ΤΕΧΝΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΠΑΤΡΩΝ
ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ
ΤΜΗΜΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΥ ΣΧΕΔΙΑΣΜΟΥ ΚΑΙ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΜΕΘΟΔΟΙ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ ΚΑΙ ΚΑΤΑΝΟΜΕΣ

ΤΣΑΠΙΚΟΥΝΗΣ ΑΝΔΡΕΑΣ
ΡΙΖΟΠΟΥΛΟΥ ΑΔΑΜΑΝΤΙΑ
ΚΟΚΚΑΛΗΣ ΒΑΣΙΛΕΙΟΣ

ΕΠΟΠΤΕΥΩΝ ΚΑΘΗΓΗΤΗΣ
ΜΠΟΥΜΠΟΥΛΗ ΑΘΑΝΑΣΙΑ

ΠΑΤΡΑ - 2012

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΙΣΑΓΩΓΗ

Ορισμός

Ως στατιστική ορίζουμε τη συστηματική απαρίθμηση και παρουσίαση αριθμητικών δεδομένων ή στοιχείων τα οποία προέρχονται από πολλές παρατηρήσεις ή μετρήσεις. Ανάλογα με το αντικείμενο στο οποίο αναφέρονται τα αριθμητικά δεδομένα η στατιστική παίρνει ιδιαίτερη ονομασία, για παράδειγμα η στατιστική επιχειρήσεων, όπου αναφέρονται αριθμητικά δεδομένα που αφορούν τις επιχειρήσεις.

Στη γλώσσα της επιστήμης, η λέξη στατιστική έχει ευρύτερη σημασία, σημαίνει την επιστήμη που έχει ως αντικείμενο όχι μόνο την συγκέντρωση και παρουσίαση αλλά ταυτόχρονα και τη μελέτη και ανάλυση των παρατηρήσεων που αναφέρεται σε ένα συγκεκριμένο γεγονός. Δηλαδή μπορούμε να πούμε ότι:

Στατιστική είναι η επιστήμη που ασχολείται με τις επιστημονικές μεθόδους συλλογής, οργάνωσης και ανάλυσης των αριθμητικών στοιχείων που αναφέρονται σε χαρακτηριστικές ιδιότητες διάφορων φαινομένων όπως κοινωνικών, οικονομικών και φυσικών. Έχει ως σκοπό τη συστηματική μελέτη των συγκεκριμένων στοιχείων για την κατάληξη σε γενικά συμπεράσματα τα οποία χρησιμεύουν στη διαδικασία της λήψης ορθών αποφάσεων.

Σύμφωνα με τον παραπάνω ορισμό παρατηρούμε ότι τα βασικά στάδια που ακολουθούμε για τη μελέτη των ιδιοτήτων των μονάδων μιας πολυπληθούς ομάδος είναι τα εξής:

1. Η συγκέντρωση των στατιστικών στοιχείων που είναι απαραίτητη για τη μελέτη του προβλήματος που ερευνούμε
2. Η μεθοδική επεξεργασία και παρουσίαση των στατιστικών στοιχείων σε μορφή αριθμητικών πινάκων και γραφικών παραστάσεων.
3. Η ανάλυση των στοιχείων αυτών και η εξαγωγή χρήσιμων συμπερασμάτων για να ληφθούν σωστές αποφάσεις.

Ιστορία της Στατιστικής

Η ρίζα της λέξης στατιστικής προέρχεται από τη λατινική λέξη Status (που σημαίνει κράτος) και δηλώνει αρχικά συλλογής στοιχείων για τις ανάγκες του κράτους (παραγωγή, πληθυσμός, κτλ). Η πρώτη απογραφή πληθυσμού έγινε στην Κίνα από τον αυτοκράτορα Υ-άο το έτος 2238π.Χ., ενώ στους Ρωμαίους η πρώτη απογραφή πληθυσμού έγινε επί Ρωμύλου (753-715π.Χ.) και η τελευταία από τον αυτοκράτορα Βεσπασιανό το 73μ.Χ.

Το 1583 γράφεται από τον Σανσοβίνο το πρώτο βιβλίο στατιστικού περιεχομένου και λίγο αργότερα εισάγεται το τον Κένρινγκ (1606-1681) η στατιστική στην ανώτερη παιδεία.

Την ίδια περίοδο ο Άγγλος αστρονόμος Χάλεϊ, χρησιμοποιώντας τα ληξιαρχικά βιβλία γεννήσεων και θανάτων, παρουσιάζει τον πρώτο πίνακα θνησιμότητας. Το ίδιο ρεύμα επεκτείνεται και στην Γερμανία όπου ο πάστορας Σίσμιλτς (1707-1767) συγκεντρώνει στοιχεία από τα ληξιαρχικά βιβλία της Πρωσίας και καταλήγει το 1741 στο συμπέρασμα ότι το ποσοστό γέννησης των αγοριών είναι 51% και των κοριτσιών 49%, ενώ τα δύο φύλα έχουν ίσα ποσοστά κατά την εποχή του γάμου. Μέχρι εκείνη την εποχή η στατιστική έχει περιγραφικό χαρακτήρα και ασχολείται κυρίως με θέματα δημογραφίας.

Μετά τον περιγραφικό χαρακτήρα της, κύρια ασχολία της στατιστικής είναι η ανάπτυξη ενός νέου κλάδου του λογισμού των πιθανοτήτων, ο οποίος προήλθε από τη μελέτη των τυχερών παιχνιδιών. Βασικοί θεμελιωτές του λογισμού των πιθανοτήτων είναι ο Μπερνούλι, ο οποίος στο βιβλίο του " Η τέχνη των προβλέψεων" διατυπώνει τον περίφημο νόμο των μεγάλων αριθμών και ο Γάλλος μαθηματικός Λαπλάς στον οποίο οφείλεται η εφαρμογή του λογισμού των πιθανοτήτων στη σπουδή των φυσικών φαινομένων με πολυσύνθετες αιτίες.

Στη νεότερη περίοδο της στατιστικής ο Βέλγος αστρονόμος Κετελέ επεκτείνει την εφαρμογή της στατιστικής στη σπουδή των φυσικών και ηθικών ιδιοτήτων του ανθρώπου και προτείνει τη σύγκληση του πρώτου διεθνούς συνεδρίου στατιστικής που έγινε στις Βρυξέλλες το 1853. Έπειτα ο Γκάλτον εφαρμόζει τη στατιστική στη βιολογία και ειδικότερα στα προβλήματα της κληρονομικότητας. Η συγκεκριμένη προσπάθεια συνεχίστηκε από τον Άγγλο μαθηματικό Πίρσον, στον οποίο οφείλεται κατά πολύ η ανάπτυξη και η θέση της στατιστικής.

Χρησιμότητα της στατιστικής

Η στατιστική χρησιμοποιείται σε όλους σχεδόν τους τομείς ανθρώπινης δραστηριότητας. Επίσης είναι απαραίτητη για τη λήψη ορθών αποφάσεων που έχουν μεγάλη σημασία, για την πρόοδο ενός κράτους ή ακόμα και μιας επιχείρησης. Για τον λόγο αυτό στις μέρες μας δεν υπάρχει κανένας τομέας ο οποίος δεν χρησιμοποιεί τις στατιστικές μεθόδους για την λήψη επιχειρηματικών αποφάσεων.

Η πιο βασική εφαρμογή της στατιστικής είναι αυτή στη δημογραφία γιατί για την μελέτη της γεννητικότητας, της θνησιμότητας, της μετανάστευσης κλπ απαιτούνται μακροχρόνιες στατιστικές παρατηρήσεις και επίπονες αναλύσεις. Επιπρόσθετα, η στατιστική εφαρμόζεται στις μέρες μας σε πολλούς κλάδους όπως στην Ιατρική, Βιολογία, Αστρονομία, Φυσική, στη μελέτη του φυσικού περιβάλλοντος, στη μελέτη των ανθρωπίνων ιδεών, στη θεωρία αποφάσεων κλπ. Τέλος η στατιστική εφαρμόζεται σε μεγάλο βαθμό και στον οικονομικό τομέα αφού η παρακολούθηση του γενικού επιπέδου των τιμών, της νομισματικής ισοτιμίας και των οικονομικών διακυμάνσεων είναι αντικείμενα στατιστικής επεξεργασίας.

Στατιστικός πληθυσμός – Έννοια στατιστικής μεταβλητής

Η λέξη πληθυσμός χρησιμοποιείται στην στατιστική για να δηλώσει το σύνολο των ατόμων ή αντικειμένων στα οποία αναφέρονται οι παρατηρήσεις μας. Τα στοιχεία του συνόλου αυτού λέγονται στατιστικές μονάδες ή άτομα του πληθυσμού. Αντικείμενο μελέτης στατιστικής δεν είναι οι μονάδες ενός πληθυσμού αλλά τα χαρακτηριστικά που περιέχουν οι μονάδες του προς μελέτη πληθυσμού.

Οι χαρακτηριστικές ιδιότητες των στατιστικών μονάδων ενός πληθυσμού με την μελέτη των οποίων ασχολείται η στατιστική ονομάζονται μεταβλητές. Οι μεταβλητές χωρίζονται σε δύο κυρίως κατηγορίες: α) στις ποιοτικές μεταβλητές, οι οποίες είναι εκείνες οι μεταβλητές των οποίων οι τιμές εκφράζονται με λέξεις και δεν μπορούν να μετρηθούν(το φύλο, οικογενειακή κατάσταση) β) στις ποσοτικές μεταβλητές οι οποίες είναι εκείνες οι μεταβλητές των οποίων οι τιμές τους είναι αριθμοί αναφερόμενοι σε συγκεκριμένες μονάδες και μπορούν να μετρηθούν (ηλικία, βάρος, θερμοκρασία).

Οι ποσοτικές μεταβλητές διακρίνονται σε ασυνεχείς και συνεχείς. Ασυνεχείς ονομάζονται οι μεταβλητές εκείνες που μπορούν να λάβουν πεπερασμένο πλήθος τιμών. Για παράδειγμα η ένδειξη ενός ζαριού είναι μια ασυνεχής τυχαία μεταβλητή γιατί το σύνολο των

τιμών της (1,2,3,4,5,6) είναι πεπερασμένο. Συνεχείς ονομάζονται οι μεταβλητές εκείνες που μπορούν να πάρουν όλες τις τιμές ενός διαστήματος, για παράδειγμα το ύψος, η ταχύτητα είναι συνεχείς μεταβλητές)

ΚΕΦΑΛΑΙΟ ΠΡΩΤΟ : ΔΕΙΓΜΑΤΟΛΗΨΙΑ

1.1. ΕΙΣΑΓΩΓΗ

Στο κεφάλαιο που ακολουθεί θα περιγράψουμε τις βασικές έννοιες της δειγματοληψίας και θα αναλύσουμε τα είδη της. Ως δειγματοληπτική έρευνα (survey sampling) ορίζεται η στατιστική έρευνα χαρακτηριστικών ενός πεπερασμένου πληθυσμού η οποία στηρίζεται σε πληροφορίες που συλλέγονται από ένα δείγμα του συγκεκριμένου πληθυσμού. Κύριο χαρακτηριστικό της δειγματοληψίας είναι ότι έχει ως αντικείμενο πεπερασμένο (πραγματικό) πληθυσμό σε αντίθεση με τους άλλους κλάδους της στατιστικής. Οι συλλεγόμενες πληροφορίες χωρίζονται σε α)ποσοτικές και αντικειμενικές για παράδειγμα όταν περιγράφουν οικονομικά χαρακτηριστικά και σε β)ποιοτικές και υποκειμενικές όταν έχουν τη μορφή γνώμης για παράδειγμα δημοσκοπήσεις πολιτικής προτίμησης.

1.2. ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ

1.2.1. *ΔΕΙΓΜΑΤΟΛΗΠΤΙΚΗ ΜΟΝΑΔΑ* (sampling units)

Κατά το σχεδιασμό της δειγματοληψίας απαραίτητη προϋπόθεση είναι ο ορισμός των μονάδων που θα αποτελέσουν τον ερευνώμενο πληθυσμό, τις οποίες καλούμε δειγματοληπτικές μονάδες. Για παράδειγμα αν θέλουμε να βρούμε όλα τα άτομα με το επίθετο "Βασιλόπουλος" που κατοικούν στο νομό Αχαΐας χρειαζόμαστε ένα τηλεφωνικό κατάλογο που περιλαμβάνει όλα τα ονόματα , τηλέφωνα και διευθύνσεις των ατόμων αυτών. Συνεπώς, ο κατάλογος αποτελεί τη δειγματοληπτική μονάδα.

Κάθε δειγματοληπτική μονάδα πρέπει να ορίζεται με ακρίβεια, διότι είναι πολύ σημαντική για τον τρόπο επιλογής ενός δείγματος.

1.2.2. *ΔΕΙΓΜΑΤΟΛΗΠΤΙΚΟ ΠΛΑΙΣΙΟ* (sampling frame)

Δειγματοληπτικό πλαίσιο καλείται το σύνολο των δειγματοληπτικών μονάδων που αντιστοιχούν στον ερευνώμενο πληθυσμό. Το δειγματοληπτικό πλαίσιο αποτελεί απαραίτητο συστατικό για την επιτυχία μιας δειγματοληπτικής έρευνας. Η επιλογή των μονάδων του δείγματος πρέπει να γίνεται προσεκτικά έτσι ώστε το δείγμα που θα επιλεγεί να είναι αντιπροσωπευτικό για να αποφευχθούν τυχόν σφάλματα και τα αποτελέσματα να είναι αξιόπιστα. Για παράδειγμα ο κατάλογος προϊόντων ενός καταστήματος ηλεκτρικών ειδών δεν μπορεί να αποτελέσει δειγματοληπτικό πλαίσιο του πληθυσμού των προϊόντων που πουλάνε όλα τα καταστήματα ηλεκτρικών ειδών και βρίσκονται στην ευρύτερη περιοχή της Ελλάδας. Ένα πλαίσιο για να είναι αποτελεσματικό πρέπει να πληροί ορισμένους όρους όπως:

α) Ο κατάλογος πλαίσιο να είναι πλήρως ενημερωμένος πριν χρησιμοποιηθεί για την επιλογή των τελικών μονάδων του δείγματος.

β) Ο κατάλογος πλαίσιο να περιέχει όλες τις δειγματοληπτικές μονάδες (δηλαδή να μην υπάρχει ελλιπής πληροφόρηση).

γ) Ο κατάλογος πλαίσιο δεν περιέχει ορισμένες δειγματοληπτικές μονάδες περισσότερο από μία φορά (δηλαδή να μην υπάρχουν πολλαπλές εγγραφές).

Βασικά είδη πλαισίου:

1. **Πλαίσιο κατάλογος** (list frame): Χωρίζεται στον πραγματικό και στον νοητό κατάλογο, όπου πραγματικός μπορεί να είναι ένας τηλεφωνικός κατάλογος και παρέχει άμεση πρόσβαση στα μέλη του πληθυσμού. Ο νοητός κατάλογος μπορεί να είναι όλα τα αυτοκίνητα που διέρχονται από συγκεκριμένο σημείο κατά τη διάρκεια κάποιου χρονικού διαστήματος.

2. **Πλαίσιο γεωγραφικής επιφάνειας** (area frame): Είναι μια ειδική περίπτωση κατάλογου-πλαίσιου όπου οι μονάδες αντιστοιχούν σε γεωγραφικές περιοχές. Οι γεωγραφικές μονάδες έχουν καλά ορισμένα φυσικά ή τεχνητά όρια που αναγνωρίζονται στο χάρτη και το μέγεθος αυτών ποικίλλει από μονάδα σε μονάδα. Σε αντίθεση με το πλαίσιο κατάλογο που παρέχει άμεση πρόσβαση το πλαίσιο γεωγραφικής επιφάνειας παρέχει έμμεση πρόσβαση στα μέλη του πληθυσμού γιατί πρώτα ένας κατάλογος γεωγραφικών μονάδων πρέπει να επιλεγεί και μετά πρέπει να σχηματιστεί κατάλογος μονάδων δειγματοληψίας.

3. **Πολλαπλά πλαίσια** (multiple frames): Όταν μόνο ένα πλαίσιο δεν επαρκεί για την πλήρη κάλυψη του πληθυσμού χρησιμοποιούνται τα συγκεκριμένα πλαίσια τα οποία καλύπτουν διαφορετικά μέρη του πληθυσμού.

Σε ορισμένες περιπτώσεις η πρόσβαση στον πληθυσμό μπορεί να κατορθωθεί μέσω ιεραρχίας πλαισίων. Μονάδες που απαρτίζουν ένα πλαίσιο σε ένα επίπεδο της ιεραρχίας διαιρούνται σε μονάδες που απαρτίζουν ένα πλαίσιο στο επόμενο επίπεδο της ιεραρχίας, τα πλαίσια αυτά χρησιμοποιούνται στην πολυσταδιακή δειγματοληψία. Για παράδειγμα σε μία έρευνα προϊόντων στο πρώτο στάδιο μπορεί να χρησιμοποιηθεί πλαίσιο καταστημάτων από το οποίο θα επιλεγεί δείγμα καταστημάτων και σε δεύτερο στάδιο πλαίσιο προϊόντων των επιλεχθέντων καταστημάτων από το οποίο θα γίνει η επιλογή δείγματος προϊόντων.

Κριτήρια επιλογής πλαισίου:

- Καταλληλότητα, πληρότητα και επικαιρότητα.
- Ευκολία συλλογής των απαιτούμενων πληροφοριών για τις μονάδες που απαρτίζουν το πλαίσιο.
- Φύση των συμπληρωματικών/ βοηθητικών πληροφοριών και αν μία αποτελεσματική δειγματοληψία μπορεί να βασιστεί σε αυτές.
- Ευκολία διαχείρισης και ενημέρωσης του πλαισίου για επαναλαμβανόμενες δειγματοληψίες.
- Κόστος δημιουργίας του πλαισίου.

1.3. ΚΑΤΑΛΟΓΟΣ Ή ΛΙΣΤΑ

Για να χρησιμοποιηθεί το δειγματοληπτικό πλαίσιο ως η πρωταρχική πηγή επιλογής του δείγματος θα πρέπει να υπάρχει ένας πραγματικός κατάλογος όλων των δειγματοληπτικών μονάδων. Άλλωστε η επιλογή του δειγματοληπτικού πλαισίου γίνεται για να εξυπηρετήσει το σκοπό αυτό. Για παράδειγμα ο κατάλογος των εργαζομένων σε μια περιοχή.

1.4. ΑΚΡΙΒΕΙΑ ΕΚΤΙΜΗΣΕΩΝ

Ως ακρίβεια ορίζεται η διαφορά μεταξύ της εκτιμήσεως που προκύπτει από ένα δείγμα και της αντίστοιχης παραμέτρου που προκύπτει μέσω της απογραφής. Όσο μικρότερο είναι το δειγματοληπτικό σφάλμα τόσο μεγαλύτερη είναι η ακρίβεια μιας εκτίμησης. Η ακρίβεια εκτιμήσεων εξαρτάται από το τυπικό σφάλμα της εκτίμησης και από το μέγεθος του δείγματος. Όσο μεγαλύτερο μέγεθος δείγματος έχουμε, τόσο μικρότερο θα είναι το τυπικό σφάλμα της εκτίμησης άρα τόσο μεγαλύτερη ακρίβεια εκτίμησης θα έχουμε. Σε σχέση με το απογραφικό μέγεθος του πληθυσμού η ακρίβεια εκτιμήσεων καθορίζεται από το δειγματοληπτικό σφάλμα, εν αντιθέσει με το αληθινό μέγεθος του πληθυσμού που καθορίζεται από όλα τα σφάλματα (δειγματοληπτικά – μη δειγματοληπτικά).

1.5. ΔΕΙΓΜΑΤΟΛΗΠΤΙΚΟ ΣΦΑΛΜΑ

Σε μια δειγματοληπτική έρευνα μπορούν να υπάρξουν κάποια λανθασμένα αποτελέσματα κάτι που συμβαίνει όσο αντιπροσωπευτικό και να είναι το δείγμα. Συνεπώς σφάλματα (errors) ορίζουμε τις αποκλίσεις των δειγματοληπτικών αποτελεσμάτων από τις πραγματικές τιμές. Το δειγματοληπτικό σφάλμα μιας εκτιμήτριας μετράται με το τυπικό σφάλμα της ή με το συντελεστή μεταβλητότητας της. Σε μια καθολική έρευνα το δειγματοληπτικό σφάλμα είναι μηδέν γιατί έχουμε ένα μοναδικό δείγμα.

<u>Πηγές δειγματοληπτικού σφάλματος</u>

i. **Μέγεθος δείγματος:** Γενικότερα μια αύξηση του μεγέθους του δείγματος έχει ως αποτέλεσμα τη μείωση του δειγματοληπτικού σφάλματος. Σε περίπτωση που στόχος της έρευνας είναι η παρατήρηση υπερπληθυσμών ή σπάνιων χαρακτηριστικών τότε γενικά απαιτείται μεγαλύτερο δείγμα.

ii. **Μέγεθος ερευνώμενου πληθυσμού:** το μέγεθος του πληθυσμού έχει διαφορετικές επιδράσεις στο μέγεθος του δειγματοληπτικού σφάλματος ανάλογα με το μέγεθος του. Έχει μικρή επίδραση για πληθυσμούς μεσαίου μεγέθους και σχεδόν ανύπαρκτη για πληθυσμούς μεγάλου μεγέθους. Επιπλέον, για πολύ μικρούς πληθυσμούς η επιρροή είναι μεγάλη και το δείγμα που απαιτείται είναι σχετικά μεγάλο.

iii. **Πληθυσμιακή διακύμανση:** Όσο μεγαλύτερη είναι η διαφορά μεταξύ των μονάδων του πληθυσμού ως προς τα ερευνώμενα χαρακτηριστικά τόσο μεγαλύτερο είναι το δειγματοληπτικό σφάλμα για δεδομένο μέγεθος δείγματος. Για παράδειγμα σε μια έρευνα ατομικού εισοδήματος το δειγματοληπτικό σφάλμα θα ήταν μεγαλύτερο σε

ένα πληθυσμό όπου τα εισοδήματα θα κυμαίνονταν από 10.000 μέχρι 200.000 ευρώ από ό,τι θα ήταν σε πληθυσμό όπου τα εισοδήματα θα κυμαίνονταν από 30.000 μέχρι 70.000 (με το ίδιο μέγεθος δείγματος και στις 2 περιπτώσεις).

iv. **Σχέδιο δειγματοληψίας και εκτιμήτρια:** Ο συνδυασμός σχεδίου δειγματοληψίας και τύπου εκτιμήτριας σχετίζεται άμεσα με το μέγεθος του δειγματοληπτικού σφάλματος. Ο σχετικός όρος αποτελεσματικότητας ενός σχεδίου δειγματοληψίας ή μίας εκτιμήτριας αναφέρεται στην εκτίμηση παραμέτρων με μικρό δειγματοληπτικό σφάλμα για δεδομένο μέγεθος δείγματος.

1.6. ΜΗ ΔΕΙΓΜΑΤΟΛΗΠΤΙΚΟ ΣΦΑΛΜΑ

Τα μη δειγματοληπτικά σφάλματα (non sampling errors) δεν οφείλονται στη δειγματοληψία αλλά είναι σφάλματα τα οποία μπορούν να γίνουν σε κάθε είδους έρευνα (καθολική ή δειγματοληπτική). Προκύπτουν κυρίως κατά την απογραφή και δεν οφείλονται στην επιλογή του δείγματος.

Τα μη δειγματοληπτικά σφάλματα περιλαμβάνουν δύο είδη σφαλμάτων τα οποία είναι:

- Τα σφάλματα μη περίληψης (non - inclusion errors).
- Τα σφάλματα μη απάντησης (non – response errors).

Τα σφάλματα μη περίληψης εμφανίζονται όταν τα μέλη του αντικειμενικού πληθυσμού (target population) δεν είναι δυνατόν να περιληφθούν στο δείγμα. Για παράδειγμα μια τηλεφωνική έρευνα δεν μπορεί να καλύψει μέλη του πληθυσμού που δεν έχουν τηλέφωνο.

Τα σφάλματα μη απάντησης περιλαμβάνουν κατά κανόνα μέλη του πληθυσμού τα οποία δεν αποδίδουν μια τιμή για μια μεταβλητή Χ που θα θέλαμε να μελετήσουμε. Τα σφάλματα αυτά προκύπτουν από διάφορους λόγους, οι οποίοι συνδέονται με τα χαρακτηριστικά των μονάδων του πληθυσμού καθώς και με τη μέθοδο με την οποία συλλέγεται η πληροφορία. Για παράδειγμα μέθοδοι συλλογής πληροφορίας είναι η συνέντευξη, το ερωτηματολόγιο και η τηλεφωνική ή ταχυδρομική συνέντευξη.

Παράγοντες που προκαλούν μη δειγματοληπτικά σφάλματα:

1. Ακαταλληλότητα του ερωτηματολογίου
2. Σφάλματα ανταπόκρισης
3. Σφάλματα ερευνητή
4. Σφάλματα επεξεργασίας

Το ερωτηματολόγιο αποτελεί τον πιο βασικό παράγοντα για την επιτυχία μια έρευνας. Εάν περιέχει πολύπλοκα ερωτήματα ή ερωτήματα τα οποία θίγουν το ερευνώμενο πρόσωπο το ερωτηματολόγιο θεωρείται ακατάλληλο. Συνεπώς οι απαντήσεις μπορεί να είναι εσφαλμένες ή μεροληπτικές.

Το σφάλμα ανταπόκρισης υφίσταται όταν το ερευνώμενο πρόσωπο μπορεί να δώσει εσφαλμένη απάντηση για διάφορους λόγους όπως να μην κατάλαβε την ερώτηση του ερευνητή, να μην θυμάται κάποιο γεγονός ή να μην επιθυμεί να δώσει σωστή απάντηση. Για παράδειγμα μπορεί να δηλώσει μικρότερη ηλικία από την πραγματική.

Επίσης ορισμένες φορές πραγματοποιούνται σφάλματα και από την πλευρά του ερευνητή, ο οποίο μπορεί να καταχωρήσει στο ερωτηματολόγιο του εσφαλμένα μια

απάντηση αν και αυτή δόθηκε σωστά από τον ανταποκρινόμενο. Για παράδειγμα ο ανταποκρινόμενος δηλώνει ως μέρος κατοικίας την Πάτρα και ο ερευνητής γράφει Αθήνα.

Τέλος, κατά το στάδιο της επεξεργασίας των στοιχείων γίνονται λάθη από τους υπαλλήλους που ασχολούνται με την εργασία αυτή.

Υπάρχει δυσκολία στο να εντοπιστούν τα μη δειγματοληπτικά σφάλματα και να εντοπιστεί η πιθανή αιτία τους. Όμως με τον ορθό σχεδιασμό μιας δειγματοληπτικής έρευνας τα μη δειγματοληπτικά σφάλματα μπορούν να περιοριστούν στο ελάχιστο. Αυτό συμβαίνει με την κατάρτιση ενός κατάλληλου ερωτηματολογίου, την επιλογή ειδικευμένων ερευνητών και τη σωστή εποπτεία στην εργασία αυτών.

Για να αυξήσουμε το βαθμό αξιοπιστίας των αποτελεσμάτων μιας έρευνας πρέπει κατά το σχεδιασμό να λάβουμε υπόψη μας τους παράγοντες σφαλμάτων που αναφέρονται παραπάνω και να προσπαθήσουμε να τα περιορίσουμε με το μικρότερο δυνατό κόστος.

1.7. ΜΕΡΟΛΗΠΤΙΚΗ ΔΕΙΓΜΑΤΟΛΗΨΙΑ

Μεροληπτική δειγματοληψία έχουμε όταν κατά την επιλογή του δείγματος από ένα πληθυσμό σε μερικές μονάδες δίνεται μεγαλύτερη ευκαιρία να συμπεριληφθούν στο δείγμα. Η επιλογή ενός τέτοιου δείγματος μπορεί να γίνει για λόγους ευκολίας ή λόγω άγνοιας του ερευνητή να αποκλείσει τη μεροληψία από το δείγμα. Ένα τέτοιο παράδειγμα δείγματος είναι η επιλογή της περιοχής της Εκάλης για την καταμέτρηση εισοδήματος. Στην περιοχή αυτή κατοικούν άτομα τα οποία έχουν υψηλό εισόδημα. Επομένως η δειγματοληπτική έρευνα σε μια τέτοια περιοχή θα είναι μεροληπτική διότι δεν δίνει ευκαιρία σε άτομα που το εισόδημα τους είναι χαμηλό να επιλεγούν στο δείγμα. Επομένως, συμπεραίνουμε ότι μια έρευνα για να είναι επιτυχής πρέπει να έχει χαμηλή ύπαρξη μεροληψίας ώστε να μην οδηγηθεί σε εσφαλμένα συμπεράσματα.

1.8. ΕΡΩΤΗΜΑΤΟΛΟΓΙΟ

Κατά το σχεδιασμό μιας έρευνας το πρώτο βήμα στη διαδικασία συλλογής στατιστικών στοιχείων είναι η κατάρτιση του ερωτηματολογίου. Το ερωτηματολόγιο αποτελεί μέσο μετασχηματισμού των σκοπών της έρευνας σε ένα εργαλείο συλλογής στοιχείων. Ο σκοπός στο σχεδιασμό του ερωτηματολογίου είναι να συλλεχθούν όλες οι απαιτούμενες πληροφορίες με όσο το δυνατόν μικρότερο σφάλμα και σε κατάλληλη μορφή για περαιτέρω επεξεργασία των στοιχείων. Ένα προβληματικό ερωτηματολόγιο μπορεί να έχει ως αποτελέσματα λανθασμένα ή παραπλανητικά στοιχεία ή ακόμα και άρνηση συμπλήρωσης του.

Πριν από το σχεδιασμό του ερωτηματολογίου λαμβάνονται υπόψη οι σχετικές προδιαγραφές του προγραμματισμού της έρευνας. Αρχικά, πρέπει να καθορίσουμε το σκοπό της έρευνας που να περιλαμβάνει τα απαιτούμενα στοιχεία και να βρούμε ένα πρόγραμμα για την περαιτέρω ανάλυση τους. Αυτό θα καθορίζει τις πληροφορίες που χρειάζονται, τις μεταβλητές της έρευνας και πως συνδέεται κάθε ζητούμενο στοιχείο με συγκεκριμένες ερωτήσεις. Έπειτα, πριν ξεκινήσει η σύνταξη συγκεκριμένων ερωτήσεων πρέπει να αποφασιστεί αν το ερωτηματολόγιο θα είναι αυτοσυμπληρούμενο ή όχι. Στην περίπτωση που το ερωτηματολόγιο είναι αυτόσυμπληρούμενο πρέπει να αποφασιστεί ο τρόπος επίδοσης του (ταχυδρομικά, με e-mail, με fax, τηλεφωνικά ή προσωπικά). Αντιθέτως, στην περίπτωση μη

αυτοσυμπληρούμενου ερωτηματολογίου πρέπει να ληφθεί υπόψη αν η συλλογή στοιχείων θα γίνει με προσωπική συνέντευξη ή τηλεφωνικά με τη βοήθεια υπολογιστή. Επειδή όμως, ο τρόπος συλλογής στοιχείων καθορίζει πως θα δομηθούν οι ερωτήσεις, η συγκεκριμένη απόφαση πρέπει να ληφθεί νωρίς στο σχεδιασμό.

Αρχές κατάρτισης ερωτηματολόγιου

Η κατάρτιση του ερωτηματολογίου πρέπει να ξεκινάει από την αρχή του σχεδιασμού της έρευνας και να ολοκληρώνεται μετά από τη διεξαγωγή μιας δοκιμαστικής έρευνας. Κατά το σχεδιασμό του ερωτηματολογίου πρώτο μας μέλημα είναι να αποφασισθεί το είδος και ο αριθμός των ερωτημάτων τα οποία πρέπει να καλύπτουν τα ενδιαφέροντα χαρακτηριστικά. Ειδικότερα πρέπει να λαμβάνονται υπόψη τα εξής:

i. Να αποφευχθούν τα μακροσκελή ερωτηματολόγια διότι είναι κουραστικά τόσο για τον ερευνητή όσο και για τον ερευνώμενο. Για τον λόγο αυτό πρέπει να τίθενται μόνο τα απαραίτητα ερωτήματα για την κάλυψη του σκοπού μας.
ii. Τα ερωτήματα πρέπει να είναι πρακτικά. Ο ερευνώμενος δεν πρέπει να ερωτάται για γεγονότα του παρελθόντος διότι θα μπορέσει να δώσει ακριβή απάντηση. Επιπρόσθετα δεν πρέπει να γίνονται ερωτήσεις για προσωπικές υποθέσεις γιατί είναι πιθανό το άτομο να δώσει μεροληπτική απάντηση.
iii. Οι ερωτήσεις πρέπει να είναι σαφείς, όσο το δυνατόν συγκεκριμένες, εύκολες να απαντηθούν και ενδιαφέρουσες για τους αποκρινόμενους.
iv. Κατά τη σύνταξη των ερωτημάτων πρέπει να λαμβάνεται υπόψη το επίπεδο εκπαίδευσης των ερευνωμένων. Τα ερωτήματα πρέπει να διατυπώνονται με τέτοιο τρόπο ώστε διευκολύνεται η επικοινωνία με τον ερευνώμενο. Για παράδειγμα όταν το ερωτηματολόγιο απευθύνεται στο ευρύ κοινό δεν πρέπει να πραγματοποιείται χρήση εξειδικευμένων όρων γιατί υπάρχουν και άτομα χαμηλού επιπέδου εκπαίδευσης.
v. Η ερώτηση πρέπει να καθορίζει πλαίσιο και χρόνο αναφοράς. Για παράδειγμα στην ερώτηση «Ποιό είναι το εισόδημα σας;», η λέξη «σας» μπορεί να αναφέρεται στο ατομικό ή οικογενειακό εισόδημα του αποκρινόμενου. Η λέξη εισόδημα μπορεί να αναφέρεται σε μισθό ή να περιλαμβάνει εισόδημα από άλλες πηγές.
vi. Η σειρά που τίθενται τα ερωτήματα επηρεάζει άμεσα το βαθμό ανταπόκρισης σε μια έρευνα. Αρχικά, πρέπει να αναφέρονται τα ερωτήματα που ενδιαφέρουν ουσιαστικά τον ερευνώμενο και αυτά που προϋποθέτουν εύκολες απαντήσεις.
vii. Επίσης, πρέπει να αποφευχθούν οι συντομογραφίες και τα αρκτικόλεξα όπως και δυσνόητες λέξεις.

1.9. ΜΕΘΟΔΟΙ ΣΥΛΛΟΓΗΣ ΣΤΟΙΧΕΙΩΝ

Η συλλογή στοιχείων είναι η διαδικασία συγκέντρωσης των απαιτούμενων πληροφοριών της έρευνας για κάθε μονάδα του δείγματος. Σημαντικό ρόλο στη δειγματοληψία παίζει η συλλογή στοιχείων διότι αποτελεί χρονοβόρα διαδικασία, απορροφά ένα μεγάλο μέρος του συνολικού κόστους της έρευνας και απαιτεί ανθρώπινο δυναμικό και υλικούς πόρους. Κάποιες από τις πιο σημαντικές μεθόδους συλλογής στατιστικών στοιχείων είναι η απογραφή και η δειγματοληπτική μέθοδος.

1.9.1. *ΑΠΟΓΡΑΦΗ*

Πρόκειται για μια καθολική μέθοδο συγκέντρωσης πληροφοριών από όλες τις στατιστικές μονάδες του πληθυσμού. Υπάρχουν πολλές μορφές απογραφών (όπως η απογραφή εμπορικών επιχειρήσεων, γεωργίας, κτηνοτροφίας), με την απογραφή του πληθυσμού να θεωρείται η σπουδαιότερη, γιατί αποτελεί την κύρια πηγή πληροφοριών πάνω στην άποψη των δημογραφικών, οικονομικών και κοινωνικών χαρακτηριστικών.

Με τη μέθοδο της απογραφής τα κυριότερα χαρακτηριστικά του πληθυσμού που μελετάμε, είναι:

i. η σύνθεση του πληθυσμού κατά ηλικία,
ii. η οικογενειακή κατάσταση (παντρεμένοι, ανύπαντροι, χωρισμένοι, χήροι),
iii. η σύνθεση κατά φύλο,
iv. η σύνθεση κατά επάγγελμα,
v. η ανεργία και η απασχόληση,
vi. η εκπαίδευση,
vii. η φυσική κίνηση του πληθυσμού.

Η απογραφή του πληθυσμού γίνεται από τη στατιστική υπηρεσία της κάθε χώρας και συνήθως η διεξαγωγή της πραγματοποιείται κάθε 10 χρόνια, στα έτη που λήγουν σε 0 ή 1. Επίσης, διαπιστώνουμε ότι οι απογραφές γίνονται κατά την άνοιξη ή το φθινόπωρο καθώς την συγκεκριμένη περίοδο παρατηρείται μειωμένη τουριστική κίνηση του πληθυσμού και λιγότερες πληθυσμιακές μετακινήσεις λόγω των γεωργικών απασχολήσεων. Απαιτείται τουλάχιστον μία μέρα για τη συμπλήρωση των απογραφικών δελτίων και προτιμάται η Κυριακή ως ημέρα διεξαγωγής.

Παρόλη τη σημασία της απογραφής ως μέθοδο συλλογής στοιχείων, παρουσιάζει τα εξής μειονεκτήματα:

i. Απαιτεί μεγάλο κόστος.
ii. Λόγω του μεγάλου πλήθους των πληροφοριών, η δημοσίευση των αποτελεσμάτων καθυστερεί.
iii. Επειδή η απογραφή δεν γίνεται από ειδικευμένο προσωπικό, μπορεί να υπάρχουν σφάλματα επομένως να οδηγηθούμε σε λανθασμένη εικόνα των χαρακτηριστικών του πληθυσμού.

Όταν μόνο ένα τμήμα του πληθυσμού μετράται τότε πρόκειται για ***μερική απογραφή*** όπως: οι γεννήσεις και οι θάνατοι σε ένα γεωγραφικό διαμέρισμα και η μετανάστευση σε ορισμένες μόνο περιοχές. Με την παραπάνω έννοια οι μερικές απογραφές διαφέρουν από τις δειγματοληπτικές έρευνες.

ΔΕΙΓΜΑΤΟΛΗΠΤΙΚΗ ΜΕΘΟΔΟΣ

Συνήθως ο πληθυσμός που θέλουμε να μετρήσουμε από την άποψη ορισμένων ιδιοτήτων αποτελείται από μεγάλο πλήθος στατιστικών μονάδων. Με τη συγκεκριμένη μέθοδο γίνεται προσπάθεια να γνωρίσουμε τις ιδιότητες ενός πληθυσμού εξετάζοντας από αυτόν μόνο ένα δείγμα. Βασική προϋπόθεση είναι η επιλογή του δείγματος να γίνει κατά τέτοιο τρόπο ώστε οι εκτιμήσεις και τα συμπεράσματα που θα λάβουμε από αυτό να έχουν ισχύ για το σύνολο του πληθυσμού στον οποίο ανήκει το δείγμα.

Η δειγματοληπτική μέθοδος έχει τα παρακάτω πλεονεκτήματα σε σύγκριση με τις απογραφές:

i. Μεγαλύτερη ταχύτητα πληροφοριών

Η απογραφή υστερεί σε σχέση με την δειγματοληπτική μέθοδο στο ότι η συλλογή και η επεξεργασία των στατιστικών στοιχείων γίνεται με πιο αργούς ρυθμούς. Αυτό έχει ιδιαίτερη σημασία όταν πρόκειται για συγκέντρωση πληροφοριών που επείγουν. Για παράδειγμα βάση των δημοσκοπήσεων προβλέπεται ποιος θα κερδίσει στις προσεχείς εκλογές.

ii. Μεγαλύτερη ακρίβεια

Στις δειγματοληπτικές έρευνες όταν ο αριθμός των μονάδων είναι μικρός, είναι εφικτό να αφιερωθεί περισσότερος χρόνος και μεγαλύτερη προσοχή στις συνεντεύξεις που παίρνουμε με σκοπό την καλύτερη εκπαίδευση και επίβλεψη στους απογραφείς ώστε να έχουμε μεγαλύτερη ακρίβεια στις πληροφορίες. Ιδιαίτερη σημασία έχει το συγκεκριμένο γεγονός, γιατί παρά τα δειγματοληπτικά σφάλματα μία δειγματοληπτική έρευνα που έχει διεξαχθεί με ορθό τρόπο μπορεί να δώσει πιο ακριβή αποτελέσματα σε σχέση με αυτά της καθολικής απογραφής.

iii. Χαμηλό κόστος

Σκοπός κάθε δειγματοληπτικής έρευνας είναι η λήψη μιας πληροφορίας με τη μεγαλύτερη δυνατή ακρίβεια και το ελάχιστο δυνατό κόστος. Είναι επόμενο ότι κοστίζει λιγότερο η συγκέντρωση και επεξεργασία πληροφοριών από μερικές μονάδες του ερευνώμενου πληθυσμού, παρά από το σύνολο του, για παράδειγμα η συλλογή πληροφοριών από 100 οικογένειες μια πόλης κοστίζει λιγότερο σε σχέση με αυτήν από 10.000 οικογένειες. Ο λόγος είναι ότι χρησιμοποιούνται λιγότεροι ερευνητές, λιγότεροι υπάλληλοι και λιγότερα έντυπα.

iv. Μεγαλύτερη ευχέρεια εφαρμογής

Εφαρμόζεται στις περιπτώσεις που η γενική απογραφή αν και είναι δυνατή θεωρείται παράλογη. Παραδείγματος χάρη, όταν ένας γιατρός θέλει να εξετάσει τα αιμοσφαίρια ενός ασθενή, δεν θα πάρει όλη την ποσότητα αίματος του ασθενούς (καθολική έρευνα), αλλά λίγα μόνο γραμμάρια (δειγματοληψία).

v. Ολοκληρωτική δύναμη εφαρμογής της γενικής απογραφής

Όταν η καθολική έρευνα είναι αδύνατο να επιτευχθεί, τότε εφαρμόζεται η δειγματοληπτική έρευνα. Για παράδειγμα, αν θέλουμε να ερευνήσουμε, αν μια θαλάσσια

περιοχή έχει πετρέλαιο τότε περιοριζόμαστε σε μία μερική έρευνα γιατί είναι αδύνατη η καθολική.

