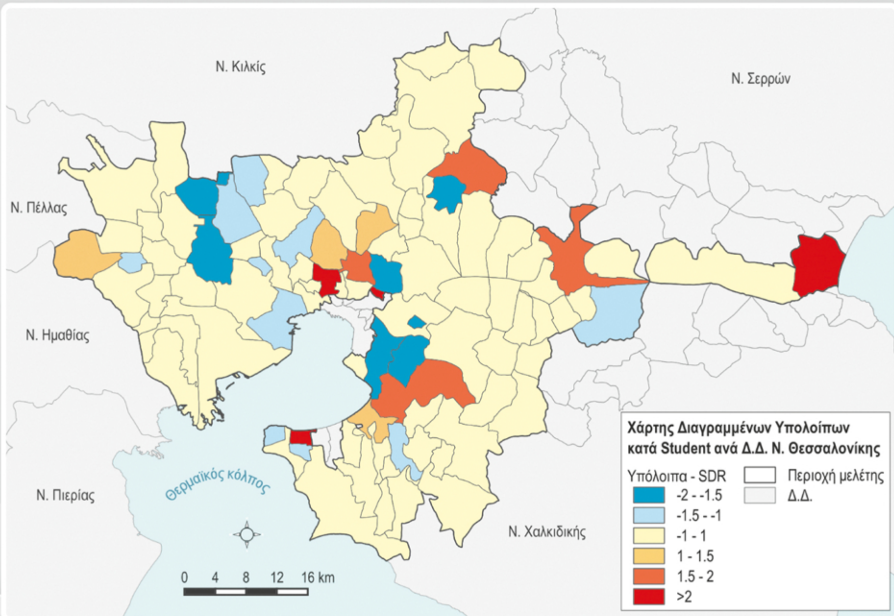
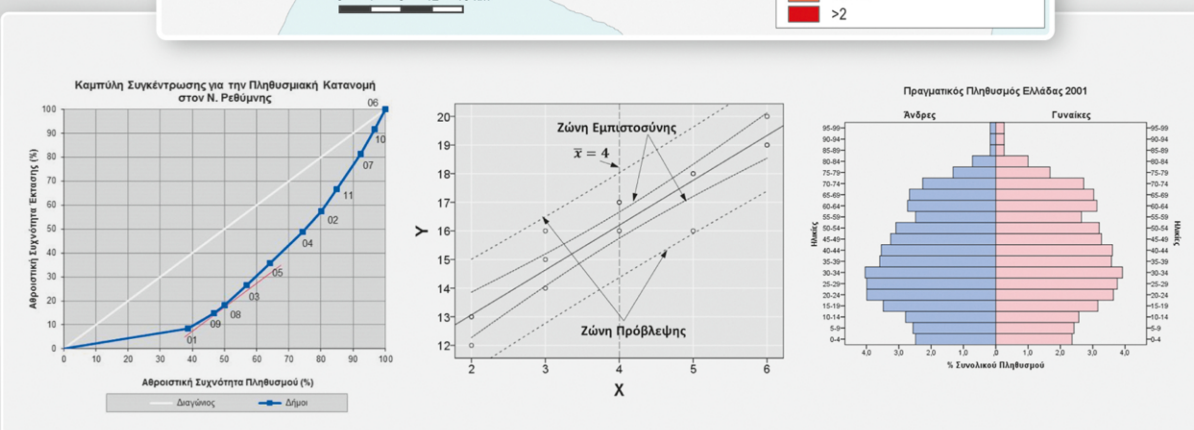
Περιστέρα Ν. Λαφαζάνη

Στατιστικές Μέθοδοι στη Γεωγραφία

θεωρία και Παραδείγματα





ΕΚΔΟΣΕΙΣ ΖΗΤΗ

ISBN 978-960-456-626-6

Φωτοστοιχειοθεσία
Εκτύπωση
Βιβλιοδεσία

Π. ΖΗΤΗ & Σια ΙΚΕ
18ο χλμ Θεσ/νίκης-Περαίας
Τ.Θ. 4171 • Περαία Θεσσαλονίκης • Τ.Κ. 570 19
Τηλ.: 2392.072.222 - Fax: 2392.072.229 • e-mail: info@ziti.gr

ΕΚΔΟΣΕΙΣ ΖΗΤΗ
ziti.gr

ΒΙΒΛΙΟΠΩΛΕΙΑ

ΘΕΣΣΑΛΟΝΙΚΗΣ:
Αρμενοπούλου 27, 546 35 Θεσσαλονίκη
Τηλ.: 2310.203.720, Fax: 2310.211.305
e-mail: sales@ziti.gr

ΑΘΗΝΩΝ:
Χαριλάου Τρικούπη 22, 106 79 Αθήνα
Τηλ.-Fax: 210.3816.650
e-mail: athina@ziti.gr

ΗΛΕΚΤΡΟΝΙΚΟ ΒΙΒΛΙΟΠΩΛΕΙΟ: www.ziti.gr

Στη μνήμη των γονιών μου

Νικολάου και Βασιλικής

Η Περιστέρα Ν. Λαφαζάνη είναι Διδάκτωρ Μηχανικός και Πτυχιούχος Φυσικός. Καθηγήτρια στο Τμήμα Αγρονόμων και Τοπογράφων Μηχανικών της Πολυτεχνικής Σχολής του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης, δίδαξε Γεωγραφία, Περιφερειακή Γεωγραφία, Γεωγραφική Οργάνωση και Ανάλυση του Ελλαδικού Χώρου, Στατιστική Ανάλυση του Χώρου, σε προπτυχιακό και μεταπτυχιακό επίπεδο.

ΠΡΟΛΟΓΟΣ

Ο Γεωγραφικός Χώρος με τα διάφορα φαινόμενα που περιλαμβάνει και εξελίσσονται σε αυτόν, εμφανίζεται «προκλητικός» και συγχρόνως «γοητευτικός» για τους ερευνητές που τον αναλύουν και τον μελετούν. Τα φυσικά, οικονομικά, δημογραφικά, πολιτιστικά, περιβαλλοντικά, κ.ά. φαινόμενα αυτού, δηλαδή τα *γεωγραφικά φαινόμενα*, εκφράζονται ή αποδίδονται από πλήθος χαρακτηριστικών, των *γεωγραφικών χαρακτηριστικών* ή *δεδομένων*, τα οποία προσδίδουν τη φυσιογνωμία αυτού του χώρου με τη δημιουργία των *χωρικών προτύπων*. Τα τελευταία συνεπάγονται διαφορετικές γεωγραφικές «συμπεριφορές» και «εκφάνσεις», οι οποίες καθίστανται σαφείς με τις *γεωγραφικές διαφοροποιήσεις* ή *περιφερειακές ανισότητες* σε μια γεωγραφική περιοχή. Η σαφήνεια των διαφοροποιήσεων και των ιδιαιτεροτήτων που υφίστανται ισοδυναμεί με τη *χαρτογράφηση* αυτών.

Τα χωρικά πρότυπα, με τη σειρά τους, εκφράζονται από πληροφορίες που προέρχονται από τα αρχικά γεωγραφικά δεδομένα. Αυτές οι πληροφορίες, μαζί με τον ιστορικό χρόνο που μεσολαβεί, αποτελούν τα «νέα» δεδομένα και δίνουν τη «νέα» οπτική του γεωγραφικού χώρου με αποτέλεσμα τη δημιουργία νέων χωρικών προτύπων, κ.ο.κ., Πρόκειται για μια συνεχή επαναλαμβανόμενη κίνηση όπου, το φυσικό και ανθρωπογενές περιβάλλον αλληλοεπιδρούν δυναμικά μεταξύ τους με αποτέλεσμα τη συνεχή μεταβολή του γεωγραφικού χώρου. Η μεταβολή αυτή δεν είναι παρά οι διαφοροποιήσεις που επέρχονται στη δομή της φυσικής, κοινωνικής, οικονομικής, πολιτιστικής και περιβαλλοντικής διάστασης αυτού, με την επίδραση και του χρόνου,

Όμως, τα γεωγραφικά δεδομένα συνιστούν ένα πλήθος ποιοτικών και ποσοτικών μεταβλητών, η επεξεργασία των οποίων αποκωδικοποιεί τη μορφή και τη συμπεριφορά των διαφόρων φαινομένων που παρατηρούνται στον γεωγραφικό χώρο. Πολύτιμος αρωγός κατά τη διαδικασία της μελέτης αυτών καθίστανται οι στατιστικές μέθοδοι. Οι μονομεταβλητές και πολυμεταβλητές στατιστικές αναλύσεις παρέχουν σχέσεις, πίνακες και διαγράμματα που βοηθούν στην ερμηνεία και την κατανόηση της συμπεριφοράς των φαινομένων, δηλαδή στην κατανόηση της κατανομής των φαινομένων. Επί πλέον, εφ’ όσον πρόκειται για συγκεκριμένο γεωγραφικό χώρο, με τη δημιουργία των σχετικών χαρτών εντοπίζονται τάσεις, συμπεριφορές και ιδιαιτερότητες των φαινομένων που ερευνώνται, όπως έχει αναφερθεί. Με άλλα λόγια, αναπαρίστανται τα χωρικά πρότυπα που δημιουργούνται από την ύπαρξη και την επίδραση αυτών των φαινομένων. Έτσι, είναι δυνατή η παρατήρηση της δυναμικής εξέλιξης ενός γεωγραφικού φαινομένου, η λήψη και η εφαρμογή διαφόρων μέτρων με σκοπό τη δυναμική ισορροπία του χώρου, γεγονός που σημαίνει μείωση των υφιστάμενων γεωγραφικών ανισοτήτων.

Οι στατιστικές επεξεργασίες που επιτελούνται έχουν ως στόχο τη δημιουργία ενός τεκμηριωμένου «γεωγραφικού υποβάθρου» έτσι ώστε, να είναι δυνατός τόσο ο εντοπισμός των τάσεων και των συμπεριφορών των φαινομένων στον χρόνο αναφοράς τους, όσο και η πρόβλεψη μελλοντικών χωρικών διαφοροποιήσεων. Έχοντας υπ’ όψη τα αναφερόμενα, η διάρθρωση του βιβλίου είναι τέτοια ώστε ο αναγνώστης – ερευνητής να μπορεί να κατανοήσει και να προσεγγίσει τη δυναμική των στατιστικών εννοιών και μεθόδων και να τις χρησιμοποιήσει με τρόπο «απλό» και «φυσικό». Αυτό που τονίζεται είναι η λογική των διαφόρων μεθόδων και η δυνατότητα εφαρμογής τους και όχι οι μαθηματικές αποδείξεις αυτών (εκτός ελαχίστων περιπτώσεων). Σε κάθε κεφάλαιο υπάρχει πλήθος παραδειγμάτων στα οποία, με αναλυτικό τρόπο (σχέσεις, πίνακες, διαγράμματα, σχήματα), καταγράφεται η εφαρμογή των αντίστοιχων στατιστικών εννοιών και μεθόδων. Το βιβλίο αυτό προέκυψε μετά από πολύχρονη έρευνα και διδασκαλία στο Τμήμα Αγρονόμων και Τοπογράφων Μηχανικών του Α.Π.Θ., σε προπτυχιακό και μεταπτυχιακό επίπεδο. Απευθύνεται σε γεωγράφους, τοπογράφους μηχανικούς, πολιτικούς μηχανικούς, χωροτάκτες, πολεοδόμους, περιβαλλοντολόγους, καθώς και σε άλλους επιστήμονες των φυσικών και κοινωνικών επιστημών, όπως και στους προπτυχιακούς και μεταπτυχιακούς φοιτητές όλων των επιστημονικών κλάδων.

Περιλαμβάνονται έντεκα κεφάλαια και ένα παράρτημα.

Το 1ο Κεφάλαιο αναφέρεται σε θέματα της περιγραφικής στατιστικής όπως: Βασικές έννοιες και ορισμοί, γραφική απεικόνιση των δεδομένων, χαρακτηριστικά ή μεταβλητές, μέτρα κεντρικής τάσης, μέτρα διασποράς, μέτρα μορφής, μέτρα συγκέντρωσης ή μέτρα ανισότητας, ακραίες ή ασυνήθιστες τιμές (οντότητες). Η κατανόηση αυτών δίνει τη δυνατότητα στον ερευνητή, με τη βοήθεια απλών στατιστικών αναλύσεων, να απεικονίσει κατανομές και να υπολογίσει σημαντικούς δείκτες (π.χ. νεότητας, γήρανσης, ανεργίας, συγκέντρωσης, εγκατάστασης, ειδίκευσης) που αφορούν τη συμπεριφορά των γεωγραφικών φαινομένων. Ειδικότερα, η χαρτογράφηση ενός δείκτη (ο οποίος είναι μία τυχαία μεταβλητή) οπτικοποιεί την κατανομή του στη γεωγραφική περιοχή αναφοράς του. Αυτό έχει ως αποτέλεσμα την τεκμηριωμένη αντίληψη της συμπεριφοράς του ή της διακύμανσής του τόσο σε συγκεκριμένο χώρο και χρόνο, όσο και διαπεριφερειακά και διαχρονικά.

Το 2ο Κεφάλαιο περιλαμβάνει βασικά θέματα της θεωρίας πιθανοτήτων. Γίνεται αναφορά: στα τυχαία φαινόμενα, στον δειγματικό χώρο, στον ορισμό της πιθανότητας, στους κανόνες που αφορούν τον λογισμό πιθανοτήτων, σε στοιχεία της συνδυαστικής, στη δειγματοληψία. Με το κεφάλαιο αυτό δίνεται η δυνατότητα για την καλύτερη κατανόηση, κυρίως, των εννοιών και των θεωρητικών κατανομών συχνοτήτων του 4ου κεφαλαίου, καθώς και των μεθόδων οι οποίες αφορούν την εκτίμηση και τη σημαντικότητα των παραμέτρων ενός πληθυσμού.

Το 3ο Κεφάλαιο αφορά το δείγμα και τον σχηματισμό του. Αναφέρονται: η εκτίμηση για το μέγεθος του δείγματος (ποσοτικά και ποιοτικά χαρακτηριστικά) και οι διαδικασίες δειγματοληψίας. Διακρίνονται οι κλασικές μέθοδοι δειγματοληψίας (πληθυσμός ή μη χωρικό δειγματοληπτικό υπόβαθρο) και οι μέθοδοι χωρικής δειγματοληψίας (χωρικό δειγματοληπτικό υπόβαθρο ή χαρτογραφικό υπόβαθρο).

Το 4ο Κεφάλαιο αναφέρεται στις τυχαίες μεταβλητές και στις θεωρητικές κατανομές αυτών ή κατανομές πιθανότητας. Περιγράφονται: οι διακριτές τυχαίες μεταβλητές και οι αντίστοιχες κατανομές τους (κατανομή Bernoulli, διωνυμική κατανομή, κατανομή Poisson, γεωμετρική κατανομή, αρνητική διωνυμική κατανομή ή καταναομή Pascal, υπεργεωμετρική κατανομή, ομοιόμορφη διακριτή κατανομή), οι συνεχείς τυχαίες μεταβλητές και οι κατανομές τους (ομοιόμορφη συνεχής κατανομή, κατανομή Γάμμα, κατανομή Βήτα, εκθετική κατανομή, κανονική κατανομή, κατανομή X_ν^2, κατανομή t_ν, κατανομή F_{ν_1,ν_2}), οι διδιάστατες τυχαίες μεταβλητές, οι πολυδιάστατες τυχαίες μεταβλητές. Το κεφάλαιο αυτό δίνει με απλό και κατανοητό τρόπο το θεωρητικό πλαίσιο ώστε από το διαθέσιμο δείγμα να είναι δυνατή η εξαγωγή συμπερασμάτων για την κατανομή του αντίστοιχου ή των αντίστοιχων πληθυσμών (φαινομένων) και για την πιθανή σχέση μεταξύ δύο τυχαίων μεταβλητών, οι οποίες είτε περιγράφουν δύο διαφορετικά φαινόμενα είτε αναφέρονται στο ίδιο φαινόμενο.

Το 5ο Κεφάλαιο αφορά τον προσδιορισμό των μεταβλητών και περιλαμβάνει: τη μέτρηση των μεταβλητών, τις κλίμακες μέτρησης, τα είδη ή τους τύπους των μεταβλητών, την «αποκωδικοποίηση» των πληροφοριών και την κωδικοποίηση των μεταβλητών. Οι απαραίτητες διαδικασίες για τη δημιουργία του πληροφοριακού υποβάθρου στη γεωγραφική έρευνα εδράζονται σε αυτό το κεφάλαιο.

Το 6ο Κεφάλαιο περιγράφει τις τεχνικές για τη συλλογή γεωγραφικών δεδομένων. Αναφέρονται: οι εξαντλητικές έρευνες (απογραφές, συνεχείς καταγραφές), οι δειγματοληπτικές έρευνες, η επιτόπια έρευνα και οι μέθοδοι επιτόπιας έρευνας οι οποίες αφορούν τη δημιουργία ενός αποτελεσματικού ερωτηματολογίου για τη συλλογή των απαιτούμενων δεδομένων. Η λειτουργικότητα του ερωτηματολογίου είναι από τα πλέον βασικά βήματα, αν όχι το βασικότερο, κατά την προετοιμασία της επιτόπιας έρευνας.

Στο 7ο Κεφάλαιο περιγράφονται οι μέθοδοι που χρησιμοποιούνται για την εκτίμηση των παραμέτρων ενός πληθυσμού, από το διαθέσιμο δείγμα, για ποσοτικές και ποιοτικές μεταβλητές. Ειδικότερα, αναφέρονται τα διαστήματα εμπιστοσύνης για τη μέση τιμή, τη διακύμανση και τη συχνότητα /αναλογία της διωνυμικής κατανομής, καθώς και τα διαστήματα εμπιστοσύνης που αφορούν: τη διαφορά των μέσων τιμών δύο πληθυσμών (δείγματα ανεξάρτητα και δείγματα εξαρτημένα), τον λόγο των διακυμάνσεων δύο κανονικών πληθυσμών, τη διαφορά των συχνοτήτων / αναλογιών δύο διωνυμικών πληθυσμών. Η κατανόηση των μεθόδων καθιστά ικανό τον ερευνητή να αναλύσει φαινόμενα ή να αντιμετωπίσει προβλήματα, σε μια γεωγραφική περιοχή, με σημαντικό βαθμό επιτυχίας.

Το 8ο Κεφάλαιο αναφέρεται στις δοκιμασίες σημαντικότητας Ι ή στους παραμετρικούς στατιστικούς ελέγχους. Εδώ, όπως και στο επόμενο κεφάλαιο, χρησιμοποιούνται οι έννοιες της Επαγωγικής Στατιστικής με στόχο την αξιολόγηση των παραμέτρων που εκτιμήθηκαν από το διαθέσιμο δείγμα. Περιγράφονται: οι δοκιμασίες συμφωνίας για την πληθυσμιακή μέση τιμή, τη διακύμανση και τη συχνότητα /αναλογία της διωνυμικής κατανομής, οι δοκιμασίες ομογένειας για την πληθυσμιακή μέση τιμή (δείγματα ανεξάρτητα και δείγματα εξαρτημένα), για την πληθυσμιακή διακύμανση, για την πληθυσμιακή συχνότητα / αναλογία, ο προσδιορισμός του ανώτατου ορίου του σφάλματος εκτίμησης για τη μέση τιμή και τη συχνότητα /αναλογία, ο προσδιορισμός του δειγματικού μεγέθους (ποσοτικές και ποιοτικές μεταβλητές· σχετική αναφορά έχει γίνει και στο 3ο Κεφάλαιο).

Στο 9ο Κεφάλαιο περιγράφονται οι δοκιμασίες σημαντικότητας ΙΙ ή οι μη παραμετρικοί στατιστικοί έλεγχοι. Πρόκειται για μεθόδους που αφορούν: δείγματα ποσοτικών μεταβλητών με άγνωστη πληθυσμιακή κατανομή, δείγματα ποιοτικών μεταβλητών, δείγματα πολύ μικρού μεγέθους. Αναφέρονται:

οι δοκιμασίες καλής προσαρμογής (δοκιμασία X^2 και δοκιμασία Kolmogorov – Smirnov), οι δοκιμασίες ανεξαρτησίας με βάση τις συχνότητες (δοκιμασία X^2), οι δοκιμασίες ομογένειας με βάση τις συχνότητες (δοκιμασία X^2 και δοκιμασία Kolmogorov – Smirnov), οι δοκιμασίες ομογένειας με βάση τα αθροίσματα τάξεων για δύο ανεξάρτητα δείγματα (δοκιμασία Mann – Whitney, δοκιμασία Robust Rank – Order) και για k – ανεξάρτητα δείγματα (δοκιμασία Kruskal – Wallis), οι δοκιμασίες ομογένειας με βάση την τυχαιότητα των δειγμάτων (δοκιμασία των ροών, δοκιμασία Wald – Wolfowitz), οι δοκιμασίες συμφωνίας με βάση τα πρόσημα των διαφορών (προσημική δοκιμασία ή Sign test, δοκιμασία Wilcoxon ή Wilcoxon Rank – sum test), οι δοκιμασίες ομογένειας με βάση τη διάμεσο για δύο ανεξάρτητα δείγματα (δοκιμασία διαμέσου ή δοκιμασία Mood) και για k – ανεξάρτητα δείγματα (επέκταση της δοκιμασίας διαμέσου), οι δοκιμασίες ομογένειας για δείγματα με ζευγαρωτές παρατηρήσεις με βάση τα πρόσημα των διαφορών (προσημική δοκιμασία ή Sign test, δοκιμασία Wilcoxon ή Wilcoxon Signed – ranks test), οι δοκιμασίες ομογένειας για δείγματα με ζευγαρωτές παρατηρήσεις με βάση τη μεταβολή συμπεριφοράς των στατιστικών ατόμων (δοκιμασία McNemar για δίτιμες ονομαστικές μεταβλητές, δοκιμασία McNemar για πολυθεματικές ονομαστικές μεταβλητές ή δοκιμασία Marginal Homogeneity), οι δοκιμασίες ομογένειας για k συσχετισμένα δείγματα (δοκιμασία Friedman ή the Friedman two – way analysis of variance by ranks, η δοκιμασία Qochran Q). Πρόκειται για τις πλέον αξιόλογες μη παραμετρικές δοκιμασίες σημαντικότητας με πολλές εφαρμογές σε γεωγραφικά φαινόμενα.

Το 10ο Κεφάλαιο αναφέρεται στην παλινδρόμηση και τη συσχέτιση, δηλαδή πραγματεύεται τη στοχαστική εξάρτηση μεταξύ δύο ή περισσότερων μεταβλητών ή πώς συνδέονται μεταξύ τους δύο ή περισσότερα γεωγραφικά φαινόμενα. Αναλύεται η απλή γραμμική παλινδρόμηση και δίνονται οι ερμηνείες των συντελεστών συσχέτισης, προσδιορισμού, διορθωμένου ή προσαρμοσμένου προσδιορισμού, καθώς και της διακύμανσης των υπολοίπων ή σφαλμάτων. Τονίζεται η σπουδαιότητα αυτών στη γεωγραφική ανάλυση και περιγράφονται οι υποθέσεις και οι στατιστικοί έλεγχοι για το απλό παλινδρομικό μοντέλο.

Η ανάλυση της πολλαπλής παλινδρόμησης περιλαμβάνει την ανάλυση της μερικής συσχέτισης και της πολλαπλής γραμμικής συσχέτισης. Υπολογίζονται και ερμηνεύονται οι συντελεστές μερικής, τμηματικής και πολλαπλής συσχέτισης, ο συντελεστής προσδιορισμού, ο συντελεστής διορθωμένου ή προσαρμοσμένου προσδιορισμού. Ερμηνεύονται οι μερικοί συντελεστές παλινδρόμησης καθώς και οι τυποποιημένοι μερικοί συντελεστές παλινδρόμησης. Επισημαίνεται, και εδώ, η σημαντικότητα της ανάλυσης των υπολοίπων προκειμένου να αξιολογηθεί η μέθοδος που εφαρμόσθηκε.

Προς τούτο, η χαρτογράφηση αυτών αναδεικνύει τα χωρικά πρότυπα που σχηματίζονται με την εφαρμογή του μοντέλου της απλής ή πολλαπλής παλινδρόμησης και πόσο σύμφωνη είναι η κατανομή τους με την πραγματική κατάσταση της περιοχής. Η εμφάνιση των χωρικών προτύπων, υφιστάμενων ή αγνώστων, δίνει τη δυνατότητα της τεκμηριωμένης ερμηνείας αυτών και συνακόλουθα της ερμηνείας του φαινομένου ή των φαινομένων που εξελίσσεται ή εξελίσσονται στον γεωγραφικό χώρο αναφοράς.

Περιγράφονται οι υποθέσεις και οι στατιστικοί έλεγχοι που αφορούν το μοντέλο πολλαπλής γραμμικής παλινδρόμησης, όπως: ο έλεγχος για την αξιοπιστία ολόκληρου του μοντέλου (γενικά), ο έλεγχος για τις παραμέτρους / τους συντελεστές παλινδρόμησης, οι έλεγχοι που αφορούν τις τιμές της τυχαίας εξαρτημένης μεταβλητής (για την προβλεπόμενη μέση τιμή και για την προβλεπόμενη εκτιμώμενη τιμή). Βεβαίως, παρόμοιοι έλεγχοι αναφέρονται και στην απλή παλινδρόμηση. Περιγράφονται βασικές σημαντικότατες έννοιες κατά τη δόμηση ενός παλινδρομικού μοντέλου όπως: η συγγραμμικότητα ή πολυσυγγραμμικότητα, το όριο ανοχής, η ετεροσκεδαστικότητα, η αυτοσυσχέτιση και η ύπαρξη ακραίων οντοτήτων και οντοτήτων επίδρασης / επιρροής.

Καταγράφονται οι διάφορες διαδικασίες ελέγχου, για την ικανοποίηση των υποθέσεων που αφορούν την ανάλυση γραμμικής παλινδρόμησης (απλής και πολλαπλής), με τη βοήθεια γενικών και ειδικών διαγραμμάτων. Αναφέρονται μέθοδοι με τις οποίες αντιμετωπίζονται οι παραβιάσεις των υποθέσεων που αφορούν ένα αξιόπιστο γραμμικό παλινδρομικό μοντέλο, καθώς και διαδικασίες απαραίτητες για την επικύρωση του «βέλτιστου» μοντέλου, το οποίο έχει επιλεγεί ως το καλύτερο από ένα πλήθος υποψηφίων μοντέλων. Οι διαδικασίες αυτές είναι: η συμμετοχή ονομαστικών μεταβλητών και το «καταλληλότερο» πλήθος ανεξάρτητων μεταβλητών (μέγεθος δείγματος και πλήθος μεταβλητών, το κριτήριο Mallows C_p, τα κριτήρια AIC, SBC, BIC, APC). Επιπροσθέτως, αναφέρονται οι τεχνικές, που συνήθως εφαρμόζονται, προκειμένου να επικυρωθεί το «βέλτιστο» μοντέλο, έτσι ώστε να επιβεβαιώνεται ακόμη καλύτερα η αξιοπιστία του (σταθερότητα και ακρίβεια πρόβλεψης).

Το μοντέλο παλινδρόμησης αποτελεί ένα αξιόλογο και αξιόπιστο εργαλείο στη γεωγραφική έρευνα και εφαρμόζεται ευρύτατα, τόσο στην ερμηνεία των φαινομένων όσο και στην πρόβλεψη των τιμών ενός φαινομένου, δηλαδή στην πρόβλεψη της μελλοντικής «συμπεριφοράς» του. Οι ιδιαιτερότητες του γεωγραφικού χώρου είναι πιθανό να επιδρούν αρνητικά στην αξιοπιστία ενός «καθολικού» παλινδρομικού μοντέλου. Προς τούτο, εφ' όσον κρίνεται αναγκαίο από τον ερευνητή, εφαρμόζεται η γεωγραφικά σταθμισμένη παλινδρόμηση. Η μέθοδος αυτή, μαζί με άλλες μεθόδους (Λογιστική παλινδρόμηση, Παραγοντική ανάλυση, Ανάλυση κατά συστάδες, Διακριτική ανάλυση, Χρονοσειρές) θα αναπτυχθεί στο Β' μέρος αυτού του πονήματος.

Εκτός από τις σχέσεις μεταξύ ποσοτικών μεταβλητών, υφίστανται σχέσεις και μεταξύ των ποιοτικών μεταβλητών που αποδίδονται με τους *μη παραμετρικούς συντελεστές συσχέτισης*. Έτσι, περιγράφονται: οι συντελεστές συσχέτισης τάξεων → Spearman r_s, Kendall τ, Gamma G, Sommer d, Kendall W (συντελεστής συμφωνίας), οι συντελεστές συσχέτισης ονομαστικών μεταβλητών → αυτοί που βασίζονται σε πίνακες συνάφειας και στη δοκιμασία X^2, οι συντελεστές → Goodman – Kruskal τ (τ_b, τ_a), Lamda λ $(\lambda_Y, \lambda_X, \lambda_{sym})$, αβεβαιότητας U (U_Y, U_X, U_{sym}), ο συντελεστής συμφωνίας Kappa κ του Cohen (συσχέτιση μεταξύ δύο ισομεγέθων στατιστικών συνόλων) και οι συντελεστές συσχέτισης «μικτών» μεταβλητών → Bi-serial r_{bis}, eta η (η_X, η_Y).

Είναι φανερό ότι στη γεωγραφική έρευνα κοινωνικών φαινομένων, οι αναφερόμενοι συντελεστές πιστοποιούν την ύπαρξη σχέσεων σημαντικών ή όχι μεταξύ αυτών, καθώς και πιθανές συσχετίσεις με άλλα φαινόμενα. Αυτό έχει ως αποτέλεσμα τη διευκόλυνση της ερμηνείας υφιστάμενων ή αγνώστων χωρικών προτύπων.

