Σχολή Διοίκησης - Οικονομίας ΤΕΙ Λάρισας
Τμήμα Λογιστικής

Σημειώσεις

ΣΤΑΤΙΣΤΙΚΗΣ

ΕΠΙΧΕΙΡΗΣΕΩΝ

Θεόδωρος Γ. Λόκκας

Μαθηματικός M.Sc.

Αν. Καθηγητής ΤΕΙ Λάρισας

Λάρισα, Απρίλιος 2013

Πίνακας περιεχομένων

Κεφάλαιο 1.

Εισαγωγικές έννοιες 4

§ 1. 1. Εισαγωγή 4

§ 1. 2. Μεταβλητές μέτρησης 6

1) Ποσοτικές μεταβλητές 6

2) Διαβαθμισμένες ή ιεραρχικές μεταβλητές 6

3) Ποιοτικές μεταβλητές 6

§ 1. 3. Διακριτές – συνεχείς μεταβλητές 6

Κεφάλαιο 2.

Περιγραφική Στατιστική 8

§ 2. 1. Παρουσίαση στατιστικών δεδομένων – Κατανομές συχνοτήτων 8

Παράδειγμα 1 8

Παράδειγμα 2 9

Παράδειγμα 3 9

§ 2. 2. Ομαδοποίηση των παρατηρήσεων 10

Παράδειγμα 4 11

§ 2. 3. Ιστογράμματα Συχνοτήτων 12

§ 2. 4. Συμβολισμοί για το άθροισμα τιμών 14

Κανόνες άθροισης 14

Βασικά συμπεράσματα 14

§ 2. 5. Μέτρα κεντρικής τάσης και διασποράς 15

2. 5. 1. Μέτρα κεντρικής τάσης 15

2. 5. 2. Μέτρα διασποράς 18

α) Το εύρος R 19

β) Η διακύμανση s^2 19

γ) Η τυπική απόκλιση 20

δ) Συντελεστής μεταβλητότητας 21

Κεφάλαιο 3.

Εισαγωγή στη θεωρία πιθανοτήτων. 23

Κατανομές πιθανοτήτων 23

§ 3. 1. Βασικές έννοιες στη θεωρία πιθανοτήτων 23

§ 3. 2. Στοιχεία θεωρίας συνόλων 24

§ 3. 3. Στοιχεία συνδυαστικής ανάλυσης 24
§ 3. 4. Ορισμοί πιθανότητας και βασικά θεωρήματα 26
1ος ορισμός Πιθανότητας γεγονότος 26
2ος ορισμός Πιθανότητας γεγονότος 27
Συμπληρωματικά – Ασυμβίβαστα γεγονότα 28
Προσθετικό θεώρημα 28
Δεσμευμένη πιθανότητα (θεώρημα Bayes) 28
Γενικό παράδειγμα 29
§ 3. 5. Τυχαίες μεταβλητές και κατανομές πιθανοτήτων 30
§ 3. 6. Διωνυμική κατανομή B(n, p) (Bernoulli) 33
§ 3. 7. Υπεργεωμετρική κατανομή 35
§ 3. 8. Κανονική κατανομή (ή κατανομή Gauss) 37
§ 3. 9. Σύγκλιση διωνυμικής προς κανονική κατανομή 40
§ 3. 10. Κατανομή Poisson 40
§ 3. 11. Κατανομή t του Student 41
§ 3. 12. Κατανομή X^2 41
§ 3. 13. Κατανομή F 42
§ 3.14. Στατιστικοί πίνακες των κατανομών t, X2 και F 42
Ασκήσεις 43

Κεφάλαιο 4.

Δειγματοληπτικές κατανομές – Εκτιμητική 46

§ 4. 1. Δειγματοληπτικές κατανομές 46
§ 4. 2. Διαστήματα εμπιστοσύνης 48
Α) Διάστημα εμπιστοσύνης για το μέσο όρο μ ενός πληθυσμού 49
Β) Διάστημα εμπιστοσύνης για την αναλογία p ενός χαρακτηριστικού σε κάποιο πληθυσμό 51
Παρατηρήσεις για τα διαστήματα εμπιστοσύνης 52
Ασκήσεις 52

Κεφάλαιο 5.

Συσχέτιση - Παλινδρόμηση 55

§ 5. 1. Συντελεστής συσχέτισης 55
§ 5. 2. Ευθεία παλινδρόμησης ή ελαχίστων τετραγώνων 58

Κεφάλαιο 6.

Έλεγχοι στατιστικών υποθέσεων 60

§ 6. 1. Γενικές αρχές 60

§ 6. 2. Σφάλματα τύπου Ι και ΙΙ. – Ισχύς ελέγχου 61
§ 6. 3. Έλεγχος συμφωνίας του μέσου όρου 62
§ 6. 4. Έλεγχος ισότητας δύο μέσων τιμών σε ανεξάρτητα δείγματα...........66
§ 6. 5. Έλεγχος ισότητας δύο διακυμάνσεων 70
§ 6. 6. Έλεγχος ισότητας δύο μέσων τιμών σε συζευγμένα δείγματα...........71
§ 6. 7. Έλεγχος συμφωνίας μιας αναλογίας 74
§ 6. 8. Έλεγχος ανεξαρτησίας δύο ποιοτικών μεταβλητών – δοκιμασία Χ2 .75
§ 6. 9. Έλεγχος ισότητας δύο αναλογιών (ποσοστών) 79
Ασκήσεις 81
Στατιστικοί Πίνακες...85
Πίνακας 1: Τυποποιημένη κανονική κατανομή Ζ..85
Πίνακας 2: Κατανομή t του Student...87
Πίνακας 3: Χ2 – Κατανομή...88
Πίνακας 4: Κατανομή F (α = 0,05) 89
Πίνακας 4: Κατανομή F (α = 0,01) 90
Βιβλιογραφία...91
Α. Ξένη 91
Β. Ελληνική 91

ΣΤΑΤΙΣΤΙΚΗ ΕΠΙΧΕΙΡΗΣΕΩΝ

Κεφάλαιο 1.

Εισαγωγικές έννοιες

§ 1. 1. Εισαγωγή

Στατιστική είναι ο κλάδος της επιστήμης που ασχολείται με την *περιγραφή, ανάλυση* και *ερμηνεία* δεδομένων με σκοπό την αντικειμενική αξιολόγηση της εγκυρότητας των συμπερασμάτων που βασίζονται στα δεδομένα.

Τα δεδομένα αυτά μπορούν να συγκεντρωθούν είτε κατόπιν παρατηρήσεων χωρίς επέμβαση του ερευνητού, είτε από κατάλληλα σχεδιασμένα πειράματα. Να σημειωθεί ότι κατά το στάδιο του σχεδιασμού των πειραμάτων, η προσφορά της στατιστικής μεθοδολογίας είναι ουσιαστική.

Βασικό στοιχείο της Στατιστικής αποτελεί η έννοια της μεταβλητής. Η έννοια αυτή αναφέρεται στην διαφοροποίηση (μεταβλητότητα) των στοιχείων μίας ομάδας ή ενός συνόλου ως προς κάποιο χαρακτηριστικό. Κάθε χαρακτηριστικό που παρουσιάζει μεταβλητότητα θα το ονομάζουμε **μεταβλητή**. Έτσι, το ύψος, το βάρος, το φύλο, το χρώμα, η ηλικία, ο δείκτης ευφυΐας, ο χρόνος αντίδρασης σε κάποιο ερέθισμα, η βαθμολογία σε κάποια εξέταση, αποτελούν μεταβλητές. Η έννοια 'δεδομένα' ή 'παρατηρήσεις' ή 'μετρήσεις', αναφέρεται ακριβώς στις τιμές των μεταβλητών. Π.χ. η τιμή 105 της μεταβλητής 'δείκτης ευφυΐας' είναι ένα δεδομένο (μία παρατήρηση). Οι μεταβλητές συμβολίζονται με κεφαλαία γράμματα όπως X, Y, Z ενώ οι τιμές τους με αντίστοιχα μικρά. Στο παραπάνω παράδειγμα π. χ., η μεταβλητή 'δείκτης ευφυΐας' συμβολίζεται με X, ενώ μία τιμή της, όπως η 105, συμβολίζεται με x.

Το σύνολο όλων των δυνατών στοιχείων για τα οποία ενδιαφερόμαστε να εξάγουμε πληροφορίες, θα το ονομάζουμε **πληθυσμό**. Ο όρος πληθυσμός αναφέρεται συνήθως σε ένα πολύ μεγάλο σύνολο ατόμων ή πραγμάτων πάνω στο οποίο παίρνει τιμές η μεταβλητή που μελετούμε. Έτσι ομιλούμε για τον πληθυσμό των ανθρώπων μιας χώρας ή μιας πόλης, ή αυτών που ανήκουν σε κάποιο κοινωνικο-οικονομικό επίπεδο, αλλά ακόμα για τον πληθυσμό ζώων, φυτών, μικροοργανισμών, ατυχημάτων, βιομηχανικών ειδών, λέξεων, μετρήσεων κ.τ.λ.

Ένας πληθυσμός λέγεται *άπειρος* όταν τα στοιχεία που τον αποτελούν είναι άπειρα και συνεπώς δεν μπορούν να καταμετρηθούν. Αντίθετα λέγεται *πεπερασμένος* όταν τα στοιχεία του μπορούν θεωρητικά να καταμετρηθούν.

Οποιοδήποτε τμήμα (σχετικά μικρό) ενός πληθυσμού θα το ονομάζουμε **δείγμα**, ενώ το πλήθος των στοιχείων από τα οποία απαρτίζεται το δείγμα το ονομάζουμε **μέγεθος** του δείγματος και το συμβολίζουμε με n. Έτσι, όταν γράφουμε ότι ένα δείγμα είναι μεγέθους $n = 40$ εννοούμε ότι το δείγμα αποτελείται από 40 στοιχεία.

Επειδή τις περισσότερες φορές είναι αδύνατη η μελέτη όλου του πληθυσμού, περιοριζόμαστε στη μελέτη ενός δείγματος παρμένου μέσα από τον πληθυσμό και κατόπιν, με τη βοήθεια κατάλληλων στατιστικών μεθόδων εξάγουμε συμπεράσματα που αφορούν τον πληθυσμό. Εάν ένα δείγμα παρθεί από ένα πληθυσμό κατά τρόπον ώστε όλα τα στοιχεία του πληθυσμού να έχουν την ίδια πιθανότητα να συμπεριληφθούν στο δείγμα, τότε αυτό ονομάζεται **τυχαίο δείγμα**. Η χρησιμοποίηση τυχαίων δειγμάτων είναι πολύ συχνή στις στατιστικές μελέτες.

Στο εξής μια ποσότητα η οποία υπολογίζεται με βάση τα *δεδομένα ενός δείγματος* θα την ονομάζουμε **στατιστικό,** ενώ **παράμετρο** θα ονομάζουμε μια ποσότητα χαρακτηριστική *ολόκληρου του πληθυσμού*. Π. χ. ο αριθμητικός μέσος όρος της ηλικίας των ατόμων ενός δείγματος παρμένου μέσα από τον πληθυσμό μιας μεγάλης πόλης, αναφέρεται ως στατιστικό, ενώ ο αριθμητικός μέσος όρος της ηλικίας όλων των κατοίκων της πόλης αναφέρεται ως παράμετρος. Τα στατιστικά υπολογίζονται για να περιγράψουν τα δεδομένα ενός δείγματος και να εκτιμήσουν ή να ελέγξουν υποθέσεις για ορισμένα χαρακτηριστικά του πληθυσμού. Συνήθως, τα στατιστικά συμβολίζονται με λατινικούς χαρακτήρες, ενώ οι παράμετροι με ελληνικούς. Έτσι, όπως θα δούμε παρακάτω, ο αριθμητικός μέσος ενός δείγματος συμβολίζεται με $\bar{x}$ και η διακύμανση με s^2, ενώ οι αντίστοιχοι παράμετροι με μ και σ^2.

Το μέρος της στατιστικής μεθοδολογίας που χρησιμοποιείται για την *συστηματική περιγραφή* των ιδιοτήτων των δειγμάτων, ή των ίδιων των πληθυσμών (όταν αυτό είναι δυνατόν), ονομάζεται **Περιγραφική Στατιστική**. Συνεπώς, οποιαδήποτε μελέτη η οποία αναφέρεται στα δεδομένα ενός κάποιου δείγματος με σκοπό την παρουσίασή τους κατά τρόπο κατανοητό και ταυτόχρονα πληροφοριακό, ανήκει στο χώρο της Περιγραφικής Στατιστικής. Αντίθετα, το μέρος των στατιστικών μεθόδων που χρησιμοποιούνται για την *εξαγωγή συμπερασμάτων* που αφορούν ιδιότητες πληθυσμών, με βάση *δειγματικά δεδομένα*, ονομάζεται **Επαγωγική Στατιστική**. Ιδιαίτερα σημαντικά κεφάλαια της Στατιστικής, όπως η *Εκτιμητική* και οι *Έλεγχοι Υποθέσεων* ανήκουν στο χώρο της Επαγωγικής Στατιστικής.

§ 1. 2. Μεταβλητές μέτρησης

Υπάρχουν τρεις βασικές κατηγορίες μεταβλητών που χρησιμοποιούνται για τη μέτρηση δεδομένων και οι οποίες κατά σειρά είναι οι εξής:

1) Ποσοτικές μεταβλητές

Οι μεταβλητές αυτές είναι μετρήσιμες και οι τιμές τους είναι πραγματικοί αριθμοί. Ποσοτικές μεταβλητές είναι το βάρος, το μήκος, ο όγκος, ο χρόνος, η πίεση, η θερμοκρασία, το εισόδημα, οι ετήσιες αποδοχές κ.λ.π. Σε δεδομένα μετρημένα με ποσοτικές μεταβλητές, μπορεί να εφαρμοστεί το μεγαλύτερο και πλουσιότερο μέρος της στατιστικής μεθοδολογίας.

2) Διαβαθμισμένες ή ιεραρχικές μεταβλητές

Οι μεταβλητές αυτές δεν είναι μετρήσιμες αλλά εκφράζουν διαβάθμιση ή ιεραρχία. Π.χ. η βαθμολογία ενός σπουδαστή ή η ποιότητα ενός προϊόντος μπορεί να χαρακτηριστεί από κακή μέχρι άριστη, με ενδιάμεση διαβάθμιση.

3) Ποιοτικές μεταβλητές

Ποιοτικές ονομάζονται οι μεταβλητές οι οποίες δεν επιδέχονται μέτρηση, αλλά εκφράζουν μια ιδιότητα ή κατηγορία. Τα αποτελέσματα των παρατηρήσεων μιας ποιοτικής μεταβλητής ταξινομούνται σε κατηγορίες, γι' αυτό οι μεταβλητές αυτές λέγονται και *κατηγορικές*. Παραδείγματα τέτοιας μεταβλητής αποτελούν το φύλο, η οικογενειακή κατάσταση, το χρώμα, το επάγγελμα, κ.τ.λ. Όταν η μεταβλητή αυτή έχει μόνο δύο κατηγορίες, όπως το φύλο (άρρεν - θήλυ), λέγεται *διχοτομική*.

§ 1. 3. Διακριτές – συνεχείς μεταβλητές

Μια ποσοτική μεταβλητή ονομάζεται **διακριτή** όταν οι τιμές της αποτελούν ένα σύνολο αριθμών μεμονωμένων και καλά προσδιορισμένων. Συνήθως είναι διαδοχικοί ακέραιοι αριθμοί. Π.χ. ο αριθμός των αγοριών σε οικογένειες των τεσσάρων παιδιών αποτελεί διακριτή μεταβλητή με δυνατές τιμές τις 0, 1, 2, 3, 4. Επίσης, ο αριθμός των σωστών απαντήσεων που δίνονται σε ερωτηματολόγιο που περιλαμβάνει δέκα ερωτήσεις με σκοπό την διερεύνηση της γνώσης των ερωτώμενων σε κάποιο θέμα, αποτελεί διακριτή μεταβλητή με δυνατές τιμές τις 0, 1, 2, … , 10. Οι διακριτές μεταβλητές ονομάζονται και *ασυνεχείς* ή *απαριθμητές*.

Όταν μια ποσοτική μεταβλητή μπορεί να πάρει οποιαδήποτε τιμή σε ένα κάποιο διάστημα τιμών ονομάζεται **συνεχής.** Έτσι, όταν μετρούμε μήκος, μεταξύ των τιμών 40cm και 41cm μπορεί να υπάρξει ένας άπειρος αριθμός τιμών μήκους. Μεταβλητές όπως το βάρος, το μήκος, η θερμοκρασία, ο χρόνος κ.λ.π. αποτελούν παραδείγματα συνεχών μεταβλητών.

Κεφάλαιο 2.

Περιγραφική Στατιστική

§ 2. 1. Παρουσίαση στατιστικών δεδομένων – Κατανομές συχνοτήτων

Η πιο στοιχειώδης μορφή παρουσίασης στατιστικών δεδομένων συνίσταται στην απλή παράθεσή τους ταξινομημένα κατά αύξουσα (ή φθίνουσα) σειρά. Μια τέτοια παρουσίαση δεδομένων ονομάζεται *στατιστική σειρά*.

Ας υποθέσουμε ότι $x_1, x_2, \ldots, x_k$ είναι οι τιμές μιας μεταβλητής *X*, που αφορά τα άτομα ενός δείγματος μεγέθους *n*, $k \leq n$. Στην τιμή x_i αντιστοιχίζεται η συχνότητα n_i δηλαδή ο φυσικός αριθμός που δείχνει πόσες φορές εμφανίζεται η τιμή x_i της μεταβλητής *X* στο σύνολο των παρατηρήσεων. Είναι φανερό ότι το άθροισμα όλων των συχνοτήτων είναι ίσο με το μέγεθος *n* του δείγματος. Δηλαδή:

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

Παράδειγμα 1

Η βαθμολογία ενός δείγματος 10 μαθητών σε ένα μάθημα, ταξινομημένη κατά αύξουσα σειρά είναι: 11, 12, 12, 13, 15, 15, 15, 16, 17, 19. Έτσι, η τιμή 12 έχει συχνότητα 2, η τιμή 15 έχει συχνότητα 3, ενώ οι άλλες τιμές 11, 13, 16, 17, 19 έχουν συχνότητα 1 γιατί εμφανίζονται μόνο μία φορά. Όπως φαίνεται, οι συχνότητες είναι ακέραιοι θετικοί αριθμοί.

Όταν τα στοιχεία είναι πολυάριθμα, είναι χρήσιμο να τα παρουσιάζουμε συνοπτικά με τη μορφή μιας **κατανομής συχνοτήτων**. Τέτοιες κατανομές συχνοτήτων μπορούμε να χρησιμοποιήσουμε είτε για ποιοτικά, είτε για ποσοτικά δεδομένα. Μια κατανομή συχνοτήτων δείχνει το πώς κατανέμονται οι συχνότητες σε σχέση με τις διάφορες διακεκριμένες κατηγορίες που μελετούμε. Ας σημειωθεί ότι, πάντοτε το άθροισμα των συχνοτήτων μας δίνει το συνολικό αριθμό των παρατηρήσεων, δηλαδή το μέγεθος του δείγματος.

Παράδειγμα 2

Κατανομή συχνοτήτων της μεταβλητής ‘επίδοση μαθητών’ με τέσσερις κατηγορίες: Ανεπαρκής, Μέτρια, Καλή, Άριστη. Η μεταβλητή ‘εφαρμόστηκε’ σε δείγμα 100 μαθητών.

Επίδοση	Συχνότητα (n_i)
Ανεπαρκής	15
Μέτρια	40
Καλή	35
Άριστη	10

Όπως παρατηρούμε, το άθροισμα των συχνοτήτων ισούται με το μέγεθος του δείγματος δηλαδή 100. Παρόμοιες είναι οι κατανομές συχνοτήτων ποιοτικών δεδομένων, όπου βέβαια για τις διάφορες κατηγορίες δεν υπάρχει θέμα ταξινόμησής τους κατά αύξουσα ή φθίνουσα σειρά.

Παράδειγμα 3

Κατανομή συχνοτήτων των σωστών απαντήσεων 700 ατόμων σε ερωτηματολόγιο με 6 ερωτήσεις. Περίπτωση διακριτής μεταβλητής.

