογιστεί από τον τύπο $\mathrm{se}(p) = \sqrt{\frac{1-f}{n}\frac{NP(1-P)}{N-1}}$ του πορίσματος 2.4. Η ανάλυση αυτή θα ήταν κατάλληλη εάν το α.τ.δ. ήταν μεγέθους 206 (206 ανεξάρτητες επιλογές δενδρυλλίων) και 61 ήταν οι 'επιτυχίες' που καταγράφηκαν μέσα στις 206 'προσπάθειες'. Αντίθετα, στα δεδομένα του Παραδείγματος, το μέγεθος δείγματος είναι 10.

## 7.3. Εκτιμητής λόγου για την πληθυσμιακή μέση τιμή ενός χαρακτηριστικού

Ένας εκτιμητής της πληθυσμιακής μέσης τιμής $\bar{Y}_1$ προκύπτει μέσω του εκτιμητή του λόγου $R$. Πιο συγκεκριμένα, έστω $Y_1$ η μεταβλητή του προβλήματος για την οποία ενδιαφερόμαστε να εκτιμήσουμε την πληθυσμιακή μέση τιμή $\bar{Y}_1$ λαμβάνοντας ένα α.τ.δ. μεγέθους $n$ από τον πληθυσμό. Ο εκτιμητής, βάσει της θεωρίας που αναπτύχθηκε για την α.τ.δ., είναι ο μέσος των μετρήσεων για τη μεταβλητή $Y_1$ στο δείγμα. Εάν για τις ίδιες παρατηρήσεις των μονάδων του δείγματος είναι διαθέσιμες και οι μετρήσεις τους σε μια δεύτερη μεταβλητή, έστω $Y_2$, η οποία σχετίζεται με την $Y_1$, τότε θα μπορούσε να προκύψει ένας εκτιμητής της $\bar{Y}_1$, κάνοντας χρήση της πληροφορίας ότι οι μεταβλητές $Y_1$ και $Y_2$ σχετίζονται και των μετρήσεων του δείγματος που είναι διαθέσιμες για την $Y_2$.

Η ιδέα πάνω στην οποία βασίζεται η εκτίμηση με τη βοήθεια της $Y_2$ μεταβλητής είναι ότι αναμένεται ο λόγος μεταξύ των δειγματικών μέσων των μεταβλητών $Y_1$ και $Y_2$ που έχουν καταμετρηθεί για ένα δείγμα να είναι 'κοντά' στον λόγο των αληθινών μέσων τιμών. Δηλαδή:

$$\frac{\bar{Y}_1}{\bar{Y}_2} \approx \frac{\hat{\bar{Y}}_1}{\hat{\bar{Y}}_2}$$

άρα:

$$\bar{Y}_1 \approx \frac{\hat{\bar{Y}}_1}{\hat{\bar{Y}}_2}\bar{Y}_2$$

Βάσει της τελευταίας σχέσης δίνεται ο ορισμός του εκτιμητή λόγου, ή λογοεκτιμητή, της πληθυσμιακής μέσης τιμής $\bar{Y}_1$.

### Ορισμός

Ο εκτιμητής λόγου της πληθυσμιακής μέσης τιμής $\bar{Y}_1$ ενός χαρακτηριστικού $Y_1$ συμβολίζεται $\bar{Y}_{1,R}$ και ορίζεται ως:

$$\hat{\bar{Y}}_{1,R} = \frac{\hat{\bar{Y}}_1}{\hat{\bar{Y}}_2}\bar{Y}_2$$

όπου $\hat{\bar{Y}}_1$ και $\hat{\bar{Y}}_2$ είναι οι δειγματικοί μέσοι των μεταβλητών $Y_1$ και $Y_2$ βάσει ενός α.τ.δ. μεγέθους $n$ και $\bar{Y}_2$ είναι η πληθυσμιακή μέση τιμή για το χαρακτηριστικό $Y_2$.

Λαμβάνοντας υπόψη το συμβολισμό και τη θεωρία για την εκτίμηση του λόγου δύο χαρακτηριστικών, όπως αναπτύχθηκε στην Παράγραφο 7.2, προκύπτει αρχικά ότι:

$$\hat{\bar{Y}}_{1,R} = \hat{R}\bar{Y}_2 \tag{7.10}$$

Συνεπώς, για την εκτίμηση της μέσης τιμής ενός χαρακτηριστικού με τη βοήθεια ενός βοηθητικού χαρακτηριστικού, αρκεί να εκτιμηθεί ο λόγος των δύο χαρακτηριστικών και ο εκτιμητής να πολλαπλασιαστεί με τον πληθυσμιακό μέσο της δεύτερης μεταβλητής, η οποία για το ζητούμενο πρόβλημα παίζει το ρόλο της βοηθητικής μεταβλητής.

## Παρατήρηση 7.1

Από τον ανωτέρω ορισμό του εκτιμητή λόγου, προκύπτει ότι η εφαρμογή του για την εκτίμηση της μέσης τιμής ενός χαρακτηριστικού είναι εφικτή, εφόσον (i) υπάρχουν διαθέσιμα στοιχεία από την έρευνα πάνω στις ίδιες δειγματοληπτικές μονάδες, για μια δεύτερη μεταβλητή, επιπλέον της κύριας, και (ii) για τη βοηθητική μεταβλητή, γνωρίζουμε την αληθινή μέση τιμή στον πληθυσμό.

## Παρατήρηση 7.2

Εάν ενδιαφερόμαστε για το σύνολο και όχι τη μέση τιμή ενός χαρακτηριστικού $Y_1$, ο εκτιμητής λόγου προκύπτει από την (7.10) και είναι:

$$\hat{Y}_{1T,R} = \hat{R}Y_{2T} \tag{7.11}$$

όπου $Y_{2T}$ είναι το σύνολο των τιμών του πληθυσμού για τη μεταβλητή $Y_2$.

Μαθηματικά, οι ιδιότητες του εκτιμητή $\hat{\bar{Y}}_{1,R}$ προκύπτουν άμεσα από τις ιδιότητες του εκτιμητή $\hat{R}$, αφού η σχέση που συνδέει τους δύο εκτιμητές είναι γραμμική. Η αναλυτική μελέτη και σύγκριση των ιδιοτήτων του $\hat{\bar{Y}}_{1,R}$ με αυτές του απλού δειγματικού μέσου $\hat{\bar{Y}}_1$ θα μας δώσει τα πλεονεκτήματα, εάν υπάρχουν, και τις προϋποθέσεις κάτω από τις οποίες η εκτίμηση $\hat{\bar{Y}}_{1,R}$ είναι πιο ακριβής σε σχέση με την $\hat{\bar{Y}}_1$. Πρακτικά, η εκτίμηση του μέσου ενός χαρακτηριστικού με τη βοήθεια ενός βοηθητικού είναι μια εκτίμηση που δεν επιβαρύνει το κόστος της έρευνας, γιατί το δείγμα είναι ένα και, συνήθως, η καταμέτρηση της δεύτερης μεταβλητής γίνεται παράλληλα με την καταμέτρηση της πρώτης (της κύριας) μεταβλητής. Επίσης, τα πληθυσμιακά στοιχεία που είναι απαραίτητα για τη βοηθητική μεταβλητή, όπως αναφέρθηκε και στην εισαγωγή, είναι πληροφορίες που αρκετά συχνά είναι διαθέσιμες ή είναι εύκολο να βρεθούν για τον πληθυσμό. Για παράδειγμα, εάν η κύρια μεταβλητή είναι η έκταση καλλιεργήσιμης γης μιας χώρας που καλλιεργείται από ένα συγκεκριμένο είδος σιτηρού, η βοηθητική μεταβλητή που μετρά τη συνολική έκταση για κάθε καλλιεργήσιμο κομμάτι γης είναι μια μεταβλητή για την οποία είναι εύκολο να γνωρίζουμε ή να αναζητήσουμε τα στοιχεία από τις επίσημες έρευνες της χώρας.

Για τη διακύμανση του εκτιμητή λόγου της μέσης τιμής ενός χαρακτηριστικού που δίνεται από τη σχέση (7.10), προκύπτει άμεσα λόγω της (7.7) ότι δίνεται από τη σχέση:

$$\mathrm{Var}\left(\hat{\bar{Y}}_{1,R}\right) = \frac{1-f}{n}(S_1^2 - 2RS_{12} + R^2S_2^2)$$

ή:

$$\mathrm{Var}\left(\hat{\bar{Y}}_{1,R}\right) = \frac{1-f}{n}\left(S_1^2 - 2R\rho S_1 S_2 + R^2 S_2^2\right) \tag{7.12}$$

και η εκτιμώμενη διακύμανση του $\hat{\bar{Y}}_{1,R}$, μέσω των (7.8) και (7.9), θα είναι:

$$\hat{\mathrm{Var}}\left(\hat{\bar{Y}}_{1,R}\right) = \frac{1-f}{n}\left(s_1^2 - 2\hat{R}\hat{S}_{12} + \hat{R}^2 s_2^2\right) \tag{7.13}$$

και:

$$\hat{\mathrm{Var}}\left(\hat{\bar{Y}}_{1,R}\right) = \frac{1-f}{n}\left(s_1^2 - 2\hat{R}\hat{\rho} s_1 s_2 + \hat{R}^2 s_2^2\right) \tag{7.14}$$

αντίστοιχα.

Το σκεπτικό πίσω από τη χρήση μιας βοηθητικής μεταβλητής στην εκτίμηση του μέσου ενός χαρακτηριστικού ήταν η βελτίωση στην ακρίβεια του εκτιμητή που λαμβάνεται από την α.τ.δ. χωρίς χρήση βοηθητικής μεταβλητής. Το αποτέλεσμα που ακολουθεί μας δίνει μια ικανή και αναγκαία συνθήκη, έτσι ώστε η εκτίμηση λόγου της μέσης τιμής (ή συνόλου) ενός χαρακτηριστικού να είναι πιο ακριβής σε σχέση με τον απλό δειγματικό μέσο όρο.

### Πρόταση 7.4

Ικανή και αναγκαία συνθήκη, ώστε ο εκτιμητής λόγου $\hat{\bar{Y}}_{1,R}$ για τη μέση τιμή της μεταβλητής $Y_1$ να είναι πιο αποτελεσματικός από τον εκτιμητή $\hat{\bar{Y}}_1$ , είναι:

$$\rho(Y_1, Y_2) > \frac{RS_2}{2S_1} \tag{7.15}$$

### Απόδειξη

Για τη σύγκριση της αποτελεσματικότητας των δύο εκτιμητών, αρκεί να συγκρίνουμε τις διακυμάνσεις τους. Για τον εκτιμητή $\hat{\bar{Y}}_1$, η διακύμανση, μέσω της α.τ.δ., είναι:

$$\mathrm{Var}\left(\hat{\bar{Y}}_1\right) = \frac{1-f}{n} S_1^2$$

ενώ για τον εκτιμητή $\hat{\bar{Y}}_{1,R}$, η προσεγγιστική διακύμανση δίνεται από την (7.12). Αφαιρώντας τις δύο διακυμάνσεις, θα είναι:

$$\mathrm{Var}\left(\hat{\bar{Y}}_1\right) - \mathrm{Var}\left(\hat{\bar{Y}}_{1,R}\right) = \frac{1-f}{n} S_1^2 - \frac{1-f}{n}\left(S_1^2 - 2\rho R S_1 S_2 + R^2 S_2^2\right) = \frac{1-f}{n}\left(2\rho R S_1 S_2 - R^2 S_2^2\right)$$

όπου $\rho = \rho(Y_1, Y_2)$. Κατά συνέπεια, θα ισχύει $\mathrm{Var}\left(\hat{\bar{Y}}_{1,R}\right) < Var\left(\hat{\bar{Y}}_1\right)$, αν και μόνο αν:

$$2\rho R S_1 S_2 - R^2 S_2^2 > 0 \text{ ή } \rho > \frac{RS_2}{2S_1}$$

που ολοκληρώνει την απόδειξη ■

Άρα, ο εκτιμητής λόγου $\hat{\bar{Y}}_{1,R}$ θα είναι πιο αποτελεσματικός από τον $\hat{\bar{Y}}_1$ (και συνεπώς θα ενδείκνυται η χρήση της βοηθητικής μεταβλητής), όταν η συσχέτιση των δύο μεταβλητών είναι μεγάλη και ξεπερνά ένα όριο που

ισούται με το κλάσμα $RS_2/2S_1$. Όσο μεγαλύτερη συνεπώς είναι η συσχέτιση, τόσο πιο πιθανό είναι να ικανοποιείται η συνθήκη, και ο $\hat{\bar{Y}}_{1,R}$ να είναι ακριβέστερος.

Αν συμβολίσουμε με $\mathrm{CV}(Y_1) = S_1/\bar{Y}_1$ και $\mathrm{CV}(Y_2) = S_2/\bar{Y}_2$ τους συντελεστές μεταβλητότητας των $Y_1$ και $Y_2$ αντίστοιχα, και λάβουμε υπόψη ότι $\bar{Y}_1 = R\bar{Y}_2$, τότε μπορούμε να γράψουμε ισοδύναμα την ικανή και αναγκαία συνθήκη της Πρότασης 7.4 ως:

$$\rho > \frac{\mathrm{CV}(Y_2)}{2\mathrm{CV}(Y_1)}$$

## Παρατήρηση 7.3

Λόγω της συνθήκης (7.15), το κριτήριο επιλογής μιας βοηθητικής μεταβλητής είναι η μέγιστη δυνατή γραμμική συσχέτιση με την κύρια μεταβλητή. Για τον λόγο αυτό, αρκετά συχνά στην πράξη θεωρούμε ως βοηθητική μεταβλητή την ίδια την υπό μελέτη μεταβλητή μετρούμενη σε προηγούμενη περίοδο ή έρευνα. Η επιλογή αυτή είναι άριστη εφόσον είναι εφικτή, γιατί εξασφαλίζεται η μεγάλη συσχέτιση μεταξύ των δύο μεταβλητών. Ο μόνος περιορισμός είναι ότι η τρέχουσα δειγματοληπτική έρευνα θα πρέπει να διεξαχθεί στο ίδιο δείγμα του πληθυσμού που έχει διεξαχθεί και η παλιότερη έρευνα.

## Παρατήρηση 7.4

Η εκτίμηση λόγου ενός χαρακτηριστικού, έχει ομοιότητες με την εκτίμηση που προκύπτει από μια δειγματοληψία με άνισες πιθανότητες. Στη δειγματοληψία με άνισες πιθανότητες, ο εκτιμητής ορίζεται ως:

$$\hat{\bar{Y}}_{1HT} = \sum_{i=1}^{n} w_i X_{1i}$$

όπου $w_i = 1/\pi_i$ και $\pi_i$ είναι η πιθανότητα επιλογής της $i$ μονάδας στο δείγμα. Στην απλή τυχαία δειγματοληψία, και συγκεκριμένα για τον απλό δειγματικό μέσο, όλα τα βάρη είναι ίσα μεταξύ τους ($w_i = n/N$). Στον εκτιμητή λόγου, ο εκτιμητής γράφεται:

$$\hat{\bar{Y}}_{1,R} = \frac{1}{\hat{\bar{Y}}_2}\bar{Y}_2\frac{1}{n}\sum_{i=1}^{n} X_{1i} = \sum_{i=1}^{n} w_i' X_{1i}$$

όπου $w_i' = \bar{Y}_2/n\hat{\bar{Y}}_2$. Μπορεί συνεπώς ο εκτιμητής λόγου να θεωρηθεί ως εκτιμητής, κατά τον υπολογισμό του οποίου διαφοροποιούνται τα βάρη ή η συμμετοχή των μετρήσεων του δείγματος. Η ίδια ιδιότητα χαρακτηρίζει και τον εκτιμητή στην περίπτωση της δειγματοληψίας με άνισες πιθανότητες.

Η διαφορά μεταξύ των δύο περιπτώσεων είναι ότι ενώ τα βάρη $w_i$ είναι προκαθορισμένα και υπολογίζονται βάσει του δειγματοληπτικού σχεδίου, τα βάρη $w_i'$ στον εκτιμητή λόγου είναι κάθε φορά εξαρτώμενα από τις τιμές του δείγματος, γιατί ο δειγματικός μέσος της βοηθητικής μεταβλητής $\hat{\bar{Y}}_2$ θα αλλάζει κάθε φορά, ανάλογα με το δείγμα.

## Παράδειγμα 7.2

Τα δεδομένα του Παραδείγματος είναι από το βιβλίο Sampling of Populations, Levy & Lemeshow (1999) σελ. 171. Τα δεδομένα αποτελούν ένα α.τ. δείγμα από τα αρχεία ενός κτηνιάτρου μικρών ζώων. Για την κάθε παρατήρηση, έγινε καταγραφή του αριθμού των επισκέψεων που έκαναν στο ιατρείο του κτηνιάτρου οι ιδιοκτήτες των ζώων για τη διάρκεια ενός έτους καθώς και το συνολικό κόστος των επισκέψεων. Σκοπός της δειγματοληψίας είναι να εκτιμηθεί το μέσο ποσό που κοστίζει στον ιδιοκτήτη ενός σκύλου ή γάτας η ετήσια ιατρική φροντίδα του ζώου. Τα δεδομένα του δείγματος μεγέθους 50 δίνονται στον Πίνακα 7.2.

| α/α ζώου | Τύπος ζώου | Αριθμ. Επισκέψεων ($Y_2$) | Συνολικό κόστος ($Y_1$) σε $ | α/α ζώου | Τύπος ζώου | Αριθμ. Επισκέψεων ($Y_2$) | Συνολικό κόστος ($Y_1$) σε $ |
|---|---|---|---|---|---|---|---|
| 1 | σκυλί | 4 | 45.14 | 26 | σκυλί | 4 | 48.3 |
| 2 | σκυλί | 5 | 50.13 | 27 | σκυλί | 5 | 54.64 |
| 3 | γάτα | 2 | 27.15 | 28 | γάτα | 3 | 21.45 |
| 4 | σκυλί | 3 | 45.8 | 29 | γάτα | 3 | 10.71 |
| 5 | γάτα | 1 | 23.4 | 30 | σκυλί | 4 | 60.57 |
| 6 | γάτα | 2 | 8.24 | 31 | σκυλί | 6 | 53.37 |
| 7 | σκυλί | 6 | 61.22 | 32 | σκυλί | 5 | 40.52 |
| 8 | γάτα | 2 | 29.9 | 33 | σκυλί | 4 | 50.26 |
| 9 | σκυλί | 5 | 56.57 | 34 | γάτα | 2 | 15.23 |
| 10 | σκυλί | 4 | 42.39 | 35 | σκυλί | 4 | 42.02 |
| 11 | γάτα | 2 | 27.24 | 36 | σκυλί | 5 | 32.78 |
| 12 | γάτα | 3 | 22.17 | 37 | γάτα | 2 | 30.21 |
| 13 | σκυλί | 6 | 39.67 | 38 | γάτα | 1 | 27.54 |
| 14 | σκυλί | 4 | 40.52 | 39 | σκυλί | 6 | 52.03 |
| 15 | σκυλί | 4 | 39.48 | 40 | σκυλί | 5 | 54.47 |
| 16 | γάτα | 1 | 7.14 | 41 | σκυλί | 5 | 46.88 |
| 17 | σκυλί | 4 | 61.82 | 42 | γάτα | 2 | 23.77 |
| 18 | γάτα | 2 | 39.88 | 43 | σκυλί | 3 | 52.48 |
| 19 | γάτα | 2 | 16.89 | 44 | σκυλί | 2 | 60.49 |
| 20 | σκυλί | 3 | 55.31 | 45 | σκυλί | 2 | 53.70 |
| 21 | σκυλί | 2 | 63.19 | 46 | σκυλί | 2 | 46.39 |
| 22 | σκυλί | 2 | 45.11 | 47 | σκυλί | 2 | 53.24 |
| 23 | σκυλί | 3 | 66.20 | 48 | γάτα | 1 | 14.18 |
| 24 | γάτα | 3 | 17.16 | 49 | σκυλί | 3 | 41.52 |
| 25 | γάτα | 3 | 28.55 | 50 | σκυλί | 2 | 39.26 |

***Πίνακας 7.2*** *Δεδομένα δείγματος 50 πελατών κτηνίατρου, για αριθμό επισκέψεων και ετήσιο κόστος.*

Ο κτηνίατρος γνωρίζει ότι συνολικά είχε 4100 επισκέψεις στο ιατρείο του τη χρονιά που πέρασε, εκ των οποίων 2515 αντιστοιχούν σε επισκέψεις πελατών που είναι ιδιοκτήτες σκύλων, και 1585 σε πελάτες με γάτες. Επίσης, γνωρίζει ότι συνολικά 'βλέπει' 850 σκυλιά και 450 γάτες. Να υπολογιστεί ο εκτιμητής του μέσου ετήσιου κόστους για ιατρική φροντίδα των ζώων (α) από το α.τ. δείγμα χωρίς χρήση της μεταβλητής $Y_2$ και (β) από το α.τ.δ. των μεταβλητών $Y_1$ και $Y_2$.

Εισάγουμε τα δεδομένα στην R, και έστω ότι visits και cost είναι τα ονόματα των στηλών με τον αριθμό των επισκέψεων και το κόστος των επισκέψεων αντίστοιχα. Για τον υπολογισμό της μέσης τιμής στην (α) περίπτωση, είναι:

```
> sum(cost)/50
[1] 39.7256
```

δηλ., κατά μέσον όρο, ο κάθε ιδιοκτήτης ενός σκύλου ή γάτας ξοδεύει 39.73 δολλάρια το χρόνο για την ιατρική του φροντίδα.

Για το τυπικό σφάλμα της εκτίμησης, λαμβάνοντας υπόψη ότι ο συνολικός αριθμός των ζώων είναι 1300 και κάνοντας χρήση των ιδιοτήτων της α.τ.δ., θα είναι:

```
> sqrt((1-50/1300)*(1/50)*var(cost))
[1] 2.221718
```

Άρα, με βάση τη μεταβλητή $Y_1$ και μόνο, εκτιμάται ότι το μέσο ετήσιο κόστος ιατρικής φροντίδας των ζώων είναι 39.73$, με τυπικό σφάλμα εκτίμησης 2.22$.

Για την περίπτωση (β), εκτός από τη μεταβλητή που δίνει το κόστος, λαμβάνουμε υπόψη και τη μεταβλητή του αριθμού επισκέψεων. Η νέα εκτίμηση της ζητούμενης μέσης τιμής, κάνοντας εφαρμογή του εκτιμητή λόγου, θα είναι:

```
> PopMeanVisits<-4100/1300
> PopMeanVisits
[1] 3.153846
> estR<-(sum(cost)/sum(visits))*PopMeanVisits
> estR
[1] 38.90945
```

Άρα, η τιμή του εκτιμητή είναι 38.91$ (αρκετά κοντά με την εκτίμηση στο (α)), και μια εκτίμηση του τυπικού σφάλματος από την (7.8) είναι:

```
> s12=(1/49)*sum((cost-mean(cost))*(visits-mean(visits)))
> R<-(sum(cost)/sum(visits))
> R
[1] 12.33714
> sqrt((1-50/1300)*(1/50)*(var(cost)-2*R*s12+R^2*var(visits)))
[1] 2.375109
```

Άρα, για τα δεδομένα του παραδείγματος, η εκτίμηση της ζητούμενης μέσης τιμής έχει μεγαλύτερη ακρίβεια εάν η εκτίμηση γίνει χωρίς χρήση της βοηθητικής μεταβλητής. Αυτό συμβαίνει, γιατί ενώ οι δύο μεταβλητές συνδέονται, δεν συνδέονται τόσο ισχυρά ώστε να ικανοποιείται η συνθήκη (7.15). Πράγματι, εκτιμώντας τις ποσότητες που εμφανίζονται στην (7.15) από τις τιμές του δείγματος προκύπτει:

```
> rho<-s12/sqrt(var(cost)*var(visits))
> rho
[1] 0.4929873
```

Άρα, μια εκτίμηση του συντελεστή συσχέτισης των δύο μεταβλητών, με βάση τα δεδομένα του δείγματος, είναι 0.493. Το δεξί μέρος της ανισότητας (7.15) είναι:

```
> R*sqrt(var(visits))/(2*sqrt(var(cost)))
[1] 0.5570925
> rho>R*sqrt(var(visits))/(2*sqrt(var(cost)))
```

```
[1] FALSE
```

Συνεπώς, η συνθήκη (7.15) για τα δεδομένα της άσκησης δεν ικανοποιείται, γιατί ο συντελεστής συσχέτισης δεν ξεπερνά την τιμή 0.557. Άρα, η χρήση της βοηθητικής μεταβλητής στο συγκεκριμένο πρόβλημα δεν βελτιώνει την εκτίμηση ως προς την ακρίβεια.