Το 11ο Κεφάλαιο πραγματεύεται την ανάλυση διακύμανσης (anova), η οποία αφορά τη σύγκριση των μέσων τιμών σε περισσότερα των δύο τυχαία και ανεξάρτητα δείγματα. Περιγράφεται η λογική αυτής της μεθόδου όταν η ανάλυση αναφέρεται: σε ένα παράγοντα, σε δύο ανεξάρτητους παράγοντες (μια οντότητα σε κάθε κελί), σε δύο παράγοντες που αλληλοεπιδρούν ($r > 1$ οντότητες σε κάθε κελί), σε τρείς παράγοντες που αλληλοεπιδρούν ($r > 1$ οντότητες σε κάθε συνδυασμό των παραγόντων). Επίσης, αναφέρονται κάποιες μέθοδοι που αφορούν τη σημαντικότητα των μέσων τιμών όταν έχει εφαρμοσθεί ανάλυση διακύμανσης με ένα παράγοντα, δηλαδή όταν απαιτούνται πολλαπλές συγκρίσεις. Προς τούτο, περιγράφονται: η μέθοδος της ελάχιστης σημαντικής διαφοράς - $L.S.D.$, η μέθοδος Bonferroni, η μέθοδος Tukey W, και η μέθοδος Scheffé S.

Το Παράρτημα περιλαμβάνει στατιστικούς πίνακες που αφορούν τους παραμετρικούς και μη παραμετρικούς ελέγχους.

Το Παράρτημα περιέχεται στο ψηφιακό αρχείο Lafazani-Appendix.pdf, το οποίο μπορείτε να προσπελάσετε στον σύνδεσμο tinyurl.com/Lafazani ή μέσω του ακόλουθου QRcode:

Θέλω να εκφράσω τις θερμές ευχαριστίες μου στην εξαιρετική συνάδελφο και φίλη Αγγελική Τσορλίνη για την πολύτιμη βοήθεια που μου προσέφερε κατά την εκπόνηση αυτού του συγγράμματος, καθώς και στον αγαπητό συνάδελφο και φίλο Μύρωνα Μυρίδη για τις χρήσιμες συμβουλές του.

Ιούλιος 2024

Περιστέρα Ν. Λαφαζάνη

Καθηγήτρια Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης

ΠΕΡΙΕΧΟΜΕΝΑ

Το Παράρτημα περιέχεται στο ψηφιακό αρχείο Lafazani-Appendix.pdf, το οποίο μπορείτε να προσπελάσετε στον σύνδεσμο tinyurl.com/Lafazani ή μέσω του QRcode:

1.

ΘΕΜΑΤΑ ΠΕΡΙΓΡΑΦΙΚΗΣ ΣΤΑΤΙΣΤΙΚΗΣ

1.1. ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ ΚΑΙ ΟΡΙΣΜΟΙ

Στατιστικό σύνολο ονομάζεται κάθε σύνολο ατόμων, αντικειμένων, περιπτώσεων, χωρικών αναφορών και οποιωνδήποτε άλλων οντοτήτων που μελετώνται με *στατιστικές μεθόδους*. Δηλαδή, πρόκειται για *σύνολο οντοτήτων που έχουν ένα ή περισσότερα κοινά χαρακτηριστικά*. Κάθε οντότητα του στατιστικού συνόλου ονομάζεται **στατιστικό στοιχείο** ή **στοιχείο** ή **άτομο** ή **μονάδα**.

Διακρίνονται δύο τύποι στατιστικών συνόλων: ο *πληθυσμός* και το *δείγμα*.

- **Πληθυσμός** είναι το σύνολο **όλων** των στατιστικών στοιχείων που έχουν συγκεκριμένη αναφορά (χωρική ή άλλη). Για παράδειγμα, το σύνολο των (πεδινών, ημιορεινών, ορεινών) οικισμών μιας γεωγραφικής περιοχής αποτελεί τον πληθυσμό των οικισμών αυτής της περιοχής. Αυτό σημαίνει ότι το παραπάνω σύνολο μπορεί να διακριθεί ως: «Ο Πληθυσμός των Οικισμών της Ελλάδας», «ο Πληθυσμός των Οικισμών της Κεντρικής Μακεδονίας», «ο Πληθυσμός των Οικισμών του νομού Θεσσαλονίκης» κ.ά.

- **Δείγμα** ονομάζεται οποιοδήποτε υποσύνολο που *λαμβάνεται τυχαία* από τον πληθυσμό. π.χ. ένα υποσύνολο των (πεδινών, ημιορεινών, ορεινών) οικισμών της Χώρας ή της Κεντρικής Μακεδονίας ή του νομού Θεσσαλονίκης, αποτελεί δείγμα των αντίστοιχων πληθυσμών.

Το πλήθος των στοιχείων ενός στατιστικού συνόλου δίνει το **μέγεθος του στατιστικού συνόλου** το οποίο, με τη σειρά του, διακρίνεται σε *μέγεθος πληθυσμού* και *μέγεθος δείγμα-*

τος. Συνήθως, οι υπό μελέτη πληθυσμοί έχουν πεπερασμένο πλήθος στοιχείων· πρόκειται, δηλαδή, για **πεπερασμένους πληθυσμούς**.

Οι διάφορες σταθερές στατιστικές ποσότητες που χρησιμοποιούνται για να περιγράψουν τον πληθυσμό ονομάζονται **παράμετροι**. Πρόκειται για αριθμητικά περιγραφικά μέτρα (μεγέθη) τα οποία υπολογίζονται από τον πληθυσμό. Όταν οι σταθερές αυτές ποσότητες αναφέρονται στο δείγμα λέγονται **στατιστικά**. Δηλαδή, *το στατιστικό για ένα δείγμα είναι ό,τι η παράμετρος για τον πληθυσμό.*

Συνήθως, *τα στατιστικά του δείγματος χρησιμοποιούνται για την εκτίμηση των παραμέτρων του πληθυσμού,* κατά το διάγραμμα[1]:

$$\begin{tikzcd}
\text{Πληθυσμός} \arrow[d, "\text{Επιλογή}"'] & \text{Παράμετροι} \arrow[l] \\
\text{Δείγμα} \arrow[r, "\text{Υπολογισμός}"] & \text{Στατιστικά} \arrow[u, "\text{Εκτίμηση}"']
\end{tikzcd}$$

Σύμφωνα με τα παραπάνω, προσδιορίζεται ο πληθυσμός, επιλέγεται ένα δείγμα τυχαία, υπολογίζονται τα στατιστικά που ενδιαφέρουν και στη συνέχεια χρησιμοποιούνται αυτά τα στατιστικά για να εκτιμηθούν οι αντίστοιχες παράμετροι του πληθυσμού.

Οι παράμετροι και τα στατιστικά αυτά, συνήθως, είναι:

- *Η μέση τιμή, η διακύμανση, η τυπική απόκλιση* και *η συχνότητα.*

  Για τον *πληθυσμό* οι ποσότητες αυτές συμβολίζονται ως: μ, σ^2, σ, p, αντιστοίχως, και συνήθως έχουν τον επιθετικό προσδιορισμό «θεωρητική» (π.χ. θεωρητική μέση τιμή κ.λ.π.).

  Για το δείγμα τα σχετικά σύμβολα είναι: $\bar{x}$ (ή m)[2], s^2, s, f, αντιστοίχως, και οι σχετικές ποσότητες μπορεί να προσδιορίζονται ως «εμπειρική» ή «παρατηρούμενη» (π.χ. εμπειρική μέση τιμή κ.λ.π.)

❑ ΠΙΝΑΚΑΣ Ή ΚΑΤΑΝΟΜΗ ΣΥΧΝΟΤΗΤΩΝ

Κάθε συλλογή ή πίνακας δεδομένων, δηλαδή κάθε συλλογή ή πίνακας αριθμητικών τιμών που προκύπτουν από μετρήσεις κάποιου χαρακτηριστικού επί των οντοτήτων ενός στατιστικού συνόλου, αποτελεί μία **κατανομή**. Εάν:

i. καθορισθούν (τα) διαστήματα τάξεων ή (οι) τάξεις *ίσου εύρους* έτσι ώστε οι παρατηρούμενες τιμές να προσεγγίζονται από τα διαστήματα αυτά,
ii. *διαταχθούν οι παρατηρούμενες τιμές κατά την αύξουσα ή φθίνουσα σειρά των διαστημάτων τάξης ή των «στρογγυλοποιημένων» τιμών τους* και
iii. καταγραφεί το **πλήθος των εμφανίσεων** ή η **εμφάνιση** κάθε τέτοιας τιμής (τάξη ή διάστημα τάξης ή «στρογγυλοποιημένη» τιμή),

τότε προκύπτει ένας *οργανωμένος πίνακας δεδομένων / τιμών.* Ο πίνακας αυτός λέγεται **Πίνακας Κατανομής Συχνοτήτων**, διότι φαίνεται *η συχνότητα με την οποία παρουσιάζεται κάθε δεδομένο / κάθε παρατηρούμενη τιμή.* Η συχνότητα αυτή πιστοποιείται είτε ως το πλήθος των εμφανίσεων (ή επαναλήψεων ή καταμετρήσεων) κάθε παρατηρούμενης τιμής, είτε ως η αναλογία του πλήθους κάθε παρατηρούμενης τιμής ως προς το συνολικό τους πλήθος.

---
[1] Κ.Β. Μπαγιάτης, 1990, σελ. 61
[2] π.χ. Α.Ι. Αναστασιάδης, 1990, σελ.17

➢ Επομένως, πολλές ή λίγες εμφανίσεις μιας τιμής *ισοδυναμούν* με μεγάλο ή μικρό πλήθος οντοτήτων που εμφανίζουν αυτή την τιμή και συνεπώς, *οι οντότητες του στατιστικού συνόλου χαρακτηρίζονται από τη μεγάλη ή μικρή συχνότητα* με την οποία εμφανίζεται (επαναλαμβάνεται) η συγκεκριμένη τιμή.

Με τη βοήθεια του πίνακα αυτού, **ορίζονται**:

- Η **εμφάνιση** κάθε τιμής x_i, η οποία εκφράζεται από το πλήθος των οντοτήτων που εμφανίζουν την τιμή αυτή. Η εμφάνιση αυτή σημειώνεται ως n_i.

$$n_i = \sum_{i=1}^{i} x_i \tag{1.1.1}$$

- Η **συχνότητα** κάθε τιμής x_i, η οποία εκφράζεται από τον λόγο του πλήθους των εμφανίσεων της τιμής αυτής προς το συνολικό πλήθος των τιμών n, που δεν είναι άλλο από το μέγεθος του στατιστικού συνόλου (δείγμα ή πληθυσμός). Η συχνότητα αυτή δίνεται ως:

$$f_i = \frac{n_i}{n} = \frac{1}{n} \cdot \sum_{i=1}^{i} x_i \tag{1.1.2}$$

  Ο πολλαπλασιασμός της συχνότητας επί 100 δίνει το **ποσοστό συχνότητας** που σημειώνεται ως:

$$f_i(\%) = 100 \cdot \frac{n_1}{n} \tag{1.1.3}$$

- Η **αθροιστική εμφάνιση** κάθε τιμής x_i, η οποία εκφράζεται από το άθροισμα των εμφανίσεων για όλες τις τιμές x_j που είναι μικρότερες ή ίσες από τη θεωρούμενη τιμή. Αυτή η αθροιστική εμφάνιση σημειώνεται ως:

$$N_i = \sum_{j \leq i} n_j \tag{1.1.4}$$

- Η **αθροιστική συχνότητα** κάθε τιμής x_i, η οποία εκφράζεται από τον λόγο της αθροιστικής εμφάνισης της τιμής αυτής προς το συνολικό πλήθος των τιμών. Με άλλη έκφραση, η αθροιστική συχνότητα κάθε τιμής x_i προκύπτει από το άθροισμα των συχνοτήτων όλων των τιμών x_j που είναι μικρότερες ή ίσες από τη θεωρούμενη τιμή. Έτσι, αυτή η αθροιστική συχνότητα σημειώνεται ως:

$$F_i = \frac{N_i}{n} = \frac{1}{n} \cdot \sum_{j \leq i} n_j = \sum_{j \leq i} f_j \tag{1.1.5}$$

  Ο πολλαπλασιασμός αυτής της τιμής επί 100 δίνει το **ποσοστό αθροιστικής συχνότητας** ή την **αθροιστική ποσοστιαία συχνότητα** που σημειώνεται ως:

$$F_i(\%) = 100 \cdot F_i = 100 \cdot \sum_{j \leq i} f_j \tag{1.1.6}$$

- Η **ποσοστιαία τάξη** για κάθε τιμή x_i του πίνακα ή της κατανομής συχνοτήτων, είναι εκείνος ο αριθμός α ο οποίος παριστάνει το ποσοστό των τιμών που είναι μικρότερες ή ίσες από τη θεωρούμενη τιμή. Με άλλα λόγια, *πρόκειται για το ποσοστό αθροιστικής συχνότητας της θεωρούμενης τιμής* και εκφράζεται ως:

$$F_i(\%) = \alpha \text{ ή } F_i = \alpha\% \tag{1.1.7}$$

Έτσι, στο Παράδειγμα 1.1.1., για την τιμή $x_i = 48$ Kgr του Πίνακα 1.1.3 η ποσοστιαία τάξη είναι η αντίστοιχη αθροιστική ποσοστιαία συχνότητα, δηλαδή η ποσότητα:
$F_{48}(\%) = 26{,}25$ ή $F_{48} = 26{,}25\%$

- Το **ποσοστιαίο σημείο**[3] (percentile), είναι εκείνη η τιμή x_i του πίνακα ή της κατανομής συχνοτήτων στην οποία αντιστοιχεί (το) *δεδομένο ποσοστό αθροιστικής συχνότητας α*.

$$P_\alpha = x_i \tag{1.1.8}$$

Από τον Πίνακα 1.1.3 φαίνεται ότι το ποσοστό αθροιστικής συχνότητας (η αθροιστική ποσοστιαία συχνότητα) 26,25% αντιστοιχεί στην τιμή $x_i = 48$ Kgr. Επομένως, η τιμή αυτή είναι (ένα) ποσοστιαίο σημείο και συμβολίζεται ως: $P_{26,25} = 48$.

Σύμφωνα με τα αναφερόμενα, αυτό το ποσοστιαίο σημείο δείχνει ότι η τιμή x_i του πίνακα συχνοτήτων ως προς την οποία είναι μικρότερες ή ίσες το 26,25% των δεδομένων τιμών, είναι η τιμή $x_i = 48$. Πράγματι, από τον Πίνακα 1.1.3. φαίνεται ότι μέχρι και την τιμή $x_i = 48$ kgr οι τιμές x_j καταγράφονται ως:

- 45 Kgr → εμφανίζεται μία φορά και αποτελεί το 1,25% του συνόλου των 80 τιμών
- 46 Kgr → εμφανίζεται τρείς φορές και αποτελεί το 3,75% του συνόλου των 80 τιμών
- 47 Kgr → εμφανίζεται έξι φορές και αποτελεί το 7,50% του συνόλου των 80 τιμών
- 48 Kgr → εμφανίζεται έντεκα φορές και αποτελεί το 13,75% του συνόλου των 80 τιμών

Επομένως, οι τιμές που κυμαίνονται από 45 Kgr έως και 48 Kgr (συνολικά) εκφράζουν το 1,25% + 3,75% + 7,50% + 13,75% = 26,25% του συνόλου των 80 τιμών, δηλαδή εκφράζουν το 26,25% των κοριτσιών του δείγματος που το βάρος τους βρίσκεται στο αναφερόμενο διάστημα τιμών.

Ειδικά ποσοστιαία σημεία που χρησιμοποιούνται είναι τα: 25%, 50%, 75% τα οποία συμβολίζονται ως $\boldsymbol{P_{25}}$, $\boldsymbol{P_{50}}$, $\boldsymbol{P_{75}}$ ή $\boldsymbol{Q_1}$, $\boldsymbol{Q_2}$, $\boldsymbol{Q_3}$, αντιστοίχως, και ονομάζονται **τεταρτημόρια** (quartiles).

☞ Είναι σαφές ότι *ο προσδιορισμός όλων των παραπάνω ποσοτήτων έχει ως στόχο την περιγραφή της κατανομής*.

ΠΑΡΑΔΕΙΓΜΑ 1.1.1.

Ο πίνακας ή η κατανομή των βαρών ενός δείγματος 80 κοριτσιών, *με προσέγγιση 1 kgr ή διαστήματα τάξης του 1 kgr*, απεικονίζεται με τον Πίνακα 1.1.1., στον οποίο τα βάρη αυτά καταγράφονται με τυχαίο τρόπο.

Σημειώνεται ότι ως «βάρος» χαρακτηρίζεται «το μέσον ενός διαστήματος τάξης», όπου *αυτά τα διαστήματα είναι ίσου μήκους*. Για παράδειγμα, το βάρος 51 kgr αντιστοιχεί στο μέσον του διαστήματος από 50,5 kgr έως 51,5 kgr. Έτσι, όλα τα βάρη που περιλαμβάνονται στο δι-

[3] Εκτενέστερη αναφορά θα γίνει στα επόμενα

άστημα αυτό, δηλαδή *όλες οι τιμές x που ικανοποιούν την ανισότητα* $50{,}5 < x \leq 51{,}5$[4] *καταγράφονται ως 51 kgr*.

Πίνακας 1.1.1. Βάρη (Kgr) 80 κοριτσιών

45	46	48	49	50	48	46	50	52	49
48	51	47	50	50	49	50	48	49	47
51	49	52	48	48	49	50	49	46	48
50	49	48	51	49	49	47	50	49	48
48	50	49	52	48	51	50	49	47	49
51	53	50	51	49	49	47	49	49	52
49	51	53	49	52	49	49	50	51	47
49	50	49	49	54	53	54	50	50	50

Στον Πίνακα 1.1.2., τα βάρη καταγράφονται κατά αύξουσα σειρά. Επισημαίνεται ότι ο πίνακας αυτός ονομάζεται *απλός πίνακας*, διότι οι τάξεις ή τα διαστήματα τάξης αναφέρονται σε ένα μόνο χαρακτηριστικό των οντοτήτων του στατιστικού συνόλου· εδώ αναφέρονται στο βάρος των κοριτσιών. Με άλλη έκφραση, *ένας απλός πίνακας παρέχει μια μονοδιάστατη κατανομή του στατιστικού συνόλου που μελετάται.* Για την αναφερόμενη κατανομή, οι τιμές βάρους καταγράφονται ως x_i και το πλήθος των οντοτήτων (των κοριτσιών) που εμφανίζουν την ίδια τιμή βάρους αναφέρεται ως n_i.

Πίνακας 1.1.2. Κατανομή εμφανίσεων για τα βάρη 80 κοριτσιών

Βάρος (kgr) x_i	**Πλήθος οντοτήτων / Πλήθος εγγραφών** n_i	**Εμφανίσεις τιμών ≡ Πλήθος εγγραφών** n_i
45	1	1
46	3	3
47	6	6
48	11	11
49	25	25
50	16	16
51	8	8
52	5	5
53	3	3
54	2	2
	Σύνολο οντοτήτων / εγγραφών: $n = \sum n_i = 80$	Σύνολο εμφανίσεων τιμών / εγγραφών: $n = \sum n_i = 80$

Μια πρώτη παρατήρηση επί του Πίνακα 1.1.2. δείχνει ότι για την πλειοψηφία των κοριτσιών, δηλαδή για ένα πλήθος 52 $(= 11 + 25 + 16)$ εγγραφών, τα βάρη κυμαίνονται από 48 kgr έως και 50 kgr, ενώ μόνο ένας μικρός αριθμός κοριτσιών εμφανίζει τιμές στα άκρα της

[4] Το διάστημα τιμών για την τιμή π.χ. 51 Kgr μπορεί να ορισθεί και ως $50{,}5 \leq x < 51{,}5$.Την ισότητα στη δεξιά πλευρά του διαστήματος χρησιμοποιούν π.χ. τα προγράμματα: Excel, SPSS, ArcMap, προκειμένου να δημιουργηθούν τα σχετικά γραφήματα καθώς και θεματικοί χάρτες.

κατανομής. Έχοντας ως βάση αυτόν τον Πίνακα, δημιουργείται ο Πίνακας 1.1.3. που ονομάζεται **Πίνακας Κατανομής Συχνοτήτων**.

Από τον Πίνακα 1.1.3. φαίνεται ότι η τιμή $x_i = 48$ kgr έχει συχνότητα $f_{48} = 11/80 = 0{,}1375$ και ποσοστό συχνότητας ή ποσοστιαία συχνότητα $f_{48}(\%) = 100 \cdot f_{48} = 13{,}75$ ή $f_{48} = 13{,}75$ %. Δηλαδή, η τιμή βάρους 48 kgr που αντιπροσωπεύει το διάστημα τιμών (47,5 ÷ 48,5] αντιστοιχεί στο 13,75% των τιμών βάρους του δείγματος των 80 κοριτσιών ή το 13,75% των κοριτσιών του δείγματος έχει βάρος που βρίσκεται στο διάστημα (47,5 ÷ 48,5] kgr.

Πίνακας 1.1.3. Πίνακας Κατανομής Συχνοτήτων

Βάρος x_i (kgr)	Εμφάνιση n_i (κορίτσια)	Συχνότητα Εμφάνισης $f_i = n_i/n$	% Συχνότητας $f_i(\%) = 100 \cdot f_i$	Αθροιστική Εμφάνιση $N_i = \sum_{j\le i} n_j$ (κορίτσια)	Αθροιστική Συχνότητα Εμφάνισης $F_i = \sum_{j\le i} f_j = \sum_{j\le i}(n_j/n)$	% Αθροιστικής Συχνότητας $F_i(\%) = 100 \cdot \sum_{j\le i} f_j$
45	1	0,0125	1,25	1	0,0125	1,25
46	3	0,0375	3,75	4	0,0500	5,00
47	6	0,0750	7,50	10	0,1250	12,50
48	11	0,1375	13,75	21	0,2625	26,25
49	25	0,3125	31,25	46	0,5750	57,50
50	16	0,2000	20,00	62	0,7750	77,50
51	8	0,1000	10,00	70	0,8750	87,50
52	5	0,0625	6,25	75	0,9375	93,75
53	3	0,0375	3,75	78	0,9750	97,50
54	2	0,0250	2,50	80	1,0000	100,00
Σύνολο	80	1,0000	100,00			

Έχει αναφερθεί ότι η τιμή 48 kgr αντιστοιχεί στο μέσον του διαστήματος (47,5 ÷ 48,5] kgr. Κατά συνέπεια, το άθροισμα:

$$13{,}75\% + 31{,}25\% + 20{,}00\% + 10{,}00\% + 6{,}25\% + 3{,}75\% + 2{,}50\% = 87{,}50\%$$

εκφράζει το ποσοστό των τιμών που είναι μεγαλύτερες των 47,5 kgr (δηλαδή το ποσοστό των κοριτσιών του δείγματος που έχουν βάρος μεγαλύτερο των 47,5 kgr).

Αντιστοίχως, το άθροισμα $1{,}25\% + 3{,}75\% + 7{,}50\% = 12{,}50\%$ εκφράζει το ποσοστό των τιμών που είναι μικρότερες ή ίσες των 47,5 kgr (δηλαδή το ποσοστό των κοριτσιών του δείγματος που έχουν βάρος μικρότερο ή ίσο των 47,5 kgr).

Η αθροιστική εμφάνιση, η αθροιστική συχνότητα και *τα ποσοστά αθροιστικής συχνότητας ή οι αθροιστικές ποσοστιαίες συχνότητες*, είναι κοινές περιγραφές που αντιστοιχούν σε διαστήματα τιμών ή σε «στρογγυλοποιημένες» τιμές. Έτσι, στην **τιμή 48 kgr** αντιστοιχούν:

- **Η αθροιστική εμφάνιση 21, $N_{48} = 21$**

  Προκύπτει ως το άθροισμα των εμφανίσεων

  $$N_{48} = \sum_{j\le i=48} n_j = n_{45} + n_{46} + n_{47} + n_{48} = 1 + 3 + 6 + 11 = 21$$

  δηλαδή, ως το άθροισμα των οντοτήτων (των κοριτσιών) που αντιστοιχούν στις τιμές βάρους 45, 46, 47, 48 Kgr. Αυτό σημαίνει ότι έχουν καταγραφεί 21 τιμές μικρότερες ή ίσες της τιμής 48 Kgr ή 21 κορίτσια έχουν βάρος μικρότερο ή ίσο των 48 Kgr.

2.

ΒΑΣΙΚΑ ΘΕΜΑΤΑ ΤΗΣ ΘΕΩΡΙΑΣ ΠΙΘΑΝΟΤΗΤΩΝ

2.1. ΕΙΣΑΓΩΓΗ

Είναι γνωστό ότι η Στατιστική Επιστήμη χρησιμοποιείται για τη μελέτη και την κατανόηση είτε φαινομένων, είτε χαρακτηριστικών / ιδιοτήτων που αναφέρονται σε διάφορους πληθυσμούς οντοτήτων, όταν *υπάρχουν πληροφορίες / δεδομένα από ένα τμήμα* είτε του φαινομένου είτε του πληθυσμού των οντοτήτων. Το τμήμα αυτό αποτελεί το λεγόμενο **δείγμα** (sample). Όμως, η *γενίκευση* της γνώσης που λαμβάνεται από τη μελέτη του δείγματος, δηλαδή η αναφορά στο αντίστοιχο φαινόμενο ή στον αντίστοιχο πληθυσμό, συνεπάγεται **αβεβαιότητα**, εφ' όσον το ζητούμενο είναι η *εξαγωγή συμπερασμάτων για το όλον από ένα τμήμα αυτού*. Η αβεβαιότητα αυτή, όταν ισχύουν ορισμένες συνθήκες, μπορεί να μετρηθεί με τη βοήθεια της **Θεωρίας Πιθανοτήτων** (Probability Theory).

Γενικά, η θεωρία πιθανοτήτων ασχολείται με **αβέβαιες διαδικασίες** (uncertain processes). Μία διαδικασία είναι αβέβαιη όταν είναι δυνατό να έχει περισσότερα από ένα αποτελέσματα και δεν είναι γνωστό με βεβαιότητα, εκ των προτέρων, ποιο αποτέλεσμα θα συμβεί. Εάν *στα δυνατά αποτελέσματα μίας αβέβαιης διαδικασίας αντιστοιχηθούν αριθμοί*, οι οποίοι ικανοποιούν ορισμένες συνθήκες, τότε είναι δυνατό, με μαθηματική λογική, να *προσδιορισθούν οι κανόνες / οι νόμοι που διέπουν τη διαδικασία*. Οι αριθμοί αυτοί είναι οι λεγόμενες **πιθανότητες** (probabilities). Με άλλα λόγια, *η θεωρία πιθανοτήτων επιτρέπει την προσομοίωση των πραγματικών καταστάσεων με μαθηματικά μοντέλα*, έτσι ώστε με τη χρησιμοποίησή τους να μπορούν να εξαχθούν συμπεράσματα.

Επειδή, ούτε η μελέτη του συνολικού πληθυσμού οντοτήτων, ούτε η παρακολούθηση ενός φαινομένου εξ ολοκλήρου είναι δυνατές, ως διαδικασίες προσέγγισης / παρατήρησης αυτών χρησιμοποιούνται:

• Η **δειγματοληψία** (sampling), δηλαδή η επιλογή δείγματος, όταν πρόκειται για πληθυσμό οντοτήτων

• Το **πείραμα**, όταν πρόκειται για φαινόμενο

Όσον αφορά την *επιλογή του δείγματος*, για να είναι αξιόπιστα τα συμπεράσματα που εξάγονται, θα **πρέπει** το δείγμα να είναι **τυχαίο** και **αντιπροσωπευτικό** του πληθυσμού από τον οποίο προέρχεται.

2.2. ΤΥΧΑΙΟ ΦΑΙΝΟΜΕΝΟ, ΤΥΧΑΙΟ ΠΕΙΡΑΜΑ Ή ΠΕΙΡΑΜΑ ΤΥΧΗΣ

2.2.1. ΤΥΧΑΙΟ ΦΑΙΝΟΜΕΝΟ

Τα φαινόμενα ή οι καταστάσεις των οποίων το αποτέλεσμα κάθε παρατήρησής τους, μετά από σειρά επαναλήψεων αυτών και με τις ίδιες υφιστάμενες συνθήκες, ***δεν*** *μπορεί να προβλεφθεί με βεβαιότητα*, λέγονται **τυχαία φαινόμενα**. Για παράδειγμα:

- Κατά τη ρίψη ενός νομίσματος δεν είναι δυνατή με βεβαιότητα η πρόβλεψη της εμφάνισης συγκεκριμένης όψης αυτού, δηλαδή εάν εμφανισθούν η «κεφαλή» ή τα «γράμματα».
- Κατά τη ρίψη ενός ζαριού δεν είναι δυνατή με βεβαιότητα η πρόβλεψη της εμφάνισης συγκεκριμένης έδρας αυτού. Ακόμη, και σε δεύτερη ρίψη του ζαριού, δεν είναι βέβαιο ότι θα εμφανισθεί η έδρα της πρώτης ρίψης.

Βεβαίως, εκτός από τα τυχαία φαινόμενα υπάρχουν και αυτά των οποίων το αποτέλεσμα κάθε παρατήρησής τους *μπορεί να προβλεφθεί με βεβαιότητα* ή με μεγάλη προσέγγιση, όσες φορές και εάν επαναληφθεί η παρατήρηση. Τα φαινόμενα αυτά ονομάζονται **προσδιοριστικά** ή **αιτιοκρατικά** (deterministic) **φαινόμενα**. Για παράδειγμα:

- Το νερό βράζει στους 100^{o} C
- Το βλήμα ενός πυροβόλου όπλου πέφτει πάντοτε στη γη, όσο μεγάλο και εάν είναι το βεληνεκές που θα διαγράψει

2.2.2. ΤΥΧΑΙΟ ΠΕΙΡΑΜΑ Ή ΠΕΙΡΑΜΑ ΤΥΧΗΣ

Το **πείραμα** (experiment) είναι μία διαδικασία που επιτρέπει την *παρατήρηση ενός φαινομένου*. Δηλαδή, με το πείραμα *μπορεί να προβλεφθεί το ίδιο αποτέλεσμα* όσες φορές και εάν παρατηρηθεί το φαινόμενο, κάτω από τις ίδιες (κατά το δυνατό) συνθήκες. Για παράδειγμα, η θέρμανση μίας μεταλλικής ράβδου προκαλεί την επιμήκυνσή της, και αυτό το (ίδιο) αποτέλεσμα είναι προβλέψιμο όσες φορές και εάν επαναληφθεί το πείραμα.