Σωστές απαντ.	Συχνότητα (n_i)
0	50
1	70
2	80
3	180
4	200
5	80
6	40

Σε ένα δείγμα 700 ατόμων δόθηκε προς απάντηση ένα σύνολο 6 ερωτήσεων, όπου κάθε ερώτηση επιδέχονταν σωστή ή λάθος απάντηση. Ο αριθμός των σωστών απαντήσεων αποτελεί διακριτή μεταβλητή με δυνατές τιμές 0, 1, 2, 3, 4, 5, 6. Όπως παρατηρούμε από τα παραπάνω αποτελέσματα της κατανομής συχνοτήτων, 50 άτομα δεν απάντησαν σωστά σε καμία ερώτηση, 70 άτομα απάντησαν σωστά μόνο σε μία ερώτηση, 80 άτομα απάντησαν σωστά σε δύο από τις 6 ερωτήσεις κ.ο.κ.

Αν διαιρέσουμε τη συχνότητα n_i με το μέγεθος n του δείγματος προκύπτει η **σχετική συχνότητα** f_i της τιμής x_i δηλαδή $f_i = \dfrac{n_i}{n}$, όπου $i = 1, 2, \ldots, k$

Για τη σχετική συχνότητα ισχύουν οι ιδιότητες:

1) $0 \le f_i \le 1$ για $i = 1, 2, \ldots, k$ αφού $0 \le n_i \le n$.

2) $f_1 + f_2 + \ldots + f_k = 1$, αφού

$$f_1 + f_2 + \ldots + f_k = \frac{n_1}{n} + \frac{n_2}{n} + \ldots + \frac{n_k}{n} = \frac{n_1 + n_2 + \ldots + n_k}{n} = \frac{n}{n} = 1 .$$

Συνήθως τις σχετικές συχνότητες τις εκφράζουμε επί τοις εκατό, οπότε συμβολίζονται με $f_i\,\%$, δηλαδή $f_i\,\% = 100\, f_i$.

Στην περίπτωση των ποσοτικών μεταβλητών, εκτός από τις συχνότητες n_i και f_i χρησιμοποιούνται συνήθως και οι λεγόμενες **αθροιστικές συχνότητες** N_i και οι **αθροιστικές σχετικές συχνότητες** F_i, οι οποίες εκφράζουν το πλήθος και το ποσοστό αντίστοιχα των παρατηρήσεων που είναι μικρότερες ή ίσες της τιμής x_i. Συχνά οι F_i πολλαπλασιάζονται επί 100, εκφραζόμενες έτσι επί τοις εκατό, δηλαδή $F_i\% = 100 F_i$.

§ 2. 2. Ομαδοποίηση των παρατηρήσεων

Όταν τα δεδομένα μας περιλαμβάνουν μεγάλο πλήθος διακεκριμένων τιμών και ειδικά όταν είναι συνεχή, ο καλύτερος τρόπος κατασκευής της κατανομής συχνοτήτων είναι η ομαδοποίησή τους σε διαδοχικές **κλάσεις** (μικρό πλήθος ομάδων) ίσου μεγέθους. Το πλήθος των δεδομένων που ανήκουν σε κάθε κλάση το ονομάζουμε *συχνότητα της κλάσης* και το συμβολίζουμε όπως και πριν με n_i. Τα άκρα των κλάσεων είναι τα όρια των κλάσεων με τη μορφή [-), όπου η δεξιά τιμή των ορίων δεν συμπεριλαμβάνεται στην κλάση. Οι παρατηρήσεις κάθε κλάσης θεωρούνται όμοιες και αντιπροσωπεύονται από τις **κεντρικές τιμές** x_i δηλαδή τα κέντρα κάθε κλάσης.

Για την ομαδοποίηση των δεδομένων υπεισέρχεται το ερώτημα της επιλογής του αριθμού των κλάσεων. Επειδή ο σκοπός της ομαδοποίησης είναι η σύμπτυξη των δεδομένων, θα πρέπει ο αριθμός των τάξεων να μην είναι ούτε πολύ μεγάλος (οπότε κάθε τάξη θα περιείχε λίγα δεδομένα και η ομαδοποίηση θα έχανε το νόημά της), αλλά ούτε και πολύ μικρός, γιατί έτσι οι πληροφορίες που θα πάρουμε από την κατανομή θα υστερούν πολύ από πλευράς ακρίβειας.

- Το πρώτο βήμα στην ομαδοποίηση των δεδομένων είναι η **εκλογή του αριθμού *κ* των ομάδων ή κλάσεων.** Ο αριθμός αυτός συνήθως ορίζεται αυθαίρετα και είναι θέμα εμπειρίας. Γενικά όμως χρησιμοποιείται ο παρακάτω εμπειρικός πίνακας:

Μέγ. δείγματος n	Αριθ. κλάσεων $κ$	Μέγ. δείγματος n	Αριθ. κλάσεων $κ$
<20	5	200 – 400	9
20 – 50	6	400 – 700	10
50 – 100	7	700 – 1000	11
100 – 200	8	>1000	12

- Το δεύτερο βήμα είναι ο προσδιορισμός του **πλάτους *c* των κλάσεων**, που είναι η διαφορά του κατώτερου από το ανώτερο όριο της κλάσης. Βέβαια, υπάρχουν περιπτώσεις όπου επιβάλλεται οι κλάσεις να έχουν άνισο πλάτος, όπως στις κατανομές εισοδήματος, ημερών απεργίας κ.τ.λ. Για ισοπλατείς κλάσεις ο ***Sturges*** πρότεινε τον τύπο: $c = \dfrac{R}{1 + 3{,}322 \cdot \log n}$ όπου R (range) είναι το εύρος του δείγματος (διαφορά της μικρότερης από τη μεγαλύτερη τιμή του συνολικού δείγματος n). Το πλάτος c των κλάσεων στρογγυλεύεται προς τα επάνω και κατά προτίμηση είναι άρτιος αριθμός. Βασικά, διαιρώντας το εύρος R δια του αριθμού των κλάσεων κ, έχουμε το πλάτος c των κλάσεων, που συμπίπτει κατά μεγάλο βαθμό με τον τύπο του *Sturges*. Υποτίθεται ότι δεν υπάρχουν ακραίες τιμές (πολύ μικρές ή μεγάλες, ως προς τις κύριες τιμές).

- Το επόμενο βήμα είναι η **κατασκευή** των κλάσεων. Ξεκινώντας από τη μικρότερη παρατήρηση, ή για πρακτικούς λόγους λίγο πιο κάτω από τη μικρότερη παρατήρηση, και προσθέτοντας κάθε φορά το πλάτος c δημιουργούμε τις κ κλάσεις. Προφανώς η μεγαλύτερη τιμή του δείγματος πρέπει να ανήκει στην τελευταία κλάση.

- Τέλος, γίνεται η **διαλογή** των παρατηρήσεων. Το πλήθος των παρατηρήσεων n_i που προκύπτουν από τη διαλογή για την κλάση i είναι η συχνότητα f_i της κλάσης αυτής, ή η συχνότητα της κεντρικής τιμής.

Παράδειγμα 4

Το ύψος (σε cm) των σπουδαστών ενός τμήματος του ΤΕΙ φαίνεται στον παρακάτω πίνακα, όπου σε παρενθέσεις είναι η μικρότερη και μεγαλύτερη τιμή $x_{\min}, x_{\max}$.

170	180	178	165	170	168	175	175	173	162
160	170	167	177	180	170	182	178	165	178
(156)	175	172	173	167	187	170	180	178	(191)
176	169	167	166	179	178	180	164	170	173

Παρατηρούμε ότι το εύρος R του δείγματος είναι R = 191 – 156 = 35. Επειδή έχουμε n = 40 παρατηρήσεις, χρησιμοποιούμε κ = 6 κλάσεις. Το πλάτος των κλάσεων είναι: $c = \dfrac{R}{\kappa} = \dfrac{35}{6} = 5{,}83 \approx 6$ (με τον τύπο του *Sturges*: c = 5,54). Αν θεωρήσουμε ως αρχή της πρώτης κλάσης το 156 θα έχουμε τον παρακάτω πίνακα κατανομών συχνοτήτων (απόλυτων, σχετικών, αθροιστικών) για τα δεδομένα του παραπάνω πίνακα:

Κλάσεις [-)	*Κεντρικές τιμές x_i*	*Συχνότητα n_i*	*Σχετική συχνότητα f_i %*	*Αθρ. συχνότητα N_i*	*Αθρ. σχετ. συχνότ. F_i%*
156 - 162	159	2	5,0	2	5,0
162 - 168	165	8	20,0	10	25,0
168 - 174	171	12	30,0	22	55,0
174 - 180	177	11	27,5	33	82,5
180 - 186	183	5	12,5	38	95,0
186 - 192	189	2	5,0	40	100,0
Σύνολο	-	40	100	-	-

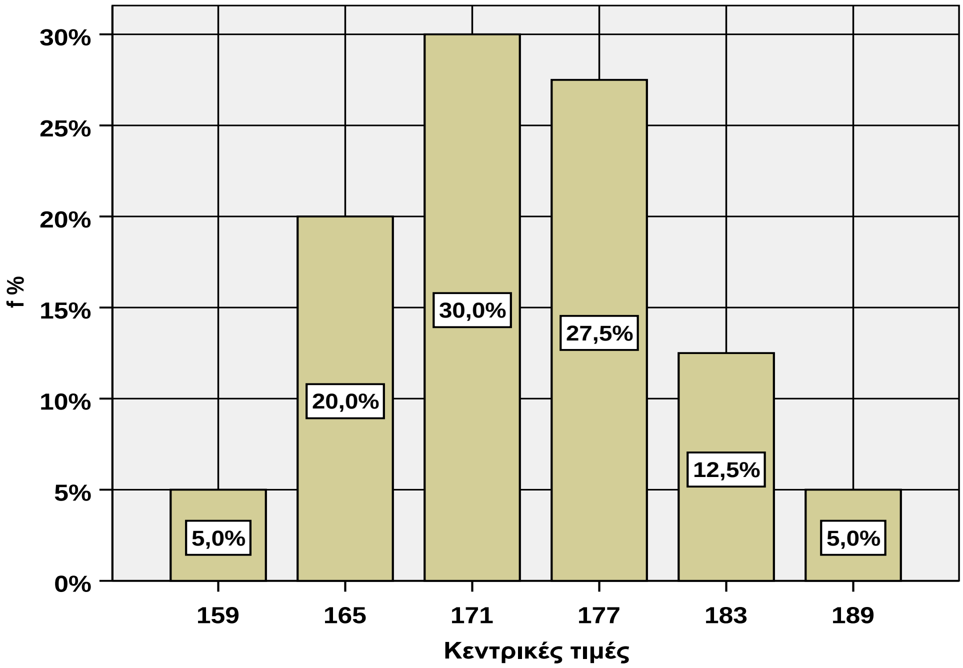
Κατά τον σχεδιασμό του πίνακα πρέπει να προσεχθούν τα εξής:

- Καμία παρατήρηση δεν μπορεί να μείνει έξω από κάποια κλάση.
- Οι κεντρικές τιμές διαφέρουν μεταξύ τους όσο και το πλάτος των κλάσεων, που εδώ είναι ίσο με 6.
- Μία παρατήρηση που συμπίπτει με το άνω άκρο μιας κλάσης θα τοποθετηθεί κατά τη διαλογή στην αμέσως επόμενη κλάση. Π.χ. ο σπουδαστής με ύψος 180 θα τοποθετηθεί στην πέμπτη κλάση [180 - 186).

§ 2. 3. Ιστογράμματα Συχνοτήτων

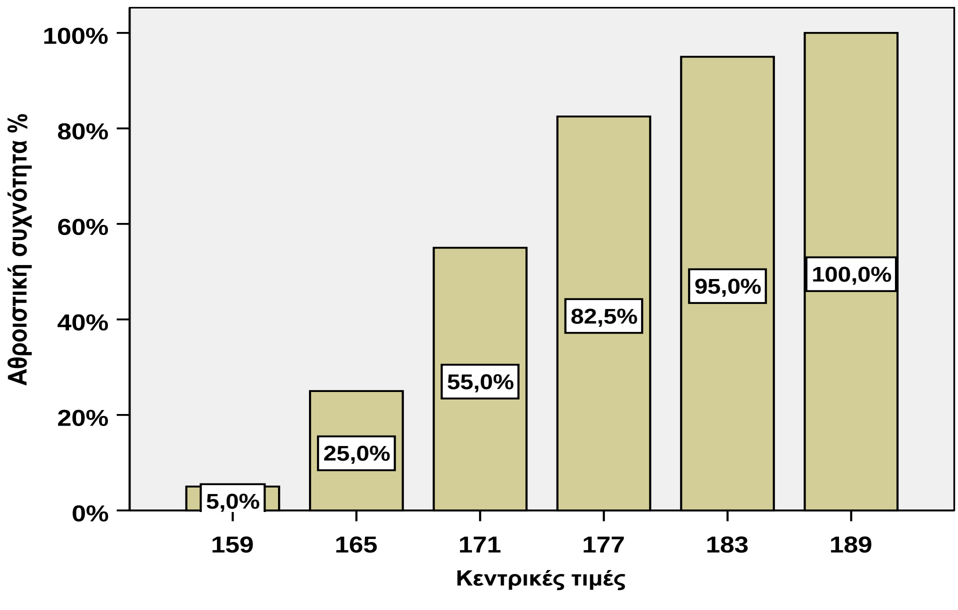
Η αντίστοιχη γραφική παράσταση ενός πίνακα συχνοτήτων με ομαδοποιημένα δεδομένα γίνεται με το λεγόμενο **ιστόγραμμα** συχνοτήτων. Στον οριζόντιο άξονα ενός συστήματος ορθογωνίων αξόνων σημειώνουμε με κατάλληλη κλίμακα τα όρια των κλάσεων. Στη συνέχεια κατασκευάζουμε διαδοχικά ορθογώνια (ιστούς), καθένα από τα οποία έχει βάση ίση με το πλάτος της κλάσης και ύψος τέτοιο ώστε το εμβαδόν του ορθογωνίου να ισούται με τη συχνότητα της κλάσης αυτής. Θεωρώντας το πλάτος c ως μονάδα μέτρησης του χαρακτηριστικού στον οριζόντιο άξονα, το ύψος κάθε ορθογωνίου είναι ίσο με τη συχνότητα της αντίστοιχης κλάσης. Επομένως, στον κατακόρυφο άξονα βάζουμε τις συχνότητες.

Ανάλογα κατασκευάζονται και τα ιστογράμματα αθροιστικών συχνοτήτων και αθροιστικών σχετικών συχνοτήτων.



Ιστόγραμμα σχετικών συχνοτήτων

(Τα ιστογράμματα έγιναν και διαμορφώθηκαν με τη βοήθεια του στατιστικού προγράμματος SPSS 14 – Statistical Package for the Social Sciences – Στατιστικό Πακέτο για τις Κοινωνικές Επιστήμες).



Ιστόγραμμα αθροιστικής συχνότητας

Το SPSS 14 και το Minitab 14 ή νεότερες εκδόσεις αυτών αποτελούν σήμερα πολύ δυνατά πακέτα για την επίλυση δύσκολων προβλημάτων της Στατιστικής.

§ 2. 4. Συμβολισμοί για το άθροισμα τιμών

Έστω X μια ποσοτική μεταβλητή. Αν θεωρήσουμε n τιμές της X, οι τιμές αυτές συμβολίζονται με $x_1, x_2, x_3, ..., x_n$ όπου οι αριθμοί 1, 2, 3, …, n βοηθούν στην ταυτοποίηση των τιμών. Οι τιμές μπορούν να εκφραστούν σύντομα με το σύνολο:

$\{x_i \,/\, i = 1, 2, ..., n\}$ όπου i είναι δείκτης που παίρνει τις τιμές από 1 έως n.

Το άθροισμα των n αυτών τιμών $x_1 + x_2 + x_3 + \ldots + x_n$ συμβολίζεται επίσης με: $\sum_{i=1}^{n} x_i$ όπου το σύμβολο Σ σημαίνει άθροισμα (διεθνές σύμβολο αθροίσματος), ενώ η αρχική και η τελική τιμή του δείκτη i γράφονται κάτω και πάνω από το Σ αντίστοιχα. Έτσι, αν έχουμε 5 τιμές, τις x_1, x_2, x_3, x_4, x_5, το άθροισμά τους συμβολίζεται με $\sum_{i=1}^{5} x_i$.

Κανόνες άθροισης

Αν c είναι μία σταθερά, τότε:

1) $\sum_{i=1}^{n} c_i = \underbrace{c + c + c + \ldots + c}_{n\ \phi o\rho \acute{\epsilon} \varsigma} = nc$, 2) $\sum_{i=1}^{n} cx_i = cx_1 + cx_2 + \ldots + cx_n = c\sum_{i=1}^{n} x_i$,

3) $\sum_{i=1}^{n} (x_i + y_i) = (x_1 + y_1) + (x_2 + y_2) + \ldots + (x_n + y_n) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$.

Βασικά συμπεράσματα

1) $\sum_{i=1}^{n} (x_i + c) = \sum_{i=1}^{n} x_i + nc$,

2) $\sum_{i=1}^{n} (x_i + c)^2 = \sum_{i=1}^{n} x_i^{\ 2} + 2c\sum_{i=1}^{n} x_i + nc^2$,

3) $\sum_{i=1}^{n} (x_i + y_i)^2 = \sum_{i=1}^{n} x_i^{\ 2} + 2\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^{\ 2}$,

4) $\sum_{i=1}^{n} \left(x_i - \frac{\sum_{i=1}^{n} x_i}{n} \right)^2 = \sum_{i=1}^{n} x_i^{\ 2} - \frac{\left(\sum_{i=1}^{n} x_i \right)^2}{n}$. Οι τρεις τελευταίες σχέσεις αποδεικνύονται βάσει της ταυτότητας: $(a \pm b)^2 = a^2 \pm 2ab + b^2$.

§ 2. 5. Μέτρα κεντρικής τάσης και διασποράς

Τα μέτρα αυτά είναι αριθμητικοί δείκτες οι οποίοι χρησιμοποιούνται για μια συνοπτική παρουσίαση των δεδομένων.

2. 5. 1. Μέτρα κεντρικής τάσης

Τα μέτρα αυτά δείχνουν το 'κέντρο' των δεδομένων δηλαδή την τιμή στην οποία τείνουν να συγκεντρωθούν τα δεδομένα, και είναι τα εξής:

α) Ο αριθμητικός μέσος όρος ή μέση τιμή

Έστω $x_1, x_2, x_3, \ldots, x_n$ οι τιμές που παίρνει η ποσοτική μεταβλητή X σε ένα δείγμα μεγέθους n. Ο αριθμητικός μέσος όρος των n αυτών τιμών ορίζεται ως εξής:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + x_3 + \ldots + x_n)$$

Χρησιμοποιώντας τα δεδομένα του παραδείγματος 1 (σελ. 4), ο μέσος όρος της βαθμολογίας των 10 μαθητών θα είναι:

$$\bar{x} = \frac{1}{10}(11+12+12+13+15+15+15+16+17+19) = \frac{145}{10} = 14{,}5\,.$$

Αν μερικές από τις τιμές εμφανίζονται περισσότερες από μία φορά, τότε ο μέσος όρος παίρνει την εξής ισοδύναμη μορφή:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{1}{n}(n_1 x_1 + n_2 x_2 + \ldots + n_k x_k)$$

Όπου n_i (i = 1, 2, … k) είναι η συχνότητα της τιμής x_i και k ο αριθμός των διαφορετικών τιμών x. Με τον παραπάνω τύπο θα είχαμε (όπου ας σημειωθεί ότι $n = \sum_{i=1}^{k} n_i$):

$$\bar{x} = \frac{1}{10}(11 + 2\cdot 12 + 13 + 3\cdot 15 + 16 + 17 + 19) = 14{,}5$$

Ο μέσος όρος των σωστών απαντήσεων των 700 ατόμων στο παράδειγμα 3 (σελ. 5) είναι σύμφωνα με τον δεύτερο τύπο ο εξής:

$$\bar{x} = \frac{1}{700}(50\cdot 0 + 70\cdot 1 + 80\cdot 2 + 180\cdot 3 + 200\cdot 4 + 80\cdot 5 + 40\cdot 6) = \frac{2210}{700} = 3{,}16\,.$$

Όταν τα δεδομένα μας είναι <u>ομαδοποιημένα σε κλάσεις</u>, για τον υπολογισμό της μέσης τιμής χρησιμοποιείται ο δεύτερος τύπος, όπου όμως η συχνότητα n_i αντικαθίσταται με τη συχνότητα f_i της κλάσης i και x_i είναι η κεντρική τιμή της κλάσης i.