Παράλληλα με τα πλεονεκτήματα της δειγματοληψίας υπάρχουν και μειονεκτήματα, τα οποία θα αναφέρουμε παρακάτω:

- Στην περίπτωση που οι μονάδες του πληθυσμού που ερευνούμε δεν εμφανίζονται συχνά τότε θεωρείται αναγκαίο να πάρουμε ένα αρκετά μεγάλο δείγμα, στο οποίο θα περιλαμβάνεται ένας ικανοποιητικός αριθμός από την κατηγορία των μονάδων που θέλουμε να μελετήσουμε, με στόχο να πετύχουμε αξιόπιστες απαιτήσεις. Σε αντίθετη περίπτωση δεν είναι δυνατό να διενεργηθεί δειγματοληπτική έρευνα. Παραδείγματος χάρη, αν επιθυμούμε να μελετήσουμε το ποσοστό των καπνιστών από άτομα μιας πόλης ηλικίας άνω των 80 ετών, πρέπει να πάρουμε ένα αρκετά μεγάλο δείγμα νοικοκυριών ώστε να περιλαμβάνεται σε αυτά ικανοποιητικός αριθμός ατόμων γιατί τα άτομα της ηλικίας αυτής αποτελούν μικρό ποσοστό στο συνολικό πληθυσμό της πόλης.

- Στην περίπτωση που ο πληθυσμός παρουσιάζει μεγάλη ανομοιογένεια τότε πρέπει να χωρισθεί σε υποπληθυσμούς, και να πάρουμε δείγμα. Σε αντίθετη περίπτωση οι εκτιμήσεις που θα πάρουμε δεν θα είναι αξιόπιστες. Για παράδειγμα εάν θέλουμε να εκτιμήσουμε τον αριθμό πορτοκαλιών σε μια περιοχή, με τη βοήθεια δειγματοληπτικής έρευνας, πρέπει να χωρίσουμε τους κατόχους σε τάξεις σύμφωνα με τον αριθμό των πορτοκαλιών που έχει στην κατοχή του ο καθένας τους και να πάρουμε δείγμα μέσα από κάθε τάξη. Εάν δεν χωρισθούν οι κάτοχοι υπάρχει κίνδυνος υποεκτίμησης ή υπερεκτίμησης του δείγματος.

- Εάν η θεωρητική διαδικασία για τον προσδιορισμό του μεγέθους του δείγματος και την εφαρμογή της κατάλληλης μεθόδου της δειγματοληψίας δεν ακολουθηθεί πιστά τότε τα αποτελέσματα που θα προκύψουν θα είναι αναξιόπιστα. Επομένως θα οδηγηθούμε σε λανθασμένα συμπεράσματα για τον συγκεκριμένο πληθυσμό.

- Τέλος, άλλο ένα βασικό μειονέκτημα της δειγματοληψίας είναι τα δειγματοληπτικά σφάλματα στα οποία αναφερθήκαμε παραπάνω.

1.9.2. *ΠΡΟΣΩΠΙΚΗ ΣΥΝΕΝΤΕΥΞΗ*

Η προσωπική συνέντευξη (personal interview) είναι από τις πιο διαδεδομένες μεθόδους συλλογής στοιχείων στατιστικού υλικού. Κυρίως χρησιμοποιείται για θέματα κοινωνικού και οικονομικού προβληματισμού. Η συγκεκριμένη μέθοδος στηρίζεται σε ένα τουλάχιστον ερευνητή (interviewer) ο οποίος προσεγγίζει τους ερωτώμενους, υποβάλλει προς αυτούς τις κατάλληλες ερωτήσεις και καταγράφει τις απαντήσεις τους με τη χρήση ενός ειδικού εντύπου ερωτηματολογίου.

Η συλλογή στοιχείων κατά την προσωπική συνέντευξη μπορεί να γίνει είτε πρόσωπο με πρόσωπο είτε με τη βοήθεια ηλεκτρονικού υπολογιστή. Σύμφωνα με την πρώτη κατηγορία, εκπαιδευμένοι συνεντευκτές επισκέπτονται τη δειγματοληπτική μονάδα για να συλλέξουν στοιχεία χρησιμοποιώντας ερωτηματολόγιο. Στην περίπτωση που η συνέντευξη γίνει με τη χρήση ηλεκτρονικού υπολογιστή η χρήση ερωτηματολογίου δεν είναι απαραίτητη αφού ο συνεντευκτής φέρει μαζί του ένα φορητό υπολογιστή με το οποίο εισάγει τις πληροφορίες κατευθείαν στη βάση δεδομένων. Η μέθοδος αυτή συμφέρει για επαναληπτικές έρευνες.

Η προσωπική συνέντευξη αποτελεί τον καλύτερο τρόπο συλλογής στατιστικών στοιχείων και έχει τα εξής πλεονεκτήματα:

- Συλλέγονται καλύτερης ποιότητας στοιχεία σε σχέση με τη μέθοδο της ταχυδρομικής αποστολής, γιατί δίνονται διευκρινήσεις στους ερωτώμενους.
- Οι ερωτώμενοι ανταποκρίνονται σχεδόν πάντα, έχουμε απόκριση του κοινού μέχρι και 100%.
- Στην περίπτωση που η συλλογή στοιχείων γίνει με την χρήση ηλεκτρονικού υπολογιστή, στοιχείων υπάρχει μεγάλη εξοικονόμηση χρόνου αφού ο ερευνητής δεν χρειάζεται να εισάγει τα στοιχεία σε μαγνητικά μέσα από το ερωτηματολόγιο που θα συμπλήρωνε με το χέρι. Επίσης περιορίζονται τα σφάλματα επεξεργασίας (κωδικογράφηση και εισαγωγή στοιχείων σε μαγνητικά μέσα).

Εν αντιθέσει, η μέθοδος της προσωπικής συνέντευξης έχει και κάποια μειονεκτήματα όπως ότι απαιτεί μεγαλύτερο κόστος σε σύγκριση με την ταχυδρομική αποστολή. Επιπρόσθετα, ορισμένες φορές η παρουσία του ερευνητή οδηγεί τον ερευνώμενο στο να δώσει μεροληπτικές απαντήσεις.

1.9.3. *ΤΑΧΥΔΡΟΜΙΚΗ ΑΠΟΣΤΟΛΗ*

Με τη μέθοδο του ταχυδρομείου έντυπα ερωτηματολόγια στέλνονται ταχυδρομικά στις μονάδες του δείγματος. Τα ερωτηματολόγια αυτά μελετούνται και συμπληρώνονται από τους παραλήπτες οι οποίοι είτε τα επιστρέφουν με το ταχυδρομείο στον αρχικό αποστολέα είτε ειδικός συνεργάτης επισκέπτεται τις μονάδες του δείγματος προκειμένου να παραλάβει τα συγκεκριμένα ερωτηματολόγια. Θεωρείται ως ο πιο εύκολος τρόπος συλλογής στατιστικών στοιχείων αλλά η εφαρμογή της μεθόδου αυτής είναι αρκετά περιορισμένη λόγο των μειονεκτημάτων της. Σε άλλες όμως περιπτώσεις, είναι μοναδικός ο τρόπος συλλογής των στοιχείων όπως για παράδειγμα συμβαίνει όταν ο ερευνώμενος πληθυσμός είναι διασκορπισμένος σε μεγάλες εκτάσεις και οι χρηματικοί πόροι περιορισμένοι.

Η συγκεκριμένη μέθοδος έχει τόσα μειονεκτήματα όσο και πλεονεκτήματα. Αρχικά τα πλεονεκτήματα είναι:

- Το κόστος συλλογής πληροφοριών είναι χαμηλό αφού ουσιαστικά καλύπτει μόνο τη δαπάνη για τους φακέλους και τα γραμματόσημα.
- Ο χρόνος που απαιτείται για να φθάσει το ερωτηματολόγιο στις μονάδες του δείγματος είναι πολύ μικρός, αφού η μετακίνηση γίνεται ταυτόχρονα προς όλους τους παραλήπτες με όλα τα διαθέσιμα μέσα (αεροπλάνο, τραίνο, πλοίο, αυτοκίνητο).
- Μείωση των σφαλμάτων που οφείλονται στην παρουσία του συνεντευκτή ή ερευνητή, οι οποίοι επηρεάζουν τους ερωτώμενους.
- Ο ερωτώμενος έχει άνεση χρόνου να συμπληρώσει το ερωτηματολόγιο, αφού δεν έχει κάποιο πρόσωπο που να του ζητάει άμεσες απαντήσεις.
- Ορισμένες ερωτήσεις προσωπικού χαρακτήρα είναι πιο εύκολο να απαντηθούν μέσω της συγκεκριμένης μεθόδου σε σχέση με την προσωπική συνέντευξη.
- Αποφυγή νέας επίσκεψης η οποία γίνεται από τον ερευνητή σε περίπτωση που ο ερωτώμενος απουσιάζει ή είναι απασχολημένος.

Από την άλλη μεριά, υπάρχουν και τα εξής μειονεκτήματα:

- Χαμηλό ποσοστό απαντήσεων αφού οι ερωτώμενοι είτε δεν συμπληρώνουν τα ερωτηματολόγια είτε δεν τα επιστρέφουν στον αποστολέα, αυτό γίνεται επειδή δεν υπάρχει ο συνεντευκτής ώστε να πιέσει τον ερωτώμενο να δώσει μια απάντηση.
- Σοβαρός λόγος αδυναμίας στη συλλογή πληροφοριών με τη συγκεκριμένη μέθοδο είναι η αγραμματοσύνη. Αν και οι αγράμματοι έχουν μειωθεί σε μεγάλο βαθμό στην Ελλάδα, η δυσχέρεια αυτή δεν παύει να υφίσταται για ορισμένα άτομα (για παράδειγμα ηλικιωμένους).
- Με την ταχυδρομική αποστολή υπάρχει δυσκολία στη συλλογή των γνωστών πρόσθετων πληροφοριών που συλλέγει ο συνεντεύκτης για τη σωστή αξιολόγηση της αξιοπιστίας του ερωτώμενου προσώπου (για παράδειγμα το ύφος που ο ερωτώμενος απαντά).

1.9.4. *ΤΗΛΕΦΩΝΙΚΗ ΣΥΝΔΙΑΛΕΞΗ*

Εάν όλες οι μονάδες του πληθυσμού που θέλουμε να εξετάσουμε διαθέτουν τηλέφωνο, τότε είναι δυνατό να προσεγγισθούν όσες από αυτές συμπεριλαμβάνονται στο δείγμα χρησιμοποιώντας την τηλεφωνική επικοινωνία. Στην περίπτωση αυτή μπορούμε να σχηματίσουμε το αναγκαίο δείγμα με τη βοήθεια του τηλεφωνικού καταλόγου. Επιπλέον μπορούμε να επιλέξουμε ένα τυχαίο δείγμα με τη χρήση του ηλεκτρονικού υπολογιστή, ο οποίος με ένα κατάλληλο πρόγραμμα μπορεί να σχηματίσει το επιθυμητό πλήθος αριθμών τηλεφώνου με τυχαία διαδικασία (random number generator).

Βασικά πλεονεκτήματα του συγκεκριμένου τρόπου συλλογής πληροφοριών είναι το πολύ χαμηλό κόστος και η τεράστια εξοικονόμηση χρόνου. Όμως υπάρχουν και αρκετά μειονεκτήματα όπως το ενδεχόμενο να μην έχουν όλα τα νοικοκυριά τηλέφωνο και να μην μπορούν να συμπεριληφθούν στο δείγμα. Ακόμα και αν περιληφθούν σε αυτό μπορεί τα άτομα να μην είναι διατεθειμένα να απαντήσουν στις ερωτήσεις του ερευνητή σκεπτόμενα ότι πρόκειται για φάρσα. Άλλο ένα μειονέκτημα είναι ότι ακόμα και αν ο ερωτώμενος είναι πρόθυμος να συνεργαστεί δεν είναι δυνατόν να χρησιμοποιηθεί μεγάλο ερωτηματολόγιο επειδή αυτό μπορεί να κουράσει το ερωτώμενο πρόσωπο. Τέλος, γνωστό είναι το πρόβλημα

της μεροληψίας που προκύπτει από τη χρήση τηλεφωνικού καταλόγου, ο οποίος δεν είναι ενημερωμένος.

1.9.5. *ΠΑΡΑΤΗΡΗΣΗ*

Η παρατήρηση αποτελεί μέθοδο συγκέντρωσης στοιχείων και ενώ χρησιμοποιείται αρκετά στις φυσικές επιστήμες είναι λιγότερο διαδεδομένη στις στατιστικές έρευνες. Χρησιμοποιείται κυρίως στην έρευνα της συμπεριφοράς του πληθυσμού μιας χώρας με την προϋπόθεση ότι εφαρμόζονται κάποιες οδηγίες, παραδείγματος χάρη, η πρόληψη των τροχαίων ατυχημάτων και η αξιολόγηση της, η οποία συμπεριλαμβάνει παρατηρήσεις στους δρόμους για τυχόν παραβιάσεις του Κώδικα Οδικής Κυκλοφορίας από τους οδηγούς αμαξωμάτων και τους πεζούς.

ΚΕΦΑΛΑΙΟ ΔΕΥΤΕΡΟ : ΠΙΘΑΝΟΤΗΤΕΣ

2.1. ΕΙΣΑΓΩΓΗ

Ο κλάδος των πιθανοτήτων και της στατιστικής είναι μια ανερχόμενη δύναμη τα τελευταία χρόνια. Αυτό προκύπτει, από τον μεγάλο αριθμό των εφαρμογών των πιθανοτήτων και της στατιστικής σε πολλούς άλλους κλάδους επιστημονικής έρευνας και προβλημάτων της καθημερινής ζωής. Η στατιστική επιχειρηματολογία χρησιμοποιείται όλο και περισσότερο από τις επιστήμες διαφόρων ειδών (φυσικές, κοινωνικές, βιολογικές) για την αντιμετώπιση και τη λύση των προβλημάτων. Επιπλέον η θεωρία των πιθανοτήτων και η στατιστική θεωρούνται απαραίτητα εργαλεία στην ιατρική, στις οικονομικές επιστήμες, στην εγκληματολογία, κλπ.

Γενικά, η λέξη πιθανότητα έχει πολλές έννοιες, πιο συχνά όμως χρησιμοποιείται για να αναφερθούμε στο ενδεχόμενο πραγματοποίησης κάποιου γεγονότος. Η θεωρία πιθανοτήτων όπως και στατιστική ασχολούνται με τυχαία φαινόμενα ή τυχαία πειράματα.

2.2. ΠΕΙΡΑΜΑ ΤΥΧΗΣ – ΔΕΙΓΜΑΤΙΚΟΣ ΧΩΡΟΣ

Κάθε πείραμα έχει ένα πιθανό σύνολο δυνατών αποτελεσμάτων τα οποία μπορούν να εμφανιστούν σε μια εκτέλεση του, το οποίο και ονομάζουμε δειγματικό χώρο.

Τα στοιχεία ενός δειγματικού χώρου ονομάζονται δειγματικά σημεία. Πιο αναλυτικά τα υποσύνολα του δειγματικού χώρου Ω λέγονται ενδεχόμενα ή γεγονότα και συμβολίζονται με κεφαλαία γράμματα Α,Β,Γ… Τα ενδεχόμενα που αποτελούνται από μόνο ένα δειγματικό σημείο καλούνται απλά ή στοιχειώδη ενδεχόμενα. Ενώ στην περίπτωση που το ενδεχόμενο Α περιέχει περισσότερα από ένα στοιχεία του δειγματικού χώρου τότε το ενδεχόμενο λέγεται σύνθετο.

Στην Θεωρία Πιθανοτήτων, ως πείραμα τύχης ορίζουμε μια διαδικασία η οποία μπορεί να επαναληφθεί θεωρητικά άπειρες φορές, κάτω από τις ίδιες συνθήκες, και στο τέλος της οποίας παρατηρούμε ορισμένα αποτελέσματα. Για παράδειγμα, ρίχνουμε ένα ζάρι (κύβο) και καταγράφουμε την ένδειξη της επάνω έδρας του. Τα δυνατά αποτελέσματα είναι: 1,2,3,4,5 ή 6 και επομένως ο αντίστοιχος δειγματικός χώρος θα είναι:

$\Omega=\{1,2,3,4,5,6\}$.

Οι αριθμοί 1,2,3,4,5,6 αποτελούν τα δειγματικά σημεία του χώρου ενώ τα σύνολα

$A_1=\{1\}$, $A_2=\{2\}$, $A_3=\{3\}$, $A_4=\{4\}$, $A_5=\{5\}$, $A_6=\{6\}$

είναι τα στοιχειώδη ενδεχόμενα του Ω. Ένα σύνθετο ενδεχόμενο δένεται από το υποσύνολο:

$B_1=\{2,4,6\}$

Το οποίο ονομάζεται σύνολο των άρτιων ενδείξεων ή το ενδεχόμενο εμφάνισης άρτιας ένδειξης. Σε περίπτωση που θέλουμε να εμφανίσουμε ένδειξη μεγαλύτερη ή ίση του 3 έχουμε:

$B_2=\{3,4,5,6\}$.

2.3. ΕΝΔΕΧΟΜΕΝΑ

Όταν ο δειγματικός χώρος Ω ενός πειράματος τύχης είναι διακριτός, τότε κάθε υποσύνολο του θεωρείται ότι είναι ένα ενδεχόμενο.

Έστω $A\subseteq\Omega$ ένα ενδεχόμενο του δειγματικού χώρου Ω. Αν το αποτέλεσμα που πήραμε σε μια επανάληψη του πειράματος ήταν ένα δειγματικό σημείο ω $\in$ στο Ω το οποίο ανήκει στο Α, τότε θα λέμε ότι το ενδεχόμενο Α συνέβη.

2.3.1. ΊΣΑ ΕΝΔΕΧΟΜΕΝΑ

Δύο ενδεχόμενα Α,Β λέγονται ίσα αν κάθε φορά που εμφανίζεται το Α, εμφανίζεται και το Β και αντίστροφα.

Για τα ίσα ενδεχόμενα χρησιμοποιούμε το συμβολισμό Α=Β. Είναι φανερό ότι η ισότητα 2 ενδεχομένων συνεπάγεται την ισχύ των σχέσεων $A\subseteq B$ και $B\subseteq A$.

2.3.2. ΈΝΩΣΗ ΕΝΔΕΧΟΜΕΝΩΝ

Ένωση δύο ενδεχομένων Α,Β λέγεται το ενδεχόμενο που πραγματοποιείται όταν πραγματοποιηθεί ένα τουλάχιστον από τα ενδεχόμενα αυτά και συμβολίζεται με $A\cup B$. Η πράξη της ένωσης μπορεί να επεκταθεί και για περισσότερα από 2 σύνολα.

2.3.3. ΤΟΜΗ ΕΝΔΕΧΟΜΕΝΩΝ

Τομή δύο ενδεχομένων Α,Β λέγεται το ενδεχόμενο που πραγματοποιείται όταν πραγματοποιηθούν ταυτόχρονα και τα δύο ενδεχόμενα αυτά και συμβολίζεται με $A \cap B$ η AB. Όπως και στην ένωση, έτσι και στην πράξη της τομής ενδεχομένων, μπορεί να επεκταθεί και περισσότερα από δύο σύμβολα.

2.3.4.ΞΕΝΑ Η ΑΣΥΜΒΙΒΑΣΤΑ ΕΝΔΕΧΟΜΕΝΑ

Αν δύο ενδεχόμενα δεν μπορούν να πραγματοποιηθούν συγχρόνως, τότε λέγονται ξένα η ασυμβίβαστα ενδεχόμενα. Δηλαδή ισχύει ότι $AB=\emptyset$ (η τομή τους είναι το αδύνατο ενδεχόμενο).

2.3.5. ΣΥΜΠΛΗΡΩΜΑΤΙΚΟ ΕΝΔΕΧΟΜΕΝΟ

Συμπληρωματικό ενδεχόμενο του ενδεχομένου Α λέγεται το ενδεχόμενο που πραγματοποιείται αν και μόνο αν δεν πραγματοποιείται το ενδεχόμενο Α και συμβολίζεται με A^C.

ΙΔΙΟΤΗΤΕΣ ΤΩΝ ΠΡΑΞΕΩΝ

1. $A \cup A = A$
2. $A \cup \emptyset = A$
3. $A \cup \Omega = \Omega$
4. $A \cup B = B \cup A$
5. $A \cup (B \cup \Gamma) = (A \cup B) \cup \Gamma$
6. $A \cup (B\Gamma) = (A \cup B)\ (A \cup \Gamma)$
7. $A \cup A' = \Omega$
8. $(A')' = A$
9. Αν $A \subseteq B$ και $B \subseteq \Gamma$ τότε $A \subseteq \Gamma$
10. Αν $A \subseteq B$ τότε $B' \subseteq A'$ και αντίστροφα.
11. Αν $A \subseteq B$ τότε AB=A και $A \cup B = B$

2.4. ΟΡΙΣΜΟΣ ΠΙΘΑΝΟΤΗΤΑΣ

Πριν αναφερθούμε εκτενώς με τον ορισμό της πιθανότητας που θα χρησιμοποιήσουμε στη συνέχεια είναι σκόπιμο να αναφέρουμε ότι δεν υπάρχει κάποιος ορισμός της πιθανότητας που να είναι αποδεκτός. Η έννοια της πιθανότητας είναι μια φιλοσοφική έννοια και έχει γίνει αφορμή διαφορετικών φιλοσοφικών θεωρήσεων και ισχυρών αντιθέσεων.

2.4.1. ΚΛΑΣΙΚΟΣ ΟΡΙΣΜΟΣ

Έστω ότι ο δειγματικός χώρος Ω ενός πειράματος είναι πεπερασμένος και όλα τα απλά στοιχειώδη ενδεχόμενα του είναι ισοπίθανα, τότε η πιθανότητα εμφάνισης ενός ενδεχομένου Α δίνεται από τον τύπο:

$$P(A) = \frac{A}{\Omega} = \frac{\text{πλήθος στοιχείων του } A}{\text{πλήθος στοιχείων του } \Omega}$$

Το Α αντιπροσωπεύει τον αριθμό των ευνοϊκών περιπτώσεων του ενδεχομένου και το Ω τον αριθμό όλων των δυνατών περιπτώσεων.

Σύμφωνα με τα παραπάνω μπορούμε να καταλήξουμε στον παρακάτω ορισμό της πιθανότητας τον οποίο πρότεινε (το 1812) ο Laplace και αναφερόμαστε σε αυτόν με την ονομασία "κλασικός ορισμός της πιθανότητας":

Πιθανότητα να πραγματοποιηθεί ένα ενδεχόμενο Α είναι ο λόγος του πλήθους των ευνοϊκών περιπτώσεων του ενδεχομένου Α προς το πλήθος όλων των δυνατών περιπτώσεων, με τον όρο ότι οι περιπτώσεις είναι ισοπίθανες.

Πολλές φορές, τα στοιχεία του ενδεχομένου Α ονομάζονται ευνοϊκές περιπτώσεις ή ευνοϊκά αποτελέσματα, ενώ τα στοιχεία του δειγματικού χώρου Ω ονομάζονται δυνατές περιπτώσεις ή δυνατά αποτελέσματα. Έτσι ο τύπος του κλασικού ορισμού της πιθανότητας παίρνει της εξής μορφή:

$$P(A) = \frac{\text{πλήθος ευνοικών αποτελεσμάτων για το ενδεχόμενο } A}{\text{πλήθος δυνατών αποτελεσμάτων}}$$

Παράδειγμα

Ρίχνουμε δυο κανονικά ζάρια. Να υπολογιστεί η πιθανότητα των παρακάτω ενδεχομένων:

α) το άθροισμα των 2 ενδείξεων να είναι μεγαλύτερο ή ίσο του 9

β)το άθροισμα των 2 ενδείξεων να είναι 5

γ) να μην εμφανιστεί σε κανένα ζάρι ο αριθμός 1

Λύση

Ο δειγματικός χώρος που αναφέρεται στη ρίψη των 2 ζαριών είναι:

B	A					
	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

α) Το ενδεχόμενο Α είναι:

Α={(3,6),(4,5),(4,6),(5,4),(5,5),(5,6),(6,3),(6,4),(6,5),(6,6)}

Η πιθανότητα του ενδεχομένου Α, σύμφωνα με τον κλασσικό ορισμό της πιθανότητας είναι:

$$P\{A\} = \frac{A}{\Omega} = \frac{10}{36} = 0{,}277$$

β) το ενδεχόμενο Β είναι:

Β={(1,4),(2,3),(3,2),(4,1)}

Η πιθανότητα του ενδεχομένου Β, σύμφωνα με τον κλασσικό ορισμό της πιθανότητας είναι:

$$P\{B\} = \frac{A}{\Omega} = \frac{4}{36} = 0{,}111$$

γ) Η πιθανότητα του ενδεχομένου Γ θα είναι:

$$P\{\Gamma\} = \frac{A}{\Omega} = \frac{25}{36} = 0{,}694$$

2.4.2. ΣΤΑΤΙΣΤΙΚΟΣ ΟΡΙΣΜΟΣ

Έστω Ω ένας δειγματικός χώρος και Α ένα ενδεχόμενο του Ω. Αν ν_A είναι ο αριθμός εμφανίσεων του ενδεχομένου Α σε ν επαναλήψεις του πειράματος, τότε χρησιμοποιούμε την προσέγγιση που πρότεινε ο Von Mises ο οποίος είναι:

$$P(A) = \lim_{n\to x} \frac{\nu_A}{\nu} = \lim_{n\to x} f_A$$

Οι ποσότητες ν_A που εμπλέκονται στον προηγούμενο ορισμό εξαρτώνται από τον ορισμό επαναλήψεως του πειράματος. Η έννοια του ορίου που εμφανίζεται στον ορισμό δεν έχει αυστηρή μαθηματική αξία αλλά αποδίδει την σταθεροποίηση της σχετικής συχνότητας σε περίπτωση αύξησης του αριθμού ν των επαναλήψεων του πειράματος.

Παράδειγμα

Από τους ελέγχους που έγιναν σε 8.000 οχήματα, διαπιστώθηκε ότι τα 800 εξέπεμπαν καυσαέρια πάνω από το νόμιμο όριο τα 400 είχαν φθαρμένα ελαστικά ενώ σε 200 διαπιστώθηκαν και οι δύο παραβάσεις. Θεωρώντας ότι οι ν=8.000 επαναλήψεις του πειράματος (του ελέγχου οχημάτων) είναι αρκετές ώστε να επιτευχθεί η σταθεροποίηση των σχετικών συχνοτήτων, να υπολογιστεί η πιθανότητα σε ένα όχημα που εκλέγεται στην τύχη.

α) να διαπιστωθεί η εκπομπή καυσαερίων πάνω από το νόμιμο όριο

β) να βρεθούν φθαρμένα ελαστικά

γ) να διαπιστωθούν και οι 2 παραβάσεις

Λύση

Ορίζουμε τα ενδεχόμενα:

A: το όχημα εκπέμπει καυσαέρια πάνω από το νόμιμο όριο

B: το όχημα έχει φθαρμένα ελαστικά

Για τις συχνότητες και σχετικές συχνότητες των ενδεχομένων A,B,AB έχουμε:

$$\nu_A = 800 \quad , \quad \nu_B = 400 \quad , \nu_{AB} = 200$$

Επομένως:

- $f_A = \frac{800}{8.000} = 0{,}1 = 10\%$
- $f_B = \frac{400}{8.000} = 0{,}05 = 5\%$
- $f_{AB} = \frac{200}{8.000} = 0.025 = 2.5\%$

Αφού υποθέσαμε ότι στις ν=8.000 επαναλήψεις έχουμε σταθεροποίηση των σχετικών συχνοτήτων οι τιμές που υπολογίσαμε θα είναι οι αντίστοιχες οριακές σχετικές συχνότητες. Άρα, έχουμε:

P(A)=0,1=10%, P(B)=0,05=5% και P(AB)=0,025=2,5%.

2.4.3. ΑΞΙΩΜΑΤΙΚΟΣ ΟΡΙΣΜΟΣ

Η αξιωματική θεμελίωση της έννοιας της πιθανότητας οφείλεται στο Ρώσο στατιστικό Kolmogorov.

Με βάση την αξιωματική θεμελίωση, η πιθανότητα ορίζεται ως μια πραγματική συνάρτηση P, με πεδίο ορισμού το σύνολο όλων των δυνατών υποσυνόλων του Ω, δηλαδή την οικογένεια των ενδεχομένων α και πεδίο τιμών ένα υποσύνολο του συνόλου των πραγματικών αριθμών $\mathbb{R}$.

Για να εκφράζει πιθανότητα η συνάρτηση P θα πρέπει να πληροί τις εξής ιδιότητες:

1) Σε κάθε ενδεχόμενο $A \in \alpha$ ή $A \in P(\Omega)$ αντιστοιχεί ο πραγματικός αριθμός P{A} που ονομάζεται πιθανότητα του ενδεχομένου A. Η πιθανότητα αυτή είναι πάντοτε $P\{A\} \geq 0$.

2) Η πιθανότητα του δειγματικού χώρου Ω είναι ίση με τη μονάδα, P(Ω)=1 δηλαδή η πιθανότητα του ενδεχομένου A περιέχεται μεταξύ του 0 και του 1, $0 \leq P\{A\} \leq 1$.

3) Αν τα ενδεχόμενα A και B είναι ασυμβίβαστα μεταξύ τους, η πιθανότητα της ένωσης αυτών είναι ίση με το άθροισμα των επιμέρους πιθανοτήτων τους.

Δηλαδή αν $A \cap B = \emptyset$ τότε $\forall$ A , B $\in$ P(Ω) θα έχουμε:

P{A+B}= P{A}+P{B} (Προσθετική ιδιότητα)

Μπορούμε επομένως να δώσουμε συνοπτικά τον παρακάτω ορισμό της αξιωματικής πιθανότητας.

Έστω Ω ένας δειγματικός χώρος για ένα πείραμα τύχης. Ας θεωρήσουμε επίσης ότι σε κάθε ενδεχόμενο Α του Ω αντιστοιχείται ένας πραγματικός αριθμός P(A). Αν η P ικανοποιεί τα επόμενα 3 αξιώματα, θα ονομάζεται πιθανότητα στο δειγματικό χώρο Ω. Ενώ ο αριθμός P(A) θα λέγεται πιθανότητα του ενδεχομένου Α.

1) P(A) $\geq$ 0 ,για κάθε ενδεχόμενο Α του Ω
2) P(Ω) = 1
3) Αν $A_1,A_2,A_3,\ldots$ είναι μια ακολουθία ξένων ανά 2 ενδεχομένων του Ω (δηλαδή $A_i\, A_j = \emptyset$ για $i \neq j$) τότε :

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Παράδειγμα

Για ένα ενδεχόμενο Α του δειγματικού χώρου Ω είναι γνωστό ότι η πιθανότητα να εμφανιστεί το Α είναι κατά 0,7 μεγαλύτερη της πιθανότητας να μην εμφανιστεί. Να υπολογιστεί η πιθανότητα εμφάνισης του Α.

Λύση

Με βάση τα δεδομένα της άσκησης ισχύει ότι:

$$P(A') = P(A) + 0{,}7$$

Σύμφωνα με τον νόμο του συμπληρωματικού ενδεχομένου έχουμε:

$$P(A') = 1 - P(A)$$

Οπότε αντικαθιστούμε το $P(A')$ και έχουμε:

$$1 - P(A) = P(A) + 0{,}7$$

Επομένως P(A) = 0,15

2.5. ΙΔΙΟΤΗΤΕΣ ΠΙΘΑΝΟΤΗΤΩΝ

1. Αν A,B είναι δύο οποιαδήποτε ενδεχόμενα του δειγματικού χώρου Ω τότε:

P(A-B) = P(AB') = P(A) – P(AB)

Στην ειδική περίπτωση που ισχύει B $\subseteq$ A έχουμε:

P(A-B) = P(A) – P(B).

2. Αν A,B είναι δύο οποιαδήποτε ενδεχόμενα του δειγματικού χώρου Ω τότε η πιθανότητα της ένωσης A $\cup$ B δίνεται από τον τύπο:

$$P(A \cup B) = P(A) + P(B) - P(AB).$$

3. $P(A^C) = 1-P\{A\}$, για κάθε $A \in P(\Omega)$.

4. Αν $B \subseteq A$ τότε $P\{B\} \leq P(A\}$ για κάθε $A , B \in P(\Omega)$.

5. $P\{\emptyset\} = 0$.

6. $0 \leq P\{A\} \leq 1$

7. $P\{A \cup B\} \leq P\{A\} + P\{B\}$

Παράδειγμα

Το ποσοστό των υποψηφίων για την εισαγωγή τους στα ΑΕΙ που παίρνουν βαθμολογία κάτω από τη βάση στο μάθημα της Έκθεσης και των Μαθηματικών είναι 20% και 30% αντίστοιχα, ενώ 10% των υποψηφίων παίρνει βαθμολογία κάτω από τη βάση και στα δύο μαθήματα. Να βρείτε το ποσοστό των υποψηφίων που παίρνει βαθμολογία κάτω από τη βάση:

α) μόνο στην Έκθεση,

β) μόνο στα Μαθηματικά,

γ) σε ένα τουλάχιστον από τα δύο μαθήματα,

δ) σε ένα ακριβώς από τα δύο μαθήματα.

Λύση

Ορίζουμε τα ενδεχόμενα:

A: Η βαθμολογία του υποψηφίου στο μάθημα της Έκθεσης είναι κάτω από τη βάση.

B: Η βαθμολογία του υποψηφίου στο μάθημα των Μαθηματικών είναι κάτω από τη βάση.

Σύμφωνα με τα δεδομένα της άσκησης έχουμε:

$P(A) = 0,20$

$P(B) = 0,30$

$P(AB) = 0,10$

α. $P(AB') = P(A) - P(AB) = 0,20 - 0,10 = 0,10 = 10\%$

β. $P(A'B) = P(B) - P(AB) = 0,30 - 0,10 = 0,20 = 20\%$

γ. $P(A \cup B) = P(A) + P(B) - P(AB) = 0,20 + 0,30 - 0,10 = 0,40 = 40\%$

δ. $P(AB' \cup A'B) = P(AB') + P(A'B) = 0,10 + 0,20 = 0,30 = 30\%$

2.6. ΘΕΩΡΗΜΑ ΤΗΣ ΠΡΟΣΘΕΣΗΣ ΤΩΝ ΠΙΘΑΝΟΤΗΤΩΝ

2.6.1. ΕΝΔΕΧΟΜΕΝΑ ΑΣΥΜΒΙΒΑΣΤΑ

Αν δύο ενδεχόμενα Α και Β τα οποία έχουν πιθανότητες πραγματοποίησης Ρ(Α) και Ρ(Β) αποκλείονται αμοιβαίως (η εμφάνιση του Α αποκλείει την εμφάνιση του Β), τότε η πιθανότητα εμφάνισης του Α ή του Β είναι ίση με το άθροισμα των επιμέρους πιθανοτήτων τους. Δηλαδή έχουμε:

$$P(A \text{ ή } B) = P(A \cup B) = P(A) + P(B)$$

Γενικά, αν Α,Β,Γ,… είναι ανά δύο ασυμβίβαστα ενδεχόμενα, τότε ισχύει η σχέση:

$$P(A \text{ ή } B \text{ ή } \Gamma \ldots) = P(A \cup B \cup \Gamma \ldots) = P(A) + P(B) + P(\Gamma) + \ldots$$

Παράδειγμα

Ρίχνουμε στον αέρα ένα νόμισμα. Ποια είναι η πιθανότητα να εμφανιστεί το ενδεχόμενο «Κεφαλή» ή το ενδεχόμενο «Γράμματα»;

Λύση

Τα ενδεχόμενα «Κεφαλή» και «Γράμματα» είναι ασυμβίβαστα, άρα η πιθανότητα να εμφανιστεί το ενδεχόμενο «Κ» ή το ενδεχόμενο «Γ» θα είναι:

$$P(\text{Κ ή } \Gamma) = P(\text{Κ}) + P(\Gamma) = {}^{1}/_{2} + {}^{1}/_{2} = 1$$

2.6.2. ΕΝΔΕΧΟΜΕΝΑ ΜΗ ΑΣΥΜΒΙΒΑΣΤΑ

Αν τώρα τα ενδεχόμενα Α και Β δεν αποκλείονται αμοιβαίως (η εμφάνιση του Α δεν αποκλείει και την ταυτόχρονη εμφάνιση και του Β), τότε η πιθανότητα εμφάνισης του Α και του Β είναι:

$$P(A \text{ ή } B) = P(A \cup B) = P(A) + P(B) - P(A \text{ και } B)$$

Παράδειγμα

Από μια τράπουλα με 52 χαρτιά, τραβάμε ένα χαρτί. Ποιά είναι η πιθανότητα το χαρτί να είναι:

i. Μπαστούνι ή Ρήγας

ii. Άσος ή Καρό

Λύση

Χαρακτηρίζουμε τα ενδεχόμενα με τα γράμματα:

Α = Άσος

Κ = Καρό

Μ = Μπαστούνι

Ρ = Ρήγας

Οι πιθανότητες για το καθένα είναι:

$P(A) = {}^{4}/_{52}$ $P(K) = {}^{13}/_{52}$ $P(M) = {}^{13}/_{52}$ $P(P) = {}^{4}/_{52}$

i. Τα ενδεχόμενα «Μπαστούνι» = Μ και «Ρήγας» = Ρ είναι μη-ασυμβίβαστα. Επομένως:

$P(M \text{ ή } P) = P(M) + P(P) - P(M \text{ και } P) = {}^{13}/_{52} + {}^{4}/_{52} - {}^{1}/_{52} = {}^{16}/_{52} = 0,308$ ή 30,8%

ii. Τα ενδεχόμενα «Άσος» = Α και «Καρό» = Κ είναι μη-ασυμβίβαστα. Άρα:

$P(A \text{ ή } K) = P(A) + P(K) - P(A \text{ και } K) = {}^{4}/_{52} + {}^{13}/_{52} - {}^{1}/_{52} = {}^{16}/_{52} = 0,308$ ή 30,8%

2.7. ΘΕΩΡΗΜΑ ΤΟΥ ΠΟΛΛΑΠΛΑΣΙΑΣΜΟΥ ΤΩΝ ΠΙΘΑΝΟΤΗΤΩΝ

2.7.1. ΕΞΑΡΤΗΜΕΝΑ ΕΝΔΕΧΟΜΕΝΑ

Δύο ενδεχόμενα Α και Β ονομάζονται εξαρτημένα (δεσμευμένα) όταν η εμφάνιση του Β εξαρτάται από την εμφάνιση του Α.

Αν δύο ενδεχόμενα Α και Β είναι εξαρτημένα, τότε η πιθανότητα να πραγματοποιηθεί και το Α και το Β είναι ίση με το γινόμενο της πιθανότητας Ρ(Β) επί τη δεσμευμένη πιθανότητα Ρ(Α/Β). Δηλαδή έχουμε:

$P(A \text{ και } B) = P(A \cap B) = P(B) \cdot P(A/B)$ ή

$P(A \text{ και } B) = P(A \cap B) = P(A) \cdot P(B/A)$

Οι παραπάνω σχέσεις εκφράζουν το γενικό νόμο του πολλαπλασιασμού των πιθανοτήτων.