Το **τυχαίο πείραμα** ή **πείραμα τύχης** (random experiment) ή **μη προσδιοριστικό πείραμα** (non deterministic experiment) είναι κάθε πείραμα που μπορεί να επαναληφθεί πολλές φορές, *με ισχύουσες τις ίδιες ή παρόμοιες συνθήκες*, και του οποίου το αποτέλεσμα, σε κάθε επανάληψη, δεν είναι πάντοτε το ίδιο και δεν μπορεί να προβλεφθεί. Δηλαδή, *υπάρχει αβεβαιότητα αναφορικά με το αποτέλεσμα που παρατηρείται κάθε φορά*, γεγονός που σημαίνει ότι τα αποτελέσματα μπορεί να είναι περισσότερα από ένα. Με άλλα λόγια, με το τυχαίο πείραμα καθορίζεται ένα σύνολο δυνατών αποτελεσμάτων, χωρίς να είναι βέβαιο ότι πρόκειται για (τα) ίδια αποτελέσματα. Παραδείγματα τυχαίων πειραμάτων είναι:

- Η ρίψη ενός νομίσματος .

- Η κλήρωση του λαχείου μία συγκεκριμένη ημέρα.
- Το πλήθος των κλήσεων που δέχεται ένα τηλεφωνικό κέντρο σε μία ημέρα.
- Η χρονοαπόσταση ενός αστικού λεωφορείου, από την αφετηρία έως το τέλος της διαδρομής του, κατά τις ώρες αιχμής της κυκλοφορίας του.
- Ο χρόνος αναμονής για να εξυπηρετηθεί ένα άτομο σε ένα πολυκατάστημα το πρωΐ του Σαββάτου.
- Οι καιρικές συνθήκες σε μία πόλη μία συγκεκριμένη ανοιξιάτικη ημέρα.
- Η κατανάλωση βενζίνης το τριήμερο του Πάσχα.

Όπως έχει αναφερθεί, και όπως συνάγεται και από τα παραπάνω παραδείγματα, ένα τυχαίο πείραμα ή πείραμα τύχης μπορεί, θεωρητικά, να επαναληφθεί άπειρες φορές, με την προϋπόθεση ότι *οι συνθήκες παραμένουν οι ίδιες*. Κάθε μία από αυτές τις επαναλήψεις ονομάζεται **δοκιμή** (trial). Κοινό γνώρισμα των πειραμάτων τύχης είναι ότι *όσο αυξάνεται το πλήθος των δοκιμών / επαναλήψεων παρατηρείται μία κανονικότητα στην εμφάνιση των αποτελεσμάτων τους*. Έτσι, εάν ένα γνήσιο νόμισμα ριφθεί πολλές φορές στον αέρα, δηλαδή το πλήθος των δοκιμών είναι μεγάλο, τότε παρατηρείται ότι περίπου στο ½ αυτού του πλήθους εμφανίζεται η όψη «κεφαλή» (περίπου το ½ του συνόλου των αποτελεσμάτων είναι «κεφαλή») και στο υπόλοιπο ½ του πλήθους των ρίψεων εμφανίζεται η όψη «γράμματα» (το υπόλοιπο ½ των συνολικών αποτελεσμάτων είναι «γράμματα»). Η κανονικότητα αυτή είναι γνωστή ως **στατιστική κανονικότητα** (statistical regularity) και *επιτρέπει τη μέτρηση της αβεβαιότητας κατά την εκτέλεση της δοκιμής*.

Για παράδειγμα, εάν ο μέγιστος χρόνος εξυπηρέτησης στην Τράπεζα Α, τις πρωϊνές ώρες 8,0 – 10,0 π.μ, είναι 15 min, τότε ένας πελάτης αισθάνεται περισσότερο ασφαλής ότι θα εξυπηρετηθεί, εάν γνωρίζει ότι κατά 90% ο μέσος χρόνος εξυπηρέτησης, γι' αυτό το χρονικό διάστημα, είναι 10 min, παρά εάν το ποσοστό αυτό είναι 20%.

Σύμφωνα με τον ορισμό του τυχαίου πειράματος, αυτό μπορεί να επαναληφθεί. Όμως, υπάρχουν και *τυχαία πειράματα που δεν μπορούν να επαναληφθούν* και τα αποτελέσματά τους είναι αβέβαια. Για παράδειγμα, η επένδυση στην παραγωγή ενός νέου προϊόντος, είναι ένα τυχαίο πείραμα. Τα τυχαία πειράματα αυτής της μορφής ονομάζονται **μοναδικά** (unique random experiments).

☞ *Στο εξής, το τυχαίο πείραμα θα αναφέρεται, απλά, ως πείραμα.*

2.3. ΔΕΙΓΜΑΤΙΚΟΣ ΧΩΡΟΣ Ή ΔΕΙΓΜΑΤΟΧΩΡΟΣ ΓΕΓΟΝΟΤΑ Ή ΕΝΔΕΧΟΜΕΝΑ

2.3.1. ΔΕΙΓΜΑΤΙΚΟΣ ΧΩΡΟΣ Ή ΔΕΙΓΜΑΤΟΧΩΡΟΣ

Το **σύνολο** όλων των δυνατών αποτελεσμάτων ενός πειράματος δημιουργεί τον λεγόμενο **δειγματικό χώρο** ή **δειγματοχώρο** (sample space) του πειράματος· ο χώρος αυτός συμβολίζεται ως $\boldsymbol{S}$ ή $\boldsymbol{\Omega}$: $S = \{s_1, s_2, \dots, s_n\}$ ή $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$, όπου: s_1, s_2, ..., s_n, ή ω_1, ω_2, ..., ω_n, είναι όλα τα δυνατά αποτελέσματα του πειράματος. Τα *στοιχεία ενός δειγματικού χώρου*, δηλαδή *όλα τα δυνατά αποτελέσματα του πειράματος*, ονομάζονται **δειγματικά σημεία** (sample points) ή **σημεία του δειγματοχώρου**. Εφ' όσον ο δειγματικός χώρος είναι ένα σύνολο, προσδιορίζεται είτε με απαρίθμηση των δειγματικών του σημείων, είτε με μία πρόταση ή ιδιότητα η οποία ικανοποιεί κάθε σημείο του. Η πρώτη μέθοδος για τον προσδιορισμό του δειγματικού χώρου λέγεται «*μέθοδος της αναγραφής*» και η δεύτερη μέθοδος λέγεται «*μέθοδος της περιγραφής*». Για παράδειγμα:

- Ένα νόμισμα, ρίπτεται 1 φορά και το αποτέλεσμα είναι είτε «κεφαλή» (K) είτε «γράμματα» (Γ). Επομένως, ο δειγματικός χώρος του πειράματος προσδιορίζεται ως:

  $S = \{K, \Gamma\}$, ή

  $S = \{0, 1\}$

  όπου: η όψη «κεφαλή» (K) *συμβολίζεται με 0* και η όψη «γράμματα» (Γ) *συμβολίζεται με 1*.

- Ένα νόμισμα ρίπτεται 2 φορές και το αποτέλεσμα κάθε φορά είναι είτε «κεφαλή» (K) είτε «γράμματα» (Γ). Τότε, ο δειγματικός χώρος του πειράματος αυτού είναι το σύνολο όλων των δυνατών εμφανίσεων είτε του $(K) \rightarrow (0)$ είτε του $(\Gamma) \rightarrow (1)$. Δηλαδή:

  $S = \{(K, K), (K, \Gamma), (\Gamma, K), (\Gamma, \Gamma)\}$, ή

  $S = \{(0,0), (0,1), (1,0), (1,1)\}$

- Σε μία κληρωτίδα υπάρχουν 10 ομοιόμορφες και αδιαφανείς μπάλες· κάθε μία από αυτές περιέχει ένα αριθμό από 0 έως 9. Από αυτή την κληρωτίδα λαμβάνεται μία μπάλα, οπότε *όλα τα αποτελέσματα είναι δυνατά*. Δηλαδή, μπορεί να εμφανισθεί είτε ο αριθμός 0, είτε ο αριθμός 1, είτε ο αριθμός 2, ..., είτε ο αριθμός 9. Κατά συνέπεια, το σύνολο όλων των δυνατών αποτελεσμάτων ή, άλλως, *ο δειγματικός χώρος του πειράματος* είναι, σύμφωνα με τη *μέθοδο της αναγραφής*:

  $S = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$

  Όμως, αυτός ο *δειγματικός χώρος* μπορεί να γραφεί και ως εξής, σύμφωνα με τη *μέθοδο της περιγραφής*:

  $S = \{s : s = ακέραιος\ αριθμός\ από\ 0\ έως\ 9\}$

Ένας δειγματικός χώρος ονομάζεται **διακριτός** (discrete) όταν έχει **πεπερασμένο** (finite) ή **άπειρο αλλά αριθμήσιμο** (countably infinite) **πλήθος στοιχείων**. Όταν ένας δειγματικός χώρος έχει **μη αριθμήσιμο πλήθος στοιχείων** λέγεται **μη διακριτός** ή **συνεχής** (continuous). Για παράδειγμα:

- Το πλήθος των μελών ενός νοικοκυριού αποτελεί ένα διακριτό δειγματικό χώρο

  $S = \{s : s = το\ μέλος\ του\ νοικοκυριού\}$

- Το ύψος των κοριτσιών ηλικίας 15 ετών αποτελεί ένα συνεχή δειγματικό χώρο

  $S = \{s : s = το\ ύψος\ ενός\ κοριτσιού\ ηλικίας\ 15\ ετών\}$

Πολλές φορές, *ο δειγματικός χώρος S απεικονίζεται γραφικά*. Σε αυτές τις περιπτώσεις συνήθως, εφ' όσον είναι δυνατό, χρησιμοποιούνται αριθμοί αντί γραμμάτων.

1ο Παράδειγμα

Ένα νόμισμα ρίπτεται 1 φορά. Τα δυνατά αποτελέσματα είναι είτε «κεφαλή» είτε «γράμματα», δηλαδή: K, Γ, ή 0, 1. Οπότε, ο δειγματικός χώρος του πειράματος είναι:

$S = \{K, \Gamma\}$ ή $S = \{0, 1\}$, με τη μέθοδο της αναγραφής, ή

$S = \{s : s = K, \Gamma\}$ ή $S = \{s : s = 0, 1\}$, με τη μέθοδο της περιγραφής

Η απεικόνισή του με μία ευθεία γραμμή φαίνεται στο Σχήμα 2.3.1.

3.

ΤΟ ΔΕΙΓΜΑ ΚΑΙ Ο ΣΧΗΜΑΤΙΣΜΟΣ ΤΟΥ

3.1. ΕΙΣΑΓΩΓΗ

Προκειμένου να μελετηθεί ο γεωγραφικός χώρος, αναλύονται και συνδυάζονται τα κάθε είδους στοιχεία που αναφέρονται σε αυτόν ή/και απορρέουν από αυτόν. Τα σύνολα όλων των σχετικών χαρακτηριστικών και πληροφοριών αποτελούν τους αντίστοιχους **πληθυσμούς** του γεωγραφικού χώρου. Η επεξεργασία όλων αυτών είναι απαραίτητη για την εξαγωγή βασικών παραμέτρων, οι οποίες, με τη σειρά τους, θα «καθοδηγήσουν» και θα προσδιορίσουν τη γεωγραφική μελέτη.

Όμως, όλα αυτά τα χαρακτηριστικά και οι πληροφορίες προέρχονται από οντότητες (άνθρωποι / χωρικές αναφορές / περιπτώσεις) του θεωρούμενου γεωγραφικού χώρου, οι οποίες έχουν κοινά ένα ή περισσότερα χαρακτηριστικά που αποδίδονται με *μετρήσεις*, *απαριθμήσεις* και *καταγραφές*. Έτσι, λοιπόν, *ως πληθυσμός ορίζεται το σύνολο των οντοτήτων με ένα ή περισσότερα κοινά χαρακτηριστικά*. Κατά συνέπεια, ο πληθυσμός συγκεκριμένων χαρακτηριστικών δεν είναι παρά ο πληθυσμός των οντοτήτων που έχουν αυτά τα χαρακτηριστικά. Για παράδειγμα:

• Ο πληθυσμός των κατοίκων ηλικίας 15 – 21 ετών στην περιφέρεια Κ. Μακεδονίας ή ο πληθυσμός των ηλικιών 15 – 21 ετών στην περιφέρεια Κ. Μακεδονίας (απαριθμήσεις).

• Ο πληθυσμός των αγροκτημάτων μεγέθους $\geq$ 10 στρμ. στην ίδια περιφέρεια (μετρήσεις),

• Ο πληθυσμός των ονομάτων των οικισμών της περιφέρειας Κ. Μακεδονίας (καταγραφές).

☞ Επισημαίνεται ότι *εφ' όσον η μελέτη πραγματεύεται τον γεωγραφικό χώρο, στους κάθε είδους πληθυσμούς που αναλύονται ενυπάρχει και η γεωγραφική – χωρική διάσταση*.

Χωρίς αμφιβολία, ο καλύτερος τρόπος για να μελετηθεί ένας πληθυσμός είναι η γνώση των χαρακτηριστικών για κάθε οντότητα αυτού ξεχωριστά, γεγονός που σημαίνει ότι θα (πρέπει να) υπάρχουν δεδομένα για όλες τις οντότητες του πληθυσμού. Κάτι τέτοιο όμως, συνήθως, ούτε εφικτό είναι ούτε αναγκαίο. Εάν, για παράδειγμα, ως ζητούμενο προκύπτει το μέσο μέγεθος των αγροκτημάτων της περιφέρειας Κ. Μακεδονίας, δηλαδή *το μέσο εμβαδόν του πληθυσμού των αγροκτημάτων της περιφέρειας Κ. Μακεδονίας*, δεν είναι αναγκαία η εμβαδομέτρηση όλων των αγροκτημάτων αυτής της γεωγραφικής περιοχής για την εξαγωγή του μέσου εμβαδού. Υπάρχουν διαδικασίες σύμφωνα με τις οποίες γίνεται εμβαδομέτρηση ενός μικρού αριθμού αγροκτημάτων και, στη συνέχεια, *προσεγγίζεται ικανοποιητικά το μέσο εμβαδόν του πληθυσμού των αγροκτημάτων*. Έτσι, γίνεται οικονομία σε χρόνο, χρήμα και προσπάθεια. Με άλλα λόγια, *εφαρμόζονται στατιστικές μέθοδοι* σε ένα μικρό *αντιπροσωπευτικό* τμήμα του πληθυσμού, το οποίο ονομάζεται **δείγμα**, και υπολογίζονται τα στατιστικά του δείγματος από τα οποία γίνεται εκτίμηση των παραμέτρων που αφορούν τον αντίστοιχο πληθυσμό. Επομένως, *απαραίτητη προϋπόθεση* για τη μελέτη ενός δείγματος και την εξαγωγή, στη συνέχεια, ασφαλών συμπερασμάτων για ολόκληρο τον πληθυσμό είναι ο *ορθός σχηματισμός του δείγματος*.

Ένα δείγμα μπορεί να λειτουργήσει ως *αξιόπιστος αντικαταστάτης* του πληθυσμού, από τον οποίο προέρχεται, όταν πληροί δύο προϋποθέσεις:

● Η *πρώτη προϋπόθεση αφορά το μέγεθος n του δείγματος*· το δείγμα πρέπει να έχει «μεγάλο» μέγεθος. Όταν το δείγμα είναι μεγάλο:

• αφ' ενός, η εφαρμογή στατιστικών μεθόδων δίνει *αποτελέσματα αρκετά ικανοποιητικά*, γεγονός που σημαίνει ότι προσεγγίζουν με μεγάλη βεβαιότητα την πραγματικότητα, και

• αφ' ετέρου, είναι δυνατή η *εφαρμογή στατιστικών μεθόδων σε ευρεία κλίμακα*.

Πολλές φορές, δεν είναι δυνατός ο σχηματισμός «μεγάλων» δειγμάτων και τα *διαθέσιμα δείγματα είναι «μικρά»*. Και σε αυτή την περίπτωση, μπορούν να εφαρμοσθούν στατιστικές μέθοδοι. Όμως, τα αποτελέσματα που προκύπτουν είναι λιγότερο ικανοποιητικά από αυτά που θα προέκυπταν εάν τα δείγματα ήταν «μεγάλα».

Κατά συνέπεια, *πρέπει το μέγεθος n του δείγματος να υπολογίζεται έτσι ώστε να θεωρείται ότι είναι «μεγάλο».*

● Η *δεύτερη προϋπόθεση είναι πολύ ουσιαστική*. *Το δείγμα πρέπει να αντιπροσωπεύει οπωσδήποτε τον πληθυσμό από τον οποίο προέρχεται*. Το γεγονός αυτό σημαίνει ότι το δείγμα πρέπει να αποδίδει τις διαφοροποιήσεις του πληθυσμού, κατά το δυνατό, αξιόπιστα. Συνεπώς, για να επιτευχθεί αυτό (θα) πρέπει *να αποφεύγεται η μεροληπτικότητα* η οποία προκύπτει όταν η επιλογή των στατιστικών οντοτήτων γίνεται από ορισμένα μόνο τμήματα του πληθυσμού και όχι από το σύνολο του πληθυσμού. Δηλαδή, η επιλογή των οντοτήτων του πληθυσμού, οι οποίες θα συμμετέχουν στο δείγμα, (θα) πρέπει να γίνεται κατά τρόπο απολύτως «αντικειμενικό» έτσι ώστε, *να εξασφαλίζεται ότι όλες οι οντότητες του πληθυσμού έχουν την ίδια πιθανότητα να περιληφθούν στο δείγμα*. Η διαδικασία αυτή αναφέρεται ως *«"τυχαία" επιλογή των οντοτήτων του πληθυσμού»*. Για παράδειγμα, ο πληθυσμός των οικισμών της χώρας εμφανίζει διαφοροποιήσεις, εφ' όσον οι οικισμοί, ανάλογα με το υψόμετρό τους, διαχωρίζονται σε πεδινούς, ημιορεινούς και ορεινούς οικισμούς. Επομένως, από αυτόν τον πληθυσμό θα γίνει τυχαία επιλογή των οικισμών με τρόπο ώστε να αποδίδονται και στο δείγμα οι υφιστάμενες πληθυσμιακές διαφοροποιήσεις.

Εάν το δείγμα δεν είναι αντιπροσωπευτικό, καθώς και στην περίπτωση που δεν είναι βέβαιο ότι είναι αντιπροσωπευτικό, τα αποτελέσματα των στατιστικών αναλύσεων είναι πολύ πιθανό να οδηγήσουν σε λανθασμένες αποφάσεις και συμπεράσματα. Επομένως, εάν η επιλογή του δείγματος γίνει με βάση την προσωπική κρίση του ερευνητή, δηλαδή είναι η περίπτωση της *υποκειμενικής δειγματοληψίας*, τότε πρόκειται για ένα μεροληπτικό δείγμα, διότι σε αυτή την περίπτωση το υποσυνείδητο, συνήθως, οδηγεί σε άλλες επιλογές. Όμως, ακόμη και εάν κατά την υποκειμενική δειγματοληψία δεν υπάρχει μεροληπτικότητα, η διαδικασία αυτή δεν μπορεί να αποδειχθεί και να τεκμηριωθεί «αντικειμενικά». Κατά συνέπεια, κατά την επιλογή ενός δείγματος από κάποιον πληθυσμό (θα) *πρέπει να ακολουθηθούν αντικειμενικές διαδικασίες επιλογής* έτσι ώστε να μειώνεται στο ελάχιστο η πιθανότητα για μεροληπτικό δείγμα.

3.2. ΕΚΤΙΜΗΣΗ ΤΟΥ ΜΕΓΕΘΟΥΣ ΤΟΥ ΔΕΙΓΜΑΤΟΣ[1]

Εάν γίνουν μετρήσεις επί όλων των οντοτήτων ενός πληθυσμού ως προς ένα ποσοτικό χαρακτηριστικό ή εάν απαριθμηθούν / καταγραφούν όλες οι οντότητες ενός πληθυσμού ως προς ένα ποιοτικό χαρακτηριστικό, τότε οποιαδήποτε παράμετρος του πληθυσμού μπορεί να υπολογισθεί με *ακρίβεια* 100%. Αυτό σημαίνει ότι για το προκύπτον αποτέλεσμα η βεβαιότητα είναι 100% ή δεν υπάρχει αβεβαιότητα, δηλαδή η αβεβαιότητα είναι 0%, ή το αποτέλεσμα είναι *ακριβές*. Είναι φανερό ότι όταν η εκτίμηση της πληθυσμιακής παραμέτρου γίνεται από ένα δείγμα, η βεβαιότητα για την πραγματική τιμή της είτε ελαττώνεται είτε αυξάνεται. Όσο μικρότερο είναι το μέγεθος του δείγματος, τόσο αυξάνεται το περιθώριο σφάλματος ή λάθους στην εκτίμηση της παραμέτρου στον πληθυσμό. Δηλαδή, αυξάνεται η αβεβαιότητα, όσον αφορά την πραγματική τιμή της πληθυσμιακής παραμέτρου, όσο ελαττώνεται το δειγματικό μέγεθος. Με άλλα λόγια, *το περιθώριο σφάλματος ή η ακρίβεια στην εκτίμηση της παραμέτρου είναι συνάρτηση του μεγέθους n του δείγματος*.

Γίνεται, λοιπόν, φανερό ότι το μέγεθος του δείγματος κατέχει πρωταγωνιστική θέση στην εκτίμηση των παραμέτρων ενός πληθυσμού. Επομένως, «όσο μεγαλύτερο είναι το δείγμα, τόσο καλύτερες είναι οι εκτιμήσεις;» Ναι, αλλά να είναι τόσο μεγάλο όσο είναι το «**ικανοποιητικό**» του **μέγεθος**. Δηλαδή, εκείνο το μέγεθος που να επιτρέπει μία «**αντικειμενικά καλή**» εκτίμηση των παραμέτρων του πληθυσμού. Μάλιστα, με τη χρησιμοποίηση ενός «ικανοποιητικού» δειγματικού μεγέθους γίνεται εξοικονόμηση σε χρόνο και σε χρήμα, παράγοντες που απαιτούνται για τον σχηματισμό ενός μεγάλου μεγέθους δείγματος. Ως «*ικανοποιητικό» μέγεθος δείγματος*, *θεωρείται το ελάχιστο μέγεθος του δείγματος που αντιστοιχεί στο περιθώριο σφάλματος* (*στη μέγιστη απόλυτη τιμή που μπορεί να έχει το σφάλμα*), το οποίο είναι δεκτό σε μία εκτίμηση. Προκειμένου να υπολογισθεί το ελάχιστο μέγεθος ενός δείγματος, θεωρούνται δύο περιπτώσεις δείγματος: δείγμα με ποσοτικό χαρακτηριστικό (μετρήσεις), δείγμα με ποιοτικό χαρακτηριστικό (απαριθμήσεις / καταγραφές).

Όταν οι παράμετροι του πληθυσμού είναι άγνωστες, απαιτείται ο σχηματισμός ενός *δοκιμαστικού ή πιλοτικού δείγματος μεγέθους* $n_0 = 30$, δηλαδή το δείγμα περιλαμβάνει 30 οντότητες από τον πληθυσμό, για να υπολογισθούν τα αντίστοιχα στατιστικά. Η διαδικασία για τον σχηματισμό του πρέπει να ακολουθεί τη διαδικασία που περιγράφεται στα επόμενα και η οποία θα εφαρμοσθεί και στο οριστικό δείγμα. Το δοκιμαστικό δείγμα, συνήθως, αποτελεί τμήμα του οριστικού δείγματος, με την προϋπόθεση ότι η επιλογή των οντοτήτων του δεν έχει επιπτώσεις στην επιλογή των υπολοίπων οντοτήτων στο οριστικό δείγμα.

Από τα παραπάνω συνάγεται ότι το μέγεθος n του δείγματος θεωρείται «μεγάλο» όταν ικανοποιείται η σχέση:

$$n \geq 30 \tag{3.2.1}$$

[1] Βλ. Κεφάλαιο 8.6.

4.

ΤΥΧΑΙΕΣ ΜΕΤΑΒΛΗΤΕΣ – ΚΑΤΑΝΟΜΕΣ

4.1. ΕΙΣΑΓΩΓΗ

Από τις πλέον βασικές έννοιες στη Στατιστική και στη θεωρία Πιθανοτήτων είναι αυτή της τυχαίας μεταβλητής.

Ως **τυχαία μεταβλητή (τ.μ.)** X (random variable), ορίζεται εκείνη η μεταβλητή της οποίας οι τιμές προκύπτουν τυχαία και μεταβάλλονται από οντότητα σε οντότητα και από «πείραμα» σε «πείραμα». Με άλλο ορισμό, **τυχαία μεταβλητή** X λέγεται η συνάρτηση που απεικονίζει το σύνολο όλων των δυνατών αποτελεσμάτων / τιμών[1] ενός «πειράματος» τύχης στο σύνολο των πραγματικών αριθμών.

Σύμφωνα με τη θεωρία πιθανοτήτων, **τυχαία μεταβλητή** X ονομάζεται εκείνη η μεταβλητή της οποίας οι δυνατές τιμές x_i χαρακτηρίζονται από συγκεκριμένες πιθανότητες εμφάνισης $f(x_i)$, αφού η τυχαία μεταβλητή είναι η μαθηματική έκφραση κάποιου γεγονότος (φαινομένου ή «πειράματος») και η συχνότητα εμφάνισης των διαφόρων δυνατών αποτελεσμάτων ή τιμών αυτού, χαρακτηρίζεται ως πιθανότητα εμφάνισης.[2] Αυτές οι πιθανότητες $\boldsymbol{f(x_i)}$, λέγεται ότι ορίζουν μία **κατανομή πιθανότητας** (probability distribution). Με άλλα λόγια, σε ένα πληθυσμό οντοτήτων είναι δυνατή η περιγραφή των τυχαίων[3] γεγονότων που συμβαίνουν σε αυτόν με μία τυχαία μεταβλητή, χωρίς να παρατηρηθεί ολόκληρος ο πληθυσμός,

[1] Ως τέτοια αποτελέσματα / τέτοιες τιμές θεωρούνται: μετρήσεις σε μια κλίμακα, συντεταγμένες σημείων, καταγραφές κοινωνικών καταστάσεων, πλήθος απασχολούμενων σε οικονομικές δραστηριότητες, κ.λ.π.

[2] Βλ. Κεφάλαιο 1.3.2.

[3] Α. Δερμάνης, τόμος 1, 1986, σελ. 57, «… γεγονότα που το αποτέλεσμά τους δεν είναι εξαρχής μονοσήμαντα ορισμένο, και που σε καθημερινή γλώσσα τα χαρακτηρίζουμε σαν **τυχαία** γεγονότα.»

αφού, όταν τεθούν οι συνθήκες στο υπό παρατήρηση δείγμα, δηλαδή όταν εκτελεσθεί το «πείραμα», η συμπεριφορά της τυχαίας μεταβλητής προσδιορίζεται από την κατανομή των πιθανοτήτων της.

Εάν σε κάποιο φαινόμενο ή «πείραμα» ενδιαφέρει η μελέτη δύο χαρακτηριστικών, τότε εισάγονται δύο τ.μ. X και Y για τη μελέτη αυτών. Εάν, όμως, ενδιαφέρει η ταυτόχρονη μελέτη και γνώση των τιμών των δύο χαρακτηριστικών, τότε εισάγεται η έννοια της **διδιάστατης τ.μ.,** $\boldsymbol{W} = (\boldsymbol{X}, \boldsymbol{Y})$. Παρομοίως, εάν ενδιαφέρει η μελέτη, ταυτοχρόνως, περισσότερων των δύο χαρακτηριστικών, χρησιμοποιούνται οι **πολυδιάστατες τ.μ.**

4.2. ΔΙΑΚΡΙΤΕΣ ΤΥΧΑΙΕΣ ΜΕΤΑΒΛΗΤΕΣ (δ.τ.μ.)

Όταν η τ.μ. X παίρνει τιμές από ένα διακριτό «πεπερασμένο» (finite) σύνολο τιμών x_i, λέγεται **διακριτή** ή **ασυνεχής τυχαία μεταβλητή** (**δ.τ.μ.**) (discrete random variable). Όταν η τ.μ. X παίρνει τιμές από κάποιο «άπειρο» αλλά αριθμήσιμο (countable finite) σύνολο τιμών καλείται **απαριθμητή τυχαία μεταβλητή**. Στην περίπτωση των δ.τ.μ., η πιθανότητα η τ.μ. X να έμφανίζει συγκεκριμένη τιμή x_i δίνεται από τη συνάρτηση $f(x_i)$ η οποία εκφράζεται ως:

$$f(x_i) = P(X = x_i) \tag{4.2.1}$$

όπου P συμβολίζει την πιθανότητα.

Η συνάρτηση $\boldsymbol{f(x_i)}$ ονομάζεται **συνάρτηση πυκνότητας πιθανότητα**ς (**σ.π.π.**) (probability density function) και έχει τις ακόλουθες ιδιότητες:

$$f(x_i) \geq 0 \quad \forall\, x_i \in R \tag{4.2.2}$$

$$\lim_{x_i \to -\infty} f(x_i) = \lim_{x_i \to +\infty} f(x_i) = 0 \tag{4.2.3}$$

$$\sum_{x_i=-\infty}^{+\infty} f(x_i) = 1 \tag{4.2.4}$$

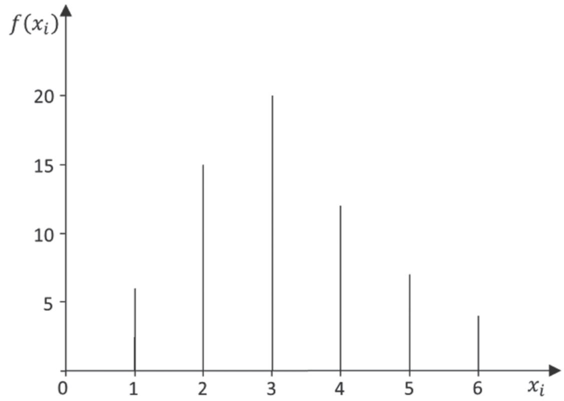
Η σχέση (4.2.4) είναι άμεση συνέπεια του ορισμού της $f(x_i)$, όπου:[4]

$$f(x_i) = \lim_{n \to \infty} f_n(x_i) = \lim_{n \to \infty} \frac{n_i}{n} \tag{4.2.5}$$

Σημειώνεται ότι η συνάρτηση $f(x_i)$ απαντάται στη βιβλιογραφία και ως **συνάρτηση πυκνότητας** (density function) ή **συνάρτηση πιθανότητας** (probability function) ή **κατανομή πιθανότητας** (probability distribution) ή **συνάρτηση μάζας πιθανότητας** (probability mass function). Η τελευταία ονομασία οφείλεται στο ότι οι πιθανότητες συγκεντρώνονται σε διακριτές τιμές της τ.μ. και δεν κατανέμονται σε ένα διάστημα. Επίσης, αναφέρεται και ως **συνάρτηση διακριτής πυκνότητας** (discrete density function), προκειμένου να ξεχωρίζει ο διακριτός χαρακτήρας της τ.μ.