Έτσι, ο μέσος όρος για το παράδειγμα 4 (σελ. 7) θα ισούται με

$$\bar{x} = \frac{1}{40}(2 \cdot 159 + 8 \cdot 165 + 12 \cdot 171 + 11 \cdot 177 + 5 \cdot 183 + 2 \cdot 189) = 173{,}25$$

Παρατήρηση

Αν αντί για κάποιο δείγμα χρησιμοποιήσουμε όλα τα στοιχεία N του πληθυσμού, στους παραπάνω τύπους αντικαθιστούμε το n με N και ο μέσος όρος δεν συμβολίζεται πια με $\bar{x}$ αλλά με μ και αποτελεί παράμετρο (βλέπε σελ. 2).

β) Η επικρατέστερη τιμή

Το μέτρο αυτό ισούται με την τιμή x που έχει την μεγαλύτερη συχνότητα. Σε αντίθεση με τη μέση τιμή, η επικρατέστερη τιμή μπορεί να μην ορίζεται με μοναδικό τρόπο. Στο παράδειγμα 1 η επικρατέστερη τιμή είναι η 15 γιατί εμφανίζεται 3 φορές, ενώ στο παράδειγμα 3 είναι η τιμή 4 γιατί έχει τη μεγαλύτερη συχνότητα (200).

Στην περίπτωση ομαδοποιημένων δεδομένων, επικρατέστερη κλάση είναι η κλάση με τη μεγαλύτερη συχνότητα. Έτσι, στο παράδειγμα 4 η επικρατέστερη κλάση είναι η [168-174) με συχνότητα 12.

γ) Η διάμεσος

Σε μια σειρά n τιμών, η διάμεσος δ ισούται (αφού πρώτα οι τιμές ταξινομηθούν κατά αύξουσα σειρά) με την τιμή που βρίσκεται στη θέση $\frac{n+1}{2}$ *όταν ο n είναι περιττός*, ή με το ημιάθροισμα των τιμών που βρίσκονται στις θέσεις $\frac{n}{2}$ και $\frac{n}{2}+1$ *όταν ο n είναι άρτιος*. Γενικά μπορούμε να πούμε ότι η διάμεσος ισούται με την τιμή που αφήνει αριστερά της και δεξιά της ίσο πλήθος τιμών, αφού πρώτα αυτές τοποθετηθούν κατά αύξουσα (ή φθίνουσα) σειρά.

Στο παράδειγμα 1 η διάμεσος δ ισούται με 15 γιατί το $n = 10$ είναι άρτιος οπότε $\frac{10}{2} = 5$ και $\frac{10}{2} + 1 = 6$. Στην 5^{η} θέση είναι ο αριθμός 15, όπως και στην 6^{η} θέση.

Οπότε, η διάμεσος ισούται με $\frac{15+15}{2}=15$. Στο παράδ. 3 η διάμεσος έχει την τιμή 3.

Στην περίπτωση <u>ομαδοποιημένων δεδομένων</u> η διάμεσος βρίσκεται με τη χρήση του εξής τύπου:

$$\text{Διάμεσος } \delta = x' + \frac{\Delta\left(\frac{n}{2} - N_{i-1}\right)}{n_i},$$

όπου x' είναι το κατώτερο ακριβές όριο[1] της κλάσης στην οποία βρίσκεται η διάμεσος, Δ είναι το 'φυσικό' πλάτος της κλάσης (βλέπε υποσημείωση), n_i η συχνότητα της i κλάσης, n ο συνολικός αριθμός των παρατηρήσεων (μέγεθος του δείγματος) και N_{i-1} η αθροιστική συχνότητα της προηγούμενης κλάσης.

Στο παράδειγμα 4 (σελ. 8), η διάμεσος βρίσκεται στην κλάση [168 – 174) (2η παρατήρηση αμέσως μετά). Για την κλάση αυτή έχουμε: x' = 167,5 ενώ Δ = 7, n = 40, n_3 = 12 (πάμε στην 3η κλάση εφόσον η φυσική διάμεσος είναι στην 20ή και 21η μέτρηση) και N_2 = 10 (αθροιστική συχνότητα της 2ης κλάσης). Συνεπώς:

$$\text{Διάμεσος } \delta = 167{,}5 + \frac{7(20-10)}{12} = 173{,}3.$$

Να σημειωθεί ότι αν ταξινομούσαμε στο παράδειγμα 4 τα ύψη κατά αύξουσα σειρά, (κάτι που για μεγάλο πλήθος είναι πολύ βαρετό), θα βρίσκαμε 173. Δηλαδή με τον παραπάνω τύπο είμαστε πολύ κοντά στην πραγματική τιμή της διαμέσου.

Παρατηρήσεις στα μέτρα κεντρικής τάσης

- Η διάμεσος δίνει λιγότερες πληροφορίες από το μέσο όρο καθότι δεν υπολογίζεται αλγεβρικά βάσει των μετρήσεων, αλλά βάσει της τάξης μεγέθους αυτών.

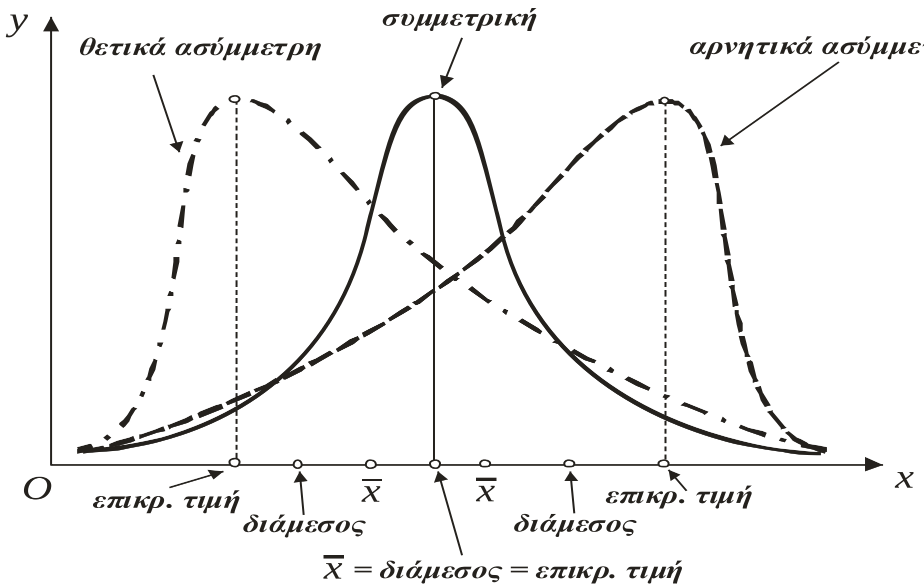
- Στην περίπτωση ομαδοποιημένων δεδομένων, η διάμεσος βρίσκεται στο διάστημα του οποίου η <u>αθροιστική συχνότητα είναι μεγαλύτερη ή ίση με n/2.</u>

---

[1] Σε μία κλάση [x,y), κατώτερο ακριβές όριο x' θεωρείται ο πρώτος αριθμός της κλάσης x, ελαττωμένος κατά μισή μονάδα του τελευταίου ψηφίου (δεκαδικού ή όχι), ενώ ανώτερο ακριβές όριο y' είναι ο δεύτερος αριθμός της κλάσης y αυξημένος κατά μισή μονάδα του τελευταίου ψηφίου (δεκαδικού ή όχι). Έτσι, το 'φυσικό' πλάτος της κλάσης προκύπτει ότι είναι η διαφορά x' - y'. Π.χ. για την κλάση [168 – 174) είναι x' = 167,5 και y' = 174,5 με πλάτος κλάσης Δ = 174,5 - 167,5 = 7, ενώ για την κλάση [5,24 – 6,38) είναι x' = 5,235 και y' = 6,385 με πλάτος κλάσης Δ = 6,385 - 5,235 = 1,15.

- Σε αντίθεση με το μέσο όρο, η διάμεσος δεν επηρεάζεται από την ύπαρξη ακραίων τιμών (ιδιαίτερα μεγάλων ή ιδιαίτερα μικρών). Χρησιμοποιείται ως μέτρο κεντρικής τάσης σε περίπτωση ασύμμετρων κατανομών.

- Η επικρατέστερη τιμή, όπως και η διάμεσος, δεν υπολογίζεται αλγεβρικά και δεν επηρεάζεται από ακραίες τιμές.



- Σε μια συμμετρική κατανομή, τα τρία μεγέθη: μέση τιμή, διάμεσος και επικρατέστερη τιμή είναι ίσα. Σε περίπτωση ασύμμετρων κατανομών τα μεγέθη αυτά δεν είναι ίσα, αλλά η διάμεσος παραμένει πάντοτε μεταξύ μέσης τιμής και επικρατέστερης τιμής, όπως φαίνεται στο παρακάτω συγκεντρωτικό σχήμα:

- Μια προσεγγιστική τέλος σχέση μεταξύ των τριών αυτών μέτρων σε περίπτωση όχι συμμετρικής κατανομής δίδεται από τον παρακάτω τύπο:

Επικρατέστερη τιμή $= \overline{x} - 3(\overline{x} -$ διάμεσος).

2. 5. 2. Μέτρα διασποράς

Τα μέτρα διασποράς (ή μεταβλητότητας) συμπληρώνουν τα μέτρα κεντρικής τάσης και αποτελούν δείκτες της διαφοροποίησης των διαφόρων μετρήσεων. Συνήθως μας δείχνουν το βαθμό απόκλισης των διαφόρων μετρήσεων από το ‘κέντρο’ τους . Τα σημαντικότερα μέτρα διασποράς είναι:

α) Το εύρος R

Το εύρος R ορίζεται ως η διαφορά μεταξύ της μεγαλύτερης και μικρότερης τιμής των δεδομένων μας: $R = x_{\max} - x_{\min}$

όπου $x_{\max}$ και $x_{\min}$ είναι η μεγαλύτερη και η μικρότερη τιμή αντίστοιχα. Έτσι, στο παράδειγμα 1 το εύρος ισούται με 19 – 11 = 8, ενώ στο παράδ. 3 ισούται με 6 – 0 = 6.

β) Η διακύμανση s²

Αποτελεί μέτρο της συνολικής απόκλισης (διασποράς) των διαφόρων τιμών από την μέση τιμή, και ορίζεται ως εξής:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2\right] \qquad (1)$$

όπου *n* είναι το μέγεθος του δείγματος.

Ένας ισοδύναμος αλγεβρικά όπως εύκολα αποδεικνύεται, αλλά πολύ πιο εύχρηστος από τον παραπάνω, είναι και ο τύπος:

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^{\,2} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right) = \frac{1}{n-1}\left(\left(x_1^{\,2} + x_2^{\,2} + \ldots + x_n^{\,2}\right) - \frac{\left(x_1 + x_2 + \ldots + x_n\right)^2}{n}\right) \qquad (2)$$

Έτσι, με τα δεδομένα του παραδείγματος 1 η διακύμανση είναι (τύπος 2):

$$s^2 = \frac{1}{9}\left(11^2 + 12^2 + 12^2 + 13^2 + \ldots + 19^2 - \frac{145^2}{10}\right) = \frac{1}{9}(2159 - 2102{,}5) = 6{,}28.$$

Όταν οι τιμές των δεδομένων μας συνοδεύονται από συχνότητες, η διακύμανση παίρνει την εξής μορφή:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2 = \frac{1}{n-1}\left(n_1(x_1 - \bar{x})^2 + n_2(x_2 - \bar{x})^2 + \ldots + n_k(x_k - \bar{x})^2\right) \quad (3)$$

όπου n_i είναι η συχνότητα της τιμής x_i και k ο αριθμός των <u>διαφορετικών τιμών</u> x.

Ας σημειωθεί ότι $n = \sum_{i=1}^{k} n_k = n_1 + n_2 + \ldots + n_k$.

Ισοδύναμος με τον παραπάνω τύπο (3) είναι και ο εξής:

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{k} n_i x_i^{\ 2} - \frac{\left(\sum_{i=1}^{k} n_i x_i\right)^2}{n} \right) = \frac{1}{n-1}\left(n_1 x_1^{\ 2} + n_2 x_2^{\ 2} + \ldots + n_k x_k^{\ 2} - \frac{\left(n_1 x_1 + n_2 x_2 + \ldots + n_k x_k\right)^2}{n} \right) \quad (4)$$

Συνεπώς, στο παράδειγμα 3 η διακύμανση είναι σύμφωνα με τον τύπο (4):

$$s^2 = \frac{1}{699}\left(50 \cdot 0^2 + 70 \cdot 1^2 + 80 \cdot 2^2 + 180 \cdot 3^2 + 200 \cdot 4^2 + 80 \cdot 5^2 + 40 \cdot 6^2 - \frac{2210^2}{700} \right) = 2{,}39$$

Σε περίπτωση ομαδοποιημένων δεδομένων σε κλάσεις, χρησιμοποιούμε τον τύπο (4), όπου το άνω όριο του αθροίσματος το *k* εκφράζει τον αριθμό των κλάσεων κ, το n_i εκφράζει την συχνότητα της *i* κλάσης και το x_i εκφράζει την κεντρική τιμή της *i* κλάσης. Στο παράδειγμα 4, η διακύμανση χρησιμοποιώντας τον τύπο (4) είναι:

$$s^2 = \frac{1}{39}\left(\begin{array}{l} 2 \cdot 159^2 + 8 \cdot 165^2 + 12 \cdot 171^2 + 11 \cdot 177^2 + 5 \cdot 183^2 + 2 \cdot 189^2 - \\ - \dfrac{\left(2 \cdot 159 + 8 \cdot 165 + 12 \cdot 171 + 11 \cdot 177 + 5 \cdot 183 + 2 \cdot 189\right)^2}{40} \end{array} \right) = 53{,}81$$

Παρατήρηση

Όταν αντί για κάποιο δείγμα μεγέθους *n* μελετούμε ολόκληρο τον πληθυσμό που απαρτίζεται από *N* στοιχεία, αντικαθιστούμε το *n* και το *n* – 1 με το *N* σε όλες τις εκφράσεις της διακύμανσης. Επίσης, η διακύμανση του πληθυσμού δεν συμβολίζεται πια με s^2 αλλά με σ^2 γιατί αποτελεί παράμετρο του πληθυσμού.

γ) Η τυπική απόκλιση

Η τυπική απόκλιση ισούται με τη θετική τετραγωνική ρίζα της διακύμανσης και συμβολίζεται με *s*. Έτσι, με την τυπική απόκλιση έχουμε ένα μέτρο διασποράς που εκφράζεται με την ίδια μονάδα μέτρησης. Σε περίπτωση που θεωρούμε όλο τον πληθυσμό, η τυπική απόκλιση αποτελεί παράμετρό του και συμβολίζεται με σ.

Στο παράδειγμα 1, η τυπική απόκλιση είναι $s = \sqrt{6{,}28} = 2{,}506$. Στο παράδειγμα 3 είναι $s = \sqrt{2{,}39} = 1{,}546$ ενώ στο παράδειγμα 4 είναι $s = \sqrt{53{,}81} = 7{,}336$.

δ) Συντελεστής μεταβλητότητας

Όταν θέλουμε να συγκρίνουμε δεδομένα με διαφορετική μονάδα μέτρησης, για να υπολογίσουμε τη διασπορά τους χρησιμοποιούμε τον συντελεστή μεταβλητότητας που είναι καθαρός αριθμός και ορίζεται ως εξής:

$$V = \frac{s}{\bar{x}} \quad ή \quad V = (\frac{s}{\bar{x}} \cdot 100)\%.$$

Είναι δηλαδή το πηλίκο της τυπικής απόκλισης με τη μέση τιμή. Ο δεύτερος τύπος εκφράζεται ως ποσοστό. Αν $\bar{x} < 0$, τότε αντί της $\bar{x}$ χρησιμοποιούμε την $|\bar{x}|$.

Έτσι, στο παράδειγμα 1 είναι $V = \dfrac{2{,}5}{14{,}5} \approx 0{,}172$ ή $17{,}2\%$, ενώ στο παράδειγμα 3

$$V = \frac{1{,}546}{3{,}16} \approx 0{,}489 \quad ή \quad 48{,}9\%.$$

Παρατηρήσεις στα μέτρα διασποράς:

- Το εύρος δίνει γενικά μια πρόχειρη εικόνα της διασποράς των τιμών και επηρεάζεται από ασυνήθιστα υψηλές ή χαμηλές τιμές.

- Η διακύμανση και η τυπική απόκλιση δίνουν με τη μεγαλύτερη ακρίβεια το μέγεθος της διασποράς. Η διακύμανση εκφράζεται με τη μονάδα μέτρησης των δεδομένων μας υψωμένη στο τετράγωνο. Έτσι, αν οι μετρήσεις μας αφορούν χρονικές περιόδους σε sec, η διακύμανση εκφράζεται σε sec^2. Αντίθετα, η τυπική απόκλιση λόγω του ότι είναι η τετραγωνική ρίζα της διακύμανσης, εκφράζεται με την ίδια μονάδα μέτρησης που χρησιμοποιούμε για τα δεδομένα μας (π.χ. σε sec) όπως εξάλλου και ο αριθμητικός μέσος.

- Ο συντελεστής μεταβλητότητας λόγω της κατασκευής του δεν έχει μονάδα μέτρησης (καθαρός αριθμός χωρίς διάσταση), γι' αυτό και χρησιμοποιείται για συγκρίσεις δεδομένων με διαφορετική μονάδα μέτρησης.

- Ας σημειωθεί τέλος, ότι, από τον ορισμό τους, οι τιμές που μπορούν να πάρουν η διακύμανση και η τυπική απόκλιση είναι πάντα θετικές, ενώ η ελάχιστη τιμή είναι το 0, που πραγματοποιείται όταν όλες οι τιμές που μελετούμε είναι ίσες.

Ασκήσεις

1) Η μέση ηλικία 18 αγοριών και 12 κοριτσιών μιας σχολικής τάξης είναι 15,4 χρόνια. Εάν η μέση ηλικία των αγοριών είναι 15,8 χρόνια, να βρεθεί η μέση ηλικία των κοριτσιών.

2) Ένα εργοστάσιο απασχολεί 5 εποχιακούς νέους στο τμήμα Α με μέσο μηνιαίο μισθό 249 ευρώ, 6 στο Β με μέσο μηνιαίο μισθό 280 ευρώ και 4 στο Γ με μέσο μηνιαίο μισθό 360 ευρώ. Ποιος είναι ο μέσος μισθός όλων των εργαζομένων;

3) Η μέση τιμή και η διάμεσος πέντε αριθμών είναι 6. Οι τρεις απ' αυτούς είναι οι 5, 8, 9. Να βρεθούν οι άλλοι δύο.

4) Η βαθμολογία 10 μαθητών σε ένα διαγώνισμα ήταν: 11, 10 13, 7, 15, 3, 11, 14, 12, 4. Να υπολογιστούν: α) η μέση τιμή, η επικρατέστερη τιμή και η διάμεσος και β) το εύρος, η τυπική απόκλιση και ο συντελεστής μεταβλητότητας.

5) Να δειχθεί ότι ο ορισμός της διακύμανσης $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ είναι ισοδύναμος αλγεβρικά με τους: $\dfrac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^{\,2} - n\bar{x}^2\right\}$ και $\dfrac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2 \Big/ n\right\}$.

6) Στο διπλανό πίνακα δίνεται η κατανομή της ηλικίας των ατόμων μιας πόλης. Να υπολογιστούν η τυπική απόκλιση και ο συντελεστής μεταβλητότητας. (Να συμπληρωθεί ο πίνακας και με άλλα απαραίτητα στοιχεία για τον ζητούμενο υπολογισμό).

Ηλικία (σε έτη)	*Συχνότητα (σε χιλιάδες)*
0 – 20	12
20 – 40	14
40 – 60	20
60 – 80	10
80 - 100	4

7) Δίνονται οι πρόσφατοι λογαριασμοί της ΔΕΗ 9 νοικοκυριών σε ευρώ: 120, 140, 160, 160, 160, 180, 200, 200, 300. Να υπολογιστούν:

α) το εύρος, η μέση τιμή, η διάμεσος, η διακύμανση, η τυπική απόκλιση και ο συντελεστής μεταβλητότητας.