Διαπιστώνοντας ότι ο εκτιμητής λόγου δεν βελτιώνει πάντα την εκτίμηση, και ότι κριτήριο για το θετικό ή όχι αποτέλεσμα είναι ο βαθμός συσχέτισης των δύο μεταβλητών, είναι άμεσο το ερώτημα τι θα συνέβαινε εάν η συσχέτιση δεν ήταν αρκετά ισχυρή ώστε να ικανοποιεί τη συνθήκη σε όλο το εύρος του πληθυσμού, αλλά ήταν ισχυρή σε υποσύνολά του. Θα μπορούσε, για παράδειγμα, να βελτιωθεί η εκτίμηση, εάν εφαρμόζαμε τον εκτιμητή λόγου σε ένα στρωματοποιημένο πληθυσμό, αντί για το ενιαίο σύνολο; Το ερώτημα αυτό έχει νόημα, επειδή πράγματι οι υπό μελέτη μεταβλητές μπορεί να έχουν πιο ισχυρή συσχέτιση σε ορισμένα στρώματα του πληθυσμού και λιγότερο ισχυρή σε άλλα. Λόγω επίσης της μεγάλης χρήσης της στρωματοποιημένης δειγματοληψίας, είναι αναγκαίο να δοθούν οι κατευθύνσεις, για τον τρόπο με τον οποίο μπορούν να συνδυαστούν οι δύο μεθοδολογίες, της εκτίμησης λόγου και της στρωματοποιημένης.

Η επόμενη παράγραφος εξετάζει την εφαρμογή του εκτιμητή λόγου για την εκτίμηση της μέσης τιμής ή το σύνολο ενός χαρακτηριστικού, στην περίπτωση ενός πληθυσμού χωρισμένου σε στρώματα. Η χρήση του εκτιμητή λόγου, σε συνδυασμό με τη στρωματοποιημένη και την εκ των υστέρων στρωματοποιημένη, είναι πάρα πολύ συχνή.

## 7.4. Εκτιμητής λόγου και στρωματοποιημένη δειγματοληψία

Η εφαρμογή του εκτιμητή λόγου σε ένα στρωματοποιημένο δειγματοληπτικό σχέδιο μπορεί να γίνει με δύο διαφορετικούς τρόπους. Στο πρώτο, υπολογίζουμε πρώτα τον εκτιμητή λόγου ανά στρώμα και στη συνέχεια τον τελικό εκτιμητή με βάση τη στρωματοποιημένη. Στον δεύτερο, εφαρμόζουμε πρώτα τη στρωματοποιημένη για την εκτίμηση των δύο μέσων τιμών των μεταβλητών και στη συνέχεια τον ορισμό του εκτιμητή λόγου.

**Α. Συνδυασμένος ή από κοινού εκτιμητής λόγου (combined ratio estimate).**

Σύμφωνα με τη μεθοδολογία εκτίμησης του συνδυασμένου εκτιμητή λόγου, εφαρμόζεται αρχικά η εκτίμηση κάτω από το στρωματοποιημένο σχέδιο για την εκτίμηση των μέσων τιμών του κύριου και του βοηθητικού χαρακτηριστικού, και στη συνέχεια εφαρμόζεται ο ορισμός του εκτιμητή λόγου. Αν με $\hat{\bar{Y}}_{1,Rc}$ συμβολίσουμε τον εκτιμητή αυτόν, η σχέση από την οποία ορίζεται αναλυτικά είναι η:

$$\hat{\bar{Y}}_{1,Rc} = \frac{\hat{\bar{Y}}_{1,st}}{\hat{\bar{Y}}_{2,st}} \bar{Y}_2$$

όπου $\hat{\bar{Y}}_{1,st}$ και $\hat{\bar{Y}}_{2,st}$ είναι οι εκτιμητές των $\bar{Y}_1$, $\bar{Y}_2$ που προκύπτουν από τη στρωματοποιημένη δειγματοληψία λαμβάνοντας υπόψη τα πληθυσμιακά βάρη των στρωμάτων, π.χ.

$$\hat{\bar{Y}}_{1,st} = \sum_h W_h \, \hat{\bar{Y}}_{1,h}$$

και $\hat{\bar{Y}}_{1,h} = (1/n_h) \sum_j X_{1j}$ ο δειγματικός μέσος για κάθε στρώμα $h$, ενώ $W_h = N_h/N$ είναι το πληθυσμιακό βάρος του στρώματος $h$.

Από τον ορισμό του εκτιμητή, παρατηρούμε ότι προκειμένου να υπολογίσουμε τον $\hat{\bar{Y}}_{1,Rc}$ χρειάζεται να γνωρίζουμε τον πληθυσμιακού μέσο της βοηθητικής μεταβλητής για τον ενιαίο πληθυσμό, και όχι ανά στρώματα. Εφαρμόζοντας τα αποτελέσματα που ισχύουν για την εκτίμηση και τη διακύμανση του εκτιμητή στην περίπτωση ενός στρωματοποιημένου δειγματοληπτικού σχεδίου (βλ. παρ. 3.2) και τον ορισμό του

εκτιμητή λόγου, υπολογίζεται η διακύμανση του εκτιμητή $\hat{\bar{Y}}_{1,Rc}$. Η Πρόταση που ακολουθεί δίνει το σχετικό αποτέλεσμα.

### Πρόταση 7.5

Η διακύμανση του από κοινού εκτιμητή λόγου $\hat{\bar{Y}}_{1,Rc}$ δίνεται προσεγγιστικά από τη σχέση:

$$\mathrm{Var}\left(\hat{\bar{Y}}_{1,Rc}\right) \cong \sum_{h=1}^{L} \frac{1-f_h}{n_h} W_h^2 \left(S_{1h}^2 - 2RS_{12,h} + R^2 S_{2h}^2\right)$$

Όπου, κατά τον συνήθη συμβολισμό, ο δείκτης $h$ αναφέρεται στην ποσότητα υπολογισμένη για κάθε στρώμα $h$ του πληθυσμού ($h = 1,2,\dots,L$).

### Απόδειξη

Η απόδειξη της Πρότασης ακολουθεί τον τρόπο απόδειξης της Πρότασης 7.3. για τον προσεγγιστικό υπολογισμό της διακύμανσης του λόγου $R$ στην απλή τυχαία. Συγκεκριμένα,

$$\mathrm{Var}\left(\hat{\bar{Y}}_{1,Rc}\right) = \mathrm{Var}\left(\frac{\hat{\bar{Y}}_{1,st}}{\hat{\bar{Y}}_{2,st}}\bar{Y}_2\right) = \bar{Y}_2^2 E\left(\hat{R}_{st} - R\right)^2$$

όπου $\hat{R}_{st}$ είναι ο λόγος των δύο χαρακτηριστικών όπως αυτός εκτιμάται εάν εφαρμόσουμε χωριστά στον αριθμητή και στον παρονομαστή την εκτίμηση της στρωματοποιημένης για τα χαρακτηριστικά $Y_1$ και $Y_2$ αντίστοιχα. Στη συνέχεια, δουλεύοντας ανάλογα με την απόδειξη της Πρότασης 7.3, αναπτύσσουμε τους λόγους $\hat{R}_{st}$ και $R$ και, μετά από πράξεις στα κλάσματα, κάνουμε την προσέγγιση αντικαθιστώντας τον πληθυσμιακό μέσο αντί για τον δειγματικό στον παρονομαστή. Αναλυτικά, είναι:

$$\bar{Y}_2^2 \mathrm{E}\left(\hat{R}_{st} - R\right)^2 = \bar{Y}_2^2 \mathrm{E}\left(\frac{\bar{X}_{1,st} - \bar{X}_{2,st}R}{\bar{X}_{2,st}}\right)^2 \cong \bar{Y}_2^2 \mathrm{E}\left(\frac{\bar{X}_{1,st} - \bar{X}_{2,st}R}{\bar{Y}_{2,st}}\right)^2 = \mathrm{E}\left(\hat{\bar{Y}}_{1,st} - \hat{\bar{Y}}_{2,st}R\right)^2$$

Άρα, για τη ζητούμενη διασπορά, το προσεγγιστικό ενδιάμεσο αποτέλεσμα είναι:

$$\mathrm{Var}\left(\hat{\bar{Y}}_{1,Rc}\right) \cong \mathrm{E}\left(\hat{\bar{Y}}_{1,st} - \hat{\bar{Y}}_{2,st}R\right)^2$$

Στη συνέχεια, όπως και στην περίπτωση του ενιαίου πληθυσμού στην Πρόταση 7.3, ορίζουμε τις παρατηρήσεις των διαφορών $Z_{hi} = Y_{1hi} - Y_{2hi}R$ για $i = 1,2,\dots,N_h$ και $h = 1,2,\dots,L$. Για τον πληθυσμό των μονάδων $Z_{hi}$, η μέση τιμή είναι μηδέν (γιατί $\bar{Z} = \bar{Y}_1 - \bar{Y}_2 R = 0$) και η ποσότητα $\hat{\bar{Z}}_{st} = \hat{\bar{Y}}_{1,st} - \hat{\bar{Y}}_{2,st}R$ είναι ο δειγματικός σταθμισμένος μέσος, τον οποίο υιοθετούμε ως εκτιμητή της μέσης τιμής του πληθυσμού στη στρωματοποιημένη. Λόγω της αμεροληψίας του εκτιμητή, θα είναι $\mathrm{E}\left(\hat{\bar{Z}}_{st}\right) = 0$ και κατά συνέπεια:

$$\mathrm{E}\left(\hat{\bar{Y}}_{1,st} - \hat{\bar{Y}}_{2,st}R\right)^2 = \mathrm{Var}\left(\hat{\bar{Y}}_{1,st} - \hat{\bar{Y}}_{2,st}R\right) = \mathrm{Var}\left(\hat{\bar{Z}}_{st}\right)$$

Τέλος, η διασπορά του εκτιμητή $\hat{\bar{Z}}_{st}$ υπολογίζεται μέσω της στρωματοποιημένου σχεδίου και είναι:

$$\mathrm{Var}\left(\hat{\bar{Y}}_{1,Rc}\right) \cong \mathrm{Var}\left(\hat{\bar{Z}}_{st}\right) = \sum_{h=1}^{L} W_h^2 \frac{1-f_h}{n_h} S_{Zh}^2$$

όπου $S_{Zh}^2$ είναι η πληθυσμιακή διασπορά των μετρήσεων $Z$ για το κάθε στρώμα $h$. Κάνοντας χρήση του ορισμού της διασποράς και αναπτύσσοντας τα $Z$ ως προς τις αρχικές μεταβλητές $Y_1$ και $Y_2$, θα είναι:

$$S_{Zh}^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(Z_{hi} - \bar{Z}_h)^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(Y_{1hi} - Y_{2hi}R\ - \bar{Y}_{1h} + R\bar{Y}_{2h})^2$$

$$= \frac{1}{N_h - 1}\sum_{i=1}^{N_h}[(Y_{1hi} - \bar{Y}_{1h}) + R(Y_{2hi} - R\bar{Y}_{2h})]^2$$

απ' όπου, εφαρμόζοντας την ταυτότητα για το τετράγωνο διωνύμου σε κάθε όρο του αθροίσματος, προκύπτει το ζητούμενο ■

Είναι σημαντικό να παρατηρήσουμε ότι στην έκφραση της διακύμανσης του από κοινού εκτιμητή $\hat{\bar{Y}}_{1,Rc}$, ο λόγος $R$ των δύο χαρακτηριστικών δεν εμφανίζεται με δείκτη $h$, δηλαδή δεν ορίζεται ανεξάρτητα ανά στρώμα, αλλά ενιαία για όλο τον πληθυσμό. Συνολικά, τα δύο βασικά χαρακτηριστικά του συνδυασμένου εκτιμητή λόγου $\hat{\bar{Y}}_{1,Rc}$ είναι:

(i) κάνει χρήση του συνολικού πληθυσμιακού μέσου για τη βοηθητική μεταβλητή $\bar{Y}_2$ και

(ii) οι ιδιότητές του εξαρτώνται από τον συνολικό, δηλ. για ολόκληρο τον πληθυσμό, λόγο των δύο χαρακτηριστικών $R$.

**Β. Ανεξάρτητος εκτιμητής λόγου (separate ratio estimate).**

Για τον υπολογισμό του ανεξάρτητου εκτιμητή λόγου του μέσου ενός χαρακτηριστικού σε ένα στρωματοποιημένο δείγμα, αρχικά υπολογίζεται ο εκτιμητής λόγου της ζητούμενης ποσότητας σε κάθε στρώμα χωριστά, και στη συνέχεια γίνεται συνδυασμός όλων των $L$ εκτιμήσεων που προκύπτουν, κάνοντας χρήση του σταθμισμένου εκτιμητή που ισχύει για την περίπτωση της στρωματοποιημένης. Αν συμβολίσουμε με $\hat{\bar{Y}}_{1,Rs}$ τον εκτιμητή, θα είναι:

$$\hat{\bar{Y}}_{1,Rs} = \sum_{h=1}^{L} W_h \frac{\hat{\bar{Y}}_{1h}}{\hat{\bar{Y}}_{2h}} \bar{Y}_{2h} = \sum_{h=1}^{L} W_h \hat{R}_h\, \bar{Y}_{2h}$$

όπου $\hat{\bar{Y}}_{1,Rh}$ είναι ο εκτιμητής λόγου (7.10) της μέσης τιμής για το $Y_1$ χαρακτηριστικό στο στρώμα $h$, και $W_h$ το βάρος του ίδιου στρώματος. Αναπτύσσοντας τον εκτιμητή $\hat{\bar{Y}}_{1,Rh}$ σύμφωνα με τον ορισμό του, προκύπτει μια ισοδύναμη έκφραση για τον ανεξάρτητο εκτιμητή λόγου $\hat{\bar{Y}}_{1,Rs}$. Θα είναι:

$$\hat{\bar{Y}}_{1,Rs} = \sum_{h=1}^{L} W_h \frac{\hat{\bar{Y}}_{1h}}{\hat{\bar{Y}}_{2h}} \bar{Y}_{2h} = \sum_{h=1}^{L} W_h \hat{R}_h\, \bar{Y}_{2h}$$

όπου $\hat{R}_h$ είναι ο εκτιμητής λόγου της μέσης τιμής του $Y_1$ χαρακτηριστικού για το στρώμα $h$ και $\bar{Y}_{2h}$ η πληθυσμιακή μέση τιμή της βοηθητικής μεταβλητής του ίδιου στρώματος.

Από τον ορισμό του εκτιμητή $\hat{\bar{Y}}_{1,Rs}$, συμπεραίνουμε ότι για τον τελικό υπολογισμό του εκτιμητή, λαμβάνεται υπόψη (i) η πληθυσμιακή μέση τιμή $\bar{Y}_{2h}$ του βοηθητικού χαρακτηριστικού για κάθε στρώμα χωριστά και (ii) η εκτίμηση του λόγου των δύο χαρακτηριστικών $\hat{R}_h$, επίσης για το κάθε στρώμα χωριστά. Τα δύο αυτά σημεία διαφοροποιούν τον ανεξάρτητο εκτιμητή λόγου από τον συνδυασμένο. Το πρώτο σημείο αποτελεί έναν πρακτικό περιορισμό. Σε ορισμένες έρευνες, ενδέχεται η πληροφορία της μέσης τιμής του βοηθητικού χαρακτηριστικού να είναι διαθέσιμη για όλον τον πληθυσμό, αλλά όχι και για κάθε στρώμα χωριστά. Το δεύτερο είναι ένα σημείο κατά το οποίο πλεονεκτεί ο ανεξάρτητος εκτιμητής λόγου σε σχέση με τον από κοινού, γιατί μπορεί να συμπεριλάβει περιπτώσεις πληθυσμών που πραγματικά ο λόγος των δύο χαρακτηριστικών διαφέρει από στρώμα σε στρώμα. Για τους πληθυσμούς αυτούς, ο ανεξάρτητος εκτιμητής

$\hat{\bar{Y}}_{1,Rs}$ θα είναι πιο ακριβής. Το αποτέλεσμα θα γίνει περισσότερο αντιληπτό μετά την Πρόταση που ακολουθεί, η οποία δίνει τη διακύμανση του εκτιμητή $\hat{\bar{Y}}_{1,Rs}$.

### Πρόταση 7.6

Η διακύμανση του ανεξάρτητου εκτιμητή λόγου $\hat{\bar{Y}}_{1,Rs}$ δίνεται προσεγγιστικά από τη σχέση:

$$\mathrm{Var}\left(\hat{\bar{Y}}_{1,Rs}\right) \cong \sum_{h=1}^{L} \frac{1-f_h}{n_h} W_h^2 \left(S_{1h}^2 - 2R_h S_{12,h} + R_h^2 S_{2h}^2\right)$$

### Απόδειξη

Από τον ορισμό του $\hat{\bar{Y}}_{1,Rs}$ και λόγω της ανεξάρτητης δειγματοληψίας ανά στρώμα σύμφωνα με το στρωματοποιημένο σχέδιο, θα έχουμε:

$$\mathrm{Var}\left(\hat{\bar{Y}}_{1,Rs}\right) = \mathrm{Var}\left(\sum_{h=1}^{L} W_h \hat{R}_h \bar{Y}_{2h}\right) = \sum_{h=1}^{L} W_h^2 \bar{Y}_h^2 \,\mathrm{Var}\left(\hat{R}_h\right)$$

Στη συνέχεια, εφαρμόζοντας την έκφραση (7.7) για την προσεγγιστική τιμή της διακύμανσης $\mathrm{Var}\left(\hat{R}_h\right)$ σε κάθε στρώμα $h$, προκύπτει άμεσα το ζητούμενο:

$$\mathrm{Var}\left(\hat{\bar{Y}}_{1,Rs}\right) = \sum_{h=1}^{L} W_h^2 \frac{1-f_h}{n_h} \left(S_{1h}^2 - 2R_h S_{12,h} + R_h^2 S_{2h}^2\right)$$

■

Συγκρίνοντας τους εκτιμητές $\hat{\bar{Y}}_{1,Rc}$ και $\hat{\bar{Y}}_{1,Rs}$ ως προς την αμεροληψία, ο εκτιμητής $\hat{\bar{Y}}_{1,Rs}$ έχει μεγαλύτερο κίνδυνο να έχει μεγάλο ποσό μεροληψίας. Αυτό προκύπτει επειδή το άνω όριο για το ποσό μεροληψίας του εκτιμητή του λόγου που δώσαμε στο Πόρισμα 7.1 ισχύει για τον κάθε επιμέρους εκτιμητή $\hat{R}_h$ ανά στρώμα. Δηλαδή:

$$\frac{\left|\mathrm{Bias}\left(\hat{R}_h\right)\right|}{\sqrt{\mathrm{Var}\left(\hat{R}_h\right)}} \le \mathrm{CV}\left(\hat{\bar{Y}}_{2h}\right)$$

$\forall\, h\ (h = 1,2,\ldots,L)$ και ο συνολικός εκτιμητής $\hat{\bar{Y}}_{1,Rs}$ θα έχει ως άνω όριο μεροληψίας $L$ φορές (χοντρικά) το ποσό μεροληψίας του ενός εκτιμητή ανά στρώμα.

Ως προς την ακρίβεια, ο ανεξάρτητος εκτιμητής λόγου $\hat{\bar{Y}}_{1,Rs}$ είναι καλύτερος από τον από κοινού εκτιμητή $\hat{\bar{Y}}_{1,Rc}$, εάν οι λόγοι $R_h$ διαφέρουν ανά στρώμα. Πράγματι:

$$\mathrm{Var}\left(\hat{\bar{Y}}_{1,Rc}\right) - \mathrm{Var}\left(\hat{\bar{Y}}_{1,Rs}\right) = \sum_{h=1}^{L} W_h^2 \frac{1-f_h}{n_h} \left(S_{2h}^2 (R^2 - R_h^2) - 2S_{12,h}(R - R_h)\right)$$

άρα για $R = R_h$, οι διακυμάνσεις είναι ίσες, ενώ, στη γενική περίπτωση που $R \neq R_h$, και για μεγάλα μεγέθη δείγματος σε κάθε στρώμα, η ποσότητα στην παρένθεση των όρων του αθροίσματος είναι θετική (για μεγαλύτερη ανάπτυξη, βλ. Cochran (1977), παράγρ. 6.12).

## Παράδειγμα 7.3

Για τα δεδομένα του Παραδείγματος 7.2, έστω ότι θεωρούμε την εκ των υστέρων στρωματοποίηση για το δείγμα με κριτήριο διαχωρισμού στρωμάτων τον τύπο του κατοικίδιου ζώου (σκύλος / γάτα). Θα εξετάσουμε πώς συγκρίνεται (α) ο εκτιμητής του μέσου ετήσιου κόστους για την ιατρική φροντίδα των ζώων χωρίς τη χρήση βοηθητικής μεταβλητής με (β) τον εκτιμητή λόγου με τη βοήθεια της μεταβλητής που μετρά τον αριθμό των επισκέψεων στον κτηνίατρο.

(α) Για την πρώτη εκτίμηση, ο εκτιμητής θα είναι ο σταθμισμένος μέσος σύμφωνα με τη στρωματοποιημένη. Τα βάρη για τον πληθυσμό υπολογίζονται από την πληροφορία ότι ο κτηνίατρος 'βλέπει' συνολικά 850 σκυλιά και 450 γάτες. Άρα, τα βάρη είναι

```
> weights<-c(850, 450)/sum(c(850, 450))
> weights
[1] 0.6538462 0.3461538
```

Στη συνέχεια, για τον υπολογισμό του εκτιμητή του μέσου κόστους ανά στρώμα, πρώτα εισάγουμε ένα νέο διάνυσμα με στοιχεία 1 και 2, που δηλώνει το στρώμα στο οποίο ανήκει η κάθε παρατήρηση του δείγματος. Ο αριθμός 1 αντιστοιχεί στο στρώμα με τύπο ζώου «σκύλος» και ο αριθμός 2 στο στρώμα «γάτα».

```
>strata<-c(1,1,2,1,2,2,1,2,1,1,2,2,1,1,1,2,1,2,2,1,1,1,1,2,2,1,1,2,
2,1,1,1,1,2,1,1,2,2,1,1,1,2,1,1,1,1,1,2,1,1)
> data<-cbind(strata, visits, cost)
> data
      strata visits  cost
 [1,]      1      4 45.14
 [2,]      1      5 50.13
 [3,]      2      2 27.15
.....................................
[48,]      2      1 14.18
[49,]      1      3 41.52
[50,]      1      2 39.26
```

Υπολογίζουμε τους μέσους ανά στρώμα και τον τελικό εκτιμητή με βάση τη στρωματοποίηση:

```
> strata1<-data[data[,'strata']==1,]
> strata2<-data[data[,'strata']==2,]
> means<-c(mean(strata1[,'cost']), mean(strata2[,'cost']))
> means
[1] 49.85844 21.71167
> sum(weights*means)
[1] 40.11532
```

Άρα, ο εκτιμητής για το μέσο ετήσιο κόστος ιατρικής φροντίδας είναι 40.11$ αντί για 39.73$ που είχαμε βρει μέσω της α.τ.δ. Οι δύο μέσες τιμές του κόστους ανά στρώμα είναι 49.86$ και 21.71$ για τα στρώματα 1 και 2 αντίστοιχα. Παρατηρούμε ότι το κόστος ιατρικής φροντίδας των κατοικίδιων σκύλων είναι πιο μεγάλο από το αντίστοιχο για τις γάτες.

Για τον υπολογισμό του τυπικού σφάλματος, χρησιμοποιούμε τον τύπο (4.2), που δίνει το σχετικό αποτέλεσμα για την εκ των υστέρων στρωματοποιημένη.

```
> var_st<-(1-50/4300)*(1/50)*sum(weights*c(var(strata1[,'cost']),
var(strata2[,'cost'])))+(1/50^2)*sum((1-
weights)*c(var(strata1[,'cost']), var(strata2[,'cost'])))
```

```
> var_st
[1] 1.449514
> se_st<-sqrt(var_st)
> se_st
[1] 1.203958
```

Άρα, κάνοντας χρήση, έστω και εκ των υστέρων, των στρωμάτων της δειγματοληψίας, το τυπικό σφάλμα της εκτίμησης βελτιώθηκε αρκετά. Η εκτίμησή του μέσω του δείγματος είναι 1.20$, ενώ το αντίστοιχο τυπικό σφάλμα κάτω από την α.τ.δ. είχε βρεθεί 2.22$, δηλ. σχεδόν διπλάσιο. Επίσης, αξίζει να σημειωθεί ότι η τιμή της διακύμανσης 1.45$ που βρέθηκε από την (4.2) θα τροποποιηθεί ελάχιστα, θα γίνει συγκεκριμένα 1.42$, εάν δεν συμπεριλάβουμε τον δεύτερο όρο στη σχέση (4.2) που αποτελεί την «ποινή» για την εκ των υστέρων, και όχι την κανονική, στρωματοποίηση.