2.7.2. ΑΝΕΞΑΡΤΗΤΑ ΕΝΔΕΧΟΜΕΝΑ

Δύο ενδεχόμενα Α και Β ονομάζονται ανεξάρτητα όταν η εμφάνιση του Α δεν αποκλείει και την εμφάνιση του Β και αντίστροφα. Όταν δύο ενδεχόμενα είναι ανεξάρτητα, τότε η πιθανότητα της ταυτόχρονης εμφάνισης του Α και του Β ισούται με το γινόμενο των ατομικών πιθανοτήτων τους.

$P(A \text{ και } B) = P(A \cap B) = P(A) \cdot P(B)$

Όταν τα Α και Β είναι ανεξάρτητα, τότε η σχέση αυτή ονομάζεται θεώρημα των σύνθετων πιθανοτήτων ή κανόνας του πολλαπλασιασμού των πιθανοτήτων.

Παράδειγμα

Από μια τράπουλα με 52 φύλλα επιλέγεται στην τύχη ένα φύλλο. Βάζουμε πίσω το φύλλο στην τράπουλα και επιλέγουμε ένα δεύτερο.

1. Ποια είναι η πιθανότητα όπως το πρώτο φύλλο είναι «Ντάμα» και το δεύτερο «Σπαθί»;
2. Ποια είναι η πιθανότητα όπως το πρώτο φύλλο είναι «Ντάμα» και το δεύτερο φύλλο είναι «Σπαθί» όταν είναι γνωστό ότι το πρώτο φύλλο που τραβήξαμε δεν το βάζουμε πίσω στην τράπουλα;

Λύση

1. Έστω Α = Ντάμα $P(A) = {}^{4}/_{52}$

   Έστω Β = Σπαθί $P(B) = {}^{13}/_{52}$

   Η πιθανότητα όπως το πρώτο χαρτί είναι «Ντάμα» και το δεύτερο είναι «Σπαθί» θα είναι:
   $P(A \text{ και } B) = P(A) \cdot P(B) = {}^{4}/_{52} \cdot {}^{13}/_{52} = 0,019$
2. Αφού δεν βάλαμε πίσω το φύλλο στην τράπουλα, η πιθανότητα να βγει «Σπαθί» (= Β) όταν είναι γνωστό ότι έχει βγει «Ντάμα» (= Α) είναι $P(B/A) = {}^{12}/_{51}$. Άρα έχουμε:

   $P(A \text{ και } B) = P(A) \cdot P(B/A) = {}^{4}/_{52} \cdot {}^{12}/_{51} = 0,018$

2.8. ΔΕΣΜΕΥΜΕΝΗ ΠΙΘΑΝΟΤΗΤΑ

Ας θεωρήσουμε ένα δοχείο το οποίο περιέχει οκτώ κόκκινες και δύο πράσινες σφαίρες. Εξάγουμε τυχαία μία από τις 10 σφαίρες και στη συνέχεια, χωρίς να επιστρέψουμε στο δοχείο τη σφαίρα που βγάλαμε, εξάγουμε μια δεύτερη.

Ας ορίσουμε τα ενδεχόμενα:

A_i: στην i εξαγωγή διαλέγουμε κόκκινη σφαίρα

B_i: στην i εξαγωγή διαλέγουμε πράσινη σφαίρα

Για i=1,2. Αφού η εξαγωγή της πρώτης σφαίρας γίνεται εντελώς τυχαία είναι φανερό ότι $P(A_i) = 8/10$ και $P(B_i) = 2/10$.

Την στιγμή της δεύτερης εξαγωγής ενώ γνωρίζουμε το πλήθος των δυνατών αποτελεσμάτων (9, όσες και οι σφαίρες που απέμειναν στο δοχείο), δεν μπορούμε να προσδιορίσουμε ακριβώς το πλήθος των ευνοϊκών αποτελεσμάτων, αφού αυτό εξαρτάται από το αποτέλεσμα της πρώτης εξαγωγής.

Μπορούμε ωστόσο να διατυπώσουμε τους ακόλουθους ισχυρισμούς:

α) Αν στην πρώτη εξαγωγή επιλέχτηκε κόκκινη σφαίρα, η πιθανότητα να επιλεγεί στη δεύτερη εξαγωγή κόκκινη είναι 7/9, ενώ η πιθανότητα να επιλεγεί κίτρινη είναι 2/9.

β) Αν στην πρώτη εξαγωγή επιλέχτηκε πράσινη σφαίρα, η πιθανότητα να επιλεγεί στη δεύτερη εξαγωγή κόκκινη είναι 8/9, ενώ η πιθανότητα να επιλεγεί κίτρινη είναι 1/9.

Οι παραπάνω πιθανότητες λέγονται δεσμευμένες πιθανότητες, αφού αναφέρονται στην εμφάνιση ή μη ενός ενδεχομένου υπό την προϋπόθεση ότι συνέβη κάποιο άλλο ενδεχόμενο. Πιο συγκεκριμένα θα γράφουμε:

$$P(A_2|A_1) = \frac{7}{9} \qquad P(B_2|A_1) = \frac{2}{9}$$

$$P(A_2|B_1) = \frac{8}{9} \qquad P(B_2|B_1) = \frac{1}{9}$$

Οι αναφερθείσες δεσμευμένες πιθανότητες μπορούν να χρησιμοποιηθούν για τον υπολογισμό των μη δεσμευμένων πιθανοτήτων $P(A_2), P(B_2)$.

Ορισμός: Έστω Ω ένας δειγματικός χώρος και Β ⊆ Ω ένα ενδεχόμενο του Ω με P(B) $\geq$ 0. Τότε, για κάθε ενδεχόμενο Α του Ω, η δεσμευμένη πιθανότητα του Α δοθέντος του Β δίνεται από τον τύπο:

$$P(A|B) = \frac{P(AB)}{P(B)}$$

Παράδειγμα

Σε μια χώρα, η πιθανότητα να ζήσει ένας άντρας τουλάχιστον 70 χρόνια είναι 0,85, ενώ η πιθανότητα να ζήσει τουλάχιστον 75 χρόνια είναι 0,80. Αν διαλέξουμε τυχαία έναν 70χρονο άντρα από τη χώρα αυτή, ποια είναι η πιθανότητα να ζήσει τουλάχιστον άλλα 5 χρόνια;

Λύση

Έστω Α,Β τα ενδεχόμενα ένας άντρας ο οποίος επιλέγεται τυχαία από τον πληθυσμό, να ζήσει περισσότερο από 75,70 χρόνια αντίστοιχα. Τότε θα έχουμε:

P(A)=0,80 P(B)=0,85

Και AB= A (γιατί A ⊆ B).

Η πιθανότητα που ζητάμε να υπολογίσουμε, είναι η δεσμευμένη πιθανότητα $P(A|B)$ και προκύπτει ως εξής:

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(A)}{P(B)} = \frac{0.80}{0.85} = 0.94$$

ΚΕΦΑΛΑΙΟ ΤΡΙΤΟ: ΕΙΔΗ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ

3.1. ΑΠΛΗ ΤΥΧΑΙΑ ΔΕΙΓΜΑΤΟΛΗΨΙΑ - ΕΙΣΑΓΩΓΗ

Υπάρχουν πολλές μέθοδοι με τις οποίες μπορεί να επιλεγεί ένα τυχαίο δείγμα. Η επιλογή της μεθόδου εξαρτάται από διάφορους παράγοντες, όπως οι σκοποί και οι προδιαγραφές της έρευνας, το διαθέσιμο πλαίσιο δειγματοληψίας, η γεωγραφική διασπορά του πληθυσμού, οι επιχειρησιακοί περιορισμοί της έρευνας και ο τρόπος ανάλυσης των στοιχείων της έρευνας από τους χρήστες.

Στην επιλογή μεθόδου ή αλλιώς σχεδιασμού τυχαίας δειγματοληψίας ο σκοπός πρέπει να είναι η ελαχιστοποίηση του δειγματοληπτικού σφάλματος των εκτιμητριών για τις πιο σημαντικές μεταβλητές της έρευνας, ελαχιστοποιώντας ταυτόχρονα τον χρόνο και το κόστος διεξαγωγής της έρευνας.

3.1.1. ΤΥΧΑΙΑ ΔΕΙΓΜΑΤΟΛΗΨΙΑ

Η τυχαία δειγματοληψία προϋποθέτει ότι η επιλογή των μονάδων του δείγματος διενεργείται με τέτοιο τρόπο ώστε όλες οι μονάδες του πληθυσμού να έχουν γνωστή πιθανότητα επιλογής μεγαλύτερη από το μηδέν. Η πιθανότητα αυτή πρέπει να εκφράζεται με ορισμένο αριθμό. Για την επιλογή ενός απλού τυχαίου δείγματος είναι αναγκαίο ένα πλαίσιο – κατάλογος όλων των μελών του πληθυσμού της δειγματοληψίας.

Η πιο βασική τεχνική δειγματοληψίας είναι η απλή τυχαία δειγματοληψία. Ένα από τα πλεονεκτήματα της είναι ότι δεν απαιτεί πρόσθετες πληροφορίες στο πλαίσιο δειγματοληψίας και χρησιμοποιεί τον πλήρη κατάλογο των μελών του πληθυσμού δειγματοληψίας μαζί με την πληροφορία εντοπισμού τους. Επιπρόσθετα, υπάρχουν εύχρηστοι τύποι για τον καθορισμό του δειγματικού μεγέθους και για εκτιμήσεις παραμέτρων και των διακυμάνσεων τους. Ωστόσο, η συγκεκριμένη τεχνική δεν χρησιμοποιεί τις εκάστοτε βοηθητικές πληροφορίες που περιέχει το πλαίσιο οι οποίες θα μπορούσαν να καταστήσουν το σχεδιασμό δειγματοληψίας πιο αποτελεσματικό. Επιπλέον αν και υπάρχει μεγάλη ευκολία εφαρμογής σε μικρούς πληθυσμούς ,η απλή τυχαία δειγματοληψία μπορεί να είναι δαπανηρή και ανέφικτη για μεγάλους πληθυσμούς επειδή όλα τα μέλη πρέπει να αναγνωριστούν και να αριθμηθούν πριν από την δειγματοληψία.

Απλή τυχαία δειγματοληψία έχουμε όταν η επιλογή n μονάδων από ένα πληθυσμό N μονάδων, έτσι ώστε κάθε δυνατό δείγμα μεγέθους n να έχει την ίδια πιθανότητα να επιλεγεί, προκειμένου όλες οι μονάδες του πληθυσμού να έχουν την ίδια ευκαιρία να συμπεριληφθούν στο δείγμα .

Αυτός ο συνδυασμός αντιστοιχεί στον εξής τύπο:

$$C_n^N = \frac{N!}{n!(N-n)!}$$

Παράδειγμα

Έστω ότι επιλέγουμε τυχαία 2 μονάδες από ένα πληθυσμό που αποτελείται από 4 μονάδες:

x_1, x_2, x_3, x_4

Τα δυνατά δείγματα είναι:

x_1x_2	x_2x_3	x_3x_4
x_1x_3	x_2x_4	
x_1x_4		

$$C_n^N = \frac{N!}{n!(N-n)!} = \frac{4!}{2!(4-2)!} = \frac{24}{4} = 6$$

Εφόσον χρησιμοποιούμε τη μέθοδο της απλής τυχαίας δειγματοληψίας πρέπει η επιλογή των ανωτέρω δειγμάτων να είναι τυχαία, δηλαδή κάθε ένας από τους συνδυασμούς x_1x_2, x_1x_3, ….., x_3x_4 να έχει πιθανότητα επιλογής :

$$\frac{n}{N} * \frac{(n-1)}{(N-1)} * \frac{(n-2)}{(N-2)} \cdots\cdots \frac{1}{(N-n+1)} = \frac{n!(N-n)!}{N!} = \frac{2}{30} = \frac{1}{15}$$

Λόγω μεγάλου μεγέθους του ερευνώμενου πληθυσμού είναι δύσκολο να καταρτίσουμε όλα τα δυνατά δείγματα, άλλωστε δεν έχει έννοια να βρεθούν τα δυνατά δείγματα αλλά η μέθοδος επιλογής του δείγματος να είναι τέτοια ώστε η πιθανότητα επιλογής σε κάθε σειρά μονάδων να είναι η ίδια.

3.1.2. ΕΠΙΛΟΓΗ ΑΠΛΟΥ ΤΥΧΑΙΟΥ ΔΕΙΓΜΑΤΟΣ

Για να επιλέξουμε ένα δείγμα n μονάδων από έναν πληθυσμό N μονάδων, πρέπει η επιλογή του δείγματος να γίνει τυχαία .Η τυχαία αυτή επιλογή εξασφαλίζεται με τις παρακάτω μεθόδους :

α)μέθοδος της κλήρωσης : Στη συγκεκριμένη μέθοδο η επιλογή των στατιστικών μονάδων γίνεται με κλήρωση. Για παράδειγμα τα ονόματα των ατόμων που αποτελούν τον ερευνώμενο πληθυσμό γράφονται σε κάρτες , τοποθετούνται σε ένα δοχείο και στη συνέχεια

εξάγεται το επιθυμητό μέγεθος δείγματος. Ανάλογα με το αν οι κάρτες αυτές επανατοποθετηθούν έχουμε δειγματοληψία χωρίς επανατοποθέτηση (sampling without replacement) η δειγματοληψία με επανατοποθέτηση(sampling with replacement).

Παράδειγμα

Ας υποθέσουμε ότι έχουμε ένα καθορισμένο πληθυσμό μεγέθους 6,παραδείγματος χάρη {1,2,3,4,5,6}.Τα δυνατά δείγματα μεγέθους 2 είναι 15:
{1,2},{1,3},{1,4},{1,5},{1,6},{2,3},{2,4},{2,5},{2,6},{3,4},{3,5},{3,6},{4,5},{4,6},{5,6}

Γενικά αν ο πληθυσμός αποτελείται από N μονάδες και ζητάμε δείγμα μεγέθους n το δείγμα του δυνατού μεγέθους είναι:

$$(N/n) = \frac{N!}{n!(N-n)!} = \frac{N(N-1)\dots(N-n+1)}{n},$$ όπου n!= 1x2x3x……xn

β)<u>μέθοδος τυχαίων αριθμών:</u> Με τη χρήση πινάκων τυχαίων αριθμών πραγματοποιείται η επιλογή των στατιστικών μονάδων του ερευνώμενου πληθυσμού. Επειδή ο πληθυσμός N είναι αρκετά μεγάλος αριθμός η εφαρμογή της μεθόδου της κλήρωσης δεν προτιμάται. Επιπλέον δεν εξασφαλίζεται η τυχαία επιλογή αφού η ανάμειξη των καρτών δεν γίνεται εύκολα.

Οι πίνακες τυχαίων αριθμών αποτελούνται από σειρές και στήλες των αριθμών 0,1,2,…,9. Ελέγχονται ως προς τη συχνότητα των συγκεκριμένων αριθμών με σκοπό την εξασφάλιση της επιλογής των τυχαίων αριθμών.

Οι πίνακες στοιχείων αριθμών χρησιμοποιούνται ως εξής:

Έστω ότι επιθυμούμε να επιλέξουμε ένα δείγμα n από έναν πληθυσμό με N μονάδες. Αρχικά, αριθμούμε τον πληθυσμό από 1 μέχρι N στη συνέχεια θα επιλέξουμε n οι οποίοι θα είναι μικρότεροι ή ίσοι από το N η επιλογή αρχίζει από κάποιον αριθμό στήλης ή σειράς και προχωρούμε είτε οριζόντια είτε κατακόρυφα μέχρι να επιλέξουμε n αριθμούς. Αριθμοί μεγαλύτεροι από το N ή αριθμοί που έχουν επιλεγεί μια φορά παραλείπονται.
Παράδειγμα

Έστω

58	62	23	87	116	50	14	3	112	40
18	59	1	127	50	103	15	28	96	72
67	114	89	20	15	66	40	67	54	24
6	56	53	110	5	189	34	53	2	54
17	58	18	55	129	94	57	51	5	193

Θέλουμε να επιλέξουμε ένα τυχαίο δείγμα 6% από ένα πληθυσμό με N=100 μονάδες. Αριθμούμε τις μονάδες πληθυσμού από 1 μέχρι 100 και επιλέγουμε 10 μονάδες χρησιμοποιώντας τον παραπάνω πίνακα, η επιλογή ξεκινά κατά τυχαίο τρόπο από κάποιο

αριθμό μιας στήλης ή σειράς. Έστω ότι επιλέγουμε την τέταρτη στήλη του πίνακα που έχει πρώτο αριθμό το 87. Επειδή N=100 θα πάρουμε τους αριθμούς οι οποίοι είναι μικρότεροι ή ίσοι από το 100. Αν προχωρήσουμε κατακόρυφα στην τέταρτη στήλη θα επιλεγούν οι αριθμοί. 87,20,55. Επειδή δεν υπάρχουν άλλοι αριθμοί μικρότεροι του 100 συνεχίζουμε την επιλογή κατά τον ίδιο τρόπο από την πέμπτη στήλη. Επομένως θα επιλεγούν οι αριθμοί 50, 15, 05.

Οι αριθμοί 127, 110, 116 και 129 οι οποίοι είναι μεγαλύτεροι του 100 παραλείπονται. Οι 6 αριθμοί που έχουν επιλεγεί με την παραπάνω διαδικασία αποτελούν τις μονάδες του δείγματος που θα ερευνηθούν από τον πληθυσμό N.

3.1.2.1. ΠΛΕΟΝΕΚΤΗΜΑΤΑ - ΜΕΙΟΝΕΚΤΗΜΑΤΑ

Πλεονεκτήματα:

- Επειδή το δείγμα είναι τυχαίο, μπορεί να γίνει πιο εύκολα ο υπολογισμός του δειγματοληπτικού σφάλματος.
- Από τη στιγμή που το δείγμα λαμβάνεται με τη μέθοδο της κλήρωσης, απομακρύνεται η πιθανότητα μεροληψίας, επομένως τα αποτελέσματα της δειγματοληψίας είναι αντικειμενικά.
- Είναι μια τεχνική εύκολη στην κατανόηση, δεδομένου του σχεδιασμού της καθιστώντας την προσιτή ακόμα και σε άτομα με χαμηλό επίπεδο μόρφωσης.

Μειονεκτήματα:

- Πρόκειται για μια αρκετά δαπανηρή μέθοδο αφού απαιτεί πλήρως ενημερωμένα πλαίσια. Με άλλα λόγια είναι απαραίτητη η κατασκευή καταλόγου και η αρίθμηση της κάθε μονάδας του συνολικού πληθυσμού.
- Η μέθοδος αυτή είναι χρονοβόρα, ειδικά όταν το σύνολο του υπό μελέτη πληθυσμού είναι πεπερασμένο, ενώ επίσης δημιουργείται μεγάλο κόστος συλλογής δεδομένων.

3.1.3. ΕΚΤΙΜΗΣΕΙΣ ΣΤΗΝ ΑΠΛΗ ΤΥΧΑΙΑ ΔΕΙΓΜΑΤΟΛΗΨΙΑ

3.1.3.1. ΕΚΤΙΜΗΣΗ ΜΕΣΟΥ ΠΛΗΘΥΣΜΟΥ

Από απλό τυχαίο δείγμα τιμών της μεταβλητής X, μεγέθους n, λαμβάνονται οι εξής τιμές:

$$x_1, x_2, x_3, \ldots\ldots, x_n$$

Με βάση τις τιμές αυτές υπολογίζεται ο μέσος όρος του δείγματος:

$$\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Η διακύμανση της μέσης τιμής του δείγματος $\bar{x}$ είναι :

$$V(\bar{x}) = E(\bar{x} - \bar{X})^2 = \frac{\sigma^2}{n}\frac{(N-n)}{N} \qquad \text{ή} \qquad V(\bar{x}) = \frac{S^2}{n}\frac{(N-n)}{N} = \frac{S^2}{n}(1\text{-}f).$$

Η διακύμανση της x_i (δηλαδή διακύμανση ενός πεπερασμένου πληθυσμού) είναι:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{X})^2}{N} \qquad \text{ή} \qquad S^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{X})^2}{N-1}, \text{ για Ν σχετικά μεγάλο.}$$

Η τυπική απόκλιση (standard deviation) είναι:

$$SE(\bar{x}) = \sqrt{V(\bar{x})} = \sqrt{\frac{\sigma^2}{n}\frac{(N-n)}{N}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{(N-n)}{N}}.$$

Η τυπική απόκλιση εκφράζει τη διασπορά του πλήθους όλων των δυνατών εκτιμήσεων $\bar{x}$ από το μέσο του πληθυσμού $\bar{X}$. Όσο μικρότερη είναι η τυπική απόκλιση τόσο μικρότερο θα είναι και το σφάλμα $\bar{x} - \bar{X}$ στην εκτίμηση του μέσου.

Η διακύμανση της εκτίμησης του συνόλου του πληθυσμού ($\hat{X} = N\bar{x}$) είναι:

$$V(\hat{X}) = E\,(\hat{X} - X)^2 = \frac{N^2\sigma^2}{n}\frac{(N-n)}{N}.$$

Το τυπικό σφάλμα (standard error) είναι:

$$SE(\hat{X}) = \sqrt{V(\hat{X})} = \sqrt{V(N\bar{x})} = \sqrt{\frac{N^2\sigma^2}{n}\frac{(N-n)}{N}} = \frac{N\sigma}{\sqrt{n}}\sqrt{\frac{(N-n)}{N}}.$$

Το σχετικό τυπικό σφάλμα είναι:

$$CV(\bar{x}) = \frac{SE(\bar{x})}{\bar{x}}\,100\%.$$

Για την εκτίμηση του συνόλου του πληθυσμού ($\hat{X} = N\bar{x}$) έχουμε:

$$CV(\hat{X}) = \frac{SE(\hat{X})}{\hat{X}}\,100\%.$$

Ο συντελεστής $\frac{(N-n)}{N}$ ονομάζεται διόρθωση πεπερασμένου πληθυσμού. Ο συντελεστής αυτός μπορεί να παραληφθεί όταν το κλάσμα δειγματοληψίας $\frac{n}{N}$ (f) είναι πολύ μικρό, άρα ο συντελεστής $\frac{(N-n)}{N} = 1 - \frac{n}{N}$ πλησιάζει τη μονάδα. Αυτό συμβαίνει όταν ο πληθυσμός είναι αρκετά μεγάλος ή όταν πρόκειται για άπειρους πληθυσμούς (δηλαδή σε περίπτωση δειγματοληψίας με επανατοποθέτηση).

Πρακτικά η διόρθωση πεπερασμένου πληθυσμού παραλείπεται όταν το κλάσμα δειγματοληψίας δεν ξεπερνά το 5%.

Παράδειγμα

Με τη μέθοδο της απλής τυχαίας δειγματοληψίας έγινε η επιλογή 40 νοικοκυριών από 800 νοικοκυριά που διαμένουν συνολικά σε μια πόλη από τη δειγματοληπτική αυτή έρευνα προέκυψε ο αριθμός μελών κατά νοικοκυριό ως εξής:

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

Να εκτιμηθούν α) το μέσο μέγεθος των νοικοκυριών, τη διακύμανση, το τυπικό σφάλμα και το σχετικό τυπικό σφάλμα β) ο συνολικός πληθυσμός της πόλης, το τυπικό σφάλμα και το σχετικό τυπικό σφάλμα του συνολικού πληθυσμού.

α) η εκτίμηση του μέσου μεγέθους των νοικοκυριών (μέση τιμή πληθυσμού) είναι: $\bar{X} = \hat{\bar{X}} = \bar{x}$ άρα έχουμε

$$\bar{x} = \frac{\sum x_i}{n} = \frac{4+3+2+\cdots+7+6}{40} = \frac{160}{40} = 4$$

Το τυπικό σφάλμα στην απλή τυχαία δειγματοληψία είναι:

$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{(N-n)}{N}}$$

Επειδή η διακύμανση του πληθυσμού σ^2 είναι άγνωστη χρησιμοποιούμε την αμερόληπτη εκτίμηση s^2 της S^2 η οποία είναι:

$$S^2 = \frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{n-1} = \frac{(4-4)^2+(3-4)^2+\cdots+(7-4)^2\,(6-4)^2}{39} = \frac{108}{39} = 2{,}77$$

Συνεπώς, το τυπικό σφάλμα είναι:

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}\sqrt{\frac{(N-n)}{N}} = \frac{1{,}66}{6{,}32}\sqrt{0{,}95} = 0{,}256$$

Το σχετικό τυπικό σφάλμα είναι:

$$CV(\bar{x}) = \frac{SE(\bar{x})}{\bar{x}} 100\% = \frac{0,256}{4} 100\% = 6,4\%$$

β) Η εκτίμηση του συνολικού πληθυσμού της πόλης είναι:

$$\hat{X} = N\bar{x} = 800 \cdot 4 = 3.200$$

Το τυπικό σφάλμα του συνολικού πληθυσμού είναι:

$$SE(N\bar{x}) = \frac{Ns}{\sqrt{n}} \sqrt{\frac{(N-n)}{N}} = \frac{800 \cdot 1,66}{6,32} \sqrt{0,95} = 204,8$$

Το σχετικό τυπικό σφάλμα είναι:

$$CV(\hat{X}) = \frac{SE(N\bar{x})}{\hat{X}} 100\% = \frac{204,8}{3200} 100\% = 6,4\%$$

3.1.3.1. ΕΚΤΙΜΗΣΗ ΠΟΣΟΣΤΟΥ Η ΑΝΑΛΟΓΙΑΣ ΠΛΗΘΥΣΜΟΥ

Ορισμένες φορές μας ενδιαφέρει να εκτιμήσουμε το ποσοστό ή την αναλογία του πληθυσμού των μονάδων που ανήκουν σε μία συγκεκριμένη τάξη, όπως για παράδειγμα είναι η εκτίμηση του ποσοστού των ανέργων σε μια πόλη.

Υποθέτουμε ότι κάθε μονάδα ενός πληθυσμού N ανήκει σε μία από τις 2 τάξεις C (τάξη στην οποία ανήκουν οι μονάδες που έχουν ένα συγκεκριμένο χαρακτηριστικό) και C΄ (τάξη στην οποία ανήκουν οι υπόλοιπες ομάδες). Αν A από τις N μονάδες του πληθυσμού ανήκουν στην τάξη C τότε το ποσοστό ή αναλογία των μονάδων της τάξης C θα είναι $P = \frac{A}{N}$ και το ποσοστό των μονάδων που ανήκουν στην τάξη C΄ θα είναι Q = 1 - P . Πρακτικά το ποσοστό P συνήθως είναι άγνωστο και περνούμε την δειγματική εκτίμηση αυτού p. Αν σε ένα δείγμα μεγέθους n βρεθούν α μονάδες του πληθυσμού οι οποίες ανήκουν στην τάξη C τότε θα έχουμε $p = \frac{\alpha}{n}$, το ποσοστό αυτό αποτελεί μια αμερόληπτη εκτίμηση του $P = \frac{A}{N}$ δηλαδή E(p) = P.

Αυτό αποδεικνύεται αν υποθέσουμε ότι η τιμή της μεταβλητής X παίρνει 2 τιμές 0 και 1 δηλαδή για κάθε μονάδα του πληθυσμού που ανήκει στη τάξη C έχουμε $x_i = 1$ ενώ για κάθε μονάδα που ανήκει στην τάξη C΄ έχουμε $x_i = 0$.

Επομένως, το σύνολο των τιμών x_i του πληθυσμού είναι:

$$X = \sum_{i=1}^{N} x_i = 1 + 0 + 1 + \ldots + 0 + 1 = A$$

Διότι ο αριθμός των τιμών $x_i = 1$ είναι ίσος με τον αριθμό των μονάδων της τάξης C. Άρα έχουμε:

$$\bar{X} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{A}{N} = P$$

Επιπλέον για το δείγμα έχουμε:

$$\sum_{i=1}^{N} x_i = 1 + 0 + 1 + \ldots + 0 + 1 = \alpha$$

και

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{a}{n} = p$$

Επειδή μπορούμε να αποδείξουμε ότι $E(\bar{x}) = \bar{X}$ τότε μπορεί να αποδειχθεί και ότι E(p) = P.

Στην περίπτωση που η μεταβλητή x_i παίρνει τιμές 1 ή 0 έχουμε:

$$\sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i = A = NP$$

και

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i = a = np$$

Επομένως, η διακύμανση του πληθυσμού σ^2 θα είναι:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{X})^2}{N} = \frac{\sum_{i=1}^{N} x_i^2 - 2\bar{X}\sum_{i=1}^{N} x_i + N\bar{X}^2}{N} = \frac{NP - 2p \cdot NP + NP^2}{N} = \frac{NP - NP^2}{N} =$$

$$P - P^2 = P(1 - P) = PQ$$

ή

$$S^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{X})^2}{N-1} = \frac{NP - NP^2}{N-1} = \frac{NP(1-P)}{N-1} = \frac{N}{N-1} PQ$$ για N σχετικά μεγάλο.

Όμοια για την αμερόληπτη εκτίμηση s^2 της S^2 έχουμε:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1} = \frac{np - np^2}{n-1} = \frac{np(1-p)}{n-1} = \frac{n}{n-1} pq$$

όπου q = 1- p

Η διακύμανση του ποσοστού p είναι:

$$X(p) = E(p - P)^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N}\right) = \frac{PQ}{n}\left(\frac{N-n}{N}\right)$$

ή

$$V(p) = \frac{S^2}{n}\left(\frac{N-n}{N}\right) = \frac{PQ\frac{N}{N-1}}{n} \cdot \frac{N-n}{N} = \frac{PQ}{n} \cdot \frac{N-n}{N-1}.$$

Η διακύμανση της εκτίμησης του συνολικού αριθμού ($\hat{A}$ = Np) των μονάδων που ανήκουν στην τάξη C είναι:

$$V(\hat{A}) = V(Np) = \frac{N^2\sigma^2}{n} \cdot \left(\frac{N-n}{N}\right) = \frac{N^2 PQ}{n} \cdot \left(\frac{N-n}{N}\right)$$

ή

$$V(\hat{A}) = V(Np) = \frac{N^2 s^2}{n} \cdot \left(\frac{N-n}{N}\right) = \frac{N^2 pq}{n} \cdot \left(\frac{N-n}{N-1}\right),$$ για Ν σχετικά μεγάλο.

Το τυπικό σφάλμα υπολογίζεται από τον τύπο:

SE(p) = $\sqrt{V(p)}$ ή SE(Np) = $\sqrt{V(Np)}$ (δίνεται από τις τετραγωνικές ρίζες των διακυμάνσεων)

Όμως επειδή πρακτικά η διακύμανση του πληθυσμού είναι άγνωστη χρησιμοποιούμε την αμερόληπτη εκτίμηση της διακύμανσης του p που προκύπτει από το δείγμα, δηλαδή

$$V(p) = \frac{s^2}{n} \cdot \left(\frac{N-n}{N}\right) = \frac{pq}{n-1} \cdot \left(\frac{N-n}{N}\right),$$ διότι $s^2 = \frac{n}{n-1} pq$.

Η εκτίμηση της διακύμανσης του $\hat{A}$ είναι:

$$V(\hat{A}) = V(Np) = \frac{N^2 s^2}{n} \cdot \frac{N-n}{N} = \frac{pq}{n-1} \cdot N(N-n).$$

Παράδειγμα

Ο πληθυσμός ενός οικισμού ανέρχεται σε 400 άτομα με τη μέθοδο της απλής τυχαίας δειγματοληψίας έγινε επιλογή 40 ατόμων. Στα 40 άτομα βρέθηκαν 16 άνδρες. Να εκτιμηθούν: α) το ποσοστό και ο συνολικός αριθμός των ανδρών στον οικισμό και β) το τυπικό σφάλμα.

α) Το ποσοστό των ανδρών στο δείγμα είναι:

$$p = \frac{\alpha}{n} = \frac{16}{40} = 0{,}4$$ ή 40%

Ο συνολικός αριθμός ανδρών στον οικισμό είναι:

$$\hat{A} = Np = 400 \cdot 0.4 = 160$$

β) Η διακύμανση της εκτίμησης του ποσοστού των ανδρών είναι:

$$V(p) = \frac{s^2}{n} \cdot \frac{N-n}{N} = \frac{pq}{n-1} \cdot \left(\frac{N-n}{N}\right) = \frac{0{,}4 \cdot 0{,}6}{39} \frac{400-40}{400} = 0.0054$$

Συνεπώς, το τυπικό σφάλμα είναι:

SE(p) = $\sqrt{0.0054}$ = 0.074

3.1.3.2. ΕΚΤΙΜΗΣΗ ΜΕΣΩΝ ΜΕΓΕΘΩΝ ΥΠΟΠΛΗΘΥΣΜΟΥ

Οι εκτιμήσεις των ερευνών πολλές φορές γίνονται κατά κατηγορία ενός πληθυσμού (υποπληθυσμού). Ακόμα, στα διαθέσιμα δειγματοληπτικά πλαίσια περιέχονται μονάδες οι οποίες δεν ανήκουν στις μονάδες του δείγματος και δεν είναι εύκολο να γίνει ο διαχωρισμός παρά μόνο μετά την επιλογή του δείγματος.

Έστω ότι ο αριθμός των μονάδων του πληθυσμού είναι N και N_j είναι ο αριθμός των μονάδων της κατηγορίας (υποπληθυσμού) που μας ενδιαφέρει. Λαμβάνοντας δείγμα μεγέθους n από τον πληθυσμό N και ο αριθμός των μονάδων του δείγματος που προκύπτει στον υποπληθυσμό N_j είναι n_j, τότε συμβολίζουμε με $x_{jκ}$ τις μετρήσεις αυτών των μονάδων (όπου κ = 1,2,…,n_j) και ο μέσος υποπληθυσμός $\bar{X}_j$για την κατηγορία j εκτιμάται από τη σχέση:

$$\bar{x}_j = \frac{\sum_{\kappa=1}^{n_j} x_{j\kappa}}{n_j}$$

Η διακύμανση της εκτίμησης $\bar{x}_j$ δίνεται από τη σχέση:

$$V(\bar{x}_j) = \left(\frac{N_j - n_j}{N_j}\right)\frac{\sum_{\kappa=1}^{n_j}(x_{j\kappa}-\bar{x}_j)^2}{n_j(n_j-1)}$$

Το τυπικό σφάλμα είναι:

$$SE(\bar{x}_j) = \sqrt{V(\bar{x}_j)}$$

Εάν η τιμή του N_j είναι άγνωστη μπορεί να χρησιμοποιηθεί το $\frac{n}{N}$ αντί του $\frac{n_j}{N_j}$ για τον υπολογισμό της διόρθωσης πεπερασμένου πληθυσμού. Η εκτίμηση του συνόλου X_j ενός υποπληθυσμού γίνεται ως εξής:

α) Εάν το N_j είναι γνωστό

$$\hat{X}_j = N_j\bar{x}_j = \frac{N_j}{n_j}\sum_{\kappa=1}^{n_j} x_{j\kappa}$$

Η διακύμανση της εκτίμησης $\hat{X}_j$ είναι :

$$V(\hat{X}_j) = V(N_j\bar{x}_j) = N_j^2 V(\bar{x}_j) = N_j^2\left(\frac{N_j - n_j}{N_j}\right)\frac{\sum_{\kappa=1}^{n_j}(x_{j\kappa}-\bar{x}_j)^2}{n_j(n_j-1)}$$

Το τυπικό σφάλμα είναι:

$$SE(\hat{X}_j) = \sqrt{V(\hat{X}_j)} = N_j\sqrt{V(\bar{x}_j)}$$

Ο συντελεστής μεταβλητότητας είναι:

$$CV(\bar{x}_j) = \frac{SE(\bar{x}_j)}{\bar{x}_j}100\%$$

β) Εάν το N_j είναι άγνωστο, έχουμε αμερόληπτη εκτίμηση και χρησιμοποιούμε τον παρακάτω τύπο:

$$\hat{X}_j = \frac{N}{n} \sum_{\kappa=1}^{n_j} x_{j\kappa}$$

Η διακύμανση της εκτίμησης $\hat{X}_j$ είναι:

$$V(\hat{X}_j) = N^2 \frac{N-n}{N} \frac{\sum_{\kappa=1}^{n_j} x_{j\kappa}^2 - \frac{1}{n}(\sum_{\kappa=1}^{n_j} x_{j\kappa})^2}{n(n-1)}$$

Το τυπικό σφάλμα είναι:

$$SE(\hat{X}_j) = \sqrt{V(\hat{X}_j)}$$

Παράδειγμα

Έστω ότι θέλουμε να εκτιμήσουμε τη μέση και τη συνολική μηνιαία δαπάνη νοσοκομειακής περίθαλψης των νοικοκυριών που έχουν άτομα ηλικίας 65 ετών και πάνω, μιας πόλης, όπου διαμένουν 1000 νοικοκυριά. Από διαθέσιμο κατάλογο των νοικοκυριών της πόλης παίρνουμε με τυχαίο τρόπο δείγμα n=100 νοικοκυριά και βρίσκουμε ότι τα νοικοκυριά που έχουν τουλάχιστον 1 άτομο ηλικίας πάνω των 65 ετών ανέρχονται σε n_j =60.Απο την έρευνα που έγινε στα 60 νοικοκυριά διαπιστώθηκε ότι οι δαπάνες κατά νοικοκυριό, για νοσοκομειακή περίθαλψη είναι:

10	9	11	15	11	14	10	13	12	12	14	12
15	12	13	14	12	13	12	14	11	13	15	14
12	14	15	12	9	15	14	11	10	15	13	13
13	15	12	14	14	9	15	9	14	14	14	10
11	10	9	10	13	10	11	10	15	12	12	9

Να εκτιμηθούν οι μέσοι και η συνολική μηνιαία δαπάνη για νοσοκομειακή περίθαλψη των νοικοκυριών της πόλης και να υπολογιστούν το τυπικό σφάλμα και ο συντελεστής μεταβλητότητας.

Η εκτίμηση της μέσης μηνιαίας δαπάνης για νοσοκομειακή περίθαλψη των νοικοκυριών της πόλης, που έχουν άτομα ηλικίας 65 ετών και πάνω είναι:

$$\bar{X}_j = \frac{\sum_{\kappa=1}^{n_j} x_{j\kappa}}{n_j} = \frac{10+9+11\ldots+112+9}{60} = \frac{739}{60} = 12{,}3 \text{ το μήνα.}$$

Εάν ο αριθμός των νοικοκυριών της πόλης που έχουν άτομα ηλικίας άνω των 65 ετών ανέρχεται σε $N_j = 650$ τότε η διακύμανση του $\bar{x}_j$ είναι:

$$V(\bar{x}_j) = \left(\frac{N_j - n_j}{N_j}\right)\frac{\sum_{k=1}^{n_j}(x_{jk} - \bar{x}_j)^2}{n_j(n_j - 1)} = \frac{650-60}{650}\cdot\frac{223}{60(60-1)} = 0.057$$

Το τυπικό σφάλμα είναι:

$$SE(\bar{x}_j) = \sqrt{V(\bar{x}_j)} = \sqrt{0,057} = 0,238$$

Ο συντελεστής μεταβλητότητας είναι:

$$CV(\bar{x}_j) = \frac{SE(\bar{x}_j)}{\bar{x}_j}100\% = \frac{0,238}{12,3}100\% = 1,93\%$$

Η εκτίμηση της συνολικής δαπάνης όλων των νοικοκυριών που έχουν άτομα ηλικίας άνω των 65 ετών της πόλης X_j είναι:

$$\hat{X}_j = N_j \cdot \bar{x}_j = 650 \cdot 12,3 = 7.995$$

Η διακύμανση της εκτίμησης $\hat{X}_j$ είναι :

$$V(\hat{X}_j) = N_j^2 V(\bar{x}_j) = 650^2 \cdot 0,057 = 24.082$$

Το τυπικό σφάλμα είναι:

$$SE(\hat{X}_j) = \sqrt{V(\hat{X}_j)} = \sqrt{24.082} = 155$$

Ο συντελεστής μεταβλητότητας είναι:

$$CV(\hat{X}_j) = \frac{SE(\hat{X}_j)}{\hat{X}_j}100\% = \frac{155}{7.995}100\% = 1,94\%$$

Εάν ο αριθμός των νοικοκυριών της πόλης που έχουν άτομα ηλικίας άνω των 65 ετών είναι άγνωστος, τότε η εκτίμηση της συνολικής δαπάνης των νοικοκυριών X_j είναι:

$$\hat{X}_j = \frac{N}{n}\sum_{k=1}^{n_j} x_{jk} = \frac{1.000}{100}\ 739 = 7.390$$

Η διακύμανση της εκτίμησης $\hat{X}_j$ είναι:

$$V(\hat{X}_j) = N^2\ \frac{N-n}{N}\ \frac{\sum_{k=1}^{n_j} x_{jk}^2 - \frac{1}{n}(\sum_{k=1}^{n_j} x_{jk})^2}{n(n-1)} = 1000^2\left(\frac{1000-100}{1000}\right)\frac{8.985 - \frac{1}{100}739^2}{100(100-1)} =$$

320.363

Το τυπικό σφάλμα είναι:

$$SE(\bar{X}_j) = \sqrt{V(\bar{X}_j)} = \sqrt{320.363} = 566$$

Ο συντελεστής μεταβλητότητας είναι:

$$CV(\bar{X}_j) = \frac{SE(\bar{X}_j)}{\bar{X}_j} 100\% = \frac{566}{7.390} 100\% = 7,65\%$$

3.2. ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΚΑΤΑ ΣΤΡΩΜΑΤΑ - ΕΙΣΑΓΩΓΗ

Όταν ο πληθυσμός που εξετάζεται δεν είναι αρκετά ομοιογενής η απλή τυχαία δειγματοληψία ενδέχεται να δώσει μη αντιπροσωπευτικό δείγμα. Σε αυτή την περίπτωση χρησιμοποιούμε την στρωματοποιημένη δειγματοληψία (stratified sampling) όπου ο ερευνώμενος πληθυσμός χωρίζεται σε ορισμένες κατηγορίες οι οποίες ονομάζονται στρώματα. Σε κάθε κατηγορία συγκεντρώνονται όσο το δυνατό πιο ομοιογενή στοιχεία του πληθυσμού έτσι ώστε να διαμορφώνεται η πιο μεγάλη δυνατή διαφοροποίηση μεταξύ των κατηγοριών αυτών, δηλαδή ένας ανομοιογενής πληθυσμός χωρίζεται σε ομοιογενείς υποπληθυσμούς. Με αυτό τον τρόπο επιτυγχάνεται μεγαλύτερη ακρίβεια στις εκτιμήσεις των χαρακτηριστικών του ερευνώμενου πληθυσμού, διότι οι μετρήσεις που προκύπτουν από ομοιογενή στρώματα διαφέρουν ελάχιστα μεταξύ τους. Για κάθε στρώμα σχηματίζουμε ένα απλό τυχαίο δείγμα και στη συνέχεια συνενώνουμε αυτά τα επιμέρους δείγματα σε ένα ενιαίο για ολόκληρο τον πληθυσμό. Αυτό το δείγμα σε σύγκριση με εκείνο της απλής τυχαία δειγματοληψίας είναι περισσότερο αντιπροσωπευτικό. Αυτό γίνεται διότι στο ενιαίο δείγμα θα συμπεριλαμβάνονται δείγματα από όλες τις κατηγορίες (στρώματα) του πληθυσμού πράγμα που είναι αδύνατο στην απλή τυχαία δειγματοληψία, όπου υπάρχει περίπτωση να παραληφθούν μονάδες που αποτελούν ένα ομοιογενές στρώμα, με αποτέλεσμα να μην έχουμε καθόλου πληροφόρηση για την κατηγορία αυτή.