[4] Βλ. Κεφάλαιο. 1.3.2.

Η γραφική απεικόνιση της κατανομής των τιμών x_i της δ.τ.μ. X γίνεται με τη βοήθεια ραβδογράμματος, Σχήμα 4.2.1. Σε αυτό, ο οριζόντιος άξονας (X) αντιπροσωπεύει τις τιμές x_i της δ.τ.μ. X και ο κατακόρυφος άξονας (Y) εκφράζει τις τιμές της συνάρτησης $f(x_i)$, δηλαδή τις τιμές της σ.π.π. για τις διάφορες τιμές x_i της δ.τ.μ. X.[5]



Σχήμα 4.2.1. Απεικόνιση της συνάρτησης πυκνότητας πιθανότητας $f(x_i)$ για την δ.τ.μ. X

4.2.1. ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΜΕΤΡΑ

Για τη δ.τ.μ. X ορίζονται οι ακόλουθες ποσότητες:

- Η **μέση τιμή μ** (mean value) ή **προσδοκία**, ως η **μέση τιμή του απείρου δείγματος**

$$\mu = \lim_{n\to\infty} \bar{x}_n = \lim_{n\to\infty} \sum_{i=1}^{n} x_i \cdot f_n(x_i) = \sum_{i=1}^{n} x_i \cdot \lim_{n\to\infty} f_n(x_i) \tag{4.2.6}$$

και λόγω της σχέσης (4.2.5) είναι

$$\mu = \sum_{i=1}^{n} x_i \cdot f(x_i) \tag{4.2.7}$$

Επίσης, η μέση τιμή μ απαντάται στη βιβλιογραφία και ως **αναμενόμενη τιμή** (expected value) ή **μαθηματική ελπίδα** (mathematical expectation) και συμβολίζεται με $\boldsymbol{EX}$ ή $\boldsymbol{E[X]}$ ή $\boldsymbol{E(X)}$. Δηλαδή, ισχύει:

$$\mu \equiv E[X] = \sum_{i=1}^{n} x_i \cdot f(x_i) \tag{4.2.8}$$

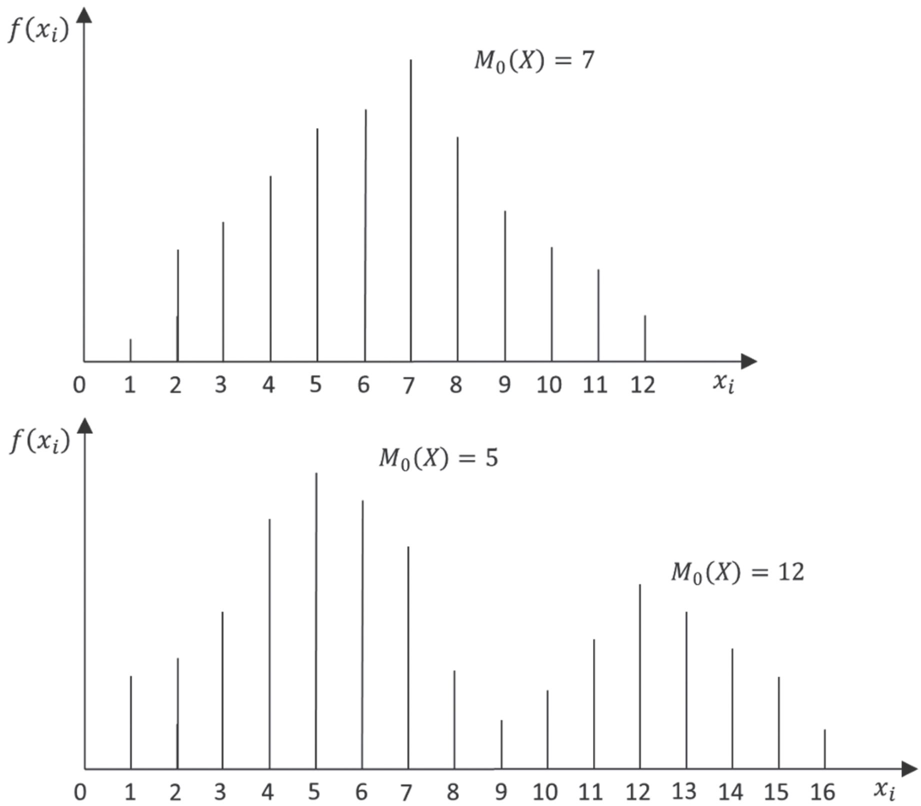
Όταν όλες οι πιθανότητες $f(x_i)$ είναι ίσες, δηλαδή κάθε τιμή της δ.τ.μ. X έχει την ίδια πιθανότητα εμφάνισης $f(x_i) = 1/n$, όπου n είναι το πλήθος των τιμών της X, ισχύει:

$$\mu \equiv E[X] = \sum_{i=1}^{n} x_i \cdot f(x_i) = \sum_{i=1}^{n} x_i \cdot \frac{1}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \bar{x} \tag{4.2.9}$$

Δηλαδή, η αναμενόμενη ή μέση τιμή συμπίπτει με τον αριθμητικό μέσο όρο ή απλά τον μέσο όρο των τιμών x_1, x_2, ..., x_n, της δ.τ.μ. X.

[5] Βλ. Κεφάλαιο. 1.2.6.

- Η **επικρατούσα τιμή**[6] ή **τύπος** (mode), M_0, μιας δ.τ.μ. X είναι εκείνη η τιμή x_i της μεταβλητής για την οποία η συνάρτηση πυκνότητας πιθανότητας $f(x_i)$ είναι μεγαλύτερη από οποιαδήποτε άλλη τιμή της σε μία γειτνίαση της $x_i = M_0(X)$. Εάν η κατανομή εμφανίζει μία ή περισσότερες επικρατούσες τιμές ονομάζεται *μονοκόρυφη*, *δικόρυφη* ή *πολυκόρυφη* κατάνομη, όπως φαίνεται στο Σχήμα 4.2.2.



Σχήμα 4.2.2. Μονοκόρυφη και δικόρυφη κατανομή μιας δ.τ.μ. X

- Η **διακύμανση** (variance) [**μεταβλητότητα** (variability)], $\boldsymbol{\sigma^2}$ ή $\boldsymbol{Var(X)}$, ως η **διασπορά** (variation) **του απείρου δείγματος**[7]

$$\sigma^2 = \lim_{n\to\infty} m^2 = \sum_{i=1}^{n} \lim_{n\to\infty} (x_i - \bar{x}_n)^2 \cdot f_n(x_i) \tag{4.2.10}$$

και

$$\sigma^2 = Var(X) = \sum_{i=1}^{n} (x_i - \mu)^2 \cdot f(x_i) \tag{4.2.11}$$

- Η **μέση τιμή συνάρτησης δ.τ.μ.**

Θεωρείται η συνάρτηση $g(X)$ της δ.τ.μ. X, η οποία αποτελεί μια νέα τυχαία μεταβλητή, έστω την τ.μ. $\boldsymbol{G}$. Σύμφωνα με τη σχέση (4.2.7), η ποσότητα $\sum_{i=1}^{n} g(x_i) \cdot f(x_i)$ εκφράζει τη **μέση τιμή της συνάρτησης** $\boldsymbol{g(X)} \equiv \boldsymbol{G}$ και συμβολίζεται ως:

---

[6] Βλ. Κεφάλαιο 1.4.3.

[7] Βλ. Κεφάλαιο 1.5.5., Α. Δερμάνης, τόμος 1, 1986, σελ. 58

$$E[g(X)] = E[G] = \mu_G = \sum_{i=1}^{n} g(x_i) \cdot f(x_i) \tag{4.2.12}$$

Η ποσότητα $E[g(X)]$ είναι γνωστή και ως **μαθηματική προσδοκία** της $g(X)$.

- Οι **ροπές**[8] (moments)

Αναφορικά με τη μορφή της συνάρτησης $g(X)$ της δ.τ.μ. X, διακρίνονται οι εξής περιπτώσεις:

I. Η συνάρτηση $g(X)$ έχει τη μορφή: $\boldsymbol{g(X) = X^r}$ με $r = 1, 2, 3, \ldots$

Η σχέση (4.2.12) δίνει:

$$E[g(X)] = E[X^r] = \sum_{i=1}^{n} x_i^r \cdot f(x_i) \tag{4.2.13}$$

Η ποσότητα $E[X^r]$ ονομάζεται **ροπή $\boldsymbol{r-}$τάξης της δ.τ.μ. $\boldsymbol{X}$ ως προς την αρχή** (rth moment about the origin) και συμβολίζεται ως:

$$\mu_r = E[X^r], \quad r = 1, 2, 3, \ldots \tag{4.2.14}$$

Για $r = 1$, προκύπτει:

$$\mu_1 = E[X] \tag{4.2.15}$$

Σύμφωνα με τις σχέσεις (4.2.7), (4.2.8), συνάγεται ότι **η ροπή 1ης τάξης συμπίπτει με τη μέση τιμή $\boldsymbol{\mu}$ της τ.μ. $\boldsymbol{X}$**. Δηλαδή,

$$\mu_1 = E[X] \equiv \mu \tag{4.2.16}$$

II. Η συνάρτηση $g(X)$ έχει τη μορφή: $\boldsymbol{g(X) = (X-k)^r}$ με $r = 1, 2, 3, \ldots$

Από τη σχέση (4.2.12) προκύπτει:

$$E[g(X)] = E[(X-k)^r] = \sum_{i=1}^{n} (x_i - k)^r \cdot f(x_i) \tag{4.2.17}$$

Η ποσότητα $E[(X-k)^r]$ ονομάζεται **ροπή $\boldsymbol{r-}$τάξης της δ.τ.μ. $\boldsymbol{X}$ ως προς $\boldsymbol{k}$** και συμβολίζεται ως:

$$\mu_r = E[(X-k)^r], \quad r = 1, 2, 3, \ldots \tag{4.2.18}$$

Στην **περίπτωση** κατά την οποία είναι $\boldsymbol{k = \mu}$, η ποσότητα $E[(X-\mu)^r]$ συμβολίζεται ως

$$\mu_r = E[(X-\mu)^r], \quad r = 1, 2, 3, \ldots \tag{4.2.19}$$

και ονομάζεται **κεντρική ροπή $\boldsymbol{r-}$τάξης της δ.τ.μ. $\boldsymbol{X}$** (rth moment about the mean).

[8] Βλ. Κεφάλαιο 1.6.

Για $r = 2$, προκύπτει:

$$\mu_2 = E[(X-\mu)^2] \tag{4.2.20}$$

Από τις σχέσεις (4.2.11), (4.2.17), (4.2.20) συνάγεται ότι **η κεντρική ροπή 2ης τάξης συμπίπτει με τη διακύμανση $\boldsymbol{\sigma^2}$ ή $\boldsymbol{Var(X)}$ της δ.τ.μ. $\boldsymbol{X}$**, δηλαδή

$$\mu_2 = E[(X-\mu)^2] = \sigma^2 = Var(X) \tag{4.2.21}$$

Η ποσότητα

$$\sigma = \sqrt{\sigma^2} = \sqrt{Var(X)} \tag{4.2.22}$$

ονομάζεται **τυπική απόκλιση** (standard deviation) **της τ.μ. $\boldsymbol{X}$**.

Για $r = 3$, προκύπτει:

$$\mu_3 = E[(X-\mu)^3] \tag{4.2.23}$$

Η σχέση (4.2.23) δίνει την κεντρική ροπή 3ης τάξης. Από τον λόγο αυτής της ροπής προς την ποσότητα σ^3 προκύπτει:

$$\alpha_3 = \frac{\mu_3}{\sigma^3} = \frac{E[(X-\mu)^3]}{\sigma^3} \tag{4.2.24}$$

Η σχέση (4.2.24) εκφράζει τον **συντελεστή** ή την **παράμετρο λοξότητας** (συμμετρία – ασυμμετρία) της κατανομής. Εάν η κατανομή είναι συμμετρική τότε θα είναι $\alpha_3 = 0$. Εάν ισχύει $\alpha_3 > 0$, τότε η κατανομή θα είναι *λοξή προς τα δεξιά*, ενώ εάν ισχύει $\alpha_3 < 0$, τότε η κατανομή θα είναι *λοξή προς τα αριστερά*.[9]

Για $r = 4$, προκύπτει:

$$\mu_4 = E[(X-\mu)^4] \tag{4.2.25}$$

Η σχέση (4.2.25) δίνει την κεντρική ροπή 4ης τάξης. Ο λόγος αυτής της ροπής προς την ποσότητα σ^4 δίνει:

$$\alpha_4 = \frac{\mu_4}{\sigma^4} = \frac{E[(X-\mu)^4]}{\sigma^4} \tag{4.2.26}$$

Η σχέση (4.2.26) εκφράζει τον **συντελεστή** ή την **παράμετρο κύρτωσης** της κατανομής. Εάν είναι $\alpha_4 > 0$, τότε η κατανομή θα είναι *λεπτόκυρτη* ή *οξύκυρτη*, εάν είναι $\alpha_4 < 0$, τότε η κατανομή θα είναι *πλατύκυρτη*, και εάν είναι $\alpha_4 = 0$, τότε η κατανομή θα είναι *μεσόκυρτη*.[10]

[9] Βλ. Κεφάλαιο 1.6.2.
[10] Βλ. Κεφάλαιο 1.6.3.

5.

ΠΡΟΣΔΙΟΡΙΣΜΟΣ ΜΕΤΑΒΛΗΤΩΝ

5.1. ΕΙΣΑΓΩΓΗ

Η μελέτη του γεωγραφικού χώρου είναι μία **σύνθετη και δυναμική διαδικασία** η οποία συναρτάται με:

1. τα *δεδομένα ή στοιχεία* τα οποία (ως πληροφορίες) είτε περιγράφουν τον γεωγραφικό χώρο αναφοράς, είτε απορρέουν από αυτόν,
2. τις *στατιστικές μεθόδους επεξεργασίας* των δεδομένων,
3. τις *στατιστικές μεθόδους ανάλυσης* των πληροφοριών που προκύπτουν από τα δεδομένα, με βάση κάποιο μαθηματικό μοντέλο,
4. την *εξαγωγή συμπερασμάτων* συμβατών με το φαινόμενο ή ενδεχομένως το «πρόβλημα» που οδήγησε στη θεωρούμενη μελέτη, καθώς και συμπερασμάτων που οδηγούν σε «νέες» θεωρήσεις του χώρου,
5. τη *στατική ή / και δυναμική παρουσίαση* των πληροφοριών με διαγράμματα, πίνακες και, **κυρίως**, χάρτες.

Οι **προκύπτουσες πληροφορίες** και ο **ιστορικός χρόνος** αποτελούν τα νέα **δεδομένα** που με τη σειρά τους θα δώσουν τη «νέα» οπτική του γεωγραφικού χώρου κ.ο.κ. Πρόκειται για μία συνεχή επαναλαμβανόμενη κίνηση όπου, το φυσικό και ανθρώπινο περιβάλλον αλληλοεπιδρούν δυναμικά μεταξύ τους, με αποτέλεσμα τη *διαρκή μεταβολή του γεωγραφικού* *χώρου*. Η μεταβολή αυτή δεν είναι παρά οι διαφοροποιήσεις που επέρχονται στη δομή της φυσικής, κοινωνικής, οικονομικής, πολιτιστικής και περιβαλλοντικής διάστασης του γεωγραφικού χώρου με την επίδραση **και** του χρόνου.

Όταν λέγεται «**μελέτη του γεωγραφικού χώρου**», εννοείται ότι αυτός ο χώρος έχει προσδιορισθεί, δηλαδή έχει «οριοθετηθεί» και, επί πλέον, έχει καθορισθεί (και) η «**μονάδα χωρικής αναφοράς**». Πρόκειται για τη *μικρότερη χωρική επιφάνεια του συγκεκριμένου γεωγραφικού χώρου όπου θα αναζητηθούν δεδομένα*. Η επιφάνεια αυτή προσδιορίζεται από το μέγεθος του προς μελέτη χώρου και από το «επίπεδο λεπτομέρειας» των δεδομένων που θα χρησιμοποιηθούν και των πληροφοριών που θα (ή «απαιτείται» να) προκύψουν. Για παράδειγμα, η μελέτη του γεωγραφικού χώρου ενός νομού, αναφορικά με τα φαινόμενα της μετανάστευσης και της ανεργίας, οριοθετεί τον νομό Α με μονάδα χωρικής αναφοράς τον οικισμό. Αυτό σημαίνει ότι *η αναζήτηση των σχετικών δεδομένων θα γίνει σε επίπεδο οικισμού*. Εάν ο γεωγραφικός χώρος είναι ένα αστικό κέντρο, το ίδιο πρόβλημα αντιμετωπίζεται καθορίζοντας ως «μονάδα χωρικής αναφοράς» το οικοδομικό τετράγωνο (Ο.Τ.) και *η αναζήτηση των σχετικών δεδομένων θα γίνει σε επίπεδο οικοδομικού τετραγώνου*. Είναι σαφές ότι το μέγεθος της χωρικής μονάδας δίνει τη δυνατότητα για τον *προσδιορισμό ιδιαιτεροτήτων* στην περιοχή αναφοράς, οι οποίες δεν είναι εμφανείς από την αρχή.

Από τα παραπάνω προκύπτει ότι, το *ίδιο φαινόμενο σε διαφορετικούς γεωγραφικούς χώρους*, είτε η διαφορετικότητα σχετίζεται με το μέγεθος είτε αφορά την «προέλευση» / «φύση» του χώρου, συναρτάται με «διαφορετικά» σύνολα δεδομένων, των οποίων *το μέγεθος και το περιεχόμενο εξαρτώνται από τη μονάδα χωρικής αναφοράς*. Καθίσταται φανερό ότι, άλλη είναι η απαίτηση για την αναζήτηση των αναγκαίων στοιχείων σε επίπεδο οικισμού και πολύ περισσότερο απαιτητική είναι η αναζήτηση αυτή σε επίπεδο Ο.Τ. *Ο ίδιος ο γεωγραφικός χώρος υποδεικνύει τα είδη και το πλήθος αυτών των δεδομένων*. Δεν τίθεται θέμα «καλύτερων» ή «χειρότερων» στοιχείων, αλλά «μεγαλύτερης» ή «μικρότερης» λεπτομέρειας στην ανάλυση των φαινομένων που απασχολούν ή στην ανάλυση του «προβλήματος» που ενδιαφέρει. Επομένως, ο *βαθμός αυτής της λεπτομέρειας*, η **κλίμακα** όπως λέγεται, καθορίζει και προσδιορίζει τα είδη και το πλήθος των δεδομένων, καθώς και την ύπαρξη γεωγραφικών ιδιαιτεροτήτων. Όμως, η μελέτη οποιουδήποτε φαινομένου, φυσικού, οικονομικού, κοινωνικού, περιβαλλοντικού κ.ά., στον γεωγραφικό χώρο, συναρτάται πάντοτε *και με την τέταρτη διάσταση*, τον **χρόνο** ή τον **ιστορικό χρόνο**, αφού κάθε φαινόμενο έχει παρελθόν που καθορίζει το παρόν και προδιαγράφει το μέλλον του.

Καθίσταται σαφές, και αποτελεί τη βάση επάνω στην οποία θα δομηθεί κάθε μελέτη του οποιουδήποτε γεωγραφικού χώρου, ότι για την «ανάγνωση» αυτού απαιτείται, ως πρώτο βήμα, κάποιο (αναγκαίο) *πλήθος δεδομένων* ή *στοιχείων που αναφέρεται σε αυτόν*. Επίσης, εάν είναι γνωστά τα δεδομένα ενός γεωγραφικού χώρου σε κάποιο (παρελθόντα) χρόνο, είναι δυνατός ο προσδιορισμός του για τον χρόνο αναφοράς. Με άλλα λόγια, *η σχέση μεταξύ γεωγραφικού χώρου και δεδομένων του είναι μία σχέση δυναμική και αμφίδρομη*. Συνεπώς, ποια είναι αυτά τα δεδομένα τα οποία, αφού αναφέρονται στον γεωγραφικό χώρο, μπορούν να λέγονται **γεωγραφικά δεδομένα** ή **γεωγραφικά στοιχεία**;

Τα δεδομένα αυτά αποτελούν ένα σύνολο ποιοτικών και ποσοτικών χαρακτηριστικών ή ιδιοτήτων που αναφέρονται στον (ή προέρχονται από τον) θεωρούμενο γεωγραφικό χώρο. Η επεξεργασία τους παράγει την **πληροφορία** ή τις **πληροφορίες**, δηλαδή τις απαντήσεις σε ερωτήματα που τίθενται και σχετίζονται με την «ανάγνωση» και ανάλυση του συγκεκριμένου χώρου στον **συγκεκριμένο χρόνο**. Για παράδειγμα, ο πληθυσμός και η έκταση αποτελούν δεδομένα για ένα οικισμό. Η επεξεργασία τους, που αναφέρεται στον λόγο αυτών των δεδομένων, δίνει την πυκνότητα κατοίκησης του οικισμού, δηλαδή την πληροφορία που απαντάει στο ερώτημα: «πόσοι κάτοικοι αντιστοιχούν ανά τετρ. χλμ.;». Όμως, η «κίνηση» του χώρου μέσα στον χρόνο είναι διαρκής. Έτσι, οι πληροφορίες αυτές αποτελούν τα δεδομένα του γεωγραφικού χώρου για κάποια άλλη μελλοντική, κοντινή ή μακρινή, χρονική αναφορά. Βεβαίως, και στον θεωρούμενο χρόνο όπου εκπονείται η μελέτη ενός χώρου, οι πληροφορίες που προκύπτουν μπορούν να χρησιμοποιηθούν ως δεδομένα για κάποιο άλλο μεταγενέστερο στάδιο της ανάλυσής του. Έτσι, για το ίδιο παράδειγμα, η πληροφορία «πυκνότητα κατοίκησης» σε ένα μελλοντικό χρόνο αποτελεί «ιστορικό» δεδομένο ή στοιχείο του οικισμού, ενώ κατά τη διάρκεια της μελέτης του οικισμού αποτελεί και πάλι δεδομένο για μία επόμενη φάση της ανάλυσης, όπως ο υπολογισμός των αναγκαίων χώρων πρασίνου με βάση την πυκνότητα κατοίκησης. Φαίνεται, λοιπόν, ότι η διαφοροποίηση σε δεδομένα ή στοιχεία και σε πληροφορίες είναι ένας *τεχνητός διαχωρισμός* ο οποίος λειτουργεί μόνο για τις διάφορες χωροχρονικές φάσεις μίας γεωγραφικής μελέτης. Επομένως, ίσως είναι καλύτερο να ειπωθεί ότι *τα στοιχεία και οι πληροφορίες του γεωγραφικού χώρου αποτελούν τα δεδομένα που «απορρέουν» από αυτόν, αποτελούν δηλαδή τα γεωγραφικά δεδομένα*.

Ένας άλλος διαχωρισμός διακρίνει τα δεδομένα σε **πρωτογενή** και **δευτερογενή**. Ως *πρωτογενή* θεωρούνται αυτά που προκύπτουν άμεσα από την παρατήρηση ή το «πείραμα» (μετρήσεις, καταγραφές / απαριθμήσεις, ερωτηματολόγια), ενώ ως *δευτερογενή* θεωρούνται αυτά που προκύπτουν από τον συνδυασμό ή την επεξεργασία των πρωτογενών. Έτσι, ως δευτερογενή θεωρούνται τα δεδομένα που διατίθενται από διάφορες Υπηρεσίες, π.χ. από την Ελληνική Στατιστική Υπηρεσία[1], ή τα δεδομένα που προκύπτουν από μελέτες που έχουν ήδη εκπονηθεί για τον γεωγραφικό χώρο ενδιαφέροντος. Φαίνεται καθαρά ότι *τα δευτερογενή δεδομένα δεν είναι παρά οι πληροφορίες της προηγούμενης παραγράφου*. Επισημαίνεται, για την αξιοπιστία της γεωγραφικής μελέτης, η ιδιαίτερη προσοχή που απαιτείται κατά τη χρησιμοποίηση δευτερογενών δεδομένων τα οποία προέρχονται από διάφορες πηγές και δεν προκύπτουν από την επεξεργασία των πρωτογενών.

Αναφορικά με την κλίμακα, δηλαδή τον επιζητούμενο βαθμό λεπτομέρειας, τα γεωγραφικά δεδομένα διακρίνονται σε **πρωτοβάθμια** και **δευτεροβάθμια**. Ως *πρωτοβάθμια* αναφέρονται τα δεδομένα που λειτουργούν σε κλίμακα ολότητας για τον γεωγραφικό χώρο ενδιαφέροντος, δηλαδή χωρίς απαίτηση λεπτομέρειας. Ως *δευτεροβάθμια* αναφέρονται αυτά στα οποία ο βαθμός λεπτομέρειας είναι αντίστοιχος της μονάδας χωρικής αναφοράς. Και αυτός ο διαχωρισμός είναι τεχνητός και αποβλέπει αφ' ενός, στην απρόσκοπτη και ελεγχόμενη συλλογή των δεδομένων και αφ' ετέρου, στη διατήρηση μίας αντιστοιχίας μεταξύ δεδομένων και γεωγραφικού χώρου.

Μία άλλη διαφοροποίηση μεταξύ των γεωγραφικών δεδομένων είναι αυτή που τα διακρίνει σε **ποιοτικά** και **ποσοτικά** ή **αριθμητικά**. *Ποιοτικά* λέγονται εκείνα τα δεδομένα που εκφράζουν ένα χαρακτηριστικό ή μία ιδιότητα και *δεν έχουν μέγεθος*. Δηλαδή, δεν μπορούν να μετρηθούν παρά μόνο να *απαριθμηθούν* ή να *περιγραφούν*. Π.χ. το φύλο των ανθρώπων, το χρώμα των ματιών, η οικογενειακή κατάσταση, το μορφωτικό επίπεδο, το επάγγελμα, η ψυχολογική κατάσταση ενός ατόμου, κ.ά. *Ποσοτικά* λέγονται εκείνα τα δεδομένα που αντιπροσωπεύουν ένα χαρακτηριστικό ή μία ιδιότητα και *έχουν μέγεθος* ή *εκφράζουν μία «ποσότητα»*. Επομένως, μπορούν να *μετρηθούν*, να πάρουν ορισμένες αριθμητικές τιμές, να αναλυθούν κατά τάξη μεγέθους, κ.λ.π. Π.χ. το ύψος, το βάρος, η έκταση των γεωργικών καλλιεργειών, το οικογενειακό εισόδημα, η ηλικία, ο αριθμός των παιδιών μίας οικογένειας, κ.ά.

Όταν λέγεται ότι *ο γεωγραφικός χώρος μεταβάλλεται*, αυτό σημαίνει ότι:

[1] Ε.Σ.Υ.Ε. ή ΕΛ.ΣΤΑΤ.

• αφ' ενός, δεν είναι ο ίδιος στις διάφορες περιφέρειές του (στις διάφορες υποδιαιρέσεις του) στον χρόνο αναφοράς, δηλαδή υφίστανται «*ενδοπεριφερειακές*» *διαφοροποιήσεις*, και

• αφ' ετέρου, αυτές οι διαφοροποιήσεις μεταβάλλονται περαιτέρω με την επίδραση (και) του χρόνου.

Κατά συνέπεια, τα δεδομένα που τον περιγράφουν ή εκπορεύονται από αυτόν *μεταβάλλονται* ή *διαφοροποιούνται από χωρική ενότητα σε χωρική ενότητα, είτε ανεξαρτήτως είτε σε σχέση (και) με τον χρόνο*. Έχει αναφερθεί ότι η αναζήτηση των δεδομένων γίνεται σε «επίπεδο χωρικής μονάδας αναφοράς». Αυτό σημαίνει ότι όλα τα χαρακτηριστικά και οι κάθε είδους πληροφορίες που αναφέρονται σε ένα γεωγραφικό χώρο προέρχονται από το «σύνολο των χωρικών μονάδων αναφοράς», στις οποίες αυτός *περιφερειοποιείται*. Το σύνολο αυτό δεν είναι άλλο από τον «*Πληθυσμό των χωρικών μονάδων αναφοράς*» *ο οποίος*, σύμφωνα με τη Στατιστική σημειολογία, *περιγράφει τον συγκεκριμένο γεωγραφικό χώρο*.