β) Αν όλοι οι λογαριασμοί αυξηθούν κατά 20% και επιβληθεί φόρος 30 ευρώ, να βρεθούν τα νέα μέτρα διασποράς των νέων λογαριασμών. Τι παρατηρείτε;

Κεφάλαιο 3.

Εισαγωγή στη θεωρία πιθανοτήτων. Κατανομές πιθανοτήτων

§ 3. 1. Βασικές έννοιες στη θεωρία πιθανοτήτων

Ένα πείραμα (ή μια διαδικασία) θα λέγεται τυχαίο, όταν πραγματοποιούμενο κάτω από τις ίδιες πρακτικά συνθήκες, δίνει αποτελέσματα τα οποία δεν μπορούμε να προβλέψουμε με βεβαιότητα. Η έλλειψη της βεβαιότητας για το τι θα συμβεί, οδηγεί στην έννοια της ‘τυχαιότητας’ που αναφέρεται στα αποτελέσματα και όχι στον τρόπο εκτέλεσης του πειράματος. Έτσι για παράδειγμα, δεν μπορούμε να προβλέψουμε από πριν το φύλο του παιδιού σε μια γέννα. Γνωρίζουμε όμως ότι θα είναι αγόρι ή κορίτσι. Ένα πείραμα τύχης συνεπώς, καθορίζεται πλήρως από τα αποτελέσματα που μπορεί να εμφανιστούν μετά την εκτέλεσή του.

Ορισμός 1: Το σύνολο των δυνατών αποτελεσμάτων ενός πειράματος τύχης λέγεται δειγματοχώρος ή **δειγματικός χώρος** και συμβολίζεται με *S*. Κάθε στοιχείο του δειγματοχώρου λέγεται δεγματοσημείο.

Ορισμός 2: **Ενδεχόμενο** ή γεγονός ονομάζεται κάθε υποσύνολο αποτελεσμάτων του δειγματικού χώρου ενός πειράματος τύχης.

Παραδείγματα

1) Για το τυχαίο πείραμα της ρίψης ενός νομίσματος ο δειγματοχώρος είναι: $S=\{\text{Κ}, \text{Γ}\}$, όπου Κ κεφάλι και Γ γράμμα είναι τα δειγματοσημεία του *S*.

2) Στη ρίψη ενός ζαριού ο δειγματοχώρος είναι: $S=\{1, 2, 3, 4, 5, 6\}$.
Το υποσύνολο Α = {άρτια πλευρά} = {2, 4, 6} είναι ένα ενδεχόμενο ή γεγονός.

3) Στο πείραμα τύχης ‘γέννηση παιδιού’ είναι $S=\{\text{Α}, \text{Κ}\}$, (Α αγόρι Κ κορίτσι).

4) Αν ριφθούν α) δύο και β) τρία νομίσματα μαζί, τότε οι δειγματικοί χώροι είναι αντίστοιχα:

$S_2=\{\text{ΚΚ}, \text{ΚΓ}, \text{ΓΚ}, \text{ΓΓ}\}$ και

$S_3=\{\text{ΚΚΚ}, \text{ΚΚΓ}, \text{ΚΓΚ}, \text{ΓΚΚ}, \text{ΚΓΓ}, \text{ΓΚΓ}, \text{ΓΓΚ}, \text{ΓΓΓ}\}$.

§ 3. 2. Στοιχεία θεωρίας συνόλων

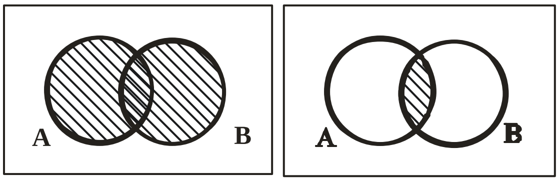
Στη θεωρία πιθανοτήτων χρησιμοποιούνται έννοιες από τη θεωρία συνόλων γι' αυτό και αναφέρονται οι σπουδαιότερες απ' αυτές.

Με τον όρο σύνολο εννοούμε μια συλλογή από διακεκριμένα αντικείμενα καλώς ορισμένα, όπως π.χ. το σύνολο των φοιτητών μιας τάξης.

Το σύνολο A λέγεται υποσύνολο του B όταν τα στοιχεία του A είναι και στοιχεία του B. Κενό σύνολο (συμβολισμός Ø) είναι αυτό που δεν έχει στοιχεία.

Ένωση συνόλων A και B (συμβολισμός $A \cup B$) λέγεται το σύνολο των στοιχείων που ανήκουν στο A ή στο B, δηλαδή τουλάχιστον σε ένα από τα δύο σύνολα.

Τομή συνόλων A και B (συμβολισμός $A \cap B$) λέγεται το σύνολο των στοιχείων που ανήκουν συγχρόνως και στο A και στο B.



AUB AΠB

Στα παραπάνω σχήματα (διαγράμματα Venn) η γραμμοσκιασμένη επιφάνεια δείχνει την ένωση και την τομή των συνόλων A και B αντίστοιχα, όπου το πλαίσιο είναι ο S.

§ 3. 3. Στοιχεία συνδυαστικής ανάλυσης

Αν ένα γεγονός μπορεί να συμβεί κατά κ τρόπους και αν για καθένα από αυτούς ένα άλλο γεγονός μπορεί να συμβεί κατά λ τρόπους, τότε και τα δύο γεγονότα μπορούν να συμβούν κατά κ·λ τρόπους. Έτσι, αν ρίξουμε ένα ζάρι, υπάρχουν 6 δυνατά αποτελέσματα. Αν ρίξουμε ένα νόμισμα, υπάρχουν 2 δυνατά αποτελέσματα. Αν ρίξουμε το ζάρι και το νόμισμα μαζί, τότε θα υπάρχουν 6·2 = 12 δυνατά αποτελέσματα (βασική αρχή της απαρίθμησης).

Μεταθέσεις

Μεταθέσεις των ν στοιχείων ενός συνόλου, λέγονται όλοι οι διαφορετικοί τρόποι με τους οποίους μπορούμε να βάλουμε σε μια σειρά τα ν αυτά στοιχεία.

Αποδεικνύεται ότι το πλήθος αυτών των μεταθέσεων είναι ν! (ονομάζεται ν παραγοντικό) και συμβολίζει το γινόμενο $1 \cdot 2 \cdot 3 \cdots ν$. Π.χ. $5! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 120$.

Διατάξεις

Διατάξεις των ν στοιχείων ανά κ (όπου κ$\leq$ν), λέγονται όλοι οι διαφορετικοί τρόποι με τους οποίους μπορούμε, από τα ν στοιχεία ενός συνόλου, να πάρουμε κ και να τα βάλουμε σε μια σειρά.

Αποδεικνύεται ότι το πλήθος των διατάξεων των ν στοιχείων ανά κ είναι:

$$\Delta_{\kappa}^{\nu} = \frac{\nu!}{(\nu-\kappa)!}.$$

Πρέπει να σημειωθεί ότι 0! = 1 (εξ ορισμού για να έχει νόημα ο παρονομαστής).

Παράδειγμα: Με πόσους τρόπους μπορεί να προκύψει η τριάδα νικητών από 8 αθλητές που παίρνουν μέρος σε αγώνα δρόμου 200 μ.;

Είναι διατάξεις 8 αθλητών ανά 3. Δηλ. $\frac{8!}{(8-3)!} = \frac{8!}{5!} = 6\cdot 7\cdot 8 = 336$ τρόπους.

Διατάξεις με επανάληψη

Στις *διατάξεις με επανάληψη*, επιτρέπεται η επανάληψη στοιχείων από τα ν ενός συνόλου μέχρι να συμπληρωθεί το πλήθος κ στοιχείων.

Αποδεικνύεται ότι το πλήθος των διατάξεων με επανάληψη είναι: $E_{\kappa}^{\nu} = \nu^{\kappa}$.

Προφανώς στις διατάξεις αυτές μπορεί να είναι κ$\geq$ν.

Παραδείγματα

1. Πόσους τετραψήφιους μπορούμε να δημιουργήσουμε με τα ψηφία 1,2, … ,9;

Για το ψηφίο των χιλιάδων έχουμε στη διάθεσή μας 9 ψηφία. Ομοίως και για τα ψηφία των εκατοντάδων, δεκάδων και μονάδων. Άρα σύμφωνα με την αρχή της απαρίθμησης θα είναι: $E_4^9 = 9\cdot 9\cdot 9\cdot 9 = 9^4 = 6561$ τετραψήφιοι.

2. Πόσες είναι οι διαφορετικές στήλες που πρέπει να συμπληρώσουμε στο ΠΡΟ-ΠΟ ώστε να είμαστε σίγουροι ότι θα πιάσουμε 13άρι;

Για να συμπληρώσουμε μια στήλη ΠΡΟ-ΠΟ, χρησιμοποιούμε τα στοιχεία 1, 2, Χ (ν = 3) και το καθένα μπορεί να χρησιμοποιηθεί μέχρι 13 φορές (κ = 13). Πρόκειται δηλ. για διατάξεις με επανάληψη των 3 στοιχείων ανά 13. Άρα:

$E_{13}^3 = 3^{13} = 1594323$ στήλες πρέπει να συμπληρωθούν για σίγουρο 13άρι.

Συνδυασμοί

Ονομάζουμε <u>συνδυασμό</u> των ν αντικειμένων ανά μ ($\mu \leq \nu$), κάθε ομάδα από μ αντικείμενα παρμένα από τα ν, <u>χωρίς να παίζει ρόλο η σειρά τους.</u> Το πλήθος των συνδυασμών των ν αντικειμένων ανά μ, αποδεικνύεται ότι ισούται με:

$$\binom{\nu}{\mu} = \frac{\nu!}{\mu!(\nu-\mu)!}.$$

Παράδειγμα: Από το σύνολο 10 ερωτήσεων καλούνται οι σπουδαστές να απαντήσουν σε 6. Με πόσους τρόπους ένας σπουδαστής μπορεί να επιλέξει τις ερωτήσεις που θα απαντήσει;

Προφανώς δεν έχει σημασία η σειρά με την οποία θα απαντηθούν οι ερωτήσεις. Άρα πρόκειται για συνδυασμούς των 10 ερωτήσεων ανά 6. Έτσι έχουμε:

$$\binom{10}{6} = \frac{10!}{6!\cdot(10-6)!} = 210.$$

§ 3. 4. Ορισμοί πιθανότητας και βασικά θεωρήματα

1ος ορισμός Πιθανότητας γεγονότος

Αν ένα πείραμα έχει πεπερασμένο δειγματικό χώρο *S*, και A είναι ένα γεγονός, τότε ορίζουμε ως πιθανότητα του γεγονότος A, αφού υποτεθεί ότι όλα τα αποτελέσματα είναι εξ' ίσου δυνατά, το πηλίκο της διαίρεσης του πλήθους των αποτελεσμάτων του γεγονότος A δια του πλήθους των αποτελεσμάτων του δειγματικού χώρου *S*.

$$P(A) = \frac{\text{πληθυκός αριθμός του } A}{\text{πληθικός αριθμός του } S}.$$

Ο ορισμός αυτός προϋποθέτει τη γνώση όλων των δυνατών αποτελεσμάτων τα οποία επιπλέον θεωρεί εξ' ίσου δυνατά.

Παραδείγματα

1) Αν ρίξουμε ένα νόμισμα, η πιθανότητα του γεγονότος A = {κεφάλι} είναι:

$P(A) = \frac{1}{2}$ (ο πληθυκός αριθμός του δειγματικού χώρου *S* είναι 2, δηλ. *S* = {K, Γ}).

2) Στο πείραμα ρίψης ενός ζαριού, η πιθανότητα του γεγονότος Α = {πλευρά 3} είναι $P(A) = \frac{1}{6}$, ενώ η πιθανότητα του γεγονότος Β = {άρτια πλευρά} = {2, 4, 6} είναι $P(B) = \frac{3}{6} = \frac{1}{2}$, (ο πληθυκός αριθμός δειγματικού χώρου *S* = {1, 2, 3, 4, 5, 6}).

2^{ος} ορισμός Πιθανότητας γεγονότος

Ο ορισμός αυτός βασίζεται στην επανάληψη ενός πειράματος τύχης με ένα μεγάλο αριθμό φορών. Ως σχετική συχνότητα εμφάνισης ενός γεγονότος Α ορίζουμε το λόγο (πηλίκο) του πλήθους *k* των εμφανίσεων του γεγονότος Α δια του πλήθους *n* των επαναλήψεων του πειράματος. Οπότε, η πιθανότητα του Α ορίζεται ως το όριο της σχετικής συχνότητας όταν το *n* είναι πάρα πολύ μεγάλο, δηλαδή θεωρητικά όταν το *n* τείνει στο άπειρο.

$$P(A) = \lim_{n \to \infty} \frac{k}{n}.$$

Παράδειγμα

Υποθέτουμε ότι ρίχνουμε ένα ζάρι 300 φορές και κατασκευάζουμε τον ακόλουθο πίνακα:

Αποτέλεσμα	Συχνότητα	Σχετική συχνότητα
1	51	51/300 = 0,170
2	54	54/300 = 0,180
3	48	48/300 = 0,160
4	51	51/300 = 0,170
5	49	49/300 = 0,163
6	47	47/300 = 0,157
		Σύνολο = 1

Η πιθανότητα του γεγονότος Α = {πλευρά 3} είναι: P(A) = 48/300 = 0,160. Ας σημειωθεί ότι η τιμή αυτή δεν ισούται με την τιμή 1/6 = 0,16666 που δίνει ο 1^{ος} ορισμός διότι το *n* είναι μόνο 300.

Για τα διάφορα γεγονότα ενός πειράματος τύχης ισχύουν τα εξής:

Αν Α είναι γεγονός τότε $0 \le P(A) \le 1$. Ειδικά για τα σύνολα Ø και *S* είναι:

$P(\varnothing) = 0$ και $P(S) = 1$. Το κενό σύνολο Ø είναι το αδύνατο γεγονός (δεν μπορεί να συμβεί), ενώ το σύνολο *S* είναι το βέβαιο γεγονός (συμβαίνει πάντοτε).

Συμπληρωματικά – Ασυμβίβαστα γεγονότα

Τα γεγονότα Α και Β για τα οποία $A \cap B = \varnothing$ ονομάζονται συμπληρωματικά, αν $A \cup B = S$. Αν Α είναι ένα γεγονός, τότε το συμπληρωματικό του συμβολίζεται με A^c. Τα γεγονότα Α και Β λέγονται ασυμβίβαστα, όταν $A \cap B = \varnothing$. Συνεπώς δύο γεγονότα λέγονται ασυμβίβαστα, όταν η πραγματοποίηση του ενός αποκλείει την πραγματοποίηση του άλλου. Ο τελευταίος ορισμός ισχύει και για περισσότερα από δύο γεγονότα.

Ισχύουν οι παρακάτω δύο προτάσεις για τα συμπληρωματικά και ασυμβίβαστα γεγονότα:

α) $P(A^c) = 1 - P(A)$ για τυχαίο σύνολο Α και

β) $P(A \text{ ή } B) = P(A \cup B) = P(A) + P(B)$ όταν τα Α, Β είναι ασυμβίβαστα..

Η πρόταση αυτή ισχύει και για περισσότερα από δύο ασυμβίβαστα γεγονότα.

Προσθετικό θεώρημα

Στη γενική περίπτωση όπου τα γεγονότα Α, Β δεν είναι ασυμβίβαστα, δηλαδή $A \cap B \neq \varnothing$ ισχύει η σχέση:

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

Τα μη ασυμβίβαστα γεγονότα μπορούν να πραγματοποιηθούν ταυτόχρονα.

Δεσμευμένη πιθανότητα (θεώρημα Bayes)

Έστω ένα πείραμα τύχης με δειγματικό χώρο *S* και έστω ότι, λόγω κάποιας επιπρόσθετης πληροφορίας, γνωρίζουμε ότι μόνο ένα γνήσιο υποσύνολο Β του χώρου *S* ($B \subset S$) αποτελεί το σύνολο των δυνατών αποτελεσμάτων του πειράματος. Έχουμε συνεπώς έναν περιορισμό του δειγματικού χώρου *S* στο γεγονός Β, οπότε το τελευταίο θα είναι βέβαιο γεγονός. Αν συμβολίσουμε με $P(A/B)$ την πιθανότητα να συμβεί το γεγονός Α όταν γνωρίζουμε ότι πραγματοποιήθηκε το γεγονός Β, τότε ορίζουμε τη δεσμευμένη πιθανότητα ή πιθανότητα υπό συνθήκη $P(A/B)$ ως εξής:

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \quad \text{όπου } P(B) \neq 0.$$

Συνεπώς για δύο μη ασυμβίβαστα γεγονότα Α, Β ισχύει: $P(A \cap B) = P(B) \cdot P(A / B)$.

Έτσι, αν η πιθανότητα να περάσει ένας φοιτητής το μάθημα Β είναι 0,7 και η πιθανότητα να περάσει το Α αφού πέρασε το Β είναι 0,8 τότε η πιθανότητα να περάσει και τα δύο μαθήματα είναι:

$$P(A\ \kappa\alpha\iota\ B) = P(A \cap B) = P(B) \cdot P(A / B) = 0{,}7 \cdot 0{,}8 = 0{,}56\,.$$

Γενικό παράδειγμα

Ένα σύνολο 5000 μαθητών ταξινομήθηκε ως προς δύο χαρακτηριστικά: το φύλο και την επίδοσή του σε κάποιο τεστ. Τα αποτελέσματα δίνονται στον παρακάτω πίνακα αποτελεσμάτων:

	Μέτρια επίδοση	Καλή επίδοση	Σύνολο
Αγόρια	2000	1000	3000
Κορίτσια	1500	500	2000
Σύνολο	3500	1500	5000

α) Να βρεθεί η πιθανότητα P(αγόρι με μέτρια επίδ. ή κορίτσι με καλή επίδ.).

β) Να βρεθεί η πιθανότητα P(αγόρι ή άτομο με καλή επίδοση).

γ) Ποια είναι η πιθανότητα, κάποιο άτομο να έχει καλή επίδοση, όταν γνωρίζουμε ότι είναι κορίτσι.

Λύση

α) Έστω τα δύο γεγονότα Α = {αγόρι με μέτρια επίδοση} και Β = {κορίτσι με καλή επίδοση}. Τα Α και Β είναι ασυμβίβαστα. Άρα:

$$PA \cup B) = P(A) + P(B) = \frac{2000}{5000} + \frac{500}{5000} = \frac{2500}{5000} = 0{,}5\,.$$

β) Έστω τα γεγονότα Α = {αγόρι} και Β = {άτομο με καλή επίδοση}. Τα δύο αυτά γεγονότα δεν είναι ασυμβίβαστα γιατί $A \cap B \neq$ Ø. Συνεπώς:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{3000}{5000} + \frac{1500}{5000} - \frac{1000}{5000} = \frac{3500}{5000} = 0{,}7\,.$$

γ) Θεωρούμε τα γεγονότα Α = {άτομο με καλή επίδοση}και Β = {κορίτσι}. Θέλουμε να βρούμε την πιθανότητα $P(A / B)$ δηλαδή την υπό συνθήκη πιθανότητα του Α γνωρίζοντας το Β. Θα είναι:

$$P(A/B)=\frac{P(A\cap B)}{P(B)}=\frac{500/5000}{2000/5000}=0{,}25$$

Ορισμός

Τα γεγονότα Α και Β ονομάζονται ανεξάρτητα όταν η πραγματοποίηση του ενός δεν εξαρτάται από την πραγματοποίηση ή όχι πραγματοποίηση του άλλου. Για τέτοια γεγονότα ισχύει:

$$P(A\ \kappa\alpha\iota\ B)=P(A\cap B)=P(A)\cdot P(B)$$

όπου $P(A\ \kappa\alpha\iota\ B)$ είναι η πιθανότητα της ταυτόχρονης πραγματοποίησης των δύο γεγονότων. Ο ορισμός και η παραπάνω σχέση ισχύουν και για περισσότερα από δύο γεγονότα δηλαδή: $P(\mathrm{A}\cap\mathrm{B}\cap\Gamma)=P(\mathrm{A})\cdot P(\mathrm{B})\cdot P(\Gamma)$.