(β) Υπολογίζουμε τον ίδιο εκτιμητή, κάνοντας χρήση της βοηθητικής μεταβλητής του αριθμού των επισκέψεων στον γιατρό, και, ταυτόχρονα, του χωρισμού των στρωμάτων. Για να επιλέξουμε μεταξύ των δύο τύπων εκτιμητή λόγου, του από κοινού και του ανεξάρτητου, εκτιμούμε το λόγο των δύο μεταβλητών για τα δύο στρώματα χωριστά.

```
> R1<-(sum(strata1[,'cost'])/sum(strata1[,'visits']))
> R1
[1] 12.86669
>  R2<-(sum(strata2[,'cost'])/sum(strata2[,'visits']))
> R2
[1] 10.56243
```

Οι δύο λόγοι διαφέρουν ανά στρώμα και για το λόγο αυτό θα υιοθετήσουμε τον ανεξάρτητο εκτιμητή λόγου του μέσου κόστους. Οι εντολές για τον υπολογισμό του είναι

```
> PopMeanVisits1<-2515/850
> PopMeanVisits2<-1585/450
> sum(weights*c(R1,R2)*c(PopMeanVisits1,PopMeanVisits2))
[1] 37.77015
```

άρα ο εκτιμητής λόγου του μέσου κόστους για το στρωματοποιημένο δείγμα είναι 37.77$ και η διακύμανσή του:

```
> n1<-dim(data[data[,'strata']==1,])[[1]]
> n2<-dim(data[data[,'strata']==2,])[[1]]
> f<-c(n1/850, n2/450)
>a1<-var(strata1[,'cost'])-2*R1*(1/(n1-1))*sum((strata1[,'cost']-
mean(strata1[,'cost']))*(strata1[,'visits']-
mean(strata1[,'visits'])))
>a2<-var(strata2[,'cost'])-2*R1*(1/(n1-1))*sum((strata2[,'cost']-
mean(strata2[,'cost']))*(strata2[,'visits']-
mean(strata2[,'visits'])))
> sum(weights^2*((1-f)/c(n1,n2))*c(a1,a2))
[1] 1.734232
```

Σύμφωνα με το παραπάνω αποτέλεσμα, ο εκτιμητής λόγου στη στρωματοποιημένη είναι πιο ακριβής από τον εκτιμητή λόγου χωρίς στρωματοποίηση, αλλά παραμένει λιγότερο ακριβής από τον στρωματοποιημένο εκτιμητή χωρίς τη χρήση βοηθητικής μεταβλητής.

## 7.5. Εκτιμητής παλινδρόμησης

Για τον εκτιμητή παλινδρόμησης, η κεντρική ιδέα παραμένει η ίδια όπως και στον εκτιμητή λόγου. Γίνεται, δηλαδή, χρήση μιας βοηθητικής μεταβλητής, η οποία συνδέεται με την κύρια μεταβλητή της έρευνας, επιδιώκοντας η επιπλέον πληροφορία να προσφέρει βελτίωση στην ακρίβεια του εκτιμητή που θα προκύψει.

Ο εκτιμητής παλινδρόμησης εφαρμόζεται όταν υπάρχει γραμμική σχέση μεταξύ των μεταβλητών (κύριας και βοηθητικής), χωρίς αναγκαστικά η ευθεία που προσαρμόζεται στα δεδομένα των δύο μεταβλητών να διέρχεται από την αρχή των αξόνων. Ο ορισμός του εκτιμητή βασίζεται στην ευθεία παλινδρόμησης που προσαρμόζεται στα σημεία $\{(Y_{11}, Y_{21}),\ (Y_{12}, Y_{22}), \dots, (Y_{1N}, Y_{2N})\}$, δηλαδή τα ζεύγη των παρατηρήσεων για τις μονάδες που πληθυσμού στις δύο μεταβλητές $Y_1$ και $Y_2$. Υποθέτοντας σταθερή κλίση της ευθείας για όλο το εύρος του πληθυσμού, το μοντέλο:

$$Y_1 = \alpha + \beta Y_2$$

είναι το γραμμικό μοντέλο που περιγράφει τη σχέση μεταξύ των μεταβλητών $Y_1$ και $Y_2$, και ισοδύναμα:

$$Y_1 - \hat{\bar{Y}}_1 = \beta(Y_2 - \hat{\bar{Y}}_2) \tag{7.16}$$

όπου $\left(\hat{\bar{Y}}_1, \hat{\bar{Y}}_2\right)$ το σημείο των δειγματικών μέσων των δύο μεταβλητών. Ο εκτιμητής παλινδρόμησης της πληθυσμιακής μέσης τιμής $\bar{Y}_1$ στη συνέχεια, ορίζεται ως η πρόβλεψη σύμφωνα με το μοντέλο (7.16) της $Y_1$ μεταβλητής, εάν η $Y_2$ μεταβλητή πάρει τιμή ίση με την πληθυσμιακή μέση τιμή της, δηλ. $Y_2 = \bar{Y}_2$. Αναλυτικά, ο εκτιμητής παλινδρόμησης (linear regression) του $\bar{Y}_1$, ας τον συμβολίσουμε με $\hat{\bar{Y}}_{1,lr}$ (με δείκτη lr, από το linear regression), δίνεται από τη σχέση:

$$\hat{\bar{Y}}_{1,lr} = \hat{\bar{Y}}_1 + \beta(\bar{Y}_2 - \hat{\bar{Y}}_2) \tag{7.17}$$

Όπως και στην περίπτωση του εκτιμητή λόγου της μέσης τιμής $\bar{Y}_1$ ενός χαρακτηριστικού, είναι απαραίτητο και για τον εκτιμητή παλινδρόμησης να γνωρίζουμε την πληθυσμιακή μέση τιμή του βοηθητικού χαρακτηριστικού $Y_2$ και επιπλέον να διαθέτουμε τα στοιχεία της έρευνας στο δείγμα, τόσο για τη μεταβλητή $Y_1$, όσο και για την $Y_2$. Ακόμα περισσότερο, στην περίπτωση του εκτιμητή παλινδρόμησης ο ορισμός (7.17) υποδηλώνει ότι είναι απαραίτητη η γνώση της κλίσης $\beta$ της ευθείας που προσαρμόζεται μεταξύ των μετρήσεων των $Y_1$ και $Y_2$ μεταβλητών στον πληθυσμό. Στην πράξη, ο συντελεστής $\beta$ στη γραμμή παλινδρόμησης εκτιμάται από τον αντίστοιχο συντελεστή, έστω $b$, όταν προσαρμόζουμε μια γραμμή παλινδρόμησης στα $n$ ζεύγη του δείγματος αντί για εκείνα του πληθυσμού. Άρα, για την περίπτωση όπου η κλίση της ευθείας παλινδρόμησης για τον πληθυσμό είναι άγνωστη, ο εκτιμητής παλινδρόμησης υπολογίζεται από τη σχέση:

$$\hat{\bar{Y}}_{1,lr} = \hat{\bar{Y}}_1 + b(\bar{Y}_2 - \hat{\bar{Y}}_2) \tag{7.18}$$

Μία τρίτη εκδοχή για τον συντελεστή $\beta$ στον ορισμό του εκτιμητή είναι όταν ο συντελεστής είναι γνωστός και προκαθορισμένος πριν από την έρευνα. Το ενδεχόμενο αυτό μπορεί να συμβεί σε έρευνες που επαναλαμβάνονται σε τακτά χρονικά διαστήματα και στις οποίες έχει διαπιστωθεί ότι ο συντελεστής $\beta$ δεν μεταβάλλεται, αλλά παραμένει σταθερός στην πάροδο του χρόνου. Αν $\beta_0$ είναι η σταθερή αυτή τιμή, τότε ο εκτιμητής παλινδρόμησης $\hat{\bar{Y}}_{1,lr}$ υπολογίζεται από τη σχέση (7.16), χρησιμοποιώντας το $\beta_0$ στη θέση του συντελεστή $\beta$.

Αν ο εκτιμητής παλινδρόμησης ορίζεται μέσω της αληθινής τιμής της κλίσης της ευθείας παλινδρόμησης $\beta$ ή μιας προκαθορισμένης τιμής $\beta_0$ που δεν εξαρτάται από τα δεδομένα της τρέχουσας έρευνας, τότε αποδεικνύεται ότι ο εκτιμητής είναι αμερόληπτος. Συνολικά, για τον εκτιμητή $\hat{\bar{Y}}_{1,lr}$ που ορίζεται από την (7.16), ισχύουν τα αποτελέσματα της Πρότασης που ακολουθεί.

### Πρόταση 7.7

Ο εκτιμητής παλινδρόμησης $\hat{\bar{Y}}_{1,lr}$ ενός α.τ. δείγματος που επιλέγεται από τον πληθυσμό είναι αμερόληπτος εκτιμητής της πληθυσμιακής μέσης τιμής $\bar{Y}_1$ και έχει διακύμανση:

$$\mathrm{Var}\left(\hat{\bar{Y}}_{1,lr}\right) = \frac{1-f}{n}(S_1^2 - 2\beta S_{12} + \beta^2 S_2^2) \tag{7.19}$$

### Απόδειξη

Για την αναμενόμενη τιμή, λόγω της α.τ.δ. και επειδή το $\beta$ είναι σταθερό και δεν εξαρτάται από το δείγμα, ισχύει:

$$\mathrm{E}\left(\hat{\bar{Y}}_{1,lr}\right) = \mathrm{E}\left(\hat{\bar{Y}}_1\right) + \beta\mathrm{E}(\bar{Y}_2 - \hat{\bar{Y}}_2) = \bar{Y}_1 + \beta(\bar{Y}_2 - \mathrm{E}(\hat{\bar{Y}}_2)) = \bar{Y}_1$$

επειδή $\mathrm{E}(\hat{\bar{Y}}_2) = \bar{Y}_2$.

Συνεπώς, ο $\hat{\bar{Y}}_{1,lr}$ είναι αμερόληπτος εκτιμητής του $\bar{Y}_1$. Για τη διακύμανση του εκτιμητή, θεωρούμε ως τιμές του δείγματος τις παρατηρήσεις $X_{1i} + \beta(\bar{Y}_2 - X_{2i}),\ i = 1,2,\dots,n$, οι οποίες, σύμφωνα με το προηγούμενο αποτέλεσμα, προέρχονται από έναν πληθυσμό με μέση τιμή $\bar{Y}_1$, και, με βάση τα αποτελέσματα που ισχύουν στην α.τ.δ. για τη διακύμανση, θα είναι:

$$\mathrm{Var}\left(\hat{\bar{Y}}_{1,lr}\right) = \frac{1-f}{n}\sum_{i=1}^{N}\frac{[Y_{1i} - \beta(Y_{2i} - \bar{Y}_2) - \bar{Y}_1]^2}{N-1} = \frac{1-f}{n}\sum_{i=1}^{N}\frac{[(Y_{1i} - \bar{Y}_1) - \beta(Y_{2i} - \bar{Y}_2)]^2}{N-1}$$

$$= \frac{1-f}{n}(S_1^2 - 2\beta S_{12} + \beta^2 S_2^2)$$

■

### Πόρισμα 7.6

Σύμφωνα με την α.τ.δ. για την επιλογή του δείγματος των παρατηρήσεων των μεταβλητών $Y_1$ και $Y_2$ ένας αμερόληπτος εκτιμητής της $\mathrm{Var}\left(\hat{\bar{Y}}_{1,lr}\right)$ είναι ο:

$$\hat{\mathrm{Var}}\left(\hat{\bar{Y}}_{1,lr}\right) = \frac{1-f}{n}(s_1^2 - 2\beta s_{12} + \beta^2 s_2^2)$$

Στην περίπτωση κατά την οποία ο εκτιμητής παλινδρόμησης ορίζεται μέσω της σχέσης (7.18), όπου ο συντελεστής $\beta$ εκτιμάται από τα δεδομένα της τρέχουσας έρευνας, αποδεικνύεται ότι ο εκτιμητής $\hat{\bar{Y}}_{1,lr}$ δεν είναι αμερόληπτος, και το ποσό μεροληψίας του είναι της τάξης $1/n$, όπου $n$ το μέγεθος του δείγματος (για την απόδειξη, βλ. Cochran, 1977, παράγραφος 7.7). Κατά συνέπεια, για μεγάλα μεγέθη δείγματος, ο εκτιμητής παλινδρόμησης μπορεί να θεωρηθεί αμερόληπτος, ακόμα και όταν γίνεται εκτίμηση του συντελεστή β από τα δεδομένα.

## 7.6. Σύγκριση του εκτιμητή παλινδρόμησης και του εκτιμητή από την α.τ.δ.

Στην παράγραφο αυτή, συγκρίνουμε ως προς την ακρίβεια τον εκτιμητή παλινδρόμησης της πληθυσμιακής μέσης τιμής (και ισοδύναμα του συνόλου) ενός χαρακτηριστικού και τον εκτιμητή που υπολογίζεται από ένα α.τ.δ. χωρίς χρήση βοηθητικής μεταβλητής. Στην περίπτωση του εκτιμητή λόγου, η παραπάνω σύγκριση έδειξε ότι ο εκτιμητής έχει βελτιωμένο τυπικό σφάλμα μόνο κάτω από προϋποθέσεις. Οι προϋποθέσεις αφορούν τη συσχέτιση μεταξύ της κύριας και της βοηθητικής μεταβλητής. Παρουσιάζει λοιπόν ενδιαφέρον το ερώτημα εάν κάτι ανάλογο ισχύει και στην περίπτωση του εκτιμητή παλινδρόμησης. Η Πρόταση που ακολουθεί απαντά στο ερώτημα και αποτελεί τη βάση για την περαιτέρω σύγκριση του εκτιμητή λόγου και του εκτιμητή παλινδρόμησης.

### Πρόταση 7.8

Ο εκτιμητής παλινδρόμησης $\hat{\bar{Y}}_{1,lr}$ έχει διακύμανση μικρότερη από τη διακύμανση του $\hat{\bar{Y}}_1$ σύμφωνα με την α.τ.δ.

### Απόδειξη

Ο πληθυσμιακός εκτιμητής ελαχίστων τετραγώνων του συντελεστή $\beta$ του μη-σταθερού όρου στο γραμμικό μοντέλο $Y_1 = \alpha + \beta Y_2$ όταν αυτό προσαρμόζεται στα $N$ ζεύγη παρατηρήσεων του πληθυσμού, ο οποίος ονομάζεται και συντελεστής γραμμικής παλινδρόμησης (linear regression coefficient) για πεπερασμένους πληθυσμούς, δίνεται από τη σχέση:

$$\beta = \frac{S_{12}}{S_2^2} = \frac{\sum_{i=1}^{N}(Y_{1i}-\bar{Y}_1)(Y_{2i}-\bar{Y}_2)}{\sum_{i=1}^{N}(Y_{2i}-\bar{Y}_2)^2}$$

Επίσης,

$$\beta = \rho\sqrt{\frac{S_1^2}{S_2^2}} \tag{7.20}$$

όπου $\rho$ ο συντελεστής συσχέτισης για πεπερασμένους πληθυσμούς. Με τη βοήθεια της τελευταίας σχέσης, η διακύμανση του εκτιμητή $\hat{\bar{Y}}_{1,lr}$ που δίνεται από την (7.19) γίνεται:

$$\mathrm{Var}\left(\hat{\bar{Y}}_{1,lr}\right) = \frac{1-f}{n}(S_1^2 - 2\beta S_{12} + \beta^2 S_2^2) = \frac{1-f}{n}(S_1^2 - 2\beta^2 S_2^2 + \beta^2 S_2^2) = \frac{1-f}{n}(S_1^2 - \beta^2 S_2^2)$$

ή:

$$\mathrm{Var}\left(\hat{\bar{Y}}_{1,lr}\right) = \frac{1-f}{n}S_1^2(1-\rho^2) \tag{7.21}$$

ενώ η διακύμανση του απλού δειγματικού μέσου όρου με βάση την α.τ.δ. είναι $\mathrm{Var}\left(\hat{\bar{Y}}_1\right) = \frac{1-f}{n}S_1^2$. Διαιρώντας τις δύο διακυμάνσεις, προκύπτει ότι:

$$\frac{\mathrm{Var}\left(\hat{\bar{Y}}_{1,lr}\right)}{\mathrm{Var}\left(\hat{\bar{Y}}_1\right)} = 1-\rho^2 \ \ (<1)$$

αφού $0 < \rho^2 < 1$.

Αυτό σημαίνει ότι $\mathrm{Var}\left(\hat{\bar{Y}}_1\right) > \mathrm{Var}\left(\hat{\bar{Y}}_{1,lr}\right)$ και, συνεπώς, ο εκτιμητής $\hat{\bar{Y}}_{1,lr}$ έχει μεγαλύτερη ακρίβεια από τον εκτιμητή $\hat{\bar{Y}}_1$ ■

Σύμφωνα με το αποτέλεσμα της Πρότασης 7.8, ανεξάρτητα από το μέγεθος της γραμμικής συσχέτισης των μεταβλητών $Y_1$ και $Y_2$, ο εκτιμητής παλινδρόμησης είναι πιο ακριβής από τον δειγματικό μέσο για τη μεταβλητή $Y_1$ που προκύπτει από μια α.τ. δειγματοληψία. Ωστόσο, το όφελος στη διακύμανση του εκτιμητή με τη χρήση της βοηθητικής μεταβλητής θα είναι τόσο μεγαλύτερο, όσο μεγαλύτερος, κατ' απόλυτη τιμή, είναι ο συντελεστής συσχέτισης μεταξύ της κύριας και της βοηθητικής μεταβλητής.

Ο εκτιμητής παλινδρόμησης υπερτερεί πάντα ως προς την ακρίβεια του απλού δειγματικού μέσου, ενώ ο εκτιμητής λόγου όχι απαραίτητα. Η σχέση μεταξύ των εκτιμητών λόγου και παλινδρόμησης, ως προς την ακρίβεια της εκτίμησης, συνοψίζεται στην Πρόταση που ακολουθεί.

### Πρόταση 7.9

Ο εκτιμητής παλινδρόμησης $\hat{\bar{Y}}_{1,lr}$ είναι πιο ακριβής από τον εκτιμητή λόγου $\hat{\bar{Y}}_{1,R}$ που κάνει χρήση της ίδιας βοηθητικής μεταβλητής. Οι δύο εκτιμητές είναι ισοδύναμοι ως προς την ακρίβεια, αν, και μόνον αν, η ευθεία παλινδρόμησης διέρχεται από την αρχή των αξόνων.

### Απόδειξη

Οι διακυμάνσεις των $\hat{\bar{Y}}_{1,lr}$ και $\hat{\bar{Y}}_{1,R}$ είναι:

$$\mathrm{Var}\left(\hat{\bar{Y}}_{1,lr}\right) = \frac{1-f}{n} S_1^2 (1-\rho^2) \quad \textit{και}$$

$$\mathrm{Var}\left(\hat{\bar{Y}}_{1,R}\right) = \frac{1-f}{n} (S_1^2 - 2\rho R S_1 S_2 + R^2 S_2^2)$$

αντίστοιχα. Θεωρώντας τη διαφορά τους, προκύπτει:

$$\mathrm{Var}\left(\hat{\bar{Y}}_{1,R}\right) - \mathrm{Var}\left(\hat{\bar{Y}}_{1,lr}\right) = \frac{1-f}{n} (\rho^2 S_1^2 - 2\rho R S_1 S_2 + R^2 S_2^2) = \frac{1-f}{n} (\rho S_1 - R S_2)^2$$

Επειδή $\rho S_1 = \beta S_2$ από την (7.20), η διαφορά διακυμάνσεων ισούται με:

$$\mathrm{Var}\left(\hat{\bar{Y}}_{1,R}\right) - \mathrm{Var}\left(\hat{\bar{Y}}_{1,lr}\right) = \frac{1-f}{n} S_2^2 (\beta - R)^2$$

Άρα:

$$\mathrm{Var}\left(\hat{\bar{Y}}_{1,R}\right) \geq \mathrm{Var}\left(\hat{\bar{Y}}_{1,lr}\right)$$

δηλ. για μεγάλα δείγματα, ο εκτιμητής παλινδρόμησης έχει μικρότερη διακύμανση σε σχέση με τον εκτιμητή λόγου. Επίσης, οι δύο διακυμάνσεις ταυτίζονται όταν $\beta = R$. Η τελευταία ισότητα ισχύει όταν η γραμμή παλινδρόμησης των δύο μεταβλητών (κύριας και βοηθητικής) διέρχεται από την αρχή των αξόνων. Πράγματι, αν η γραμμή παλινδρόμησης είναι $Y_1 = \beta Y_2$ τότε για το σημείο των μέσων $(\bar{Y}_1, \bar{Y}_2)$ ισχύει $\bar{Y}_1 = \beta \bar{Y}_2$, απ' όπου $\beta = \frac{\bar{Y}_1}{\bar{Y}_2} = R$ ■

Για τον εκτιμητή της διακύμανσης του εκτιμητή παλινδρόμησης, όπως και στην περίπτωση του εκτιμητή λόγου, αντικαθιστούμε τις πληθυσμιακές ποσότητες π.χ. $S_1^2$ και $\rho^2$στην έκφραση (7.21) με τις αντίστοιχες δειγματικές, που υπολογίζονται με τη βοήθεια των στοιχείων του δείγματος.

## Παράδειγμα 7.4

Συνεχίζοντας το Παράδειγμα 7.2, εάν υπολογίσουμε τον εκτιμητή παλινδρόμησης για το μέσο ετήσιο κόστος ιατρικής φροντίδας των κατοικίδιων ζώων, αυτός προκύπτει

```
> lm(cost ~ visits)
```

```
Call:
lm(formula = cost ~ visits)

Coefficients:
(Intercept)       visits
     22.148        5.459
```

Άρα, ο συντελεστής $\beta$, όπως εκτιμάται από τα δεδομένα, είναι 5.46. Ισοδύναμα:

```
> beta<-s12/var(visits)
> beta
[1] 5.458748
```

Η τιμή του εκτιμητή παλινδρόμησης του μέσου κόστους είναι:

```
> cost_lr<-mean(cost)+beta*(PopMeanVisits-mean(visits))
> cost_lr
[1] 39.36448
```

Δηλαδή, το μέσο κόστος κάνοντας χρήση της μεταβλητής του αριθμού επισκέψεων, εκτιμάται ως 39.364$ ενώ χωρίς χρήση της βοηθητικής, ο απλός δειγματικός μέσος για τα δεδομένα είναι 39.726$. Για τη διακύμανση, υπολογίζουμε αρχικά τον συντελεστή συσχέτισης $\rho$, και στη συνέχεια εφαρμόζουμε τη σχέση (7.21).

```
> rho<-s12/sqrt(var(cost)*var(visits))
> var_cost_lr<-(1-50/1300)*(1/49)*var(cost)*(1-rho^2)
> var_cost_lr
[1] 3.812648
```

Άρα, η διακύμανση του εκτιμητή παλινδρόμησης είναι 3.81, ενώ η διακύμανση του εκτιμητή χωρίς χρήση βοηθητικής μεταβλητής, όπως υπολογίστηκε στο Παράδειγμα 7.2, είναι 4.936.

## 7.7. Εκτιμητής παλινδρόμησης και στρωματοποιημένη δειγματοληψία

Για τον εκτιμητή παλινδρόμησης, όταν πρόκειται να εφαρμοστεί σε στρωματοποιημένο πληθυσμό, υπάρχουν οι δύο ανάλογοι τρόποι εφαρμογής που μελετήσαμε και στην περίπτωση της εκτίμησης λόγου: Η από κοινού ή συνδυασμένη εκτίμηση και η ανεξάρτητη. Εν συντομία, ο εκτιμητής για τις δύο δυνατές εφαρμογές του είναι:

**Α. Συνδυασμένος ή από κοινού εκτιμητής παλινδρόμησης (combined regression estimate).**

Ακριβώς όπως και στην περίπτωση του συνδυασμένου εκτιμητή λόγου, αρχικά εφαρμόζεται η εκτίμηση κάτω από το στρωματοποιημένο σχέδιο για την εκτίμηση των μέσων τιμών του κύριου και του βοηθητικού χαρακτηριστικού, και στη συνέχεια εφαρμόζεται ο ορισμός του εκτιμητή παλινδρόμησης. Αν συμβολίσουμε τον εκτιμητή με $\hat{\bar{Y}}_{1,lr,c}$, θα είναι:

$$\hat{\bar{Y}}_{1,lr,c} = \hat{\bar{Y}}_{1,st} + \beta(\bar{Y}_2 - \hat{\bar{Y}}_{2,st})$$

όπου $\hat{\bar{Y}}_{1,st}$ και $\hat{\bar{Y}}_{2,st}$ είναι οι εκτιμητές που προκύπτουν από τη στρωματοποιημένη δειγματοληψία.