Όμως, μέσα σε κάθε γεωγραφικό χώρο, όπως έχει αναφερθεί, λαμβάνουν χώρα διάφορα φαινόμενα (φυσικά, οικονομικά, κοινωνικά, περιβαλλοντικά, πολιτιστικά, κ.ά.), η μελέτη των οποίων συναρτάται άμεσα με αυτόν και γι' αυτό αναφέρονται ως **γεωγραφικά φαινόμενα**. Συνάγεται, λοιπόν, ότι «*μελέτη γεωγραφικού χώρου*» και «*μελέτη γεωγραφικού φαινομένου*» είναι *έννοιες ταυτόσημες*, εφ' όσον το πρώτο προϋποθέτει το δεύτερο και τανάπαλιν. Με άλλα λόγια, *φαινόμενα και γεωγραφικός χώρος βρίσκονται σε διαρκή δράση και ανάδραση μεταξύ τους*. Κατά συνέπεια, *με τον όρο γεωγραφικά δεδομένα εννοούνται τα κάθε είδους χαρακτηριστικά και ιδιότητες που αφορούν τα γεωγραφικά φαινόμενα ή τα φαινόμενα ενός γεωγραφικού χώρου ή ένα γεωγραφικό χώρο, όμως η «μονάδα αναζήτησης των δεδομένων» δεν είναι πάντοτε χωρική*. Για παράδειγμα, εάν ζητείται να μελετηθεί η συμπεριφορά των οικονομικών μεταναστών σε μία επαρχία και σε ένα αστικό κέντρο, «*η μονάδα αναζήτησης δεδομένων*» *είναι ο κάθε οικονομικός μετανάστης*, ξεχωριστά στην επαρχία και ξεχωριστά στο αστικό κέντρο. Τα δύο, *διακριτά μεταξύ τους, σύνολα* αποτελούν, αφ' ενός, τον Πληθυσμό των Οικονομικών Μεταναστών στην Επαρχία και, αφ' ετέρου, τον Πληθυσμό των Οικονομικών Μεταναστών στο Αστικό Κέντρο. Γενικά, πρόκειται για *δύο διαφορετικά μεταξύ τους σύνολα δεδομένων με διαφορετικές χωρικές αναφορές*. Καθίσταται, λοιπόν, σαφές ότι στο στατιστικό σύνολο Πληθυσμός: χωρικών ενοτήτων, προσώπων, αντικειμένων, περιπτώσεων και γενικά ατόμων ή οντοτήτων, *ως* «*μονάδες αναζήτησης δεδομένων*» *λειτουργούν τα άτομα ή οι οντότητες του Πληθυσμού*. Αντιστοίχως, εάν το στατιστικό σύνολο είναι ένα τυχαίο και αντιπροσωπευτικό Δείγμα του Πληθυσμού, τις «*μονάδες αναζήτησης δεδομένων*» *αποτελούν τα άτομα ή οι οντότητες του Δείγματος*.

Αφού οι οντότητες ενός στατιστικού συνόλου είναι διαφορετικές μεταξύ τους, είναι σαφές ότι και τα δεδομένα που αναφέρονται σε αυτές, δηλαδή τα χαρακτηριστικά και οι ιδιότητες που απορρέουν από αυτές τις οντότητες, διαφέρουν μεταξύ τους ή, όπως λέγεται, «παίρνουν ή έχουν διαφορετικές τιμές» για κάθε οντότητα του στατιστικού συνόλου. Με άλλα λόγια, *οι τιμές των χαρακτηριστικών και των ιδιοτήτων ενός στατιστικού συνόλου μεταβάλλονται από οντότητα σε οντότητα* (αυτό όμως δεν αποκλείει το γεγονός να υπάρχουν δύο ή περισσότερες οντότητες με ίδιες τις τιμές κάποιων δεδομένων). Έτσι, προκύπτει ένας άλλος διαχωρισμός των δεδομένων που τα διακρίνει σε **μεταβλητές** και **σταθερές**.

5.2. ΜΕΤΡΗΣΗ ΜΕΤΑΒΛΗΤΩΝ

Ορίζεται ως *μεταβλητή* (variable) το χαρακτηριστικό / η ιδιότητα που παίρνει διαφορετικές τιμές για τις οντότητες ενός στατιστικού συνόλου ή έχει διαφορετικές τιμές στην ίδια στατιστική οντότητα μετά από επαναλήψεις «μετρήσεων». Αντιθέτως, εάν το χαρακτηριστικό /

η ιδιότητα δεν εμφανίζει μεταβολές στις τιμές του / της για κάθε οντότητα του στατιστικού συνόλου ή στις διάφορες επαναληπτικές «μετρήσεις» για την ίδια οντότητα, δηλαδή έχει *μία* μόνο *τιμή*, ονομάζεται *σταθερά* (constant). Για παράδειγμα, μεταβλητές είναι το φύλο των ανθρώπων, η ηλικία, το μορφωτικό επίπεδο, το εισόδημα, ο τόπος καταγωγής, το θρήσκευμα, η νοημοσύνη, οι πολιτικές προτιμήσεις, το μέγεθος των νοικοκυριών, το επάγγελμα, το μέγεθος του πληθυσμού, το ποσοστό γεννήσεων, η πυκνότητα κατοίκησης, η αναλογία δασκάλων σε μαθητές, η οικογενειακή κατάσταση κ.ά. σε ορισμένη γεωγραφική περιοχή σε κάποια χρονική περίοδο. Το χαρακτηριστικό – δείκτης: «αριθμός νοσοκομειακών ιατρών ανά κάτοικο», αποτελεί μία σταθερά κατά τη γεωγραφική ανάλυση ενός νομού, διότι έχει την ίδια τιμή για κάθε κάτοικο του νομού. Η ιδιότητα ή το χαρακτηριστικό «Υπηκοότητα» που αναφέρεται στους πολίτες της Ελλάδας είναι μία σταθερά, διότι έχει την ίδια τιμή (Ελληνική Υπηκοότητα) για κάθε πολίτη της Ελλάδας.

Επίσης, θα πρέπει να σημειωθεί ότι, ένα χαρακτηριστικό που είναι μεταβλητή για ένα στατιστικό σύνολο, είναι πιθανό να είναι σταθερά για κάποιο άλλο και το αντίστροφο. Έτσι, εάν το στατιστικό σύνολο είναι ο Ελλαδικός Χώρος με χωρική μονάδα αναφοράς τον νομό, τότε ο δείκτης «αριθμός νοσοκομειακών ιατρών ανά κάτοικο» είναι μία μεταβλητή, διότι έχει διαφορετική τιμή για κάθε νομό. Παρομοίως, η ιδιότητα «Υπηκοότητα» για τους πολίτες της Ευρώπης είναι μία μεταβλητή διότι παίρνει διαφορετικές τιμές αναφορικά με τις χώρες που απαρτίζουν την Ευρώπη (Ελληνική Υπηκοότητα, Γαλλική, Γερμανική, Ισπανική κ.λ.π.).

Είναι φανερό, για παράδειγμα, ότι οι κάτοικοι του Δήμου Α (θα) διαφέρουν μεταξύ τους ως προς το φύλο, το βάρος, το ύψος, την οικογενειακή κατάσταση, το εισόδημα, την καταγωγή, τις πολιτικές πεποιθήσεις κ.λ.π. Ο κάθε κάτοικος απεικονίζει μία «*ποσότητα*» *πληροφοριών που μεταβάλλεται από άτομο σε άτομο*, δηλαδή οι *σχετικές μεταβλητές διαφοροποιούνται* μεταξύ των κατοίκων για τον Δήμο Α σε δεδομένο χρόνο. Συνεπώς, για να είναι δυνατός ο εντοπισμός αυτών των μεταβολών, θα πρέπει να βρεθεί ένας τρόπος σύμφωνα με τον οποίο *θα μπορεί να αποδοθεί το περιεχόμενο κάθε μεταβλητής* σε μορφή που θα διευκολύνει τη σύγκριση των μεταβολών μεταξύ των κατοίκων του Δήμου Α ή μεταξύ των οντοτήτων ενός στατιστικού συνόλου. Αυτό σημαίνει ότι θα πρέπει να «**μετρηθεί**» η «ποσότητα» της πληροφορίας που περιλαμβάνει κάθε μεταβλητή, γεγονός που επιτυγχάνεται είτε με τη βοήθεια του κατάλληλου *οργάνου* είτε με την κατάλληλη *κωδικοποίηση*, έτσι ώστε να δοθούν «**τιμές**» στις μεταβλητές για κάθε οντότητα αναφοράς τους.

Ως «**μέτρηση μεταβλητής**» εννοείται κάθε εμπειρική διαδικασία με τη βοήθεια της οποίας, σύμφωνα με ορισμένους κανόνες, αποδίδεται (ή αντιπροσωπεύεται) με σύμβολα η «ποσότητα» της πληροφορίας που περικλείει μία μεταβλητή για την κάθε στατιστική οντότητα. Επομένως, *με τη μέτρηση αντιστοιχίζονται* **«τιμές»** *από ένα γνωστό σύνολο συμβόλων, του οποίου τα στοιχεία είναι είτε αριθμοί* (όπως το σύνολο των ακεραίων, των φυσικών, των πραγματικών αριθμών), *είτε άλλα σύμβολα* (όπως π.χ. το χρώμα των ματιών, οι ΟΤΑ[2] της χώρας, τα διάφορα θρησκεύματα κ.λ.π.), *στη μεταβλητή ενδιαφέροντος για την κάθε οντότητα του στατιστικού συνόλου*. Προφανώς, η αντιστοίχιση αυτή πρέπει να είναι, κατά το δυνατόν, *σαφής* και *ακριβής* και αυτό μπορεί να εξασφαλισθεί με το **κατάλληλο** «**εργαλείο μέτρησης**».

Το εργαλείο μέτρησης είναι κατάλληλο όταν παρέχει **Αξιοπιστία** (reliability) και **Εγκυρότητα** (validity), δηλαδή όταν:

- Είναι **ικανό** *να διακρίνει διαφορές* στην τάξη μεγέθους που ενδιαφέρει τον ερευνητή. Π.χ. δεν είναι δυνατή η μέτρηση ή η έκφραση του βάρους σε gr όταν τα σταθμά είναι σε kgr.

[2] Οργανισμός Τοπικής Αυτοδιοίκησης (Ο.Τ.Α.)

6.

ΤΕΧΝΙΚΕΣ ΣΥΛΛΟΓΗΣ ΓΕΩΓΡΑΦΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

6.1. ΕΙΣΑΓΩΓΗ

Όπως έχει αναφερθεί, η μελέτη του γεωγραφικού χώρου συναρτάται με την ανάλυση και τη σύνθεση των κάθε είδους στοιχείων και πληροφοριών που απορρέουν από αυτόν σε συγκεκριμένο ιστορικό χρόνο. Είναι προφανές ότι *καμιά στατιστική και χαρτογραφική επεξεργασία δεν μπορεί να εφαρμοσθεί εάν δεν είναι διαθέσιμα τα γεωγραφικά δεδομένα*, δηλαδή όλα τα στοιχεία και οι πληροφορίες φυσικού, δημογραφικού, οικονομικού, κοινωνικού, πολιτιστικού, περιβαλλοντικού κ.ά. χαρακτήρα, που αναφέρονται στον γεωγραφικό χώρο ενδιαφέροντος. Ο ιστορικός χρόνος εκφράζει την απαραίτητη παράμετρο, την απαραίτητη συνθήκη, η γεωγραφική μελέτη να μπορεί να λειτουργεί συγχρόνως: i. στο επίπεδο του παρόντος, με τεκμηριωμένες προβλέψεις για το μέλλον, και ii. στο επίπεδο του παρελθόντος, με τεκμηριωμένες απόψεις για το παρόν. Οι στατιστικές επεξεργασίες, που μεταφράζονται σε πίνακες και διαγράμματα, και οι χαρτογραφικές απεικονίσεις των φαινομένων αναφοράς, καθιστούν τον «αφηρημένο» ή μακρινό γεωγραφικό χώρο μία οντότητα προσιτή και διαχειρίσιμη.

Επομένως, το *απαραίτητο υπόβαθρο* επάνω στο οποίο θα θεμελιωθεί βήμα – βήμα μία γεωγραφική μελέτη οποιουδήποτε χαρακτήρα, είναι η ύπαρξη των απαιτούμενων γεωγραφικών δεδομένων, τόσο κατά τον χρόνο της γεωγραφικής έρευνας όσο και σε παλαιότερες χρονικές περιόδους. Όμως, αυτή η διαθεσιμότητα των γεωγραφικών δεδομένων θέτει το μεγάλο πρόβλημα της συλλογής τους.

Ένας γεωγραφικός χώρος συνίσταται από μία ή περισσότερες μονάδες χωρικής αναφοράς, ανάλογα με το είδος και την κλίμακα της μελέτης. Για παράδειγμα, ένας νομός αποτελείται από δήμους και κοινότητες, δηλαδή από ΟΤΑ. Ως μονάδα χωρικής αναφοράς του νομού θεωρείται η επιφάνεια του ΟΤΑ. Μπορεί, όμως, για άλλου είδους μελέτη, ως μονάδα χωρικής αναφοράς του ίδιου νομού, να θεωρηθεί ο οικισμός. Συνάγεται, λοιπόν, ότι *εάν τα γεωγραφικά δεδομένα έχουν χωρική αναφορά, η συλλογή τους συναρτάται με το είδος της χωρικής αναφοράς. Τα προς αναζήτηση δεδομένα ανήκουν στην ίδια κατηγορία, διαφέρουν όμως στη λεπτομέρεια*. Έτσι, για παράδειγμα, η απασχόληση στον δευτερογενή τομέα είναι ένα δεδομένο που ανήκει στην κατηγορία των οικονομικών χαρακτηριστικών. Αυτό το δεδομένο μπορεί να εμφανίζεται ως:

● Δείκτης δευτερογενή τομέα σε επίπεδο νομού,

● Πλήθος και αναλογία εργαζόμενων σε συγκεκριμένους κλάδους του δευτερογενή τομέα σε επίπεδο ΟΤΑ, ένας ή περισσότεροι δείκτες,

● Ηλικίες, πλήθος, αναλογία και κλάδος για τους εργαζόμενους στον δευτερογενή τομέα σε επίπεδο αστικού κέντρου, όπου επί πλέον υπάρχει και μπορεί να εμφανίζεται και χαρτογραφικά η χωρική συγκέντρωση ή διασπορά αυτών των ατόμων.

Κατά συνέπεια, η γεωγραφική έρευνα προσεγγίζει το πλήθος των συλλεγόμενων δεδομένων ανάλογα με την κλίμακα της γεωγραφικής μελέτης, η οποία συναρτάται άμεσα με τη μονάδα χωρικής αναφοράς. Τα γεωγραφικά δεδομένα, *είτε χωρικά είτε περιγραφικά*, έχουν πάντα συγκεκριμένη χωρική αναφορά: κράτος, περιφέρεια, νομός, ΟΤΑ, οικισμός, συνοικία, αγρόκτημα, κ.ά.

Οι διάφορες μέθοδοι συλλογής δεδομένων, μπορούν να διακριθούν σε δύο μεγάλες κατηγορίες:[1]

1. Στην *κατηγορία των εξαντλητικών ερευνών*
2. Στην *κατηγορία των δειγματοληψιών*

6.2. ΕΞΑΝΤΛΗΤΙΚΕΣ ΕΡΕΥΝΕΣ

Η κατηγορία αυτή περιλαμβάνει, ως μεθόδους συλλογής δεδομένων, τις *απογραφές* και τις *συνεχείς καταγραφές*.

6.2.1. ΑΠΟΓΡΑΦΕΣ

Η απογραφή διενεργείται όταν το ζητούμενο είναι η συλλογή πληροφοριών για ορισμένα χαρακτηριστικά, ορισμένες μεταβλητές, *ολόκληρου του πληθυσμού των οντοτήτων για τις οποίες απαιτούνται δεδομένα*. Επί πλέον, με την απογραφή γίνεται προσπάθεια να διερευνηθούν και οι διαρθρωτικές σχέσεις που υπάρχουν μεταξύ των διαφόρων μεταβλητών. Έτσι, για παράδειγμα, όταν γίνεται *γεωργική απογραφή* καταγράφονται: οι εκτάσεις που καλλιεργούνται, οι αγραναπαύσεις, τα είδη των καλλιεργειών σε έκταση και παραγωγή, ο αριθμός και τα είδη των γεωργικών μηχανημάτων, κ.λ.π., του αγροκτήματος που απογράφεται. Όταν γίνεται *δημογραφική απογραφή*, συγκεντρώνονται τα δεδομένα που αναφέρονται στο φύλο, στην ηλικία, στο μορφωτικό επίπεδο, στο επάγγελμα, στην οικογενειακή κατάσταση, κ.λ.π., του ατόμου που απογράφεται. Στην πρώτη περίπτωση, ο πληθυσμός των οντοτήτων είναι ο *πληθυσμός των αγροκτημάτων της χώρας*· στη δεύτερη περίπτωση, ο πληθυσμός των οντοτήτων είναι ο *πληθυσμός των κατοίκων της χώρας*. Βεβαίως, και στις δύο περιπτώ-

[1] Γ. Παπαδημητρίου, τεύχος 1, 1990, σελ. 25

σεις καταγράφεται και η χωρική αναφορά της απογραφόμενης οντότητας έτσι ώστε, ανάλογα με το είδος και την απαιτούμενη λεπτομέρεια της γεωγραφικής έρευνας, η μονάδα χωρικής αναφοράς μπορεί να είναι το αγρόκτημα, ο οικισμός, ο ΟΤΑ, ο νομός, κ.λ.π. Δηλαδή, *ο πληθυσμός που απογράφεται μπορεί να έχει και μπορεί να μην έχει χωρικά (με την έννοια της γεωμετρίας) χαρακτηριστικά, όμως οι οντότητές του έχουν πάντα χωρική μονάδα αναφοράς.*

Μία απογραφή έχει συγκεκριμένο χρόνο αναφοράς: την ημέρα ή το χρονικό διάστημα που διενεργείται. Προφανώς, σε μία απογραφή συγκεντρώνονται στοιχεία και πληροφορίες για πληθυσμούς οντοτήτων ολόκληρης της χώρας. Πρόκειται, με άλλα λόγια, για εθνικές απογραφές. Έτσι, εκτός από τον πληθυσμό των κατοίκων, απογράφεται ο πληθυσμός των σχολείων, ο πληθυσμός των εγκαταστάσεων δευτερογενή τομέα, ο πληθυσμός των καταστημάτων τριτογενή τομέα, κ.ά. Επειδή με την απογραφή συλλέγονται δεδομένα σε συγκεκριμένο και προσδιορισμένο χρονικό διάστημα, είναι απαραίτητη η κατάλληλη προπαρασκευή για την επιτυχή έκβασή της, ενώ επί πλέον οι απαιτούμενες δαπάνες είναι υψηλές. Αυτός είναι και ο λόγος που οι εθνικές απογραφές διενεργούνται κάθε 5 ή 10 έτη, ανάλογα με το περιεχόμενό τους, δηλαδή ανάλογα με το είδος των απογραφόμενων πληθυσμών. Η πιστότητα και η πληρότητα των συλλεγόμενων δεδομένων θεωρείται υψηλή εφ' όσον, βεβαίως, οι παράμετροι που έχουν τεθεί κατά τη συλλογή τους είναι σύμφωνες με τους κανόνες της Στατιστικής επιστήμης και δεν εξυπηρετούν άλλες σκοπιμότητες. Για παράδειγμα, (μπορεί να) εμφανίζονται ως κατοικημένες κάποιες περιοχές, οι οποίες στην πραγματικότητα λειτουργούν ως θερινά θέρετρα, προκειμένου να απορροφηθούν οι ανάλογοι πόροι από τις αντίστοιχες νομαρχιακές αυτοδιοικήσεις.

Η αξιολόγηση και η επεξεργασία των δεδομένων μίας καθ' όλα «σωστής» απογραφής, δίνει τα περισσότερο άρτια και αντικειμενικά αποτελέσματα που μπορούν να επιτευχθούν. Όμως, και σε μία τέτοια απογραφή υπάρχουν μειονεκτήματα:

1. Το *μεγάλο κόστος*. Για να επιτευχθεί μία απογραφή απαιτείται σωστή προεργασία και ικανός αριθμός απογραφέων. Στην Ελλάδα, κάθε δεκαετία διενεργείται η απογραφή του πληθυσμού της και κάθε πενταετία γίνεται η απογραφή των βιομηχανιών και βιοτεχνιών της.
2. Είναι *απαραίτητη η εκπαίδευση του μεγάλου πλήθους των απογραφέων* προκειμένου:
   i. να μηδενισθούν τα προσωπικά σφάλματα
   ii. να τελειώσει η απογραφή για τον καθένα μέσα στο συγκεκριμένο χρονικό διάστημα, έτσι ώστε «να μην απογράφονται σπίτια και άνθρωποι από μακριά»
3. Ο *μεγάλος όγκος των στοιχείων και πληροφοριών* που προκύπτουν από μία τέτοια απογραφή, δεν επιτρέπει τη σύντομη δημοσίευση των αποτελεσμάτων έτσι ώστε, όταν μπορούν αυτά να διατεθούν να μην είναι, πια, επίκαιρα.

Επειδή, η σύνθεση και η δομή του πληθυσμού, καθώς και άλλα δημογραφικά δεδομένα, μόνο με την απογραφή μπορούν να διαπιστωθούν, τα πληθυσμιακά αποτελέσματα είναι τα πρώτα που δημοσιεύονται. Χαρακτηρίζονται ως «προσωρινά αποτελέσματα», κατά την πρώτη δημοσίευσή τους, και ως «τελικά αποτελέσματα», στη συνέχεια.

6.2.2. ΣΥΝΕΧΕΙΣ ΚΑΤΑΓΡΑΦΕΣ

Με τις συνεχείς καταγραφές εγγράφονται οι πληροφορίες αναφορικά με μία ή περισσότερες μεταβλητές. Με άλλα λόγια, καταγράφονται οι τιμές μίας ή περισσοτέρων μεταβλητών που ενδιαφέρουν, καθώς και οι μεταβολές αυτών των τιμών κάθε φορά που εμφανίζονται.

Παραδείγματα συνεχών καταγραφών αποτελούν:

7.

ΕΚΤΙΜΗΣΕΙΣ ΜΕ ΒΑΣΗ ΤΟ ΔΕΙΓΜΑ

7.1. ΕΙΣΑΓΩΓΗ

Από τα πλέον βασικά προβλήματα που καλείται να επιλύσει η Στατιστική Επιστήμη, είναι αυτά που αναφέρονται στην *εκτίμηση των χαρακτηριστικών παραμέτρων ενός πληθυσμού οντοτήτων, από πληροφορίες που περιλαμβάνει ένα δείγμα του*. Στην **πράξη**, προκειμένου να γίνει η εκτίμηση μίας παραμέτρου του πληθυσμού, επιλέγεται εκείνη η στατιστική μέθοδος η οποία παρέχει, για τη συγκεκριμένη περίπτωση, τη μεγαλύτερη δυνατή ακρίβεια και αξιοπιστία. Αυτό σημαίνει ότι η εκλογή της μεθόδου γίνεται έτσι ώστε *να ληφθούν, από το δείγμα, οι περισσότερες κατά το δυνατό πληροφορίες με τη μικρότερη αβεβαιότητα*. Η όλη διαδικασία είναι γνωστή ως **Εκτιμητική** και οι χρησιμοποιούμενες μέθοδοι ονομάζονται **Μέθοδοι Εκτίμησης**.

7.2. ΕΚΤΙΜΗΣΗ ΣΕ ΣΗΜΕΙΟ ΚΑΙ ΕΚΤΙΜΗΣΗ ΔΙΑΣΤΗΜΑΤΟΣ

Οι μέθοδοι εκτίμησης διακρίνονται σε δύο είδη:

■ Η **πρώτη** μέθοδος εκτιμά, με βάση το διατιθέμενο δείγμα, μία τιμή $\hat{\theta}$ της άγνωστης παραμέτρου θ του πληθυσμού και ονομάζεται **εκτίμηση σε σημείο** ή **σημειακή εκτίμηση** (point estimation). Με άλλα λόγια, στην περίπτωση αυτή, θεωρείται ως τιμή της παραμέτρου θ στον πληθυσμό η τιμή της αντίστοιχης παραμέτρου $\hat{\theta}$ στο δείγμα, δηλαδή η τιμή του αντίστοιχου στατιστικού $\hat{\theta}$. Έτσι, ο υπολογισμός του στατιστικού δίνει την περίπου και όχι ακριβή τιμή της πληθυσμιακής παραμέτρου. Προφανώς, οι εκτιμήσεις / τα στατιστικά $\hat{\theta}$ διαφέρουν από δείγμα σε δείγμα *ίσου μεγέθους*.

Για παράδειγμα, από ένα δείγμα 18 ανδρών προέκυψε ότι η μέση τιμή του ύψους τους είναι 174,25 cm και η τυπική απόκλιση είναι 4,06 cm. Έτσι, η εκτίμηση της μέσης τιμής του ύψους στον αντίστοιχο πληθυσμό μπορεί να εκφρασθεί ως:

- «Μία εκτίμηση για το μέσο ύψος του πληθυσμού των ανδρών είναι 174,25 cm»
- «Μία εκτίμηση για το μέσο ύψος του πληθυσμού των ανδρών είναι 174,25 cm και το μέγεθος του δείγματος είναι $n = 18$»
- «Μία εκτίμηση για το μέσο ύψος του πληθυσμού των ανδρών είναι 174,25 cm όταν το μέγεθος του δείγματος είναι $n = 18$ και η τυπική απόκλιση είναι 4,06 cm»

Όπως φαίνεται, ενώ η μέθοδος είναι η ίδια (εκτίμηση σε σημείο), το πλήθος των πληροφοριών είναι διαφορετικό για τις τρείς εκφράσεις του μέσου ύψους των ανδρών στον πληθυσμό. Δηλαδή, με την εκτίμηση σε σημείο, η τιμή μιας παραμέτρου είναι «περίπου ίση με την τιμή της στο δείγμα», και αυτό αποτελεί μειονέκτημα της μεθόδου.

▪ Η **δεύτερη** μέθοδος εκτιμά ένα διάστημα μέσα στο οποίο θα βρίσκεται η πραγματική τιμή της παραμέτρου θ του πληθυσμού και ονομάζεται **εκτίμηση σε διάστημα** ή **εκτίμηση διαστήματος**. Αυτό σημαίνει ότι, με την εκτίμηση διαστήματος, προσδιορίζεται ένα διάστημα τιμών $b \leq \theta \leq c$ μέσα στο οποίο θεωρείται ότι βρίσκεται η *άγνωστη παράμετρος* θ με μια *συγκεκριμένη πιθανότητα*. Με άλλα λόγια, για την εκτίμηση ενός διαστήματος χρειάζεται αφ' ενός, η γνώση ενός *αμερόληπτου εκτιμητή* για την παράμετρο θ και, αφ' ετέρου, η γνώση της *συνάρτησης πυκνότητας πιθανότητας αυτής της παραμέτρου, δηλαδή της τυχαίας μεταβλητής* θ. Έτσι, για το προηγούμενο παράδειγμα, η εκτίμηση της μέσης τιμής του ύψους στον πληθυσμό των ανδρών εκφράζεται ως:

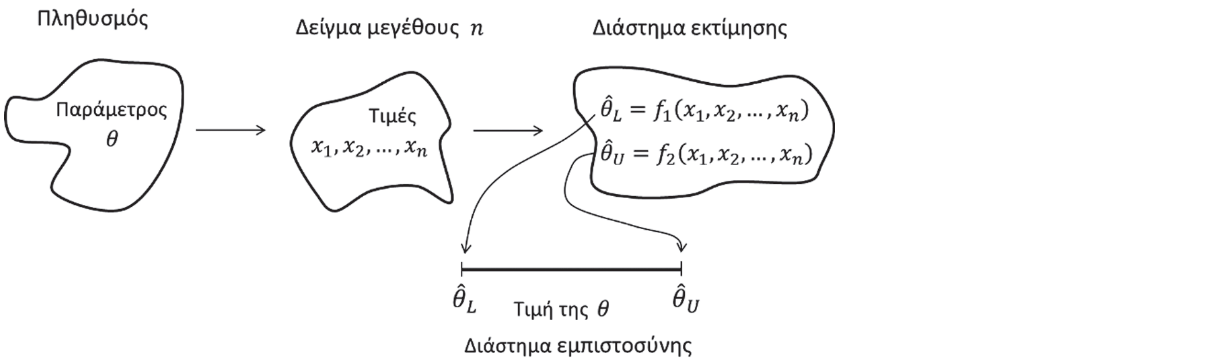
- «Η μέση τιμή μ του ύψους των ανδρών στον πληθυσμό βρίσκεται μεταξύ των τιμών 172, 43 cm και 177,21 cm ή στο διάστημα [172,43 cm ÷ 177,21 cm], με πιθανότητα 95% ».

Όπως φαίνεται, με την εκτίμηση σε διάστημα, υπάρχει «ένας βαθμός βεβαιότητας» ότι η πραγματική τιμή της παραμέτρου που ενδιαφέρει βρίσκεται εντός συγκεκριμένου διαστήματος. Αυτός ο «βαθμός βεβαιότητας», δηλαδή η πιθανότητα, δεν είναι παρά ο λεγόμενος **συντελεστής** ή **επίπεδο εμπιστοσύνης** (confidence level) $\boldsymbol{P = 1 - \alpha}$ που συναρτάται με το προτεινόμενο διάστημα το οποίο, για τον ίδιο λόγο, ονομάζεται **διάστημα εμπιστοσύνης (δ.ε.)** (confidence interval)[1].

Αυτό δείχνει ότι, αντί να ειπωθεί: «υπάρχει βεβαιότητα κατά 95% ότι...», μπορεί να ειπωθεί: «σε 95% διάστημα εμπιστοσύνης (δ.ε.) ....». *Όσο μεγαλύτερος είναι ο συντελεστής εμπιστοσύνης τόσο μεγαλύτερο διάστημα εμπιστοσύνης προκύπτει*, δηλαδή τόσο μεγαλύτερη έκταση έχει το διάστημα τιμών $[b, c]$. και συνεπώς *μειώνεται η ακρίβεια της εκτίμησης*. Έτσι, ένα 99% δ.ε. έχει μεγαλύτερη έκταση από ένα δ.ε. 95%, γεγονός που σημαίνει ότι υπάρχει μεγαλύτερη βεβαιότητα για την εκτίμηση της παραμέτρου μέσα στο πρώτο διάστημα και, συγχρόνως, υπάρχει μικρότερη ακρίβεια στην εκτίμηση αυτή. Επομένως, το ιδανικό θα ήταν ένα μικρό διάστημα με ένα υψηλό συντελεστή εμπιστοσύνης. Με άλλα λόγια, *το διάστημα εμπιστοσύνης είναι ένα διάστημα που κατασκευάζεται από τις πληροφορίες του διατιθέμενου δείγματος και το διάστημα αυτό καλύπτει την πραγματική αλλά άγνωστη τιμή της παραμέτρου με πιθανότητα $P = 1 - \alpha$ την οποία καθορίζει ο ερευνητής*. Στο Σχήμα 7.2.1. απεικονίζεται η διαδικασία για την εκτίμηση του διαστήματος εμπιστοσύνης μιας πληθυσμιακής παραμέτρου θ.