Έτσι, αν η πιθανότητα να βρούμε άτομο με ομάδα αίματος Β είναι 0,14 η δε πιθανότητα για Rhesus αρνητικό είναι 0,15, η πιθανότητα κάποιο άτομο να έχει συγχρόνως και τα δύο αυτά χαρακτηριστικά είναι:

$$P(A\ \kappa\alpha\iota\ B)=P(A)\cdot P(B)=0{,}14\cdot 0{,}15=0{,}021.$$

Η παραπάνω σχέση ισχύει γιατί η ομάδα αίματος είναι ανεξάρτητη από τον παράγοντα Rhesus.

§ 3. 5. Τυχαίες μεταβλητές και κατανομές πιθανοτήτων

Θεωρούμε τις οικογένειες με τρία παιδιά και παρατηρούμε το φύλο τους (από το μικρότερο παιδί στο μεγαλύτερο). Ο δειγματικός χώρος θα είναι ο εξής :

$$S=\{KKK,\ AKK,\ KAK,\ KKA,\ AAK,\ AKA,\ KAA,\ AAA\}$$

γιατί αυτό είναι το σύνολο όλων των δυνατών περιπτώσεων τριών παιδιών, όπου Α αγόρι και Κ κορίτσι.

Έστω X ο αριθμός των αγοριών στις παραπάνω οικογένειες. Οι δυνατές τιμές που μπορεί να πάρει το X είναι 0, 1, 2, 3. Δηλαδή

$$X(KKK)=0$$

$$X(AKK)=X(KAK)=X(KKA)=1$$

$$X(AAK)=X(AKA)=X(KAA)=2$$

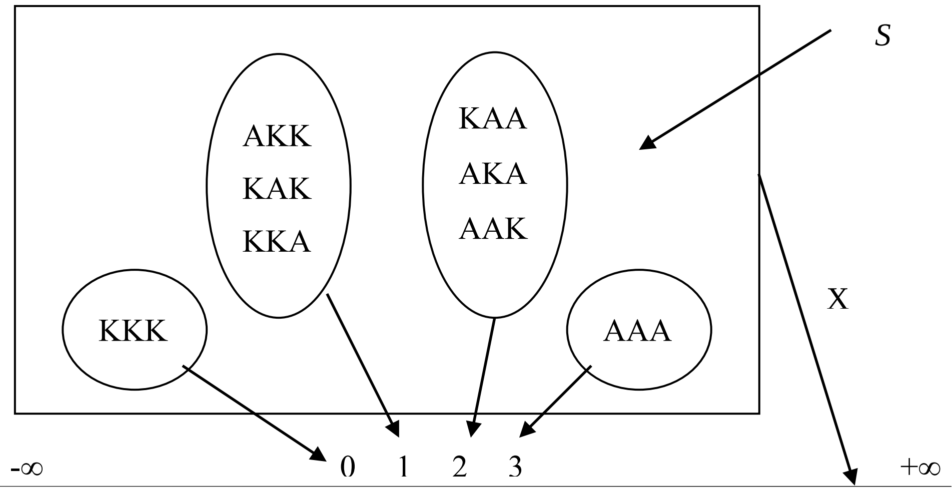
$$X(AAA)=3$$

Με την παραπάνω διαδικασία καθορίσαμε έναν κανόνα έτσι ώστε σε κάθε στοιχείο του δειγματοχώρου S να αντιστοιχεί ένας αριθμός του συνόλου {0, 1, 2, 3} που είναι υποσύνολο των πραγματικών αριθμών.

Αν σε ένα πείραμα τύχης μετρούμε μήκος ή χρόνο ή επίδοση ή δείκτη νοημοσύνης, ο δειγματοχώρος *S* είναι το σύνολο όλων των δυνατών αποτελεσμάτων της μέτρησης που κάναμε. Και στην περίπτωση αυτή μπορούμε να αντιστοιχίσουμε σε κάθε στοιχείο s του *S* το ίδιο το s. Από τα παραπάνω συνάγουμε τον ορισμό μιας τυχαίας μεταβλητής.

Ορισμός: Η τυχαία μεταβλητή είναι μία συνάρτηση με πεδίο ορισμού το δειγματοχώρο S και πεδίο τιμών τους πραγματικούς αριθμούς. Είναι δηλαδή ο κανόνας που αντιστοιχεί σε κάθε στοιχείο του δειγματοχώρου S έναν πραγματικό αριθμό.

Στο παράδειγμα που προαναφέραμε, η τυχαία μεταβλητή Χ παίρνει τις τιμές 0, 1, 2, 3 και το παρακάτω σχήμα δείχνει την αντιστοιχία των υποσυνόλων του *S* με το σύνολο Χ των πραγματικών αριθμών.



Η τυχαία μεταβλητή Χ λέγεται *διακριτή* όταν το σύνολο των τιμών της είναι διακριτό (όπως στο παράδειγμά μας) και *συνεχής* όταν μπορεί να πάρει οποιαδήποτε τιμή σε κάποιο διάστημα τιμών, όπως π.χ. η βαθμολογία, η θερμοκρασία, ο χρόνος κ.τ.λ.

Ας υπολογίσουμε τώρα στο παράδειγμά μας τις πιθανότητες με τις οποίες η τυχαία μεταβλητή Χ παίρνει τις τιμές 0, 1, 2, 3.

Σύμφωνα με τον 1ο ορισμό πιθανότητας θα είναι:

$$P(X=0)=\frac{1}{8},\quad P(X=1)=\frac{3}{8},\quad P(X=2)=\frac{3}{8},\quad P(X=3)=\frac{1}{8}.$$

Παρατηρούμε ότι 1/8 + 3/8 + 3/8 + 1/8 = 1. Γενικά, αν η τυχαία μεταβλητή Χ παίρνει τις τιμές $x_1, x_2, ..., x_n$, τότε θα είναι πάντοτε:

$$P(X=x_1)+P(X=x_2)+\ldots+P(X=x_n)=1$$

Ορισμός

Η συνάρτηση $f(x) = P(X = x)$ ονομάζεται **συνάρτηση πιθανότητας** της διακριτής τυχαίας μεταβλητής X, αν για κάθε *x* ισχύει:

α) $f(x) \geq 0$ και ειδικότερα $0 \leq f(x) \leq 1$, β) $\sum_x f(x) = 1$.

Το σύνολο των τιμών της X και των αντίστοιχων πιθανοτήτων $P(X = x)$ συνιστούν μια κατανομή πιθανότητας ή αλλιώς μια διακριτή θεωρητική κατανομή. Οι θεωρητικές κατανομές αποτελούν μια ιδεαλιστική μορφή των κατανομών σχετικών συχνοτήτων που είδαμε στην Περιγραφική Στατιστική και που αναφέρονταν στα δεδομένα κάποιου δείγματος.

Ορισμός

Η **αθροιστική συνάρτηση κατανομής** $F(x)$ μιας τυχαίας μεταβλητής με συνάρτηση πιθανότητας $f(x)$ ορίζεται ως εξής: $F(x) = P(X \leq x)$.

Παρατήρηση

Όταν η τυχαία μεταβλητή είναι συνεχής, αποδεικνύεται ότι $P(X = x) = 0$ και ισχύουν ανάλογοι ορισμοί για τη συνάρτηση πιθανότητας και την αθροιστική συνάρτηση κατανομής (με ολοκληρώματα) που δεν θα τις μελετήσουμε. Να σημειωθεί μόνο, ότι στις συνεχείς τυχαίες μεταβλητές, λόγω της $P(X = x) = 0$ ενδιαφερόμαστε πάντα για πιθανότητες διαστημάτων τιμών και όχι συγκεκριμένων τιμών *x*.

Παράδειγμα

Είχαμε υπολογίσει προηγουμένως για κάθε τιμή της X, δηλαδή για τις τιμές 0, 1, 2, 3 τις αντίστοιχες πιθανότητες, ορίζοντας έτσι την κατανομή πιθανότητας της X. Ας υπολογίσουμε τώρα για κάθε *x* = 0, 1, 2, 3 τις τιμές της αθροιστικής συνάρτησης κατανομής. Θα είναι:

$$F(0) = P(X \leq 0) = f(0) = 1/8$$

$$F(1) = P(X \leq 1) = f(0) + f(1) = 4/8$$

$$F(2) = P(X \leq 2) = f(0) + f(1) + f(2) = 7/8$$

$$F(3) = P(X \leq 3) = f(0) + f(1) + f(2) + f(3) = 8/8 = 1$$

Από το παράδειγμα αυτό φαίνεται η ομοιότητα της $F(x)$ με την κατανομή αθροιστικών σχετικών συχνοτήτων που είδαμε στην Περιγραφική Στατιστική.

Θα μελετήσουμε στη συνέχεια μερικές (απ' τις πολλές) χρήσιμες κατανομές.

§ 3. 6. Διωνυμική κατανομή B(n, p) (Bernoulli)

Ένα τυχαίο πείραμα (ή τυχαία διαδικασία) ονομάζεται διωνυμικό όταν ισχύουν οι εξής συνθήκες:

1) Το πείραμα αποτελείται από n επαναλήψεις.

2) Κάθε επανάληψη του πειράματος τύχης έχει δύο δυνατά αποτελέσματα που θα ονομάζουμε 'επιτυχία' και 'αποτυχία'.

3) Η πιθανότητα 'επιτυχίας' που θα συμβολίζουμε με p παραμένει σταθερή σε όλες τις επαναλήψεις.

4) Οι επαναλήψεις είναι ανεξάρτητες μεταξύ τους ως προς το αποτέλεσμά τους.

Ο αριθμός των 'επιτυχιών' ενός διωνυμικού πειράματος στις n επαναλήψεις, ορίζει μια διωνυμική τυχαία μεταβλητή η οποία έχει ως δυνατές τιμές τις $x = 0, 1, 2, 3, \ldots, \nu$. Έτσι, αν επαναλάβουμε 10 φορές την ρίψη ενός πειράματος και θεωρήσουμε ως 'επιτυχία' να έρθει η πλευρά 'κεφάλι', η διωνυμική τυχαία μεταβλητή έχει ως δυνατές τιμές τις $x = 0, 1, 2, 3, 4, \ldots, 10$ γιατί στις 10 φορές μπορεί καμία φορά να μην έρθει η πλευρά 'κεφάλι', ή να έρθει μόνο μία ή μόνο δύο κ.τ.λ. ή τέλος να έρθει και τις 10 φορές 'κεφάλι'.

Η διωνυμική τυχαία μεταβλητή Χ ακολουθεί τη *διωνυμική κατανομή* $B(n, p)$ με συνάρτηση πιθανότητας:

$$f(x) = P(X = x) = \frac{n!}{x!(n-x)!} \cdot p^x \cdot (1-p)^{n-x}, \text{ όπου } x = 0, 1, 2, 3, \ldots, n.$$

Η τιμή $P(X = x)$ ισούται με την πιθανότητα να συμβούν x 'επιτυχίες' στις n επαναλήψεις. Υπενθυμίζεται ότι $n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n$ δηλαδή είναι το γινόμενο των αριθμών $1, 2, 3, \ldots, n$.

Παράδειγμα

Η πιθανότητα σωστής απάντησης σε μια συγκεκριμένη ερώτηση είναι 0,7. Αν στην ερώτηση αυτή υποβληθούν 10 τυχαία άτομα, ποια είναι η πιθανότητα να απαντήσουν σωστά: 1) ακριβώς 6 άτομα, 2) τουλάχιστον 9 άτομα, 3) το πολύ 2 άτομα και 4) από 5 έως 7 άτομα.

Λύση

Πρόκειται για διωνυμικό πείραμα, στο οποίο ο αριθμός των επαναλήψεων είναι $n = 10$, ενώ η πιθανότητα 'επιτυχίας', δηλαδή σωστής απάντησης, είναι $p = 0,7$.

1) Είναι: $\begin{cases} P(X=6) = \dfrac{10!}{6!(10-6)!} \cdot 0,7^6 \cdot (1-0,7)^{10-6} = \\ = \dfrac{1\cdot2\cdot3\cdot4\cdot5\cdot6\cdot7\cdot8\cdot9\cdot10}{1\cdot2\cdot3\cdot4\cdot5\cdot6\cdot1\cdot2\cdot3\cdot4} \cdot 0,7^6 \cdot 0,3^4 = 210 \cdot 0,7^6 \cdot 0,3^4 = 0,2 \end{cases}$

2) Τουλάχιστον 9 άτομα, άρα 9 ή 10. Είναι: $P(X \geq 9) = P(X=9) + P(X=10)$.

$$P(X=9) = \frac{10!}{9!\cdot1!} \cdot 0,7^9 \cdot 0,3^1 = 10 \cdot 0,7^9 \cdot 0,3 = 0,121$$

$$P(X=10) = \frac{10!}{10!\cdot0!} \cdot 0,7^{10} \cdot 0,3^0 = 1 \cdot 0,7^{10} \cdot 1 = 0,7^{10} = 0,028.$$

Συνεπώς $P(X \geq 9) = 0.121 + 0,028 = 0,149$.

3) Είναι $P(X \leq 2) = P(X=0) + P(X=1) + P(X=2)$. Έτσι:

$$P(X=0) = \frac{10!}{0!\cdot10!} \cdot 0,7^0 \cdot 0,3^{10} = 1 \cdot 1 \cdot 0,3^{10} = 0,3^{10} = 0,0000059$$

$$P(X=1) = \frac{10!}{1!\cdot10!} \cdot 0,7^1 \cdot 0,3^9 = 10 \cdot 0,7 \cdot 0,3^9 = 0,00013778$$

$$P(X=2) = \frac{10!}{2!\cdot8!} \cdot 0,7^2 \cdot 0,3^8 = 45 \cdot 0,7^2 \cdot 0,3^8 = 0,0014467 .$$

Συνεπώς $P(X \leq 2) = 0,0000059 + 0,00013778 + 0,0014467 = 0,00159$.

4) Είναι $P(5 \leq X \leq 7) = P(X=5) + P(X=6) + P(X=7)$. Οπότε:

$$P(X=5) = \frac{10!}{5!\cdot5!} \cdot 0,7^5 \cdot 0,3^5 = 252 \cdot 0,7^5 \cdot 0,3^5 = 0,103$$

$P(X=6) = 0,2$ από το 1^{o} ερώτημα

$$P(X=7) = \frac{10!}{7!\cdot3!} \cdot 0,7^7 \cdot 0,3^3 = 120 \cdot 0,7^7 \cdot 0,3^3 = 0,267 .$$

Συνεπώς $P(5 \leq X \leq 7) = 0,103 + 0,2 + 0,267 = 0,57$.

Αποδεικνύεται ότι η <u>μέση τιμή</u> και η <u>διακύμανση</u> της διωνυμικής κατανομής ισούνται με: $\mu = n \cdot p$ και $\sigma^2 = n \cdot p \cdot (1-p)$ αντίστοιχα.

Έτσι, στο παραπάνω παράδειγμα, η μέση τιμή των σωστών απαντήσεων και η διακύμανση θα είναι αντίστοιχα:

$\mu = n \cdot p = 10 \cdot 0,7 = 7$ και $\sigma^2 = n \cdot p \cdot (1-p) = 10 \cdot 0,7 \cdot 0,3 = 2,1$.

Τονίζεται ακόμη, ότι η διωνυμική κατανομή είναι διακριτή γιατί οι τιμές που παίρνει η τυχαία μεταβλητή Χ είναι, όπως είδαμε, 0, 1, 2, 3, …, *n*.

§ 3. 7. Υπεργεωμετρική κατανομή

Αν στη διωνυμική κατανομή δεν ισχύει η συνθήκη 4, δηλαδή οι επαναλήψεις δεν είναι ανεξάρτητες μεταξύ τους ως προς το αποτέλεσμά τους, τότε η διωνυμική κατανομή αλλάζει μορφή.

Ας υποθέσουμε ότι ένα κιβώτιο περιέχει *N* σφαίρες και έστω ότι υπάρχουν *k* λευκές και *N* – *k* μαύρες σφαίρες. Αν πραγματοποιήσουμε *n* **ανεξάρτητες** εξαγωγές σφαιρών από το κιβώτιο, (δηλαδή με επανατοποθέτηση κάθε φορά της σφαίρας στο κιβώτιο), τότε ο αριθμός των εξαγόμενων λευκών σφαιρών του δείγματος είναι μια τυχαία μεταβλητή, η οποία ακολουθεί τη διωνυμική κατανομή.

Αν τώρα η εξαγωγή των *n* μονάδων του δείγματος γίνεται **χωρίς επανατοποθέτηση,** τότε το περιεχόμενο του κιβωτίου μεταβάλλεται μετά από κάθε εξαγωγή σφαίρας και οι εξαγωγές **δεν είναι πλέον ανεξάρτητες.** Ο αριθμός των περιεχόμενων λευκών σφαιρών στο δείγμα είναι μια ασυνεχής τυχαία μεταβλητή, η οποία δεν ακολουθεί τη διωνυμική κατανομή, αλλά τη λεγόμενη *υπεργεωμετρική κατανομή*.

Αν η τυχαία μεταβλητή *X* παριστάνει το πλήθος των λευκών σφαιρών στο δείγμα των *n* εξαγωγών, τότε, η πιθανότητα το δείγμα *n* σφαιρών να περιέχει *x* λευκές σφαίρες, υπολογίζεται με βάση τον τύπο:

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

όπου $x = 0, 1, 2, 3, \ldots, n$, N = πλήθος σφαιρών στο κιβώτιο, n = πλήθος σφαιρών στο δείγμα, k = πλήθος λευκών σφαιρών στο δείγμα.

Η υπεργεωμετρική κατανομή πιθανότητας εφαρμόζεται σε περιπτώσεις κατά τις οποίες η επιλογή των μονάδων του δείγματος γίνεται **χωρίς επανατοποθέτηση**, ενώ η διωνυμική κατανομή, όπως αναφέρθηκε, εφαρμόζεται σε περιπτώσεις που η επιλογή των μονάδων του δείγματος γίνεται με επανατοποθέτηση. Αν τώρα ο αριθμός *N* είναι πολύ μεγάλος και το μέγεθος του δείγματος *n* αναλογικά μικρό, τότε οι πιθανότητες που υπολογίζονται με την υπεργεωμετρική κατανομή δεν διαφέρουν πολύ από τις πιθανότητες που υπολογίζονται με τη διωνυμική κατανομή.

Αποδεικνύεται ότι η <u>μέση τιμή</u> και η <u>διακύμανση</u> της υπεργεωμετρικής κατανομής ισούνται με:

$\mu = n \cdot p$ και $\sigma^2 = n \cdot p \cdot (1-p)\dfrac{N-n}{N-1}$ δηλαδή μέση τιμή της υπεργεωμετρικής κατανομής συμπίπτει με τη μέση τιμή της διωνυμικής κατανομής, ενώ η διακύμανσή της είναι μικρότερη από αυτή της διωνυμικής κατανομής.

Η υπεργεωμετρική κατανομή έχει περιορισμένη εφαρμογή, γιατί ο υπολογισμός των πιθανοτήτων απαιτεί κοπιώδεις υπολογισμούς. Χρησιμοποιείται όμως στο Στατιστικό Έλεγχο Ποιότητας, για την αποδοχή ή απόρριψη προϊόντων που υποβάλλονται σε έλεγχο ποιότητας κατά μεγάλες παρτίδες.

Παράδειγμα

Σ’ ένα κιβώτιο υπάρχουν 10 ηλεκτρικοί λαμπτήρες, από τους οποίους οι 4 είναι καμένοι. Βγάζουμε από το κιβώτιο διαδοχικά (χωρίς επανατοποθέτηση) 5 λαμπτήρες. Ποια είναι η πιθανότητα να υπάρχουν στο δείγμα 2 καμένοι λαμπτήρες;

Εδώ έχουμε $N = 10$, $k = 4$, $n = 5$ και $x = 2$.