Η διακύμανση του συνδυασμένου, ή από κοινού, εκτιμητή παλινδρόμησης $\hat{\bar{Y}}_{1,lr,c}$ δίνεται από τη σχέση:

$$\mathrm{Var}\left(\hat{\bar{Y}}_{1,Rc}\right) \cong \sum_{h=1}^{L} \frac{1-f_h}{n_h} W_h^2 \left(S_{1h}^2 - 2\beta S_{12,h} + \beta^2 S_{2h}^2\right)$$

**Β. Ανεξάρτητος εκτιμητής παλινδρόμησης (separate regression estimate).**

Για τον υπολογισμό του ανεξάρτητου εκτιμητή παλινδρόμησης, αρχικά υπολογίζεται ο εκτιμητής παλινδρόμησης σε κάθε στρώμα χωριστά, και στη συνέχεια γίνεται ο συνδυασμός όλων των επιμέρους $L$ εκτιμήσεων που έχουν προκύψει, κάνοντας χρήση του σταθμισμένου εκτιμητή που ισχύει για τη στρωματοποιημένη. Αν συμβολίσουμε τον εκτιμητή με $\hat{\bar{Y}}_{1,lr,s}$, θα είναι:

$$\hat{\bar{Y}}_{1,lr,s} = \sum_{h=1}^{L} W_h \hat{\bar{Y}}_{1,lr,h} = \sum_{h=1}^{L} W_h \left[\hat{\bar{Y}}_{1,st,h} + \beta_h (\bar{Y}_{2h} - \hat{\bar{Y}}_{2,st,h})\right]$$

όπου $\hat{\bar{Y}}_{1,lr,h}$ είναι ο εκτιμητής παλινδρόμησης της μέσης τιμής για το $Y_1$ χαρακτηριστικό στο στρώμα $h$, $\hat{\bar{Y}}_{1,st,h}$ και $\hat{\bar{Y}}_{2,st,h}$ οι δειγματικοί μέσοι των δύο χαρακτηριστικών για το στρώμα $h$, $\beta_h$ ο συντελεστής παλινδρόμησης για το στρώμα $h$, και $\bar{Y}_{2h}$ ο πληθυσμιακός μέσος της βοηθητικής μεταβλητής για το στρώμα $h$.

Όπως είναι φανερό από τον ορισμό του εκτιμητή, για την εφαρμογή του απαιτείται να γνωρίζουμε τις πληθυσμιακές μέσες τιμές $\bar{Y}_{2h}$ για κάθε στρώμα $h$ του πληθυσμού χωριστά. Παράλληλα, ο συντελεστής $\beta$ εμφανίζεται με δείκτη $h$, δηλ. ορίζεται ανεξάρτητα από στρώμα σε στρώμα. Αυτό μπορεί να αποτελέσει ένα πλεονέκτημα για τον εκτιμητή παλινδρόμησης, επειδή μπορεί να καλύψει με καλύτερο τρόπο περιπτώσεις πληθυσμών για τους οποίους η γραμμή παλινδρόμησης των δύο μεταβλητών δεν έχει σταθερή κλίση σε όλο το εύρος του πληθυσμού, αλλά διαφέρει από στρώμα σε στρώμα.

Η διακύμανση του ανεξάρτητου εκτιμητή παλινδρόμησης $\hat{\bar{Y}}_{1,lr,s}$, στην περίπτωση που ο συντελεστής $\beta_h$, ή ισοδύναμα ο συντελεστής συσχέτισης $\rho_h$, ανά στρώμα, είναι γνωστός ή προκαθορισμένος, δίνεται από τη σχέση:

$$\mathrm{Var}\left(\hat{\bar{Y}}_{1,lr,s}\right) \cong \sum_{h=1}^{L} \frac{1-f_h}{n_h} W_h^2 \left(S_{1h}^2 - 2\beta_h S_{12h} + \beta_h^2 S_{2h}^2\right)$$

Όπως και στα προηγούμενα, οι εκτιμητές των διακυμάνσεων των εκτιμητών παλινδρόμησης στη στρωματοποιημένη υπολογίζονται εκτιμώντας τις πληθυσμιακές ποσότητες που εμφανίζονται στις εκφράσεις των διακυμάνσεων με τη βοήθεια του δείγματος.

Αποδεικνύεται ότι οι διακυμάνσεις των δύο εκτιμητών παλινδρόμησης για ένα στρωματοποιημένο σχήμα ταυτίζονται όταν $\forall\, h,\ \beta_h = \beta$, δηλαδή ο συντελεστής παλινδρόμησης είναι ίδιος για όλα τα στρώματα, ενώ, όταν οι συντελεστές $\beta_h$ διαφέρουν, ο ανεξάρτητος εκτιμητής παλινδρόμησης $\hat{\bar{Y}}_{1,lr,s}$ είναι πιο αποτελεσματικός.

## Παράδειγμα 7.5

Οι φοιτητές του 3ου έτους σε ένα Τμήμα Μαθηματικών προσέρχονται για τις εξετάσεις στο μάθημα της Στατιστικής. Συνολικά εξετάζονται 210 φοιτητές. Επιλέγουμε ένα αναλογικό στρωματοποιημένο δείγμα μεγέθους 30 φοιτητών με βοηθητική μεταβλητή τον μέσο όρο (μ.ο.) της μέχρι τώρα βαθμολογίας τους στα μαθήματα που έχουν εξεταστεί επιτυχώς. Τα όρια χωρισμού των δύο στρωμάτων είναι: 1ο στρώμα οι φοιτητές με μ.ο. βαθμολογίας κάτω από 6 και 2ο στρώμα οι φοιτητές με μ.ο. 6 ή παραπάνω. Τα στοιχεία του μ.ο. της βαθμολογίας του κάθε φοιτητή είναι διαθέσιμα από τη Γραμματεία του Τμήματος, καθώς επίσης και η πληροφορία ότι το 60% των φοιτητών ανήκει στο 1ο στρώμα. Ο μ.ο. της γενικής βαθμολογίας όλων των φοιτητών του Τμήματος είναι 7.4.

Για ένα δείγμα μεγέθους 30 και βάσει των πληθυσμιακών βαρών των στρωμάτων, επιλέγονται 18 φοιτητές με μ.ο. βαθμολογίας κάτω από 6 και 12 φοιτητές με μ.ο. 6 ή παραπάνω. Για τους φοιτητές του δείγματος, ζητάμε από τον εξεταστή τη βαθμολογία τους στο μάθημα της Στατιστικής.

Συμβολίζουμε με **y1** τη βαθμολογία στη Στατιστική και με **y2** τη γενική βαθμολογία, με **a** το πρώτο στρώμα και με **b** το δεύτερο.

Οι βαθμοί των φοιτητών του 1ου στρώματος, μαζί με τον γενικό μ.ο. τους είναι:

| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **y1a** | 7.00 | 0.00 | 4.00 | 7.50 | 0.00 | 0.00 | 2.00 | 2.50 | 3.00 | 2.00 | 3.50 | 2.50 | 5.00 | 1.50 | 3.50 | 7.00 | 2.50 | 4.00 |
| **y2a** | 5.67 | 5.20 | 5.38 | 5.95 | 5.25 | 5.40 | 5.74 | 5.53 | 5.90 | 5.60 | 5.39 | 5.80 | 5.24 | 5.60 | 5.85 | 5.76 | 5.20 | 5.35 |

Οι βαθμοί του 2ου στρώματος είναι, αντίστοιχα:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **y1b** | 7.50 | 9.50 | 7.00 | 6.50 | 7.50 | 7.00 | 8.50 | 5.00 | 8.50 | 9.00 | 8.00 | 10.00 |
| **y2b** | 8.21 | 9.20 | 8.65 | 6.23 | 8.20 | 6.90 | 8.80 | 6.21 | 8.57 | 8.30 | 8.50 | 8.76 |

Θα υπολογίσουμε την εκτίμηση της μέσης βαθμολογίας των φοιτητών στη Στατιστική με εκτιμητή παλινδρόμησης, χρησιμοποιώντας τη μεταβλητή του μέσου όρου (μ.ο.) της γενικής βαθμολογίας των φοιτητών ως βοηθητική μεταβλητή. Θα υπολογιστούν: (i) ο συνδυασμένος και (ii) ο ανεξάρτητος εκτιμητής παλινδρόμησης και τέλος, (iii) οι δύο εκτιμητές θα συγκριθούν ως προς την ακρίβεια.

Για τον συνδυασμένο εκτιμητή παλινδρόμησης, εκτιμούμε τον συντελεστή παλινδρόμησης βάσει του συνολικού δείγματος και εφαρμόζουμε τον ορισμό του $\hat{\bar{Y}}_{1,lr,c}$. Τα πληθυσμιακά βάρη των δύο στρωμάτων είναι 0.60 και 0.40 αντίστοιχα. Θα είναι αναλυτικά:

```
> y1<-c(y1a,y1b)
> y2<-c(y2a,y2b)
> lm(y1~y2)

Call:
lm(formula = y1 ~ y2)
Coefficients:
(Intercept)           y2
     -6.408        1.751
```

Άρα, ο εκτιμητής του συνολικού συντελεστή παλινδρόμησης, δηλαδή για τους φοιτητές χωρίς τον χωρισμό τους σε στρώματα, είναι 1.751. Με βάση το αποτέλεσμα αυτό, τα βάρη των στρωμάτων και τον ορισμό του εκτιμητή παλινδρόμησης $\hat{\bar{Y}}_{1,lr,c}$, θα είναι:

```
> beta<-1.751
> y1_lr_c<-(.6*mean(y1a)+.4*mean(y1b)) + beta*(7.4-
(.6*mean(y2a)+.4*mean(y2b)) )
> y1_lr_c
[1] 6.547689
```

Ο συνδυασμένος εκτιμητής παλινδρόμησης για τη μέση βαθμολογία στο μάθημα της Στατιστικής είναι, επομένως, 6.55. Η εκτίμηση της διακύμανσης του εκτιμητή υπολογίζεται από τον τύπο της $\text{Var}\left(\hat{\bar{Y}}_{1,lr,c}\right)$ εκτιμώντας τις ποσότητες του πληθυσμού που εμφανίζονται και είναι άγνωστες, από τις αντίστοιχες του δείγματος. Θα είναι αναλυτικότερα:

```
> f1=f2=30/210
> n1<-18
> n2<-12
>  var_c<-((1-f1)/n1)*.6^2*(var(y1a)-
2*beta*cov(y1a,y2a)+beta^2*var(y2a))+((1-f2)/n2)*.4^2*(var(y1b)-
2*beta*cov(y1b,y2b)+beta^2*var(y2b))
> var_c
[1] 0.09084781
> sqrt(var_c)
[1] 0.3014097
```

Άρα, το τυπικό σφάλμα της εκτίμησης της μέσης βαθμολογίας στη Στατιστική για την περίπτωση (i) είναι 0.30.

Για τον ανεξάρτητο εκτιμητή παλινδρόμησης, χρειάζεται να ζητήσουμε από τη Γραμματεία τα στατιστικά για τους μ.ο. των φοιτητών σε κάθε στρώμα. Έστω ότι ο μ.ο. των φοιτητών στο 1$^{ο}$ στρώμα είναι 5.38 και ο μ.ο. της γενικής βαθμολογίας των φοιτητών στο 2$^{ο}$ στρώμα είναι 7.62. Υπολογίζουμε στη συνέχεια τους συντελεστές παλινδρόμησης ανά στρώμα και εφαρμόζουμε τον ορισμό του $\hat{\bar{Y}}_{1,lr,s}$. Συγκεκριμένα:

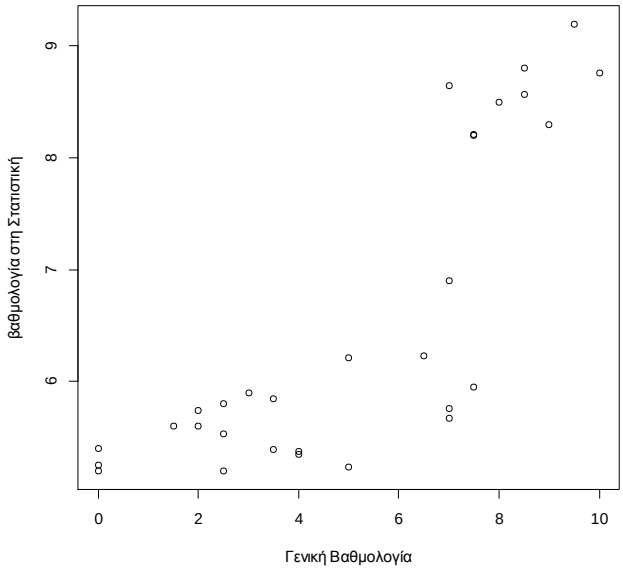
```
> lm(y1a~y2a)
Call:
lm(formula = y1a ~ y2a)

Coefficients:
(Intercept)          y2a
    -19.149        4.029
> beta1<-4.029

> lm(y1b~y2b)
Call:
lm(formula = y1b ~ y2b)
Coefficients:
(Intercept)          y2b
     -1.140        1.115
> beta2<-1.115

> y1_lr_s<-.6*(mean(y1a)+beta1*(5.38-mean(y2a)))+
.4*(mean(y1b)+beta2*(7.62-mean(y2b)))
> y1_lr_s
[1] 4.461951
```

Άρα, η εκτίμηση με την εφαρμογή του ανεξάρτητου στρωματοποιημένου εκτιμητή παλινδρόμησης είναι 4.46.

Για την τυπική του απόκλιση, λαμβάνοντας υπόψη τους εκτιμητές των πληθυσμιακών ποσοστήτων ανά στρώμα, θα έχουμε:

```
>  var_s<-((1-f1)/n1)*.6^2*(var(y1a)-
2*beta1*cov(y1a,y2a)+beta1^2*var(y2a))+((1-f2)/n2)*.4^2*(var(y1b)-
2*beta2*cov(y1b,y2b)+beta2^2*var(y2b))
> sqrt(var_s)
[1] 0.2835768
```

Το τυπικό σφάλμα του ανεξάρτητου εκτιμητή παλινδρόμησης προκύπτει λοιπόν 0.28. Η γραφική παράσταση των μετρήσεων του δείγματος για τις δύο μεταβλητές είναι η ακόλουθη:




***Σχήμα 7.1*** *Γραφική παράσταση των βαθμολογιών του δείγματος φοιτητών.*

## 7.8. Εκτιμητές παλινδρόμησης και λόγου για περισσότερες από μία βοηθητικές μεταβλητές

Η επέκταση του προβλήματος της εκτίμησης λόγου ή παλινδρόμησης για περισσότερα από ένα βοηθητικά χαρακτηριστικά είναι άμεση. Έστω ότι $Y_1$ είναι το κύριο χαρακτηριστικό και $Y_1^{'}$, $Y_2^{'}$, ..., $Y_p^{'}$ , $p$ βοηθητικές μεταβλητές για τις οποίες είναι εφικτή η καταγραφή των απαντήσεων κατά τη διάρκεια της έρευνας και έστω επίσης ότι γνωρίζουμε τις πληθυσμιακές μέσες τιμές τους $\bar{Y}_1^{'}$, $\bar{Y}_2^{'}$, ..., $\bar{Y}_p^{'}$ αντίστοιχα. Ο πολυμεταβλητός εκτιμητής λόγου (multivariate ratio estimate) $\hat{\bar{Y}}_{1,MR}$ της μέσης τιμής του χαρακτηριστικού $Y_1$ ορίζεται ως:

$$\hat{\bar{Y}}_{1,MR} = w_1 \frac{\hat{\bar{Y}}_1}{\hat{\bar{Y}}_1^{'}} \bar{Y}_1^{'} + w_2 \frac{\hat{\bar{Y}}_1}{\hat{\bar{Y}}_2^{'}} \bar{Y}_2^{'} + \cdots + w_p \frac{\hat{\bar{Y}}_1}{\hat{\bar{Y}}_p^{'}} \bar{Y}_p^{'}$$

$$= w_1 \hat{R}_1 \bar{Y}_1^{'} + w_2 \hat{R}_2 \bar{Y}_2^{'} + \cdots + w_p \hat{R}_p \bar{Y}_p^{'}$$

όπου $\hat{R}_i$ $(i = 1,2, \ldots, p)$, είναι οι εκτιμώμενοι από το δείγμα λόγοι του κύριου χαρακτηριστικού με το καθένα από τα $Y_i^{'}$ βοηθητικά και $w_1$,$w_2$, ..., $w_p$ είναι βάρη, τέτοια ώστε $\sum_{i=1}^{p} w_i = 1$. Τα βάρη καθορίζονται από τον ερευνητή, ο οποίος τα επιλέγει είτε υποκειμενικά, ανάλογα με τη σπουδαιότητα της κάθε βοηθητικής μεταβλητής ως προς την κύρια, είτε βάσει των εκτιμήσεων των λόγων $\hat{R}_i$. Ένα κριτήριο καθορισμού των

βαρών μπορεί να είναι, για παράδειγμα, να δίνεται μεγαλύτερο βάρος σε βοηθητικές μεταβλητές με μεγαλύτερο $\hat{R}$, και αντίστροφα. Ένας άλλος τρόπος να καθοριστούν τα βάρη είναι να προκύψουν ως η λύση του προβλήματος ελαχιστοποίησης της τελικής διασποράς του εκτιμητή ως συνάρτησης των βαρών:

$$\min_{w}\left\{\mathrm{Var}\left(\hat{\bar{Y}}_{1,MR}\right)\right\}$$

Αντίστοιχα, και αρκετά άμεσα, προκύπτει η επέκταση του εκτιμητή παλινδρόμησης για την περίπτωση πολλών βοηθητικών μεταβλητών. Η επέκταση είναι ταυτόσημη με την επέκταση του προβλήματος της απλής γραμμικής παλινδρόμησης στην **πολλαπλή γραμμική (multiple linear regression)**. Συγκεκριμένα, υποθέτουμε το μοντέλο:

$$Y_1 = \alpha + \beta_1 Y_1^{'} + \beta_2 Y_2^{'} + \cdots + \beta_p Y_p^{'}$$

για τον πληθυσμό, και ο πολυμεταβλητός εκτιμητής παλινδρόμησης (multivariate regression estimate) του μέσου της $Y_1$, έστω $\hat{\bar{Y}}_{1,Mlr}$ , ορίζεται ως:

$$\hat{\bar{Y}}_{1,Mlr} = \hat{\bar{Y}}_1 + \beta_1\left(\bar{Y}_1^{'} - \hat{\bar{Y}}_1^{'}\right) + \beta_2\left(\bar{Y}_2^{'} - \hat{\bar{Y}}_2^{'}\right) + \cdots + \beta_p\left(\bar{Y}_p^{'} - \hat{\bar{Y}}_p^{'}\right)$$

## Παράδειγμα 7.6

Ένα τηλεοπτικό κανάλι έρχεται σε συμφωνία με ένα ερευνητικό ινστιτούτο, για την παραγωγή των προβλέψεων των αποτελεσμάτων που θα πάρει το κόμμα Α στις εκλογές. Το ινστιτούτο θα πρέπει να δώσει στο κανάλι κάποιους πρώτους εκτιμητές για το ποσοστό του κόμματος Α στις 8 το βράδυ των εκλογών, ενώ οριστικά αποτελέσματα για τον πληθυσμό θα είναι γνωστά μετά τις 10 το βράδυ. Το ινστιτούτο, για τον σκοπό αυτό, επιλέγει τυχαία $m$ exit polls από εκλογικά τμήματα, και στις 7:50 είναι σε θέση να γνωρίζει το $t_i$ $(i = 1,2,..,m)$, που είναι ο αριθμός των ψήφων που πήρε το κόμμα Α στην πλήρη καταμέτρηση του exit poll του τμήματος $i$. Έστω ότι $M = 500$ είναι συνολικά τα εκλογικά τμήματα, τα οποία έχουν ίσο αριθμό ψηφοφόρων $K = 1000$. Για την έρευνα, επιλέγονται $m = 15$ τμήματα και οι τιμές των $t_i$ του δείγματος με τις θετικές απαντήσεις των ψηφοφόρων για το κόμμα Α είναι:

| **Exit poll $i$** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $t_i$ | 302 | 336 | 278 | 305 | 290 | 250 | 340 | 319 | 306 | 325 | 295 | 282 | 293 | 314 | 340 |

Με βάση τα δεδομένα, θέλουμε να απαντήσουμε στα ερωτήματα:

(i) να δοθεί η εκτίμηση ποσοστού του κόμματος Α που θα προσφέρει το ινστιτούτο στις 8 το βράδυ βάσει των exit polls των τμημάτων, καθώς κι ένα εκτιμώμενο τυπικό σφάλμα της εκτίμησης.

(ii) το ινστιτούτο διαθέτει τα στοιχεία για τον αριθμό των ψηφοφόρων του κόμματος Α στα ίδια εκλογικά τμήματα στις προηγούμενες εκλογές. Επίσης, είναι γνωστό ότι το κόμμα Α στις προηγούμενες εκλογές είχε ποσοστό επιτυχίας 26%. Πως θα τροποποιηθεί ο εκτιμητής και το τυπικό σφάλμα του προηγούμενου ερωτήματος, εάν το ινστιτούτο εφαρμόσει α) εκτιμητή λόγου και β) εκτιμητή παλινδρόμησης για το ποσοστό; Τα στοιχεία των προηγουμένων εκλογών είναι:

| **Exit poll $i$** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $t_i$ | 245 | 287 | 252 | 236 | 197 | 168 | 245 | 210 | 230 | 215 | 195 | 202 | 210 | 235 | 265 |

**Απάντηση**

(i) Για την εκτίμηση του ποσοστού στην περίπτωση της τυχαίας επιλογής των εκλογικών τμημάτων κάνουμε χρήση της θεωρίας για την κατά ομάδες δειγματοληψία σε ένα στάδιο με ίσο μέγεθος ομάδων και ίση πιθανότητα επιλογής. Έστω $t$ το διάνυσμα με τις ψήφους που συγκέντρωσε το κόμμα Α. Εισάγουμε αρχικά τα δεδομένα και στη συνέχεια υπολογίζουμε τον εκτιμητή από τη σχέση (6.2). Θα είναι αναλυτικά:

```
> t<-c(302, 336, 278, 305, 290, 250, 340, 319, 306, 325, 295, 282,
293, 314, 340)
> p<-sum(t)/(14*1000)
> p
[1] 0.3267857
```

Άρα, το ποσοστό που θα συγκεντρώσει το κόμμα Α στις εκλογές εκτιμάται σε 32.68%. Η εκτιμώμενη διακύμανση του εκτιμητή υπολογίζεται από τη σχέση (6.3). Η εφαρμογή του τύπου στα δεδομένα δίνει:

```
> K<-1000
> m<-15
> f<-15/500
> pi<-t/1000
> var_a<-(1-f)*(1/m)*var(pi)
> se_a<-sqrt(var_a)
> se_a
[1] 0.006393635
```

Κατά συνέπεια, το τυπικό σφάλμα του εκτιμητή του ποσοστού για το πρώτο ερώτημα, όπου δεν γίνεται χρήση βοηθητικού χαρακτηριστικού, εκτιμάται ότι είναι 0.64%.

(ii) Έστω ότι $t0$ είναι το διάνυσμα των ψήφων του κόμματος Α κατά τις προηγούμενες εκλογές. Ο δειγματικός εκτιμητής του ποσοστού του κόμματος Α στις προηγούμενες εκλογές, βάσει του δείγματος των εκλογικών τμημάτων, υπολογίζεται όπως και στο προηγούμενο ερώτημα.

```
>t0<-c(245,287,252,236,197,168,245,210,230,215,195,202,210,235,265)
> p0<-sum(t0)/(14*1000)
> p0
[1] 0.2422857
```

α) Ο εκτιμητής λόγου του ποσοστού του κόμματος Α στις τωρινές εκλογές είναι:

```
> p_ratio<-(p/p0)*.26
> > p_ratio
[1] 0.3506781
```

άρα, η εκτίμηση λόγου του ποσοστού είναι 35.06% και το τυπικό του σφάλμα εκτιμάται με τη βοήθεια της έκφρασης (7.15), ή ισοδύναμα από την (7.6). Θα είναι

```
> pi<-t/1000
> pi0<-t0/1000
> R<-p/p0
> var_2R<-(1-f)*(1/m)*sum((pi-R*pi0)^2)/14
```

```
> sqrt(var_2R)
[1] 0.007659517
```

Το τυπικό σφάλμα του εκτιμητή λόγου του ποσοστού είναι 0.77%, αντί για 0.64% στο πρώτο ερώτημα, άρα ο εκτιμητής λόγου του ζητούμενου ποσοστού είναι λιγότερο ακριβής.

β) Για τον εκτιμητή παλινδρόμησης, εκτιμούμε αρχικά την κλίση b της γραμμής παλινδρόμησης που προσαρμόζεται στα δεδομένα. Θα είναι

```
> lm(pi~pi0)

Call:
lm(formula = pi ~ pi0)

Coefficients:
(Intercept)          pi0
     0.1769       0.5663
```

άρα ο συντελεστής της ανεξάρτητης μεταβλητής του μοντέλου εκτιμάται σε 0.56 και, με τη βοήθειά του, υπολογίζεται ο εκτιμητής παλινδρόμησης από τη σχέση (7.17). Αναλυτικά:

```
> b<-0.56
> p_lr<-p+b*(p0-.26)
> p_lr
[1] 0.3168657
```

Συνεπώς, η εκτίμηση του ποσοστού του κόμματος Α είναι 31.69%. Το εκτιμώμενο τυπικό σφάλμα της εκτίμησης υπολογίζεται από τη σχέση (7.21).

```
> rho<-cor(pi, pi0)
> rho
[1] 0.6919364
> var_2lr<-var_a*(1-rho^2)
> sqrt(var_2lr)
[1] 0.004615939
```

Άρα ο εκτιμητής παλινδρόμησης του ποσοστού έχει τυπικό σφάλμα 0.46% και είναι συνεπώς πιο ακριβής από όλους τους προηγούμενους εκτιμητές.