Παράδειγμα

Υποθέτουμε ότι θέλουμε να εκτιμήσουμε το χρόνο τον οποίο διαθέτει ο μέσος σπουδαστής της σχολής Α για μελέτη των μαθημάτων του. Επειδή έχουμε παρατηρήσει ότι ο χρόνος μεταξύ σπουδαστών του ίδιου έτους σπουδών μικρές διαφορές (σε αντίθεση με το χρόνο μεταξύ σπουδαστών διαφορετικού έτους σπουδών) χωρίζουμε τους σπουδαστές της σχολής Α σε τόσα στρώματα όσα είναι τα έτη σπουδών. Στον πίνακα που ακολουθεί οι 6.000 σπουδαστές της σχολής χωρίζονται στα τέσσερα έτη σπουδών (στρώματα) ως εξής:

Έτος σπουδών (στρώμα)	Πληθυσμός σπουδαστών κατά έτος σπουδών (πληθυσμός κατά στρώματα)
Α	1.800
Β	700
Γ	2.500
Δ	1.000
Σύνολο	6.000

Σχηματίζουμε δείγμα από κάθε στρώμα (έτος σπουδών) με κλάσμα δειγματοληψίας 30% (το ίδιο κλάσμα για όλα τα στρώματα). Το συνολικό δείγμα θα αποτελείται από 1.800 σπουδαστές και θα απαρτίζεται από επιμέρους δείγματα των τεσσάρων στρωμάτων ως εξής:

Έτος σπουδών (στρώμα)	Δείγμα σπουδαστών κατά έτος σπουδών (κατά στρώματα)
Α	540
Β	210
Γ	750
Δ	300
Σύνολο	1.800

Συμπεραίνουμε ότι, το δείγμα των 1.800 σπουδαστών αντιπροσωπεύει τους σπουδαστές όλων των ετών των σπουδών. Με τη μέθοδο της απλής τυχαίας δειγματοληψίας θα μπορούσαμε να χρησιμοποιήσουμε το αντίστοιχο δείγμα που αποτελείται μόνο από σπουδαστές του έτους Α ή Β ή Γ ή Δ.

3.2.1. ΠΛΕΟΝΕΚΤΗΜΑΤΑ – ΜΕΙΟΝΕΚΤΗΜΑΤΑ

Πλεονεκτήματα της στρωματοποιημένης δειγματοληψίας:

- Τα δεδομένα σε κάθε στρώμα είναι περισσότερο ομοιογενή από ότι σε ολόκληρο το πληθυσμό, με συνέπεια τη μικρότερη διασπορά στην εκτίμηση των παραμέτρων εξασφαλίζοντας μεγαλύτερη ακρίβεια.
- Το κόστος της δειγματοληψίας μπορεί να μειωθεί σημαντικά, αφού είναι δυνατόν να περιορισθεί το μέγεθος του δείγματος στους επιμέρους ομοιογενείς υποπληθυσμούς.
- Το ενδιαφέρον για εκτιμήσεις κάθε υποπληθυσμού ξεχωριστά.
- Η στρωματοποίηση δίνει τη δυνατότητα αντιμετώπισης των διαφορών που υπάρχουν μεταξύ πληθυσμιακών ομάδων. Για παράδειγμα, η συμπεριφορά των μεγάλων επιχειρήσεων, στο σύνολο των οικονομικών τους δραστηριοτήτων, είναι διαφορετική από αυτή των μικρών επιχειρήσεων. Οι απαιτήσεις των ατόμων που ζουν σε συλλογικές κατοικίες (νοσοκομεία, οικοτροφεία) είναι διαφορετικές από αυτές των ατόμων που ζουν σε κανονικές κατοικίες.

Μειονεκτήματα της στρωματοποιημένης δειγματοληψίας:

- Η διαδικασία επιλογής του δείγματος είναι πιο σύνθετη.
- Η στρωματοποίηση προϋποθέτει περισσότερη προκαταρκτική πληροφόρηση γύρω από το σύνολο του πληθυσμού.
- Το σύνολο του πληθυσμού πρέπει να ορίζεται βάσει των χαρακτηριστικών της στρωματοποίησης.

3.3. ΚΑΤΑΝΟΜΗ ΔΕΙΓΜΑΤΟΣ ΣΤΑ ΣΤΡΩΜΑΤΑ

Με την επιλογή του απαραίτητου μεγέθους δείγματος σε κάθε στρώμα, σκοπός μας είναι να πετύχουμε εκτιμήτριες με μικρές διακυμάνσεις στο ελάχιστο δυνατό κόστος. Επομένως η κατανομή του συνολικού δείγματος μεταξύ των στρωμάτων πρέπει να γίνεται με τέτοιο τρόπο ώστε να επιτυγχάνεται καθορισμένο μέγεθος πληροφοριών στο ελάχιστο κόστος.

Με βάση το σκοπό αυτό για την κατανομή δείγματος μεταξύ των στρωμάτων, πρέπει να λαμβάνονται υπόψη οι παρακάτω παράγοντες:

α) Ο συνολικός αριθμός μονάδων σε κάθε στρώμα,

β) η μεταβλητότητα των μονάδων σε κάθε στρώμα και

γ) το κόστος κατά δειγματοληπτική μονάδα σε κάθε στρώμα.

Η επιλογή του αριθμού των δειγματοληπτικών μονάδων σε κάθε στρώμα επιδρά στο μέγεθος της πληροφόρησης. Για παράδειγμα ένα δείγμα 100 μονάδων από πληθυσμό 2000 μονάδων δίνει περισσότερες πληροφορίες, από ένα δείγμα 100 μονάδων που λαμβάνεται από πληθυσμό 80000 μονάδων. Συνεπώς από στρώματα με μεγαλύτερους πληθυσμούς πρέπει να λαμβάνεται μεγαλύτερο δείγμα.

Η μεταβλητότητα επιδρά στην αξιοπιστία των εκτιμήσεων άρα πρέπει να λαμβάνεται μεγαλύτερο δείγμα από στρώματα που παρουσιάζουν μεγαλύτερη ανομοιογένεια.

Στην περίπτωση που το κατά μονάδα κόστος μιας δειγματοληπτικής έρευνας διαφοροποιείται από στρώμα σε στρώμα, πρέπει να πάρουμε μικρότερο δείγμα από το στρώμα με υψηλό κόστος, έτσι ώστε να κρατήσουμε το κόστος της δειγματοληπτικής έρευνας σε χαμηλά επίπεδα, με την προϋπόθεση ότι δεν επηρεάζεται η αξιοπιστία των αποτελεσμάτων.

Η κατανομή του συνολικού δείγματος στα στρώματα είναι αναγκαίο να γίνεται με μια από τις παρακάτω μεθόδους, προκειμένου να πετύχουμε εκτιμήτριες με μικρές διακυμάνσεις στο ελάχιστο δυνατό κόστος:

α) αναλογική κατανομή δείγματος και

β) άριστη κατανομή δείγματος.

Αρκετά συχνά η κατανομή του δείγματος γίνεται αυθαίρετα στα διάφορα στρώματα, δηλαδή το δείγμα που επιλέγεται σε ένα στρώμα είναι ανεξάρτητο από οποιοδήποτε δείγμα άλλου στρώματος.

3.3.1. ΑΝΑΛΟΓΙΚΗ ΚΑΤΑΝΟΜΗ ΔΕΙΓΜΑΤΟΣ

Όταν σε όλα τα στρώματα το κόστος κατά μονάδα δείγματος και η διακύμανση είναι ίδια τότε εφαρμόζεται η αναλογική κατανομή δείγματος (proportional allocation). Με την αναλογική κατανομή το δείγμα λαμβάνεται αναλογικά προς το μέγεθος των στρωμάτων του ερευνώμενου πληθυσμού, δηλαδή με σταθερό κλάσμα δειγματοληψίας το οποίο δίνει αυτοσταθμιζόμενο δείγμα:

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \cdots = \frac{n_\kappa}{N_\kappa} = \frac{n}{N}$$

Επομένως, ο τύπος προσδιορισμού του δείγματος στο στρώμα i είναι:

$$\frac{n_i}{n} = \frac{N_i}{N} \quad \text{ή} \quad n_i = n\frac{N_i}{N}$$

Παράδειγμα

Δίνεται η κατανομή των σπουδαστών του ΤΕΙ Πατρών, σύμφωνα με τα στοιχεία της απογραφής ενός έτους, κατά ειδικότητα (μέγεθος).

Στρώμα	Ειδικότητα φοιτητών (μέγεθος)	Αριθμός σπουδαστών
1^o	1-2	933

2ο	3-5	883
3ο	6-9	173
4ο	10-19	49
5ο	20-29	12
6ο	30-49	6
7ο	50-99	6
8ο	100-199	4
	Σύνολο	2.066

Έστω ότι θέλουμε να πάρουμε δείγμα 10% με τη μέθοδο της στρωματοποιημένης δειγματοληψίας για να εκτιμήσουμε το μέσο αριθμό φοιτητών. Λαμβάνοντας υπόψη μας ότι το κόστος κατά μονάδα δείγματος είναι το ίδιο και ότι η διακύμανση του αριθμού φοιτητών είναι περίπου ίδια σε όλα τα στρώματα, να κατανεμηθεί το δείγμα στα οκτώ στρώματα.

Λύση

Εφόσον η διακύμανση του αριθμού των φοιτητών είναι περίπου ίδια σε όλα τα στρώματα και το κατά μονάδα κόστος του δείγματος είναι ίδιο, η κατανομή του συνολικού δείγματος θα γίνει αναλογικά, επομένως θα εφαρμόσουμε τη σχέση $\mathbf{n_i = n\frac{N_i}{N}}$ (αφού θέλουμε δείγμα $^{n}/_{N} = 10\%$ του αριθμού σπουδαστών άρα για το πρώτο στρώμα δείγματος έχουμε 933* 0,1), οπότε θα έχουμε τον παρακάτω πίνακα:

Στρώμα	Ειδικότητα φοιτητών (μέγεθος)	Δείγμα $\mathbf{n_i = n\frac{N_i}{N}}$
1ο	1-2	93

2ο	3-5	88
3ο	6-9	17
4ο	10-19	5
5ο	20-29	1
6ο	30-49	1
7ο	50-99	1
8ο	100-199	1
	Σύνολο	207

3.3.2. ΑΡΙΣΤΗ ΚΑΤΑΝΟΜΗ ΔΕΙΓΜΑΤΟΣ

Όταν το ερευνώμενο δείγμα του πληθυσμού παρουσιάζει σε μερικά στρώματα μεγαλύτερη μεταβλητότητα είναι αναγκαίο στα στρώματα αυτά να ληφθεί μεγαλύτερο δείγμα του πληθυσμού, προκειμένου να έχουμε μεγαλύτερη ακρίβεια εκτιμήσεων. Στην περίπτωση που το κόστος συλλογής πληροφοριών είναι μεγαλύτερο σε μερικά στρώματα, χρειαζόμαστε μικρότερο δείγμα προκειμένου να έχουμε μείωση του κόστους της έρευνας. Επομένως όταν η διακύμανση της μεταβλητής που εξετάζουμε και το κόστος συλλογής πληροφοριών διαφοροποιούνται από στρώμα σε στρώμα, οι μονάδες δείγματος πρέπει να κατανέμονται στα στρώματα κατά τέτοιο τρόπο ώστε η διακύμανση της εκτίμησης να ελαχιστοποιηθεί για ένα καθορισμένο κόστος ή το κόστος να ελαχιστοποιηθεί για μια ορισμένη τιμή της διακύμανσης της εκτίμησης.

Η συγκεκριμένη κατανομή του δείγματος ονομάζεται άριστη κατανομή (optimum allocation) και επιτυγχάνεται αν το κλάσμα δειγματοληψίας στα διάφορα στρώματα ληφθεί υπόψη αναλογικά προς την τυπική απόκλιση και αντιστρόφως ανάλογα προς την τετραγωνική ρίζα του κόστους κατά μονάδα, δηλαδή με κλάσμα δειγματοληψίας:

$$\frac{n_1}{N_1 S_1/\sqrt{C_1}} = \frac{n_2}{N_2 S_2/\sqrt{C_2}} = \cdots = \frac{n_\kappa}{N_\kappa S_\kappa/\sqrt{C_\kappa}} = \frac{n}{\sum (N_i S_i/\sqrt{C_i})}$$

όπου C_i = το κόστος κατά μονάδα και

S_i = η τυπική απόκλιση στο στρώμα i, οπότε ο τύπος προσδιορισμού του δείγματος στο στώμα i είναι:

$$\frac{n_i}{N_i S_i/\sqrt{C_i}} = \frac{n}{\Sigma(N_i S_i/\sqrt{C_i})} \quad \text{ή} \quad n_i = n\frac{N_i S_i/\sqrt{C_i}}{\Sigma(N_i S_i/\sqrt{C_i})} = n\frac{W_i S_i/\sqrt{C_i}}{\Sigma(W_i S_i/\sqrt{C_i})}$$

Εάν το κόστος κατά μονάδα είναι σταθερό στα στρώματα του ερευνώμενου πληθυσμού, τότε η διακύμανση της εκτίμησης ελαχιστοποιείται, αν το κλάσμα δειγματοληψίας στα στρώματα , λαμβάνεται αναλογικά προς την τυπική απόκλιση, δηλαδή λαμβάνοντας κλάσμα δειγματοληψίας:

$$\frac{n_1}{N_1 S_1} = \frac{n_2}{N_2 S_2} = \dots = \frac{n_\kappa}{N_\kappa S_\kappa} = \frac{n}{\Sigma N_i S_i}$$

Οπότε ο τύπος προσδιορισμού του δείγματος στο στρώμα i είναι:

$$\frac{n_i}{N_i S_i} = \frac{n}{\Sigma N_i S_i} \quad \text{ή} \quad n_i = n\frac{N_i S_i}{\Sigma N_i S_i} = \frac{W_i S_i}{\Sigma W_i S_i}$$

Η ειδική αυτή περίπτωση της άριστης κατανομής ονομάζεται κατανομή Neyman (Neyman allocation).

Παράδειγμα 1

Δίνεται η στρωματοποίηση των επιχειρήσεων μιας πόλης σύμφωνα με το κεφάλαιο τους και η τυπική απόκλιση κατά στρώμα.

Στρώμα	Κεφάλαιο σε χιλιάδες ευρώ	Αριθμός επιχειρήσεων N_i	Τυπική Απόκλιση S_i
1ο	Μέχρι 999	1.000	150
2ο	1.000-1.999	800	300
3ο	2.000-2.999	500	700
4ο	3.000 και άνω	100	1.100
	Σύνολο	1.400	

Λαμβάνοντας υπόψη ότι το κόστος κατά μονάδα είναι σταθερό στα τέσσερα στρώματα, να κατανεμηθεί δείγμα 200 μονάδων στα στρώματα, έτσι ώστε να ελαχιστοποιηθεί η διακύμανση της εκτίμησης.

Λύση

Επειδή η διακύμανση της εξεταζόμενης μεταβλητής διαφέρει μεταξύ των στρωμάτων, ενώ το κόστος κατά μονάδα είναι ίδιο για την ελαχιστοποίηση της διακύμανσης της εκτίμησης εφαρμόζεται η ειδική περίπτωση της άριστης κατανομής (κατανομή Neyman) για να κατανεμηθεί το δείγμα 200 μονάδων σε τέσσερα στρώματα. Με τη βοήθεια της σχέσης

$$n_i = n \frac{N_i S_i}{\sum N_i S_i}$$

φτιάχνουμε τον παρακάτω πίνακα υπολογισμών, όπου στην τελευταία στήλη φαίνεται η κατανομή των 200 μονάδων στα τέσσερα στρώματα.

Στρώμα	Κεφάλαιο σε χιλιάδες ευρώ	$N_i S_i$	$\frac{N_i S_i}{\sum N_i S_i}$	$n_i = n \frac{N_i S_i}{\sum N_i S_i}$
				$n_1 = 200 \cdot 0{,}176 = 35$
1^o	Μέχρι 999	150.000	0,176	$n_2 = 200 \cdot 0{,}282 = 57$
2^o	1.000-1.999	240.000	0,282	$n_3 = 200 \cdot 0{,}411 = 82$
3^o	2.000-2.999	350.000	0,411	$n_4 = 200 \cdot 0{,}129 = 26$
4^o	3.000 και άνω	110.000	0,129	
	Σύνολο	850.000		200

Παράδειγμα 2

Δίνεται η στρωματοποίηση 15.000 καταστημάτων του κλάδου ένδυσης σύμφωνα με τις ετήσιες πωλήσεις αυτών, η τυπική απόκλιση και το κόστος κατά μονάδα των στρωμάτων.

Στρώμα	Ετήσιες πωλήσεις σε τόνους	Αριθμός καταστημάτων N_ι	Τυπική απόκλιση S_i	Κόστος κατά μονάδα C_i
1^o	Μέχρι 99	7.000	75	200
2^o	100-199	4.000	180	430
3^o	200-399	2.500	340	600

4ο	400 και άνω	1.500	450	820
	Σύνολο	15.000		

Να κατανεμηθεί δείγμα 500 μονάδων στα τέσσερα στρώματα, με σκοπό την ελαχιστοποίηση της διακύμανσης της εκτίμησης.

Λύση

Επειδή η διακύμανση της εξεταζόμενης μεταβλητής και το κόστος κατά μονάδα διαφέρουν μεταξύ των στρωμάτων, για την ελαχιστοποίηση της διακύμανσης της εκτίμησης εφαρμόζεται η ειδική περίπτωση της άριστης κατανομής (κατανομή Neyman) για να κατανεμηθεί το δείγμα 500 μονάδων σε τέσσερα στρώματα. Με τη βοήθεια της σχέσης

$$n_i = n\frac{N_i S_i/\sqrt{C_i}}{\Sigma(N_i S_i/\sqrt{C_i})}$$

φτιάχνουμε τον παρακάτω πίνακα υπολογισμών, όπου στην τελευταία στήλη φαίνεται η κατανομή των 500 μονάδων στα τέσσερα στρώματα.

Στρώμα	$N_i S_i$	$\sqrt{C_i}$	$\frac{N_i S_i}{\sqrt{C_i}}$	$\frac{N_i S_i/\sqrt{C_i}}{\Sigma(N_i S_i/\sqrt{C_i})}$	$n_i = n\frac{N_i S_i/\sqrt{C_i}}{\Sigma(N_i S_i/\sqrt{C_i})}$
1ο	525.000	14,1	37.234	0,285	n_1=500·0,285=143
2ο	720.000	20,7	34.782	0,266	n_2=500·0,266=133
3ο	850.000	24,4	34.836	0,267	n_3=500·0,267=134
4ο	675.000	28,6	23.601	0,180	n_4=500·0,180=90
Σύνολο			130.453		500

3.3.3. ΕΚΤΙΜΗΣΗ ΜΕΣΟΥ ΚΑΙ ΣΥΝΟΛΙΚΟΥ ΠΛΗΘΥΣΜΟΥ

Ο μέσος του πληθυσμού είναι:

$$\bar{X} = \frac{\sum_{i=1}^{\kappa}\sum_{j=1}^{N_i} X_{ij}}{N} = \frac{\sum_{i=1}^{\kappa} N_i \bar{X_i}}{N} = \sum_{i=1}^{\kappa} N_i \bar{X_i}$$

δηλαδή ο μέσος σταθμικός των μέσων στρωμάτων, με στάθμιση W_i για κάθε στρώμα i.

Επειδή μέσα σε κάθε στρώμα η δειγματοληψία είναι απλή τυχαία, ο μέσος δείγματος $\bar{x}$ είναι αμερόληπτη εκτιμήτρια του μέσου $\bar{X_i}$ του στρώματος i συνεπώς μια αμερόληπτη εκτίμηση του μέσου $\bar{X}$ του πληθυσμού είναι ο μέσος σταθμικός των εκτιμήσεων των i στρωμάτων και συμβολίζεται με $\bar{x_{st}}$. Ο δείκτης st σημαίνει στρωματοποιημένη δειγματοληψία, δηλαδή:

$$\bar{x}_{st} = \frac{1}{N}\left(N_1\bar{x_1} + N_2\bar{x_2} + \cdots + N_\kappa\bar{x_\kappa}\right) = \frac{\sum_{i=1}^{\kappa} N_i\bar{X_i}}{N} = \sum_{i=1}^{\kappa} N_i\bar{X_i}$$

Η διακύμανση της εκτίμησης του $\bar{x_{st}}$ εξαρτάται από τον τρόπο κατανομής του δείγματος στα στρώματα. Άρα έχουμε τις εξής περιπτώσεις:

α) εάν το δείγμα που επιλέγεται σε ένα στρώμα είναι ανεξάρτητο από οποιοδήποτε δείγμα άλλου στρώματος τότε η αμερόληπτη εκτίμηση της διακύμανσης του $\bar{x_{st}}$ είναι:

$$V(\bar{x}_{st}) = \frac{1}{N^2}\left[N_1^2 V(\bar{x}_1) + N_2^2 V(\bar{x}_2) + \ \ldots + N_\kappa^2 V(\bar{x}_\kappa)\right] =$$

$$= \frac{1}{N^2}\sum_{i=1}^{\kappa} {N_i}^2 V(\bar{x}_i) = \frac{1}{N^2}\sum_{i=1}^{\kappa} {N_i}^2 \frac{S_i^2}{n_i}\left(\frac{N_i - n_i}{N_i}\right) =$$

$$= \sum_{i=1}^{\kappa} {W_i}^2 \frac{S_i^2}{n_i}\left(\frac{N_i - n_i}{N_i}\right)$$

όπου $V(\bar{x}_i) = \frac{S_i^2}{n_i}\left(\frac{N_i - n_i}{N_i}\right)$ και $SE(\bar{x}_{st}) = \sqrt{V(\bar{x}_{st})}$

Αν η S_i^2 είναι άγνωστη παίρνουμε την αμερόληπτη εκτίμηση της s_i^2, άρα η αμερόληπτη εκτίμηση της διακύμανσης του $\bar{x}_{st}$ είναι:

$$V(\bar{x}_{st}) = \sum_{i=1}^{\kappa} {W_i}^2 \frac{s_i^2}{n_i}\left(\frac{N_i - n_i}{N_i}\right) = \sum_{i=1}^{\kappa} \frac{{W_i}^2 s_i^2}{n_i} - \sum_{i=1}^{\kappa} \frac{{W_i}^2 s_i^2}{N}$$

όπου $s_i^2 = \frac{\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x_i}\right)^2}{n_i - 1}$

β) Στην περίπτωση που έχουμε αναλογική κατανομή του δείγματος, η διακύμανση της εκτίμησης του $\bar{x}_{st}$ συμβολίζεται με $V(\bar{x}_{st})_{pr}$ αντί $V(\bar{x}_{st})$ το οποίο συμβολίζει την διακύμανση της εκτίμησης του $(\bar{x}_{st})$ όταν η κατανομή του δείγματος γίνει αυθαίρετη .

Στην περίπτωση αυτή η διακύμανση της εκτίμησης του $\bar{x}_{st}$ προκύπτει από τη σχέση:

$$\sum_{i=1}^{\kappa} W_i^2 \frac{S_i^2}{n_i}\left(\frac{N_i - n_i}{N_i}\right)$$

αντικαθιστώντας το n_i με $n\frac{N_i}{N}$.

Άρα έχουμε:

$$V(\bar{x}_{st})_{pr} = \frac{1}{n}\left(\frac{N-n}{N}\right)\sum_{i=1}^{k} w_i \; S_i^2$$

και $$SE(\bar{x}_{st})_{pr} = \sqrt{V(\bar{x}_{st})_{pr}}$$

Παράλληλα προκύπτει το σχετικό τυπικό σφάλμα:

$$CV(\bar{x}_{st})_{pr} = \frac{SE(\bar{x}_{st})_{pr}}{\bar{x}_{st}}$$

Εάν οι διακυμάνσεις έχουν την ίδια τιμή S_w^2 σε όλα τα στρώματα, τότε η διακύμανση του $\bar{x}_{st}$ βρίσκεται από τη σχέση:

$$V(\bar{x}_{st})pr = \frac{S_w^2}{n}\left(\frac{N-n}{N}\right)$$

Όταν η S_i^2 είναι άγνωστη τότε η εκτίμηση της διακύμανσης του $\bar{x}_{st}$ υπολογίζεται από την αμερόληπτη εκτίμηση αυτής s_i^2,άρα :

$$V(\bar{x}_{st})_{pr} = \frac{1}{n}\left(\frac{N-n}{N}\right)\sum_{i=1}^{\kappa} W_i \; s_i^2$$

γ) Στην περίπτωση που έχουμε άριστη κατανομή του δείγματος η διακύμανση της εκτίμησης του $\bar{x}_{st}$ συμβολίζεται με $V(\bar{x}_{st})_{Ne}$ όταν το κόστος κατά μονάδα είναι ίδιο σε όλα τα στρώματα (κατανομή Neyman) ενώ στην περίπτωση που το κόστος κατά μονάδα διαφέρει μεταξύ των στρωμάτων συμβολίζεται με $V(\bar{x}_{st})_{opt}$.

Η διακύμανση της εκτίμησης του $\bar{x}_{st}$ όταν έχουμε άριστη κατανομή με σταθερό κόστος κατά μονάδα είναι:

i) $V(\bar{x}_{st})_{Ne} = \frac{\left(\sum_{i=1}^{\kappa} W_i S_i\right)^2}{n} - \frac{\sum_{i=1}^{\kappa} W_i S_i^2}{N}$ ενώ στην περίπτωση που έχουμε άριστη κατανομή με διαφορετικό κόστος κατά μονάδα είναι:

ii) $SE(\bar{x}_{st})_{opt} = \frac{(W_i S_i \sqrt{C_i})(W_i S_i / \sqrt{C_i})}{n} - \frac{\sum_{i=1}^{\kappa} W_i S_i^2}{N}$

Ακολουθεί η εκτίμηση του συνολικού πληθυσμού Y:

$\hat{X} = \sum_{j=1}^{n_j} x_{ij} \frac{N_i}{n_i}$ όπου $\frac{N_i}{n_i}$ ο συντελεστής αναγωγής, δηλαδή το αντίστροφο του κλάσματος της δειγματοληψίας για κάθε στρώμα.

Στην περίπτωση που το δείγμα κατανέμεται αυθαίρετα στα διάφορα στρώματα, η διακύμανση της εκτίμησης του συνολικού πληθυσμού Y είναι:

$$V(\hat{X}) = V(N\bar{x}_{st}) = N^2 V(\bar{x}_{st}) = \sum_{i=1}^{\kappa} N_i^2 \frac{S_i^2}{n_i} \left(\frac{N_i - n_i}{N_i}\right)$$

Όταν η κατανομή του δείγματος είναι αναλογική, η διακύμανση του συνολικού πληθυσμού είναι:

$$V(\hat{X})_{pr} = V(N\bar{x}_{st})_{pr} = N^2 V(\bar{x}_{st})_{pr} = \frac{N-n}{n} \sum_{i=1}^{\kappa} N_i S_i^2$$

Συνεπώς σε περίπτωση άριστης κατανομής με διαφορετικό κόστος κατά μονάδα η διακύμανση της εκτίμησης του συνολικού πληθυσμού Y είναι:

$$V(\bar{x}_{st})_{Ne} = N^2 (\bar{x}_{st})_{Ne} \text{ και } V(\hat{X})_{opt} = N^2 V(\bar{x}_{st})_{opt}$$

Τα τυπικά σφάλματα είναι αντίστοιχα:

$$V(\hat{X})_{Ne} = \sqrt{V(\hat{X})_{Ne}} \text{ και } SE(\hat{X})_{opt} = \sqrt{V(\hat{X})_{opt}}$$

Παράδειγμα

Υποθέτουμε ότι παίρνουμε ένα δείγμα 10% από μια πόλη με τη μέθοδο της τυχαίας δειγματοληψίας, με σκοπό την εκτίμηση του μέσου αριθμού απασχολούμενων στα καταστήματα κατασκευής επίπλων. Από την απογραφή προκύπτει η κατανομή των καταστημάτων ως προς το μέγεθος απασχόλησης:

Στρώμα	Τάξεις απασχολούμενων	Αριθμός καταστημάτων
1ο	1-2	50
2ο	3-5	30
3ο	6-9	20
Σύνολο		100

Γνωρίζοντας ότι το κόστος κατά μονάδα δείγματος είναι ίδιο σε όλα τα στρώματα η κατανομή του δείγματος πραγματοποιείται αναλογικά, επομένως παίρνουμε 5 μονάδες από το ένα στρώμα, 3 από το δεύτερο και 2 από το τρίτο. Από την έρευνα που έγινε, βρέθηκαν οι παρακάτω απασχολούμενοι κατά κατάστημα στα 3 στρώματα:

$1^{ο}$ στρώμα: 1,2,1,2,1

$2^{ο}$ στρώμα: 3,5,3

$3^{ο}$ στρώμα: 9,6

Να εκτιμηθούν: α) ο μέσος αριθμός απασχολουμένων στα καταστήματα κατασκευής επίπλων της πόλης, το τυπικό σφάλμα και το σχετικό τυπικό σφάλμα β) το σύνολο των απασχολουμένων στα καταστήματα επίπλων της πόλης, το τυπικό σφάλμα και το σχετικό τυπικό σφάλμα.

Λύση

α) Οι αριθμητικοί μέσοι και οι διακυμάνσεις κατά στρώμα υπολογίζονται ως εξής:

$$\bar{X}_1 = \frac{\sum_{j=1}^{N_1} X_{1j}}{N} = \frac{1+2+1+2+1}{5} = \frac{7}{5} = 1{,}4$$

$$\bar{X}_2 = \frac{\sum_{j=2}^{N_i} X_{2j}}{N} = \frac{3+5+3}{3} = \frac{11}{3} = 3{,}66$$

$$\bar{X}_3 = \frac{\sum_{j=3}^{N_i} X_{1j}}{N} = \frac{9+6}{2} = \frac{15}{2} = 7{,}5$$

$$s_1^2 = \frac{\sum_{j=1}^{n_1}(x_{1j} - \bar{x}_1)^2}{n_1 - 1} = \frac{(1-1{,}4)^2 + (2-1{,}4)^2 + \cdots + (1-1{,}4)^2}{5-1} = 0{,}3$$

$$s_2^2 = \frac{\sum_{j=1}^{n_2}(x_{2j} - \bar{x}_2)^2}{n_2 - 1} = \frac{(3-3{,}66)^2 + (5-3{,}66)^2 + (3-3{,}66)^2}{3-1} = 1{,}33$$

$$s_3^2 = \frac{\sum_{j=1}^{n_3}(x_{3j} - \bar{x}_3)^2}{n_3 - 1} = \frac{(9-7{,}5)^2 + (6-7{,}5)^2 +}{2-1} = 4{,}5$$

Η εκτίμηση του μέσου αριθμού απασχολουμένων στα καταστήματα κατασκευής επίπλων είναι:

$$\bar{x}_{st} = \frac{1}{N}\left(N_1\overline{x_1} + N_2\overline{x_2} + \cdots + N_\kappa\overline{x_\kappa}\right) = \frac{1}{100}(50 \cdot 1{,}4 + 30 \cdot 3{,}66 + 20 \cdot 7{,}5) = 3{,}29$$

Η διακύμανση του μέσου $\bar{x}_{st}$ στην αναλογική κατανομή είναι:

$$V(\bar{x}_{st})_{pr} = \frac{1}{n}\left(\frac{N-n}{N}\right)\sum_{i=1}^{\kappa} w_i \; S_i^2 =$$

$$= \frac{1}{n}\left(\frac{N-n}{N}\right)\left(\frac{N_1}{N}s_1^2 + \frac{N_2}{N}s_2^2 + \frac{N_3}{N}s_3^2\right) =$$

$$= \frac{1}{10}\left(\frac{100-10}{100}\right)\left(\frac{50}{100}\cdot 0{,}3 + \frac{30}{100}\cdot 1{,}33 + \frac{20}{100}\cdot 4{,}5\right) = 0{,}13$$

Το τυπικό σφάλμα είναι:

$$SE(\bar{x}_{st})_{pr} = \sqrt{V(\bar{x}_{st})_{pr}} = \sqrt{0{,}13} = 0{,}36$$

Το σχετικό τυπικό σφάλμα είναι:

$$CV(\bar{x}_{st})_{pr} = \frac{SE(\bar{x}_{st})_{pr}}{\bar{x}_{st}} = \frac{0{,}36}{3{,}29} = 0{,}109 \quad \text{ή} \quad 10{,}9\%$$

β) Η εκτίμηση του συνολικού αριθμού απασχολουμένων στα καταστήματα επίπλων είναι:

$$\hat{X} = \sum_{j=1}^{n_j} x_{ij}\,\frac{N_i}{n_i} = \sum_{j=1}^{n_1} x_{1j}\,\frac{N_1}{n_1} + \sum_{j=1}^{n_2} x_{2j}\,\frac{N_2}{n_2} + \sum_{j=1}^{n_3} x_{3j}\,\frac{N_3}{n_3} =$$

$$= 7\cdot\frac{50}{5} + 11\cdot\frac{30}{3} + 15\frac{20}{2} = 330$$

Η διακύμανση της εκτίμησης του συνόλου των απασχολημένων είναι:

$$V(\hat{X})_{pr} = V(N\bar{x}_{st})_{pr} = N^2 V(\bar{x}_{st})_{pr} = 100^2\cdot 0{,}13 = 1.300$$

Το τυπικό σφάλμα της εκτίμησης $\hat{X}$ είναι:

$$SE(\hat{x}_{st})_{pr} = \sqrt{V(\hat{x}_{st})_{pr}} = \sqrt{1300} = 36{,}1$$

Το σχετικό τυπικό σφάλμα είναι:

$$CV(\hat{x}_{st})_{pr} = \frac{SE(\hat{x}_{st})_{pr}}{\hat{x}_{st}} = \frac{36{,}1}{330} = 0{,}109 \text{ ή } 10{,}9\%$$

3.3.4. ΕΚΤΙΜΗΣΗ ΠΟΣΟΣΤΟΥ Η ΑΝΑΛΟΓΙΑΣ ΠΛΗΘΥΣΜΟΥ

Για την εκτίμηση κατά στρώματα του ποσοστού ή της αναλογίας των μονάδων ενός πληθυσμού οι οποίες ανήκουν σε μια συγκεκριμένη τάξη C, κατηγοριοποιούμε τον πληθυσμό σε στρώματα. Υποθέτουμε ότι A_i από τις N_i μονάδες του στρώματος i , ανήκουν στην τάξη C, τότε το ποσοστό των μονάδων θα είναι: $P_i = \frac{A_i}{N_i}$.

Το ποσοστό των μονάδων που δεν ανήκουν στην τάξη C είναι: $Q_i = 1-P_i$.

Στις περισσότερες περιπτώσεις το ποσοστό P_i είναι άγνωστο και παίρνουμε τη δειγματική εκτίμηση αυτού $p_i = \frac{a_i}{n_i}$.

Το πλήθος των μονάδων της τάξης C εντός του στρώματος i εκφράζεται από το γινόμενο $N_i p_i$. Επομένως το άθροισμα δίνεται από τη σχέση :

$$N_1 p_1 + N_2 p_2 + \cdots + N_\kappa p_\kappa$$

Η εκτίμηση του ποσοστού P του πληθυσμού προκύπτει από τη διαίρεση της παραπάνω σχέσης με το σύνολο του πληθυσμού N:

$$p_{st} = \frac{1}{N} \sum_{i=1}^{\kappa} N_i p_i$$

Η διακύμανση S_i^2 του στρώματος i είναι:

$$S_i^2 = \frac{N_i}{N_i - 1} P_i Q_i$$

Οπότε η διακύμανση της εκτίμησης του p_{st} είναι:

$$V(p_{st}) = \sum_{i=1}^{\kappa} W_i^2 \frac{N_i - n_i}{N_i - 1} \cdot \frac{P_i Q_i}{n_i}$$

Επειδή όμως οι όροι $1/N_i$ θεωρούνται αμελητέοι τότε καταλήγουμε στη σχέση:

$$V(p_{st}) = \sum_{i=1}^{\kappa} W_i^2 \frac{N_i - n_i}{N_i} \cdot \frac{P_i Q_i}{n_i} \quad \text{και} \quad SE(p_{st}) = \sqrt{V(p_{st})}$$

Όταν η διακύμανση του p_{st} υπολογίζεται από στοιχεία δείγματος τότε έχουμε:

$$V(p_{st}) = \sum_{i=1}^{\kappa} W_i^2 \frac{N_i - n_i}{N_i} \cdot \frac{p_i q_i}{n_{i-1}} \quad \text{όπου} \quad q_i = 1 - p_i$$

Στην περίπτωση αναλογικής κατανομής του δείγματος η διακύμανση της εκτίμησης p_{st} είναι:

$$V(p_{st})_{pr} = \frac{1}{n}\left(\frac{N-n}{N}\right) \sum_{i=1}^{\kappa} W_i P_i Q_i \quad \text{και} \quad SE(p_{st})_{pr} = \sqrt{V(p_{st})_{pr}}$$

Σε περίπτωση που η κατανομή είναι άριστη, η διακύμανση της εκτίμησης p_{st} είναι:

α) $n_i = n \frac{N_i \sqrt{P_i Q_i}}{\sum N_i \sqrt{P_i Q_i}}$ (κατανομή Neyman)

β) $n_i = n \frac{N_i \sqrt{P_i Q_i / C_i}}{\sum N_i \sqrt{P_i Q_i / C_i}}$ (Άριστη κατανομή με διαφορετικό κόστος κατά μονάδα όπου $P_i Q_i$ επέχει τη θέση της διακύμανσης S_i^2.

Παράδειγμα

Έστω ότι θέλουμε να εκτιμήσουμε το ποσοστό των εταιρειών μιας πόλης, οι οποίες προτιμούν την αγορά ενός προϊόντος Α έναντι ενός άλλου προϊόντος Β. Χωρίζουμε τις εταιρείες της πόλης σε 4 στρώματα ανάλογα με το ύψος του κεφαλαίου και παίρνουμε δείγμα 90 μονάδων, εφαρμόζοντας την αναλογική κατανομή. Από τη δειγματοληπτική έρευνα προκύπτουν τα παρακάτω αποτελέσματα:

Στρώματα	Αριθμός εταιρειών N_i	Δείγμα n_i	Αριθμός εταιρειών που προτιμούν το προϊόν Α	p_i
1ο	150	20	12	0,60
2ο	90	25	17	0,68
3ο	230	32	8	0,25
4ο	130	13	11	0,84
	600	90		

Να εκτιμηθεί το ποσοστό των εταιρειών που προτιμούν το προϊόν Α και να υπολογιστεί το τυπικό σφάλμα.