Αντικαθιστούμε τα δεδομένα αυτά στον παραπάνω τύπο και βρίσκουμε:

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

Αν τώρα εφαρμόσουμε τη διωνυμική κατανομή (δηλαδή βγάζουμε από το κιβώτιο με επανατοποθέτηση 5 λαμπτήρες), τότε θα είναι:

$$p = \frac{4}{10},\ 1-p = \frac{6}{10},\ n = 5 \text{ και } x = 2 \,.$$

Έτσι, αν αντικαταστήσουμε τα δεδομένα αυτά στον τύπο της διωνυμικής κατανομής της προηγούμενης παραγράφου βρίσκουμε:

$$P(X=2) = \binom{5}{2}\left(\frac{4}{10}\right)^2\left(\frac{6}{10}\right)^3 = 0{,}346 \,.$$

Παρατηρούμε ότι με τη διωνυμική κατανομή βρήκαμε μικρότερη πιθανότητα από την πραγματική που βρέθηκε με την υπεργεωμετρική κατανομή.

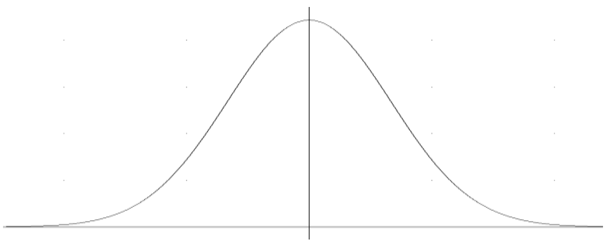
§ 3. 8. Κανονική κατανομή (ή κατανομή Gauss)

Η κανονική κατανομή παίζει σημαντικότατο ρόλο στη στατιστική ανάλυση. Κατ' αρχή αποδεικνύεται ότι αρκετές κατανομές πιθανοτήτων συγκλίνουν κάτω από ορισμένες συνθήκες προς την κανονική κατανομή. Επίσης, η χρήση πολλών στατιστικών μεθόδων διευκολύνεται, αν οι υπό μελέτη μεταβλητές κατανέμονται κανονικά. Αλλά υπάρχουν και πολλές μεταβλητές, όπως το I.Q. που ακολουθούν προσεγγιστικά κανονική κατανομή.

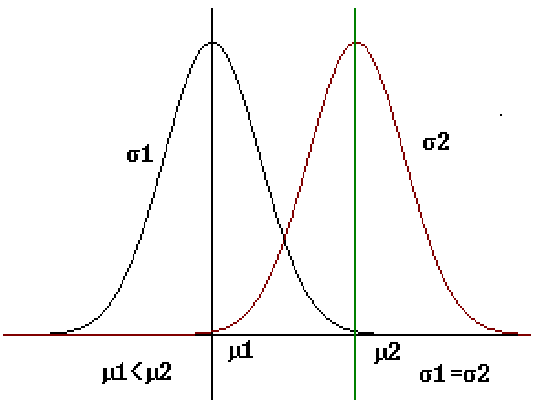
Η κανονική κατανομή, σε αντίθεση με τη διωνυμική, είναι συνεχής, δηλαδή η τυχαία μεταβλητή Χ παίρνει θεωρητικά οποιαδήποτε τιμή σε κάποιο διάστημα όπως το (-∞, +∞). Η εξίσωση της καμπύλης δίνεται από την παρακάτω σχέση:

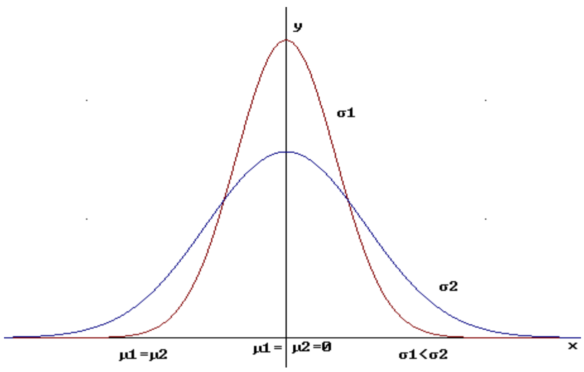
$y = \dfrac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$, όπου π = 3,14159 και e = 2,71828, ενώ μ και σ^2 είναι παράμετροι και ισούνται αντίστοιχα με τη μέση τιμή και τη διακύμανση της κατανομής, και καθορίζουν πλήρως την κατανομή. Γι' αυτό η κανονική κατανομή συμβολίζεται διεθνώς με: $N(\mu,\sigma^2)$ (από την λέξη Normal κανονική).

Η μορφή της κατανομής είναι χαρακτηριστική και φαίνεται στο παρακάτω σχήμα:



Στα επόμενα δύο σχήματα που ακολουθούν, φαίνεται η επίδραση της μεταβολής των παραμέτρων μ και σ στη μορφή της καμπύλης της κανονικής κατανομής.





Βασικές ιδιότητες της κανονικής κατανομής

1) Η κανονική καμπύλη έχει μορφή κωνοειδή και είναι συμμετρική ως προς τον κάθετο άξονα στο σημείο $x = \mu$.

2) Στο σημείο $x = \mu$ η κατανομή εμφανίζει μέγιστο.

3) Η κανονική καμπύλη πλησιάζει ασυμπτωτικά τον οριζόντιο άξονα όσο απομακρυνόμαστε από την κορυφή της.

4) Το εμβαδόν μεταξύ της κανονικής καμπύλης και του οριζόντιου άξονα είναι πάντοτε ίσο με 1.

Οι πιθανότητες που ενδιαφερόμαστε να υπολογίσουμε με χρήση της $N(\mu,\sigma^2)$ είναι πιθανότητες διαστημάτων επειδή η κατανομή είναι συνεχής (παρ. σελ. 26). Έτσι ενδιαφερόμαστε για την πιθανότητα αριστερά ή δεξιά μιας τιμής x ή ανάμεσα σε δύο τιμές x_1 και x_2. Για τον υπολογισμό αυτών των πιθανοτήτων απαιτείται η χρήση ειδικών πινάκων. Επειδή όμως για τις διάφορες τιμές των μ και σ ορίζονται διαφορετικές κανονικές κατανομές, κάνουμε πάντοτε χρήση συγκεκριμένου πίνακα. Ο πίνακας αυτός αναφέρεται σε μια ειδική μορφή της κανονικής κατανομής για την οποία είναι μ = 0 και σ^2 = 1 και ονομάζεται <u>τυποποιημένη κανονική κατανομή</u> ή Ζ-κατανομή Ν(0, 1).

Μπορεί να αποδειχθεί ότι αν η τυχαία μεταβλητή Χ ακολουθεί κανονική κατανομή με παραμέτρους μ και σ, τότε η τυχαία μεταβλητή $Z = \dfrac{X-\mu}{\sigma}$ (1) ακολουθεί τυποποιημένη κανονική κατανομή που συμβολίζεται όπως αναφέραμε με Ν(0, 1). Για τις διάφορες τιμές του Ζ ο πίνακας 1 (σελ. 74) της κανονικής κατανομής Ν(0, 1) δίνει τις πιθανότητες P(Z), όπου η πιθανότητα P(Z) εκφράζει το εμβαδόν κάτω από την καμπύλη της Ν(0, 1) <u>και αριστερά</u> της καθέτου στο σημείο Ζ.

Έτσι, για να λύσουμε ένα πρόβλημα με την κανονική κατανομή $N(\mu,\sigma^2)$, μετασχηματίζουμε τις τιμές Χ σε τιμές Ζ χρησιμοποιώντας τον παραπάνω τύπο (1), και επιλύουμε το πρόβλημα με τη χρήση του πίνακα 1 (σελ. 74) της Ν(0, 1).

Παράδειγμα

Έχει αποδειχθεί ότι ο χρόνος σε min επίλυσης κάποιου συγκεκριμένου προβλήματος ακολουθεί, σε επίπεδο πληθυσμού, κανονική κατανομή με μέση τιμή μ = 5 min και τυπική απόκλιση σ = 1,5 min. Να υπολογιστούν τα ποσοστά του πληθυσμού που επιλύουν το πρόβλημα αυτό:

α) σε χρόνο μικρότερο από 4,25 min

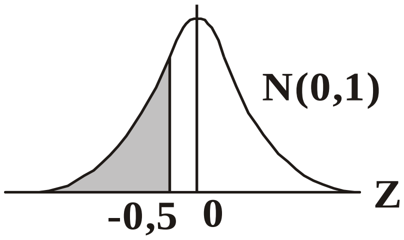
β) σε χρόνο μεγαλύτερο από 6,5 min

γ) σε χρόνο μεταξύ 3,5 min και 7,25 min.

Λύση

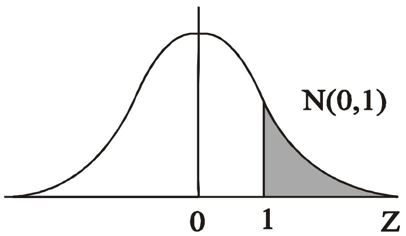
α) Θα μετατρέψουμε την τιμή X = 4,25 που είναι τιμή της κανονικής κατανομής $N(5, 1{,}5^2)$, σε τιμή της τυποποιημένης κανονικής κατανομής N(0, 1). Είναι:

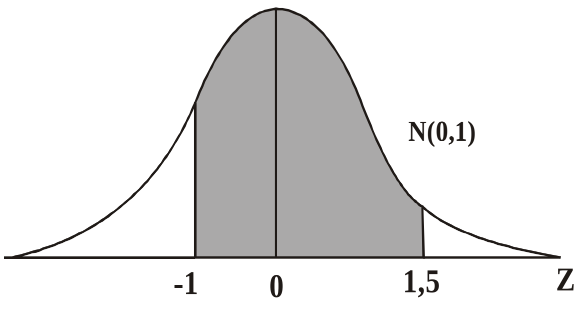
$Z = \frac{4{,}25-5}{1{,}5} = -0{,}5$. Ενδιαφερόμαστε για την πιθανότητα αριστερά της τιμής 4,25



που σημαίνει σε Z, τιμές αριστερά της τιμής -0,5. Από τον πίνακα 1 της N(0,1) βρίσκουμε ότι η πιθανότητα είναι 0,3085. Συνεπώς, ένα ποσοστό 30,85% του πληθυσμού επιλύουν το πρόβλημα σε χρόνο κάτω των 4,25 λεπτών. (σκιαγραφημένο μέρος).

β) Όμοια θα έχουμε $Z = \frac{6{,}5-5}{1{,}5} = 1$. Η πιθανότητα δεξιά της τιμής 6,5 θα ισούται με την πιθανότητα δεξιά της Z = 1. Από τον πίνακα της N(0,1) έχουμε ότι η πιθανότητα αριστερά της Z = 1 είναι 0,8413. Συνεπώς δεξιά της Z = 1 είναι 1 – 0,8413 = 0,1587 δηλαδή 15,87%. (σκιαγραφημένο μέρος).





γ) Εδώ έχουμε δύο τιμές για το Z. Συνεπώς:

$$Z_1 = \frac{3{,}5-5}{1{,}5} = -1 \quad \text{και}$$

$$Z_2 = \frac{7{,}25-5}{1{,}5}1{,}5 \ .$$

Η πιθανότητα αριστερά του Z_2 = 1,5 από τον πίνακα 1 της N(0,1) είναι 0,9332 ενώ αριστερά του Z_1 = – 1 είναι 0,1587. Άρα η πιθανότητα ανάμεσα στις δύο τιμές είναι η διαφορά τους $Z_2 - Z_1 = 0{,}9332 - 0{,}1587 = 0{,}7745$ ή 77,45%.

Παρατήρηση

Από τον πίνακα της N(0,1) και λόγω του ότι: $Z = \dfrac{X - \mu}{\sigma}$ συμπεραίνεται ότι σε μια κανονική κατανομή N(μ, σ^2):

1) το 68,27 % των μετρήσεων βρίσκονται στην περιοχή $\mu \pm \sigma$,

2) το 95,45 % των μετρήσεων βρίσκονται στην περιοχή $\mu \pm 2\sigma$,

3) το 99,73 % των μετρήσεων βρίσκονται στην περιοχή $\mu \pm 3\sigma$.

§ 3. 9. Σύγκλιση διωνυμικής προς κανονική κατανομή

Έστω X τυχαία μεταβλητή που ακολουθεί τη διωνυμική κατανομή B(*n, p*). Γνωρίζουμε (παράγρ. 3.6) ότι η μέση τιμή και η διακύμανση της διωνυμικής κατανομής είναι αντίστοιχα $\mu = n \cdot p$ και $\sigma^2 = n \cdot p \cdot (1 - p)$. Αποδεικνύεται ότι, όταν το μέγεθος *n* του δείγματος είναι μεγάλο και η πιθανότητα *p* είναι κοντά στο 0,5 τότε η τυχαία μεταβλητή $Z = \dfrac{X - np}{\sqrt{n \cdot p \cdot (1 - p)}}$ ακολουθεί προσεγγιστικά την τυποποιημένη κανονική κατανομή N(0,1). Η προσέγγιση είναι ικανοποιητική όταν $n{\cdot}p$ και $n{\cdot}(1 - p) > 5$.

§ 3. 10. Κατανομή Poisson

Υπάρχουν πολλές περιπτώσεις όπου, η πιθανότητα *p* της διωνυμικής κατανομής να συμβεί ένα ενδεχόμενο είναι πολύ μικρή ($p \le 0{,}10$), ενώ το *n* είναι μεγάλος αριθμός ($n > 50$) έτσι, ώστε ο μέσος όρος $\mu = n{\cdot}p$ να είναι μικρός αριθμός ($0 \le \mu \le 10$). Στις περιπτώσεις αυτές η τυχαία μεταβλητή X αντί της διωνυμικής κατανομής *B*(*n, p*) ακολουθεί την κατανομή **Poisson** με συνάρτηση πιθανότητας:

$$f(x) = P_x = P(X = x) = e^{-\lambda} \cdot \frac{\lambda^x}{x!}.$$

όπου $x = 0, 1, 2, 3, \ldots$, $e = 2{,}71828$ (βάση των νεπέρειων λογαρίθμων) και λ η παράμετρος που είναι ίση με το μέσο όρο, δηλαδή $\lambda = \mu = n{\cdot}p$.

Η τυχαία μεταβλητή *X* της κατανομής Poisson είναι ασυνεχής και παίρνει μόνο ακέραιες τιμές, ενώ το ιστόγραμμά της παρουσιάζει θετική ασυμμετρία που τείνει να γίνει συμμετρική όσο αυξάνεται το λ.

Η κατανομή Poisson εφαρμόζεται σε περιπτώσεις *σπάνιων γεγονότων*, όπως:

- αριθμό τηλεφωνικών κλήσεων που δέχεται ένα τηλεφωνικό κέντρο,

- αριθμό ημερήσιων αυτοκινητιστικών δυστυχημάτων σε μια εθνική οδό,
- αριθμό ελαττωματικών προϊόντων σε ισοπληθή δείγματα,
- αριθμό τυπογραφικών λαθών ανά σελίδα
- αριθμό αφίξεων σε ταμεία, οχημάτων σε διόδια, κ.λ.π.

Παράδειγμα

Κατά τη μαζική παραγωγή ομοειδών προϊόντων τα 5 στα 1000 τεμάχια είναι ελαττωματικά. Τα παραγόμενα προϊόντα συσκευάζονται σε κιβώτια των 100 τεμαχίων. Ποια η πιθανότητα να έχουμε: α) μηδέν ελαττωματικά, β) το πολύ δύο ελαττωματικά και γ) τουλάχιστον τρία ελαττωματικά προϊόντα;

Είναι $p = 5/1000 = 0{,}005$, $n = 100$ και $\lambda = n{\cdot}p = 100{\cdot}0{,}005 = 0{,}5$. Έτσι:

α) $P_0 = P(X = 0) = e^{-0,5} \cdot 0{,}5^0 / 0! = 0{,}6065$,

β) $P(X \le 2) = P_0 + P_1 + P_2$ όπου $P_1 = P(X = 1) = e^{-0,5} \cdot 0{,}5^1 / 1! = 0{,}3033$ και

$P_2 = P(X = 2) = e^{-0,5} \cdot 0{,}5^2 / 2! = 0{,}0758$. Άρα: $P(X \le 2) = 0{,}9856$.

γ) $P(X \ge 3) = P_3 + P_4 + \ldots + P_{100} = 1 - (P_0 + P_1 + P_2) = 1 - 0{,}9856 = 0{,}0144$.

§ 3. 11. Κατανομή t του Student

Η κατανομή t είναι συνεχής, δηλαδή η τυχαία μεταβλητή t μεταβάλλεται κατά συνεχή τρόπο από -∞ έως +∞ και εξαρτάται από μια παράμετρο *k* (που είναι αριθμός ακέραιος και θετικός), η οποία ονομάζεται «αριθμός των βαθμών ελευθερίας». Η μέση τιμή της κατανομής t είναι $\mu = 0$ για $k > 1$, ενώ η διακύμανση είναι μεγαλύτερη από 1 ($\sigma^2 > 1$). Όσο όμως ο αριθμός των βαθμών ελευθερίας αυξάνει, τόσο η διαφορά από το 1 μικραίνει.

Η κατανομή t είναι συμμετρική και κωδωνοειδής. Αποδεικνύεται ότι για μεγάλο αριθμό βαθμών ελευθερίας, η κατανομή t συγκλίνει προς την τυποποιημένη κανονική κατανομή Z. Η προσέγγιση είναι ικανοποιητική για $k > 30$.

§ 3. 12. Κατανομή X^2

Η τυχαία μεταβλητή είναι συνεχής και μεταβάλλεται από 0 έως +∞. Όπως η κατανομή t έτσι και η X^2 εξαρτάται από μια ακέραια θετική παράμετρο *k* που ονομάζεται «αριθμός βαθμών ελευθερίας». Η μέση τιμή της κατανομής είναι μ = k, ενώ η διακύμανση $\sigma^2 = 2k$.

Η κατανομή X^2 παρουσιάζει θετική ασυμμετρία. Αποδεικνύεται τέλος, ότι μια κατανομή X^2 με ένα βαθμό ελευθερίας ταυτίζεται με το τετράγωνο της τυποποιημένης κανονικής κατανομής Ζ.

§ 3. 13. Κατανομή F

Η τυχαία μεταβλητή μεταβάλλεται κατά συνεχή τρόπο από 0 έως $+\infty$ και εξαρτάται από δύο παραμέτρους ακέραιες και θετικές τις k_1 και k_2 που ονομάζονται «αριθμοί βαθμών ελευθερίας». Όπως η X^2 έτσι και η F έχει μορφή θετικά ασύμμετρη. Αποδεικνύεται ότι κάθε κατανομή F με 1 και k_2 αντίστοιχα βαθμούς ελευθερίας ισούται με το τετράγωνο της κατανομής t με k_2 βαθμούς ελευθερίας. Επίσης, αν η μεταβλητή Χ ακολουθεί κατανομή F με k_1 και k_2 βαθμούς ελευθερίας, τότε η μεταβλητή 1/Χ ακολουθεί κατανομή F με k_2 και k_1 βαθμούς ελευθερίας. Τέλος, αν κάθε μια από τις ανεξάρτητες τυχαίες μεταβλητές X_1 και X_2 ακολουθεί την κατανομή X^2 με k_1 και k_2 βαθμούς ελευθερίας αντίστοιχα, τότε και η τυχαία μεταβλητή $\frac{X_1^2 / k_1}{X_2^2 / k_2}$ ακολουθεί την κατανομή F με k_1 και k_2 βαθμούς αντίστοιχα.

§ 3.14. Στατιστικοί πίνακες των κατανομών t, X2 και F

Όπως για την τυποποιημένη κανονική κατανομή Ζ (πίνακας 1 σελ. 78), έτσι και για τις τρεις αυτές κατανομές υπάρχουν κατάλληλοι στατιστικοί πίνακες που επιτρέπουν την επίλυση προβλημάτων σχετικών με τις κατανομές αυτές όπως θα δούμε στα επόμενα κεφάλαια.

Συγκεκριμένα ο πίνακας 2 για την t κατανομή Student (σελ. 79) δίνει, σε σχέση με τον αριθμό των βαθμών ελευθερίας, τις θετικές τιμές t οι οποίες αφήνουν δεξιά τους πιθανότητες ίσες με 0,10, 0,05, 0,025, 0,01 και 0,005.

Ο πίνακας 3 της κατανομής X^2 (σελ. 76) δίνει, σε σχέση με τον αριθμό των βαθμών ελευθερίας, τις τιμές της μεταβλητής που αφήνουν δεξιά τους πιθανότητες ίσες με 0,995, 0,99, 0,975, 0,95, 0,05, 0,025, 0,01 και 0,005.