### Απόδειξη Πρότασης 7.2

Για την απόδειξη του αποτελέσματος, γίνεται χρήση του αναπτύγματος σε σειρά Taylor της ποσότητας $\hat{R} - R$. Θα είναι:

$$\hat{R} - R = \frac{\bar{X}_1 - \bar{X}_2 R}{\bar{Y}_2}\left(\frac{\bar{Y}_2}{\bar{X}_2}\right) = \frac{\bar{X}_1 - \bar{X}_2 R}{\bar{Y}_2}\left(1 + \frac{\bar{X}_2 - \bar{Y}_2}{\bar{Y}_2}\right)^{-1}$$

$$= \frac{\bar{X}_1 - \bar{X}_2 R}{\bar{Y}_2}\left[1 - \frac{\bar{X}_2 - \bar{Y}_2}{\bar{Y}_2} + \left(\frac{\bar{X}_2 - \bar{Y}_2}{\bar{Y}_2}\right)^2 - \cdots\right]$$

$$\hat{R} - R = \frac{\bar{X}_1}{\bar{X}_2} - R = \frac{\bar{X}_1 - \bar{X}_2 R}{\bar{X}_2} = \frac{\bar{X}_1 - \bar{X}_2 R}{\bar{Y}_2}\left(\frac{\bar{Y}_2}{\bar{X}_2}\right) = \frac{\bar{X}_1 - \bar{X}_2 R}{\bar{Y}_2}\left(1 + \frac{\bar{X}_2 - \bar{Y}_2}{\bar{Y}_2}\right)^{-1}$$

$$= \frac{\bar{X}_1 - \bar{X}_2 R}{\bar{Y}_2}\left[1 - \frac{\bar{X}_2 - \bar{Y}_2}{\bar{Y}_2} + \left(\frac{\bar{X}_2 - \bar{Y}_2}{\bar{Y}_2}\right)^2 - \cdots\right]$$

Αν κρατήσουμε τους δύο πρώτους όρους του αναπτύγματος, το ποσό μεροληψίας του $\hat{R}$ είναι κατά προσέγγιση:

$$\text{Bias}(\hat{R}) = \text{E}(\hat{R} - R) = \text{E}\left(\frac{\bar{X}_1 - \bar{X}_2 R}{\bar{Y}_2}\right) - \text{E}\left[\left(\frac{\bar{X}_1 - \bar{X}_2 R}{\bar{Y}_2}\right)\left(\frac{\bar{X}_2 - \bar{Y}_2}{\bar{Y}_2}\right)\right]$$

$$= 0 - \frac{1}{\bar{Y}_2^2}\text{E}[(\bar{X}_1 - \bar{X}_2 R)(\bar{X}_2 - \bar{Y}_2)] = -\frac{1}{\bar{Y}_2^2}\{\text{E}[\bar{X}_1(\bar{X}_2 - \bar{Y}_2)] + R\text{E}[\bar{X}_2(\bar{X}_2 - \bar{Y}_2)]\}$$

$$= \frac{1}{\bar{Y}_2^2}[R[\text{E}(\bar{X}_2^2) - \bar{Y}_2^2] - \text{Cov}(\bar{X}_1, \bar{X}_2)] = \frac{1-f}{n\bar{Y}_2^2}(RS_2^2 - \rho S_1 S_2)$$

γιατί $\text{Var}(\bar{X}_2) = \text{E}(\bar{X}_2^2) - \bar{Y}_2^2 = \frac{1-f}{n} S_2^2$ και $\text{Var}(\bar{X}_1) = \frac{1-f}{n} S_1^2$ ■

# Βιβλιογραφικές Αναφορές


Barnett, V. (2002). *Sample survey: Principles and methods* (3rd Edition). London: Arnold.

Cochran, W. G. (1977). *Sampling Techniques* (3rd Edition). New York: John Wiley and Sons.

Levy P.S. and Lemeshow, S. (1999). *Sampling of populations. Methods and applications* (3rd Edition). New York: John Wiley and Sons.

Lohr, S. L. (2010). *Sampling: Design and Analysis* (2nd edition). Boston: Brooks/Cole, Cengage Learning.

# Κεφάλαιο 8 - ΑΣΚΗΣΕΙΣ

**8.1.** Από έναν πεπερασμένο πληθυσμό μεγέθους $N$ να εκλεγεί, με τη βοήθεια ενός προγράμματος παραγωγής τυχαίων αριθμών, ένα τυχαίο δείγμα μεγέθους $n$ = 10, α) όταν $N$ = 100 και β) όταν $N$ = 500.

**8.2.** Ο παρακάτω πίνακας είναι το σχεδιάγραμμα μιας τάξης. Σε κάθε θέση, εμφανίζεται το φύλο του μαθητή (Α=αγόρι, Κ=κορίτσι), το ύψος (σε cm - το πραγματικό ύψος προκύπτει αν στην τιμή που υπάρχει στον πίνακα προσθέσουμε 100) και το βάρος (σε kg).

Επιλέγοντας ένα δείγμα μεγέθους $n$ = 7, να εκτιμήσετε:

α) το ποσοστό των αγοριών,

β) το μέσο ύψος των μαθητών και

γ) το μέσο βάρος των μαθητών.

Για κάθε μία από τις παραπάνω περιπτώσεις, να δοθεί το τυπικό σφάλμα των εκτιμήσεων.

δ) Ποιο θα πρέπει να είναι το μέγεθος του δείγματος το οποίο θα πρέπει να επιλέξουμε από τον πληθυσμό των μαθητών, έτσι ώστε οι προηγούμενες εκτιμήτριες να μη διαφέρουν περισσότερο από 5% των αντίστοιχων πραγματικών τιμών, με πιθανότητα 0,95;

[Οι πληθυσμιακές τιμές για σύγκριση είναι: ποσοστό αγοριών $P_A$ = 0.62

μέσο ύψος μαθητών $\bar{Y}_{Υψος}$ = 71,09 cm, διασπορά ύψους $S^2_{Υψος}$ = 52,92 cm$^2$

μέσο βάρος μαθητών $\bar{Y}_{Βάρος}$ = 67,95 kg, διασπορά βάρους $S^2_{Βάρος}$ = 117,9 kg$^2$].

| | **Α** | **Β** | **Γ** | **Δ** | **Ε** | **Ζ** |
|---|---|---|---|---|---|---|
| **1** | Α 70 54 | Κ 62 50 | Κ 70 58 | Α 73 68 | Α 77 72 | Κ 64 62 |
| **2** | Κ 60 58 | Α 74 68 | Κ 70 70 | Κ 62 57 | Α 74 71 | Κ 67 55 |
| **3** | Α 75 85 | Α 70 74 | Α 88 81 | Α 76 70 | Κ 60 62 | Α 78 72 |
| **4** | Α 85 80 | Α 76 73 | Α 71 90 | Α 79 72 | Α 70 58 | Κ 69 57 |
| **5** | Α 70 54 | Α 73 70 | Α 78 70 | Κ 55 49 | Α 74 88 | Κ 60 55 |
| **6** | Α 70 54 | Α 71 65 | Κ 60 54 | Α 75 80 | Α 72 69 | Α 82 85 |
| **7** | Α 70 54 | Γ 65 60 | Α 80 81 | Γ 68 60 | Α 80 87 | Γ 65 63 |

***Πίνακας 8.1*** *Δεδομένα για την άσκηση 8.2*

**8.3.** Να αποδειχθεί ότι στην απλή τυχαία δειγματοληψία χωρίς επανατοποθέτηση από ένα πεπερασμένο πληθυσμό μεγέθους *N*, η συσχέτιση μεταξύ δύο οποιωνδήποτε μονάδων ισούται με -1/(*N*-1).

**8.4.** Να δειχθεί ότι στην απλή τυχαία δειγματοληψία χωρίς επανατοποθέτηση από ένα πεπερασμένο πληθυσμό μεγέθους *N*, η πιθανότητα εκλογής μιας οποιασδήποτε μονάδας είναι σταθερή και ίση με 1/*N*.

**8.5.** Σε ένα Internet Café με 98 υπολογιστές, ο ιδιοκτήτης επέλεξε στην τύχη ένα δείγμα 8 υπολογιστών και μέτρησε τους χρόνους χρήσης του διαδικτύου από κάθε χρήστη. Οι χρόνοι αυτοί (σε λεπτά) είναι οι εξής:

4,2 5,1 7,9 3,8 5,3 4,6 5,1 4,1.

Να εκτιμήσετε τον μέσο χρόνο χρήσης του διαδικτύου στο συγκεκριμένο Internet Café και να δώσετε το τυπικό σφάλμα της εκτίμησής σας.

**8.6.** Ένα γραφείο της ΔΕΗ γνωρίζει ότι 400 καταναλωτές δεν έχουν πληρώσει τους λογαριασμούς τους. Για να εκτιμήσει το συνολικό ποσό που οφείλουν οι πελάτες αυτοί, επέλεξε στην τύχη 10 από αυτούς και βρήκε ότι τα οφειλόμενα ποσά (σε δεκάδες ευρώ) ήταν τα εξής:

33, 32, 52, 43, 40, 41, 45, 42, 39, 28.

Να εκτιμήσετε το ολικό οφειλόμενο ποσό και να δώσετε το τυπικό σφάλμα της εκτίμησής σας.

**8.7.** Για την εκτίμηση της μέσης τιμής ενός πεπερασμένου πληθυσμού μεγέθους *N*, συνεχίζουμε να επιλέγουμε μονάδες με ίσες πιθανότητες και με επανατοποθέτηση, έως ότου πετύχουμε *d* διακεκριμένες μονάδες. Αν *n* είναι ο συνολικός αριθμός των μονάδων που επιλέξαμε και $k_r$ η συχνότητα εμφάνισης της *r* διακεκριμένης μονάδας στο δείγμα, να δειχθεί ότι:

α) οι εκτιμητές $\bar{y}_n = \sum_{r=1}^{d} \frac{k_r y_r}{n}$ και $\bar{y}_d = \sum_{r=1}^{d} \frac{y_r}{d}$ είναι αμερόληπτοι

β) $\mathrm{Var}(\bar{y}_n) = \mathrm{E}\frac{1}{n}\sigma_y^2$

γ) $\mathrm{E}(n) = N(\frac{1}{N} + \frac{1}{N-1} + L + \frac{1}{N-d-1})$

δ) $\mathrm{E}\left(\frac{1}{n}\right) > \frac{1}{\mathrm{E}(n)} > (N-d)[d(N-1)]^{-1}$

Με βάση τα παραπάνω, ή με οποιοδήποτε άλλο τρόπο, να δειχθεί ότι:

$$\mathrm{Var}(\bar{y}_n) \geq \mathrm{Var}(\bar{y}_d).$$

**8.8.** Ένας πληθυσμός αποτελείται από *N* μονάδες. Η τιμή τού υπό μελέτη χαρακτηριστικού για την πρώτη μονάδα είναι γνωστή και ίση με $y_1$. Από τις υπόλοιπες *N* - 1 μονάδες, επιλέγουμε, χωρίς επανατοποθέτηση, ένα τυχαίο δείγμα μεγέθους *n*. Να δειχθεί ότι ο εκτιμητής $y_1 + (N-1)\bar{y}_n$ έχει μικρότερη διακύμανση από τον εκτιμητή $N\bar{y}_n$, ο οποίος βασίζεται επίσης σε ένα τυχαίο δείγμα χωρίς επανατοποθέτηση, αλλά από ολόκληρο τον πληθυσμό. (Το δεύτερο δείγμα δηλαδή, δεν λαμβάνει υπόψη του την επιπλέον πληροφορία για το $y_1$).

**8.9.** α) Για κάποιο θέμα συγκεντρώθηκαν υπογραφές σε 676 σελίδες. Κάθε σελίδα έχει αρκετό χώρο για 42 υπογραφές, αλλά αρκετά φύλλα έχουν μικρότερο αριθμό υπογραφών. Ο αριθμός των υπογραφών ανά σελίδα, σε ένα τυχαίο δείγμα 50 φύλλων, παρουσιάζεται στον επόμενο πίνακα.

| **Αριθμός υπογραφών $Y_i$** | **Συχνότητα $f_i$** |
|---|---|
| 42 | 23 |
| 41 | 4 |
| 36 | 1 |
| 32 | 1 |
| 29 | 1 |
| 27 | 2 |
| 23 | 1 |
| 19 | 1 |
| 16 | 2 |
| 15 | 2 |
| 14 | 1 |
| 11 | 1 |
| 10 | 1 |
| 9 | 1 |
| 7 | 1 |
| 6 | 3 |
| 5 | 2 |
| 4 | 1 |
| 3 | 1 |

***Πίνακας 8.2*** *Πλήθος υπογραφών σε 50 τυχαία επιλεγμένες σελίδες*

Να εκτιμήσετε τον συνολικό αριθμό των υπογραφών και να δώσετε ένα 95% διάστημα εμπιστοσύνης.

β) Μετά από την επιλογή του δείγματος, ο αριθμός των σελίδων με πλήρη αριθμό υπογραφών (δηλαδή 42) μετρήθηκε και βρέθηκε ίσος με 326. Κάνοντας χρήση της πληροφορίας αυτής, να βρείτε έναν βελτιωμένο εκτιμητή του συνολικού αριθμού των υπογραφών που συγκεντρώθηκαν, και να υπολογίσετε την τυπική του απόκλιση.

**8.10.** Σε έναν πληθυσμό με $N = 6$, οι τιμές του πληθυσμού είναι 8, 3, 1, 11, 4 και 7. (i) Υπολογίστε τη δειγματική μέση τιμή για όλα τα δυνατά δείγματα μεγέθους 2 και δείξτε ότι ο εκτιμητής αυτός εκτιμά αμερόληπτα την πληθυσμιακή μέση τιμή. Επιβεβαιώστε ότι η διακύμανση του εκτιμητή δίνεται από τη σχέση $\frac{(1-f)}{n}S^2$. (ii) Για τον ίδιο πληθυσμό, να υπολογισθεί η τιμή τού $s^2$ για όλα τα δυνατά δείγματα μεγέθους 3 και να επαληθευθεί ότι $E(s^2) = S^2$.

**8.11.** Ένα απλό τυχαίο δείγμα 30 νοικοκυριών επιλέχθηκε τυχαία, από μία περιοχή μιας πόλης που περιέχει 14.848 νοικοκυριά. Ο αριθμός των ατόμων σε κάθε νοικοκυριό στο δείγμα έχει ως 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

Να εκτιμήσετε τον συνολικό αριθμό των ατόμων στην περιοχή και να υπολογίσετε την πιθανότητα ο εκτιμητής αυτός να μη διαφέρει περισσότερο από ±10% από την αληθινή τιμή.

**8.12.** Σε μια μελέτη για την πιθανή χρήση της δειγματοληψίας στην ελάττωση των απαραίτητων εργασιών για την αποθήκευση ενός συνόλου προϊόντων, η τιμή των προϊόντων αυτών, σε κάθε ένα από τα 36 ράφια της αποθήκης, μετρήθηκε και, σε στρογγυλοποιημένη μορφή στο πλησιέστερο ευρώ, έχει ως εξής:

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.

Ο εκτιμητής της συνολικής αξίας, ο οποίος θα βασίζεται σε ένα τυχαίο δείγμα, θα πρέπει να είναι σωστός στα ±200 ευρώ με πιθανότητα 1 στα 20. Ένας ειδικός συνιστά ότι μια απλή τυχαία δειγματοληψία με μέγεθος δείγματος 12 θα ήταν ικανοποιητική. Συμφωνείτε με τη γνώμη του;

[Δίνεται ότι $\sum y = 2.138$ και $\sum y^2 = 131.682$].

**8.13.** Δύο οδοντίατροι, οι κ.κ. Α και Β, θέλουν να εκτιμήσουν την κατάσταση των δοντιών στα 200 παιδιά ενός χωριού. Ο κ. Α επιλέγει ένα τυχαίο δείγμα 20 παιδιών και μετράει τον αριθμό των χαλασμένων δοντιών στο καθένα. Τα αποτελέσματα έχουν ως εξής:

| **Πλήθος. χαλασμένων δοντιών/παιδί** | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Πλήθος Παιδιών** | 8 | 4 | 2 | 2 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |

***Πίνακας 8.3*** *Δεδομένα για την άσκηση 8.13*

Ο κ. Β, χρησιμοποιώντας τις ίδιες ιατρικές τεχνικές, εξετάζει και τα 200 παιδιά και καταγράφει εκείνα που δεν είχαν κανένα χαλασμένο δόντι. Ο αριθμός των παιδιών αυτών ήταν συνολικά 60. Να εκτιμήσετε τον συνολικό αριθμό των παιδιών του χωριού με χαλασμένα δόντια (α) χρησιμοποιώντας μόνο τα αποτελέσματα του κ. Α, (β) χρησιμοποιώντας τα αποτελέσματα του κ. Α και του κ. Β. (γ) Είναι οι εκτιμητές αυτοί αμερόληπτοι; (δ) Ποιος από τους δύο εκτιμητές περιμένετε να είναι πιο ακριβής και γιατί;

**8.14.** Σε έναν πληθυσμό, είναι γνωστό ότι $\bar{Y}$ = 19 και $S^2 = 85{,}6$. Στην περίπτωση της απλής τυχαίας δειγματοληψίας, πόσο θα πρέπει να είναι το μέγεθος του δείγματος, έτσι ώστε η τιμή του εκτιμητή μας να διαφέρει από την $\bar{Y}$ το πολύ 10%, με πιθανότητα 1 στα 20;

**8.15.** Σε μία περιοχή υπάρχουν 4000 σπίτια. Θέλουμε να εκτιμήσουμε α) το ποσοστό των ατόμων που έχουν ιδιόκτητο σπίτι, με τυπικό σφάλμα όχι μεγαλύτερο του 2% και β) το ποσοστό των οικογενειών που έχουν περισσότερα από δύο αυτοκίνητα, με τυπικό σφάλμα όχι μεγαλύτερο του 1%. Για τον πληθυσμό αυτόν, είναι γνωστό ότι το αληθινό ποσοστό αυτών που έχουν ιδιόκτητο σπίτι βρίσκεται μεταξύ 45 και 65%, ενώ αυτό των οικογενειών με περισσότερα από δύο αυτοκίνητα, μεταξύ 5 και 10%. Αν χρησιμοποιήσουμε απλή τυχαία δειγματοληψία, πόσο μεγάλο θα πρέπει να είναι το δείγμα, ώστε να ικανοποιεί και τις δύο απαιτήσεις;

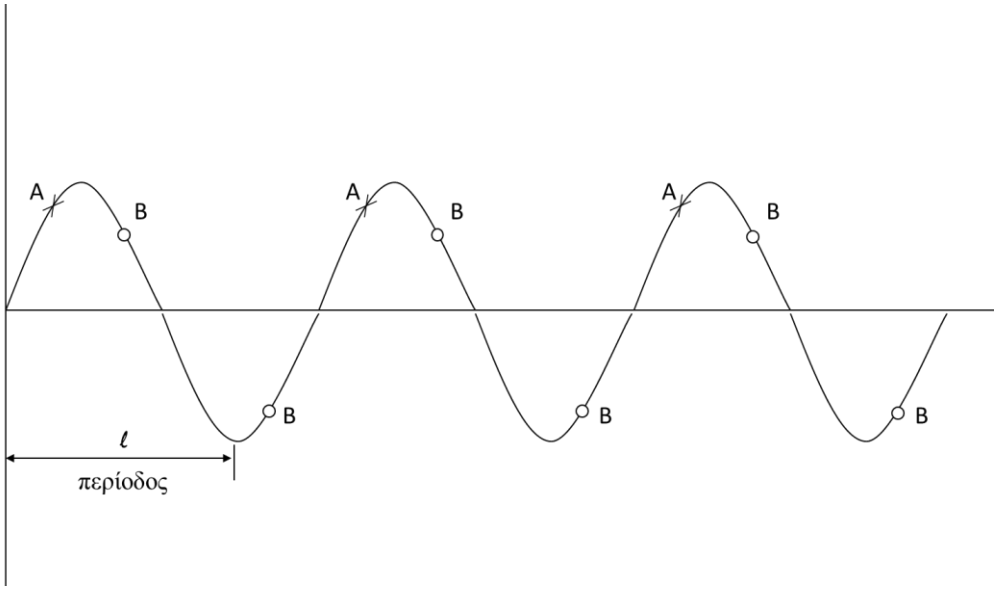
**8.16.** Από έναν κατάλογο με 3042 ονόματα και διευθύνσεις, επιλέξαμε ένα τυχαίο (χωρίς επανατοποθέτηση) δείγμα 200 ονομάτων. Στο δείγμα αυτό, βρέθηκε ότι 38 διευθύνσεις ήταν λάθος. Να εκτιμήσετε τον συνολικό αριθμό των λανθασμένων διευθύνσεων στον κατάλογο αυτό, και να δώσετε και το τυπικό σφάλμα της εκτίμησής σας.

**8.17.** (*Πληθυσμοί με γραμμική τάση*) Ας υποθέσουμε ότι οι διαδοχικές τιμές του υπό μελέτη χαρακτηριστικού *y*, σε έναν πεπερασμένο πληθυσμό μεγέθους *N*, αυξάνονται σύμφωνα με το μοντέλο

$$y_i = \mu + i\theta$$

όπου $\mu$ και $\theta$ είναι σταθερές και το *i* μεταβάλλεται από 1 έως *N*. Να δειχθεί ότι στην περίπτωση αυτή, η συστηματική είναι καλύτερη από την απλή τυχαία δειγματοληψία χωρίς επανατοποθέτηση. [Η έκφραση "καλύτερη" αναφέρεται στις διακυμάνσεις των εκτιμητών].

**8.18.** (*Πληθυσμοί με περιοδικότητα*) Υποθέτουμε ότι οι τιμές ενός πεπερασμένου πληθυσμού παρουσιάζουν περιοδικότητα (π.χ. απλής ημιτονοειδούς μορφής). Ένα γράφημα ενός τέτοιου πληθυσμού είναι το επόμενο.




***Σχήμα 8.1*** *Περιοδική μεταβλητότητα*

Από τον πληθυσμό αυτόν επιλέγουμε δύο συστηματικά δείγματα. Στο πρώτο (η περίπτωση που αντιστοιχεί στο Α) το *k* είναι ίσο με την περίοδο $\ell$ της κύμανσης. Στο δεύτερο (που αντιστοιχεί στο δείγμα Β) το *k* είναι ίσο με το μισό της περιόδου. Να σχολιάσετε ποιο κατά τη γνώμη σας δείγμα είναι καλύτερο και γιατί.

**8.19.** Η διεύθυνση ενός κινηματογράφου θέλει να εκτιμήσει το ποσοστό p των θεατών που είχαν θετικές κρίσεις για μια συγκεκριμένη ταινία. Για το σκοπό αυτό, από το σύνολο των θεατών που παρακολούθησαν τη συγκεκριμένη ταινία στη διάρκεια μιας εβδομάδας, εκλέγει ένα τυχαίο δείγμα κατά την έξοδο των θεατών, με τον εξής τρόπο. Ανά δέκα άτομα, εκλέγει ένα και ρωτάει τη γνώμη του για την ταινία. Αν υποθέσουμε ότι στη διάρκεια μιας εβδομάδας 2.000 άτομα παρακολούθησαν την ταινία, να εκτιμήσετε το ποσοστό των θεατών που έχουν θετικές κρίσεις και να δώσετε ένα 95% διάστημα εμπιστοσύνης. Στη συνέχεια, υπολογίστε το μέγεθος του δείγματος που απαιτείται για την εκτίμηση του p αν το σφάλμα της εκτίμησης δεν πρέπει να υπερβαίνει το 0.01. [Με $y_i = 1$ παριστάνεται η απάντηση του i-οστού θεατή που έχει θετικές κρίσεις για την ταινία και με $y_i = 0$ η απάντηση του i-οστού θεατή που έχει αρνητικές κρίσεις. Δίνεται ότι συνολικά 232 θεατές είχαν θετικές κρίσεις για την ταινία.]