---

[1] Βλ. Κεφάλαιο 4.6.6.



Σχήμα 7.2.1. Διάστημα εμπιστοσύνης για την εκτίμηση μιας πληθυσμιακής παραμέτρου θ.
(Πηγή: J.E. Barber & G.M. Barber, 1996, p. 259, ιδία επεξεργασία)

Συνήθως, στα περισσότερα προβλήματα εκτίμησης, προκαθορίζονται είτε 95% δ.ε. είτε 99 % δ.ε., δηλαδή, προκαθορίζονται μεγάλοι συντελεστές εμπιστοσύνης ή μεγάλες πιθανότητες 95% ή 99%, ανάλογα με το πρόβλημα. Γενικά, εάν $P = 1 - \alpha$ είναι μια υψηλή πιθανότητα, τότε μπορεί να βρεθεί ένα $100 \cdot P = 100 \cdot (1 - \alpha)\%$ δ.ε. για την παράμετρο που εκτιμάται από το δείγμα. Όπως είναι γνωστό, η τιμή $\boldsymbol{\alpha = 1 - P}$ ονομάζεται **στάθμη** ή **επίπεδο σημαντικότητας (σ.σ.)** (level of significance) του διαστήματος εμπιστοσύνης και εκφράζει την πιθανότητα η παράμετρος που έχει εκτιμηθεί να βρίσκεται *εκτός του διαστήματος εμπιστοσύνης*. Με άλλα λόγια, *η πιθανότητα α εκφράζει το σφάλμα που μπορεί να γίνει κατά την εκτίμηση του διαστήματος εμπιστοσύνης*, αφού η τιμή της προκαθορίζει την πιθανότητα ώστε η τιμή της παραμέτρου να βρίσκεται μέσα σε ένα διάστημα εμπιστοσύνης. Γι' αυτό και ονομάζεται στάθμη ή επίπεδο σημαντικότητας. Δηλαδή, *η στάθμη σημαντικότητας (σ.σ.) α προσδιορίζει (και) τη σημαντικότητα του σφάλματος εκτίμησης του δ.ε.*

Εάν π.χ. είναι $\alpha = 0{,}05$ τότε: $1 - \alpha = P = 0{,}95$ και $100 \cdot (1 - \alpha)\% = 100 \times (1 - 0{,}05)\% = 95\%$ δ.ε. για μία παράμετρο του πληθυσμού. Δηλαδή, *η τιμή της παραμέτρου που εκτιμήθηκε βρίσκεται στο διάστημα εμπιστοσύνης με πιθανότητα 95% και μόνο με πιθανότητα 5% η τιμή της παραμέτρου δεν βρίσκεται μέσα σ' αυτό.*

Στην **πράξη**, προσδιορίζεται από πριν η στάθμη σημαντικότητας α για το διάστημα εμπιστοσύνης της παραμέτρου και στη συνέχεια καθορίζεται το $100 \cdot (1 - \alpha)\%$ δ.ε.

☞ Επισήμανση 7.2.1.

Προφανώς, διαφορετικά μεταξύ τους δείγματα, ίσου μεγέθους n από τον ίδιο πληθυσμό, αποδίδουν διαφορετικές τιμές της δειγματικής παραμέτρου $\hat{\theta}$ και, επομένως, δημιουργούνται διαφορετικά διαστήματα εμπιστοσύνης για την παράμετρο θ. Η γνώση της κατανομής της δειγματικής $\hat{\theta}$ καθιστά δυνατό τον προσδιορισμό των ορίων b και c αυτών των διαστημάτων για όλα τα πιθανά δείγματα, έτσι ώστε οποιοδήποτε *καθορισμένο* ποσοστό αυτών των διαστημάτων να περιλαμβάνει την παράμετρο θ. Επομένως, εάν οι τιμές b και c υπολογίζονται με τρόπο ώστε π.χ. το 95% όλων των πιθανών διαστημάτων, σε επαναλαμβανόμενα δείγματα, να περιέχουν την παράμετρο θ, τότε *το ποσοστό αυτό είναι ίσο με την πιθανότητα* να επιλεγεί ένα από αυτά τα διαστήματα, δηλαδή *να επιλεγεί ένα δείγμα από αυτά* στα οποία το διάστημα εμπιστοσύνης που δημιουργείται περιλαμβάνει την παράμετρο θ. Έτσι, αυτό το διάστημα που υπολογίσθηκε από το τυχαία επιλεγμένο δείγμα, χαρακτηρίζεται ως 95% *διάστημα εμπιστοσύνης*. Με άλλα λόγια, υπάρχει 95% βεβαιότητα ότι το διάστημα που έχει υπολογισθεί περιλαμβάνει την πληθυσμιακή παράμετρο θ.

Συνεπώς, *συντελεστής ή επίπεδο εμπιστοσύνης $P = 1 - \alpha$ σημαίνει ότι, εάν ληφθούν όλα τα πιθανά τυχαία δείγματα ίδιου μεγέθους n από ένα πληθυσμό, τότε στο δειγματικό στατιστικό $\hat{\theta}$ αντιστοιχεί ένα διαφορετικό διάστημα εμπιστοσύνης, όμως σε οποιοδήποτε καθορισμένο* *ποσοστό* $100 \cdot P = 100 \cdot (1-\alpha)$ *αυτών των διαστημάτων θα βρίσκεται η άγνωστη πληθυσμιακή παράμετρος* θ.

7.3. ΕΚΤΙΜΗΤΕΣ

7.3.1. ΕΙΣΑΓΩΓΗ

Ο προσδιορισμός της τιμής κάποιας παραμέτρου θ στον πληθυσμό της τυχαίας μεταβλητής X, δεν είναι πάντα εφικτός στην πράξη εξ αιτίας τεχνικών και οικονομικών λόγων. Προς τούτο, ακολουθείται η *διαδικασία της δειγματοληψίας*. Αυτό σημαίνει ότι ο προσδιορισμός της τιμής της παραμέτρου που ενδιαφέρει, θα γίνει με βάση την τιμή που παίρνει η παράμετρος αυτή μόνον από τα στοιχεία του διατιθέμενου δείγματος, και από την τιμή αυτή θα εξαχθούν συμπεράσματα, «*θα γίνει εκτίμηση*», όπως λέγεται, για ολόκληρο τον πληθυσμό. Η εκτίμηση αυτή γίνεται με τη βοήθεια συναρτήσεων που λέγονται «**Εκτιμήτριες Συναρτήσεις**» ή, απλούστερα, «**Εκτιμητές**» (estimators). Πρόκειται για *μαθηματικές σχέσεις οι οποίες προσδιορίζουν τη διαδικασία για τον υπολογισμό του στατιστικού* $\widehat{\theta}$.

Θεωρείται ο πληθυσμός μιας τ.μ. X με *πεπερασμένο* πλήθος οντοτήτων N. Σχηματίζονται *όλα τα δυνατά δείγματα μεγέθους* n, με $n < N$, τα οποία συμβολίζονται ως $\delta_{j,n}$, όπου j είναι η αρίθμηση του δείγματος και n είναι το μέγεθός του, δηλαδή το σύνολο των οντοτήτων που περιλαμβάνει. Το πλήθος m όλων αυτών των δειγμάτων $\delta_{j,n}$ υπολογίζεται με δύο διαφορετικούς τρόπους οι οποίοι συναρτώνται με το είδος της δειγματοληψίας. Εάν η *δειγματοληψία είναι χωρίς επανατοποθέτηση*[2], το πλήθος m προκύπτει ως:

$$m = \binom{N}{n} \tag{7.3.1a}$$

Εάν η *δειγματοληψία είναι με επανατοποθέτηση*[3], το πλήθος m υπολογίζεται από τη σχέση:

$$m = N^n \tag{7.3.1b}$$

Τα δείγματα αυτά αποτελούν το σύνολο $\Delta_{m,n} = \{\delta_{1,n}, \delta_{2,n}, \dots, \delta_{j,n}, \dots, \delta_{m,n}\}$.

Έστω η τ.μ. X η οποία, για το πλήθος των οντοτήτων N του πληθυσμού της, παίρνει τις τιμές: x_1, x_2, ..., x_N. Οι τιμές της τ.μ. X που αντιστοιχούν στο δείγμα $\delta_{j,n}$ αυτού του πληθυσμού, είναι:

$$x_{j1}, x_{j2}, \dots, x_{jn}, \quad \text{με } j = 1, 2, \dots, \binom{N}{n} \quad \text{ή} \quad j = 1, 2, \dots, N^n \tag{7.3.2}$$

Εάν η συνάρτηση $g(X)$ της τυχαίας μεταβλητής X, η οποία αναφέρεται στον πληθυσμό, *συμβολίζεται ως* $E_N[g(X)]$, τότε η *αντίστοιχη συνάρτηση* της ίδιας τ.μ. X η οποία αναφέρεται στο δείγμα, ονομάζεται (όπως έχει αναφερθεί) «**Εκτιμήτρια Συνάρτηση**» ή «**Εκτιμητής**» της συνάρτησης $E_N(g(X))$ και συμβολίζεται ως $\boldsymbol{E_n(g(X))}$[4]: Προφανώς, $E_N(g(X)) \equiv g(X) \longrightarrow$ «Εκτίμηση» στον Πληθυσμό ≡ Πραγματική τιμή στον Πληθυσμό.

---

[2] Ν. Φαρμάκης, 1992, σελ. 12
[3] Κ.Β. Μπαγιάτης, 1990, σελ. 63, 66
[4] Ν. Φαρμάκης, 1992, σελ. 13

8.

ΔΟΚΙΜΑΣΙΕΣ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ Ι : ΠΑΡΑΜΕΤΡΙΚΟΙ ΣΤΑΤΙΣΤΙΚΟΙ ΕΛΕΓΧΟΙ

8.1. ΓΕΝΙΚΑ

Με τον όρο *στατιστικές δοκιμασίες* ή *στατιστικοί έλεγχοι* ή *δοκιμασίες υποθέσεων* ή *δοκιμασίες / έλεγχοι σημαντικότητας*, αναφέρονται εκείνες οι στατιστικές διαδικασίες οι οποίες ως στόχο έχουν τον έλεγχο των τιμών των παραμέτρων μίας τυχαίας μεταβλητής, οι οποίες εκτιμήθηκαν από ένα *τυχαίο* επιλεγμένο δείγμα του αντίστοιχου πληθυσμού. Όπως έχει αναφερθεί, από ένα πληθυσμό οντοτήτων σχηματίζεται ένα *τυχαίο* δείγμα και θεωρείται κάποιο χαρακτηριστικό του πληθυσμού, δηλαδή κάποια μεταβλητή αυτού, η οποία συνήθως *ακολουθεί κατανομή γνωστής μορφής*. Από την ανάλυση του δείγματος υπολογίζονται τα λεγόμενα *στατιστικά* (δηλαδή, οι τιμές των παραμέτρων της τυχαίας μεταβλητής στο δείγμα), από τα οποία προκύπτουν, για τις αντίστοιχες παραμέτρους του πληθυσμού, είτε οι σημειακές εκτιμήσεις τους, είτε τα διαστήματα εμπιστοσύνης τους για συγκεκριμένους συντελεστές εμπιστοσύνης[1]. Πολλές φορές, όμως, είναι δυνατό, μετά από τον υπολογισμό ενός στατιστικού να γίνεται κάποια «*υπόθεση*» *σχετικά με την τιμή που εμφανίζει η αντίστοιχη παράμετρος στον πληθυσμό.* Αυτό συνεπάγεται ότι γίνεται έλεγχος εάν η υποτιθέμενη τιμή συμφωνεί με το στατιστικό ή εάν η εμφανιζόμενη διαφορά μεταξύ στατιστικού και παραμέτρου οφείλεται στην τύχη που επέδρασε στον σχηματισμό του δείγματος, γεγονός που σημαίνει ότι *η τιμή της πληθυσμιακής παραμέτρου βρίσκεται μέσα στα όρια που προσδιορίζονται από τα αντίστοιχα διαστήματα εμπιστοσύνης*.

[1] Βλ. Κεφάλαιο 7

Σε αυτές τις περιπτώσεις, γίνεται μία **βασική υπόθεση** η οποία ονομάζεται **μηδενική υπόθεση** ή **υπόθεση μηδέν**, H_0, και γίνεται έλεγχος αυτής. Η μέθοδος που ακολουθείται για τον έλεγχο αυτό, λέγεται **δοκιμασία** [**έλεγχος** (test)] **υποθέσεων** ή **δοκιμασία σημαντικότητας**. Σύμφωνα με τη μέθοδο αυτή, υπάρχει πάντα μία συγκεκριμένη τιμή για την παράμετρο ενδιαφέροντος, και το ερώτημα που τίθεται είναι «*κατά πόσο συμφωνεί η τιμή της παραμέτρου που εκτιμήθηκε από το δείγμα με τη συγκεκριμένη τιμή*».

Προκειμένου να εφαρμοσθεί μία δοκιμασία ή ένα test σημαντικότητας, ακολουθούνται τα εξής βήματα:

1. Ορίζεται η **μηδενική υπόθεση** ή **υπόθεση μηδέν**, $\boldsymbol{H_0}$,

2. Ορίζεται η **εναλλακτική υπόθεση**, $\boldsymbol{H_1}$,

3. Ορίζεται η **δοκιμασία** ή το **tes**t που θα εφαρμοσθεί,

4. Ορίζεται η **απορριπτική περιοχή** $\boldsymbol{R}$ της υπόθεσης H_0,

5. Εξάγονται τα **συμπεράσματα**.

Οι δοκιμασίες σημαντικότητας διακρίνονται σε *τρείς βασικές κατηγορίες*:

- Της σύγκρισης με μία γνωστή εκ των προτέρων τιμή ή *δοκιμασίες συμφωνίας*
- Της σύγκρισης δύο δειγμάτων ή *δοκιμασίες ομογένειας*
- Της ανεξαρτησίας δύο μεταβλητών

■ **Η δοκιμασία της σύγκρισης με μία εκ των προτέρων γνωστή τιμή** ή **δοκιμασία συμφωνίας**, συνίσταται στο να διαπιστωθεί εάν η τιμή του στατιστικού μίας τυχαίας μεταβλητής, δηλαδή η τιμή της δειγματικής παραμέτρου, διαφέρει ή όχι από την τιμή της παραμέτρου στον πληθυσμό, η οποία (τιμή) είναι ήδη «*γνωστή*». Αυτό το τελευταίο σημαίνει ότι είτε η πληθυσμιακή παράμετρος είναι γνωστή εκ των προτέρων, είτε καθορίζεται αυτή με βάση μία υπόθεση η οποία ζητείται να επαληθευθεί.

■ **Η δοκιμασία της σύγκρισης δύο δειγμάτων** ή η **δοκιμασία ομογένειας**, συνίσταται στο να διαπιστωθεί εάν δύο τυχαία *ανεξάρτητα* δείγματα, δηλαδή δύο δείγματα όπου ο σχηματισμός του καθενός δεν εξαρτάται από τον σχηματισμό του άλλου, είναι ή όχι *ομογενή ως προς ορισμένο χαρακτηριστικό των πληθυσμών από τους οποίους επιλέγονται* ή είναι *ομογενή ως προερχόμενα από τον ίδιο πληθυσμό.* Στις δοκιμασίες αυτές συγκρίνονται τα στατιστικά των δύο δειγμάτων. Για παράδειγμα, γίνεται μία έρευνα σχετικά με ένα κοινωνικό θέμα. Το ένα δείγμα περιλαμβάνει άνδρες και το άλλο εργαζόμενες γυναίκες. Ζητείται να διαπιστωθεί εάν είναι ή όχι συγκρίσιμα τα αποτελέσματα της έρευνας (*1η περίπτωση ομογένειας*). Τα δύο δείγματα λαμβάνονται από τον ίδιο πληθυσμό, π.χ. τις εργαζόμενες γυναίκες, και ζητείται και πάλι να συγκριθούν τα δύο αποτελέσματα (*2η περίπτωση ομογένειας*).

■ **Η δοκιμασία της ανεξαρτησίας δύο τυχαίων μεταβλητώ**ν (οι μεταβλητές μπορεί να είναι είτε ποσοτικές είτε ποιοτικές) συνίσταται στο να διαπιστωθεί *εάν υπάρχει ή όχι στοχαστική εξάρτηση* μεταξύ δύο τυχαίων μεταβλητών του δείγματος που έχει επιλεγεί. Οι σχετικές μέθοδοι αναπτύσσονται στο Κεφάλαιο 10.

● Οι στατιστικές δοκιμασίες συμφωνίας και ομογένειας διακρίνονται σε *παραμετρικές* και *μη παραμετρικές*.

- **Παραμετρικές** (parametric tests), ονομάζονται οι δοκιμασίες οι οποίες εφαρμόζονται όταν η κατανομή της τυχαίας μεταβλητής στον πληθυσμό «υπακούει» σε ορισμένους νόμους, δηλαδή είναι γνωστής μορφής. Συνήθως, η τ.μ. ακολουθεί την κανονική ή τη διωνυμική κατανομή. Βεβαίως, ορισμένες από τις δοκιμασίες αυτές μπορούν να εφαρμοσθούν και σε πληθυσμιακές κατανομές που δεν είναι γνωστής μορφής, αρκεί σε αυτή την περίπτωση τα αντίστοιχα δείγματα να είναι μεγάλα.

- **Μη παραμετρικές** (non-parametric tests), λέγονται οι δοκιμασίες για την εφαρμογή των οποίων δεν γίνεται καμιά υπόθεση αναφορικά με την κατανομή της τ.μ. στον πληθυσμό και το μέγεθος του δείγματος είναι μικρό.

Για την εφαρμογή μίας δοκιμασίας σημαντικότητας δίνονται οι εξής ορισμοί:

■ **Ορισμος 8.1.1.**

Αρχική ή **μηδενική υπόθεση** ή **υπόθεση μηδέν**, $\boldsymbol{H_0}$ (null hypothesis), ονομάζεται η υπόθεση που γίνεται αρχικά. Όταν η δοκιμασία είναι *δοκιμασία συμφωνίας*, η H_0 αφορά μία συγκεκριμένη τιμή της παραμέτρου. Όταν η δοκιμασία είναι *δοκιμασία ομογένειας*, η H_0 αφορά τη διαφορά των τιμών της παραμέτρου στους δύο πληθυσμούς ή στον ίδιο πληθυσμό. Γενικά, η H_0 *διατυπώνεται έτσι ώστε να αμφισβητηθεί*. Με άλλα λόγια, το συμπλήρωμα ή το αντίθετο της υπόθεσης για την οποία εφαρμόζεται η στατιστική δοκιμασία αποτελεί τη μηδενική υπόθεση, ή η μηδενική υπόθεση είναι το πλαίσιο αναφοράς ως προς το οποίο θα κριθούν τα δειγματικά αποτελέσματα. Για παράδειγμα, εάν το ζητούμενο είναι να αποφασισθεί ότι η μέθοδος διδασκαλίας Α είναι καλύτερη από τη μέθοδο Β, υποτίθεται ότι δεν υπάρχει διαφορά μεταξύ των δύο μεθόδων. Αυτή η υπόθεση είναι η μηδενική υπόθεση.

■ **Ορισμος 8.1.2.**

Εναλλακτική υπόθεση, $\boldsymbol{H_1}$ (alternative hypothesis), ονομάζεται η δεύτερη υπόθεση, συμπληρωματική ή αντίθετη της πρώτης, ως προς την οποία ελέγχεται η H_0.

Πίνακας 8.1.1. Μορφή Στατιστικής Δοκιμασίας

I.	$H_0: \mu = 15$ ή $\mu_0 = 15$ (Μονόπλευρος έλεγχος προς τα δεξιά ή Δεξιόπλευρος $H_1: \mu > 15$ ή $\mu_1 > 15$ έλεγχος – δεξιά «ουρά»)
	Είναι γνωστή η μέση τιμή του πληθυσμού, $\mu_0 = 15$, (αρχική υπόθεση H_0) και ελέγχεται η εναλλακτική υπόθεση H_1, βάσει των δεδομένων του δείγματος, μήπως η μέση τιμή του πληθυσμού είναι μεγαλύτερη του 15, $\mu_1 > 15$.
II.	$H_0: \mu = 15$ ή $\mu_0 = 15$ (Μονόπλευρος έλεγχος προς τα αριστερά ή Αριστερόπλευρος $H_1: \mu < 15$ ή $\mu_1 < 15$ έλεγχος – αριστερή «ουρά»)
	Ελέγχεται η εναλλακτική υπόθεση H_1, μήπως η μέση τιμή του πληθυσμού είναι μικρότερη του 15, $\mu_1 < 15$.
III.	$H_0: \mu = 15$ ή $\mu_0 = 15$ (Δίπλευρος ή Αμφίπλευρος έλεγχος - δεξιά ή αριστερή $H_1: \mu \neq 15$ ή $\mu_1 \neq 15$ «ουρά»)
	Ελέγχεται η εναλλακτική υπόθεση H_1, μήπως η μέση τιμή του πληθυσμού είναι διαφορετική του 15, $\mu_1 \neq 15$. Δηλαδή, στην περίπτωση αυτή δεν ενδιαφέρει εάν η τιμή είναι μεγαλύτερη ή μικρότερη, αλλά ότι προκύπτει διαφορετική από την αρχική γνωστή τιμή. *Αυτό σημαίνει ότι οι έλεγχοι για $\mu_1 > 15$ και $\mu_1 < 15$ γίνονται ταυτοχρόνως*.

Πάντοτε γίνεται δεκτή η μία υπόθεση και απορρίπτεται η άλλη. Η αποδοχή ή η απόρριψη της H_0 γίνεται με τη βοήθεια μίας στατιστικής δοκιμασίας. Σε όλες τις στατιστικές δοκιμασίες η H_0 ελέγχεται ως προς την εναλλακτική H_1, *με υποθέσεις του τύπου*: «μεγαλύτερη από» ή «μικρότερη από» ή «διάφορη από». Έτσι, εάν είναι θ η παράμετρος του πληθυσμού που θα ελεγχθεί και θ_0 είναι μία συγκεκριμένη τιμή αυτής, τότε η μηδενική υπόθεση διατυπώνεται ως $H_0 : \theta = \theta_0$ και η εναλλακτική υπόθεση διατυπώνεται ως $H_1 : \theta > \theta_0$ ή $\theta < \theta_0$ ή $\theta \neq \theta_0$. Για παράδειγμα, εάν η παράμετρος θ είναι η μέση τιμή μ της κατανομής μίας τυχαίας μεταβλητής X με $\mu = \mu_0 = 15$, η στατιστική δοκιμασία μπορεί να έχει τη μορφή του Πίνακα 8.1.1.

Όταν ο έλεγχος αφορά τη μεγαλύτερη ή μικρότερη τιμή της παραμέτρου σε σχέση με την αρχική H_0 (περιπτώσεις I, II), τότε η **δοκιμασία** ή το **test** λέγονται **μονόπλευρη / ο** ή **μονής κατεύθυνσης** (one-sided ή one-tailed). Όταν ο έλεγχος αφορά, απλά, τη διαφορετικότητα των δύο υποθέσεων, της αρχικής H_0 και της εναλλακτικής H_1 (περίπτωση III), τότε η **δοκιμασία** ή το **test** λέγονται **δίπλευρη / ο** ή **αμφίπλευρη / ο** ή **διπλής κατεύθυνσης** (two-sided ή two-tailed).

■ **Ορισμος 8.1.3.**

Λάθος ή **σφάλμα τύπου I** (type 1 error), λέγεται το λάθος ή το σφάλμα *να απορριφθεί η υπόθεση H_0* ενώ είναι *σωστή* ή *αληθής*. Η πιθανότητα να συμβεί αυτό το λάθος συμβολίζεται με το ελληνικό γράμμα $\boldsymbol{\alpha}$, και δεν είναι άλλη από τη γνωστή **στάθμη** ή **επίπεδο σημαντικότητας** (**σ.σ.**) (significance level). Πρόκειται για την ανώτατη τιμή της πιθανότητας α ώστε να συμβεί λάθος τύπου I.

Όσο μικρότερο είναι ένα λάθος τύπου I, τόσο περισσότερο συνεπής είναι η δοκιμασία που εφαρμόζεται. *Συνήθως, η πιθανότητα για λάθος τύπου I, δηλαδή η σ.σ. α, έχει τις τιμές 0,01* ή *0,05* ή *0,10*, και επιλέγεται από τον ερευνητή. Επισημαίνεται ότι όσο ελαττώνεται η στάθμη σημαντικότητας (σ.σ.) α, τόσο διευρύνεται η περιοχή όπου γίνεται αποδεκτή η H_0, οπότε χρειάζονται περισσότερες αποδείξεις για την απόρριψη της. Έτσι, για παράδειγμα, δειγματικά δεδομένα που είναι επαρκή για την απόρριψη της H_0 σε σ.σ. $\alpha = 0{,}05$ είναι δυνατό να είναι ανεπαρκή για την απόρριψη της H_0 σε σ.σ. $\alpha = 0{,}01$. Ενδείκνυται, λοιπόν, σε μία τέτοια περίπτωση, όταν βρίσκεται ένα στατιστικώς σημαντικό[2] αποτέλεσμα στην ανώτατη επιτρεπτή στάθμη σημαντικότητας, π.χ. $\alpha = 0{,}05$, να γίνεται έλεγχος και για τις κατώτερες στάθμες σημαντικότητας που διατίθενται, π.χ. $\alpha = 0{,}01$ ή 0,001, και να επιλέγεται εκείνη που εξακολουθεί να παρέχει στατιστικώς σημαντικά αποτελέσματα, αλλά με τη μικρότερη πιθανότητα σφάλματος.

■ **Ορισμος 8.1.4.**

Λάθος ή **σφάλμα τύπου II** (type II error), λέγεται το λάθος ή το σφάλμα να γίνει *αποδεκτή η υπόθεση H_0* ενώ στην πραγματικότητα *είναι λανθασμένη*. Η πιθανότητα να συμβεί αυτό το λάθος συμβολίζεται με το ελληνικό γράμμα $\boldsymbol{\beta}$.

Το λάθος τύπου II εξαρτάται από τις άγνωστες τιμές της παραμέτρου και συνεπώς είναι δύσκολο να υπολογισθεί. Η υπόθεση H_0 μπορεί να απορριφθεί χωρίς προβλήματα, όταν όμως γίνει αποδεκτή θα πρέπει να υπολογισθεί και η τιμή της πιθανότητας β, για να έχει μεγαλύτερη βεβαιότητα η απόφαση που θα ληφθεί. Γι' αυτόν το λόγο, η *σωστή διατύπωση για την αποδοχή της H_0* σε σ.σ. α είναι: «τα δεδομένα του δείγματος δεν επιτρέπουν την απόρριψη της H_0 σε σ.σ. α» ή «η H_0 μπορεί να είναι αληθής σε σ.σ. α», οπότε δεν είναι υποχρεωτικός ο υπολογισμός της πιθανότητας β. Με άλλα λόγια, θα πρέπει να καταστεί σαφές ότι η αποδοχή της H_0 είναι το αποτέλεσμα μίας ανεπαρκούς απόδειξης για την απόρριψη αυτής, οπότε το γεγονός αυτό δεν συνεπάγεται ότι η H_0 είναι αποδεκτή, δηλαδή είναι αληθής.

[2] Βλ. Ορισμούς 8.1.7., 8.1.9.

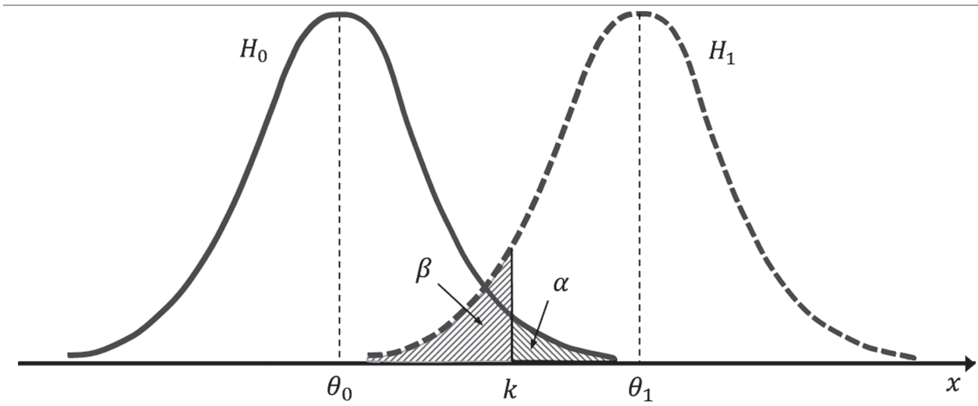
■ **ΟΡΙΣΜΟΣ 8.1.5.**

Ισχύς (power) μίας στατιστικής δοκιμασίας ονομάζεται η πιθανότητα *να απορριφθεί η* H_0 *όταν πραγματικά είναι λανθασμένη*. Η πιθανότητα αυτή συμβολίζεται με το ελληνικό γράμμα $\boldsymbol{\gamma}$ και ισχύει $\boldsymbol{\gamma = 1 - \beta}$.

Πίνακας 8.1.2. Αποφάσεις για την αποδοχή ή απόρριψη της H_0

Υπόθεση H_0	Απόφαση σχετικά με την Υπόθεση H_0	
	Αποδοχή H_0	Απόρριψη H_0
H_0 Σωστή (Αληθής)	Ναι με πιθανότητα $1-\alpha$	Ναι → Λάθος ή σφάλμα τύπου Ι με πιθανότητα α
H_0 Λανθασμένη (Ψευδής)	Ναι → Λάθος ή σφάλμα τύπου ΙΙ με πιθανότητα β	Ναι με πιθανότητα $\gamma = 1-\beta$

Όπως φαίνεται από το Σχήμα 8.1.1., όσο αυξάνεται η πιθανότητα $P = 1-\alpha$ [δηλαδή, όσο δεξιότερα μετακινείται η τιμή k της κατανομής H_0 και συνεπώς ελαττώνεται η στάθμη σημαντικότητας α, $P(\theta \geq k \mid H_0) = \alpha$], τόσο περισσότερο αυξάνεται η πιθανότητα β της κατανομής H_1, $P(\theta \leq k \mid H_1) = \beta$, δηλαδή τόσο περισσότερο αυξάνεται η πιθανότητα να γίνει αποδεκτή ως σωστή η αρχική υπόθεση H_0, γεγονός που σημαίνει ότι υπάρχει λάθος τύπου ΙΙ. Κατά συνέπεια, στην προσπάθεια να αποφευχθεί ένα λάθος τύπου Ι, αυξάνεται αντιστοίχως η πιθανότητα να γίνει λάθος τύπου ΙΙ.