Τέλος ο πίνακας 4 της κατανομής F (σελίδες 77 και 78) δίνει, σε σχέση με τον αριθμό των βαθμών ελευθερίας k_1 και k_2, τις τιμές της μεταβλητής δεξιά των οποίων οι πιθανότητες είναι 0,05 και 0,01. Ο πίνακας αυτός δίνει τις τιμές που είναι πάντοτε μεγαλύτερες από 1.

Ασκήσεις

1) Ρίχνουμε ένα ιδανικό νόμισμα τρεις φορές. Ποια είναι η πιθανότητα να εμφανιστούν α) δύο κεφάλια, β) λιγότερα από δύο κεφάλια, γ) ένα κεφάλι ή δύο κεφάλια δ) τρία κεφάλια.

2) Ρίχνουμε ένα ιδανικό ζάρι. Ποια η πιθανότητα να εμφανιστεί πάνω στην έδρα α) ζυγός αριθμός, β) αριθμός μεγαλύτερος του 4, γ) αριθμός μικρότερος του 5.

3) Από μια τράπουλα με 52 χαρτιά τραβάμε ένα χαρτί. Ποια είναι η πιθανότητα το χαρτί να είναι α) Σπαθί η Ρήγας, β) Άσσος ή Καρό.

4) Σε ένα μικτό σχολείο με 100 μαθητές, τα αγόρια είναι 60, οι μαθητές που γνωρίζουν ξένη γλώσσα είναι 10 και τα αγόρια που γνωρίζουν ξένη γλώσσα είναι 7. Κληρώνουμε έναν μαθητή. Ποια η πιθανότητα ο μαθητής που κληρώθηκε να είναι γλωσσομαθής όταν είναι γνωστό ότι η κλήρωση έδειξε αγόρι.

5) Από μια τράπουλα με 52 φύλλα τραβάμε στην τύχη ένα φύλλο. Ποια είναι η πιθανότητα το φύλλο να είναι «Σπαθί», όταν είναι γνωστό ότι το φύλλο που τραβήξαμε είναι «Φιγούρα»

6) Αν οι πινακίδες κυκλοφορίας αυτοκινήτων περιέχουν τρία γράμματα που ακολουθούνται από έναν τετραψήφιο αριθμό, πόσες τέτοιες πινακίδες μπορούμε να σχηματίσουμε;

7) Σε ένα κοινοβούλιο με τρία κόμματα, το καθένα αντιπροσωπεύεται με 180, 90 και 30 βουλευτές. Πόσοι τρόποι υπάρχουν για τον σχηματισμό μιας 30/ μελούς επιτροπής στην οποία θα μετέχουν 18, 9 και 3 βουλευτές από κάθε κόμμα αντίστοιχα;

8) Στο παιχνίδι του ΛΟΤΤΟ έχουμε να επιλέξουμε 6 από τους 49 αριθμούς 1, 2, 3, … , 49. Με πόσους τρόπους μπορούμε να επιλέξουμε τους αριθμούς αυτούς;

9) Ένα ζαχαροπλαστείο διαθέτει 4 διαφορετικά είδη παγωτού (βανίλια, σοκολάτα, φράουλα, κεράσι). Με πόσους διαφορετικούς τρόπους μπορούμε να φτιάξουμε ένα χωνάκι από 3 μπάλες παγωτού αν: α) όλες οι μπάλες είναι διαφορετικές, β) δύο από τις μπάλες είναι ίδιες;

10) Σε ένα παιχνίδι τύχης «φρουτάκι» τραβάμε το μοχλό, με αποτέλεσμα, σε καθεμιά από τις τρεις θέσεις, να εμφανίζεται με την ίδια πιθανότητα ένα φρούτο (μήλο, κεράσι, ή μπανάνα). Κερδίζουμε 5 €, αν εμφανιστεί και στις τρεις θέσεις το ίδιο φρούτο. α) Αν παίξουμε μια φορά, ποια είναι η πιθανότητα να κερδίσουμε; β) Αν το κάθε παιχνίδι κοστίζει 2 €, σε 360 παιχνίδια, τι ποσό αναμένεται να κερδίσει η εταιρεία που έχει το παιχνίδι;

	Επιτυχόντες	Αποτυχόντες	Σύνολο
Αγόρια	90	360	450
Κορίτσια	30	120	150
Σύνολο	120	480	600

11) Στον διπλανό πίνακα περιέχονται τα αποτελέσματα 600 σπουδαστών που εξετάστηκαν στο μάθημα της Στατιστικής. Υποθέτουμε ότι τα χαρακτηριστικά «επίδοση» και «φύλο» είναι ανεξάρτητα μεταξύ τους. Αν από το σύνολο των σπουδαστών επιλεγεί τυχαία ένας σπουδαστής, να υπολογιστούν οι πιθανότητες, ο σπουδαστής να είναι α) επιτυχών και αγόρι β) αποτυχών και αγόρι γ) επιτυχούσα και κορίτσι και δ) αποτυχούσα και κορίτσι

12) Ρίχνουμε δύο ιδανικά ζάρια. Να βρεθεί η πιθανότητα α) το άθροισμα των ενδείξεων να είναι 7, β) το άθροισμα των ενδείξεων να είναι μεγαλύτερο του 9, γ) το άθροισμα των ενδείξεων να είναι μικρότερο ή ίσο του 8 και μεγαλύτερο ή ίσο του 5.

13) Προϊόν συσκευάζεται σε κιβώτια των 6 τεμαχίων. Παίρνουμε στην τύχη 500 κιβώτια και διαπιστώνουμε τον αριθμό των ελαττωματικών σε κάθε κιβώτιο.

Τα αποτελέσματα φαίνονται στον παρακάτω πίνακα:

ελαττωματικά (x_i)	0	1	2	3	4	5	6
αριθμός κιβωτίων (f_i)	65	145	160	90	25	10	5

α) Να υπολογιστεί ο μέσος όρος και το ποσοστό p (πιθανότητα) των ελαττωματικών τεμαχίων

β) Επιλέγεται στην τύχη ένα κιβώτιο από τα 500. Ποια η πιθανότητα να βρούμε στο κιβώτιο: i) ακριβώς δύο ελαττωματικά προϊόντα, ii) δύο ή τρία ελαττωματικά, iii) τέσσερα ή περισσότερα ελαττωματικά, iv) λιγότερα των τεσσάρων ελαττωματικά, v) το πολύ δύο ελαττωματικά, vi) τουλάχιστον δύο ελαττωματικά προϊόντα.

14) Ρίχνουμε ένα νόμισμα 10 φορές. Ποια είναι η πιθανότητα να εμφανιστεί το ενδεχόμενο «κεφάλι» από 3 έως 6 φορές;

15) Σ’ ένα σχολείο έγινε εμβολιασμός 2000 μαθητών. Η πιθανότητα να ασθενήσει ένας μαθητής από την αντίδραση του εμβολίου είναι 0,001. Ποια η πιθανότητα, από τους 2000 μαθητές να ασθενήσουν α) ακριβώς 2 μαθητές, β)το πολύ δύο μαθητές, γ) περισσότεροι από δύο μαθητές;

16) Η βαθμολογία 80.000 υποψηφίων που έλαβαν μέρος στις πανελλήνιες εξετάσεις προσεγγίζει την κανονική κατανομή με μέση τιμή 14 και τυπική απόκλιση 2. α) Να υπολογιστεί το ποσοστό και ο αριθμός των ατόμων που πήραν βαθμολογία από 13 μέχρι 16, β) μεγαλύτερη από 18, γ) μικρότερη από 10. δ) Να υπολογιστεί εκείνη η βαθμολογία των υποψηφίων, έτσι ώστε το 95% των βαθμολογιών να φτάνουν μέχρι την βαθμολογία αυτή.

17) Μηχανή που γεμίζει αυτόματα μπουκάλια μπύρας των 500 ml, έχει ρυθμιστεί έτσι ώστε το περιεχόμενο Χ να ακολουθεί κανονική κατανομή με μέση τιμή 480 ml και τυπική απόκλιση σ = 1,2 ml. Ποια είναι η πιθανότητα για ένα μπουκάλι που παίρνουμε με τυχαίο τρόπο, να έχει περιεχόμενο λιγότερο από 478 ml.

18) Είναι γνωστό ότι το ύψος του πληθυσμού των παιδιών στην προσχολική ηλικία ακολουθεί την κανονική κατανομή με μέση τιμή 100 cm και τυπική απόκλιση 10. Να βρεθεί το ποσοστό των παιδιών που το ύψος τους είναι μεταξύ 95 και 105 cm.

19) Η πιθανότητα επιτυχίας σ’ ένα μάθημα είναι 80%. Στις εξετάσεις του μαθήματος αυτού πήραν μέρος 15 σπουδαστές. Ποια είναι η πιθανότητα, το πολύ τρεις απ’ αυτούς να μην περάσουν το μάθημα.

Κεφάλαιο 4.

Δειγματοληπτικές κατανομές – Εκτιμητική

§ 4. 1. Δειγματοληπτικές κατανομές

Έστω ότι από έναν άπειρο αριθμό πληθυσμού παίρνουμε τυχαία δείγματα μεγέθους *n*. Αν συμβολίσουμε με $\bar{x}_1$ το μέσο όρο του πρώτου δείγματος, με $\bar{x}_2$ το μέσο όρο του δευτέρου δείγματος κ.ο.κ. οι τιμές $\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots$ μπορούν να θεωρηθούν τιμές μιας τυχαίας μεταβλητής $\overline{X}$. Η τυχαία αυτή μεταβλητή ακολουθεί μια κατανομή πιθανοτήτων που ονομάζεται δειγματοληπτική κατανομή του μέσου όρου.

Αν ο πληθυσμός από τον οποίο παίρνουμε τα δείγματα ακολουθεί κανονική κατανομή $N(\mu, \sigma^2)$ (παράγραφος 3.8), τότε η δειγματοληπτική κατανομή της $\overline{X}$ ακολουθεί και αυτή κανονική κατανομή $N(\mu, \sigma^2/n)$ δηλαδή έχει ως μέσο όρο τον μέσο όρο του πληθυσμού μ και διακύμανση σ^2/n, όπου σ^2 είναι η διακύμανση του πληθυσμού. Η τυπική απόκλιση της δειγματοληπτικής κατανομής ισούται προφανώς με $\sigma/\sqrt{n}$ και ονομάζεται τυπικό σφάλμα.

Ένα πολύ βασικό θεώρημα της Στατιστικής, το κεντρικό οριακό θεώρημα, δείχνει ότι και αν ακόμα η κατανομή του πληθυσμού δεν είναι κανονική, η δειγματοληπτική κατανομή του μέσου όρου ακολουθεί προσεγγιστικά την κανονική κατανομή αρκεί να πάρουμε δείγματα μεγάλου μεγέθους *n*. Όσο πιο πολύ αποκλίνει η κατανομή του πληθυσμού από την κανονική, τόσο μεγαλύτερο μέγεθος δειγμάτων απαιτείται. Αφού λοιπόν η $\overline{X}$ ακολουθεί την $N(\mu, \sigma^2/n)$, η τυχαία μεταβλητή $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ ακολουθεί την τυποποιημένη κανονική κατανομή (βλ. παράγρ. 3.8).

Γενικότερα, όπως οι μέσοι όροι των δειγμάτων, έτσι και άλλα στατιστικά ακολουθούν δειγματοληπτικές κατανομές.

Μια άλλη δειγματοληπτική κατανομή είναι π.χ. η δειγματοληπτική κατανομή μιας αναλογίας. Έστω ένας άπειρος πληθυσμός τον οποίο μελετούμε ως προς ένα χαρακτηριστικό. Καθένα από τα άτομα του πληθυσμού κατέχει ή δεν κατέχει το χαρακτηριστικό αυτό. Έστω *p* η αναλογία του χαρακτηριστικού μέσα στον πληθυσμό. Το

p αποτελεί προφανώς παράμετρο του πληθυσμού. Παίρνουμε από τον πληθυσμό τυχαία δείγματα μεγέθους n και έστω $f_1 = \frac{x_1}{n}$, $f_2 = \frac{x_2}{n}$, $f_3 = \frac{x_3}{n}$,... οι παρατηρούμενες αναλογίες του χαρακτηριστικού στα δείγματα αυτά. Ας σημειωθεί ότι $x_1, x_2, x_3, \ldots$ είναι αντίστοιχα ο αριθμός των ατόμων που κατέχουν το χαρακτηριστικό στα δείγματα. Οι τιμές $f_1, f_2, f_3, \ldots$ μπορούν να θεωρηθούν τιμές μιας τυχαίας μεταβλητής της $f = \frac{X}{n}$. Η τυχαία μεταβλητή X είναι μια διωνυμική μεταβλητή γιατί σε ένα τυχαίο δείγμα μεγέθους n, η τιμή της δίνει τον αριθμό 'επιτυχιών' (βλ. παράγρ. 3.6).

Γνωρίζουμε ακόμα ότι η τυχαία μεταβλητή X έχει μέσο όρο $\mu = n \cdot p$ και τυπική απόκλιση $\sigma = \sqrt{n \cdot p \cdot (1-p)}$. Αποδεικνύεται ότι η τυχαία μεταβλητή $f = \frac{X}{n}$ έχει μέσο όρο $\mu = p$ και τυπική απόκλιση $\sigma = \sqrt{\frac{p \cdot (1-p)}{n}}$.

Όπως η τυχαία μεταβλητή $Z = \frac{X - n \cdot p}{\sqrt{n \cdot p \cdot (1-p)}}$ ακολουθεί την τυποποιημένη κανονική κατανομή N(0, 1) (βλ. παράγρ. 3.8), έτσι και η δειγματοληπτική κατανομή της f ακολουθεί κανονική κατανομή με παραμέτρους, όπως είδαμε, $\mu = p$ και $\sigma = \sqrt{\frac{p \cdot (1-p)}{n}}$ με την προϋπόθεση ότι, το μέγεθος του δείγματος n να είναι μεγάλο και το p κοντά στο 0,5. Συνεπώς η τυχαία μεταβλητή $Z = \frac{f - p}{\sqrt{\frac{p \cdot (1-p)}{n}}}$ ακολουθεί προσεγγιστικά την τυποποιημένη κανονική κατανομή N(0, 1).

Θα ονομάζουμε <u>αναμενόμενη τιμή</u> ενός στατιστικού τον μέσο όρο της δειγματοληπτικής κατανομής του στατιστικού, ενώ τυπικό σφάλμα την τυπική απόκλιση της κατανομής αυτής.

Είδαμε πιο πάνω ότι οι αναμενόμενες τιμές των στατιστικών $\overline{X}$ και f είναι αντίστοιχα μ και p δηλαδή ο μέσος όρος του πληθυσμού και η αναλογία του χαρακτηριστικού στον πληθυσμό. Ας αναφερθεί τέλος, ότι για την δειγματοληπτική κατανομή του στατιστικού s^2 δηλαδή των διακυμάνσεων, η αναμενόμενη τιμή ισούται με σ^2 δηλαδή με τη διακύμανση του πληθυσμού.

§ 4. 2. Διαστήματα εμπιστοσύνης

Ένας από τους λόγους που υπολογίζονται τα στατιστικά ($\overline{X}, s^2, f, ...$) είναι για να εκτιμήσουν τις αντίστοιχες παραμέτρους του πληθυσμού ($\mu, \sigma^2, p, ...$) των οποίων οι τιμές είναι άγνωστες.

Ένα στατιστικό ονομάζεται αμερόληπτη εκτιμήτρια μιας παραμέτρου, αν η αναμενόμενη τιμή της ισούται με την παράμετρο. Τα στατιστικά $\overline{X}, s^2, f$ είναι αμερόληπτες εκτιμήτριες των παραμέτρων μ, σ^2, p αντίστοιχα, γιατί όπως είδαμε, οι αναμενόμενες τιμές τους είναι ίσες με τις παραμέτρους αυτές. Συνεπώς, η καλύτερη εκτίμηση των άγνωστων παραμέτρων μ, σ^2, p είναι αντίστοιχα ο μέσος όρος $\overline{X}$, η διακύμανση s^2 και η αναλογία f ενός τυχαίου δείγματος. Τα $\overline{X}, s^2, f$ ονομάζονται σημειακές εκτιμήσεις των παραμέτρων μ, σ^2, p . Η ακρίβεια μιας εκτίμησης μπορεί να μετρηθεί με το τυπικό σφάλμα. Έτσι, όπως είδαμε προηγουμένως, το τυπικό σφάλμα της δειγματοληπτικής κατανομής του στατιστικού $\overline{X}$ είναι $\frac{\sigma}{\sqrt{n}}$. Φαίνεται συνεπώς ότι η ακρίβεια της εκτίμησης αυξάνει όσο αυξάνει το μέγεθος του δείγματος *n*, γιατί τότε το σφάλμα μικραίνει.

Ένας άλλος τρόπος εκτίμησης που συμπληρώνει αυτόν της σημειακής εκτίμησης είναι τα διαστήματα εμπιστοσύνης. Με το διάστημα εμπιστοσύνης επιχειρείται η ανεύρεση των ορίων ενός διαστήματος για το οποίο να υπάρχει μεγάλη πιθανότητα να περιέχει την άγνωστη παράμετρο. Τα όρια του διαστήματος ονομάζονται όρια εμπιστοσύνης, ενώ η πιθανότητα με την οποία το διάστημα περιέχει την παράμετρο ονομάζεται επίπεδο εμπιστοσύνης. Το επίπεδο εμπιστοσύνης καθορίζεται συνήθως σε 0,95 ή 0,99. Τα διαστήματα εμπιστοσύνης κατασκευάζονται γύρω από την σημειακή εκτίμηση της άγνωστης παραμέτρου. Πρέπει να διευκρινιστεί ότι όταν λέμε, ότι το διάστημα εμπιστοσύνης έχει πιθανότητα π.χ. 0,99 να περιέχει την άγνωστη παράμετρο, αυτό σημαίνει, ότι παίρνοντας 100 τυχαία δείγματα μέσα από τον πληθυσμό και κατασκευάζοντας ανάλογο αριθμό διαστημάτων, τα 99 από αυτά αναμένεται να περιέχουν την παράμετρο αυτή. Παρακάτω περιγράφονται τα διαστήματα εμπιστοσύνης για τις παραμέτρους μ και *p* δηλαδή για το μέσο όρο και την αναλογία ενός χαρακτηριστικού σε κάποιο πληθυσμό.

Α) Διάστημα εμπιστοσύνης για το μέσο όρο μ ενός πληθυσμού

Έστω πληθυσμός που κατανέμεται κανονικά με άγνωστο μέσο όρο μ και άγνωστη τυπική απόκλιση σ. Παίρνουμε μέσα από τον πληθυσμό τυχαίο δείγμα μεγέθους n. Έστω $\bar{x}$ και s ο μέσος όρος και η τυπική απόκλιση αυτού του δείγματος αντίστοιχα. Είχε αναφερθεί παραπάνω ότι το επίπεδο εμπιστοσύνης καθορίζει την πιθανότητα να κατασκευαστεί διάστημα που να περιέχει την άγνωστη παράμετρο. Αν συνεπώς αυτή η πιθανότητα είναι 0,95, η τιμή 1 – 0,95 = 0,05 δίνει την πιθανότητα, το διάστημα να μην περιέχει την παράμετρο. Όμοια, αν το επίπεδο εμπιστοσύνης είναι 0,99, τότε 1 – 0,99 = 0,01 δίνει αντίστοιχα την πιθανότητα, το διάστημα να μην περιέχει την παράμετρο. Θα συμβολίζουμε από εδώ και πέρα με α την πιθανότητα, το διάστημα να μην περιέχει την παράμετρο. Συνεπώς $\alpha = 0{,}05$ ή $\alpha = 0{,}01$.

Μετά από όλα αυτά, το διάστημα εμπιστοσύνης για τον άγνωστο μέσο όρο μ ενός πληθυσμού ευρίσκεται ως εξής:

<u>1) στην περίπτωση μεγάλου δείγματος (n ≥ 30):</u>

i) Το 95% των ορίων εμπιστοσύνης δίδεται από τον τύπο:

$$\bar{x} \pm 1{,}96\, s/\sqrt{n}$$

δηλαδή το κατώτερο όριο είναι $\bar{x} - 1{,}96\, s/\sqrt{n}$ και το ανώτερο $\bar{x} + 1{,}96\, s/\sqrt{n}$.

ii) Το 99% των ορίων εμπιστοσύνης δίδεται από τον τύπο:

$$\bar{x} \pm 2{,}58\, s/\sqrt{n}$$

δηλαδή το κατώτερο όριο είναι $\bar{x} - 2{,}58\, s/\sqrt{n}$ και το ανώτερο $\bar{x} + 2{,}58\, s/\sqrt{n}$.