**8.20.** Σκοπός μιας έρευνας ήταν η εκτίμηση του μέσου αριθμού ατόμων ανά αυτοκίνητο που διέρχεται από τον Ισθμό κατά τη διάρκεια μιας ώρας, σε περίοδο αιχμής. Η εκτίμηση αυτή βασίστηκε σε ένα τυχαίο δείγμα αυτοκινήτων και μετρήθηκε ο αριθμός των επιβατών σε κάθε ένα από αυτά. Η λήψη του δείγματος έγινε με επαναλαμβανόμενη συστηματική δειγματοληψία, παίρνοντας 10 δείγματα των 8 αυτοκινήτων το καθένα. Από προηγούμενες χρονιές, είναι γνωστό ότι περίπου 400 αυτοκίνητα διέρχονται από τον Ισθμό μέσα σε 1 ώρα σε περίοδο αιχμής. Με βάση τα ακόλουθα δεδομένα, να εκτιμήσετε τον μέσο αριθμό επιβατών ανά αυτοκίνητο και να δώσετε ένα 95% διάστημα εμπιστοσύνης για την εκτίμησή σας.

| **Δείγμα** | **# επιβατών $y_i$** | | | | | | | | $\bar{y}$ |
|---|---|---|---|---|---|---|---|---|---|
| **1** | 4 | 5 | 3 | 6 | 1 | 4 | 4 | 3 | 3.75 |
| **2** | 5 | 3 | 4 | 2 | 4 | 2 | 3 | 4 | 3.38 |
| **3** | 2 | 4 | 6 | 2 | 3 | 2 | 1 | 3 | 2.88 |
| **4** | 6 | 4 | 6 | 7 | 2 | 3 | 2 | 7 | 4.62 |
| **5** | 4 | 5 | 7 | 4 | 2 | 6 | 2 | 6 | 4.50 |
| **6** | 7 | 6 | 4 | 4 | 3 | 6 | 7 | 5 | 5.25 |
| **7** | 3 | 3 | 2 | 3 | 6 | 5 | 6 | 8 | 4.50 |
| **8** | 2 | 6 | 2 | 5 | 5 | 4 | 4 | 5 | 4.12 |
| **9** | 2 | 6 | 3 | 6 | 4 | 4 | 5 | 4 | 4.25 |
| **10** | 6 | 5 | 4 | 6 | 3 | 3 | 5 | 3 | 4.38 |

***Πίνακας 8.4*** *Δεδομένα για την άσκηση 8.20*

**8.21.** Σε ένα εργοστάσιο που παράγει βίδες, χρησιμοποιείται συστηματική δειγματοληψία για την εκτίμηση του μέσου μήκους τους. Για τον σκοπό αυτό, λαμβάνεται 1 βίδα κάθε 50 βίδες που παράγονται και μετριέται το μήκος της. Αν σε 1 ώρα παράγονται 1800 βίδες, εκτιμήστε το μέσο μήκος τους χρησιμοποιώντας τα δεδομένα του ακόλουθου Πίνακα, και δώστε ένα 99% διάστημα εμπιστοσύνης για την εκτίμηση. Τι μέγεθος δείγματος απαιτείται για την εκτίμηση του μέσου, όταν το σφάλμα είναι 0.03;

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 12,00 | 11,97 | 12,01 | 12,01 | 11,80 | 12,03 | 11,91 | 11,98 | 12,03 | 11,98 | 12,00 | 11,83 |
| 11,87 | 12,01 | 11,98 | 11,87 | 11,90 | 11,88 | 12,05 | 11,87 | 11,91 | 11,93 | 11,94 | 11,89 |
| 11,72 | 11,93 | 11,95 | 11,97 | 11,93 | 12,05 | 11,85 | 11,98 | 11,87 | 12,05 | 12,02 | 12,04 |

***Πίνακας 8.5*** *Δεδομένα για την άσκηση 8.21 - Μήκος Βιδών*

**8.22.** Σε έρευνα που έγινε για την εκτίμηση του ολικού αριθμού τουριστών που επισκέπτονται ένα αρχαιολογικό μουσείο σε μια περίοδο 180 ημερών, χρησιμοποιήθηκε συστηματική δειγματοληψία με περίοδο δειγματοληψίας τις 10 μέρες. 0 αριθμός των ξένων επισκεπτών κάθε δέκατη μέρα, με τυχαίο ξεκίνημα την τρίτη μέρα, δίνεται στον παρακάτω πίνακα. Ζητείται να εκτιμηθεί ο ολικός αριθμός τουριστών που επισκέφθηκαν το αρχαιολογικό μουσείο στη συγκεκριμένη χρονική περίοδο, και να δοθεί ένα 95% διάστημα εμπιστοσύνης για την εκτίμηση αυτή.

| Ημέρα | # τουριστών | Ημέρα | # τουριστών |
|---|---|---|---|
| 3 | 160 | 93 | 410 |
| 13 | 350 | 103 | 270 |
| 23 | 245 | 113 | 330 |
| 33 | 225 | 123 | 500 |
| 43 | 180 | 133 | 120 |
| 53 | 270 | 143 | 110 |
| 63 | 340 | 153 | 150 |
| 73 | 420 | 163 | 280 |
| 83 | 210 | 173 | 290 |

***Πίνακας 8.6*** *Δεδομένα για την άσκηση 8.22*

**8.23.** Στον παρακάτω πίνακα, δίνεται ο αριθμός των χαμένων ημερών εργασίας από ασθένεια, για τους 162 υπαλλήλους μιας εταιρείας. Η εταιρεία θέλει να γνωρίζει τον μέσο αριθμό των ημερών άδειας λόγω ασθένειας. Για τον σκοπό αυτό, αποφασίζει να πάρει ένα τυχαίο συστηματικό δείγμα μεγέθους $n = 18$. Με βάση ένα τέτοιο δείγμα, να εκτιμήσετε τον μέσο αριθμό των χαμένων ημερών εργασίας λόγω ασθένειας, και να δώσετε ένα 95% διάστημα εμπιστοσύνης για τον εκτιμητή αυτόν.

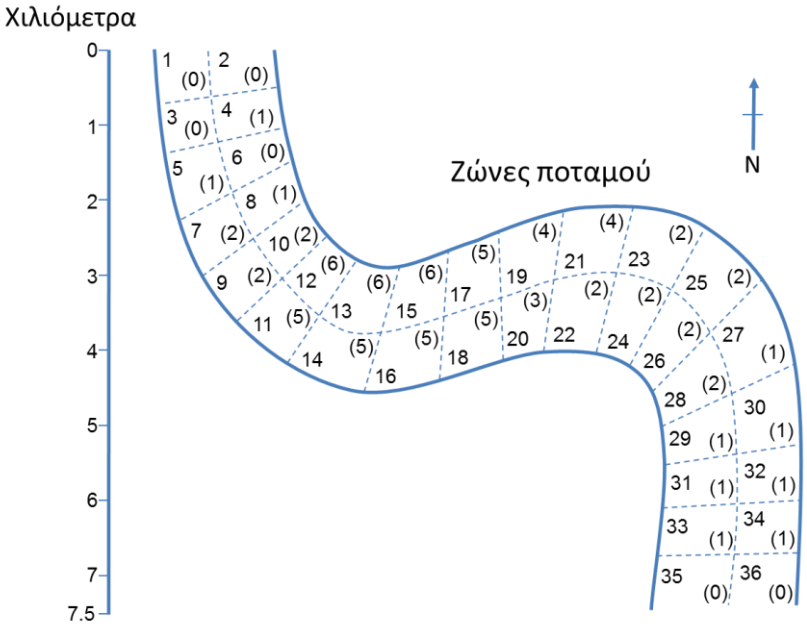
| Κωδ. Εργ. | Ημέρ. Άδειας | Κωδ. Εργ. | Ημέρ. Άδειας | Κωδ. Εργ. | Ημέρ. Άδειας | Κωδ. Εργ. | Ημέρ. Άδειας |
|---|---|---|---|---|---|---|---|
| 1 | 7 | 42 | 5 | 83 | 3 | 123 | 6 |
| 2 | 6 | 43 | 3 | 84 | 5 | 124 | 3 |
| 3 | 10 | 44 | 6 | 85 | 4 | 125 | 9 |
| 4 | 11 | 45 | 11 | 86 | 0 | 126 | 9 |
| 5 | 3 | 46 | 6 | 87 | 11 | 127 | 6 |
| 6 | 8 | 47 | 5 | 88 | 3 | 128 | 5 |
| 7 | 0 | 48 | 5 | 89 | 4 | 129 | 4 |
| 8 | 5 | 49 | 0 | 90 | 11 | 130 | 1 |
| 9 | 8 | 50 | 8 | 91 | 0 | 131 | 1 |
| 10 | 4 | 51 | 1 | 92 | 6 | 132 | 11 |
| 11 | 7 | 52 | 10 | 93 | 1 | 133 | 3 |
| 12 | 13 | 53 | 7 | 94 | 9 | 134 | 5 |
| 13 | 4 | 54 | 9 | 95 | 6 | 135 | 9 |
| 14 | 5 | 55 | 8 | 96 | 0 | 136 | 5 |
| 15 | 2 | 56 | 2 | 97 | 3 | 137 | 1 |

| Κωδ. Εργ. | Ημέρ. Άδειας | Κωδ. Εργ. | Ημέρ. Άδειας | Κωδ. Εργ. | Ημέρ. Άδειας | Κωδ. Εργ. | Ημέρ. Άδειας |
|---|---|---|---|---|---|---|---|
| 16 | 0 | 57 | 9 | 98 | 6 | 138 | 15 |
| 17 | 7 | 58 | 9 | 99 | 0 | 139 | 2 |
| 18 | 17 | 59 | 8 | 100 | 12 | 140 | 10 |
| 19 | 5 | 60 | 6 | 101 | 11 | 141 | 8 |
| 20 | 6 | 61 | 5 | 102 | 6 | 142 | 2 |
| 21 | 1 | 62 | 3 | 103 | 1 | 143 | 6 |
| 22 | 7 | 63 | 9 | 104 | 3 | 144 | 14 |
| 23 | 9 | 64 | 6 | 105 | 2 | 145 | 10 |
| 24 | 3 | 65 | 3 | 106 | 5 | 146 | 8 |
| 25 | 8 | 66 | 3 | 107 | 3 | 147 | 7 |
| 26 | 9 | 67 | 4 | 108 | 12 | 148 | 9 |
| 27 | 4 | 68 | 9 | 109 | 1 | 149 | 1 |
| 28 | 8 | 69 | 5 | 110 | 7 | 150 | 2 |
| 29 | 4 | 70 | 8 | 111 | 9 | 151 | 6 |
| 30 | 17 | 71 | 5 | 112 | 6 | 152 | 4 |
| 31 | 6 | 72 | 11 | 113 | 6 | 153 | 6 |
| 32 | 9 | 73 | 5 | 114 | 3 | 154 | 3 |
| 33 | 9 | 74 | 9 | 115 | 4 | 155 | 1 |
| 34 | 5 | 75 | 8 | 116 | 2 | 156 | 8 |
| 35 | 8 | 76 | 7 | 117 | 5 | 157 | 0 |
| 36 | 5 | 77 | 6 | 118 | 10 | 158 | 3 |
| 37 | 8 | 78 | 4 | 119 | 10 | 159 | 2 |
| 38 | 5 | 79 | 3 | 120 | 15 | 160 | 8 |
| 39 | 8 | 80 | 9 | 121 | 5 | 161 | 0 |
| 40 | 0 | 81 | 5 | 122 | 5 | 162 | 15 |
| 41 | 3 | 82 | 5 | | | | |

***Πίνακας 8.7*** *Δεδομένα για την άσκηση 8.23*

**8.24.** Ας υποθέσουμε ότι πρόκειται να κάνουμε μια μελέτη για τον προσδιορισμό της περιεκτικότητας των νερών ενός ποταμού σε μια καρκινογόνο ουσία. Ο έλεγχος αυτός αφορά ένα τμήμα του ποταμού, που φαίνεται στο επόμενο σχήμα. Για τον σκοπό της μελέτης, χωρίσαμε το τμήμα αυτό του ποταμού σε 36 ζώνες, όπως παρουσιάζεται στο σχήμα, και στη συνέχεια επιλέξαμε ένα τυχαίο συστηματικό δείγμα, με

βήμα 1 στα 4. Στις επιλεγμένες ζώνες μετρήσαμε την ποσότητα (σε χιλιοστόγραμμα ανά λίτρο) της καρκινογόνου ουσίας. (Οι τιμές αυτές, για όλες τις ζώνες, δίνονται στο σχήμα εντός των παρενθέσεων).




***Σχήμα 8.2*** *Δεδομένα για την άσκηση 8.24*

Να εκτιμηθεί η μέση περιεκτικότητα των νερών του ποταμού στη συγκεκριμένη καρκινογόνο ουσία και να δοθεί ένα 90% διάστημα εμπιστοσύνης για την τιμή αυτή.

**8.25.** Μια διαφημιστική εταιρεία θέλει να διαφημίσει ένα προϊόν στην τηλεόραση. Για τον σκοπό αυτό, θέλει να εκτιμήσει πόσες ώρες την εβδομάδα οι κάτοικοι μιας πόλης παρακολουθούν τηλεόραση. Για τη διεξαγωγή της έρευνας, αποφασίζει να χρησιμοποιήσει στρωματοποιημένη δειγματοληψία. Έτσι, η πόλη χωρίζεται σε τρία στρώματα: το κέντρο (Κ), τα κοντινά προάστια (ΚΠ) και τα απομακρυσμένα προάστια (ΑΠ). Στους καταλόγους είναι εγγεγραμμένες 155 οικογένειες στο στρώμα Κ, 62 στο στρώμα ΚΠ και 93 στο στρώμα ΑΠ. Από τα τρία αυτά στρώματα, επιλέγουμε στην τύχη τυχαία δείγματα 8, 4 και 6 οικογενειών, αντίστοιχα. Οι ώρες παρακολούθησης τηλεόρασης, σε εβδομαδιαία βάση, για κάθε οικογένεια στο δείγμα, είναι

| | |
|---|---|
| Κέντρο (Κ): | 10, 20, 12, 16, 30, 13, 4, 9 |
| Κοντινά προάστια (ΚΠ): | 36, 12, 1, 10 |
| Απομακρυσμένα προάστια (ΑΠ): | 22, 12, 2, 31, 21, 14 |

***Πίνακας 8.8*** *Δεδομένα για την άσκηση 8.25*

Να εκτιμηθεί ο μέσος χρόνος παρακολούθησης τηλεόρασης των κατοίκων της συγκεκριμένης πόλης και να δοθεί το τυπικό σφάλμα εκτίμησης.

**8.26.** Με βάση τα δεδομένα της προηγούμενης άσκησης, να εκτιμηθεί το ποσοστό των κατοίκων της πόλης αυτής που παρακολουθούν τηλεόραση περισσότερο από 15 ώρες εβδομαδιαίως.

**8.27.** Να βρεθούν τα 95% διαστήματα εμπιστοσύνης α) για τον μέσο εβδομαδιαίο χρόνο παρακολούθησης τηλεόρασης στην άσκηση 4.1 και β) για το ποσοστό των κατοίκων της πόλης που παρακολουθούν τηλεόραση περισσότερο από 15 ώρες στην άσκηση 4.2.

**8.28.** Χωρίζουμε τον πληθυσμό μιας πόλης σε τρία στρώματα, με βάση το εισόδημα. Στον πίνακα που ακολουθεί, $\bar{Y}_h$ και $S_h^2$ είναι, αντίστοιχα, το μέσο εισόδημα (σε χιλιάδες ευρώ) και η διακύμανση του $h$ στρώματος ($h$ = 1,2,3). Ακόμη, $W_h$ είναι το βάρος κάθε στρώματος και $P_h$ το ποσοστό των ατόμων, στο $h$ στρώμα, με εισόδημα μικρότερο από 3000€.

| Στρώμα | $W_h$ | $\bar{Y}_h$ | $S_h^2$ | $P_h$ |
|---|---|---|---|---|
| **1** | 0,35 | 3,1 | 4 | 0,54 |
| **2** | 0,55 | 3,9 | 11 | 0,39 |
| **3** | 0,10 | 7,8 | 128 | 0,24 |
| **Για ολόκληρο τον πληθυσμό** | 1,00 | 4 | 21,98 | 0,43 |

***Πίνακας 8.9*** *Δεδομένα για την άσκηση 8.28*

α) Χρησιμοποιώντας απλή τυχαία δειγματοληψία με μέγεθος δείγματος $n$ = 1000, να εκτιμήσετε το μέσο εισόδημα του πληθυσμού και το ποσοστό των ατόμων με εισόδημα μικρότερο από 3000 ευρώ. Για κάθε σας εκτίμηση, να δώσετε και την εκτιμώμενη τυπική της απόκλιση.

β) Να εκτιμηθούν οι ίδιες ποσότητες του ερωτήματος (α), αν χρησιμοποιήσουμε στρωματοποιημένη τυχαία δειγματοληψία με αναλογικό καταμερισμό και $n$ = 1000. Στην περίπτωση αυτή, ποιο είναι το σχετικό κέρδος ως προς την περίπτωση (α);

γ) Να βρεθεί ο βέλτιστος καταμερισμός του δείγματος $n$ = 1000 στα τρία στρώματα, έτσι ώστε να ελαχιστοποιείται η διακύμανση του εκτιμητή του μέσου εισοδήματος του πληθυσμού. Για τον συγκεκριμένο καταμερισμό του δείγματος, να βρείτε τις διακυμάνσεις των εκτιμητών στο ερώτημα (β) και να τις συγκρίνετε με αυτές στα ερωτήματα (α) και (β). [Να αγνοηθεί η διόρθωση του πεπερασμένου πληθυσμού].

**8.29.** Στην άσκηση 8.25, υποθέτουμε ότι οι διακυμάνσεις σε κάθε στρώμα είναι γνωστές και ίσες με $S_K^2$ = 25, $S_{K\Pi}^2$ = 225, $S_{A\Pi}^2$ = 100. Να βρεθούν το ολικό μέγεθος του δείγματος και ο καταμερισμός του, έτσι ώστε με ελάχιστο κόστος να εκτιμήσουμε τον πληθυσμιακό μέσο χρόνο παρακολούθησης τηλεόρασης (σε ώρες) στην πόλη αυτή, δοθέντος ότι η διακύμανση του εκτιμητή μας θέλουμε να είναι ίση με τη μονάδα.

**8.30.** Ένας δειγματολήπτης έχει δύο στρώματα και θα ήθελε, για λόγους ευκολίας, να έχει *n1 = n2*, αντί να χρησιμοποιήσει τον καταμερισμό Neyman. Αν με $\text{Var}(\bar{y}_{st})$ και $\text{Var}_{\text{Neyman}}(\bar{y}_{st})$ συμβολίσουμε τις διακυμάνσεις του εκτιμητή μας στην περίπτωση όπου *n1 = n2* και στον καταμερισμό Neyman, αντίστοιχα, να δειχθεί ότι η κλασματική αύξηση της διακύμανσης δίνεται από τη σχέση

$$\frac{\text{Var}(\bar{y}_{st}) - \text{Var}_{\text{Neyman}}(\bar{y}_{st})}{\text{Var}_{\text{Neyman}}(\bar{y}_{st})} = \left(\frac{r-1}{r+1}\right)^2,$$

όπου $r = n1/n2$.

**8.31.** Τα αποτελέσματα από μία απλή τυχαία δειγματοληψία χωρίς επανατοποθέτηση με $n$ = 1000 μπορούν να ταξινομηθούν σε τρία στρώματα, με $\bar{y}_h$ = 10,2, 12,6 και 17,1, $s_h^2$ =10,82 (το ίδιο και για τα τρία στρώματα) και $s^2$ = 17,66. Οι εκτιμητές των $W_h$ είναι οι $w_h$ =0,5, 0,3 και 0,2. Είναι γνωστό ότι τα βάρη αυτά δεν είναι ακριβή, αλλά πιστεύεται ότι είναι σωστά με ένα σφάλμα 5%. Αυτό σημαίνει ότι στη χειρότερη περίπτωση, τα πραγματικά βάρη θα είναι είτε $W_h$ = 0,525, 0,285 και 0,190 ή $W_h$ = 0,475, 0,285 και 0,210. Με βάση τα στοιχεία αυτά, θα προτείνατε τη στρωματοποιημένη δειγματοληψία έναντι της απλής τυχαίας; [Όπου απαιτείται, θεωρούμε ότι $\bar{y}_h = \bar{Y}_h$ και $s_h^2 = S_h^2$].

**8.32.** Τα δεδομένα που ακολουθούν προέκυψαν από μία στρωματοποιημένη δειγματοληψία με τέσσερα στρώματα. Τα $\bar{y}_h$ είναι οι μέσες τιμές τού υπό μελέτη χαρακτηριστικού σε κάθε στρώμα.

| Στρώμα | $N_h$ | $W_h$ | $\bar{y}_h$ | $s_h^2$ | $n_h$ |
|---|---|---|---|---|---|
| **1** | 19.850 | 0,8032 | 4,1 | 34,8 | 3.000 |
| **2** | 3.250 | 0,1315 | 13,0 | 92,2 | 600 |
| **3** | 1.007 | 0,0407 | 25,0 | 174,2 | 340 |
| **4** | 606 | 0,0245 | 38,2 | 320,4 | 230 |
| **Σύνολα** | **24.713** | **0,9999** | | | **4.170** |

***Πίνακας 8.10*** *Δεδομένα για την άσκηση 8.32*

α) Να εκτιμήσετε το κέρδος στην ακρίβεια που οφείλεται στη στρωματοποίηση.

β) Να συγκρίνετε αυτό το αποτέλεσμα με το κέρδος στην ακρίβεια την οποία θα πετυχαίναμε αν χρησιμοποιούσαμε αναλογικό καταμερισμό.

**8.33.** Ένας ερευνητής ενδιαφέρεται να εκτιμήσει τον μέσο αριθμό των εφημερίδων που αγοράζουν οι οικογένειες σε μια κοινότητα 4.000 κατοίκων. Η κοινότητα χωρίστηκε σε 400 συστάδες με 10 οικογένειες η κάθε μία. Από τις συστάδες αυτές, επιλέξαμε στην τύχη και χωρίς επανατοποθέτηση τέσσερεις συστάδες και πήραμε τα αποτελέσματα που δίνονται στον επόμενο πίνακα.

| Συστάδα | Αριθμός εφημερίδων που αγόρασε η οικογένεια | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **1** | 1 | 2 | 1 | 3 | 3 | 2 | 1 | 4 | 1 | 1 |
| **2** | 1 | 3 | 2 | 2 | 3 | 1 | 4 | 1 | 1 | 2 |
| **3** | 2 | 1 | 1 | 1 | 1 | 3 | 2 | 1 | 3 | 1 |
| **4** | 1 | 1 | 3 | 2 | 1 | 5 | 1 | 2 | 3 | 1 |

***Πίνακας 8.11*** *Δεδομένα για την άσκηση 8.33*

Με βάση τα στοιχεία αυτά, να εκτιμηθεί ο μέσος αριθμός των εφημερίδων ανά οικογένεια.

**8.34.** Μια εταιρεία, η οποία διαθέτει υπηρεσιακά αυτοκίνητα στους υπαλλήλους της, θέλει να εκτιμήσει τον μέσο αριθμό χιλιομέτρων τα οποία έχουν διανυθεί από κάθε αυτοκίνητο. Η εταιρεία έχει 12 υποκαταστήματα και ο αριθμός των αυτοκινήτων $N_i$, καθώς και οι μέσες τιμές και διακυμάνσεις ($\bar{Y}_i$ και $S_i^2$) του αριθμού των διανυθέντων χιλιομέτρων τον τελευταίο χρόνο (σε χιλιάδες χιλιόμετρα) σε κάθε υποκατάστημα, έχουν ως εξής:

| Υποκατάστημα | $N_i$ | $\bar{Y}_i$ | $S_i^2$ |
|---|---|---|---|
| 1 | 6 | 24,32 | 5,07 |
| 2 | 2 | 27,06 | 5,53 |
| 3 | 11 | 27,6 | 6,24 |
| 4 | 7 | 28,01 | 6,59 |
| 5 | 8 | 27,56 | 6,21 |

| Υποκατάστημα | $N_i$ | $\overline{Y}_i$ | $S_i^2$ |
|---|---|---|---|
| 6 | 14 | 29,07 | 6,12 |
| 7 | 6 | 32,03 | 5,97 |
| 8 | 2 | 28,41 | 6,01 |
| 9 | 2 | 28,91 | 5,74 |
| 10 | 5 | 25,55 | 6,78 |
| 11 | 12 | 28,58 | 5,87 |
| 12 | 6 | 27,27 | 5,38 |

***Πίνακας 8.12*** *Δεδομένα για την άσκηση 8.34*

Επιλέγοντας τυχαία τέσσερα (4) από τα παραπάνω υποκαταστήματα και παίρνοντας όλα τα αυτοκίνητα σε αυτά, να εκτιμήσετε τον μέσο αριθμό των διανυθέντων χιλιομέτρων, χρησιμοποιώντας ως εκτιμητή λόγου $\hat{\overline{Y}}_{cl,r}$. Στη συνέχεια, να επιλέξετε ένα τυχαίο δείγμα μεγέθους 27 (τόσος είναι ο αναμενόμενος συνολικός αριθμός των αυτοκινήτων για τα τέσσερα υποκαταστήματα που επιλέξαμε) και να υπολογίσετε την αποτελεσματικότητα της απλής τυχαίας δειγματοληψίας ως προς αυτή της κατά συστάδες.