Λύση

Η εκτίμηση του ποσοστού των εταιρειών της πόλης που προτιμούν το προϊόν Α είναι:

$$p_{st} = \frac{1}{N}\sum_{i=1}^{\kappa} N_i p_i = \frac{1}{600}(150 \cdot 0{,}60 + 90 \cdot 0{,}68 + 230 \cdot 0{,}25 + 130 \cdot 0{,}84) = 0{,}529 \text{ ή } 52{,}9\%$$

Η διακύμανση της εκτίμησης p_{st} είναι:

$$V(p_{st})_{pr} = \frac{1}{n}\left(\frac{N-n}{N}\right)\sum_{i=1}^{\kappa} W_i \frac{n_i}{n_i - i} p_i\, q_i =$$

$$\frac{1}{90}\left(\frac{600-90}{600}\right)\left[\left(\frac{150}{600}\cdot\frac{20}{19}\cdot 0{,}60 \cdot 0{,}40\right) + \left(\frac{90}{600}\cdot\frac{25}{24}\cdot 0{,}68 \cdot 0{,}32\right) + \left(\frac{230}{600}\cdot\frac{32}{31}\cdot 0{,}25 \cdot 0{,}75\right) + \left(\frac{130}{600}\cdot\frac{13}{12}\cdot 0{,}84 \cdot 0{,}16\right)\right] =$$

0,00192

Το τυπικό σφάλμα της εκτίμησης p_{st} είναι:

$$SE(p_{st})_{pr} = \sqrt{V(p_{st})_{pr}} = \sqrt{0{,}00192} = 0{,}043.$$

3.3.5. ΣΤΡΩΜΑΤΟΠΟΙΗΣΗ ΜΕΤΑ ΤΗ ΣΥΛΛΟΓΗ ΤΟΥ ΔΕΙΓΜΑΤΟΣ

Μερικές φορές η στρωματοποίηση του ερευνώμενου πληθυσμού είναι αδύνατο να πραγματοποιηθεί πριν τη συλλογή του δείγματος επειδή δεν μπορούμε να γνωρίζουμε εκ των προτέρων σε ποιο στρώμα ανήκει μια δειγματοληπτική μονάδα. Στην περίπτωση αυτή είναι αναγκαία η συλλογή απλού τυχαίου δείγματος το οποίο διαιρείται σε στρώματα μετά τη συλλογή των μονάδων και μεταχειρίζεται σα να ήταν στρωματοποιημένο δείγμα. Με την προϋπόθεση ότι το κλάσμα $\frac{N_i}{N}$ είναι γνωστό ο μέσος του πληθυσμού $\bar{X}$ προσδιορίζεται από την εκτίμηση του $\bar{x}_{st}$ όμως αν το $\frac{N_i}{N}$ είναι γνωστό και το δείγμα μεγάλο ($n_i \geq 20$ σε κάθε στρώμα) τότε η μέθοδος της στρωματοποίησης μετά τη συλλογή δείγματος έχει περίπου την ακρίβεια της αναλογικής στρωματοποιημένης δειγματοληψίας.

Ο μέσος του πληθυσμού $\bar{X}$ εκτιμάται από τη σχέση: $\bar{x}_{st} = \sum_{i=1}^{\kappa} W_i \bar{x}_i$

Η διακύμανση της εκτίμησης $\bar{x}_{st}$ δίνεται από τη σχέση:

$$V(\bar{x}_{st})_{post} = \frac{1}{n}\left(\frac{N-n}{N}\right)\sum_{i=1}^{\kappa} W_i S_i^2 = \frac{1}{n^2}\sum_{i=1}^{\kappa}(1-W_i) S_i^2$$

Ο τύπος αυτός προκύπτει από το άθροισμα της διακύμανσης του $\bar{x}_{st}$ στην αναλογική στρωματοποιημένη δειγματοληψία, όπου είναι ο πρώτος όρος και του δεύτερου όρου που παρουσιάζει την αύξηση της διακύμανσης η οποία οφείλεται στη στρωματοποίηση μετά τη συλλογή του δείγματος. συγκρίνοντας του δύο όρους παρατηρούμε ότι ο δεύτερος είναι μικρότερος σε σχέση με τον πρώτο, όταν το n είναι μεγάλο, με αποτέλεσμα η στρωματοποίηση μετά τη συλλογή του δείγματος να δίνει περίπου την ίδια ακρίβεια που δίνει η αναλογική στρωματοποιημένη δειγματοληψία.

Μια αμερόληπτη εκτιμήτρια της διακύμανσης είναι:

$$V(\bar{x}_{st})_{post} = \sum_{i=1}^{\kappa} W_i^2 \cdot \frac{s_i^2}{n_i}$$

Παράδειγμα

Παίρνουμε δείγμα από τις ηλικίες των κατοίκων μιας περιοχής. Από προηγούμενη έρευνα γνωρίζουμε ότι 43% είναι άνδρες και το 57% γυναίκες. Λαμβάνοντας από το σύνολο των κατοίκων ένα απλό τυχαίο δείγμα N=200, προκύπτει ότι:

Άνδρες	Γυναίκες
$n_1 = 80$	$n_2=120$
$\bar{x}_1=55$	$\bar{x}_2=61$
$s_1^2=111$	$s_2^2=89$

Να εκτιμηθούν:

Α) η μέση ηλικία των κατοίκων,

Β)το τυπικό σφάλμα και το σχετικό τυπικό σφάλμα.

Λύση

Α) Η εκτίμηση της μέσης ηλικία των κατοίκων της περιοχής :

$$\bar{x}_{st} = \sum_{i=1}^{\kappa} W_i \bar{x}_i = W_1\bar{x}_1 + W_2\bar{x}_2 = 0{,}43 \cdot 55 + 0{,}57 \cdot 61 = 58{,}42$$

Η διακύμανση της εκτίμησης $\bar{x}_{st}$ είναι:

$$V(\bar{x}_{st})_{post} = \sum_{i=1}^{\kappa} W_i^2 \cdot \frac{s_i^2}{n_i} = W_1^2 \cdot \frac{s_1^2}{n_1} + W_2^2 \cdot \frac{s_2^2}{n_2} = 0{,}43^2 \cdot \frac{111}{80} + 0{,}57^2 \cdot \frac{89}{120} = 0{,}496$$

Το τυπικό σφάλμα τις εκτίμησης $\bar{x}_{st}$ είναι:

$$SE(\bar{x}_{st}) = \sqrt{V(\bar{x}_{st})_{post}} = \sqrt{0{,}496} = 0{,}704$$

Και το σχετικό τυπικό σφάλμα είναι:

$$CV(\bar{x}_{st}) = \frac{SE(\bar{x}_{st})}{\bar{x}_{st}} = \frac{0{,}704}{58{,}42} = 0{,}012$$ ή 1,2%.

3.3.6. ΣΥΓΚΡΙΣΗ ΑΠΛΗΣ ΤΥΧΑΙΑΣ ΚΑΙ ΣΤΡΩΜΑΤΟΠΟΙΗΜΕΝΗΣ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ

Η σύγκριση λοιπόν των δύο ειδών δειγματοληψίας, από άποψη ακρίβειας έχει ως αποτέλεσμα τη μικρότερη διασπορά για την εκτιμήτρια της μέσης τιμής του πληθυσμού κατά τη στρωματοποιημένη δειγματοληψία. Δεν ισχύει πάντα όμως ότι οποιοδήποτε στρωματοποιημένο δείγμα δίνει μικρότερη διασπορά από ένα απλό τυχαίο δείγμα. Στην περίπτωση που οι τιμές $n_1,n_2,\ldots,n_k$ διαφέρουν από αυτές του βέλτιστου καταμερισμού, τότε η εκτίμηση της μέσης τιμής μπορεί να έχει μεγαλύτερη διασπορά. Ο πληθυσμός διαχωρίζεται σε στρώματα σύμφωνα με παράγοντες οι οποίοι σχετίζονται με το χαρακτηριστικό που εξετάζουμε. Όμως, αυτό δεν είναι πάντοτε δυνατό δεδομένου ότι πολλά χαρακτηριστικά δεν είναι εύκολο να ταυτοποιηθούν για κάθε μονάδα του πληθυσμού.

Καταλήγοντας, παρατηρούμε ότι η στρωματοποιημένη δειγματοληψία χρησιμοποιείται ευρύτατα διότι ωφελούμαστε σε ακρίβεια έστω αν είναι ελάχιστη και επίσης η στρωματοποιημένη σε πολλές περιπτώσεις είναι από μόνη της έτοιμη, αφού ο πληθυσμός είναι από μόνος του χωρισμένος σε στρώματα.

3.4. ΣΥΣΤΗΜΑΤΙΚΗ ΔΕΙΓΜΑΤΟΛΗΨΙΑ - ΕΙΣΑΓΩΓΗ

Η συστηματική δειγματοληψία (systematic sampling) είναι μια μέθοδος επιλογής δείγματος η οποία χρησιμοποιείται συχνά στην πράξη και για την εφαρμογή της απαιτείται κατάλληλο δειγματοληπτικό πλαίσιο. Όλες οι δειγματοληπτικές μονάδες πρέπει να είναι καταχωρημένες στο κατάλογο αυτό με σκοπό όλες οι μονάδες του ερευνώμενου πληθυσμού να έχουν την ίδια ευκαιρία επιλογής. Επιπλέον η συγκεκριμένη μέθοδος δειγματοληψίας έχει την δυνατότητα να εφαρμοστεί και σε άγνωστους πληθυσμούς, όπου το δειγματοληπτικό πλαίσιο δεν είναι αναγκαίο. Για παράδειγμα στην περίπτωση που επιθυμούμε να εκτιμήσουμε τον αριθμό των ατόμων που εισέρχονται σε ένα ηλεκτρονικό κατάστημα μπορούμε να εφαρμόσουμε συστηματική δειγματοληψία.

Η εφαρμογή της συστηματικής δειγματοληψίας γίνεται ως εξής:

Υποθέτουμε ότι οι μονάδες ενός πληθυσμού είναι N οι οποίες έχουν καταχωρηθεί σε ένα κατάλογο και αριθμούνται από ένα 1 έως N και θέλουμε να επιλέξουμε n μονάδες.

Αρχικά υπολογίζουμε το διάστημα της δειγματοληψίας (sampling interval) το οποίο είναι: $\lambda=\frac{N}{n}$.

Στη συνέχεια επιλέγουμε τυχαία έναν αριθμό μεταξύ των πρώτων λ αριθμών του καταλόγου, δηλαδή από το 1 ως το λ. Ο αριθμός αυτός ονομάζεται τυχαίο ξεκίνημα (random start). Εάν i είναι ο επιλεγόμενος αριθμός μεταξύ 1 και λ, τότε οι μονάδες του πληθυσμού που θα επιλεγούν στο δείγμα είναι αυτές που έχουν αύξοντες αριθμούς:

$i, i + \lambda, i + 2\lambda, \ldots., i + (n - 1)\lambda$

Συνεπώς οι επιλεγόμενες μονάδες θα είναι:

$x_i, x_{i+\lambda}, x_{i+2\lambda}, \ldots, x_{i+(n-1)\lambda}$

Παράδειγμα

Έστω ότι θέλουμε να επιλέξουμε 50 μαθητές ενός σχολειού από 500 που διαθέτει συνολικά, για τη διενέργεια μιας έρευνας. Οι μαθητές αυτοί είναι καταχωρημένοι ονομαστικά σε ένα κατάλογο με αύξουσα αρίθμηση.

Αρχικά υπολογίζουμε το διάστημα δειγματοληψίας

$$\lambda = \frac{N}{n} = \frac{500}{50} = 10$$

Επιλέγουμε τυχαία έναν αριθμό από το 1 έως το 10. Έστω ο αριθμός 7 είναι το τυχαίο ξεκίνημα τότε στο δείγμα θα επιλεγούν οι μαθητές που έχουν αύξοντες αριθμούς : 7, 17, 27, 37, … , 487 , 497.

Στη πραγματικότητα κατά τη συστηματική δειγματοληψία ο πληθυσμός χωρίζεται σε λ ισοπληθείς ομάδες των n μονάδων η καθεμία. Δηλαδή έχουμε N=n λ. Στο παρακάτω πίνακα δίνονται τα λ δυνατά συστηματικά δείγματα.

Συστηματικά δείγματα	Επιλεγόμενες μονάδες
1	x_1 $x_{1+\lambda}$ $x_{1+2\lambda}$ … $x_{1+(n-1)\lambda}$
2	x_2 $x_{2+\lambda}$ $x_{2+2\lambda}$ … $x_{2+(n-1)\lambda}$
…	… … … … …
i	x_i $x_{i+\lambda}$ $x_{i+2\lambda}$ … $x_{i+(n-1)\lambda}$
…	… … … … ….
λ	x_λ $x_{2\lambda}$ $x_{3\lambda}$ $x_{n\lambda}$

Η πιθανότητα να επιλεγεί ένα από τα λ συστηματικά δείγματα είναι $\frac{1}{\lambda}$.Άρα σύμφωνα με τη μέθοδο της συστηματικής δειγματοληψίας, κάθε μονάδα του πληθυσμού έχει την ίδια ευκαιρία να επιλεγεί στο δείγμα.

3.4.1. ΠΛΕΟΝΕΚΤΗΜΑΤΑ - ΜΕΙΟΝΕΚΤΗΜΑΤΑ

Πλεονεκτήματα συστηματικής δειγματοληψίας

I. Η εφαρμογή της συστηματικής δειγματοληψίας είναι ευκολότερη για τους ερευνητές και επιφέρει μικρότερο αριθμό σφαλμάτων συλλογής.
II. Εφαρμόζοντας τη συστηματική δειγματοληψία, περιορίζεται η μεροληπτική επιλογή των δειγματικών μονάδων (ειδικότερα όταν η επιλογή του δείγματος γίνεται απευθείας από τον ερευνητή κατά τη διενέργεια της έρευνας).
III. Η συστηματική δειγματοληψία δίνει περισσότερες πληροφορίες σε δεδομένο κατά μονάδα κόστους εφόσον ένα συστηματικό δείγμα είναι διεσπαρμένο με μεγαλύτερη ομοιογένεια στον ερευνώμενο δείγμα.
IV. Με τη συστηματική δειγματοληψία ομαδοποιείται ο πληθυσμός αφού βάσει του καταλόγου επιλέγεται ένα στοιχείο από κάθε ομάδα για το δείγμα. Συνεπώς τα στοιχεία του πληθυσμού αντιπροσωπεύονται με την ίδια αναλογία στο δείγμα, με αποτέλεσμα το δείγμα να είναι περισσότερο αντιπροσωπευτικό.

Μειονεκτήματα συστηματικής δειγματοληψίας

I. Είναι αναγκαία η ύπαρξη και ο σχηματισμός καταλόγου με το σύνολο των στοιχείων Ν του πληθυσμού.
II. Αν το λ είναι ίσο με την περίοδο ή πολλαπλάσιο της τότε έχουμε περιοδικότητα στις τιμές των μονάδων του πληθυσμού όσον αφορά τη σειρά εμφάνισής τους στη λίστα με αποτέλεσμα να παρέχουν την ίδια πληροφορία.

3.4.2. ΕΚΤΙΜΗΣΗ ΤΟΥ ΜΕΣΟΥ ΚΑΙ ΤΟΥ ΣΥΝΟΛΙΚΟΥ ΠΛΗΘΥΣΜΟΥ

Για την εκτίμηση του μέσου συνολικού πληθυσμού $\bar{X}$ από ένα συστηματικό δείγμα χρησιμοποιούμε τον παρακάτω τύπο:

$$\bar{x}_{sy} = \frac{\sum_{j=1}^{n} x_{ij}}{n}$$

Η εκτίμηση του $\bar{x}_{sy}$ είναι αδύνατο να εκτιμηθεί από τα στοιχεία ενός μόνο συστηματικού δείγματος, εκτός αν ο πληθυσμός είναι τυχαίος, δηλαδή όταν οι μονάδες του είναι διατεταγμένες τυχαία. Όταν το N είναι μεγάλο η διακύμανση του $\bar{x}_{sy}$ είναι περίπου ίση με τη διακύμανση του $\bar{x}$ που βασίζεται στην απλή τυχαία δειγματοληψία, οπότε μπορούμε να χρησιμοποιήσουμε τον τύπο:

$$V(\bar{x}_{sy}) = \frac{s^2}{n}\left(\frac{N-n}{N}\right)$$

Η διακύμανση είναι:

$$s^2 = \frac{\sum_{j=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Η εκτίμηση του συνολικού πληθυσμού είναι:

$$\hat{X} = N\bar{x}_{sy}$$

ή

$$\hat{X} = \sum_{j=1}^{n} x_{ij}\frac{N}{n}$$

Όπου $\frac{N}{n}$ ο συντελεστής αναγωγής, δηλαδή το αντίστροφο του κλάσματος δειγματοληψίας.

Η διακύμανση του $\hat{X}$ είναι:

$$V(\hat{X}) = V(N\bar{x}_{sy}) = N^2 V(\bar{x}_{sy}) = N^2\frac{s^2}{n}\left(\frac{N-n}{N}\right)$$

και

$$SE(\hat{X}) = SE(N\bar{x}_{sy}) = \sqrt{V(N\bar{x}_{sy})}$$

Παράδειγμα

Θέλουμε να εκτιμήσουμε κατά μέσο όρο τον αριθμό των ατόμων που εισέρχονται σε ένα κατάστημα και το σύνολο των ατόμων που εισέρχονται κατά τη διάρκεια 2 μηνών (δηλαδή 60 ημερών). Λαμβάνοντας υπόψη το κόστος, αποφάσισε να καταμετρά τα άτομα που εισέρχονται στο κατάστημα κάθε 10 ημέρες. Ο αριθμός των ατόμων που εισέρχονται κάθε δέκατη μέρα, με τυχαίο ξεκίνημα την έκτη ημέρα, δίνεται από τον παρακάτω πίνακα.

Ημέρα	Αριθμός ατόμων που εισέρχονται στο κατάστημα
6	18
16	25
26	23
36	12
46	29
56	14

Να εκτιμηθούν: α) ο μέσος όρος των ατόμων που εισέρχονται ημερησίως στο κατάστημα, β) το σύνολο των ατόμων που εισέρχονται κατά τη διάρκεια 60ημερών και να υπολογιστούν το τυπικό σφάλμα και το σχετικό τυπικό σφάλμα.

Λύση

α) Η εκτίμηση του μέσου όρου των ατόμων που εισέρχονται ημερησίως στο κατάστημα είναι:

$$\bar{x}_{sy} = \frac{\sum_{j=1}^{n} x_{ij}}{n} = \frac{18+25+23+12+29+14}{6} = \frac{121}{6} = 20{,}16$$

Η διακύμανση του $\bar{x}_{sy}$ είναι:

$$s^2 = \frac{\sum_{j=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(18-20{,}16)^2 + (25-20{,}16)^2 + \cdots + (14-20{,}16)^2}{5} = 43{,}76$$

και

$$V(\bar{x}_{sy}) = \frac{s^2}{n}\left(\frac{N-n}{N}\right) = \frac{43{,}76}{6}\left(\frac{60-6}{60}\right) = 6{,}56$$

άρα το τυπικό σφάλμα της εκτίμησης είναι:

$$SE(\bar{x}_{sy}) = \sqrt{V(\bar{x}_{sy})} = \sqrt{6{,}56} = 2{,}56$$

και το σχετικό τυπικό σφάλμα είναι:

$$CV(\bar{x}_{sy}) = \frac{SE(\bar{x}_{sy})}{\bar{x}_{sy}} 100\% = \frac{2,56}{20,16} 100\% = 12,6$$

β) Η εκτίμηση του συνόλου των ατόμων Χ που εισέρχονται στο κατάστημα κατά τη διάρκεια 60 ημερών είναι:

$$\hat{X} = N\bar{x}_{sy} = 60 \cdot 20,16 = 1.209,6$$

Η διακύμανση της εκτίμησης $\hat{X}$ είναι:

$$V(N\bar{x}_{sy}) = N^2 \frac{s^2}{n}\left(\frac{N-n}{N}\right) = 60^2 \frac{43,76}{6}\left(\frac{60-6}{60}\right) = 23.630,4$$

Το τυπικό σφάλμα της εκτίμησης είναι:

$$SE(\hat{X}) = SE(N\bar{x}_{sy}) = \sqrt{V(N\bar{x}_{sy})} = \sqrt{23.630,4} = 153,72$$

και το σχετικό τυπικό σφάλμα είναι:

$$CV(\hat{X}) = \frac{SE(N\bar{x}_{sy})}{\hat{X}} 100\% = \frac{153,72}{1.209,6} 100\% = 12.7\%$$

3.4.3. ΕΚΤΙΜΗΣΗ ΠΟΣΟΣΤΟΥ ΕΝΟΣ ΠΛΗΘΥΣΜΟΥ

Στην περίπτωση τυχαίου πληθυσμού, η εκτίμηση του ποσοστού του πληθυσμού είναι αντίστοιχη με εκείνη της απλής τυχαίας δειγματοληψίας δηλαδή έχουμε:

$$p = \frac{\sum_{i=1}^{n} x_i}{n}$$

Όταν η διακύμανση του πληθυσμού σ^2 =PQ είναι άγνωστη, μια αμερόληπτη εκτίμησης της διακύμανση του ποσοστού p είναι:

$$V(p) = \frac{s^2}{n}\left(\frac{N-n}{N}\right) = \frac{pq}{n-1}\left(\frac{N-n}{N}\right)$$

Παράδειγμα

Με τη μέθοδο της συστηματικής δειγματοληψίας επιλέξαμε 400 εργαζόμενους από 3.000 που εργάζονται σε μια επιχείρηση. Από την δειγματοληπτική έρευνα προέκυψε ότι 150 εργαζόμενοι είναι ικανοποιημένοι από τις συνθήκες εργασίας. Να εκτιμηθεί το ποσοστό των εργαζομένων οι οποίοι είναι ικανοποιημένοι και το τυπικό σφάλμα.

Λύση

Το ποσοστό των εργαζομένων που είναι ικανοποιημένοι είναι:

$$p = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{150}{400} = 0{,}375 \text{ ή } 37{,}5\%$$

Η εκτίμηση της διακύμανσης του p είναι:

$$V(p) = \frac{pq}{n-1}\left(\frac{N-n}{N}\right) = \frac{0{,}375 \cdot 0{,}625}{399}\left(\frac{3.000-400}{3.000}\right) = 0{,}0005$$

Το τυπικό σφάλμα της εκτίμησης είναι:

$$SE(p) = \sqrt{V(p)} = \sqrt{0.0005} = 0.022$$

3.4.4. ΕΠΑΝΑΛΑΜΒΑΝΟΜΕΝΗ ΣΥΣΤΗΜΑΤΙΚΗ ΔΕΙΓΜΑΤΟΛΗΨΙΑ

Μια εναλλακτική μέθοδος που μπορούμε να χρησιμοποιήσουμε με σκοπό την εξασφάλιση μιας ικανοποιητικής διακύμανσης $V(\bar{x}_{gy})$ είναι η επαναλαμβανόμενη συστηματική δειγματοληψία (repeated systematic sampling). Η συγκεκριμένη μέθοδος προϋποθέτει τη συλλογή περισσοτέρων του ενός συστηματικών δειγμάτων. Πιο αναλυτικά, στη θέση ενός δείγματος μεγέθους n με διάστημα δειγματοληψίας $\lambda = \frac{N}{n}$ επιλέγουμε n_r συστηματικά δείγματα μεγέθους $\frac{n}{n_r}$ με διάστημα δειγματοληψίας $\lambda' = \lambda n_r$. Για κάθε ένα από αυτά τα συστηματικά δείγματα το τυχαίο ξεκίνημα επιλέγεται με τυχαίο τρόπο μεταξύ των πρώτων λ' αριθμών (1-λ'). Για παράδειγμα στην περίπτωση που έχουμε έναν πληθυσμό μεγέθους 720 και από αυτό επιλέξουμε ένα συστηματικό δείγμα μεγέθους n =90, με διάστημα δειγματοληψίας λ= $\frac{720}{90} = 8$ και με τυχαίο ξεκίνημα έναν αριθμό ανάμεσα στο 1 και στο 8 τότε θα επιλεγούν 90 μονάδες. Εάν όμως δεν θέλουμε να επιλέξουμε ένα συστηματικό δείγμα για την επιλογή των 90 μονάδων παίρνουμε 10 (δηλαδή n_r =10) με διάστημα δειγματοληψίας $\lambda' = \lambda n_r = 8 \cdot 10 = 80$. Έπειτα επιλέγουμε 10 τυχαίους αριθμούς μεταξύ του 1 και του 80. Έστω ότι οι αριθμοί αυτοί είναι: 9,27,38,46,53,59,64,68,72,79 οι οποίοι αποτελούν το τυχαίο ξεκίνημα για τα 10 συστηματικά δείγματα, όπως φαίνεται στον ακόλουθο πίνακα. Ο υπολογισμός του δεύτερου αριθμού σε κάθε βήμα μπορεί να βρεθεί αν προσθέσουμε στο τυχαίο ξεκίνημα τον αριθμό 80 ενώ ανάλογα γίνεται για τις υπόλοιπες περιπτώσεις.

Τυχαίο ξεκίνημα	Δεύτερος αριθμός δείγματος	Τρίτος αριθμός δείγματος	……..	Ένατος αριθμός δείγματος
9	89	169		649
27	107	187		667
38	118	198	…	678
46	126	206		686
53	133	213		693

59	139	219		699
64	144	224		704
68	146	228		707
72	152	232		712
79	159	239		719

Η εκτίμηση του μέσου πληθυσμού $\bar{X}$ χρησιμοποιώντας n_r συστηματικά δείγματα δίνεται από την εξής σχέση:

$$\bar{x}_r = \frac{1}{n_r}\sum_{i=1}^{n_r} \bar{x}_i$$

Όπου $\bar{x}_i$ ο μέσος του i συστηματικού δείγματος (i=1,2,…,n_r). Ο δείκτης r σημαίνει ότι χρησιμοποιείται επαναλαμβανόμενη συστηματική δειγματοληψία.

Η εκτίμηση της διακύμανσης του μέσου $\bar{x}_r$ δίνεται από τη παρακάτω σχέση:

$$V(\bar{x}_r) = \frac{N-n}{N}\frac{\sum_{i=1}^{n_r}(\bar{x}_i - \bar{x}_r)^2}{n_r(n_r-1)}$$

Το τυπικό σφάλμα είναι:

$$SE(\bar{x}_r) = \sqrt{V(\bar{x}_r)}$$

Η εκτίμηση του συνολικού πληθυσμού είναι:

$$\hat{X} = N\bar{x}_r$$

Η διακύμανση του $\hat{X}$ είναι:

$$V(\hat{X}) = N^2 V(\bar{x}_r) = N(N-n)\frac{\sum_{i=1}^{n_r}(\bar{x}_i - \bar{x}_r)^2}{n_r(n_r-1)}$$

Το τυπικό σφάλμα είναι:

$$SE(\hat{X}) = \sqrt{V(\hat{X})}$$

Παράδειγμα

Με την μέθοδο της επαναλαμβανόμενης συστηματικής δειγματοληψίας έγινε επιλογή 40 ομάδων αθλητών από 800 ομάδες αθλητών που διαμένουν στο ολυμπιακό χωριό. Για την επιλογή του δείγματος χρησιμοποιήθηκαν 10 συστηματικά δείγματα των 4 ομάδων το κάθε ένα. Με διάστημα δειγματοληψίας $\lambda' = \lambda n_r = 20 \cdot 10 = 200$ (όπου $\lambda = \frac{N}{n} = \frac{800}{40} = 20$), επιλέξαμε 10 τυχαίους αριθμούς μεταξύ 1 και 200: 25,67,98,110,124,138, 159,168,184,199 οι οποίοι αποτελούν το τυχαίο ξεκίνημα για τα 10 συστηματικά δείγματα. Από την διενέργεια της δειγματοληπτικής έρευνας προέκυψε ο αριθμός αθλητών κατά ομάδα και κατά συστηματικό δείγμα ως εξής:

Τυχαίο ξεκίνημα	Αριθμός αθλητών ομάδων	Δεύτερος αριθμός δείγματος	Αριθμός αθλητών ομάδων	Τρίτος αριθμός δείγματος	Αριθμός αθλητών ομάδων	Τέταρτος αριθμός δείγματος	Αριθμός αθλητών ομάδων
25	2	225	6	425	5	625	5
67	4	267	3	467	2	667	2
98	3	298	5	498	6	698	1
110	5	310	2	510	4	710	3
124	1	324	1	524	7	724	4
138	5	338	7	538	1	738	3
159	4	359	3	559	3	759	2
168	2	368	2	568	2	768	1
184	1	384	1	584	2	784	6
199	3	399	5	599	4	799	2

Να εκτιμηθούν

Α) Το μέσο μέγεθος των ομάδων, το τυπικό σφάλμα και το σχετικό τυπικό σφάλμα.

Β) Ο συνολικός πληθυσμός των αθλητών, το τυπικό σφάλμα και το σχετικό τυπικό σφάλμα.

Λύση

Α) Η εκτίμηση του μέσου μεγέθους των ομάδων είναι:

$$\bar{X}_r = \frac{1}{n_r}\sum_{i=1}^{n_r} \bar{X}_i = \frac{1}{10}\left(\frac{2+6+5+5}{4} + \frac{4+3+2+2}{4} + \cdots + \frac{3+5+4+2}{4}\right) = 3{,}25$$

Η διακύμανση του μέσου $\bar{x}_r$ είναι:

$$V(\bar{x}_r) = \frac{N-n}{N}\frac{\sum_{i=1}^{n_r}(\bar{x}_i-\bar{x}_r)^2}{n_r(n_r-1)} = \left(\frac{800-40}{800}\right)\frac{(4,5-3,25)^2+(2,75-3,25)^2+....(3,5-3,25)^2}{10(10-1)} = 0,059$$

Επομένως το τυπικό σφάλμα είναι:

$$SE(\bar{x}_r) = \sqrt{V(\bar{x}_r)} = \sqrt{0,059} = 0,242$$

Το σχετικό τυπικό σφάλμα:

$$CV(\bar{x}_r) = \frac{SE(\bar{x}_r)}{\bar{x}_r}\ 100\% = \frac{0,059}{3,25}\ 100\% = 1,81\%$$

Β) Η εκτίμηση του συνολικού πληθυσμού των αθλητών είναι:

$$\hat{X} = N\bar{x}_r = 800 \cdot 3,25 = 2600$$

Η διακύμανση του $\hat{X}$ είναι:

$$V(\hat{X}) = N^2 V(\bar{x}_r) = 800^2 \cdot 0,059 = 37.760$$

Επομένως το τυπικό σφάλμα του $\hat{X}$ είναι:

$$SE(\hat{X}) = \sqrt{V(\hat{X})} = \sqrt{37.760} = 194,31$$

Το σχετικό τυπικό σφάλμα είναι:

$$CV(\hat{X}) = \frac{SE(\hat{X})}{\hat{X}}\ 100\% = \frac{194,31}{2600}\ 100\% = 7,35\%$$

3.4.5. ΣΥΓΚΡΙΣΗ ΣΥΣΤΗΜΑΤΙΚΗΣ ΚΑΙ ΣΤΡΩΜΑΤΟΠΟΙΗΜΕΝΗΣ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ

Η ειδοποιός διαφορά μεταξύ αυτών των δυο ειδών δειγματοληψίας βρίσκεται στο γεγονός ότι στη συστηματική δειγματοληψία οι μονάδες έχουν την ίδια σχετική θέση στο στρώμα, ενώ στη στρωματοποιημένη η θέση των μονάδων καθορίζεται τυχαία. Επομένως το συστηματικό δείγμα είναι πιο ομοιόμορφα κατανεμημένο στον πληθυσμό με αποτέλεσμα την παροχή ακριβέστερων εκτιμήσεων από ένα στρωματοποιημένο τυχαίο δείγμα.

3.5. ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΚΑΤΑ ΟΜΑΔΕΣ

Η δειγματοληψία κατά ομάδες χρησιμοποιείται όταν δεν διαθέτουμε δειγματοληπτικό πλαίσιο με τις αρχικές μονάδες του πληθυσμού που συγκεντρώνουν το ενδιαφέρον της ερευνάς μας. Για παράδειγμα στην περίπτωση που επιθυμούμε να πάρουμε δείγμα από τις εύπορες οικογένειες του πληθυσμού της Ελλάδας, δεν μπορούμε να σχηματίσουμε από τέτοιου είδους οικογένειες αφού δεν υπάρχει κατάλληλος κατάλογος όλων των ευπόρων οικογενειών της Ελλάδος. Επίσης η δειγματοληψία κατά ομάδες έχει τη δυνατότητα να εφαρμοστεί όταν δεν υπάρχουν τα απαιτούμενα δειγματοληπτικά πλαίσια, ενώ αντίθετα υπάρχουν πλαίσια για ομάδες στοιχείων. Παρότι θα μπορούσαμε να θεωρήσουμε κάθε ομάδα ως ένα στρώμα η δειγματοληψία κατά ομάδες διαφέρει από την δειγματοληψία κατά στρώματα. Ενώ στη στρώματοποιημένη δειγματοληψία επιλέγουμε τυχαίο δείγμα από κάθε στρώμα, στη δειγματοληψία κατά ομάδες παίρνουμε τυχαία ορισμένες ομάδες και στη συνέχεια όλες τις μονάδες που έχει κάθε ομάδα του δείγματος αυτού. Άρα, στη δειγματοληψία κατά ομάδες είναι αναγκαία η δημιουργία ομάδων με τέτοιο τρόπο ώστε μέσα σε κάθε μια να υπάρχει τόση ανομοιογένεια όση σε ολόκληρο τον πληθυσμό (διότι τότε κάθε ομάδα θα αντιπροσωπεύει πιστά ολόκληρο τον πληθυσμό).

3.5.1. ΠΛΕΟΝΕΚΤΗΜΑΤΑ - ΜΕΙΟΝΕΚΤΗΜΑΤΑ

Πλεονεκτήματα δειγματοληψίας κατά ομάδες

Η δειγματοληψία κατά ομάδες είναι μια αποτελεσματική μέθοδος η οποία εφαρμόζεται συνήθως όταν το πλήθος των στοιχείων του πληθυσμού είναι γεωγραφικά εντοπισμένο, αλλά άγνωστο. Στην περίπτωση αυτή καμία άλλη δειγματοληψία από αυτές που προαναφέραμε δεν είναι δυνατόν να χρησιμοποιηθούν. Μπορεί επίσης να εφαρμοστεί όταν δεν υπάρχουν απαιτούμενα δειγματοληπτικά πλαίσια για όλα τα στοιχεία του πληθυσμού, ενώ αντίθετα υπάρχουν πλαίσια για ομάδες στοιχείων. Η έρευνα διευκολύνεται με τη συγκέντρωση του δείγματος σε ομάδες και παράλληλα έχει χαμηλότερο κόστος σε σχέση με τις άλλες μεθόδους. Επιπλέον ο χρόνος διεξαγωγής της έρευνας περιορίζεται σημαντικά.

Μειονεκτήματα δειγματοληψίας κατά ομάδες

Η ομαδοποίηση του πληθυσμού έχει σαν αποτέλεσμα οι μονάδες κάθε ομάδας διασποράς των εκτιμήσεων να δίνει λιγότερο αντιπροσωπευτικό δείγμα του πληθυσμού από όσο θα ήταν αν εφαρμόζαμε κάποια άλλη μέθοδο, ερευνώντας το ίδιο μέθοδος του δείγματος. Συνεπώς ομαδοποιώντας το δείγμα διαπιστώνουμε μείωση της αποτελεσματικότητας.

3.5.2. ΕΚΤΙΜΗΣΗ ΤΟΥ ΜΕΣΟΥ ΚΑΙ ΣΥΝΟΛΙΚΟΥ ΠΛΗΘΥΣΜΟΥ

Η δειγματοληψία κατά ομάδες είναι μια απλή τυχαία δειγματοληψία στην οποία κάθε δειγματοληπτική μονάδα αποτελείται από μια ομάδα στοιχείων του πληθυσμού, συνεπώς η εκτίμηση του μέσου πληθυσμού και του συνολικού πληθυσμού γίνεται με όμοιο τρόπο όπως στην απλή τυχαία δειγματοληψία. Η αμερόληπτη εκτίμηση του μέσου $\bar{X}$ του πληθυσμού συμβολίζεται με $\bar{x}_c$, όπου ο δείκτης c σημαίνει ότι έχουμε δειγματοληψία κατά ομάδες (cluster sampling) δηλαδή:

$$\bar{x}_c = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} m_i},$$

όπου x_i : συνολική τιμή της μεταβλητής x στην i ομάδα και m_i : ο αριθμός των στοιχείων στην ομάδα i.

Η διακύμανση της εκτίμησης του $\bar{x}_c$ είναι:

$$V(\bar{x}_c) = \frac{N-n}{Nn\bar{M}^2} S^2$$

Όπου $\bar{M} = \frac{M}{N}$: μέσο μέγεθος ομάδων στον πληθυσμό, $S^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{X}m_i)^2}{N-1}$, ενώ εάν το S^2 είναι άγνωστο χρησιμοποιούμε την παρακάτω εκτιμήτρια: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_c m_i)^2}{n-1}$

Το τυπικό σφάλμα είναι:

$$SE(\bar{x}_c) = \sqrt{V(\bar{x}_c)}$$

Η εκτίμηση του συνολικού πληθυσμού είναι:

$$\hat{X} = M\bar{x}_c$$

ή

$$\hat{X} = \sum_{i=1}^{n} x_i \frac{\sum_{i=1}^{N} m_i}{\sum_{i=1}^{n} m_i}$$

Η διακύμανση της εκτίμησης του συνολικού πληθυσμού X δίνεται από τη σχέση:

$$V(\hat{X}) = V(M\bar{x}_c) = M^2 V(\bar{x}_c) = \frac{N(N-n)}{n} S^2,$$

όπου M ο αριθμός των στοιχείων στον πληθυσμό.

Το τυπικό σφάλμα:

$$SE(\hat{X}) = SE(M\bar{x}_c) = \sqrt{V(M\bar{x}_c)}$$

Στην περίπτωση που ο αριθμός των στοιχείων M είναι άγνωστος η εκτίμηση του συνολικού πληθυσμού δίνεται από τη σχέση:

$$\hat{X} = N\bar{x}_t$$

Η διακύμανση της εκτίμησης του συνολικού πληθυσμού είναι:

$$V(N\bar{x}_t) = N^2 V(\bar{x}_t) = N^2 \left(\frac{N-n}{Nn}\right) \frac{\sum_{i=1}^{n}(x_i - \bar{x}_t)^2}{n-1}$$

Το τυπικό σφάλμα είναι:

$$SE(N\bar{x}_t) = \sqrt{V(N\bar{x}_t)}$$

Παράδειγμα

Θέλουμε να εκτιμήσουμε το μηνιαίο εισόδημα των νοικοκυριών μιας πόλης, όπου ο αριθμός των νοικοκυριών είναι 4.200. Επειδή δεν υπάρχουν διαθέσιμοι κατάλογοι νοικοκυριών θα εφαρμόσουμε δειγματοληψία κατά ομάδες. Συνεπώς κάθε οικοδομικό τετράγωνο του διαγράμματος της πόλης αποτελεί μια ομάδα νοικοκυριών. Συνολικά υπάρχουν 250 οικοδομικά τετράγωνα. Συνολικά επιλέχθηκαν 8 νοικοκυριά με τυχαίο τρόπο. Ακολουθεί πίνακας με τα αποτελέσματα που προέκυψαν από την έρευνα.

Ομάδα I (οικοδομικό τετράγωνο)	***Αριθμός νοικοκυριών*** m_i	***Συνολικό εισόδημα κατά ομάδα*** x_i ***(σε €)***
1	14	1000
2	18	1700
3	23	2100
4	11	2450
5	27	2750
6	12	1800
7	19	2600
8	22	2200
Σύνολο	146	16.600

Να εκτιμηθούν το μέσο μηνιαίο εισόδημα των νοικοκυριών της πόλης και το συνολικό εισόδημα όλων των νοικοκυριών της πόλης και να υπολογιστούν το τυπικό σφάλμα, το σχετικό τυπικό σφάλμα.