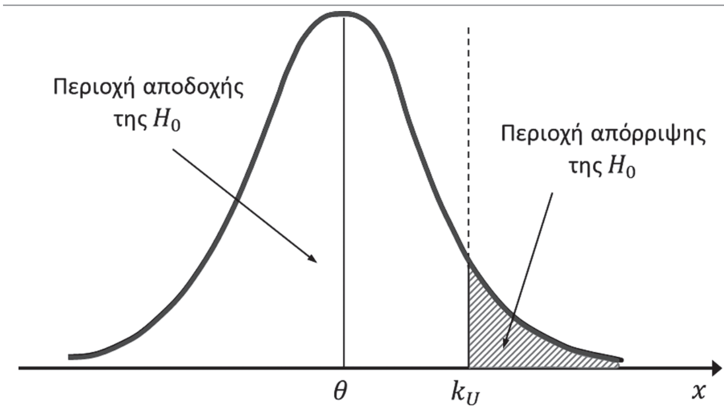
Σχήμα 8.1.1. Η σχέση μεταξύ των λαθών ή σφαλμάτων τύπου Ι (πιθανότητα α) και τύπου ΙΙ (πιθανότητα β)

Όπως γίνεται αντιληπτό, για *συγκεκριμένο συντελεστή εμπιστοσύνης* $1-\alpha$, *η πιθανότητα* β *ελαττώνεται όσο απομακρύνεται η τιμή της παραμέτρου* θ_1 (εναλλακτική υπόθεση H_1) *από την τιμή* θ_0 (αρχική ή μηδενική υπόθεση H_0). Δηλαδή, η τιμή της β εξαρτάται από το πόσο κοντά ή μακρυά βρίσκεται η τιμή της παραμέτρου θ_0 από την πραγματική της τιμή.

Φαίνεται, λοιπόν, ότι η πιθανότητα β εξαρτάται και από τις δύο υποθέσεις (μηδενική και εναλλακτική), ενώ η πιθανότητα α εξαρτάται μόνο από την H_0. Έτσι, *προκαθορίζεται η πιθανότητα* α και η προσπάθεια αφορά την ελαχιστοποίηση της πιθανότητας β. Βεβαίως, *οι τιμές των* α *και* β *μπορούν να ελαττωθούν όταν αυξηθεί το μέγεθος του δείγματος*, με δεδομένο ότι το ζητούμενο, σε μία δοκιμασία σημαντικότητας, είναι τόσο το λάθος τύπου Ι όσο και το λάθος τύπου ΙΙ να έχουν συγχρόνως μικρές πιθανότητες εμφάνισης, δηλαδή οι τιμές α και β να είναι συγχρόνως μικρές.

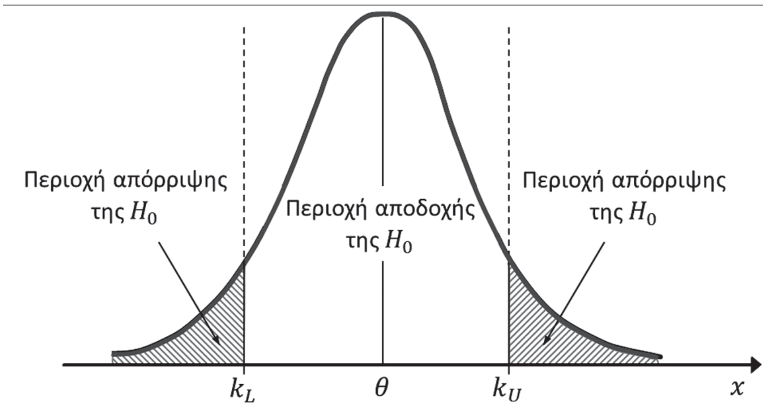
■ **ΟΡΙΣΜΟΣ 8.1.6.**

Απορριπτική περιοχή ή **περιοχή απόρριψης** (region of rejection) ή **κρίσιμη περιοχή** (critical region) ή **περιοχή σημαντικότητας** (region of significance) **της H_0**, σε μία *μονόπλευρη δοκιμασία*, ονομάζεται η περιοχή τιμών της παραμέτρου στην οποία απορρίπτεται η μηδενική υπόθεση H_0 και γίνεται αποδεκτή η εναλλακτική H_1. *Πρόκειται για την περιοχή τιμών οι οποίες είναι μικρότερες ή μεγαλύτερες από την τιμή k (k_L ή k_U) της παραμέτρου που αντιστοιχεί στην πιθανότητα α*. Η τιμή αυτή είναι γνωστή ως *εκατοστιαίο σημείο* ή *κρίσιμη τιμή* ή *τιμή κρίσης* ή *κριτική τιμή* ή *οριακή τιμή*, και η θέση της στον άξονα τιμών της τ.μ. βρίσκεται είτε στο αριστερό (τιμή k_L) είτε στο δεξιό τμήμα (τιμή k_U) της κατανομής. Έτσι, η σχετική διατύπωση για τη δοκιμασία είναι είτε *μονόπλευρος έλεγχος προς τα αριστερά*, $H_1: \theta < \theta_0$, *(αριστερόπλευρος έλεγχος ή αριστερή «ουρά» της κατανομής)*, είτε *μονόπλευρος έλεγχος προς τα δεξιά*, $H_1: \theta > \theta_0$, *(δεξιόπλευρος έλεγχος ή δεξιά «ουρά» της κατανομής)*.



Σχήμα 8.1.2. Περιοχή αποδοχής και περιοχή απόρριψης της H_0 για μονόπλευρο έλεγχο.

Όπως φαίνεται, Σχήμα 8.1.2., με τη μονόπλευρη δοκιμασία, η περιοχή τιμών ή το πεδίο ορισμού της παραμέτρου διαιρείται σε δύο διαστήματα, στην περιοχή απόρριψης και στην **περιοχή αποδοχής** (region of acceptance) ή **περιοχή μη σημαντικότητας** (region of nonsignificance), η οποία περιλαμβάνει όλες τις υπόλοιπες τιμές για τις οποίες γίνεται αποδεκτή η H_0.



Σχήμα 8.1.3. Περιοχή αποδοχής και περιοχές απόρριψης της H_0 για δίπλευρο έλεγχο.

Όταν η δοκιμασία είναι *δίπλευρη* ή *αμφίπλευρη*, τότε οι **απορριπτικές περιοχές** ή **περιοχές απόρριψης** ή **κρίσιμες περιοχές** ή **περιοχές σημαντικότητας της H_0** είναι δύο και το διάστημα τιμών της παραμέτρου διαιρείται σε τρία τμήματα, όπως φαίνεται στο Σχήμα 8.1.3. Το τμήμα που βρίσκεται μεταξύ των κρίσιμων τιμών (ο έλεγχος γίνεται συγχρόνως για μεγαλύτερη και μικρότερη τιμή) αποτελεί την *περιοχή αποδοχής* και τα άλλα δύο τμήματα, προς τα αριστερά και τα δεξιά των κρίσιμων τιμών k_L και k_U, αντιστοίχως, αποτελούν τις *περιοχές απόρριψης (αριστερή και δεξιά «ουρά της κατανομής»*.

Επισημαίνεται ότι *ανάλογα με την παράμετρο η οποία ελέγχεται, επιλέγεται και η αντίστοιχη θεωρητική κατανομή* οπότε είναι γνωστή και η μορφή της καμπύλης της. Οι κατανομές αυτές είναι π.χ. η *τυπική κανονική κατανομή*, η *κατανομή* t_ν, η *κατανομή* X_ν^2, η *κατανομή* F_{ν_1,ν_2} κ.ά. Έτσι, στην περίπτωση της δίπλευρης δοκιμασίας, για τις δύο πρώτες κατανομές οι περιοχές απόρριψης αντιστοιχούν σε δύο συμμετρικά ακραία τμήματα της κατανομής, ενώ για τις επόμενες δύο οι περιοχές απόρριψης είναι ασύμμετρες μεταξύ τους.

■ **ΟΡΙΣΜΟΣ 8.1.7.**

Σημαντικός/ή (Significant): πρόκειται για όρο των στατιστικών δοκιμασιών ο οποίος δηλώνει ότι το αποτέλεσμα είναι διαφορετικό από αυτό που αναμένεται.

■ **ΟΡΙΣΜΟΣ 8.1.8.**

Τιμή Πιθανότητας $\boldsymbol{P}$ ($P -$ Value) σε μία δοκιμασία σημαντικότητας, ονομάζεται η πιθανότητα σύμφωνα με την οποία η δειγματική τιμή είναι τόσο μεγάλη ή και μεγαλύτερη από την πραγματική τιμή της παραμέτρου, όταν η H_0 είναι αληθής. Όσο μικρότερη είναι η τιμή P τόσο ισχυρότερο είναι το αποτέλεσμα, έναντι της H_0, που παράγεται από τα δειγματικά δεδομένα.

Ή

Τιμή Πιθανότητας $\boldsymbol{P}$ ($P -$ Value) είναι η ελάχιστη τιμή της στάθμης σημαντικότητας (σ. σ.) α στην οποία η παρατηρούμενη δειγματική τιμή είναι *σημαντική*, ώστε να μπορεί να απορριφθεί η H_0.

■ **ΟΡΙΣΜΟΣ 8.1.9.**

Στατιστική Σημαντικότητα (Statistical Significance) μίας στατιστικής δοκιμασίας:

Εάν για την πιθανότητα P ισχύει $P < \alpha$, τότε λέγεται ότι τα δεδομένα είναι **στατιστικώς σημαντικά σε σ.σ. $\boldsymbol{\alpha}$**. Με άλλα λόγια, *η στατιστική σημαντικότητα μίας δοκιμασίας συναρτάται με την απόρριψη της* H_0.

■ **ΟΡΙΣΜΟΣ 8.1.10.**

Έλεγχος Σημαντικότητας μίας μηδενικής υπόθεσης H_0:

Όταν δεν προκαθορίζεται η στάθμη σημαντικότητας (σ.σ.) α και υπολογίζεται η τιμή πιθανότητας P ($P -$ value) από τα δειγματικά δεδομένα τότε, *ανάλογα με την επιθυμητή σ.σ.* α, απορρίπτεται η H_0 εφ' όσον είναι $P < \alpha$.

● Οι στατιστικές δοκιμασίες αφορούν συνήθως τις εξής παραμέτρους: τη μέση τιμή μ και τη διακύμανση σ^2, εάν πρόκειται για ποσοτικές τυχαίες μεταβλητές, καθώς και τη συχνότητα / αναλογία p, εάν πρόκειται για ποιοτικές τυχαίες μεταβλητές.

9.

ΔΟΚΙΜΑΣΙΕΣ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ ΙΙ : ΜΗ ΠΑΡΑΜΕΤΡΙΚΟΙ ΣΤΑΤΙΣΤΙΚΟΙ ΕΛΕΓΧΟΙ

9.1. ΓΕΝΙΚΑ

Στις δοκιμασίες σημαντικότητας που έχουν αναφερθεί μέχρι τώρα, γίνονται οι βασικές παραδοχές ότι οι τιμές της μεταβλητής (ή των μεταβλητών) στο δείγμα προέρχονται από πληθυσμό που ακολουθεί, κυρίως, κανονική κατανομή με ίσες διακυμάνσεις και ότι το μέγεθος του δείγματος είναι επαρκές και αξιόπιστο. Δηλαδή, είναι γνωστές η μορφή και οι παράμετροι της κατανομής και, έχοντας ως βάση αυτά τα δεδομένα, διατυπώνεται η μηδενική υπόθεση H_0. Τι γίνεται, όμως, στην περίπτωση κατά την οποία δεν είναι γνωστή η κατανομή των τιμών του δείγματος; Όταν, λοιπόν, είναι *άγνωστη η κατανομή των τιμών του δείγματος*, προκειμένου να διατυπωθούν οι στατιστικοί έλεγχοι αναπτύχθηκαν *τεχνικές και μέθοδοι οι οποίες δεν προϋποθέτουν τη γνώση των παραμέτρων κάποιας υποτιθέμενης κατανομής*. Για τον λόγο αυτό, οι δοκιμασίες που εφαρμόζονται ονομάζονται **μη παραμετρικές** (non parametric) ή **απαραμετρικές** ή **ελεύθερες κατανομών** (distribution-free) **δοκιμασίες**, ενώ οι δοκιμασίες που βασίζονται σε γνωστές κατανομές ονομάζονται **παραμετρικές δοκιμασίες**.

Εκτός, όμως, από το πλεονέκτημα της μη πλήρωσης των προϋποθέσεων για την κατανομή των τιμών του δείγματος, ένας από τους λόγους που προβάλλεται για τη χρήση των μη παραμετρικών τεχνικών, συναρτάται (και) με τη θεώρηση ότι *οι μη παραμετρικοί έλεγχοι βασίζονται στην τάξη των οντοτήτων, ή στις οντότητες που περιέχουν κάποια διάταξη, και όχι στη «μαθηματική τιμή» της μεταβλητής* σε κάθε οντότητα.

Με άλλα λόγια, οι μη παραμετρικές δοκιμασίες εφαρμόζονται όταν:

1. Οι μεταβλητές είναι ποιοτικές [κλίμακες μέτρησης: ονομαστική και τακτική]
2. Οι μεταβλητές είναι ποσοτικές [κλίμακες μέτρησης: (ισο)διαστημάτων και αναλογική],

όμως είναι άγνωστη η κατανομή του πληθυσμού.

Από τα παραπάνω συνάγεται ότι όταν οι τιμές των μεταβλητών, ποιοτικών ή/και ποσοτικών, δεν υπάρχουν ξεχωριστά αλλά *δίνονται μόνο με τη μορφή πλήθους εμφανίσεων που αντιστοιχούν σε κατηγορίες*, τότε εφαρμόζονται οι μη παραμετρικές δοκιμασίες.

Επίσης, οι μη παραμετρικές μέθοδοι μπορούν να χρησιμοποιηθούν ως προπαρασκευαστικές για αντίστοιχες παραμετρικές μεθόδους. Για παράδειγμα[1], με μια μη παραμετρική μέθοδο ελέγχεται εάν ένα δείγμα προέρχεται από πληθυσμό που ακολουθεί κανονική κατανομή και στη συνέχεια εφαρμόζεται ένας παραμετρικός έλεγχος για τη μέση τιμή ή τη διακύμανση.

Δύο επί πλέον λόγοι που συνηγορούν στη χρήση των μη παραμετρικών στατιστικών ελέγχων, συναρτώνται με:

α. το μέγεθος του δείγματος $(n < 6)$[2] ή $(n \leq 7)$[3], και

β. τους υπολογισμούς, που είναι απλούστεροι από αυτούς των παραμετρικών ελέγχων

Ο πρώτος λόγος μπορεί να είναι έγκυρος, αν και υπάρχει η άποψη ότι δεν πρέπει να επιχειρείται η εξαγωγή συμπερασμάτων για ένα πληθυσμό από τόσο μικρό δείγμα. Αναφορικά με τον δεύτερο λόγο, η χρήση των Η/Υ στους υπολογισμούς μηδενίζει το αναφερόμενο επιχείρημα.

Όπως είναι γνωστό, η μέση τιμή επηρεάζεται από τις ακραίες τιμές της κατανομής. Κατά συνέπεια, και οι παραμετρικές δοκιμασίες που βασίζονται στη μέση τιμή επηρεάζονται από τις τιμές αυτές με αποτέλεσμα την ύπαρξη σφάλματος, τόσο στα συμπεράσματα όσο και στη λήψη αποφάσεων, από την ανάλυση του δείγματος. Για παράδειγμα, οι τιμές 2, 8, 9, 13, 18, 21, 25, 30, 32, ενός δείγματος, έχουν μέση τιμή $\bar{x} = 17{,}56$ και διάμεσο $M_d = 18$. Εάν από λάθος η τιμή 32 γραφόταν 132, τότε η μέση τιμή θα ήταν $\bar{x} = 28{,}67$ ενώ η διάμεσος θα παρέμενε 18. Είναι φανερό ότι το συμπέρασμα που θα προκύψει από την ανάλυση του δείγματος θα είναι λανθασμένο στη δεύτερη περίπτωση.

Κατά την εφαρμογή των παραμετρικών δοκιμασιών συγκρίνονται η μέση τιμή και η συχνότητα μιας μεταβλητής στο δείγμα ή στα δείγματα, με τις αντίστοιχες ποσότητες του πληθυσμού από τον οποίο προέρχονται το δείγμα ή τα δείγματα. Όμως, *η σύγκριση αυτή είναι ατελής* διότι είναι δυνατό η μέση τιμή στο δείγμα να συμφωνεί με τη μέση τιμή στον πληθυσμό, αλλά η κατανομή των τιμών της μεταβλητής να είναι πολύ διαφορετική στο δείγμα από ότι στον πληθυσμό. Για παράδειγμα, είναι δυνατό το δείγμα νοικοκυριών μιας γεωγραφικής περιοχής, στο οποίο υπερτερούν νοικοκυριά με 1 και 6 μέλη, και ο αντίστοιχος πληθυσμός της ίδιας περιοχής, όπου πλειοψηφούν τα νοικοκυριά με 3 και 4 μέλη, να δίνουν την ίδια μέση τιμή στη μεταβλητή «μέγεθος νοικοκυριού». Επομένως, η σχετική παραμετρική δοκιμασία θα δείχνει συμφωνία, όμως η διασπορά των νοικοκυριών ως προς το μέγεθός τους διαφέρει στο δείγμα και στον αντίστοιχο πληθυσμό και αυτό δεν φαίνεται από την παραμετρική δοκιμασία.

Βεβαίως, η χρήση των μη παραμετρικών δοκιμασιών σε τιμές που «από τη φύση τους» είναι αριθμητικές έχει το μειονέκτημα ότι χάνεται μια σημαντική ποσότητα αριθμητικής πληροφορίας. Για παράδειγμα, εάν τα υψόμετρα τριών οικισμών είναι 150μ., 305μ. και 832μ., με

[1] Φ. Κολυβά-Μαχαίρα, Ε. Μπόρα-Σέντα, 1999, σελ. 405
[2] Κ.Β. Μπαγιάτης, 1990, σελ. 280
[3] Χ. Ζαχαροπούλου, τόμος Α', 2001, σελ. 641

την κατάταξή τους σε: «πεδινός οικισμός», «ημιορεινός οικισμός», «ορεινός οικισμός», αντιστοίχως, χάνεται πληροφορία. Στην καθημερινή ζωή, όταν θέλουμε να συγκρίνουμε μεγέθη μεταξύ τους, συνήθως, η σύγκριση αφορά μέσες τιμές και όχι κατανομές. Κατά συνέπεια, οι μη παραμετρικοί έλεγχοι ίσως είναι λιγότερο κατανοητοί από τους παραμετρικούς και αυτό αποτελεί ένα ακόμη μειονέκτημα.

Συνοψίζοντας, τα κύρια *πλεονεκτήματα* των μη παραμετρικών ελέγχων σε σχέση με τους παραμετρικούς είναι:

1. Οι εφαρμογές τους είναι πιο σύντομες και σχετικά απλούστερες.
2. Εφαρμόζονται και σε ποιοτικές και σε ποσοτικές μεταβλητές.
3. Εφαρμόζονται σε πολύ μικρά δείγματα.
4. Δεν γίνονται, συνήθως, υποθέσεις για τις κατανομές που ακολουθούν οι αρχικοί πληθυσμοί, αλλά και όταν γίνονται κάποιες, αυτές είναι εξαιρετικά ασθενείς.

▶ *Το κυριότερο μειονέκτημα των μη παραμετρικών ελέγχων είναι η πολυπλοκότητα της διαδικασίας για τον προσδιορισμό των διαστημάτων εμπιστοσύνης*.

☞ Επισημαίνεται ότι σε δείγμα όπου ισχύουν οι βασικές παραμετρικές προϋποθέσεις, οι παραμετρικές δοκιμασίες εμφανίζονται περισσότερο ισχυρές από τις αντίστοιχες μη παραμετρικές. Μάλιστα[4], για να επιτευχθεί η ίδια ισχύς, μία μη παραμετρική δοκιμασία θα απαιτήσει μεγαλύτερο μέγεθος δείγματος από ό,τι η αντίστοιχη παραμετρική. Βεβαίως, σοβαρές αποκλίσεις από την κανονικότητα καθιστούν τη μη παραμετρική διαδικασία περισσότερο αποτελεσματική από την αντίστοιχη παραμετρική. Κατά συνέπεια, όταν υπάρχει επιλογή μεταξύ παραμετρικών και μη παραμετρικών δοκιμασιών, σαφώς προτιμώνται οι πρώτες.

Σε αυτό το Κεφάλαιο θα παρουσιασθούν οι περισσότερο συνηθισμένες μη παραμετρικές δοκιμασίες, όπως: δοκιμασίες καλής προσαρμογής, δοκιμασίες ανεξαρτησίας, δοκιμασίες ομογένειας και δοκιμασίες συμφωνίας, που στηρίζονται σε διάφορες μεθόδους.

9.2. ΔΟΚΙΜΑΣΙΕΣ ΚΑΛΗΣ ΠΡΟΣΑΡΜΟΓΗΣ

Οι **δοκιμασίες καλής προσαρμογής** (goodness-of-fit tests) είναι εκείνες οι οποίες ως στόχο έχουν να δείξουν *εάν και πόσο καλά η κατανομή μίας μεταβλητής στο δείγμα «προσαρμόζεται» στην αντίστοιχη κατανομή του πληθυσμού*. Εδώ, χρησιμοποιείται ο όρος «προσαρμογή» και όχι «συμφωνία» που αναφέρεται στις παραμετρικές δοκιμασίες, για να φαίνεται σαφέστερα ότι πρόκειται για τη *σύγκριση ολόκληρων των κατανομών.* Αυτό σημαίνει ότι ελέγχεται εάν οι «τιμές» της μεταβλητής στο δείγμα προσαρμόζονται σε κάποια θεωρητική κατανομή η οποία αναμένεται, σύμφωνα με τις υποθέσεις και τη θεωρία. Σημειώνεται και πάλι ότι ο όρος «τιμές» δεν ανταποκρίνεται στις «μαθηματικές τιμές» που προκύπτουν από τη μέτρηση (με γνωστές μονάδες μέτρησης) των πληροφοριών που περιλαμβάνει η μεταβλητή στο δείγμα. Πρόκειται για «τιμές» που προκύπτουν από απαριθμήσεις, απαντήσεις, καταγραφές και κατηγοριοποιήσεις των πληροφοριών που περιλαμβάνει μία μεταβλητή και παρατηρούνται στο δείγμα. *Υποτίθεται ότι στον πληθυσμό ισχύει μία συγκεκριμένη κατανομή πιθανότητας* και, με τη διαδικασία της καλής προσαρμογής, ελέγχεται κατά πόσο η κατανομή αυτή είναι ένα καλό μοντέλο «ανάγνωσης» για το δείγμα. Η σχετική απόφαση λαμβάνεται εφ' όσον η κατανομή πιθανότητας, που προσδιορίζεται στη μηδενική υπόθεση H_0, εξασφαλίζει μια καλή προσαρμογή για τα δεδομένα του δείγματος, δηλαδή *δεν υπάρχει σημαντική στατιστική διαφορά μεταξύ των δύο συνόλων*.

[4] R.E. Walpole, R.H. Myers, S.L. Myers, 1998, p. 610

10.

ΠΑΛΙΝΔΡΟΜΗΣΗ ΚΑΙ ΣΥΣΧΕΤΙΣΗ

10.1. ΕΙΣΑΓΩΓΗ

Πολλές φορές στις εφαρμογές στατιστικών μελετών και ειδικότερα όταν εξετάζονται οικονομικά, κοινωνικά, πολιτικά, γεωγραφικά κ.ά. φαινόμενα, χρειάζεται να διερευνηθεί η **ταυτόχρονη συμπεριφορά** δύο ή περισσότερων **ποσοτικών** μεταβλητών. Η ταυτόχρονη συμπεριφορά συνάδει με το γεγονός ότι εξετάζεται εάν οι τιμές της μίας μεταβλητής συνδέονται με τις τιμές των υπολοίπων ή/και εάν έχουν επίδραση στη διαμόρφωση των τιμών των υπολοίπων και τανάπαλιν. Η συμπεριφορά αυτή αποδίδεται ως ζεύγος τιμών κάθε μεταβλητής με όλες τις υπόλοιπες για κάθε οντότητα αναφοράς. Στη συνέχεια, προσδιορίζεται κάποια μαθηματική σχέση, εφ' όσον υπάρχει τέτοια, η οποία εκφράζει αυτή την αλληλοεπίδραση.

Έτσι, σε ένα στατιστικό σύνολο (πληθυσμός ή δείγμα) π.χ. δύο ποσοτικών μεταβλητών X και Y, σε κάθε στατιστικό στοιχείο / σε κάθε οντότητα αυτού αντιστοιχεί ένα διατεταγμένο ζεύγος τιμών (x, y). Το ζητούμενο είναι *ο προσδιορισμός συναρτησιακής σχέσης*[1] μεταξύ των

[1] **Συναρτησιακή** λέγεται εκείνη η σχέση όπου **σε κάθε τιμή x_i της X αντιστοιχεί μία και μόνον τιμή y_i της Y**. Δηλαδή, οι τιμές των X, Y συνδέονται με τη σχέση $y = f(x)$. Σχέσεις αυτής της μορφής απαντώνται μόνον στις θετικές επιστήμες. Έτσι, για παράδειγμα, η πλευρά ενός τετραγώνου (μεταβλητή X) και το εμβαδόν του (μεταβλητή Y) συνδέονται με συναρτησιακή σχέση, εφ' όσον η γνώση οποιασδήποτε τιμής της πρώτης μεταβλητής καθορίζει και την αντίστοιχη μοναδική τιμή της δεύτερης μεταβλητής,σύμφωνα με την $y = x^2$, .

Κατά συνέπεια, στις κοινωνικές, οικονομικές και άλλες επιστήμες του ανθρωπογεωγραφικού χώρου, δεν υφίστανται σχέσεις συναρτησιακής μορφής μεταξύ δύο μεταβλητών, διότι σε τιμή της μίας μεταβλητής δεν αντιστοιχεί μία και μόνον τιμή της άλλης. Για παράδειγμα, δύο άνδρες / γυναίκες με το ίδιο ύψος δεν έχουν συνήδύο μεταβλητών, εάν βεβαίως υπάρχει, καθορίζοντας τη μία από τις δύο ως **ανεξάρτητη** (independent) μεταβλητή ή μεταβλητή **πρόβλεψης** ή **προβλέπουσα** μεταβλητή (predictor) ή μεταβλητή **εισόδου** (input) ή μεταβλητή **παλινδρομούσα** (regressor) ή **παράγοντα** (factor) ή **ερμηνευτική** (explanatory) μεταβλητή, και την άλλη ως **εξαρτημένη** (dependent) μεταβλητή ή μεταβλητή **αποτελέσματος** (outcome) ή μεταβλητή **εξόδου** (output) ή μεταβλητή **απόκρισης / ανταπόκρισης** (response) ή μεταβλητή **κριτήριο** (criterion) ή **ερμηνευόμενη** (explained) μεταβλητή.

Διακρίνονται δύο περιπτώσεις:

1. Ο καθορισμός της ανεξάρτητης μεταβλητής υποδεικνύεται από αυτό το ίδιο το πρόβλημα που διερευνάται, ή, ισοδύναμα, από αυτό το ίδιο το φαινόμενο του οποίου μελετάται η συμπεριφορά.

Προκειμένου να μελετηθεί, για παράδειγμα, η σχέση μεταξύ της ηλικίας των ανθρώπων από 0 έως 18 ετών (μεταβλητή X) και του ύψους τους (μεταβλητή Y), καθορίζεται ως *ανεξάρτητη μεταβλητή X η ηλικία* και ως *εξαρτημένη μεταβλητή Y το ύψος*. Στην περίπτωση αυτή, η ανεξάρτητη μεταβλητή, δηλαδή η ηλικία, είναι «**ελεγχόμενη**» μεταβλητή, με την έννοια ότι εξετάζεται το ύψος ανθρώπων 0,0 ετών (μόλις έχουν γεννηθεί), 0,5 ετών, 1,0 έτους κ.ο.κ. για να βρεθεί, εάν υπάρχει, η μαθηματική σχέση που εκφράζει το ύψος (εξαρτημένη μεταβλητή Y) ως συνάρτηση της ηλικίας (ανεξάρτητη μεταβλητή X). Είναι προφανές ότι η σχέση που υπάρχει μεταξύ των δύο ποσοτικών μεταβλητών είναι *σχέση αιτίας – αποτελέσματος*.

Τα προβλήματα αυτά στα οποία:

- υπάρχει **σαφής διάκριση** *μεταξύ της ανεξάρτητης και της εξαρτημένης μεταβλητής*,
- *οι τιμές* που παίρνει η ανεξάρτητη μεταβλητή *καθορίζονται από τον ερευνητή* με **ακρίβεια**,
- η εξαρτημένη μεταβλητή, της οποίας οι τιμές προφανώς συνδέονται με τις τιμές της ανεξάρτητης ως το αποτέλεσμα μίας υφιστάμενης αιτίας, είναι **τυχαία μεταβλητή**, γεγονός που σημαίνει ότι μπορεί να πάρει για κάθε «ελεγχόμενη» τιμή της ανεξάρτητης μεταβλητής *μία οποιαδήποτε τιμή μεταξύ ορισμένων «λογικών» ορίων*,

ονομάζονται **προβλήματα παλινδρόμησης** ή **προβλήματα εξάρτησης**.