**8.35.** Σκοπός μιας έρευνας είναι να εκτιμήσει το συνολικό μηνιαίο εισόδημα των οικογενειών μιας πόλης. Είναι γνωστό ότι η πόλη αυτή αποτελείται από 415 οικοδομικά τετράγωνα. Θεωρώντας τα οικοδομικά τετράγωνα ως συστάδες, επιλέγουμε στην τύχη και χωρίς επανατοποθέτηση 25 συστάδες. Σε κάθε συστάδα, ζητούμε από όλα τα ενήλικα άτομα να μας πουν: 1ον) το συνολικό εισόδημά τους (σε χρηματικές μονάδες, χ.μ.) και 2ον) αν είναι ενοικιαστές ή όχι. Τα αποτελέσματα της έρευνας δίνονται στον επόμενο πίνακα. Εάν είναι γνωστό ότι στην πόλη αυτή υπάρχουν 2500 ενήλικες, να εκτιμηθεί α) το μέσο μηνιαίο εισόδημα, καθώς και β) το συνολικό μηνιαίο εισόδημα των κατοίκων και γ) το ποσοστό των ενοικιαστών.

| Συστάδα $i$ | Αριθμός Ενηλίκων $N_i$ | Ολικό Εισόδημα $y_i$ | Αριθμός Ενοικιαστών $M_i$ |
|---|---|---|---|
| 1 | 8 | 96.000 | 4 |
| 2 | 12 | 121.000 | 7 |
| 3 | 4 | 42.000 | 1 |
| 4 | 5 | 65.000 | 3 |
| 5 | 6 | 52.000 | 3 |
| 6 | 6 | 40.000 | 4 |
| 7 | 7 | 75.000 | 4 |
| 8 | 5 | 65.000 | 2 |
| 9 | 8 | 45.000 | 3 |
| 10 | 3 | 50.000 | 2 |
| 11 | 2 | 85.000 | 1 |

| Συστάδα $i$ | Αριθμός Ενηλίκων $N_i$ | Ολικό Εισόδημα $y_i$ | Αριθμός Ενοικιαστών $M_i$ |
|---|---|---|---|
| 12 | 6 | 43.000 | 3 |
| 13 | 5 | 54.000 | 2 |
| 14 | 10 | 49.000 | 5 |
| 15 | 9 | 53.000 | 4 |
| 16 | 3 | 50.000 | 1 |
| 17 | 6 | 32.000 | 4 |
| 18 | 5 | 22.000 | 2 |
| 19 | 5 | 45.000 | 3 |
| 20 | 4 | 37.000 | 1 |
| 21 | 6 | 51.000 | 3 |
| 22 | 8 | 30.000 | 3 |
| 23 | 7 | 39.000 | 4 |
| 24 | 3 | 47.000 | 0 |
| 25 | 8 | 41.000 | 3 |
| **Σύνολο** | **151** | **1.329.000** | **72** |

***Πίνακας 8.13*** *Δεδομένα για την άσκηση 8.35*

**8.36.** Ένας βιοτέχνης παιδικών ενδυμάτων θέλει να εκτιμήσει την αξία των εμπορευμάτων που βρίσκονται στην αποθήκη του. Η αποθήκη του έχει 48 ράφια. Από τα ράφια αυτά, εκλέγει στην τύχη και χωρίς επανατοποθέτηση, 10 ράφια και μετράει το πλήθος των ενδυμάτων ($N_i$) και τη συνολική τους αξία ($y_i$) σε χιλιάδες ευρώ. Με βάση τα αποτελέσματα στον παρακάτω πίνακα, να εκτιμηθεί η συνολική αξία των εμπορευμάτων του βιοτέχνη.

| **Ράφι** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|---|
| $N_i$ | 42 | 27 | 38 | 63 | 72 | 12 | 24 | 14 | 32 | 41 |
| $y_i$ | 83 | 62 | 45 | 112 | 96 | 58 | 75 | 58 | 67 | 80 |

***Πίνακας 8.14*** *Δεδομένα για την άσκηση 8.36*

# Παράρτημα Ι - ΕΙΣΑΓΩΓΗ ΣΤΗΝ R

Η R είναι μια ολοκληρωμένη σουίτα για διαχείριση δεδομένων, στατιστικούς υπολογισμούς και δημιουργία γραφημάτων. Αποτελεί μέρος του GNU *project* και είναι παρόμοια της σουίτας S, η οποία έχει αναπτυχθεί στα Bell Laboratories (πρώην AT&T, νυν Lucent Technologies) από τους J. Chambers και συνεργάτες. Η σουίτα R περιλαμβάνει τη γλώσσα προγραμματισμού R και το σχετικό γραφικό περιβάλλον και θεωρείται μια εναλλακτική υλοποίηση της S. Οι περισσότερες ρουτίνες της S είναι εκτελέσιμες και στην R, παρά τις αρκετές διαφορές μεταξύ των δύο.

Η R είναι μια εφαρμογή Ανοικτού Κώδικα (*Open Source*), διαθέσιμη ως Ελεύθερο Λογισμικό (*Free Software*), δίχως κόστος δικαιωμάτων κτήσης ή χρήσης και διέπεται από τους όρους του *Free Software Foundation GNU General Public License*. Είναι εκτελέσιμη σε ένα μεγάλο εύρος από UNIX πλατφόρμες και παρόμοια συστήματα (όπως FreeBSD και Linux), σε Windows, αλλά και σε Mac OS.

Η R περιλαμβάνει:

- αποτελεσματικό σύστημα διαχείρισης και αποθήκευσης δεδομένων,
- πλήρες σύνολο λειτουργιών για υπολογισμούς με πολυδιάστατα δεδομένα (`array, matrix, vector`),
- διευρυμένο και συνεκτικό σύνολο ενδιάμεσων εργαλείων για ανάλυση δεδομένων,
- λειτουργίες γραφικών για *on-screen* εμφάνιση, αποθήκευση ή εκτύπωση και
- καλώς ορισμένη, απλή και αποτελεσματική γλώσσα προγραμματισμού για *user-defined functions*.

Η R είναι σχεδιασμένη στη βάση μιας γλώσσας προγραμματισμού, γεγονός που δίνει μεγάλη επεκτασιμότητα, καθώς επιτρέπει στους χρήστες να προσθέσουν επιπλέον λειτουργικότητα με τη δημιουργία νέων, σύγχρονων εργαλείων (*packages*). Μεγάλο μέρος του ίδιου του συστήματος είναι δομημένο πάνω στη γλώσσα R ("διάλεκτο" της S), γεγονός που διευκολύνει τους χρήστες στην κατανόηση και την επέκτασή της. Επιπλέον, για εργασίες με μεγάλες υπολογιστικές απαιτήσεις, μπορούν να χρησιμοποιηθούν άλλες γλώσσες, όπως C, C++ και Fortran, για τη βελτίωση των επιδόσεών της. Τέλος, οι χρήστες μπορούν να γράψουν κώδικα στη C και να χειρίζονται αντικείμενα R απευθείας.

## Εκδόσεις R

Οι λειτουργικές εκδόσεις της R, η οποία υποστηρίζεται από μια ομάδα προγραμματιστών ανά τον κόσμο, αριθμούν κάποιες δεκάδες, λόγω του ότι συνηθίζεται η επίσημη έκδοση της R να ανανεώνεται δύο φορές το χρόνο (Απρίλιο και Οκτώβριο - τρέχουσα έκδοση 3.2.2, 2015-08-14, Fire Safety). Ωστόσο, η R δεν έχει αλλάξει δραματικά τα τελευταία χρόνια, δεδομένου ότι μια νέα έκδοση περιλαμβάνει μικρές διορθώσεις σφαλμάτων (*bugs*), καθώς και μερικές βελτιώσεις και ενδεχόμενα κάποιες προσθήκες νέων λειτουργιών. Έχουν γίνει και ορισμένες αλλαγές στη γλώσσα προγραμματισμού, αλλά οι περισσότερες από αυτές είναι σε λειτουργίες που δεν επηρεάζουν τον μέσο χρήστη. Προφανώς, δεν είναι ανάγκη κάποιος να χρησιμοποιεί την έκδοση της R η οποία χρησιμοποιήθηκε κατά τη συγγραφή του συγκεκριμένου βιβλίου, καθώς τα αποτελέσματα που παράγονται από διαφορετικές εκδόσεις θα πρέπει να είναι, αν όχι ταυτόσημα, πάρα πολύ παρόμοια.

## Λήψη και Εγκατάσταση της R

Η R είναι δυνατόν να εγκατασταθεί σε όλες τις (κύριες) πλατφόρμες ηλεκτρονικών υπολογιστών. Για τους χρήστες λειτουργικού συστήματος Mac ή Windows, η λήψη και εγκατάσταση της R είναι μια απλοποιημένη, τυπική διαδικασία εγκατάστασης λογισμικού. Για τους χρήστες Linux/Unix, είναι απαραίτητα τα δικαιώματα *sudo* (ή ο *root* κωδικός) και συνιστάται η χρήση συστημάτων *port management*, όπως το Yum, για την απλοποίηση των διαδικασιών εγκατάστασης και αναβάθμισης.

Η λήψη των απαραίτητων αρχείων, σε κάθε περίπτωση, γίνεται μέσω της ιστοσελίδας, http://www.r-project.org, από το σύνδεσμο "CRAN" (κάτω από το "Download"). Ο σύνδεσμος "CRAN" οδηγεί σε μια λίστα ιστοτόπων, ταξινομημένους κατά χώρα, από τους οποίους επιλέγεται ο πλησιέστερος γεωγραφικά για την ταχύτερη λήψη. Στη νέα σελίδα που ανοίγει, μπορεί κανείς να επιλέξει για λήψη εκτελέσιμο αρχείο που αντιστοιχεί στο λειτουργικό σύστημα που τον ενδιαφέρει. Όταν το αρχείο αποθηκευτεί τοπικά, απλή εκτέλεσή του θα ανοίξει έναν οδηγό εγκατάστασης, ο οποίος θα καθοδηγήσει τον χρήστη στην ολοκλήρωση της εγκατάστασης.

Υπάρχουν ορισμένες λεπτομέρειες που πρέπει να λάβει κανείς υπόψη, ανάλογα με το λειτουργικό σύστημα, για παράδειγμα προβλήματα δικαιωμάτων σε Windows Vista και XP, ή ακόμα συμβατότητα νέων εκδόσεων R με παλαιότερες εκδόσεις Mac OS. Πιθανές απαντήσεις σε τέτοια προβλήματα βρίσκονται αναρτημένες στη διεύθυνση https://cran.r-project.org/faqs.html

## Εισαγωγή Δεδομένων στην R με πληκτρολόγηση

Στην περίπτωση μικρών σε μέγεθος συνόλων δεδομένων, η πληκτρολόγησή τους μπορεί να γίνει κατευθείαν και μάλιστα η R παρέχει περισσότερους από έναν τρόπους για να επιτευχθεί κάτι τέτοιο.

### Εισαγωγή Δεδομένων με εντολές R

Η εντολή c χρησιμοποιείται κατά κόρον για τη δημιουργία διανυσμάτων (vector),

```
> age <- c(25, 27, 26, 29, 31)
```

Εναλλακτικά, μπορεί να χρησιμοποιηθεί η εντολή scan,

```
> salary <- scan()
1: 18700000 14626720 14137500 13980000 12916666
6:
```

Συνηθίζεται τα διανύσματα ενός *data set* να εντάσσονται σε ένα data frame,

```
> top.salaries <- data.frame(age, salary)
> top.salaries
  age   salary
1  25 18700000
2  27 14626720
3  26 14137500
4  29 13980000
5  31 12916666
```

### Εισαγωγή Δεδομένων με το Edit GUI

Η εισαγωγή δεδομένων σε διανύσματα δεν είναι και τόσο εύχρηστη όταν αυτά προέρχονται από το πρωτογενές υλικό. Η R, μέσω ενός επεξεργαστή (*data editor*), δίνει τη δυνατότητα εισαγωγής (και κυρίως επεξεργασίας) δεδομένων πολλών μεταβλητών. Με την εντολή edit, ο *data editor* ανοίγει για εισαγωγή

(ή/και επεξεργασία) δεδομένων. Για παράδειγμα, για να εισαγάγει κανείς δεδομένα, μπορεί να χρησιμοποιήσει την εντολή:

```
> new.dt <- edit(data.frame())
```

ή για να επεξεργαστεί κανείς το *top.salaries* `data frame`:

```
> top.salaries <- edit(top.salaries)
```

Να σημειωθεί ότι θα πρέπει να γίνει εκχώρηση των αποτελεσμάτων (*output*) της εντολής `edit` σε ένα αντικείμενο, διαφορετικά, η όποια επεξεργασία γίνει θα χαθεί. Για την αποφυγή απωλειών επεξεργασίας, συστήνεται η χρήση της εντολής `fix`, η οποία αποθηκεύει τις όποιες επεξεργασίες στο ίδιο αντικείμενο δίχως την ανάγκη εκχώρησης:

```
> fix(top.salaries)
```

Ένας εναλλακτικός τρόπος χρήσης της εντολής `fix` για τα Microsoft Windows γίνεται μέσω της παραθυριακής επιλογής "Data Editor..." από το βασικό μενού "Edit". Μεταξύ λειτουργικών συστημάτων, υπάρχουν μικρές διαφοροποιήσεις στην εμφάνιση και στις λειτουργίες του *data editor*, ο οποίος, γενικά, δεν συνιστάται για εργασίες με πολλά δεδομένα.

## Εισαγωγή Δεδομένων από Εξωτερικά Αρχεία

Ένα από τα πλεονεκτήματα της `R` είναι η ευκολία με την οποία εισάγονται δεδομένα από διαφορετικά προγράμματα. Η `R` μπορεί να προσπελάσει δεδομένα από αρχεία κειμένου (*text*), από άλλα στατιστικά προγράμματα, ακόμα και από υπολογιστικά φύλλα (*spreadsheets*). Δεν χρειάζεται καν ένα αντίγραφο του αρχείου τοπικά, απλά μπορεί κανείς να καθορίσει ένα αρχείο σε μια URL διεύθυνση, και η `R` θα προσπελάσει το αρχείο διαδικτυακά.

### Text Files

Τα περισσότερα *text* αρχεία δεδομένων έχουν παρόμοια μορφοποίηση, όπου κάθε γραμμή αντιστοιχεί σε μία παρατήρηση (εγγραφή) και ένα σύνολο στηλών που αντιστοιχούν στις μεταβλητές. Οι γραμμές διακρίνονται με το χαρακτήρα "Enter", ενώ οι στήλες μπορούν να διακρίνονται από διάφορους χαρακτήρες (*delimiter*), όπως το διάστημα, ο χαρακτήρας "tab", το κόμμα, η τελεία, κτλ. και χαρακτηρίζονται αντιστοίχως (*space-delimited, tab-delimited, comma-delimited, dot-delimited*, κτλ.). Ενίοτε, οι στήλες διακρίνονται από τη θέση τους σε κάθε γραμμή (*fixed-width*).

### Delimited files

Η `R` περιλαμβάνει μια οικογένεια εντολών για την εισαγωγή *delimited text* αρχείων, βάσει της εντολής `read.table`. Η πλήρης σύνταξη με το σύνολο των ορισμάτων της εντολής δίνεται παρακάτω. Ωστόσο, για την εφαρμογή της, αρκούν συνήθως 2 - 3 ορίσματα, κυρίως το όρισμα `sep,` με το οποίο καθορίζεται ο οριοθέτης των τιμών του αρχείου (*separator* / *delimiter*), και το `header,` με το οποίο καθορίζεται αν το αρχείο περιέχει ονόματα μεταβλητών (*header*). Μια πλήρης περιγραφή των διαθέσιμων ορισμάτων για τη `read.table` είναι διαθέσιμη στο βοηθητικό αρχείο (*documentation*) της εντολής, πληκτρολογώντας `help(read.table)` ή `?read.table` στην `R`.

```
read.table(file, header = FALSE, sep = "", quote = "\"'",
           dec = ".", numerals = c("allow.loss", "warn.loss",
"no.loss"),
           row.names, col.names, as.is = !stringsAsFactors,
           na.strings = "NA", colClasses = NA, nrows = -1,
           skip = 0, check.names = TRUE, fill = !blank.lines.skip,
           strip.white = FALSE, blank.lines.skip = TRUE,
           comment.char = "#",
           allowEscapes = FALSE, flush = FALSE,
           stringsAsFactors = default.stringsAsFactors(),
           fileEncoding = "", encoding = "unknown", text, skipNul =
FALSE)
```

Για παράδειγμα,

```
data <- read.table(file = "C:/Users/Laptop/prices.txt",
sep = "\t", header = TRUE)
data <- read.table("C:/Users/Laptop/prices.txt", TRUE, "\t")
```

Οι δύο προηγούμενες εντολές έχουν ακριβώς το ίδιο αποτέλεσμα: διαβάζουν, δηλαδή, το *text* αρχείο *prices.txt* και το εκχωρούν στο `data.frame` *data*. Στην πρώτη περίπτωση, χρησιμοποιούνται τα ονόματα των ορισμάτων (σε αυθαίρετη σειρά), ενώ στη δεύτερη, παραλείπονται, διατηρείται όμως η σειρά με την οποία εμφανίζονται στην πλήρη σύνταξη της εντολής. Με το όρισμα `file` (ή το πρώτο στη σειρά όρισμα) καθορίζεται η ακριβής διεύθυνση του αρχείου των δεδομένων στον ηλεκτρονικό υπολογιστή, ή ακόμα και διαδικτυακά. Η διεύθυνση πρέπει να βρίσκεται μέσα σε εισαγωγικά, "...", και για τη σύνταξή της να χρησιμοποιούνται κάθετες (*slash*, `"/"`) ή διπλές αντίστροφες κάθετες (*backslash* `"\\"`). Είναι δυνατόν να χρησιμοποιηθούν όλα σχεδόν τα είδη *text* επεκτάσεων (π.χ. *.txt*, *.dat*, *.asc*, κτλ.) ως *input* αρχείο στην `R`, όπου κάθε γραμμή του αρχείου θεωρείται ως μια παρατήρηση και κάθε στήλη μια μεταβλητή. Η εντολή `read.table` είναι πολυ-λειτουργική και μπορεί να διαβάσει δεδομένα από αρχεία με πολλά διαφορετικά χαρακτηριστικά.

Συνηθέστερα, ωστόσο, τα δεδομένα βρίσκονται σε αρχεία λογιστικών φύλλων (*.xls*, *.xlsx* αρχεία). Σε αυτή την περίπτωση, μπορεί κανείς να εξαγάγει τα δεδομένα σε μορφή *.txt* αρχείου και να τα εισαγάγει στην `R` με την εντολή `read.table` ή, προτιμότερα, σε μορφή *.csv* αρχείου με τη χρήση της εντολής `read.csv`, κατάλληλης για *comma-separated* αρχεία. Σε αυτή την κατηγορία υπάρχει και η εντολή `read.delim` για *tab-delimited* αρχεία. Και στις δύο περιπτώσεις, δεν χρειάζεται να καθοριστεί το σχετικό όρισμα `sep`. (Εξαιρώντας όσους χρησιμοποιούν το κόμμα ως *decimal point*, οι εντολές `read.csv2` και `read.delim2` μπορούν εξίσου να χρησιμοποιηθούν).

## Fixed-width files

Για να προσπελαστούν *fixed-width text* αρχεία, μπορεί κάποιος να χρησιμοποιήσει την εντολή `read.fwf`:

```
read.fwf(file, widths, header = FALSE, sep = "\t",
         skip = 0, row.names, col.names, n = -1,
         buffersize = 2000, ...)
```

Μια περιεκτική περιγραφή των ορισμάτων και αυτής της εντολής υπάρχει στο βοηθητικό αρχείο (*documentation)* της εντολής, και εμφανίζεται όταν κανείς πληκτρολογήσει `help(read.fwf)` ή `?read.fwf`.

## Εναλλακτικές Εντολές για Εισαγωγή Δεδομένων

Εναλλακτικά της εντολής `read.table`, διατίθενται οι εντολές `readLines`, η οποία δημιουργεί διάνυσμα χαρακτήρων (`mode character`) και η εντολή `scan`, η οποία πέρα από τη λειτουργία εισαγωγής δεδομένων με πληκτρολόγηση, επιτρέπει να αναγνωσθεί το περιεχόμενο αρχείου στην `R`.

```
readLines(con = stdin(), n = -1L, ok = TRUE, warn = TRUE,
          encoding = "unknown", skipNul = FALSE)

scan(file = "", what = double(), nmax = -1, n = -1, sep = "",
     quote = if(identical(sep, "\n")) "" else "'\"", dec = ".",
     skip = 0, nlines = 0, na.strings = "NA",
     flush = FALSE, fill = FALSE, strip.white = FALSE,
     quiet = FALSE, blank.lines.skip = TRUE, multi.line = TRUE,
     comment.char = "", allowEscapes = FALSE,
     fileEncoding = "", encoding = "unknown", text, skipNul = FALSE)
```

Περισσότερες πληροφορίες, στο *documentation* της κάθε εντολής.

## Άλλα Προγράμματα

Παρ' ότι σχεδόν όλα τα προγράμματα μπορούν να εξάγουν δεδομένα σε *text* αρχεία, η διαδικασία είναι συχνά περιττή, δεδομένου ότι η `R` μπορεί να διαβάσει αρχεία διαφορετικών προγραμμάτων (π.χ. Minitab, Stata, SPSS), μέσω του πακέτου `foreign`. Μια λίστα εντολών για ανάγνωση (και εγγραφή) αρχείων διαφορετικών προγραμμάτων δίνεται στον πίνακα που ακολουθεί.

| File format | Reading | Writing |
|---|---|---|
| ARFF | read.arff | write.arff |
| DBF | read.dbf | write.dbf |
| Stata | read.dta | write.dta |
| Epi Info | read.epiinfo | |
| Minitab | read.mtp | |
| Octave | read.octave | |
| S3 binary files, data.dump files | read.S | |
| SPSS | read.spss | |
| SAS Permanent Dataset | read.ssd | |
| Systat | read.sys stat | |
| SAS XPORT File | read.xport | |

Εκτός από το foreign, υπάρχουν και άλλα πακέτα της R που έχουν ως στόχο είτε την απλούστευση της διαδικασίας εισαγωγής, είτε την εισαγωγή/εξαγωγή δεδομένων στην/από την R, από/σε περισσότερο εξειδικευμένα προγράμματα (π.χ. το R.matlab πακέτο για εισαγωγή και αποθήκευση αρχείων σε μορφή Matlab).

## Βασικές Πράξεις στην R

Απλές πράξεις, όπως η πρόσθεση και ο πολλαπλασιασμός, καθώς και άλλες βασικές αριθμητικές πράξεις, εκτελούνται με τη χρήση αριθμητικών τελεστών. Έτσι, με τους τυπικούς τελεστές +, −, *, /, ^ επιτυγχάνονται η πρόσθεση, η αφαίρεση, ο πολλαπλασιασμός, η διαίρεση και η ύψωση σε δύναμη, αντίστοιχα. Το σύμβολο %/% χρησιμοποιείται για την ακέραια διαίρεση, και το σύμβολο %% για το υπόλοιπο της ακέραιης διαίρεσης. Οι βασικές πράξεις έχουν αριθμητικό αποτέλεσμα. Η R χρησιμοποιεί τον τυπικό κανόνα σειράς των πράξεων BODMAS.

Για λογικές πράξεις χρησιμοποιούνται οι λογικοί τελεστές <, <=, >, => για ανισότητες και ανισοϊσότητες, και οι τελεστές ==, != για ισότητα και άρνηση ισότητας, αντίστοιχα. Επιπλέον, το σύμβολο & χρησιμοποιείται για την τομή, το | για την ένωση και το! για την άρνηση. Οι λογικές πράξεις έχουν αποτέλεσμα TRUE ή FALSE, που αντιστοιχούν στις αριθμητικές τιμές 1 και 0.

Επιπρόσθετα, στην R είναι διαθέσιμη μια σειρά από βασικές μαθηματικές συναρτήσεις, όπως abs, sign, log, sqrt, exp, sin, cos, tan με τις αντίστοιχες προφανείς ερμηνείες. Οι μαθηματικές σταθερές, επίσης, pi, e και i είναι διαθέσιμες μέσω των εντολών pi, exp(1) και 1i, ενώ υπάρχουν και οι ειδικές τιμές NA για τις ελλείπουσες τιμές (*not available*), Inf για το άπειρο (*infinity*), και NaN για τις απροσδιοριστίες (*not a number*). Τέλος, το σύμβολο # είναι χρήσιμο για σχολιασμό, καθώς η R δεν αναγνωρίζει οτιδήποτε έπεται αυτού.

## Δομές Δεδομένων

Στην R υπάρχουν πέντε δομές δεδομένων (ή αντικείμενα δεδομένων): το διάνυσμα (vector), ο πίνακας (matrix), ο πολυεπίπεδος πίνακας (array), η λίστα (list) και η ορθογώνια λίστα δεδομένων (data frame).