Λύση

Η εκτίμηση του μέσου μηνιαίου εισοδήματος των νοικοκυριών της πόλης είναι:

$$\bar{x}_c = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} m_i} = \frac{16.600}{146} = 113{,}69$$

Το M βρίσκεται από τον παρακάτω τύπο:

$$\bar{M} = \frac{M}{N} = \frac{4200}{250} = 16{,}8$$

Φτιάχνουμε τον παρακάτω πίνακα υπολογισμών:

Ομάδα i	x_i	$\bar{x}_c m_i$	$(x_i - \bar{x}_c m_i)^2$
1	1.000	1.592	350.464
2	1.700	2.046	119.716
3	2.100	2.615	265.225
4	2.450	1.251	1.437.601
5	2.750	3.070	102.400
6	1.800	1.364	190.096
7	2.600	2.160	193.600
8	2.200	2.501	90.601
Σύνολο	16.600		2.749.703

Η διακύμανση του $\bar{x}_c$ είναι:

$$V(\bar{x}_c) = \frac{N - n}{Nn\bar{M}^2} \frac{\sum_{i=1}^{n} (x_i - \bar{x}_c m_i)^2}{n - 1} = \frac{250 - 8}{250 \cdot 8 \cdot 16{,}8^2} \frac{2.749.703}{8 - 1} = 168{,}4$$

Το τυπικό σφάλμα είναι:

$$SE(\bar{x}_c) = \sqrt{V(\bar{x}_c)} = \sqrt{168{,}4} = 12{,}9$$

Το σχετικό τυπικό σφάλμα είναι:

$$CV(\bar{x}_c) = \frac{SE(\bar{x}_c)}{\bar{x}_c} = \frac{12.9}{113.69} = 0{,}113 \text{ ή } 11{,}3\%$$

Η εκτίμηση του συνολικού εισοδήματος όλων των νοικοκυριών της πόλης Χ είναι:

$$\hat{X} = \sum_{i=1}^{n} x_i \frac{\sum_{i=1}^{N} m_i}{\sum_{i=1}^{n} m_i} = 16.600 \cdot \frac{4200}{146} = 477.534$$

η διακύμανση της εκτίμησης $\hat{X}$ είναι:

$$V(\hat{X}) = V(M\bar{x}_c) = \frac{N(N-n)}{n} \frac{\sum_{i=1}^{n}(x_i - \bar{x}_c m_i)^2}{n-1} = \frac{250(250-8)}{8} \frac{2.749.703}{8-1} = 2{,}97\delta\iota\sigma$$

Το τυπικό σφάλμα είναι:

$$SE(\hat{X}) = SE(M\bar{x}_c) = \sqrt{V(M\bar{x}_c)} = \sqrt{2{,}97\delta\iota\sigma} = 54.504$$

Και το σχετικό τυπικό σφάλμα είναι:

$$CV(\hat{X}) = \frac{SE(\hat{X})}{\hat{X}} = \frac{54.504}{477.534} = 0{,}114 \text{ ή } 11{,}4\%$$

3.5.3. ΕΚΤΙΜΗΣΗ ΠΟΣΟΣΤΟΥ ΕΝΟΣ ΠΛΗΘΥΣΜΟΥ

Όπως και στην απλή τυχαία δειγματοληψία έτσι και στη μέθοδο της δειγματοληψίας κατά ομάδες η εκτίμηση ποσοστού ενός πληθυσμού γίνεται από το ποσοστό των στοιχείων του δείγματος που ανήκουν σε μια συγκεκριμένη κατηγορία την οποία εξετάζουμε. Επομένως έχουμε:

$$p = \frac{\sum_{i=1}^{n} \alpha_i}{\sum_{i=1}^{n} m_i}$$

Η διακύμανση της εκτίμησης του p είναι:

$$V(p) = \frac{(N-n)}{Nn\bar{M}^2} s^2$$

όπου

$$s^2 = \frac{\sum_{i=1}^{n}(\alpha_i - pm_i)^2}{n-1}$$

Το τυπικό σφάλμα είναι:

$$SE(p) = \sqrt{V(p)}$$

Παράδειγμα

Έστω ότι θέλουμε να εκτιμήσουμε το ποσοστό των καταναλωτικών δανείων που ανήκουν στην Εμπορική Τράπεζα. Από προηγούμενη έρευνα γνωρίζουμε ότι στον νομό Λακωνίας δόθηκαν 9.000 δάνεια και ότι ο αριθμός των καταστημάτων ανέρχεται σε 100. Από το σύνολο των 100 τραπεζών επιλέγησαν οι 4. Από την έρευνα προέκυψαν τα παρακάτω αποτελέσματα.

Ομάδα i (αριθμός καταστημάτων)	Αριθμός δανείων m_i	Αριθμός καταναλωτικών δανείων α_i
1	11	6
2	15	9
3	21	8
4	17	11
Σύνολο	64	34

Να εκτιμηθεί το ποσοστό των καταναλωτικών δανείων και να υπολογιστεί το τυπικό σφάλμα.

Λύση

$$p=\frac{\sum_{i=1}^{n}\alpha_i}{\sum_{i=1}^{n}m_i}=\frac{34}{64}=0{,}53 \text{ ή } 53\%$$

το M είναι γνωστό και είναι:

$$\bar{M}=\frac{M}{N}=\frac{9000}{100}=90$$

Φτιάχνουμε τον πίνακα υπολογισμών:

Ομάδα i	m_i	α_i	pm_i	$(\alpha_i - pm_i)^2$
1	11	6	5,83	0,03
2	15	9	7,95	1,10
3	21	8	11,13	9,80
4	17	11	9,01	3,96
Σύνολο	64			14,89

Η διακύμανση του p είναι:

$$V(p)=\frac{N-n}{NnM^2}\ \frac{\sum_{i=1}^{n}(\alpha_i-pm_i)^2}{n-1}=\frac{100-4}{100\cdot4\cdot90^2}\ \frac{14,89}{4-1}=0,000147$$

Το τυπικό σφάλμα είναι:

$$SE(p)=\sqrt{V(p)}=\sqrt{0,000147}=0,012$$

3.5.4. ΣΥΓΚΡΙΣΗ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ ΚΑΤΑ ΟΜΑΔΕΣ -ΣΤΡΩΜΑΤΟΠΟΙΗΜΕΝΗΣ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ

Αν και οι δύο αυτές μέθοδοι μοιάζουν πάρα πολύ ως προς τον διαμερισμό του πληθυσμού, διαφέρουν τόσο ως προς τη διαδικασία επιλογής δείγματος όσο και ως προς τους λόγους διαμερισμού του πληθυσμού σε στρώματα ή ομάδες. Στην δειγματοληψία κατά ομάδες επιλέγεται ένα δείγμα υποπληθυσμών (ομάδων) ενώ αντίθετα στην στρωματοποιημένη επιλέγεται ένα δείγμα του πληθυσμού από κάθε στρώμα. Ενώ ο στόχος της στρωματοποιημένης δειγματοληψίας είναι η βελτίωση της αποτελεσματικότητας, οι σπουδαιότεροι λόγοι της δειγματοληψίας κατά ομάδες είναι αποκλειστικά πρακτικοί. Πιο συγκεκριμένα, σκοπός μας είναι το χαμηλό κόστος της έρευνας και η δυνατότητα επιλογής δείγματος.

3.6. ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΠΟΣΟΣΤΩΝ - ΕΙΣΑΓΩΓΗ

Δειγματοληψία ποσοστών (quota sampling) ονομάζουμε το δειγματοληπτικό σχέδιο που μοιάζει με τη δειγματοληψία κατά στρώματα, αλλά η επιλογή των μονάδων μέσα σε κάθε στρώμα δεν γίνεται τυχαία. Πρόκειται για μια κατευθυνόμενη μέθοδο αφού η επιλογή του δείγματος γίνεται με υποκειμενικά κριτήρια. Δηλαδή ο ερευνητής είναι ελεύθερος να διενεργήσει την έρευνα κατά τον τρόπο που αυτός νομίζει ότι είναι ο καλύτερος , ώστε να προκύψουν αξιόπιστα συμπεράσματα. Βέβαια πάντα υπάρχει ο κίνδυνος το δείγμα να είναι μεροληπτικό και να δοθούν παραπλανητικά αποτελέσματα. Η επιδίωξη του ερευνητή είναι ο σχηματισμός στρωμάτων με τη μεγαλύτερη δυνατή εσωτερική ομοιογένεια και τη μεγαλύτερη δυνατή διαφορά μεταξύ τους. Επίσης η δειγματοληψία ποσοστών εφαρμόζεται κυρίως σε έρευνες αγοράς και γενικότερα έρευνες κοινής γνώμης.

Παράδειγμα

Έστω ότι θέλουμε να καθορίσουμε το δείγμα για την διενέργεια μιας έρευνας κοινής γνώμης σε μια πόλη. Η κατανομή των ερευνώμενων ατόμων κατά φύλο, ηλικία και απασχόληση δίνεται στον παρακάτω πίνακα, σύμφωνα με στοιχεία που προέκυψαν από την τελευταία απογραφή πληθυσμού.

Ηλικία	Εργαζόμενοι		Άνεργοι		Σύνολο
	Άνδρες	Γυναίκες	Άνδρες	Γυναίκες	
15-30	300	200	40	30	570
31-60	600	500	40	60	1200
61-	100	100	20	10	230
Σύνολο	1000	800	100	100	2000

Να καθοριστεί δείγμα 10% των ατόμων κατά φύλο, ηλικία, απασχόληση με τη μέθοδο δειγματοληψίας ποσοστών.

Λύση

Καταρτίζουμε κατανομή του δείγματος όπου ο αριθμός των ατόμων είναι ανάλογος με τα άτομα κατανομής του πληθυσμού, δηλαδή το 10% του πληθυσμού.

Ηλικία	Εργαζόμενοι		Άνεργοι		Σύνολο
	Άνδρες	Γυναίκες	Άνδρες	Γυναίκες	
15-30	30	20	4	3	57
31-60	60	50	4	6	120
61-	10	10	2	1	23
Σύνολο	100	80	10	10	200

3.6.1. ΠΛΕΟΝΕΚΤΗΜΑΤΑ – ΜΕΙΟΝΕΚΤΗΜΑΤΑ

Πλεονεκτήματα δειγματοληψίας ποσοστών

1. Η δειγματοληψία ποσοστών είναι σχετικά λιγότερο δαπανηρή από την τυχαία δειγματοληψία. Αυτό συμβαίνει επειδή το φαινόμενο της μη ανταπόκρισης δεν αποτελεί πρόβλημα αφού ο συνεντεύκτης έχει τη δυνατότητα να αντικαταστήσει αμέσως τον ερωτώμενο ο οποίος είτε δεν έχει χρόνο είτε αρνείται να συνεργαστεί. Επίσης, η

μετακίνηση των συνεντευκτών είναι αρκετά περιορισμένη διότι τείνουν να επιλέγουν τα πρώτα άτομα που συναντούν με τα επιθυμητά χαρακτηριστικά.
2. Απαιτείται σχετικά λίγος χρόνος για την ολοκλήρωση της έρευνας. Συνεπώς προσφέρεται ως δειγματοληπτικό σχέδιο για έρευνες που πρέπει να διεξαχθούν μέσα σε μικρό χρονικό διάστημα.
3. Είναι δυνατό να διεξαχθεί όταν δεν μπορεί να πραγματοποιηθεί τυχαία λήψη της μονάδος του δείγματος επειδή το αναγκαίο δειγματοληπτικό πλαίσιο είναι ανύπαρκτο.
4. Παρουσιάζονται λιγότερα διοικητικά προβλήματα, αυτό οφείλεται στην έλλειψη των προβλημάτων που προκύπτουν τόσο από τις αρνήσεις για συνεργασία όσο και από την επανάληψη των συνεντεύξεων.

Μειονεκτήματα δειγματοληψίας ποσοστών

1. Η έλλειψη τυχαιότητας κατά την επιλογή των μονάδων του δείγματος δεν επιτρέπει να υπολογιστούν τα τυπικά σφάλματα των εκτιμήσεων της δειγματοληψίας ποσοστών. Συγκεκριμένα το συνολικό δείγμα διαιρείται σε άλλα ανεξάρτητα δείγματα που αντλούνται από τον ίδιο πληθυσμό. Τα συγκεκριμένα δείγματα προσεγγίζονται ως δείγματα ποσοστών από τους συνεντεύκτες.
2. Η δειγματοληψία ποσοστών δεν είναι δυνατόν να απαλλαγεί από την μεροληψία επιλογής των μονάδων του δείγματος που ίσως δημιουργούν οι συνεντεύξεις. Η προσωπική επιλογή ενδέχεται να είναι μεροληπτική σε μεγάλο βαθμό.
3. Είναι θεωρητικά δύσκολο να ελεγχθεί ο συνεντεύκτης τη στιγμή που κάνει το έργο εντοπισμού των μονάδων του δείγματος και συλλογής των πληροφοριών.

3.7. ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΜΕ ΣΤΑΘΕΡΑ ΔΕΙΓΜΑΤΑ

Η δειγματοληψία με σταθερά δείγματα (panels) είναι μια μέθοδος η οποία εφαρμόζεται στις περιπτώσεις που συγκεντρώνονται πληροφορίες ανά τακτά χρονικά διαστήματα. Ουσιαστικά, εξετάζουμε ένα δείγμα κάθε φορά που θέλουμε να μελετήσουμε την κατάσταση του υπό έρευνα φαινομένου. Η επιλογή του δείγματος γίνεται με τυχαίο τρόπο, με οποιαδήποτε μέθοδο (όπως η απλή τυχαία δειγματοληψία, στρωματοποιημένη).

Η δειγματοληψία με σταθερά δείγματα είναι πολύ χρήσιμη, ιδιαίτερα αν στόχος μας είναι να επισημάνουμε μεταβολές στη ζήτηση ή τις προτιμήσεις ορισμένων εμπορευμάτων. Επιπλέον είναι χρήσιμη για την παρακολούθηση της ακροαματικότητας ή θεαματικότητας των μέσων μαζικής ενημέρωσης καθώς και για τη μελέτη της διαχρονικής πορείας συγκεκριμένων φαινομένων (πχ ανεργία). Χρησιμοποιείται ακόμα σε έρευνες αγοράς όταν θέλουμε να μελετήσουμε την εξέλιξη της οικονομικής κατάστασης των επιχειρήσεων. Συχνά, η μέθοδος χρησιμοποίησης σταθερών δειγμάτων εφαρμόζεται για να διαπιστωθεί αν ορισμένες μεταβλητές παρουσιάζουν εξάρτηση στο χρόνο.

3.7.1. ΠΛΕΟΝΕΚΤΗΜΑΤΑ – ΜΕΙΟΝΕΚΤΗΜΑΤΑ

Πλεονεκτήματα δειγματοληψίας με σταθερά δείγματα

Α)Απαιτείται μικρότερος χρόνος για το σχεδιασμό μιας έρευνας αφού χρησιμοποιείται το ίδιο δείγμα.

Β)Περιορίζονται τα μη δειγματοληπτικά σφάλματα, αφού οι ερευνητές με την επίσκεψη στις ίδιες μονάδες δύο ή περισσότερες φορές αποκτούν εμπειρία.

Γ)Έρευνες με αντικείμενο την επίπτωση των ασθενειών της παιδικής ηλικίας στη ζωή των ενηλίκων διεξάγονται καλύτερα, αφού παρακολουθούμε ένα σταθερό δείγμα προσώπων.

Δ)Υπάρχει δυνατότητα να εξεταστεί ο ερευνώμενος πληθυσμός σε επιθυμητά χρονικά διαστήματα.

Ε)Η εκτίμηση των διαφορών ενός φαινομένου που γίνεται με σταθερά βήματα έχει σχετικά μεγαλύτερη στατιστική ακρίβεια.

Μειονεκτήματα δειγματοληψίας με σταθερά δείγματα

Α)Μείωση του ποσοστού ανταπόκρισης, η οποία οφείλεται στο ότι ο ερευνώμενος πληθυσμός κουράζεται και δεν είναι πρόθυμος να συνεργαστεί με τον ερευνητή.

Β)Μεγάλος αριθμός ατόμων που δέχονται να συνεργαστούν σε σταθερό δείγμα παύουν να είναι ενεργά μέλη του δείγματος αυτού.

Γ)Ένα σταθερό δείγμα ενδέχεται να χάσει την αντιπροσωπευτικότητα του από τη συχνή παροχή των πληροφοριών που του ζητούμε.

3.8. ΕΠΙΦΑΝΕΙΑΚΗ ΔΕΙΓΜΑΤΟΛΗΨΙΑ

Η επιφανειακή δειγματοληψία (area sampling) είναι η μέθοδος με την οποία επιλέγουμε ένα δείγμα με βάση γεωγραφικά κριτήρια. Για παράδειγμα από το χάρτη μια πόλης επιλέγεται με τυχαίο τρόπο ένας αριθμός οικοδομικών τετραγώνων, όπου μπορούν να ερευνηθούν τα καταστήματα ή οι επιχειρήσεις που βρίσκονται στις επιφάνειες αυτές.

Η επιφανειακή δειγματοληψία διεξάγεται συνήθως σε τρία ή τέσσερα στάδια (φάσεις). Για παράδειγμα, έστω ότι θέλουμε να σχηματίσουμε τυχαίο δείγμα οικογενειών μιας πόλης, με απώτερο σκοπό τον υπολογισμό των ατόμων του πληθυσμού τα οποία είναι άνεργα. Το δείγμα μας είναι δυνατό να σχηματισθεί με δειγματοληψία σε τρία στάδια. Στο πρώτο στάδιο διεξάγουμε επιφανειακή δειγματοληψία και επιλέγουμε τυχαίο δείγμα οικοδομικών τετραγώνων από το οικοδομικό σχέδιο αυτής της πόλης. Κατά το δεύτερο στάδιο σχηματίζουμε τυχαίο δείγμα οικοδομών, οι οποίες βρίσκονται σε όλα τα οικοδομικά τετράγωνα του δείγματος του πρώτου σταδίου. Τέλος, στο τρίτο στάδιο παίρνουμε τυχαίο δείγμα οικογενειών, οι οποίες ζουν σε όλες τις οικοδομές του δείγματος δευτέρου σταδίου.

Η επιλογή του δείγματος των οικοδομικών τετραγώνων στο παραπάνω παράδειγμα, συνήθως γίνεται με πιθανότητες ανάλογες προς το μέγεθος κάθε οικοδομικού τετραγώνου. Για να γίνει όμως αυτό, υπάρχουν κάποιες προϋποθέσεις όπως η δυνατότητα υπολογισμού του μεγέθους κάθε οικοδομικού τετραγώνου, δηλαδή του αριθμού των ατόμων που ζουν σε κάθε οικοδομικό τετράγωνο.

Η επιφανειακή δειγματοληψία μπορεί να είναι απλή τυχαία ή τυχαία κατά στρώματα. Παραδείγματος χάρη, τα μεγάλα και τα μικρά οικοδομικά τετράγωνα στο προηγούμενο παράδειγμα ομαδοποιούνται σε δύο αντίστοιχα στρώμα και η επιφανειακή δειγματοληψία οικοδομών γίνεται μέσα από τα δύο αυτά στρώματα.

Η χρησιμότητα της επιφανειακής δειγματοληψίας είναι φανερή στις περιπτώσεις έρευνας επιθεμάτων τα οποία αναφέρονται σε οικοδομές, οικόπεδα κλπ. Το ίδιο δειγματοληπτικό σχέδιο και στις περιπτώσεις έρευνας θεμάτων που αναφέρονται σε ανθρώπινους πληθυσμούς (άτομα, οικογένειες). Τέλος, η επιφανειακή δειγματοληψία είναι ιδιαίτερα χρήσιμη όταν δεν υπάρχουν διαθέσιμοι κατάλληλοι κατάλογοι (άτομα, οικογένειες, καταστήματα), για να χρησιμοποιηθούν ως δειγματοληπτικά πλαίσια για τη διενέργεια μια έρευνας.

3.9. ΔΙΣΤΑΔΙΑΚΗ ΔΕΙΓΜΑΤΟΛΗΨΙΑ

Η δισταδιακή δειγματοληψία (two-stage sampling) είναι η μέθοδος με την οποία επιλέγουμε το δείγμα σε δύο στάδια. Αρχικά, ο ερευνώμενος πληθυσμός χωρίζεται σε ομάδες από τις οποίες γίνεται η επιλογή ενός τυχαίου δείγματος και στη συνέχεια μέσα από κάθε ομάδα που έχει επιλεγεί στο δείγμα, επιλέγεται ένας αριθμός στοιχείων. Κατά το πρώτο στάδιο επιλέγεται με τυχαίο τρόπο ένα δείγμα μονάδων (πχ επιχειρήσεις) και οι ομάδες αυτές οι οποίες αποτελούν τις μονάδες δείγματος στο στάδιο αυτό ονομάζονται πρωτογενείς μονάδες. Στο δεύτερο στάδιο επιλέγουμε με τυχαίο τρόπο έναν αριθμό στοιχείων (πχ καταστήματα) μέσα από κάθε πρωτογενή μονάδα που έχει επιλεγεί στο δείγμα. Τα στοιχεία αυτά αποτελούν τις μονάδες δείγματος κατά το δεύτερο στάδιο και ονομάζονται δευτερογενείς μονάδες.

Όπως και στην περίπτωση της δειγματοληψίας κατά ομάδες, μέσω της συγκεκριμένης μεθόδου μειώνονται ταυτόχρονα ο χρόνος διεξαγωγής της έρευνα και το κόστος της. Επιπλέον, κατά την εφαρμογή της δισταδιακής δειγματοληψίας δεν θεωρούνται απαραίτητα δειγματοληπτικά πλαίσια για το σύνολο των μονάδων του ερευνώμενου πληθυσμού παρά μόνο για τις ομάδες που έχουν επιλεγεί στο δείγμα.

Επιπρόσθετα, βασική προϋπόθεση είναι ο κατάλληλος διαχωρισμός του δείγματος σε ομάδες. Ένα κύριο στοιχείο που πρέπει να ληφθεί υπόψη για το σωστό διαχωρισμό είναι η ομοιογένεια των ομάδων ως προς το ερευνώμενο χαρακτηριστικό. Γενικά, στις ομάδες μεγάλου μεγέθους όπου και απαιτείται μεγαλύτερο δείγμα, τα χαρακτηριστικά παρουσιάζουν μεγαλύτερη ανομοιογένεια σε σχέση με αυτές του μικρού μεγέθους οι οποίες είναι περισσότερο ομοιογενείς.

Σύμφωνα με τα παραπάνω αν θέλουμε να μελετήσουμε το εισόδημα των νοικοκυριών που διαμένουν σε μια πόλη και χωρίσουμε τα νοικοκυριά σε μεγάλες ομάδες (συνοικίες) ή σε μικρότερες ομάδες (οικοδομικά τετράγωνα), προκύπτει ότι τα νοικοκυριά που θα διαμένουν σε μια συνοικία θα παρουσιάζουν μεγαλύτερη ανομοιογένεια ως προς το εισόδημα σε σχέση με αυτά που διαμένουν στο ίδιο οικοδομικό τετράγωνο. Συνεπώς, ο διαχωρισμός των νοικοκυριών σε ομάδες κατά συνοικία απαιτεί μεγαλύτερο δείγμα από ότι σε ομάδες κατά οικοδομικό τετράγωνο.

3.10. ΤΡΙΣΤΑΔΙΑΚΗ ΔΕΙΓΜΑΤΟΛΗΨΙΑ

Στην περίπτωση που οι έρευνες είναι ευρεία κλίμακας, η εφαρμογή της δειγματοληψίας σε τρία ή περισσότερα στάδια είναι αναγκαία. Η συγκεκριμένη μέθοδος ονομάζεται

τρισταδιακή δειγματοληψία (three-stage sampling) ή πολυσταδιακή δειγματοληψία (multi-stage sampling).

Χωρίζουμε τον πληθυσμό που επιθυμούμε να μελετήσουμε σε συγκεκριμένο αριθμό ομάδων (clusters), από τις οποίες επιλέγουμε ορισμένες εφαρμόζοντας τυχαία δειγματοληψία. Αν τώρα από τις ομάδες του δείγματος αυτού επιλέξουμε όλες τις μονάδες που περιέχουν, τότε εφαρμόζουμε τη μέθοδο της δειγματοληψίας κατά ομάδες. Όμως, αν πάρουμε ορισμένες μόνο τότε διεξάγουμε δειγματοληψία σε τρία στάδια. Κατά το πρώτο στάδιο επιλέγουμε τυχαίο δείγμα μονάδων πρώτου σταδίου (ομάδες- πόλεις), στο δεύτερο στάδιο παίρνουμε δείγμα μονάδων δευτέρου σταδίου από το σύνολο των μονάδων που περιέχονται στις ομάδες πρώτου σταδίου που έχουν ήδη επιλεγεί (δείγμα οικοδομικών τετραγώνων από τις πόλεις του δείγματος). Κατά το τρίτο στάδιο επιλέγουμε δείγμα μονάδων τρίτου σταδίου από εκείνες που περιέχονται στις μονάδες δευτέρου σταδίου που έχουμε επιλέξει (δείγμα οικογενειών από τα οικοδομικά τετράγωνα του δευτέρου σταδίου που έχουν επιλεχθεί). Είναι δυνατό στο τρίτο και τελευταίο στάδιο του παραδείγματος μας να επιλέξουμε όλες τις οικογένειες που ζουν στα οικοδομικά τετράγωνα του δείγματος, όπως το διαμορφώσαμε στο δεύτερο στάδιο.

Η δειγματοληψία κάθε σταδίου διεξάγεται με τυχαίο τρόπο, δηλαδή εφαρμόζοντας τυχαία επιλογή.

3.11. ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΑΠΟ ΚΥΡΙΑ ΔΕΙΓΜΑΤΑ

Η δειγματοληψία από κύρια δείγματα χρησιμοποιείται όταν υποχρεούμαστε να αντλήσουμε διαδοχικά δείγματα από τον ίδιο πληθυσμό. Στην περίπτωση μεγάλου μεγέθους του πληθυσμού και παράλληλα εάν η προσέγγιση των μονάδων του δείγματος είναι δυσχερής, είναι απαραίτητο να σχηματιστεί με ιδιαίτερη προσοχή ένα κύριο δείγμα από το οποίο θα αντληθούν όλα τα αναγκαία διαδοχικά δείγματα. Τα διαδοχικά δείγματα προκύπτουν πάντα από το κύριο δείγμα και όχι από τον πληθυσμό, δηλαδή όλα τα ουσιαστικά χαρακτηριστικά του πληθυσμού πρέπει να αντιπροσωπεύονται αποτελεσματικά στο κύριο δείγμα. Επιπλέον το μέγεθος του κυρίου δείγματος πρέπει να είναι αρκετά μεγάλο έτσι ώστε να μπορεί να παράγει ολόκληρο πλήθος από διαφορετικά δείγματα.

Ας υποθέσουμε ότι παρακολουθούμε δειγματοληπτικά την κίνηση των εμπορικών καταστημάτων όλης της χώρας με τυχαία δειγματοληψία επιλέγουμε ορισμένες πόλεις της χώρας και από το σύνολο των καταστημάτων που λειτουργούν σε αυτές σχηματίζουμε το αντίστοιχο κύριο δείγμα καταστημάτων. Κάθε φορά που θα χρειαστεί να καταμετρήσουμε τον πληθυσμό των καταστημάτων θα σχηματίσουμε δείγμα από τα καταστήματα του κύριου δείγματος. Μπορούμε σε κάθε νομό του δείγματος να εγκαταστήσουμε από ένα ερευνητή έτσι ώστε να εξοικειωθεί με τις ειδικές συνθήκες των καταστημάτων του νομού του, έτσι η έρευνα θα ολοκληρωθεί σε σύντομο χρονικό διάστημα και θα έχουμε υψηλά επίπεδα αποτελεσματικότητας.

Τα κύρια δείγματα είναι χρήσιμα όταν αντιπροσωπεύουν σταθερά τον πληθυσμό. Συνεπώς δεν είναι σκόπιμο να σχηματίζονται κύρια δείγματα από έντονα μεταβαλλόμενο πληθυσμό. Παράλληλα δεν υπάρχουν και εντελώς αμετάβλητοι πληθυσμοί (απόλυτα σταθεροί). Επομένως η τροποποίηση των κύριων δειγμάτων είναι αναγκαία δηλαδή αυτά ξανασχηματίζονται από τον πληθυσμό στον οποίο αναφέρονται. Η δειγματοληψία από κύρια δείγματα δεν πρέπει να εφαρμόζεται σε πληθυσμούς που μεταβάλλονται με την πάροδο του χρόνου, διότι τα δείγματα δεν αντιπροσωπεύουν πλέον το κύριο δείγμα.

3.12. ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΜΕ ΥΠΕΡΤΙΘΕΜΕΝΑ ΔΕΙΓΜΑΤΑ

Ας υποθέσουμε ότι για τους σκοπούς της ερευνάς μας είναι απαραίτητη η χρήση δείγματος μεγέθους 100 μονάδων μπορούμε να σχηματίσουμε το συγκεκριμένο δείγμα με τον εξής τρόπο: αντλούμε δείγμα 10 δειγματοληπτικών μονάδων από τον πληθυσμό τον οποίο καλύπτει η ερευνά μας. Έπειτα παίρνουμε νέο δείγμα 10 μονάδων από τον ίδιο πληθυσμό με τον ίδιο τρόπο. Συνεχίζουμε με την ίδια μέθοδο αντλώντας ισοπληθή δείγματα σταματάμε όταν έχουμε καταρτίσει 10 δείγματα 10 μονάδων δηλαδή συνολικό δείγμα 100 μονάδων.

Βέβαια αντί να παίρνουμε ανεξάρτητα δείγματα των 10 μονάδων, μπορούμε να επιλέξουμε δείγματα των 20 μονάδων αλλά στη συγκεκριμένη περίπτωση θα απαιτηθούν 5 ισοπληθή δείγματα αυτού του μεγέθους ώστε το συνολικό δείγμα να είναι και πάλι 100 μονάδων το ίδιο αποτέλεσμα έχουμε με ισοπληθή δείγματα των 5 ή των 50 μονάδων.

Αυτά τα ανεξάρτητα δείγματα που σχηματίζονται από τον ίδιο πληθυσμό με την εφαρμογή του ίδιου βασικού δειγματοληπτικού σχεδίου (για παράδειγμα με απλή τυχαία δειγματοληψία ή με δειγματοληψία κατά ομάδες) ονομάζονται υπερτιθέμενα ή ομοιότυπα δείγματα. Η συγκεκριμένη μέθοδος δειγματοληψίας εφαρμόζεται στις εξής περιπτώσεις:

α) Ενδεχομένως, το ενιαίο συνολικό δείγμα απαιτεί σχετικά πολύ χρόνο ώστε να σχηματιστεί ενώ τα υπερτιθέμενα δείγματα ως μικρότερα σχηματίζονται σε πολύ μικρότερο χρόνο ή με σχετικά λιγότερες δυσχέρειες. Συνεπώς, πριν σχηματιστεί ολόκληρο το ενιαίο δείγμα χρησιμοποιούνται οι πληροφορίες έστω και ορισμένων από τα υπερτιθέμενα δείγματα για να συναχθούν συμπεράσματα για το εξεταζόμενο θέμα. Όμως, η στατιστική ακρίβεια των εκτιμήσεων θα είναι περιορισμένη σε σχέση με την ακρίβεια των εκτιμήσεων που προκύπτουν από το συνολικό δείγμα.

β) Αν τα υπερτιθέμενα δείγματα από διαφορετικού συνεντεύκτες (ένας για κάθε δείγμα), η συστηματική διαφορά που θα παρατηρηθεί μεταξύ των δειγμάτων ως προς οποιοδήποτε χαρακτηριστικό που καλύπτει η έρευνα, είναι δυνατό να χρησιμοποιηθεί ως έκφραση των διαφοροποιητικών επιδράσεων που οφείλονται στα μη δειγματοληπτικά σφάλματα, τα οποία με τη σειρά τους οφείλονται σε συγκεκριμένους συνενετεύκτες. Βασική προϋπόθεση είναι όλα τα επιμέρους δείγματα να σχηματίζονται από ολόκληρο τον πληθυσμό και όχι από διαφορετικό στρώμα το καθένα.

γ) Η δειγματοληψία με υπερτιθέμενα δείγματα χρησιμοποιείται για ευκολότερη εκτίμηση του τυπικού σφάλματος. Συγκεκριμένα, αν τα υπερτιθέμενα δείγματα είναι N σε αριθμό, και από τον καθένα τους προκύπτει η εκτίμηση X_i (όπου i = 1,2,…..,N), η γενική εκτίμηση της X θα είναι:

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_N}{N}$$

Η διακύμανση των εκτιμήσεων X δίνεται από τη σχέση:

$$S_x^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_N - \overline{X})^2}{N - 1}$$

Το τυπικό σφάλμα της μεταβλητής X υπολογίζεται ως εξής:

$$\sqrt{\frac{S_x^2}{N}} = \sqrt{\frac{\Sigma(X_i - \overline{X})^2/(N-1)}{N}}$$

Παράδειγμα

Έστω από ορισμένο πληθυσμό σχηματίζονται 5 ανεξάρτητα δείγματα και ότι η ιδιότητα Α βρίσκεται με τις εξής αναλογίες στα δείγματα αυτά: 29% (1^o), 25% (2^o), 34% (3^o), 25% (4^o) και 40%(5^o). Από τα στοιχεία αυτά υπολογίζουμε:

$$\overline{X} = (29 + 25 + 34 + 25 + 40)/5 = 30{,}6$$

$$S_x^2 = \frac{[(29-30{,}6)^2 + (25-30{,}6)^2 + (34-30{,}6)^2 + (25-30{,}6)^2 + (40-30{,}6)^2]}{5-1}$$

$$= \frac{165{,}2}{4} = 41{,}3$$

Άρα το τυπικός σφάλμα της Χ είναι:

$$\sqrt{S_x^2/N} = \sqrt{41{,}3/5} = \sqrt{8{,}26} = 2{,}87$$

ΚΕΦΑΛΑΙΟ ΤΕΤΑΡΤΟ: ΚΑΤΑΝΟΜΕΣ

4.1. ΔΙΑΚΡΙΤΕΣ ΚΑΤΑΝΟΜΕΣ

Οι μεταβλητές χωρίζονται σε ασυνεχείς (διακριτές) και συνεχείς. Για μια διακριτή μεταβλητή κατανομή πιθανοτήτων (probability distribution) ονομάζεται η καταγραφή όλων των δυνατών τιμών της μεταβλητής με τις αντίστοιχες πιθανότητες εμφάνισης τους. Για παράδειγμα η κατανομή πιθανότητας του αριθμού των παιδιών ανά οικογένεια, σύμφωνα με την τελευταία απογραφή είναι:

Αριθμός παιδιών ανά οικογένεια (X)	**Πιθανότητα (P)**
0	0,15
1	0,25
2	0,35

3	0,15
4	0,07
5	0,02
6	0,01
Σύνολο	1,00

Στο παραπάνω παράδειγμα το άθροισμα των πιθανοτήτων είναι 1, αφού οι τιμές της μεταβλητής Χ καλύπτουν ολόκληρο το δειγματικό χώρο.

Βασικοί παράμετροι για κάθε κατανομή πιθανοτήτων είναι ο μέσος όρος και η τυπική απόκλιση.

4.1.1. ΔΙΩΝΥΜΙΚΗ ΚΑΤΑΝΟΜΗ

Η διωνυμική κατανομή (binomial distribution) είναι μια ασυνεχής κατανομή με πολλές εφαρμογές στην καθημερινή μας ζωή. Από τις ασυνεχείς κατανομές θεωρείται η σπουδαιότερη γιατί τα περισσότερα φαινόμενα μπορούν να μελετηθούν με τη βοήθεια της. Την διωνυμική κατανομή ανακάλυψε ο Bernoulli το 1700. Παριστάνουμε με p την πιθανότητα να πραγματοποιηθεί σε μια δοκιμή Bernoulli ένα ενδεχόμενο Κ σε μια δοκιμή Bernoulli και με q = 1 – p την πιθανότητα να μην πραγματοποιηθεί το Κ. Το p καλείται πιθανότητα επιτυχίας ενώ το q καλείται πιθανότητα αποτυχίας. Αν σε καθεμία από τις επιλογές ενός πειράματος τα ποσοστά p και q είναι τα ίδια και αν σε κάθε δοκιμή έχουμε μόνο δύο δυνατά ενδεχόμενα (να πραγματοποιηθεί ή όχι το Κ) και οι δοκιμές είναι ανεξάρτητες μεταξύ τους, τότε το φαινόμενο που εξετάζουμε λέμε ότι ακολουθεί τη διωνυμική κατανομή. Αν επαναλάβουμε το πείραμα τύχης n φορές και ζητήσουμε την πιθανότητα να πραγματοποιηθεί το ενδεχόμενο Κ, ακριβώς x τότε μια ικανοποιητική λύση του προβλήματος είναι να πραγματοποιηθεί το ενδεχόμενο Κ στις x πρώτες δοκιμές και να μην πραγματοποιηθεί στις υπόλοιπες n-x δοκιμές. Σύμφωνα με το θεώρημα του Πολλαπλασιασμού των πιθανοτήτων, η πιθανότητα πραγματοποίησης του συγκεκριμένου ενδεχομένου είναι:

$$\underbrace{p \cdot p \cdot p \dots p}_{x \text{ φορές}} \cdot \underbrace{q \cdot q \cdot q \dots q}_{(n-x) \text{ φορές}} = p^x \cdot q^{n-x}$$

Το πλήθος των δυνατών θέσεων μπορεί να βρεθεί αν υπολογίσουμε κατά πόσους τρόπους από τις n δοκιμές μπορούμε να ξεχωρίσουμε x τέτοιες θέσεις. Επομένως η ζητούμενη πιθανότητα P_x δηλαδή το ενδεχόμενο Κ να συμβεί x φορές στις n δοκιμές, δίνεται από τον τύπο:

$$P(X = x) = P_x = \binom{n}{x} p^x \cdot q^{n-x} = \frac{n!}{x!\,(n-x)!} p^x \cdot q^{n-x}$$

Όπου η τυχαία μεταβλητή P παριστάνει το πλήθος των επιτυχιών σε n δοκιμές και x = 0,1,2,3,…..,n.

Παράδειγμα

Η πιθανότητα να κλαπεί ένα αυτοκίνητο το οποίο βρίσκεται σταθμευμένο στη διάρκεια της νύχτας σε μια κακόφημη συνοικία είναι 60%. Αν σε ένα δρόμο είναι σταθμευμένα 10 αυτοκίνητα, ποια είναι η πιθανότητα:

α) να μην κλαπεί κανένα

β) να κλαπούν το πολύ 3

γ) να κλαπούν τουλάχιστον 7

Λύση

Αν συμβολίσουμε με X τον αριθμό των αυτοκινήτων που κλέβονται (ανάμεσα στα 10 που βρίσκονται σταθμευμένα στο συγκεκριμένο δρόμο) θα έχουμε:

$$P(X = x) = \binom{10}{x} 0{,}6^x \cdot 0{,}4^{10-x}, \quad x = 0,1,\dots,10$$

Άρα:

α) $P(X = 0) = \binom{10}{0} 0{,}6^0 \cdot 0{,}4^{10-0} = 0{,}0001 = 0{,}01\%$

β) $P(X \leq 3)$ = P (X=0) + P (X=1) + P (X=2) + P (X=3) = 0,000+0,0002+0,011+0,042= 0,055 = 5,5 %

γ) P (X $\geq$ 7) = P (X=7) + P (X=8) + P (X=9) + P (X=10) = 0,215+0,121+0,040+0,006=0,382=38,2%

Η διωνυμική κατανομή εφαρμόζεται όταν το μέγεθος του δείγματος είναι μικρότερο του 50 (n < 50) και η πιθανότητα εμφάνισης του εξεταζόμενου ενδεχομένου είναι σταθερή και μεγαλύτερη του 10 % (p > 10).

4.1.1.1. ΒΑΣΙΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΔΙΩΝΥΜΙΚΗΣ ΚΑΤΑΝΟΜΗΣ

Η μορφή της διωνυμικής κατανομής εξαρτάται από τα p και q. Αν p = q =1/2 τότε η διωνυμική κατανομή είναι συμμετρική ανεξάρτητα από το n. Αν όμως p $\neq$ q , τότε η κατανομή είναι ασυμμετρική. Γενικά όσο αυξάνει το n, τόσο η κατανομή τείνει να γίνει συμμετρική.

Τα βασικά συστατικά μέτρα της διωνυμικής κατανομής είναι:

α) ο μέσος αριθμητικός: μ = n · p

β) η διακύμανση και η μέση απόκλιση τετραγώνου:

$$\sigma^2 = n \cdot p \cdot q \quad \text{και} \quad \sigma = \sqrt{n \cdot p \cdot q}$$

γ) ο συντελεστής ασυμμετρίας: $\beta_1 = \frac{(q-p)^2}{n \cdot p \cdot q}$ όσο αυξάνεται το n τόσο ο β_1 τείνει προς το μηδέν

δ) ο συντελεστής κύρτωσης: $\beta_2 = 3 + \frac{1-6p \cdot q}{n \cdot p \cdot q}$ όσο αυξάνεται το n τόσο ο β_2 τείνει προς το 3.