Από τα παραπάνω συνάγεται ότι *στα προβλήματα παλινδρόμησης, εκτός της αλληλοεπίδρασης, ζητείται επί πλέον (και) η μαθηματική σχέση, κυρίως γραμμική, που συνδέει τις δύο μεταβλητές*. Επομένως, οι τιμές της εξαρτημένης μεταβλητής διαμορφώνονται / μεταβάλλονται αναφορικά με τις τιμές της ανεξάρτητης.

Ώστε:

- **ΠΑΛΙΝΔΡΟΜΗΣΗ :**
  1. **ΑΝΕΞΑΡΤΗΤΗ ΜΕΤΑΒΛΗΤΗ «ΕΛΕΓΧΟΜΕΝΗ»**
  2. **ΕΞΑΡΤΗΜΕΝΗ ΤΥΧΑΙΑ ΜΕΤΑΒΛΗΤΗ**
  3. **ΣΥΝΑΡΤΗΣΙΑΚΗ ΣΧΕΣΗ**

θως και το ίδιο βάρος. Προφανώς υφίσταται σχέση μεταξύ των δύο μεταβλητών, λιγότερο ή περισσότερο έντονη, όμως αυτή η σχέση δεν είναι συναρτησιακή. Οι σχέσεις αυτού του είδους λέγονται, όπως σημειώνεται και στα επόμενα, **στοχαστικές** ή **στατιστικές** σχέσεις. Με άλλα λόγια, **στοχαστική ή στατιστική** λέγεται εκείνη η σχέση όπου **σε κάθε τιμή x_i της X αντιστοιχεί μία τιμή y_i της Y η οποία μπορεί να είναι ίση ή διάφορη της $f(x_i)$**. Ειδικότερα, *αυτού του είδους η σχέση εκτιμά μια μέση τιμή*. Έτσι, προκειμένου να περιγραφεί η αλληλοεπίδραση μεταξύ δύο ποσοτικών μεταβλητών με τη βοήθεια μαθηματικής εξίσωσης, χρησιμοποιούνται οι αντίστοιχες συναρτησιακές σχέσεις. (π.χ. Ο. Παπαδήμας – Χ. Κοίλιας, 2001, σελ. 244-245, Κ.Δ. Τραχανάς, Α.Χ. Τσεβάς, 1998, σελ. 195, J.F. Hair et al,1992, p. 31) και

2. **Δεν είναι προφανές ποια από τις μεταβλητές είναι η ανεξάρτητη και ποια η εξαρτημένη. Δηλαδή, οι δύο μεταβλητές είναι τυχαίες και συμμετέχουν εξ ί-σου στο φαινόμενο που παρατηρείται.**

Για παράδειγμα, η αναζήτηση της σχέσης που συνδέει τις ηλικίες των συζύγων σε ένα ζευγάρι, αντικατοπτρίζει ένα φαινόμενο της κατηγορίας αυτής. Στην αναφερόμενη περίπτωση, ως ανεξάρτητη μεταβλητή μπορεί να θεωρηθεί τόσο η ηλικία του άνδρα όσο και αυτή της γυναίκας, χωρίς κανένα περιορισμό, αφού και οι δύο ηλικίες είναι τυχαίες μεταβλητές.[2] Ένα άλλο παράδειγμα αφορά τη σχέση μεταξύ της απασχόλησης και της ανεργίας σε μια γεωγραφική περιοχή. Ποιά από τις δύο τυχαίες μεταβλητές μπορεί να θεωρηθεί ως εξαρτημένη ή ως ανεξάρτητη μεταβλητή;

Τα προβλήματα αυτά στα οποία και οι δύο μεταβλητές είναι τυχαίες και οι οποίες είναι δυνατό είτε να αλληλοεπηρεάζονται, είτε καμιά από τις δύο να μην επηρεάζει την άλλη[3], όμως οι δύο μεταβλητές συνδέονται / συσχετίζονται μεταξύ τους, ονομάζονται **προβλήματα συσχέτισης**. Εδώ, *δεν αναζητείται ο τρόπος εξάρτησης της μίας μεταβλητής από την άλλη*, δηλαδή η μαθηματική σχέση που ενδεχομένως τις συνδέει, αλλά ζητείται να προσδιορισθεί ένας **συντελεστής** που θα δείχνει τη μεγάλη ή μικρή σχέση μεταξύ των δύο μεταβλητών, δηλαδή τη μεταξύ τους **«ένταση συσχέτισης»**.

- Με την παλινδρόμηση αναλύεται π.χ. πώς η δαπάνη για την αγορά γάλακτος εξαρτάται από τον αριθμό των παιδιών μιας οικογένειας.
- Με τη συσχέτιση διαπιστώνεται π.χ. εάν και πώς συνδέονται η κατανάλωση γάλακτος και η κατανάλωση άλλων γαλακτοκομικών προϊόντων σε μια οικογένεια.
- Με την παλινδρόμηση αναλύεται π.χ. πώς η απασχόληση (ανεργία) εξαρτάται από τους θετικούς (αρνητικούς) ρυθμούς ανάπτυξης σε μια γεωγραφική περιοχή.
- Με τη συσχέτιση π.χ. διαπιστώνεται πώς συνδέονται η απασχόληση και η ανεργία σε μια γεωγραφική περιοχή.

Ώστε:

- **ΣΥΣΧΕΤΙΣΗ :**
  1. **ΤΥΧΑΙΕΣ ΜΕΤΑΒΛΗΤΕΣ**
  2. **ΑΛΛΗΛΟΕΠΗΡΕΑΖΟΝΤΑΙ Ή ΟΧΙ**
  3. **ΣΥΝΤΕΛΕΣΤΗΣ ΣΥΣΧΕΤΙΣΗΣ**

10.2. ΣΤΟΧΑΣΤΙΚΗ Ή ΣΤΑΤΙΣΤΙΚΗ ΕΞΑΡΤΗΣΗ

10.2.1. ΓΕΝΙΚΑ

Κατά τη μελέτη φαινομένων όπου αναλύονται, συγχρόνως, περισσότερες της μιας τυχαίες μεταβλητές (τ.μ.), εκείνο που διερευνάται αρχικά είναι η ύπαρξη ή όχι αλληλοεπίδρασης / αλληλοεξάρτησης μεταξύ τους.

Όπως είναι γνωστό[4], σε δείγμα n οντοτήτων όπου *μελετούνται ταυτοχρόνως* δύο τ.μ. *Χ και Υ, εξετάζεται η συμπεριφορά του πλήθους των τιμών* y_j, *ή* y, *που εμφανίζονται σε ζεύγη* (x_i, y_i) *ή* (x, y), ***για κάθε ξεχωριστή συγκεκριμένη τιμή*** $\boldsymbol{x_i}$ ***ή*** $\boldsymbol{x}$. Δηλαδή, για κάθε τιμή x_i

[2] Σ. Κουνιάς κ.ά., 1985, σελ. 252
[3] με την έννοια της μεταβολής των τιμών της μιας ή/και της άλλης
[4] Βλ. Κεφάλαια 4.7., 4.8.

εμφανίζεται ένα σύνολο τιμών $y_j = \{y_1, y_2, \cdots, y_j\}$ με τη μορφή (x_i, y_1), (x_i, y_2), $\cdots$, (x_i, y_j), *και όχι μία μοναδική τιμή* y_i. Σε αυτή την περίπτωση, οι τ.μ. X και Y λέγονται **εξαρτημένες μεταξύ τους** ή **στοχαστικά εξαρτημένες**. Όταν, όμως, σε κάθε τιμή x_i αντιστοιχεί *μία μοναδική τιμή* y_i, τότε οι τ.μ. X *και* Y ονομάζονται **ανεξάρτητες μεταξύ τους** ή **στοχαστικά ανεξάρτητες**.

Το στατιστικό μέγεθος που δείχνει τη ***γραμμική σχέση*** *μεταξύ των δύο τ.μ. X* και *Y και συγχρόνως «μετράει» την ένταση αυτής της σχέσης* είναι **ο συντελεστής συσχέτισης ρ_{XY}** ή **ο συντελεστής θεωρητικής συσχέτισης ρ**, ο οποίος δίνεται από τη σχέση:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)} \cdot \sqrt{Var(Y)}} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{Var(X)} \cdot \sqrt{Var(Y)}} = \rho \qquad (10.2.1)$$

Αποδεικνύεται ότι ο συντελεστής ρ_{XY} ικανοποιεί την ανισότητα

$$-1 \leq \rho_{XY} \leq +1 \qquad (10.2.2)$$

Από την παραπάνω σχέση φαίνεται ότι εάν:

i. $0 < \rho_{XY} \leq +1$, οι δύο τ.μ. X και Y χαρακτηρίζονται ως *θετικά συσχετισμένες*

ii. $-1 \leq \rho_{XY} < 0$, οι δύο τ.μ. X και Y χαρακτηρίζονται ως *αρνητικά συσχετισμένες*

iii. $\rho_{XY} = 0$, οι δύο τ.μ. X και Y χαρακτηρίζονται ως *ασυσχέτιστες μεταξύ τους*

☞ *Όταν* $\rho_{XY} = 0$, δηλαδή όταν οι τ.μ. X και Y είναι ασυσχέτιστες, *δεν σημαίνει ότι οι X και Y είναι (και) στοχαστικά ανεξάρτητες*. Στην περίπτωση αυτή *μπορεί να υπάρχει στοχαστική εξάρτηση, όμως οι δύο μεταβλητές συνδέονται με σχέση μη γραμμική.*

Όπως φαίνεται από τη σχέση (10.2.1), ο παρονομαστής είναι πάντα ο ίδιος. Κατά συνέπεια, η τιμή του συντελεστή ρ_{XY} εξαρτάται από το μέγεθος της συνδιακύμανσης σ_{XY} η οποία εκφράζεται ως το άθροισμα (ολοκλήρωμα) των γινομένων $(x-\mu_X) \cdot (y-\mu_Y)$. *Η μέγιστη τιμή της ποσότητας σ_{XY} για ένα στατιστικό σύνολο, επιτυγχάνεται όταν η σειρά κατάταξης των τιμών της X, για το σύνολο των οντοτήτων, είναι η ίδια με τη σειρά κατάταξης των τιμών της Y.* Δηλαδή, όταν μεγάλες τιμές της X $(x > \mu_X)$ εμφανίζονται συχνότερα (σε ζεύγη) με μεγάλες τιμές της Y $(y > \mu_Y)$ και, αντιστοίχως, μικρές τιμές της X συσχετίζονται συχνότερα (σε ζεύγη) με μικρές τιμές της Y. Τότε οι όροι $(x-\mu_X)$ και $(y-\mu_Y)$ εμφανίζονται, συχνότερα, είτε και οι δύο θετικοί, είτε και οι δύο αρνητικοί. Κατά συνέπεια, το γινόμενο $(x-\mu_X) \cdot (y-\mu_Y)$ εμφανίζεται συχνότερα θετικό και ο συντελεστής συσχέτισης ρ_{XY} έχει θετική τιμή. Με ανάλογο τρόπο, όταν μεγάλες τιμές της X $(x > \mu_X)$ εμφανίζονται συχνότερα με μικρές τιμές της Y $(y < \mu_Y)$, ενώ μικρές τιμές της X $(x < \mu_X)$ εμφανίζονται συχνότερα με μεγάλες τιμές της Y $(y > \mu_Y)$, τότε το γινόμενο $(x-\mu_X) \cdot (y-\mu_Y)$ εμφανίζεται συχνότερα αρνητικό και ο συντελεστής συσχέτισης ρ_{XY} έχει αρνητική τιμή.[5]

▶ Να σημειωθεί ότι όταν η συνδιακύμανση δύο μεταβλητών σε ένα στατιστικό σύνολο είναι μεγαλύτερη από τη συνδιακύμανση των ίδιων ή άλλων μεταβλητών σε ένα άλλο στατιστικό σύνολο, αυτό *δεν συνεπάγεται ότι (και) η υφιστάμενη σχέση μεταξύ των μεταβλητών είναι εντονότερη στο πρώτο στατιστικό σύνολο από εκείνη του δευτέρου*.[6]

[5] Α. Δερμάνης, 1986, τόμος.1, σελ.78-79
[6] Γ. Παπαδημητρίου, 1990, τεύχος 1, σελ .369

Ως παραδείγματα τυχαίων μεταβλητών που παρουσιάζουν τα τρία είδη συσχέτισης αναφέρονται τα ακόλουθα:

- *Το ύψος και το βάρος των ανθρώπων είναι δύο θετικά συσχετισμένες τυχαίες μεταβλητές.* Συνήθως, οι υψηλότεροι άνθρωποι είναι και βαρύτεροι.
- *Η απασχόληση και η ανεργία σε μια γεωγραφική περιοχή είναι δύο αρνητικά συσχετισμένες τυχαίες μεταβλητές.* Όπου υπάρχει μεγάλο ποσοστό απασχόλησης, το ποσοστό ανεργίας είναι χαμηλό.
- *Το ύψος και το μηνιαίο εισόδημα των ανθρώπων είναι δύο ασυσχέτιστες τυχαίες μεταβλητές.* Με κανένα τρόπο δεν επηρεάζει η μία μεταβλητή την άλλη.

● Εάν οι τυχαίες μεταβλητές X και Y *συνδέονται πλήρως* με σχέση της μορφής

$$Y = b_0 + b \cdot X \quad \text{ή} \quad y = b_0 + b \cdot x \tag{10.2.3}$$

δηλαδή με *απλή γραμμική σχέση*, τότε:[7]

$$\mu_Y = E[b_0 + b \cdot X] = E[b_0 + b \cdot x] = b_0 + b \cdot \mu_X \tag{10.2.4}$$

Είναι:

$$\sigma_{XY} = E[(X-\mu_X)\cdot(Y-\mu_Y)] = E[(X-\mu_X)\cdot(b_0 + b\cdot X - b_0 - b\cdot\mu_X)] =$$

$$= E[(X-\mu_X)\cdot(X-\mu_X)\cdot b] = b\cdot E[(X-\mu_X)^2] = b\cdot\sigma_X^2$$

και επομένως:

$$\sigma_{XY} = E[(X-\mu_X)\cdot(Y-\mu_Y)] \equiv E[(x-\mu_X)\cdot(y-\mu_Y)] = b\cdot\sigma_X^2 \tag{10.2.5}$$

Παρομοίως, ισχύει:

$$\sigma_Y^2 = E[(Y-\mu_Y)^2] = E[(b_0 + b\cdot X - b_0 - b\cdot\mu_X)^2] = E[b^2\cdot(X-\mu_X)^2] =$$

$$= b^2\cdot E[(X-\mu_X)^2] = b^2\cdot\sigma_X^2$$

Οπότε:

$$\sigma_Y^2 = E[(Y-\mu_Y)^2] \equiv E[(y-\mu_Y)^2] = b^2\cdot\sigma_X^2 \tag{10.2.6}$$

και συνεπώς

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\cdot\sigma_Y} = \frac{b\cdot\sigma_X^2}{\sigma_X|b|\sigma_X} = \frac{b}{|b|} = \begin{cases} +1, & b>0 \\ -1, & b<0 \end{cases} \tag{10.2.7}$$

Δηλαδή, *εάν υπάρχει πλήρης αλληλοεξάρτηση της μορφής (10.2.3) μεταξύ δύο μεταβλητών, ο συντελεστής συσχέτισης αυτών παίρνει τη μέγιστη $(+1)$ ή ελάχιστη τιμή του (-1), ανάλογα με το πρόσημο της ποσότητας b.*[8]

[7] Βλ. Κεφάλαιο 4.4.
[8] Α. Δερμάνης, 1986, τόμος.1, σελ.79, 81

11.

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

11.1. ΕΙΣΑΓΩΓΗ

Σε προηγούμενο κεφάλαιο αναφέρθηκαν οι μέθοδοι για τη σύγκριση των *μέσων τιμών* μίας *ποσοτικής τ.μ. X* σε δύο τυχαία και ανεξάρτητα δείγματα. Τι γίνεται, όμως, εάν πρέπει να συγκριθούν οι μέσες τιμές σε περισσότερα, των δύο, δείγματα; Για την αντιμετώπιση τέτοιων προβλημάτων χρησιμοποιείται μία μέθοδος που ονομάζεται **Ανάλυση Διακύμανσης** ή **Ανάλυση Διασποράς** (Analysis of Variance – ANOVA). Η ονομασία αυτή οφείλεται στο γεγονός ότι *η μέθοδος έχει ως στόχο να διαμερίσει τη συνολική διακύμανση ενός στατιστικού συνόλου σε επιμέρους συνιστώσες*. Κάθε συνιστώσα αντιστοιχεί σε μία συγκεκριμένη πηγή (παράγοντα) μεταβλητότητας, η οποία είναι υπεύθυνη για την παρατηρούμενη διακύμανση. Έτσι, αξιολογείται η συμμετοχή και η βαρύτητα κάθε πηγής στη συνολική μεταβλητότητα των δεδομένων.

Όπως είναι γνωστό, οι τιμές ενός φαινομένου – *ποσοτική μεταβλητή* επηρεάζονται από ένα πλήθος παραγόντων – ανεξάρτητων μεταβλητών. Προκειμένου να ελεγχθεί η επίδραση ενός μόνον παράγοντα στο θεωρούμενο φαινόμενο, εξομοιώνονται όλοι οι υπόλοιποι. Για παράδειγμα, η απόσταση – *ποσοτική μεταβλητή* που διανύει ένα Ι.Χ. αυτοκίνητο, με την ίδια ποσότητα βενζίνης, εξαρτάται από: τον τύπο του αυτοκινήτου, την ικανότητα του οδηγού, τις καιρικές συνθήκες, την ποιότητα του οδοστρώματος κ.ά. παράγοντες – *ποιοτικές μεταβλητές*. Για να ελεγχθεί εάν ο παράγοντας – ποιοτική μεταβλητή «τύπος αυτοκινήτου» επηρεάζει την κατανάλωση βενζίνης και συνεπώς τον αριθμό των χλμ. που διανύονται, εξομοιώνονται όλοι οι υπόλοιπο παράγοντες που έχουν αναφερθεί.

Η επίδραση του παράγοντα στο θεωρούμενο φαινόμενο, δηλαδή στον πληθυσμό των οντοτήτων που το περιγράφουν, έχει ως αποτέλεσμα τη δημιουργία k ανεξάρτητων κατηγοριών, οι οποίες θεωρούνται ως k ανεξάρτητοι (υπο)πληθυσμοί. Για παράδειγμα, ο πληθυσμός των μισθών των εργαζομένων σε μία επιχείρηση, ποσοτική τ.μ. $X =$ «Μισθός εργαζομένου», αναφορικά με τον παράγοντα – ποιοτική μεταβλητή «Επίπεδο σπουδών», διακρίνεται στις εξής $k = 5$ κατηγορίες: «Απόφοιτος Λυκείου», «Πτυχιούχος Ανώτερης Σχολής», «Πτυχιούχος Πανεπιστημίου», «Μεταπτυχιακός τίτλος», «Διδακτορικός τίτλος». Αυτές οι κατηγορίες θεωρούνται ως k διακριτοί ανεξάρτητοι (υπο)πληθυσμοί. Οπότε, για να ελεγχθεί εάν ο παράγοντας «επίπεδο σπουδών» επιδρά στον πληθυσμό των εργαζομένων, όσον αφορά τη διαμόρφωση του μισθού τους, θα *πρέπει να ελεγχθούν οι μέσες τιμές των μισθών* στις αντίστοιχες κατηγορίες, δηλαδή στους αντίστοιχους (υπο)πληθυσμούς.

Είναι προφανές ότι οι κατηγορίες των εργαζομένων της συγκεκριμένης επιχείρησης, αποτελούν δείγματα των αντίστοιχων (υπο)πληθυσμών. Έτσι, με την ανάλυση διακύμανσης ελέγχονται *όλες οι διαφορές* $\bar{x}_i - \bar{x}_j$ *μεταξύ των δειγματικών μέσων τιμών* $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$, *στα αντίστοιχα δείγματα* $\delta_1, \delta_2, \ldots, \delta_k$. Οι διαφορές αυτές, οι οποίες αποδίδουν τις *διαφορές μεταξύ των δειγμάτων* δ_i, *θεωρούνται στατιστικά σημαντικές ή όχι, σε σχέση με τις διακυμάνσεις*, δηλαδή τις αποκλίσεις των δειγματικών τιμών από τις αντίστοιχες μέσες τιμές, *μέσα στα δείγματα*.

Με άλλα λόγια, εάν θεωρηθεί ότι δεν υφίσταται μεταβλητότητα μεταξύ των δειγματικών μέσων τιμών, δηλαδή μεταξύ των δειγμάτων, αυτό δεν συνεπάγεται ότι οι δειγματικές τιμές είναι ίδιες· ενδέχεται να είναι διαφορετικές λόγω του τυχαίου σχηματισμού των δειγμάτων. Επομένως, *υφίσταται μεταβλητότητα μέσα σε κάθε δείγμα η οποία μπορεί να εκτιμηθεί*. Κατά συνέπεια, εάν οι μεταβλητότητες μεταξύ των δειγμάτων και του εσωτερικού τους είναι περίπου ίσες, τότε δεν υφίστανται διαφορές μεταξύ των αντίστοιχων πληθυσμιακών μέσων τιμών. Εάν όμως, η μεταβλητότητα μεταξύ των δειγμάτων είναι μεγαλύτερη από τη μεταβλητότητα μέσα στα δείγματα, τότε οι μέσες πληθυσμιακές τιμές διαφέρουν μεταξύ τους. Σημειώνεται ότι η ανάλυση διακύμανσης διέπεται από την ίδια λογική, είτε εξετάζεται η επίδραση ενός παράγοντα είτε εξετάζεται η επίδραση περισσότερων παραγόντων στη διαμόρφωση ενός φαινομένου.

11.2. Η ΛΟΓΙΚΗ ΤΟΥ ΚΡΙΤΗΡΙΟΥ ΤΗΣ ΑΝΑΛΥΣΗΣ ΔΙΑΚΥΜΑΝΣΗΣ

Για την εφαρμογή της μεθόδου απαιτούνται τρεις **βασικές υποθέσεις**:

- **1η υπόθεση**

Όλα τα δείγματα είναι **τυχαία** και **ανεξάρτητα** και έχουν: μέγεθος n_i, μέση τιμή $\bar{x}_i$ και διακύμανση s_i^2. Οι οντότητες σε κάθε δείγμα, αλλά και μεταξύ των δειγμάτων, είναι ανεξάρτητες μεταξύ τους.

- **2η υπόθεση**

Οι αντίστοιχοι πληθυσμοί της ποσοτικής τυχαίας μεταβλητής X ακολουθούν **κανονική κατανομή**. Δηλαδή είναι: $\boldsymbol{X_i \sim N(\mu_i, \sigma_i^2)}$.

- **3η υπόθεση**

Όλοι οι πληθυσμοί έχουν την *ίδια* διακύμανση σ^2. Δηλαδή είναι: $\boldsymbol{\sigma_i^2 = \sigma^2}$.

Επομένως, προκειμένου να ελεγχθεί ότι τα k δείγματα προέρχονται από τον ίδιο πληθυσμό ή από πληθυσμούς με την ίδια διακύμανση, θα πρέπει να ικανοποιούνται οι παραπάνω υποθέσεις. Οπότε, η μηδενική υπόθεση διατυπώνεται ως:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \tag{11.2.1}$$

Είναι γνωστό[1] ότι για τη σύγκριση δύο δειγμάτων των οποίων οι αντίστοιχοι πληθυσμοί ακολουθούν κανονική κατανομή και έχουν ίσες διακυμάνσεις, $\sigma_1^2 = \sigma_2^2 = \sigma^2$, η ποσότητα / το κριτήριο που χρησιμοποιείται για τη σύγκριση αυτή είναι είτε η τ.μ. Z, είτε η τ.μ. T, ανάλογα με τα μεγέθη των δειγμάτων. Δηλαδή:

$$Z \equiv Z_{\bar{X}_1-\bar{X}_2} = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim N(0,1) \tag{11.2.2}$$

και

$$T \equiv T_{\bar{X}_1-\bar{X}_2} = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t_\nu \equiv t_{n_1+n_2-2} \tag{11.2.3}$$

όπου, η ποσότητα:

$$s^2 \equiv s_p^2 = \frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2} \tag{11.2.4}$$

ή

$$s^2 \equiv s_p^2 = \frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2}{n_1 + n_2 - 2} \tag{11.2.5}$$

είναι η *εκτίμηση της κοινής πληθυσμιακής διακύμανσης* $\sigma_1^2 = \sigma_2^2 = \sigma^2$, όταν αυτή δεν είναι γνωστή.

Εάν, για παράδειγμα, υπήρχαν $k = 4$ δείγματα, οι συγκρίσεις μεταξύ των μέσων τιμών ανά δύο θα έδιναν:

$H_{01} : \mu_1 = \mu_2, \quad H_{02} : \mu_1 = \mu_3, \quad H_{03} : \mu_1 = \mu_4,$

$H_{04} : \mu_2 = \mu_3, \quad H_{05} : \mu_2 = \mu_4, \quad H_{06} : \mu_3 = \mu_4$

Όπως φαίνεται, δημιουργείται ένα πλήθος $m = \binom{k}{2} = \binom{4}{2} = 6$ συγκρίσεων ανά δύο, δηλαδή 6 διαδοχικών δοκιμασιών· όσο αυξάνεται ο αριθμός των προς σύγκριση δειγμάτων, τόσο αυξάνεται και το πλήθος των σχετικών δοκιμασιών. Όμως, σε αυτές τις συγκρίσεις, το *βασικό μειονέκτημα* δεν είναι το πλήθος αυτών. *Όσο αυξάνεται το πλήθος των δοκιμασιών, τόσο αυξάνεται και η πιθανότητα να απορριφθεί η μηδενική υπόθεση, ενώ είναι σωστή*, δηλαδή να συμβεί λάθος τύπου Ι.

Έτσι, για το θεωρούμενο παράδειγμα η H_0 απορρίπτεται εάν, με την εφαρμογή διαδοχικών δοκιμασιών, διαπιστωθεί ότι ισχύει τουλάχιστον μία από τις σχέσεις:

[1] Βλ. Κεφάλαιο 7.5.

$H_{11} : \mu_1 \neq \mu_2, \quad H_{12} : \mu_1 \neq \mu_3, \quad H_{13} : \mu_1 \neq \mu_4,$

$H_{14} : \mu_2 \neq \mu_3, \quad H_{15} : \mu_2 \neq \mu_4, \quad H_{16} : \mu_3 \neq \mu_4$

Για κάθε μία δοκιμασία, από τις παραπάνω, η πιθανότητα να γίνει λάθος τύπου Ι είναι α. Επομένως, η πιθανότητα αποδοχής της H_0, σε κάθε δοκιμασία, είναι $P(H_{0i}) = 1 - \alpha = 1 - 0{,}05 = 0{,}95$ σε σ.σ. $\alpha = 0{,}05$. Άρα, η πιθανότητα να γίνει λάθος τύπου Ι στο σύνολο των 6 διαδοχικών δοκιμασιών, θα προκύπτει ως μία συνάρτηση του λάθους τύπου Ι που μπορεί να γίνει στην κάθε μία από αυτές. Δηλαδή:

$$P(\text{συνολικό λάθος τύπου } I) = 1 - P(H_{01} \cap H_{02} \cap \ldots \cap H_{06}) =$$

$$= 1 - P(H_{01}) \cdot P(H_{02}) \cdot \ldots \cdot P(H_{06}) =$$

$$= 1 - 0{,}95 \times 0{,}95 \times \ldots \times 0{,}95 = 1 - 0{,}95^6 =$$

$$= 1 - 0{,}735092 = 0{,}264908 = 26{,}49\%$$

Όπως φαίνεται, η χρήση διαδοχικών δοκιμασιών για τον έλεγχο της $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ έχει ως αποτέλεσμα μεγαλύτερο συνολικό λάθος τύπου Ι για την απόρριψή της, όταν αυτή είναι αληθής, αναφορικά με το λάθος τύπου Ι κάθε διακριτής (μεμονωμένης) δοκιμασίας. Γενικά, η συνολική πιθανότητα για τη λανθασμένη απόρριψη της H_0, υπολογίζεται σύμφωνα με τη σχέση: $1 - (1 - \alpha)^m$, όπου m είναι το πλήθος των δυνατών διαδοχικών συγκρίσεων και $P(H_{0i}) = 1 - \alpha$.

Όταν, λοιπόν, το πλήθος των δειγμάτων είναι $k > 2$, η *εκτίμηση της κοινής διακύμανσης* $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$, θα δίνεται από τον συνδυασμό των επιμέρους k δειγματικών διακυμάνσεων s_i^2, ο οποίος δεν είναι άλλος από τη *σταθμισμένη μέση τιμή* αυτών, ως προς τους αντίστοιχους *βαθμούς ελευθερίας*:

$$s_e^2 \equiv s^2 = \frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2 + \cdots + (n_k - 1) \cdot s_k^2}{n_1 + n_2 + \cdots + n_k - k} \tag{11.2.6}$$

ή

$$s_e^2 \equiv s^2 = \frac{\sum_{j=1}^{n_1}(x_{1j} - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_{2j} - \bar{x}_2)^2 + \cdots + \sum_{j=1}^{n_k}(x_{kj} - \bar{x}_k)^2}{n_1 + n_2 + \cdots + n_k - k} \tag{11.2.7}$$

ή

$$s_e^2 \equiv s^2 = \frac{1}{\sum_{i=1}^{k}(n_i - 1)} \cdot \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2 \tag{11.2.8}$$

Αποδεικνύεται ότι:

$$E[s_e^2] = \sigma^2 \tag{11.2.9}$$

Δηλαδή, η ποσότητα $\boldsymbol{s_e^2}$ είναι η (συνολική) **διακύμανση των τιμών $\mathbf{x_{ij}}$ μέσα σε όλα τα δείγματα $\boldsymbol{\delta_i}$**, αναφορικά με τη μέση τιμή $\bar{x}_i$ κάθε δείγματος. Με άλλα λόγια, είναι η διακύμανση των σφαλμάτων (error) και αποτελεί ένα *αμερόληπτο εκτιμητή της πληθυσμιακής διακύμανσης* σ^2.