### vector

Το vector είναι μια διατεταγμένη συλλογή στοιχείων ιδίου τύπου (*mode numeric, character, logical*) και αποτελεί κύριο στοιχείο για την R. Τα στοιχεία ενός vector μπορούν να έχουν όνομα. Κάθε αριθμός που εισάγεται στην R *console* θεωρείται διάνυσμα. Συνηθίζεται να σχηματίζονται vector με την εντολή c(...) (το c προέρχεται από το "*combine*"):

```
> x <- c(0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89)
> x
 [1]  0  1  1  2  3  5  8 13 21 34 55 89

> names(x) <- c("no.1", "no.2", "no.3", "no.4", "no.5", "no.6",
"no.7", "no.8", "no.9", "no.10", "no.11", "no.12")

> x
 no.1  no.2  no.3  no.4  no.5  no.6  no.7  no.8  no.9 no.10 no.11
no.12
```

```
    0     1     1     2     3     5     8    13    21    34    55
89
```

Αναφορά σε στοιχείο (ή στοιχεία) ενός `vector` γίνεται χρησιμοποιώντας τα σύμβολα αριστερής και δεξιάς αγκύλης "`[ ]`" και τη θέση του στοιχείου στο διάνυσμα (ή το όνομά του, εάν έχει, σε εισαγωγικά "`"`"):

```
> x[8]
no.8
  13

> x["no.3"]
no.3
   1

> smpl <- c(1, 4, 5, 7, 12)
> x[smpl]
 no.1  no.4  no.5  no.7 no.12
    0     2     3     8    89

> x[-c(2, 5)]
 no.1  no.3  no.4  no.6  no.7  no.8  no.9 no.10 no.11 no.12
    0     1     2     5     8    13    21    34    55    89
```

**matrix**

Ένα `matrix` για την `R` είναι απλά ένα `vector` δύο διαστάσεων. Είναι, δηλαδή, μια ταξινομημένη, σε δύο διαστάσεις (γραμμές / στήλες), συλλογή στοιχείων ιδίου τύπου. Για τον σχηματισμό `matrix`, χρησιμοποιείται κυρίως η εντολή `matrix(...)`.

```
> m <- matrix(x, nrow = 3, ncol = 4)
> m
     [,1] [,2] [,3] [,4]
[1,]    0    2    8   34
[2,]    1    3   13   55
[3,]    1    5   21   89
```

Οι γραμμές και οι στήλες ενός πίνακα μπορούν εξίσου να έχουν ονόματα, χρησιμοποιώντας την εντολή `dimnames`, είτε ως όρισμα της εντολής `matrix`, είτε ως ξεχωριστή εντολή:

```
> dimnames(m) <- list(c("row.1", "row.2", "row.3"), c("col.1",
"col.2", "col.3", "col.4"))
> m
      col.1 col.2 col.3 col.4
row.1     0     2     8    34
row.2     1     3    13    55
row.3     1     5    21    89
```

Αναφορά σε στοιχείο (ή στοιχεία) ενός `matrix`, όπως και παραπάνω, γίνεται με χρήση των αγκυλών `[ , ]` και της θέσης του στοιχείου (γραμμή και στήλη, διαχωρισμένες με κόμμα).

```
> m[2,3]
[1] 13

> m[2:3, 3:4]
      col.3 col.4
row.2    13    55
row.3    21    89

> m[, 2]
row.1 row.2 row.3
    2     3     5

> m[, 2, drop = FALSE]
      col.2
row.1     2
row.2     3
row.3     5

> m["row.3",]
col.1 col.2 col.3 col.4
    1     5    21    89
```

**array**

Ένα `array` είναι ένα πολυδιάστατο διάνυσμα. Πρόκειται για μια γενίκευση του `matrix` και φέρει τις ίδιες ιδιότητες. Για τον σχηματισμό `array` χρησιμοποιείται κυρίως η εντολή `array(...)`.

```
> a <- array(x, dim = c(2, 3, 2))
> a
, , 1

     [,1] [,2] [,3]
[1,]    0    1    3
[2,]    1    2    5

, , 2

     [,1] [,2] [,3]
[1,]    8   21   55
[2,]   13   34   89
```

Αναφορά σε στοιχείο (ή στοιχεία) ενός `array` γίνεται αναλόγως,

```
> a[1, 3, 2]
[1] 55

> a[1, 2:3, 1:2]
     [,1] [,2]
[1,]    1   21
[2,]    3   55
```

**`list`**

Για περισσότερο σύνθετες δομές δεδομένων, η `R` χρησιμοποιεί το αντικείμενο `list`. Ένα `list` στην `R` μπορεί να περιέχει μια ετερογενή συλλογή αντικειμένων, όπως `vector`, `matrix`, ακόμα και άλλο `list`. Όπως και για το `vector`, κάθε στοιχείο του `list` μπορεί να έχει όνομα. Για τον σχηματισμό `list` χρησιμοποιείται η εντολή `list(...)`.

```
> l <- list(vec = x, mat = m, array = a, "fibonacci")
> l
$vec
 no.1  no.2  no.3  no.4  no.5  no.6  no.7  no.8  no.9 no.10 no.11
no.12
    0     1     1     2     3     5     8    13    21    34    55
89

$mat
      col.1 col.2 col.3 col.4
row.1     0     2     8    34
row.2     1     3    13    55
row.3     1     5    21    89

$array
, , 1

     [,1] [,2] [,3]
[1,]    0    1    3
[2,]    1    2    5

, , 2

     [,1] [,2] [,3]
[1,]    8   21   55
[2,]   13   34   89


[[4]]
```

```
[1] "fibonacci"
```

Αναφορά σε στοιχείο (ή στοιχεία) ενός `list` μπορεί να γίνει με περισσότερους από έναν τρόπους. Αρχικά, με χρήση (μονών) αγκυλών `[ ]` και τη θέση του στοιχείου (ή το όνομά του, εάν υπάρχει, σε εισαγωγικά `" "`), οπότε και ανακτάται μια υπό-λίστα (αντικείμενο `list`):

```
> l[1]
$vec
 no.1  no.2  no.3  no.4  no.5  no.6  no.7  no.8  no.9 no.10 no.11
no.12
    0     1     1     2     3     5     8    13    21    34    55
89

> l[4]
[[1]]
[1] "fibonacci"
```

Διαφορετικά, με χρήση διπλών αγκυλών `[[ ]]`, οπότε ανακτάται το αντικείμενο της θέσης (ή του ονόματος):

```
> l[[2]]
      col.1 col.2 col.3 col.4
row.1     0     2     8    34
row.2     1     3    13    55
row.3     1     5    21    89

> l[['array']]
, , 1

     [,1] [,2] [,3]
[1,]    0    1    3
[2,]    1    2    5

, , 2

     [,1] [,2] [,3]
[1,]    8   21   55
[2,]   13   34   89

> l[[4]]
[1] "fibonacci"
```

Τέλος, με χρήση του ονόματος του στοιχείου, με το σύμβολο `$`

```
> l$mat
```

```
      col.1 col.2 col.3 col.4
row.1     0     2     8    34
row.2     1     3    13    55
row.3     1     5    21    89
```

**data frame**

Τέλος, ένα `data frame` είναι ένα `list` με τη δομή ενός πίνακα, δηλαδή με γραμμές και στήλες. Ένα `data frame` είναι περίπου σαν ένα `matrix`, διατηρώντας ωστόσο τις ιδιότητες του `list`. Για τον σχηματισμό `data.frame` χρησιμοποιείται η εντολή `data.frame(...)`.

```
> age <- c(23,43,34,26,56)
> weight <- c(60, 85, 52, 54, 81)
> smoke <- c('Yes','No','No','Yes','Yes')
> dfrm <- data.frame(age,weight,smoke)
> dfrm
  age weight smoke
1  23     60   Yes
2  43     85    No
3  34     52    No
4  26     54   Yes
5  56     81   Yes
```

Μπορεί κάποιος να ανακτήσει στοιχεία του `data frame`, όπως ακριβώς σε ένα αντικείμενο `list`, χρησιμοποιώντας δηλαδή το σύμβολο `$`, μονές ή διπλές αγκύλες:

```
> dfrm$age
[1] 23 43 34 26 56

> dfrm['smoke']
  smoke
1   Yes
2    No
3    No
4   Yes
5   Yes

> dfrm[[2]]
[1] 60 85 52 54 81
```

## Βασικές Πράξεις με `Vectors`

Ένα από τα μεγαλύτερα πλεονεκτήματα της `R` είναι οι πράξεις με `vectors`. Οι αριθμητικές και οι λογικές πράξεις εφαρμόζονται σε κάθε ένα στοιχείο του `vector` (*element-wise*). Οι περισσότερες εντολές λειτουργούν με βάση αυτή τη λογική (*vectorization*), με το παραγόμενο αποτέλεσμα να διατηρεί την αρχική δομή των δεδομένων:

```
> x <- x[1:7]
> x
no.1 no.2 no.3 no.4 no.5 no.6 no.7
   0    1    1    2    3    5    8

> x + 2
no.1 no.2 no.3 no.4 no.5 no.6 no.7
   2    3    3    4    5    7   10

> x > 4
 no.1  no.2  no.3  no.4  no.5  no.6  no.7
FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE

> 2^x
no.1 no.2 no.3 no.4 no.5 no.6 no.7
   1    2    2    4    8   32  256

> log(x)
     no.1      no.2      no.3      no.4      no.5      no.6      no.7
     -Inf 0.0000000 0.0000000 0.6931472 1.0986123 1.6094379 2.0794415
```

Το *vectorization* έχει δύο μεγάλα πλεονεκτήματα, την ευκολία χρήσης και την ταχύτητα εκτέλεσης.

Η λογική των πράξεων με `vectors` επεκτείνεται και μεταξύ `vectors`, όταν αυτά έχουν το ίδιο μήκος. Όταν τα `vectors` έχουν διαφορετικά μήκη, η `R` εφαρμόζει το *Recycling Rule*. Με βάση αυτόν τον κανόνα, το παραγόμενο αποτέλεσμα διατηρεί το μήκος του μεγαλύτερου διανύσματος, και το μικρότερο διάνυσμα "ανακυκλώνεται" όσες φορές χρειάζεται για να εκτελεστεί η αριθμητική πράξη. Η `R` παράγει προειδοποιητική ειδοποίηση για την ανισότητα των μηκών, όταν το μήκος του μεγαλύτερου `vector` δεν είναι ακέραιο πολλαπλάσιο του μικρότερου.

Όπως γίνεται αντιληπτό, όλη η δουλειά στην `R` γίνεται από τις εντολές (συνήθως προεγκατεστημένες) οι οποίες χρησιμοποιούνται κατά κόρον και αποτελούν δομικά στοιχεία αυτής. Στην ουσία πρόκειται για προγραμματισμένες συναρτήσεις (*functions*), της γενικής δομής:

```
f(argument1, argument2, ...)
```

όπου `f` το όνομα της εντολής και `argument1, argument2, . . .` τα ορίσματά της. Ωστόσο, υπάρχουν και *functions* με τη μορφή τελεστών, όπως για παράδειγμα ο τελεστής `:` (*colon*), ο οποίος χρησιμοποιείται για τη δημιουργία απλών αριθμητικών ακολουθιών:

```
> 5:25
 [1]  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
```

Μεταξύ των πλέον διαδεδομένων *functions* είναι και η γενίκευση του τελεστή *colon*, η εντολή `seq()`, η οποία έχει πέντε ορίσματα και χρησιμοποιείται για τη δημιουργία περισσότερο σύνθετων αριθμητικών ακολουθιών:

```
> seq(5, 25)
 [1]  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
```

```
> seq(5, 20, by = 5)
[1]  5 10 15 20
```

Άξιες αναφοράς είναι επίσης οι εντολές `rep()`, για την επανάληψη δομών της `R`, και `paste()`, για την ένωση διανυσμάτων σε ένα *character* διάνυσμα:

```
> rep(exp(1), 5)
[1] 2.718282 2.718282 2.718282 2.718282 2.718282

> paste("row", 1:5, sep = ".")
[1] "row.1" "row.2" "row.3" "row.4" "row.5"
```

Τέλος, συχνής χρήσης τυγχάνουν και οι εντολές `order()` και `rank()`. Η πρώτη χρησιμοποιείται για τη διάταξη (με αύξουσα ή φθίνουσα σειρά) διανυσμάτων και η δεύτερη επιστρέφει τις τάξεις (*ranks*) των στοιχείων ενός διανύσματος,

```
> order(dfrm$weight)
[1] 3 4 1 5 2

> dfrm$weight[order(dfrm$weight)]
[1] 52 54 60 81 85

> rank(dfrm$weight)
[1] 3 5 1 2 4
```

# Παράρτημα ΙΙ - ΠΑΚΕΤΑ ΣΤΗΝ R

Ένα πακέτο (package) της R αποτελείται από ένα σύνολο εντολών, βοηθητικών αρχείων (*help files*) και ενδεχόμενα αρχείων δεδομένων, τα οποία διανέμονται μαζί. Προφανώς, όλες οι εντολές ενός πακέτου σχετίζονται μεταξύ τους. Τα πακέτα αποτελούν τα αντίστοιχα των *modules* της Perl, των *libraries* στις C/C++ και των *classes* της Java και στοχεύουν στην επεξεργασία δεδομένων, την εφαρμογή στατιστικών αναλύσεων ή/και τη δημιουργία γραφημάτων.

Ως προεπιλογή, η R κατά την εκκίνηση φορτώνει τα βασικά προς χρήση, από τα άμεσα διαθέσιμα, πακέτα. Πέραν αυτών, η R προσφέρει ένα πολύ μεγάλο αριθμό πακέτων για στατιστικές αναλύσεις και δημιουργία γραφικών:

(i) ορισμένα περιλαμβάνονται στην R κατά την αρχική εγκατάσταση και απλά χρειάζεται να πει κανείς στην R ότι θέλει να τα χρησιμοποιήσει,

(ii) πολλά περισσότερα είναι διαθέσιμα μέσω του διαδικτύου και

(iii) μπορεί κανείς να δημιουργήσει και τα δικά του πακέτα.

## Επισκόπηση των πακέτων

Για να κάνει κανείς χρήση ενός πακέτου στην R, θα πρέπει πρώτα να βεβαιωθεί ότι είναι εγκατεστημένο τοπικά (*local library*). Έπεται το φόρτωμα των πακέτων κατά την τρέχουσα συνεδρία της R (*current session*). Ο λόγος για τον οποίο πρέπει κανείς να φορτώνει τα πακέτα στην R για να τα χρησιμοποιήσει είναι, πρώτον, ότι το *help system* της R καθυστερεί όταν έχουν προστεθεί αρκετά πακέτα στον μηχανισμό αναζήτησης. Επιπλέον, είναι πιθανό δύο πακέτα να έχουν αντικείμενα του ιδίου ονόματος, κάτι που δημιουργεί ζητήματα συμβατότητας (*conflicts*). Φορτώνοντας μόνο πακέτα που χρειάζεται κανείς, ελαχιστοποιεί την πιθανότητα αυτή.

### Λίστα Άμεσα Διαθέσιμων Πακέτων

Η λίστα των πακέτων που φορτώνονται αυτόματα κατά την έναρξη της R δίνεται με την εντολή getOption, χρησιμοποιώντας το όρισμα defaultPackages,

```
> getOption("defaultPackages")
[1] "datasets"  "utils"     "grDevices" "graphics"  "stats"
"methods"
```

Με την εντολή αυτ, παραλείπεται το πακέτο base το οποίο υλοποιεί μια σειρά από βασικές λειτουργίες για τη γλώσσα R και το οποίο φορτώνεται πάντα.

Εναλλακτικά, μπορεί κανείς να δει τη λίστα των πακέτων που είναι ήδη φορτωμένα, χρησιμοποιώντας την εντολή .packages σε παρενθέσεις,

```
> (.packages())
[1] "stats"     "graphics"  "grDevices" "utils"     "datasets"
"methods"
[7] "base"
```

Για την πλήρη λίστα των διαθέσιμων πακέτων, χρησιμοποιείται το όρισμα all.available,

```
> (.packages(all.available=TRUE))
  [1] "BB"              "boot"            "car"             "class"
  [5] "cluster"         "codetools"       "coin"            "colorspace"
  [9]  ...
[125] "survival"        "tcltk"           "tools"
"translations"
[129] "utils"
```

Επίσης, η εντολή `library()` (χωρίς ορίσματα) ανοίγει ένα παράθυρο, με τη λίστα των διαθέσιμων πακέτων, διανθισμένη με σύντομη περιγραφή για το κάθε ένα από αυτά.

## Διαθέσιμα Κατά Την Εγκατάσταση Πακέτα

Η R συνοδεύεται από έναν αριθμό πακέτων κατά την εγκατάσταση. Μερικά από αυτά τα πακέτα (όπως τα `base`, `graphics`, `grDevices`, `methods` και `utils`) υλοποιούν βασικές λειτουργίες της γλώσσας και του περιβάλλοντος R. Άλλα πακέτα προσφέρουν κοινά στατιστικά εργαλεία (όπως τα `cluster`, `nnet` και `stats`). Άλλα πακέτα υλοποιούν εξειδικευμένα γραφήματα (`grid` και `lattice`), περιλαμβάνουν σύνολα δεδομένων, ή περιλαμβάνουν άλλες χρήσιμες/συχνά χρησιμοποιούμενες εντολές. Στις περισσότερες περιπτώσεις, τα πακέτα της αρχικής εγκατάστασης είναι υπεραρκετά. Για επιπλέον πακέτα, κανείς θα πρέπει να ακολουθήσει τη διαδικασία αναζήτησης / εγκατάστασης πακέτων.

## Αναζήτηση Πακέτων

Χιλιάδες πακέτα της R είναι διαθέσιμα διαδικτυακά. Η μεγαλύτερη πηγή άντλησης πακέτων είναι το CRAN (Comprehensive R Archive Network), αλλά ορισμένα πακέτα είναι διαθέσιμα σε εναλλακτικές όπως το Bioconductor και το R-Forge. Το CRAN υποστηρίζεται από το R Foundation (ο μη-κερδοσκοπικός οργανισμός που επιβλέπει και τη σουίτα R). Εκεί θα βρει κανείς έναν μεγάλο αριθμό πακέτων, που καλύπτουν ένα μεγάλο εύρος από διαφορετικές εφαρμογές.

Το CRAN φιλοξενείται σε πολλούς ιστοτόπους ανά τον κόσμο, από τους οποίους επιλέγεται ο πλησιέστερος γεωγραφικά ιστότοπος για να κατεβάσει κανείς πακέτα, ελαχιστοποιώντας με τον τρόπο αυτό τους χρόνους λήψης και μειώνοντας το φόρτο του R Foundation *server*. Το Bioconductor είναι ένα *open-source project* δημιουργίας εργαλείων ανάλυσης, και το R-Forge είναι ένας ιστότοπος, ο οποίος περιέχει εν εξελίξει, κατά κανόνα, *projects*. Πάντως, τα περισσότερα R *projects* επιλέγουν να χρησιμοποιούν το CRAN για να κοινοποιήσουν τα πακέτα τους.

## Διαδικτυακή Αναζήτηση Πακέτων R

Η R δεν προσφέρει κάποιο εξειδικευμένο εργαλείο περιγραφικής αναζήτησης πακέτων. Ένας εύκολος, ωστόσο, τρόπος αναζήτησης πακέτου είναι μέσω μιας διαδικτυακής μηχανής αναζήτησης. Για παράδειγμα, αναζητώντας "R package stratification survey populations" μπορεί να οδηγηθεί κανείς στο πακέτο `stratification`, το οποίο περιέχει την εντολή `strata.geo`. Μπορεί κανείς, επιπλέον, να περιηγηθεί στο σύνολο των διαθέσιμων πακέτων στους ιστότοπους των CRAN, Bioconductor και R-Forge.

## Εγκατάσταση Πακέτων στην R

Όταν θα έχει καταλήξει κανείς στο πακέτο που θέλει να εγκαταστήσει, ο πιο εύκολος τρόπος είναι να το κάνει μέσα από την R.

### Windows και Linux GUIs

Η εγκατάσταση πακέτων μέσω του Windows GUI είναι σχετικά άμεση.

- Από προεπιλογή, η R είναι προγραμματισμένη να λαμβάνει πακέτα από το "CRAN" και το "CRAN (extra)". Για να εμπλουτίσει κανείς τη λίστα των πακέτων, επιλέγει "Select repositories" από το βασικό μενού Packages, όπου μπορεί να αυξήσει τις επιλογές του. (Προαιρετικά).
- Από το βασικό μενού Packages, επιλέγεται "Install package(s)".
- Εάν είναι η πρώτη φορά που γίνεται εγκατάσταση πακέτου κατά τη διάρκεια της τρέχουσας R εφαρμογής (session), η R θα ζητήσει να επιλεχθεί ιστότοπος από όπου θα γίνει η λήψη.
- Επιλέγεται το όνομα του πακέτου προς εγκατάσταση και, ακολούθως, OK.
- Η R θα κατεβάσει και θα εγκαταστήσει το πακέτο που επιλέχθηκε.

Σημειώνεται ότι κατά την εγκατάσταση πακέτων, ενδέχεται να προκύψουν ορισμένα ζητήματα, ανάλογα με τα δικαιώματα του κάθε χρήστη. Για να ολοκληρωθεί η εγκατάσταση, θα πρέπει η R να εκτελείται ως Administrator. Στο Mac OS X, υπάρχει ένα ελαφρώς διαφορετικό UI για την εγκατάσταση πακέτων, όπου από το βασικό μενού Package and Data, επιλέγεται το Package Installer. Στην επάνω αριστερή γωνία του νέου παραθύρου, υπάρχει η επιλογή της κατηγορίας των πακέτων, και επιλέγοντας το κουμπί Get List εμφανίζεται το διαθέσιμο σύνολο των πακέτων.

### R console

Μπορεί κανείς να εγκαταστήσει στην R πακέτα κατευθείαν από την R *console*. Για αυτή τη διαδικασία είναι διαθέσιμο ένα σύνολο εντολών, όπως `installed.packages`, `available.packages`, `old.packages`, `new.packages`, `download.packages`, `install.packages`, `remove.packages`, `update.packages`, `setRepositories`. Περισσότερες πληροφορίες για τις εντολές μπορεί να ανακτήσει κανείς από τα αντίστοιχα *help files*. Ένα απλό παράδειγμα εγκατάστασης του πακέτου `stratification` θα είχε το ακόλουθο *outcome*:

```
> install.packages('stratification')
Installing package into ‘C:/Users/Laptop/Documents/R/win-library/3.1’
(as ‘lib’ is unspecified)
trying URL
'http://cran.rstudio.com/bin/windows/contrib/3.1/stratification_2.2-
5.zip'
Content type 'application/zip' length 621991 bytes (607 Kb)
opened URL
downloaded 607 Kb

package ‘stratification’ successfully unpacked and MD5 sums checked
Warning: cannot remove prior installation of package ‘stratification’

The downloaded binary packages are in
```

```
C:\Users\Laptop\AppData\Local\Temp\RtmpmIzdPo\downloaded_packages
```

Η ανωτέρω διαδικασία εγκατέστησε το πακέτο σε προεπιλεγμένο σημείο, το οποίο καθορίζει η μεταβλητή `.Library`. Εάν κανείς επιθυμεί να αφαιρέσει το συγκεκριμένο πακέτο, μπορεί να χρησιμοποιήσει την εντολή `remove.packages`. Χρειάζεται ωστόσο να καθοριστεί και πού βρίσκεται εγκατεστημένο το πακέτο.

```
> remove.packages('stratification',.Library)
```

## Φόρτωση Πακέτων

Ως προεπιλογή, όλα τα άμεσα διαθέσιμα πακέτα δεν φορτώνονται αυτόματα στην `R`. Κατά την εκκίνηση του προγράμματος, επτά (περίπου) βασικά πακέτα φορτώνονται. Εάν κανείς επιχειρήσει να τρέξει μια εντολή από ένα πακέτο το οποίο δεν έχει φορτωθεί, η `R` εμφανίζει το μήνυμα:

```
> # try to use strata.geo before loading it
> adjust <- strata.geo(x=USbanks, CV=0.01, Ls=4,
alloc=c(0.35,0.35,0))
Error: could not find function "strata.geo"
```

Συνεπώς, και σε αντίθεση με την εγκατάσταση ενός πακέτου η οποία γίνεται άπαξ, η χρησιμοποίηση ένας πακέτου στην `R` απαιτεί τη φόρτωσή του κάθε φορά που πρόκειται να χρησιμοποιηθεί (εκτός αυτών που φορτώνονται ως προεπιλογή). Για να φορτώσει κανείς ένα πακέτο στην `R`, μπορεί να χρησιμοποιήσει την εντολή `library()`. Για παράδειγμα:

```
> library(stratification)
```

Υπάρχει και η επιλογή της εντολής `require()`, η οποία έχει ελαφρώς διαφορετικά ορίσματα. Για περισσότερα, στο `R` *help file*.

Τέλος, υπάρχει και η επιλογή της χρήσης του GUI, όπου μπορεί κανείς να περιηγηθεί και να επιλέξει πακέτο, από το νέο παράθυρο που ανοίγει επιλέγοντας "Load Package..." στο βασικό μενού Packages. Στο Mac OS X η επιλογή είναι "Package Manager" από το βασικό μενού "Packages & Data", όπου ο *Package Manager* UI, επιτρέπει να διακρίνει ποια πακέτα είναι φορτωμένα, ποια μπορούν να φορτωθούν ακόμα και να περιηγηθεί στο *help file* ενός πακέτου.