4.1.1.2. ΠΡΟΣΑΡΜΟΓΗ ΔΙΩΝΥΜΙΚΗΣ ΚΑΤΑΝΟΜΗΣ ΣΕ ΕΜΠΕΙΡΙΚΗ ΚΑΤΑΝΟΜΗ

Η διωνυμική κατανομή μπορεί να προσαρμοστεί σε μια εμπειρική δειγματοληπτική κατανομή συχνοτήτων με αντικειμενικό σκοπό την εκτίμηση των ιδιοτήτων του αγνώστου πληθυσμού από τον οποίο προέρχεται. Η προσαρμογή μιας διωνυμικής κατανομής σε μια εμπειρική κατανομή περιέχει τρείς φάσεις:

α) Προσδιορισμός της αριθμητικής τιμής της παραμέτρου p του αγνώστου πληθυσμού με τη βοήθεια ενός δείγματος

β) Υπολογισμός των θεωρητικών συχνοτήτων

γ) Έλεγχος της προσαρμογής της διωνυμικής κατανομής σε μια εμπειρική κατανομή.

Ο μέσος αριθμητικός της εμπειρικής κατανομής δίνεται από τον τύπο:

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i}$$

Παράδειγμα

Προϊόν συσκευάζεται σε κιβώτια των 6 τεμαχίων. Παίρνουμε στην τύχη 500 κιβώτια και διαπιστώνουμε τον αριθμό των ελαττωματικών σε κάθε κιβώτιο. Μας δίνεται ο παρακάτω πίνακας.

Αριθμός ελαττωματικών (x_i)	0	1	2	3	4	5	6

Αριθμός κιβωτίων (f_i)	65	145	160	90	25	10	5

Ζητείται:

Να υπολογιστούν τα βασικά συστατικά στοιχεία της διωνυμικής κατανομής.

Λύση

Για να βρούμε το μέσο αριθμητικό της εμπειρικής κατανομής πρέπει να βρεθεί το γινόμενο $x_i f_i$, οπότε φτιάχνουμε μια επιπλέον στήλη στον πίνακα και ο πίνακας μετατρέπεται ως εξής:

x_i	0	1	2	3	4	5	6	**Σύνολο**
f_i	65	145	160	90	25	10	5	500
$x_i f_i$	0	145	320	270	100	50	30	915

Συνεπώς έχουμε:

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{915}{500} = 1{,}8$$

Δηλαδή σε κάθε κουτί των 6 τεμαχίων, 1,8 τεμάχια κατά μέσο όρο είναι ελαττωματικά. Επομένως, από τη σχέση $\bar{x}$ (ή μ) = n · p έχουμε:

$$p = \frac{\bar{x}}{n} = \frac{1.8}{6} = 0.30$$

Άρα το μέσο ποσοστό των ελαττωματικών είναι 30%. Από τη σχέση p+q=1 έχουμε:

q = 1- p = 1- 0,30 = 0,70

Άρα το μέσο ποσοστό των μη ελαττωματικών προϊόντων είναι 70%.

Τα βασικά στατιστικά μέτρα της διωνυμικής κατανομής είναι:

i. Μέσος αριθμητικός: $\bar{x} = n \cdot p = 6 \cdot 0{,}3 = 1{,}8$
ii. Μέση απόκλιση τετραγώνου:
$\sigma = \sqrt{n \cdot p \cdot q} = \sqrt{6 \cdot 0{,}3 \cdot 0{,}7} = 1{,}12$
iii. Ασυμμετρία: $\beta_1 = \frac{(q-p)^2}{n \cdot p \cdot q} = \frac{(0{,}7-0{,}3)^2}{6 \cdot 0{,}3 \cdot 0{,}70} = 0{,}13$
iv. Κύρτωση: $\beta_2 = 3 + \frac{1-6p \cdot q}{n \cdot p \cdot q} = 3 + \frac{1-6 \cdot 0{,}3 \cdot 0{,}7}{6 \cdot 0{,}3 \cdot 0{,}7} = 3{,}21$

4.1.2. *ΚΑΤΑΝΟΜΗ BERNOULLI*

Οι δοκιμές Bernoulli των οποίων η ονομασία οφείλεται στον σουηδό μαθηματικό James Bernoulli (1654-1705) οδηγούν στον ορισμό της απλούστερης ίσως διακριτής τυχαίας μεταβλητής. Με τον συγκεκριμένο όρο αναφερόμαστε σε ένα πείραμα με δύο δυνατά αποτελέσματα. Το ένα αποτέλεσμα θα ονομάζεται επιτυχία (Ε) και το άλλο αποτυχία (Α).

Η πιο κλασσική περίπτωση μιας δοκιμής Bernoulli είναι το πείραμα της ρίψης ενός νομίσματος, αν αυτό που επιθυμούμε είναι η ένδειξη «κεφαλή» τότε ως επιτυχία Ε θεωρείται το αποτέλεσμα αυτό και ως αποτυχία Α η ένδειξη «γράμματα». Σύμφωνα με τα παραπάνω ο δειγματικός χώρος του πειράματος που παράγει μια δοκιμή Bernoulli είναι:

Ω = { E,A}

Αν συμβολίσουμε με p,q τις δυνατότητες εμφάνισης των ενδεχομένων {E,A} αντίστοιχα τότε θα έχουμε:

p + q = 1 $\quad 0 \leq p \leq 1 \quad 0 \leq q \leq 1$

Ορισμός: Έστω X ο αριθμός των επιτυχιών σε μια δοκιμή Bernoulli με πιθανότητα επιτυχίας p και αποτυχίας q. Η κατανομή της τυχαίας μεταβλητής X καλείται κατανομή Bernoulli με παράμετρο p.

4.1.3. ΚΑΤΑΝΟΜΗ POISSON

Η κατανομή Poisson είναι μια διακριτή κατανομή πιθανοτήτων με πολλές πρακτικές εφαρμογές. Συνήθως χρησιμοποιείται για να περιγράψει καταστάσεις “ουρών”, για παράδειγμα άφιξη πελατών σε τράπεζες. Αποτελεί την προέκταση της διωνυμικής κατανομής σε περιπτώσεις που το μέγεθος του δείγματος (παρατηρήσεις ή επαναλήψεις του πειράματος) είναι πολύ μεγάλο (n > 50), ενώ παράλληλα η τιμή του p (πιθανότητα να συμβεί ένα ενδεχόμενο είναι πολύ μικρή (p < 0,10).

Στις περιπτώσεις αυτές για τον υπολογισμό των διαφόρων πιθανοτήτων εφαρμόζεται ο τύπος που διατύπωσε το 1837 ο Poisson:

$$P(X = x) = P_x = \frac{e^{-\lambda} \cdot \lambda^x}{x!}$$ όπου x = 0,1,2,3,… , e = 2,71828

4.1.3.1. ΒΑΣΙΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ

Η κατανομή Poisson είναι συνεχής, δηλαδή η τυχαία μεταβλητή x παίρνει μόνο ακέραιες τιμές 0,1,2,3,… Τα βασικά στατιστικά μέτρα είναι:

i. Ο μέσος αριθμητικός: $\mu = \lambda = n \cdot p$
ii. Η διακύμανση και η μέση απόκλιση τετραγώνου αντίστοιχα: $\sigma^2 = \lambda$ και $\sigma = \sqrt{\lambda}$
iii. Συντελεστής ασυμμετρίας: $\beta_1 = \frac{1}{\lambda}$

Η κατανομή παρουσιάζει θετική ασυμμετρία και τείνει να γίνει συμμετρική ($\beta_1 = 0$) όταν αυξάνεται το λ.

iv. Συντελεστής κύρτωσης: $\beta_2 = 3 + \frac{1}{\lambda}$

Η κατανομή είναι λεπτόκυρτη αλλά όταν αυξάνεται το λ τότε το β_2 τείνει προς το 3 (μεσόκυρτη).

Στις πρακτικές εφαρμογές για τον υπολογισμό των διάφορων πιθανοτήτων χρησιμοποιούμε τον τύπο:

$$P_{x+1} = \frac{\lambda}{x+1} \cdot P_x \qquad x = 0,1,2,3,$$

Η κατανομή Poisson σε περιπτώσεις σπάνιων γεγονότων. Ενδεικτικά αναφέρουμε τις ακόλουθες περιπτώσεις:

α) Η κατανομή του αριθμού των τηλεφωνικών κλήσεων τις οποίες δέχεται ένα τηλεφωνικό κέντρο ανά λεπτό για μεγάλο αριθμό λεπτών.

β) Η κατανομή του αριθμού των παραγόμενων ελαττωματικών προϊόντων τα οποία υπάρχουν σε ισοπληθή δείγματα για μεγάλο αριθμό δειγμάτων.

γ) Η κατανομή του αριθμού των ημερήσιων αυτοκινητιστικών δυστυχημάτων σε μια εθνική οδό για μεγάλο αριθμό ημερών.

Παράδειγμα

Κατά τη διάρκεια του δεύτερου Παγκοσμίου πολέμου το Λονδίνο διαιρέθηκε σε 200 οικοδομικά τετράγωνα και καταγράφηκε ο αριθμός των ιπτάμενων βομβών που έπεσαν σε κατοικημένα σπίτια και βρέθηκαν τα εξής αποτελέσματα:

Αριθμός βομβών κατά τετράγωνο	0	1	2	3	4	5
Αριθμός οικοδομικών τετραγώνων	46	71	48	23	9	3

Να προσδιοριστούν τα βασικά χαρακτηριστικά της κατανομής Poisson. Για να βρούμε το μέσο αριθμητικό της εμπειρικής κατανομής, πρέπει να βρεθεί το γινόμενο $x_i f_i$ το οποίο βρίσκεται στον παρακάτω πίνακα.

x_i	f_i	$x_i f_i$
0	46	0
1	71	71
2	48	96
3	23	69

4	9	36
5	3	15
Σύνολο	200	287

Επομένως έχουμε:

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{287}{200} = 1{,}44$$

Τα βασικά στατιστικά μέτρα είναι:

μ = λ = 1,44 (μέσος αριθμητικός)

σ = $\sqrt{\lambda} = \sqrt{1{,}44} = 1{,}2$ (μέση απόκλιση τετραγώνου)

$\beta_1 = \frac{1}{\lambda} = \frac{1}{1{,}44} = 0{,}69$ θετική ασυμμετρία (συντελεστής ασυμμετρίας)

$\beta_2 = 3 + \frac{1}{\lambda} = 3 + 0{,}69 = 3{,}69$ λεπτόκυρτη (συντελεστής κύρτωσης)

4.1.4. ΥΠΕΡΓΕΩΜΕΤΡΙΚΗ ΚΑΤΑΝΟΜΗ

Ας υποθέσουμε ότι ένα κιβώτιο περιέχει N σφαίρες και έστω ότι υπάρχουν k πράσινες και N - k κόκκινες σφαίρες. Αν πραγματοποιήσουμε n ανεξάρτητες εξαγωγές σφαιρών από το κιβώτιο (με επανατοποθέτηση), τότε ο αριθμός των εξαγόμενων πράσινων σφαιρών του δείγματος είναι μια συνεχής τυχαία μεταβλητή η οποία ακολουθεί τη διωνυμική κατανομή. Στην περίπτωση που η εξαγωγή των n μονάδων του δείγματος πραγματοποιηθεί χωρίς επανατοποθέτηση, τότε μετά από κάθε εξαγωγή σφαίρας το περιεχόμενο του κιβωτίου μεταβάλλεται και οι εξαγωγές δεν είναι πλέον ανεξάρτητες. Ο αριθμός των πράσινων σφαιρών στο δείγμα είναι μια ασυνεχής τυχαία μεταβλητή, η οποία δεν ακολουθεί τη διωνυμική κατανομή αλλά τη λεγόμενη υπεργεωμετρική.

Αν η τυχαία μεταβλητή X αντιπροσωπεύει το πλήθος των πράσινων σφαιρών στο δείγμα των n εξαγωγών, τότε η πιθανότητα το δείγμα n σφαιρών να περιέχει x πράσινες σφαίρες υπολογίζεται με βάση τον παρακάτω τύπο:

$$P(X = x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$$

Όπου x = 0,1,2,3,…,n

N = πλήθος σφαιρών στο κιβώτιο

n = πλήθος σφαιρών στο δείγμα

k = πλήθος πράσινων σφαιρών στο δείγμα.

Η συνάρτηση ονομάζεται υπεργεωμετρική κατανομή πιθανότητας και εφαρμόζεται σε εκείνες τις περιπτώσεις κατά τις οποίες η επιλογή των μονάδων του δείγματος γίνεται χωρίς

επανατοποθέτηση εν αντιθέσει με την διωνυμική κατανομή όπου η επιλογή των μονάδων του δείγματος γίνεται με επανατοποθέτηση. Αν ο αριθμός N είναι πολύ μεγάλος και το μέγεθος του δείγματος n σχετικά μικρό, τότε οι πιθανότητες που υπολογίζονται με την υπεργεωμετρική κατανομή σχεδόν ταυτίζονται με αυτές της διωνυμικής κατανομής. Επιπλέον η υπεργεωμετρική κατανομή χρησιμοποιείται αρκετά στον ποιοτικό στατιστικό έλεγχο. Έχει περιορισμένη εφαρμογή διότι απαιτούνται αρκετοί υπολογισμοί για την εύρεση των πιθανοτήτων. Τέλος χρησιμοποιείται στο στατιστικό έλεγχο ποιότητας για την αποδοχή ή απόρριψη προϊόντων που υποβάλλονται σε έλεγχο της ποιότητας κατά μεγάλες παρτίδες.

Η υπεργεωμετρική κατανομή εξαρτάται από τις εξής τρεις παραμέτρους:

1. Το μέγεθος N του πληθυσμού που ερευνούμε
2. Το μέγεθος n του δείγματος που εξετάζουμε
3. Το ποσοστό $\mathbf{p} = \frac{N_1}{N}$

Ο μέσος αριθμητικός υπεργεωμετρικής και διωνυμικής κατανομής βρίσκεται από τον ίδιο τύπο: $\mu = n \cdot p$.

Η διακύμανση και η μέση απόκλιση τετραγώνου είναι:

$$\sigma^2 = n \cdot p \cdot q \cdot \frac{N-n}{N-1} \quad , \quad \sigma = \sqrt{n \cdot p \cdot q \cdot \frac{N-n}{N-1}}$$

Παράδειγμα

Έστω ότι σε ένα κιβώτιο υπάρχουν 10 ηλεκτρικοί λαμπτήρες από τους οποίους οι 4 είναι καμένοι. Βγάζουμε από το κιβώτιο διαδοχικά (χωρίς επανατοποθέτηση 5 λαμπτήρες). Ποιά είναι η πιθανότητα να υπάρχουν στο δείγμα 2 καμένοι λαμπτήρες;

Λύση

Έχουμε: N = 10, k = 4, n = 5, x = 2.

Επομένως:

$$P(X=2) = \frac{\binom{4}{2}\binom{10-4}{5-2}}{\binom{10}{5}} = \frac{\binom{4}{2}\binom{6}{3}}{\binom{10}{5}} = \frac{30}{63} = 0,476$$

4.1.5. ΓΕΩΜΕΤΡΙΚΗ ΚΑΤΑΝΟΜΗ

Η γεωμετρική κατανομή έχει σχέση με το πλήθος των δοκιμών Bernoulli που θα πρέπει να εκτελεστούν μέχρι να παρατηρηθεί ορισμένος αριθμός επιτυχιών. Για το λόγο αυτό αναφέρεται συνήθως ως κατανομή του χρόνου αναμονής.

Ορισμός: Θεωρούμε μια ακολουθία (ανεξάρτητων) δοκιμών Bernoulli με πιθανότητα επιτυχίας p και αποτυχίας q=1-p σταθερή για όλες τις δοκιμές. Έστω x ο αριθμός των δοκιμών μέχρι την εμφάνιση της πρώτης επιτυχίας. Η κατανομή της τυχαίας κατανομής X καλείται γεωμετρική κατανομή με παράμετρο p.

Το ενδεχόμενο {X = x}, x ≥ 1 σημαίνει ότι στις δοκιμές 1,2,…,x-1 εμφανίστηκε αποτυχία α και στην επόμενη δοκιμή εμφανίστηκε επιτυχία ε. Άρα το ενδεχόμενο αυτό περιέχει ένα και μοναδικό αποτέλεσμα το:

$$\underbrace{\alpha \quad \alpha \quad \dots \quad \alpha}_{x-1 \text{ φορές}} \; \varepsilon$$

Αφού οι δοκιμές είναι ανεξάρτητες, η πιθανότητα εμφάνισης του αποτελέσματος αυτού θα είναι:

$$\underbrace{q \quad q \quad \dots \quad q}_{x-1 \text{ φορές}} \; p = q^{x-1} \cdot p$$

Επομένως η συνάρτηση πιθανότητας της γεωμετρικής κατανομής δίνεται από τον τύπο: f(x) = P(X = x) = $q^{x-1} \cdot p$ x = 1,2,…

Για τη συνάρτηση της γεωμετρικής κατανομής μπορούμε να πούμε ότι:

$$F(t) = \begin{cases} 0, & t < 1 \\ \sum_{x=1}^{|t|} f(x), & t \geq 1 \end{cases}$$

Και παρατηρώντας ότι για |t| = k καταλήγουμε στον εξής τύπο:

$$F(t) = \begin{cases} 0, & t < 1 \\ 1 - q^{|t|}, & t \geq 1 \end{cases}$$

Αν η τυχαία μεταβλητή X ακολουθεί τη γεωμετρική κατανομή με παράμετρο p τότε μ = $\frac{1}{p}$ (μέση τιμή) $\sigma^2 = \frac{q}{p^2}$ (διακύμανση).

Παράδειγμα

Ένα ζάρι ρίχνεται συνεχώς μέχρι να εμφανιστεί άσος. Ποια είναι η πιθανότητα να συμβεί αυτό:

α) στη δέκατη ρίψη

β) πριν από τη δέκατη ρίψη

γ) μετά από τη δέκατη ρίψη.

Λύση

Αν συμβολίσουμε με X τον αριθμό των ρίψεων μέχρι να εμφανιστεί για πρώτη φορά άσος, η τυχαία μεταβλητή X ακολουθεί τη γεωμετρική κατανομή με p = $^1/_6$ άρα:

$f(x) = P(X = x) = \left(\frac{5}{6}\right)^{x-1} \cdot {}^1/_6$ x = 1,2,…

και

$$F(t) = \begin{cases} 0, & t < 1 \\ 1 - \left(\frac{5}{6}\right)^{\lfloor t \rfloor}, & t \geq 1 \end{cases}$$

Με βάση τους τύπους έχουμε:

α) $P(X=10) = f(10) = \left(\frac{5}{6}\right)^9 \cdot \frac{1}{6} = 3.2\%$

β) $P(X<10) = P(X \leq 9) = F(9) = 1 - \left(\frac{5}{6}\right)^9 = 80,6\%$

γ) $P(X>10) = 1 - P(X \leq 10) = 1 - F(10) = 1 - \left[1 - \left(\frac{5}{6}\right)^{10}\right] = \left(\frac{5}{6}\right)^{10} =$ 16,2%

4.1.6. ΑΡΝΗΤΙΚΗ ΔΙΩΝΥΜΙΚΗ ΚΑΤΑΝΟΜΗ

Όπως και η γεωμετρική κατανομή έτσι και η αρνητική διωνυμική αποκαλείται κατανομή του χρόνου αναμονής. Επίσης, η συγκεκριμένη κατανομή πολλές φορές φέρει την ονομασία κατανομή του Pascal. Μια άμεση γενίκευση της γεωμετρικής κατανομής προκύπτει αν εξετάσουμε το χρόνο αναμονής μέχρι την r επιτυχία, όπου r είναι ένας θετικός ακέραιος αριθμός.

Ορισμός: Θεωρούμε μια ακολουθία (ανεξάρτητων) δοκιμών Bernoulli με πιθανότητα επιτυχίας p και αποτυχίας q=1-p σταθερή για όλες τις δοκιμές. Έστω X ο αριθμός των δοκιμών μέχρι την εμφάνιση της r επιτυχίας. Η κατανομή της τυχαίας μεταβλητής X καλείται αρνητική διωνυμική κατανομή, με παραμέτρους r και p.

Το σύνολο τιμών της τυχαίας μεταβλητής X είναι $R_x = \{r, r+1, r+2, \ldots\}$. Η συνάρτηση πιθανότητας της αρνητικής διωνυμικής κατανομής δίνεται από τον τύπο: $f(x) = P(X = x) = \binom{x-1}{r-1} q^{x-r} \cdot p^r$ $\quad x = r, r+1, \ldots$

Επομένως προκύπτει, ότι η γεωμετρική κατανομή συμπίπτει με την ειδική περίπτωση r = 1 της αρνητικής διωνυμικής.

Αν η τυχαία μεταβλητή X ακολουθεί την αρνητική διωνυμική κατανομή με παραμέτρους r και p, τότε:

$\mu = \frac{r}{p}$ μέσος

$\sigma^2 = \frac{rq}{p^2}$ διακύμανση.

Η διαφορά μεταξύ διωνυμικής και κατανομής του Pascal είναι:

Στην διωνυμική κατανομή έχουμε έναν αριθμό n δοκιμών και προσπαθούμε να υπολογίσουμε την πιθανότητα να έχουμε στις n δοκιμές x επιτυχίες. Αντιθέτως, στην αρνητική διωνυμική κατανομή ορίζουμε a τον αριθμό των επιτυχιών και προσπαθούμε να προσδιορίσουμε την πιθανότητα του αριθμού των δοκιμών που απαιτούνται για να έχουμε x επιτυχίες. Όπως αναφέρθηκε και παραπάνω η ειδική περίπτωση r = 1 δίνει τη γεωμετρική κατανομή.

Παράδειγμα

Στον τελικό των play-offs έχουν προκριθεί 2 ομάδες α, β οι οποίες παίζουν μεταξύ τους μέχρι κάποια να συμπληρώσει 3 νίκες. Ας υποθέσουμε ότι οι αγώνες αποτελούν ανεξάρτητες δοκιμές και ότι η πιθανότητα να κερδίσει έναν αγώνα η ομάδα α είναι 0,48 , ενώ η πιθανότητα να κερδίσει η ομάδα β είναι 0,52 (δεν υπάρχει πιθανότητα ισοπαλίας). Ποια είναι η πιθανότητα να χρειαστεί να γίνουν 5 παιχνίδια μέχρι να αναδειχθεί ο πρωταθλητής.

Λύση

Ορίζουμε τις τυχαίες μεταβλητές

X_1 : αριθμός παιχνιδιών μέχρι να συμπληρώσει 3 νίκες η ομάδα α

X_2 : αριθμός παιχνιδιών μέχρι να συμπληρώσει 3 νίκες η ομάδα β.

Άρα, έχουμε:

$f_1(x) = P(X_1 = x) = \binom{x-1}{2} (0,48)^3 (0,52)^{x-3}$ $\quad x = 3, 4, \ldots$

$f_2(x) = P(X_2 = x) = \binom{x-1}{2} (0,52)^3 (0,48)^{x-3}$ $\quad x = 3, 4, \ldots$

Η πιθανότητα που ζητάμε είναι η πιθανότητα της ένωσης $A \cup B$ των ξένων ενδεχομένων.

A: Η ομάδα α αναδεικνύεται πρωταθλήτρια σε 5 παιχνίδια

B: Η ομάδα β αναδεικνύεται πρωταθλήτρια σε 5 παιχνίδια

Επομένως:

$$P(A \cup B) = P(A) + P(B) = P(X_1 = 5) + P(X_2 = 5) = f_1(5) + f_2(5)$$
$$= \binom{4}{2} (0{,}48)3\,(0{,}52)x - 3 + \binom{4}{2} (0{,}52)3\,(0{,}48)x - 3 = 37\%$$

4.2. ΣΥΝΕΧΕΙΣ ΚΑΤΑΝΟΜΕΣ

Εάν μια μεταβλητή είναι συνεχής, τότε δεν έχει νόημα να αναζητούμε την πιθανότητα να έχει μια συγκεκριμένη τιμή, αφού οι πιθανές τιμές είναι άπειρες. Επομένως, στις συνεχείς μεταβλητές δεν υπάρχουν απλές πιθανότητες αλλά αυτό που γνωρίζουμε ή προσπαθούμε να προσεγγίσουμε είναι η μαθηματική συνάρτηση της κατανομής f(X).Η συνάρτηση μια συνεχούς μεταβλητής ονομάζεται συνάρτηση πυκνότητας πιθανότητας (probability density function). Δηλαδή είναι η συνάρτηση που εκφράζει την πυκνότητα (συχνότητα) εμφάνισης των παρατηρήσεων στα διάφορα διαστήματα τιμών.

4.2.1. ΚΑΝΟΝΙΚΗ ΚΑΤΑΝΟΜΗ

Η κανονική κατανομή είναι η πιο σπουδαία κατανομή της θεωρίας πιθανοτήτων και της στατιστικής, κυρίως λόγω της ευρείας χρησιμότητας της σε ένα μεγάλο πλήθος εφαρμογών. Μερικοί από τους λόγους που εξηγούν την εξέχουσα θέση της είναι οι εξής:

- Πολλά πληθυσμιακά χαρακτηριστικά (ύψος, βάρος κλπ) περιγράφονται ικανοποιητικά από την κανονική κατανομή
- Τυχαία σφάλματα που εμφανίζονται σε διάφορες μετρήσεις ακολουθούν την κανονική κατανομή. Για αυτό το λόγο, η κανονική κατανομή αναφέρεται συχνά και ως κατανομή σφαλμάτων
- Το άθροισμα και ο μέσος όρος μεγάλου αριθμού παρατηρήσεων τυχαίας μεταβλητής ακολουθεί κατά προσέγγιση κανονική κατανομή, ανεξαρτήτως από το ποια κατανομή ακολουθούν οι αρχικές παρατηρήσεις
- Μεγάλος αριθμός κατανομών (διακριτές και συνεχείς) μπορούν κάτω από ορισμένες συνθήκες να προσεγγισθούν από την κανονική κατανομή.

Για πρώτη φορά η κανονική κατανομή χρησιμοποιήθηκε το 1733 από το γάλλο μαθηματικό Abraham De Moivre, σαν μια οριακή μορφή της διωνυμικής κατανομής. Όταν το μέγεθος του δείγματος n αυξάνει (τείνει προς το άπειρο) τότε το διωνυμικό ανάπτυγμα $(p+q)^n$ προσεγγίζει μια συμμετρική ομαλή καμπύλη που ονομάζεται κανονική κατανομή. Η ανακάλυψη της κανονικής κατανομής οφείλεται στο γάλλο μαθηματικό Laplace (1749-1727) και στο γερμανό μαθηματικό και αστρονόμο Gauss (1777-1855), οι οποίοι εργάστηκαν ανεξάρτητα ο ένας από τον άλλο. Η ανακάλυψη της κανονικής κατανομής προήλθε από την μελέτη των σφαλμάτων παρατηρήσεως στη φυσική και στην αστρονομία για αυτό ονομάζεται και μελέτη των σφαλμάτων. Ο άγγλος στατιστικός Karl Pearson απέδειξε ότι η κανονική καμπύλη ήταν μία από τις πολλές μορφές καμπυλών που εμφανίζονται στην φύση. Εκτός από την τεράστια σημασία που έχει η κανονική κατανομή στη στατιστική, αποτελεί το θεμέλιο στο οποίο στηρίχθηκε η θεωρία της νεώτερης στατιστικής.

4.2.1.1. ΓΕΝΙΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΤΗΣ ΚΑΝΟΝΙΚΗΣ ΚΑΤΑΝΟΜΗΣ:

Η συνάρτηση πιθανότητας της κανονικής κατανομής είναι:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty < x < +\infty$$

όπου:

- μ ο μέσος αριθμητικός της τυχαίας μεταβλητής X
- σ η τυπική απόκλιση της τυχαίας μεταβλητής X
- π = 3,14
- e = 2,71

Η κανονική κατανομή έχει συντελεστή ασυμμετρίας του Pearson $\beta_1 = 0$ και κύρτωση $\beta_2 = 3$. Επιπρόσθετα, ο μέσος αριθμητικός, η διάμεσος και η επικρατούσα τιμή συμπίπτουν. Αν διαπιστώσουμε ότι μια κατανομή συχνοτήτων έχει $\beta_1 = 0$ και $\beta_2 = 3$, τότε συμπεραίνουμε ότι η κατανομή συχνοτήτων ακολουθεί κανονική κατανομή.

Ο μέσος αριθμητικός δίνεται από τον εξής τύπο:

$$\mu = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Ενώ ο τύπος της διακύμανσης είναι:

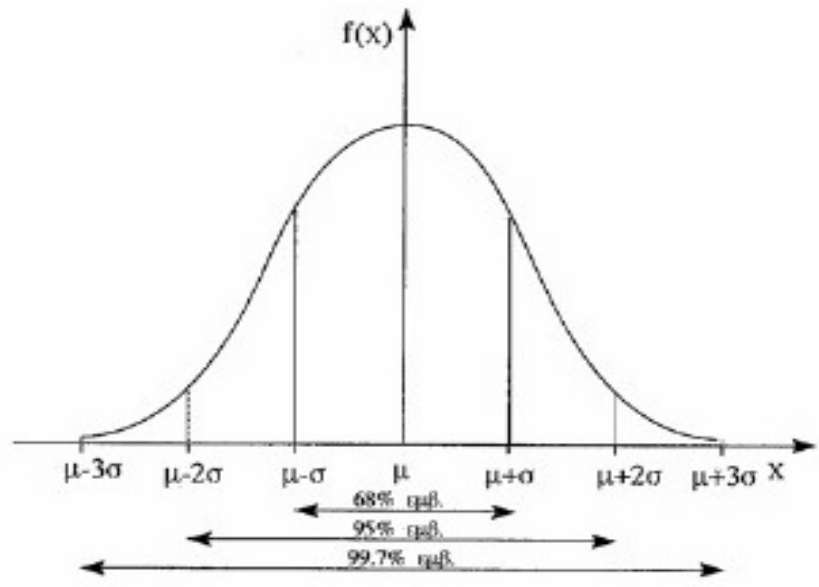
$$\sigma^2 = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx - \mu^2$$

Ορισμός: αν έχουμε μια συνεχή τυχαία μεταβλητή X που μπορεί να πάρει τιμές από -∞ μέχρι +∞ και αν η συνάρτηση πιθανότητας της X δίνεται από τι σχέση $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ τότε λέμε ότι η τυχαία μεταβλητή ακολουθεί την κανονική κατανομή με μέσο αριθμητικό μ και διακύμανση σ^2.

Για να είναι η παραπάνω σχέση συνάρτηση πυκνότητας πρέπει να πληρούνται οι παρακάτω προϋποθέσεις:

1) $f(x) \geq 0$

2) $\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = 1$



Σύμφωνα με το παραπάνω διάγραμμα ένα ποσοστό σχετικά κοντά στο 68% των τιμών ενός κανονικού βρίσκονται σε απόσταση το πολύ μιας τυπικής απόκλισης από τη μέση τιμή μ (μ – σ ≤ x ≤ μ + σ), ενώ περίπου 95% σε απόσταση δύο τυπικών αποκλίσεων από το μ και περίπου 99,7% σε απόσταση τριών αποκλίσεων από το μ. Τα αποτελέσματα αυτά είναι πάρα πολύ χρήσιμα για τη δημιουργία διαστημάτων εμπιστοσύνης και για τον έλεγχο στατιστικών υποθέσεων.

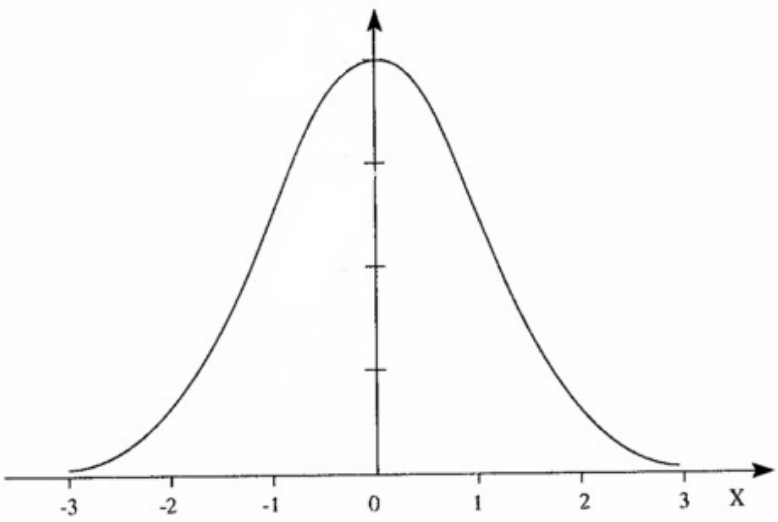
Τυποποιημένη κανονική κατανομή

Οι κανονικές κατανομές είναι άπειρες όσα και τα δυνατά ζεύγη των παραμέτρων μ και σ. Κάθε φορά που ορίζουμε ένα συγκεκριμένο συνδυασμό μ και σ έχουμε και μια διαφορετική κανονική κατανομή. Για παράδειγμα, διαφορετική είναι η κατανομή της διάρκειας των μπαταριών (μ=191 ώρες, σ=5 ώρες) σε σχέση με τη κατανομή του βάρους των νεογνών (μ=3000 γραμμάρια,σ=400 γραμμάρια). Αυτό σημαίνει ότι υπάρχουν άπειρες κανονικές κατανομές, αφού άπειρα είναι και τα χαρακτηριστικά που ακολουθούν τον κανονικό νόμο. Συνεπώς χρησιμοποιούμε μια συγκεκριμένη κατανομή για να τυποποιήσουμε τα δεδομένα, με σκοπό τη μείωση των προβλημάτων στον υπολογισμό των πιθανοτήτων των διαφόρων κατανομών. Η παραπάνω κατανομή καλείται τυποποιημένη κανονική κατανομή.

Η συνάρτηση πυκνότητας πιθανότητας της τυποποιημένης κανονικής κατανομής δίνεται από τη σχέση:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad ,-\infty < x < \infty$$

Η κανονική κατανομή που έχει μέσο αριθμητικό 0 [E(x) = μ = 0] και διακύμανση 1 [Var(x) = σ_x^2 = 1] λέγεται τυποποιημένη κανονική κατανομή και συμβολίζεται με N(0,1). Επιπλέον είναι ανεξάρτητη από τη μονάδα μέτρησης της μεταβλητής X αφού δεν εκφράζεται σε καμία μονάδα μέτρησης.



Στην περίπτωση που η κανονική κατανομή είναι τυποποιημένη τότε στο παραπάνω διάγραμμα φαίνεται ότι μεταξύ ±1 η καμπύλη περιλαμβάνει το 68% περίπου των περιπτώσεων, ενώ στα διαστήματα ±2 και ±3 η καμπύλη περιλαμβάνει το 95,4% και το 99,7% των περιπτώσεων αντίστοιχα.

4.2.1.2. ΠΡΟΣΕΓΓΙΣΗ ΤΗΣ ΔΙΩΝΥΜΙΚΗΣ ΚΑΤΑΝΟΜΗΣ ΜΕ ΤΗΝ ΚΑΝΟΝΙΚΗ ΚΑΤΑΝΟΜΗ

Το βασικότερο πλεονέκτημα της κανονικής κατανομής είναι ότι μπορεί να προσεγγίσει με μεγάλη ακρίβεια τις ασυνεχείς κατανομές, αν και ως συνεχής κατανομή περιγράφει μόνο συνεχείς μεταβλητές. Για παράδειγμα η διωνυμική κατανομή τείνει προς την κανονική, όταν το δείγμα n είναι μεγαλύτερο από 20. Αντιθέτως, σε δείγματα μικρότερου μεγέθους η πιθανότητα p πλησιάζει στο 0,5.

Αν έχουμε μια τυχαία μεταβλητή που ακολουθεί την διωνυμική κατανομή, δηλαδή έχουμε:

$$P\{X = x_i\} = \binom{n}{x} p^x q^{n-x} \quad , \quad x = 0,1,2,3,\dots,n$$

Αν n είναι αρκετά μεγάλο και ταυτόχρονα κανένα από τα p και q δεν τείνει προς το μηδέν τότε η πιθανότητα $P\{X = x\}$, δίνεται με σχετικά μεγάλη προσέγγιση από την τυποποιημένη κανονική κατανομή με μεταβλητή:

$$Z = \frac{X - np}{\sqrt{npq}}$$

Όπου np ο μέσος και npq η διακύμανση της διωνυμικής κατανομής. Όταν το x βρίσκεται κοντά στη μέση τιμή np και το p είναι περίπου 0,5 τόσο η προσέγγιση μεγαλώνει και στην περίπτωση που np και npq είναι μεγαλύτερο του 5, τότε η προσέγγιση είναι καλή.

Επειδή τυχαία μεταβλητή x της διωνυμικής κατανομής είναι ασυνεχής πρέπει να μετατραπεί σε συνεχή ώστε να μπορέσουμε να χρησιμοποιήσουμε τα εμβαδά της κανονικής κατανομής. Για να γίνει αυτό προσθέτουμε ή αφαιρούμε με τον αριθμό 0,5 και έχουμε:

$$P\{x_1 \le X \le x_2\} = F\left(\frac{x_2+0,5-np}{\sqrt{npq}}\right) - F\left(\frac{x_1-0,5-np}{\sqrt{npq}}\right)$$

Παράδειγμα

Έστω ότι μια τυχαία μεταβλητή X ακολουθεί τη διωνυμική κατανομή με n = 250 και p =0,3. Ζητείται η πιθανότητα $P\{60 \le X \le 80\}$ με βάση την κανονική κατανομή σαν προσέγγιση της διωνυμικής.

Λύση

Η πιθανότητα $P\{60 \le X \le 80\}$ της ασυνεχούς μεταβλητής της διωνυμικής κατανομής είναι ίση με την πιθανότητα $P\{59,5 \le X \le 80,5\}$ της συνεχούς μεταβλητής με μέσο αριθμητικό μ = np = 250*0,3 = 75 και τυπική απόκλιση $\sigma = \sqrt{npq} = \sqrt{250 * 0,3 * 0,7} = 7,24$

Άρα η ζητούμενη πιθανότητα σύμφωνα με την τυποποιημένη κανονική κατανομή θα είναι:

$$P\{60 \le X \le 80\} = P\{59,5 \le X \le 80,5\} = P\left\{\frac{59,5-75}{7,24} \le \frac{X-np}{\sqrt{npq}} \le \frac{80,5-75}{7,24}\right\} =$$
$$P\{-2,14 \le Z \le 0,76\} = F(0,76) - F(-2,14) = F(0,76) - [1 - F(-2,14)] =$$
$$0,7764 - [1 - 0,9838] = 0,7764 - 0,0162 = 0,7602$$

4.2.2. ΕΚΘΕΤΙΚΗ ΚΑΤΑΝΟΜΗ

Εάν έχουμε μια συνεχής τυχαία μεταβλητή X που έχει συνάρτηση πυκνότητας πιθανότητας

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0 \\ 0, & x < 0 \end{cases}$$

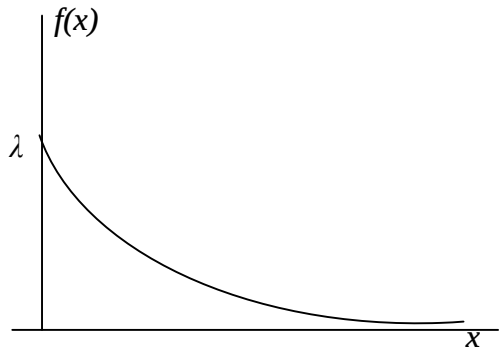
τότε λέμε ότι η τυχαία μεταβλητή X ακολουθεί την εκθετική κατανομή με παράμετρο λ > 0.

Η συνάρτηση f είναι μη αρνητική και ταυτόχρονα έχουμε:

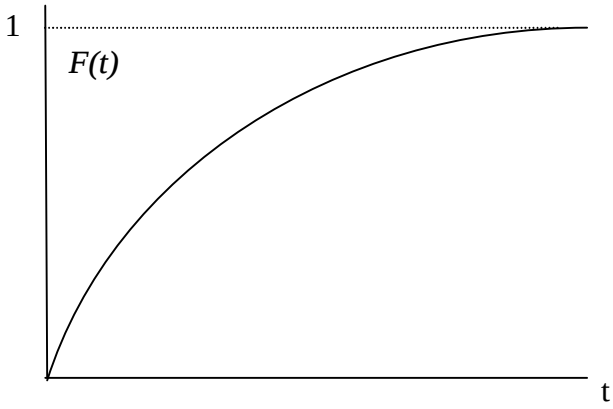
$$\int_{-\infty}^{\infty} f(x)dx = \int_0^{\infty} \lambda e^{-\lambda x} dx = \left[-e^{-\lambda x}\right]_0^{\infty} = 1$$

Η συνάρτηση κατανομής F δίνεται από τη σχέση:

$$F(t) = \begin{cases} 0, & t < 0 \\ 1 - e^{-\lambda t}, & t \ge 0 \end{cases}$$



Αυτή είναι η γραφική παράσταση της συνάρτησης πυκνότητας.



Αυτή είναι η γραφική παράσταση της συνάρτησης κατανομής της F.

Όταν η τυχαία μεταβλητή X ακολουθεί την εκθετική κατανομή με παράμετρο λ>0, τότε η μέση τιμή της X έχει τον εξής τύπο:

μ= E(X) = $\frac{1}{\lambda}$ ενώ η διακύμανση της X δίνεται από τον τύπο:

σ^2 = V(X) = $\frac{1}{\lambda^2}$

Παράδειγμα

Ας υποθέσουμε ότι ένας συγκεκριμένος εστιακός χώρος προκαλεί σεισμούς εντάσεως μεγαλύτερης των πέντε βαθμών της κλίμακας Richter με ρυθμό δύο σεισμούς ανά έτος.

Α) Να βρεθεί η κατανομή των ενδιάμεσων χρόνων μεταξύ διαδοχικών σεισμών έντασης τουλάχιστον πέντε βαθμών.
Β) Ποιος είναι ο μέσος χρόνος αναμονής μεταξύ διαδοχικών σεισμών έντασης τουλάχιστον πέντε βαθμών;
Γ) Να βρεθεί η πιθανότητα ο επόμενος σεισμός έντασης τουλάχιστον πέντε βαθμών να συμβεί μετά από ένα χρόνο και όχι αργότερα από τρία χρόνια από σήμερα.

Λύση

Α) Οι ενδιάμεσοι χρόνοι μεταξύ των διαδοχικών εμφανίσεων του ενδεχομένου θα ακολουθούν την εκθετική κατανομή με παράμετρο λ=2. Επομένως αν Χ είναι η τυχαία μεταβλητή που περιγράφει αυτούς τους χρόνους τότε η συνάρτηση πυκνότητας f δίνεται από τον παρακάτω τύπο:

$$f(x) = \begin{cases} 2e^{-2x} & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Ενώ η συνάρτηση κατανομής F της Χ έχει τον εξής τύπο:

$$F(t) = \begin{cases} 0, & t < 0 \\ 1 - e^{-2t}, & t \geq 0 \end{cases}$$

Β) Ο μέσος χρόνος αναμονής για το ενδεχόμενο ισούται με:

$$\mathbf{E(X)} = \frac{1}{\lambda} = \frac{1}{2} = 0{,}5\ \chi\rho\acute{o}\nu\iota\alpha$$

Γ) Η ζητούμενη πιθανότητα είναι:
$P(1 < X \leq 3) = F(3) – F(1) = (1 - e^{-6}) – (1 - e^{-2}) = e^{-2} - e^{-6} = 13\%$.

4.2.3. *ΚΑΤΑΝΟΜΗ ΓΑΜΜΑ*

Έστω Χ μια συνεχής τυχαία μεταβλητή με συνάρτηση πυκνότητας

$$f(x) = \begin{cases} \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Όπου λ και α μεγαλύτερο του μηδενός και πραγματικοί αριθμοί. Η κατανομή της τυχαίας μεταβλητής Χ καλείται κατανομή Γάμμα με παραμέτρους λ και α.

Αν η τυχαία μεταβλητή Χ ακολουθείται η κατανομή γάμμα με παραμέτρους λ και α, τότε η μέση τιμή ισούται με:

$$\mathbf{E(X)} = \frac{\alpha}{\lambda}$$

Ενώ η διακύμανση της Χ είναι:

$$V(X) = \frac{\alpha}{\lambda^2}$$

Η συνεχής κατανομή με συνάρτηση πυκνότητας

$$f(x) = \begin{cases} \frac{\lambda^{\nu}}{(\nu-1)!} x^{\nu-1} e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Καλείται κατανομή ***Erlang*** με παραμέτρους ν και λ προς τιμήν του Δανού μαθηματικού Erlang ο οποίος πρώτος την χρησιμοποίησε για να μελετήσει προβλήματα σχετικά που αφορούσαν το τηλεπικοινωνιακό δίκτυο της χώρας. Στην περίπτωση που ν = 1 η κατανομή Erlang συμπίπτει με την εκθετική κατανομή. Τέλος αν θελήσουμε να γενικεύσουμε την κατανομή Erlang τότε το κύριο πρόβλημα που καλούμαστε να αντιμετωπίσουμε είναι η ύπαρξη $(\nu-1)!$ η οποία έχει νόημα μόνο για ακέραιες θετικές τιμές του ν. Η λύση του προβλήματος επιτυγχάνεται με τη χρήση της κατανομής γάμμα που αναφέρεται πιο πάνω.

4.2.4. ΚΑΤΑΝΟΜΗ ΒΗΤΑ

Έστω ότι έχουμε μια συνεχής τυχαία μεταβλητή με συνάρτηση:

$$f(x) = \begin{cases} \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, & 0 < x < 1 \\ 0, & x \leq 0 \text{ και } x \geq 1 \end{cases}$$

όπου α και β > 0 και πραγματικοί αριθμοί. Η συνάρτηση κατανομής F(X) δίνεται από τον παρακάτω τύπο:

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$$

Η κατανομή της τυχαίας μεταβλητής X ονομάζεται κατανομή βήτα με παραμέτρους α και β. Η ποσότητα $B(\alpha,\beta)$ που εμφανίζεται παραπάνω καλείται συνάρτηση βήτα και η σχέση της με τη συνάρτηση γάμμα είναι η παρακάτω:

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\cdot\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Αν η τυχαία μεταβλητή X ακολουθεί την κατανομή βήτα με παραμέτρους α και β τότε η μέση τιμή της X είναι:

$$\mathrm{E(X)} = \frac{\alpha}{\alpha+\beta}$$

Ενώ η διακύμανση του X δίνεται από τον εξής τύπο:

$$\mathrm{V(X)} = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$$

Παράδειγμα

Το ποσοστό των γνώσεων που συγκρατεί ένας φοιτητής κατά την παρακολούθηση ενός συγκεκριμένου μαθήματος, περιγράφεται από μια συνεχή τυχαία μεταβλητή X που ακολουθεί την κατανομή βήτα με παραμέτρους α = 6 και β = 2.

α) Να βρεθεί η συνάρτηση κατανομής f(x)

β) Ποιο είναι το μέσο ποσοστό γνώσεων που συγκρατεί ο φοιτητής;

γ) Ποια είναι η διακύμανση της τυχαίας μεταβλητής X;

Λύση

Η συνάρτηση f της τυχαίας μεταβλητής X είναι:

$$f(x) = \begin{cases} \dfrac{x^{6-1}(1-x)^{2-1}}{B(6,2)}, & 0 < x < 1 \\ 0, & x \leq 0 \; και \; x \geq 1 \end{cases}$$

και

$$B(6,2) = \frac{\Gamma(6)\cdot\Gamma(2)}{\Gamma(6+2)} = \frac{5!\,1!}{7!} = \frac{1}{6\cdot 7} = \frac{1}{42}$$

άρα θα έχουμε:

$$f(x) = \begin{cases} 42x^5(1-x), & 0 < x < 1 \\ 0, & x \leq 0 \; και \; x \geq 1 \end{cases}$$

α) για τη συνάρτηση κατανομής της X θα έχουμε:

$$F(t) = \int_{-\infty}^{t} f(x)dx = \begin{cases} 0, & x < 0 \\ \int_0^t 42(x^5 - x^6), & 0 \leq x < 1 \\ 1, & x \geq 1 \end{cases}$$

$$= \begin{cases} 0, & x < 0 \\ x^6(7-6x), & 0 \leq x < 1 \\ 1, & x \geq 1 \end{cases}$$

β) η μέση τιμή είναι:

$$\mathrm{E(X)} = \frac{\alpha}{\alpha+\beta} = \frac{6}{6+2} = \frac{6}{8} = 0{,}75 \text{ ή } 75\%$$

Η διακύμανση θα είναι:

$$\mathrm{V(X)} = \frac{6\cdot 2}{(6+2+1)(6+2)^2} = \frac{12}{9\cdot 64} = 0{,}002$$

4.3. ΚΑΤΑΝΟΜΕΣ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ - ΕΙΣΑΓΩΓΗ

Ο βασικός σκοπός της ανάλυσης των δεδομένων είναι να χρησιμοποιήσουμε τις εκτιμήσεις των παραμέτρων του δείγματος όπως ο μέσος και το ποσοστό για να εκτιμήσουμε τις αντίστοιχες παραμέτρους του πληθυσμού. Λόγω του μεγάλου μεγέθους του πληθυσμού, ο ερευνητής περιορίζεται σε ένα δείγμα μικρότερου μεγέθους. Στην περίπτωση που οι παρατηρήσεις του δείγματος αφορούν ποσοτικά χαρακτηριστικά τότε ακολουθεί η εκτίμηση του μέσου ενώ αντίθετα η εκτίμηση του ποσοστού αφορά ποιοτικά χαρακτηριστικά.

Μερικά ερωτήματα όπως πόσο κοντά στην πραγματική τιμή της παραμέτρου του πληθυσμού βρίσκεται η εκτίμηση του δείγματος και ποια είναι η ακρίβεια της είναι αναπάντητα. Αν γνωρίζαμε την παράμετρο του πληθυσμού είναι προφανές ότι δεν θα διεξαγόταν έρευνα. Όμως η προσέγγιση της κατανομής των εκτιμήσεων που προκύπτουν από τα τυχαία δείγματα είναι δυνατή. Δηλαδή μπορούμε να προσεγγίσουμε το δειγματικό χώρο όλων των δυνατών εκτιμήσεων που είναι πιθανές να προκύψουν από όλα τα τυχαία δείγματα ίδιου μεγέθους. Αυτό ονομάζεται κατανομή δειγματοληψίας και δείχνει πως κατανέμονται οι εκτιμήσεις που προκύπτουν από όλα τα πιθανά δείγματα.

4.3.1. ΚΑΤΑΝΟΜΗ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ ΜΕΣΟΥ ΑΡΙΘΜΗΤΙΚΟΥ

Η βαθμολογία πέντε φοιτητών στο μάθημα της Λογιστικής είναι η παρακάτω:

5,6,7,8,9

Έστω ότι θέλουμε να εξετάσουμε τη μέση βαθμολογία των πέντε φοιτητών με τη βοήθεια του τυχαίου δείγματος.

Η μέση βαθμολογία: $\mu = \frac{\Sigma x_i}{N} = \frac{35}{5} = 7$

του συνολικού πληθυσμού στην πραγματικότητα δεν είναι γνωστή και θέλουμε να την εκτιμήσουμε με τη βοήθεια ενός δείγματος με μέγεθος n = 2. Επιλέγουμε με τυχαίο τρόπο δύο βαθμούς και έστω ότι έχουμε επιλέξει τους βαθμούς 5 και 7. Συνεπώς ο μέσος αριθμητικός του δείγματος είναι: $\bar{x} = \frac{5+7}{2} = 6$.

Όμως είναι λάθος να πούμε ότι ο μέσος αριθμητικός του πληθυσμού είναι μ = 6. Αυτό προκύπτει γιατί σε περίπτωση που επιλέξουμε δύο άλλους βαθμούς τότε θα έχουμε ένα νέο αποτέλεσμα. Για παράδειγμα αν πάρουμε τις τιμές 6 και 9 θα έχουμε: $\bar{x} = \frac{6+9}{2} = 7,5$.

Επομένως, αν επιλέξουμε τυχαία τους βαθμούς από το δείγμα (τυχαία δειγματοληψία) είναι πολύ πιθανή η ύπαρξη σφάλματος. Για την λύση του προβλήματος είναι αναγκαίο να έχουμε υπόψη μας το αποτέλεσμα που προκύπτει αν επιλεγούν όλα τα δυνατά δείγματα. Τα δυνατά δείγματα με μέγεθος n = 2 και με επανατοποθέτηση θα είναι τα παρακάτω:

(3,3), (3,4), (3,5), (3,6), (3,7)
(4,3), (4,4), (4,5), (4,6), (4,7)
(5,3), (5,4), (5,5), (5,6), (5,7)
(6,3), (6,4), (6,5), (6,6), (6,7)
(7,3), (7,4), (7,5), (7,6), (7,7)

Αν από κάθε δείγμα υπολογιστεί ο μέσος αριθμητικός, προκύπτει μια σειρά από εκτιμήσεις της αντίστοιχης παραμέτρου του πληθυσμού.

Αν ολόκληρη η σειρά των μέσων αριθμητικών των παραπάνω δειγμάτων ταξινομηθεί κατά συχνότητες, προκύπτει μια κατανομή συχνοτήτων που καλείται κατανομή δειγματοληψίας του μέσου αριθμητικού.

4.3.1.1. ΙΔΙΟΤΗΤΕΣ ΚΑΤΑΝΟΜΗ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ ΜΕΣΟΥ ΑΡΙΘΜΗΤΙΚΟΥ

Η κατανομή συχνοτήτων του μέσου αριθμητικού θα είναι:

1. Στην περίπτωση που ο πληθυσμός από τον οποίο έγινε η επιλογή του δείγματος είναι κανονικός τότε και η κατανομή δειγματοληψίας είναι κανονική, ανεξάρτητα από το μέγεθος των ισοπληθών δειγμάτων. Όμως βασική προϋπόθεση είναι ο αριθμός των δειγμάτων να είναι μεγάλος. Αντίθετα αν ο πληθυσμός από τον οποίο πήραμε τα δείγματα δεν είναι κανονικός τότε η κατανομή δειγματοληψίας του μέσου αριθμητικού θα τείνει να είναι κανονική όσο το μέγεθος του δείγματος αυξάνει.

2. Ο μέσος αριθμητικός της κατανομής ισούται με το μέσο αριθμητικό του πληθυσμού από τον οποίο έγινε η επιλογή των δειγμάτων
   $\mathbf{E}(\bar{x}) = \mu$
   Επιπρόσθετα αν υπολογιστούν και ταξινομηθούν κατά συχνότητα οι διακυμάνσεις όλων των δειγμάτων προκύπτει μια κατανομή που λέγεται κατανομή δειγματοληψίας της διακύμανσης. Η διακύμανση του δείγματος, αν ο μέσος του πληθυσμού είναι άγνωστος δίνεται από τον παρακάτω τύπο:
   $$S_{\bar{x}}^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

Παράδειγμα

Έστω ότι το καθαρό μηνιαίο εισόδημα του πληθυσμού 5 επιχειρήσεων (N = 5) απεικονίζεται στον παρακάτω πίνακα.

Οικογένειες	Καθαρό μηνιαίο εισόδημα
A	100
B	210

Γ	80
Δ	90
Ε	70

Το μέσο μηνιαίο εισόδημα του πληθυσμού των 5 οικογενειών θα είναι:

$$\mu = \frac{100+210+80+90+70}{5} = 110$$

Η διακύμανση του πληθυσμού θα είναι:

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$
$$= \frac{(100-110)^2 + (210-110)^2 + (80-110)^2 + (90-110)^2 + (70-110)^2}{5}$$
$$= 2.599,98$$

Και η τυπική απόκλιση θα είναι: $\sigma = \sqrt{2.599,98}$

Τα δυνατά δείγματα για n = 2 χωρίς επανατοποθέτηση των στοιχείων είναι τα εξής: ΑΒ,ΑΓ,ΑΔ,ΑΕ,ΒΓ,ΒΔ,ΒΕ,ΓΔ,ΓΕ,ΔΕ.

Η μέση αριθμητική των παραπάνω δειγμάτων παρουσιάζονται στον παρακάτω πίνακα:

Δείγματα	Μέσοι δειγμάτων
ΑΒ	(100+210):2 = 155
ΑΓ	(100+80):2 = 90
ΑΔ	(100+90):2 = 95
ΑΕ	(100+70):2 = 85
ΒΓ	(210+80):2 = 145
ΒΔ	(210+90):2 = 150
ΒΕ	(210+70):2 = 140
ΓΔ	(80+90):2 = 85
ΓΕ	(80+70):2 = 75
ΔΕ	(90+70):2 = 80

Σύνολο	1.100

Ο μέσος των μέσων των δειγμάτων θα είναι:

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{10} \bar{x}_i}{10} = \frac{1.100}{10} = 110 = \mu$$

Παρατηρούμε ότι ο μέσος των μέσων όλων των δειγμάτων είναι ίσος με το μέσο του πληθυσμού. Για τον υπολογισμό της διακύμανσης της κατανομής των μέσων των δειγμάτων έχουμε:

$$\sigma_{\bar{x}}^2 = \frac{1}{10}[(155-110)^2 + (90-110)^2 + (95-110)^2 + \ldots + (80-110)^2] = 975$$

$$\sigma_{\bar{x}} = \sqrt{975} = 31{,}2$$

Αν πάρουμε στη συνέχεια n=4 τότε τα δυνατά δείγματα θα είναι:

ΑΒΓΔ, ΑΒΓΕ,ΑΒΔΕ,ΑΓΔΕ,ΒΓΔΕ

Οι μέσοι των δειγμάτων και τα συγκεκριμένα δείγματα παρουσιάζονται αναλυτικά παρακάτω:

Δείγματα	Μέσοι δειγμάτων
ΑΒΓΔ	(100+210+80+90):4= 120
ΑΒΓΕ	(100+210+80+70):4= 115
ΑΒΔΕ	(100+210+90+70):4= 117,5
ΑΓΔΕ	(100+80+90+70):4= 85
ΒΓΔΕ	(210+80+90+70):4= 112,5
Σύνολο	550

Επομένως, ο μέσος των μέσων δειγμάτων θα είναι:

$$\bar{\bar{X}} = \frac{\sum_{i=1}^{5} \bar{x}_i}{5} = \frac{550}{5} = 110 = \mu$$

Η τυπική απόκλιση θα είναι:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} = \frac{50.99}{\sqrt{4}} \cdot \sqrt{\frac{5-4}{5-1}} = 12{,}74$$

Σύμφωνα με τα παραπάνω παρατηρούμε ότι όσο αυξάνεται το μέγεθος του δείγματος, τόσο οι μέσοι των δειγμάτων συγκεντρώνονται κοντά στο μέσο του πληθυσμού.

4.3.2. ΚΑΤΑΝΟΜΗ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ ΠΟΣΟΣΤΟΥ – ΑΝΑΛΟΓΙΑΣ

Ορισμένες φορές σκοπός μας είναι η εκτίμηση του ποσοστού των μονάδων που παρουσιάζουν μια ορισμένη ιδιότητα σε ένα πληθυσμό για παράδειγμα το ποσοστό των αγοριών που φοιτούν σε μια σχολή. Οι πληθυσμοί μπορούν να διαιρεθούν σε δύο κατηγορίες από τις οποίες η μία κατηγορία παρουσιάζει μια ιδιότητα που μας ενδιαφέρει, ενώ η άλλη δεν εμφανίζει το χαρακτηριστικό που θεωρείται σημαντικό για εμάς. Οι συγκεκριμένοι πληθυσμοί καλούνται διχοτομικοί πληθυσμοί.

Αν σε ένα πληθυσμό η αναλογία των μονάδων που εμφανίζουν ένα χαρακτηριστικό συμβολίζεται με P, τότε η αναλογία των μονάδων που δεν περιλαμβάνουν το χαρακτηριστικό αυτό είναι Q = 1 – P.

Για να κάνουμε μία εκτίμηση του ποσοστού του πληθυσμού που παρουσιάζει μια συγκεκριμένη ιδιότητα, επιλέγουμε ένα τυχαίο δείγμα n και προσδιορίζουμε το ποσοστό των μονάδων $p = \frac{\kappa}{n}$ που παρουσιάζουν την συγκεκριμένη ιδιότητα.

Αν επιλέξουμε όλα τα δυνατά δείγματα αυτού του μεγέθους τότε θα έχουμε:

$$n_1 = n_2 = n_3 = \cdots = n_m = n$$

Και αντίστοιχα: $p_1 = \frac{\kappa_1}{n_1}$, $p_2 = \frac{\kappa_2}{n_2}$, $p_3 = \frac{\kappa_3}{n_3}$, . . . , $p_m = \frac{\kappa_m}{n_m}$

Η παραπάνω κατανομή ονομάζεται κατανομή δειγματοληψίας του ποσοστού. Ο μέσος αριθμητικός του ποσοστού της κατανομής δειγματοληψίας είναι: $E(p) = P$

Το τυπικό σφάλμα του P θα είναι: $\sigma_p = \sqrt{\frac{P(1-P)}{n}}$.

Εάν ο πληθυσμός δεν είναι άπειρος, τότε το τυπικό σφάλμα θα είναι:

$$\sigma_p = \sqrt{\frac{pq}{n} \cdot \frac{N-n}{n-1}}$$

Η κατανομή δειγματοληψίας του ποσοστού ακολουθεί την κανονική κατανομή, αν n > 30.

Έστω ότι σε κάθε αντικείμενο ενός συνόλου αντικειμένου αντιστοιχούμε την τιμή 1 ή 0, ανάλογα με το αν το αντικείμενο έχει κάποια ιδιότητα που μας ενδιαφέρει ή όχι. Παίρνουμε τότε ένα πληθυσμό Bernoulli (p) όπου p είναι η αναλογία των αντικειμένων που έχουν χαρακτηριστική ιδιότητα (αναλογία των τιμών “1” στον πληθυσμό).

Έστω x , x_2, …, x_n ένα τυχαίο δείγμα από ένα πληθυσμό Bernoulli (p). Το πλήθος των τιμών “1” (επιτυχιών) στο δείγμα είναι η τυχαία μεταβλητή $Y = \sum_{i=1}^{n} x_i$ που ως γνωστό έχει τη διωνυμική κατανομή με παραμέτρους n και p. Η τυχαία μεταβλητή $\hat{p} = \frac{Y}{n}$ λέγεται

δειγματική αναλογία και χρησιμοποιείται στη στατιστική συμπερασματολογία για την παράμετρο p του πληθυσμού όταν αυτή είναι άγνωστη.

Από τον ορισμό της στατιστικής συμπερασματολογίας $\hat{\mathbf{p}}$ φαίνεται ότι $\hat{\mathbf{p}} = \bar{x}$ δηλαδή η δειγματική αναλογία είναι στην ουσία μια δειγματική μέση τιμή.

Η μέση τιμή και το τυπικό σφάλμα της δειγματικής αναλογίας δίνονται από τους παρακάτω τύπους:

$$\mu_{\hat{p}} = p \text{ (μέση τιμή)}, \qquad \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} \text{ (τυπικό σφάλμα)}$$

Για μεγάλες τιμές του n μπορούμε να προσεγγίσουμε τη δειγματοληπτική κατανομή του $\hat{p}$ με τη $N\left(p, \sqrt{\frac{pq}{n}}\right)$ κατανομή.

Παράδειγμα

Υποθέτουμε ότι το 45% των πολιτών εγκρίνουν την οικονομική πολιτική της κυβέρνησης. Ποια είναι η πιθανότητα σε ένα τυχαίο δείγμα 1000 πολιτών το ποσοστό αυτό να βρίσκεται μεταξύ 45% και 50%;

Λύση

Επειδή το δείγμα είναι πολύ μεγάλο θα χρησιμοποιήσουμε την κανονική κατανομή. Άρα: $\mu_{\bar{p}} = p = 0{,}45$

$$\sigma_{\bar{p}}^2 = \frac{p(1-p)}{n} = \frac{0{,}45(1-0{,}45)}{1000} = 0{,}0002475$$

$$\sigma_{\bar{p}} = 0{,}0157$$

Η τιμή της πιθανότητας θα είναι:

$$P(0{,}45 < \bar{p} < 0{,}50) = P\left[\frac{0{,}45-0{,}45}{0{,}0157} < \frac{\bar{p}-\mu_{\bar{p}}}{\sigma_{\bar{p}}} < \frac{0{,}50-0{,}45}{0{,}0157}\right]$$
$$= P\left(\frac{0}{0{,}0157} < z < \frac{0{,}05}{0{,}0157}\right) = P(0 < z < 3{,}18) = P(z < 3{,}18) - P(z < 0)$$
$$= \Phi(3{,}18) - \Phi(0) = 0{,}99926 - 0{,}5 = 0{,}49926$$

4.3.3. ΚΑΤΑΝΟΜΗ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ ΔΙΑΚΥΜΑΝΣΗΣ

Η διακύμανση S^2 ενός τυχαίου δείγματος $x_1, x_2, \ldots, x_n$ δίνεται από τη σχέση:

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

Η S^2 χρησιμεύει στην στατιστική συμπερασματολογία για την άγνωστη διασπορά του πληθυσμού. Η διακύμανση όπως και η μέση τιμή είναι ανεξάρτητες τυχαίες μεταβλητές.

Η κατανομή της τυχαίας μεταβλητής:

$$\frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma^2}$$

είναι x_i^2 με ν = n-1.

Για την εύρεση της μέσης τιμής και διασποράς για κανονικό πληθυσμό, παρατηρούμε ότι $s^2 = \frac{V\sigma^2}{n-1}$ όπου V είναι μία τυχαία μεταβλητή με τη x_{n-1}^2 κατανομή. Κάνοντας χρήση των δύο σχέσεων E(U) = ν και Var(U) = 2ν έχουμε:

$$E(s^2) = \frac{\sigma^2}{n-1} E(V) = \sigma^2$$

και

$$Var(s^2) = \frac{\sigma^2}{(n-1)^2} Var(V) = \frac{2\sigma^4}{n-1}$$

Η πρώτη σχέση είναι γενική και ισχύει ακόμα και στην περίπτωση που ο πληθυσμός δεν είναι κανονικός. Λόγω αυτής της ιδιότητας λέμε ότι η s^2 είναι αμερόληπτη εκτιμήτρια της διασποράς σ^2 του πληθυσμού.

Παράδειγμα

Σε μία φαρμακοβιομηχανία, μια μηχανή είναι ρυθμισμένη γεμίζει αμπούλες ενέσεων με ορισμένη δόση κάποιου υγρού παρασκευάσματος (φαρμάκου). Το ίδιο σημαντική με τη μέση τιμή των ποσοτήτων του φαρμάκου που αφήνει η μηχανή στις αμπούλες είναι και η διασπορά σ^2 αυτών των ποσοτήτων. Αν η τιμή του σ^2 είναι μεγάλη πολλές αμπούλες θα περιέχουν σημαντικά μικρότερη ή μεγαλύτερη δόση φαρμάκου από την καθορισμένη. Οι τιμές του σ^2 που θεωρούνται επιτρεπτές είναι $\sigma^2 \leq 1(\text{mgr})^2$ και οι προδιαγραφές της μηχανής ικανοποιούν αυτή τη συνθήκη. Έστω ότι ένας ελεγκτής επιλέγει στην τύχη 20 αμπούλες, ζυγίζει το περιεχόμενο τους και βρίσκει $s^2 = 1.6(\text{mgr})^2$. Πρέπει να γίνει νέα ρύθμιση της μηχανής;

Λύση

Για να βρεθεί λύση πρέπει να βρούμε την πιθανότητα να παρατηρήσουμε σε ένα τυχαίο πείραμα την τιμή του s^2 που βρήκε ο ελεγκτής. Αφού οι αμπούλες είναι 20 τότε και ο αριθμός των παρατηρήσεων είναι 20. Υποθέτουμε ότι ο πληθυσμός των ποσοτήτων του φαρμάκου που αφήνει η μηχανή στις αμπούλες είναι κανονικός N(μ,σ) και επομένως η τυχαία μεταβλητή V = (19) s^2/σ^2 έχει την x_{19}^2 κατανομή (τύπος x_{n-1}^2). Έχουμε:

$$P(S^2 \geq 1.6|\sigma^2 = 1) = P\left(V \geq \frac{(19)(1.6)}{1}\right) = P(V > 30.4)$$

Μέσω του πίνακα των κατανομών βλέπουμε ότι η τιμή 30.4 είναι μεγαλύτερη από την τιμή $x^2_{0.05,19} = 30{,}144$.

Επομένως, η ζητούμενη πιθανότητα είναι λίγο μικρότερη από 0,05. Επειδή η πιθανότητα αυτή είναι μικρή έχουμε σημαντικές ενδείξεις ότι η τιμή $\sigma^2 = 1$ που υποθέσαμε δεν είναι σωστή αλλά είναι $\sigma^2 > 1$. Για αυτό πρέπει να κάνουμε νέα ρύθμιση της μηχανής.

4.3.4. ΚΑΤΑΝΟΜΗ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ ΔΙΑΦΟΡΩΝ

Αν υποθέσουμε ότι έχουμε 2 πληθυσμούς και στη συνέχεια πάρουμε όλα τα δυνατά ζεύγη δειγμάτων (n_1 n_2) και από κάθε ζεύγος δείγματος υπολογίσουμε τους μέσους $\bar{x}_1$ και $\bar{x}_2$ τότε θα προκύψουν τόσες διαφορές $(\bar{x}_1 - \bar{x}_2)$ όσα και τα ζεύγη των δειγμάτων, άρα θα πάρουμε μια δειγματοληπτική κατανομή των διαφορών των μετρήσεων των 2 δειγμάτων. Αν συμβολίσουμε το μέσο αριθμητικό με μ και τη διακύμανση με σ^2 και ο πληθυσμός είναι άπειρος τότε η μέση τιμή των διαφορών $\bar{x}_1 - \bar{x}_2$ θα ισούται με τη διαφορά $\mu_1 - \mu_2$, δηλαδή:

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$$

Η τυπική απόκλιση της κατανομής της δειγματοληψίας της διαφοράς των δύο μέσων με δεδομένο ότι οι διακυμάνσεις και των δύο πληθυσμών είναι γνωστές θα είναι:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Στην περίπτωση που ο πληθυσμός είναι πεπερασμένος αλλά και πάλι οι διακυμάνσεις είναι γνωστές τότε το τυπικό σφάλμα εκτίμησης ισούται με:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \cdot \sqrt{\frac{N - n}{N - 1}}$$

Παράδειγμα

Έστω ότι εξετάζουμε τους φοιτητές δύο σχολών Α και Β ως προς το ύψος τους. Υποθέτουμε ότι η κατανομή των υψών τους είναι κανονική με μ_1 = 170cm , σ_1^2 = 3 και μ_2 = 169cm και σ_2^2 = 2. Παίρνουμε ένα τυχαίο δείγμα 60 φοιτητών από το τη σχολή Α και ένα τυχαίο δείγμα 50 φοιτητών από τη σχολή Β. Να βρεθεί η πιθανότητα η διαφορά των μέσων τιμών των υψών στα δείγματα να είναι μεταξύ 1cm και 2cm.

Λύση

Επειδή η κατανομή των υψών είναι κανονική ισχύουν οι προϋποθέσεις του Κεντρικού Οριακού Θεωρήματος και η δειγματική κατανομή της διαφοράς των μέσων τιμών μπορεί να μελετηθεί με βάση την κανονική κατανομή δηλαδή:

$$\bar{x}_A - \bar{x}_B \sim N\left(\mu_A - \mu_B, \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}\right)$$

Οπότε $P(1 \le \bar{x}_A - \bar{x}_B \le 2) =$

$$= P\left[\frac{1-(170-169)}{\sqrt{\frac{3}{60}+\frac{2}{50}}} \le \frac{(\bar{x}_A - \bar{x}_B)-(\mu_A-\mu_B)}{\sqrt{\frac{\sigma_A^2}{n_A}+\frac{\sigma_B^2}{n_B}}} \le \frac{2-(170-169)}{\sqrt{\frac{3}{60}+\frac{2}{50}}}\right] =$$

$$= P\left[\frac{1-1}{\sqrt{0.09}} \le z \le \frac{2-1}{\sqrt{0.09}}\right] =$$

$$= P(0 \le z \le 3{,}03) = P(z \le 3{,}03) - P(z \le 0) =$$

$$\Phi(3{,}03) - \Phi(0) = 0{,}99904 - 0{,}5 = 0{,}49904$$

4.3.5. ΚΑΤΑΝΟΜΗ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ ΔΙΑΦΟΡΑΣ ΤΩΝ ΠΟΣΟΣΤΩΝ

Έστω ότι έχουμε δύο πληθυσμούς Α και Β και οι αντίστοιχες αναλογίες – ποσοστά είναι P_A και P_B. Αν πάρουμε όλα τα τυχαία δείγματα n_A και n_B από τους εκάστοτε πληθυσμούς Α και Β και υπολογίσουμε τα ποσοστά P_A και P_B και την διαφορά τους P_A - P_B , τότε θα προκύψουν τόσες διαφορές όσες και τα ζεύγη των δειγμάτων. Αν τις διαφορές αυτές τις κατατάξουμε σε μια κατανομή συχνοτήτων, η συγκεκριμένη κατανομή καλείται κατανομή δειγματοληψίας διαφορών δύο ποσοστών και περιλαμβάνει τα παρακάτω χαρακτηριστικά:

Αν τα n_A , n_B είναι μεγάλος αριθμός τότε η δειγματική κατανομή ακολουθεί κατά προσέγγιση την κανονική.

$$\mu_{p_A - p_B} = P_A - P_B$$

$$\sigma^2_{p_A - p_B} = \frac{P_A(1-P_A)}{n_A} + \frac{P_B(1-P_B)}{n_B}$$

$$P_A - P_B \sim N\left[P_A - P_B, \frac{P_A(1-P_A)}{n_A} + \frac{P_B(1-P_B)}{n_B}\right]$$

Παράδειγμα

Τα ποσοστά των ψηφοφόρων που στηρίζουν ένα κόμμα είναι 35% και 30% για τα αστικά κέντρα και τις αγροτικές περιοχές αντίστοιχα. Πήραμε 1000 ψηφοφόρους από τα αστικά κέντρα και 800 από τις αγροτικές περιοχές. Ποια είναι η πιθανότητα η διαφορά ποσοστών των δύο δείγμα των να είναι τουλάχιστον διπλάσια από τη διαφορά των ποσοστών των πληθυσμών;

Λύση

$\mu_{P_A - P_B} = P_A - P_B$ = 0,35 – 0,30 = 0,05

$$\sigma^2_{P_A - P_B} = \frac{P_A(1-P_A)}{n_A} + \frac{P_B(1-P_B)}{n_B} = \frac{0,35(1-0,35)}{1000} + \frac{0,30(1-0,30)}{800} = \frac{0,2275}{1000} + \frac{0,21}{800} =$$

$$0,0002275 + 0,0002625 = 0,00049$$

=

Άρα $\sigma_{P_A - P_B}$ = 0,022

Ζητάμε την πιθανότητα

$$P[P_A - P_B \geq 2 \cdot (0.35 - 0.30)] = P[P_A - P_B \geq 0.1] =$$

$$P\left[\frac{(P_A - P_B) - \mu_{P_A - P_B}}{\sigma_{P_A - P_B}} \geq \frac{0.1 - 0.05}{0.022}\right] = P(z \geq 2.27) = 1 - P(z \geq 2.27) =$$

$$1 - \Phi(2,27) = 1 - 0,98840 = 0,0116$$

4.3.6. Η ΔΕΙΓΜΑΤΙΚΗ ΚΑΤΑΝΟΜΗ ΤΟΥ ΛΟΓΟΥ ΔΥΟ ΔΙΑΣΠΟΡΩΝ

Σε περίπτωση που θέλουμε να συγκρίνουμε τις διασπορές δύο πληθυσμών χρησιμοποιούμε τον λόγο των διασπορών τόσο των τυχαίων όσο και των ανεξάρτητων δειγμάτων τα οποία προκύπτουν από την επιλογή των πληθυσμών Α και Β αντιστοίχως.

Έστω ότι έχουμε δύο πληθυσμούς Α και Β που ο καθένας ακολουθεί κανονική κατανομή με διακύμανση σ_A^2 και σ_B^2 αντίστοιχα. Από τους παραπάνω πληθυσμούς επιλέγουμε τυχαία και ανεξάρτητα δείγματα μεγέθους n_A και n_B, ενώ οι διακυμάνσεις τους αντιστοιχούν σε s_A^2 και s_B^2.

Τότε η τυχαία μεταβλητή $F = \frac{\frac{s_A^2}{\sigma_A^2}}{\frac{s_B^2}{\sigma_B^2}} = \frac{s_A^2 \cdot \sigma_B^2}{s_B^2 \cdot \sigma_A^2}$ ακολουθεί την F κατανομή με $n_A - 1, n_B - 1$.

Οι συγκεκριμένοι παράμετροι ($n_A - 1, n_B - 1$) αναφέρονται ως βαθμοί ελευθερίας για τον αριθμητή και τον παρανομαστή. Επιπλέον μπορούμε να συνδέσουμε την F κατανομή με την x^2 . Αν οι μεταβλητές X και Y ακολουθούν την x^2 κατανομή τότε η μεταβλητή $F = \frac{\frac{x}{n_1}}{\frac{y}{n_2}}$ ακολουθεί την F κατανομή με n_1 , n_2 βαθμούς ελευθερίας.

Παράδειγμα

Έχει διαπιστωθεί ότι η διακύμανση στον αριθμό ατυχημάτων που προκαλούν οι άντρες οδηγοί είναι διπλάσια της αντίστοιχης διακύμανσης των γυναικών. Αν πάρουμε τυχαία δείγματα από $n_A = 61$ άντρες και $n_Γ = 121$ γυναίκες. Ποία είναι η πιθανότητα η διακύμανση στο δείγμα των αντρών να είναι μικρότερη από το τριπλάσιό της διακύμανσης στο δείγμα των γυναικών;

Λύση

Από την εκφώνηση γνωρίζουμε ότι:

$\sigma_A^2 = 2\sigma_Γ^2$

s_A^2 ,$s_Γ^2$ οι διακυμάνσεις των δύο δειγμάτων

Ζητάμε την πιθανότητα:

$$P(s_A^2 < 3s_Γ^2) = P\left(\frac{s_A^2}{s_Γ^2} < 3\right) = P\left(\frac{\frac{s_A^2}{\sigma_A^2}}{\frac{s_Γ^2}{\sigma_A^2}} < 3\right) = P\left(\frac{\frac{s_A^2}{\sigma_A^2}}{\frac{s_Γ^2}{2\sigma_Γ^2}} < 3\right) = P\left(\frac{\frac{s_A^2}{\sigma_A^2}}{\frac{s_Γ^2}{\sigma_Γ^2}} < \frac{3}{2}\right)$$

$$= P\left(F_{n_A-1, n_B-1} < 1{,}5\right) = P(F_{60,120} < 1{,}5)$$

Με βάση τους πίνακες για βαθμούς ελευθερίας 60 και 120 έχουμε 1,5 κοντά στο 1,53 και P = 0,975.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- ΣΤΑΤΙΣΤΙΚΕΣ ΜΕΘΟΔΟΙ ΤΕΥΧΟΣ 1 ΕΚΔΟΣΕΙΣ ΣΥΜΜΕΤΡΙΑ (ΠΑΤΡΑ 1999) Ι.Α. ΚΟΥΤΡΟΥΒΕΛΗ
- ΟΡΓΑΝΩΣΗ ΚΑΙ ΔΙΕΞΑΓΩΓΗ ΔΕΙΓΜΑΤΟΛΗΠΤΙΚΩΝ ΕΡΕΥΝΩΝ ΠΑΝΑΓΙΩΤΗ Θ.ΤΖΟΡΤΖΟΠΟΥΛΟΥ ΑΘΗΝΑ (2004 – 2005)
- ΕΙΣΑΓΩΓΗ ΣΤΗ ΣΤΑΤΙΣΤΙΚΗ ΣΚΕΨΗ (ΤΟΜΟΣ 2)ΕΙΣΑΓΩΓΗ ΣΤΙΣ ΠΙΘΑΝΟΤΗΤΕΣ ΚΑΙ ΣΤΗ ΣΤΑΤΙΣΤΙΚΗ ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑ Ι.ΠΑΝΑΡΕΤΟΣ, Ε.ΞΕΚΑΛΑΚΗ
- ΣΤΑΤΙΣΤΙΚΗ ΠΕΤΡΟΣ ΚΙΟΧΟΣ (INTERBOOKS)
- ΕΙΣΑΓΩΓΗ ΣΤΙΣ ΠΙΘΑΝΟΤΗΤΕΣ ΘΕΩΡΙΑ ΚΑΙ ΕΦΑΡΜΟΓΕΣ ΚΟΥΤΡΑΣ
- ΣΤΑΤΙΣΤΙΚΗ – ΜΕΘΟΔΟΙ ΑΝΑΛΥΣΗΣ ΓΙΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΕΣ ΑΠΟΦΑΣΕΙΣ ΧΑΛΙΚΙΑΣ
- ΣΤΑΤΙΣΤΙΚΗ ΕΠΙΧΕΙΡΗΣΕΩΝ – ΣΥΓΧΡΟΝΗ ΕΚΔΟΤΙΚΗ
- ΠΙΘΑΝΟΤΗΤΕΣ ΣΤΟΧΑΣΤΙΚΩΝ ΑΝΕΛΙΞΕΩΝ ($3^{η}$ ΕΚΔΟΣΗ) Ε.ΞΕΚΑΛΑΚΗ ΚΑΙ Ι.ΠΑΝΑΡΕΤΟΥ (1993)
- ΘΕΩΡΙΑ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ ΚΑΙ ΕΦΑΡΜΟΓΕΣ ΧΑΡΙΣΗΣ ΚΩΣΤΑΣ Ι. ΚΑΙ ΚΙΟΧΟΣ ΠΕΤΡΟΣ (INTERBOOKS 1997)