<u>2) στην περίπτωση μικρού δείγματος (n < 30):</u>

Στην περίπτωση αυτή χρησιμοποιούμε την κατανομή t για να πάρουμε τα όρια εμπιστοσύνης που δίνονται από τον τύπο:

$$\bar{x} \pm t_{\alpha/2}\, s/\sqrt{n}\ , \qquad (1)$$

δηλαδή το κατώτερο όριο είναι $\bar{x} - t_{\alpha/2} \dfrac{s}{\sqrt{n}}$ και το ανώτερο όριο είναι $\bar{x} + t_{\alpha/2} \dfrac{s}{\sqrt{n}}$,

όπου $t_{\alpha/2}$ είναι η τιμή της κατανομής t του Student (παράγρ. 3.7) που αφήνει δεξιά της πιθανότητα ίση με $\alpha / 2$ για n – 1 βαθμούς ελευθερίας. Δηλαδή, αν το επίπεδο εμπιστοσύνης είναι 0,95, τότε α = 0,05 άρα $\alpha / 2$ = 0,025.

Ενώ αν είναι 0,99, τότε α = 0,01 άρα $\alpha / 2$ = 0,005. Συνεπώς οι τιμές $t_{\alpha/2}$ εντοπίζονται από τον πίνακα της κατανομής t προσέχοντας τις στήλες με την ένδειξη 0,025 ή 0,005 για τους ανάλογους βαθμούς ελευθερίας φυσικά.

Παράδειγμα

Σε τυχαίο δείγμα 20 μαθητών μετρήθηκε η επίδοσή τους στο μάθημα Α. Βρέθηκε ότι ο μέσος όρος ήταν $\bar{x} = 6{,}5$ και η τυπική απόκλιση $s = 1{,}1$. Θέλουμε να κατασκευάσουμε ένα διάστημα εμπιστοσύνης που με πιθανότητα 1) 0,95 και 2) 0,99 να περιέχει τον άγνωστο μέσο όρο μ της επίδοσης όλου του πληθυσμού των μαθητών στο μάθημα Α.

Λύση

Υποθέτουμε κατ' αρχήν ότι σε επίπεδο πληθυσμού η επίδοση στο συγκεκριμένο μάθημα κατανέμεται κανονικά. Είναι n = 20, $\bar{x} = 6{,}5$ και s = 1,1.

1) Αν το επίπεδο εμπιστοσύνης είναι 0,95 τότε α = 0,05 και $\alpha/2$ = 0,025. Οι βαθμοί ελευθερίας είναι n – 1 δηλαδή 20 – 1 = 19. Από τον πίνακα κατανομής t για 19 βαθμούς ελευθερίας και από τη στήλη 0,025 βρίσκουμε ότι $t_{\alpha/2} = t_{0{,}025} = 2{,}093$.

Κατώτερο όριο διαστήματος: $\bar{x} - t_{\alpha/2}\dfrac{s}{\sqrt{n}} = 6{,}5 - 2{,}093\dfrac{1{,}1}{\sqrt{20}} = 6{,}5 - 0{,}515 = 5{,}985$.

Ανώτερο όριο διαστήματος: $\bar{x} + t_{\alpha/2}\dfrac{s}{\sqrt{n}} = 6{,}5 + 2{,}093\dfrac{1{,}1}{\sqrt{20}} = 6{,}5 + 0{,}515 = 7{,}015$.

Συνεπώς το διάστημα εμπιστοσύνης είναι το εξής: [5,985 , 7,015]

2) Αν το επίπεδο εμπιστοσύνης είναι 0,99 τότε α = 0,01 άρα $\alpha/2$ = 0,005. Για 19 βαθμούς ελευθερίας ο πίνακας IV δίνει $t_{\alpha/2} = t_{0{,}005} = 2{,}861$.

Κατώτερο όριο διαστήματος: $6{,}5 - 2{,}861\dfrac{1{,}1}{\sqrt{20}} = 6{,}5 - 0{,}704 = 5{,}796$.

Ανώτερο όριο διαστήματος: $6{,}5 + 2{,}861\dfrac{1{,}1}{\sqrt{20}} = 6{,}5 + 0{,}704 = 7{,}204$.

Επομένως το διάστημα εμπιστοσύνης είναι: [5,796 , 7,204].

Β) Διάστημα εμπιστοσύνης για την αναλογία p ενός χαρακτηριστικού σε κάποιο πληθυσμό

Έστω πληθυσμός τον οποίο μελετούμε ως προς ένα χαρακτηριστικό και έστω p η άγνωστη αναλογία του χαρακτηριστικού μέσα στον πληθυσμό. Σε τυχαίο δείγμα n ατόμων βρίσκουμε ότι x άτομα κατέχουν το χαρακτηριστικό, ενώ $n – x$ άτομα δεν το κατέχουν. Συνεπώς η αναλογία των ατόμων που κατέχουν το χαρακτηριστικό στο δείγμα είναι $f = x/n$. Θεωρούμε ότι το p δεν απέχει πολύ από την τιμή 0,5 και x και $n – x$ είναι μεγέθους πάνω από 20. Κάτω από αυτές τις συνθήκες, ένα διάστημα εμπιστοσύνης για την παράμετρο p δίνεται από τον εξής τύπο:

$$f \pm Z_{1-\alpha/2}\sqrt{f\cdot(1-f)/n}, \qquad (2)$$

με κατώτερο όριο: $f - Z_{1-\alpha/2}\sqrt{f\cdot(1-f)/n}$ και ανώτερο: $f + Z_{1-\alpha/2}\sqrt{f\cdot(1-f)/n}$,

όπου $Z_{1-\alpha/2}$ είναι η τιμή της τυποποιημένης κανονικής κατανομής Ν(0, 1) που αφήνει αριστερά της πιθανότητας $1 – \alpha/2$. Πιο συγκεκριμένα, αν το επίπεδο εμπιστοσύνης είναι 0,95 τότε $\alpha = 1 – 0,95 = 0,05$. Συνεπώς $1 – \alpha/2 =1 – 0,025 = 0,975$.
Από τον πίνακα της Ν(0, 1) βρίσκουμε ότι η τιμή $Z_{0,975}$ είναι η 1,96. Αν το επίπεδο εμπιστοσύνης είναι 0,99 τότε $\alpha = 0,01$ άρα $1 – \alpha/2 = 1 – 0,005 = 0,995$. Από τον ίδιο πίνακα βρίσκουμε ότι η τιμή $Z_{0,995}$ είναι η 2,58.

Παράδειγμα

Σε τυχαίο δείγμα 1000 ατόμων βρέθηκε ότι 400 άτομα τάχθηκαν εναντίον της προβολής κάποιου τηλεοπτικού μηνύματος, ενώ οι υπόλοιποι 600 υπέρ. Να κατασκευαστεί διάστημα εμπιστοσύνης που με πιθανότητα 1) 0,95 και 2) 0,99 να περικλείει την άγνωστη αναλογία p των ατόμων που σε επίπεδο πληθυσμού είναι εναντίον της προβολής.

Λύση

Η αναλογία των ατόμων του δείγματος που κατέχουν το χαρακτηριστικό (κατά της προβολής) είναι $f = \frac{x}{n} = \frac{400}{1000} = 0,4$ (εναντίον). Επίσης $1 – f = 0,6$ (υπέρ). Είναι επίσης $x = 400$, $n – x = 600$ και το p όχι μακριά από το 0,5 διότι η σημειακή εκτίμηση του p δηλαδή το f είναι 0,4. Συνεπώς οι προϋποθέσεις εγκυρότητας πληρούνται.

1) Για επίπεδο εμπιστοσύνης 0,95 είναι $Z_{1-\alpha/2}$ = 1,96. Άρα, κατώτερο όριο διαστήματος : $f - Z_{1-\alpha/2}\sqrt{f\cdot(1-f)/n} = 0{,}4 - 1{,}96\sqrt{0{,}4\cdot 0{,}6/1000} = 0{,}4 - 0{,}03 = 0{,}37$ και ανώτερο: $f + Z_{1-\alpha/2}\sqrt{f\cdot(1-f)/n} = 0{,}4 + 1{,}96\sqrt{0{,}4\cdot 0{,}6/1000} = 0{,}4 + 0{,}03 = 0{,}43$.
Συνεπώς το ζητούμενο διάστημα εμπιστοσύνης είναι: [0,37 , 0,43].

2) Για επίπεδο εμπιστοσύνης 0,99 είναι $Z_{1-\alpha/2}$ = 2,58. Άρα, κατώτερο όριο: $f - Z_{1-\alpha/2}\sqrt{f\cdot(1-f)/n} = 0{,}4 - 2{,}58\sqrt{0{,}4\cdot 0{,}6/1000} = 0{,}4 - 0{,}04 = 0{,}36$, ανώτερο: $f + Z_{1-\alpha/2}\sqrt{f\cdot(1-f)/n} = 0{,}4 + 2{,}58\sqrt{0{,}4\cdot 0{,}6/1000} = 0{,}4 + 0{,}04 = 0{,}44$.
Συνεπώς το ζητούμενο διάστημα εμπιστοσύνης είναι: [0,36 , 0,44].

Παρατηρήσεις για τα διαστήματα εμπιστοσύνης

1) Τα διαστήματα για τις παραμέτρους μ και *p* έχουν, λόγω κατασκευής τους, ως κέντρο τις σημειακές εκτιμήσεις των μ και *p* δηλαδή τα $\bar{x}$ και *f*.

2) Για δείγμα σταθερού μεγέθους, όσο αυξάνει το επίπεδο εμπιστοσύνης, τόσο τα διαστήματα εμπιστοσύνης έχουν μεγαλύτερο πλάτος και το αντίθετο. Αυτό φαίνεται και από τα δύο προηγούμενα παραδείγματα, όπου το επίπεδο εμπιστοσύνης ήταν 0,95 και 0,99. Επειδή ενδιαφερόμαστε να έχουμε διαστήματα με μεγάλο επίπεδο εμπιστοσύνης και μικρό πλάτος, πρέπει να παίρνουμε δείγματα μεγάλου μεγέθους, οπότε το πλάτος μικραίνει όπως προκύπτει από τους τύπους (1) και (2).

3) Για μεγάλου μεγέθους δείγματα ($n \geq 30$), τα διαστήματα εμπιστοσύνης για κανονική κατανομή και σε ποσοστά 68,27%, 95,45% και 99,73% δίνονται από τους τύπους $\bar{x} \pm s/\sqrt{n}$, $\bar{x} \pm 2s/\sqrt{n}$, $\bar{x} \pm 3s/\sqrt{n}$ αντίστοιχα (βλέπε σελίδα 32).

3) Για μικρού μεγέθους δείγματα, τα όρια των διαστημάτων εμπιστοσύνης για την αναλογία *p* ενός χαρακτηριστικού δίνονται από ειδικό πίνακα.

Ασκήσεις

1) Από δείγμα 100 κιβωτίων φρούτων, τα 5 κιβώτια είχαν βάρος 61 κιλά, τα 18 κιβώτια 64 κιλά, τα 42 είχαν 67 κιλά, τα 27 είχαν 70 και τα 8 κιβώτια 73 κιλά. Να υπολογιστούν τα όρια εμπιστοσύνης για τη μέση τιμή μ του βάρους των κιβωτίων του πληθυσμού α) για το 95%, β) για το 99% και γ) για το 99,73% όλου του πληθυσμού.

2) Μια μηχανή κατασκευάζει μπίλιες για ρουλεμάν. Οι 200 μπίλιες ενός τυχαίου δείγματος από την παραγωγή έχουν διαμέτρους με μέση τιμή 0,824 cm και τυπική απόκλιση 0,042 cm.Να υπολογιστούν τα 95% και 99% όρια εμπιστοσύνης για τη μέση τιμή της διαμέτρου του πληθυσμού.

3) Έχει βρεθεί ότι ο χρόνος αντιδράσεως σε εξωτερικούς ερεθισμούς έχει τυπική απόκλιση 0,05 sec. Πόσο μεγάλο πρέπει να είναι το δείγμα των μετρήσεων, για να είμαστε βέβαιοι α) 95% και β) 99% ότι το σφάλμα στην εκτίμηση του μέσου χρόνου αντιδράσεως δεν υπερβαίνει το 0,01 sec;

4) Δέκα μετρήσεις της διαμέτρου μιας σφαίρας έδωσαν μέση τιμή 4,38 cm και τυπική απόκλιση 0,06 cm. Να υπολογιστούν α) τα 95% και β) τα 99% όρια εμπιστοσύνης για την πραγματική διάμετρο της σφαίρας.

5) Από 100 τυχαίους ψηφοφόρους που ρωτήθηκαν σε μια περιοχή, οι 55 είπαν ότι θα ψηφήσουν έναν ορισμένο υποψήφιο. Να υπολογιστούν α) τα 95%, β) τα 99% και γ) τα 99,73% όρια εμπιστοσύνης για την αναλογία των ψήφων που θα πάρει ο υποψήφιος αυτός κατά τις εκλογές σ' όλη την περιοχή.

6) Σε 40 ρίψεις ενός νομίσματος ήρθε 24 φορές 'κεφάλι'. Να υπολογιστούν α) τα 95% και β) τα 99,73% όρια εμπιστοσύνης για την αναλογία των αποτελεσμάτων 'κεφάλι' σε πολύ μεγάλο αριθμό ρίψεων του νομίσματος.

7) Η μέση τιμή και η τυπική απόκλιση των μέγιστων φορτίων που άντεξαν 60 συρματόσχοινα ενός τύπου είναι 11,09 και 0,73 τόνοι αντίστοιχα. Να υπολογιστούν τα α) 95% και β) 99% όρια εμπιστοσύνης για τη μέση τιμή αντοχής των συρματόσχοινων του τύπου αυτού.

8) Από 250 παξιμάδια η μέση τιμή και η τυπική απόκλιση των εσωτερικών διαμέτρων τους είναι 0,72642 και 0,00058 cm αντίστοιχα. Υπολογίστε τα α) 95%, β) 99%, γ) 98% και δ) 90% όρια εμπιστοσύνης για τη μέση διάμετρο του πληθυσμού των παξιμαδιών.

9) Η τυπική απόκλιση της ζωής ενός τύπου λυχνιών είναι 100 ώρες. i) Πόσο μεγάλο πρέπει να είναι ένα δείγμα ώστε να είμαστε α) 95%, β) 99%, γ) 90% και δ) 99,73% βέβαιοι ότι το σφάλμα στην εκτίμηση της μέσης ζωής δεν υπερβαίνει τις 20 ώρες; ii) Πόσο μεγάλο πρέπει να είναι το δείγμα στις προηγούμενες περιπτώσεις, ώστε το σφάλμα στην εκτίμηση της μέσης ζωής να μην υπερβαίνει τις 10 ώρες;

10) Δώδεκα μετρήσεις της αντοχής ενός σπάγκου έδωσαν μέση τιμή μέγιστου φορτίου 7,38 κιλά και τυπική απόκλιση 1,24 κιλά. Υπολογίστε τα α) 95% και β) 99% όρια εμπιστοσύνης για το πραγματικό μέγιστο φορτίο.

11) Ένα κουτί περιέχει κόκκινες και άσπρες σφαίρες σε άγνωστη αναλογία. Παίρνουμε με επανατοποθέτηση ένα τυχαίο δείγμα με 60 σφαίρες. Εάν το 70% από αυτές είναι κόκκινες, υπολογίστε τα α) 95%, β) 99% και γ) 99,73% όρια εμπιστοσύνης για την αναλογία των κόκκινων σφαιρών στο κουτί.

Κεφάλαιο 5.

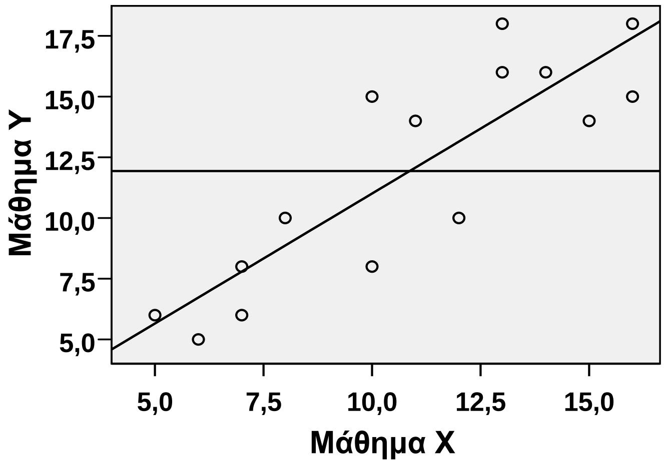
Συσχέτιση - Παλινδρόμηση

§ 5. 1. Συντελεστής συσχέτισης

Έστω ότι θέλουμε να μελετήσουμε το βαθμό συσχέτισης μεταξύ των επιδόσεων σε δύο μαθήματα Χ και Υ, δεκαπέντε μαθητών. Στον παρακάτω πίνακα εμφανίζονται οι επιδόσεις αυτές και υπολογίζονται ορισμένες ποσότητες χρήσιμες για τη συνέχεια.

Α/Α	*X*	*Y*	*X²*	*Y²*	*XY*
1	16	18	256	324	288
2	5	6	25	36	30
3	16	15	256	225	240
4	13	18	169	324	234
5	10	15	100	225	150
6	13	16	169	256	208
7	8	10	64	100	80
8	15	14	225	196	210
9	7	8	49	64	56
10	12	10	144	100	120
11	11	14	121	196	154
12	6	5	36	25	30
13	14	16	196	256	224
14	7	6	49	36	42
15	10	8	100	64	80
Σύνολο	***163***	***179***	***1959***	***2427***	***2146***

Η μελέτη της συσχέτισης μπορεί να αρχίσει με την κατασκευή του διαγράμματος διασποράς. Σε ένα σύστημα καθέτων αξόνων, κάθε τιμή (x_i, y_i) $i = 1, 2,\ldots,15$ αναπαριστάνεται με ένα σημείο με συντεταγμένες (x_i, y_i). Στο επόμενο σχήμα παρουσιάζεται το διάγραμμα διασποράς για τις μετρήσεις του παραπάνω παραδείγματος. Το σύνολο των 15 σημείων ονομάζεται σμήνος σημείων. Ο οριζόντιος άξονας περιλαμβάνει τις μετρήσεις για το μάθημα Χ, ενώ ο κάθετος τις μετρήσεις για το μάθημα Υ. Επίσης, η οριζόντια γραμμή εκφράζει το μέσο όρο του μαθήματος Υ που είναι 11,93 ενώ ο μέσος όρος του Χ είναι 10,87. Η μελέτη του σχήματος αυτού δείχνει ότι γενικά τα άτομα με μεγάλη επίδοση στο μάθημα Χ, έχουν και μεγάλη επίδοση στο μάθημα Υ, ενώ τα άτομα με μικρή επίδοση στο μάθημα Χ, έχουν επίσης μικρή επίδοση στο μάθημα Υ. Το σχήμα δείχνει ακόμα ότι η συσχέτιση μεταξύ των Χ και Υ έχει μορφή που τείνει προς τη γραμμική, δηλαδή η συσχέτιση μπορεί να αναπαρασταθεί αρκετά ικανοποιητικά με μία ευθεία που λέγεται <u>ευθεία παλινδρόμησης</u> και είναι η πλάγια ευθεία. Για τον τρόπο υπολογισμού της θα αναφερθούμε στην επόμενη παράγραφο.



Για να μετρήσουμε το βαθμό της συσχέτισης μεταξύ δύο μεταβλητών, που οι τιμές τους εκφράζονται ως ποσοτικές μεταβλητές, χρησιμοποιούμε το συντελεστή συσχέτισης Pearson. Ο συντελεστής συσχέτισης συμβολίζεται με r όταν χρησιμοποιείται ως στατιστικό, ενώ με ρ όταν αποτελεί παράμετρο. Ο τύπος του είναι ο εξής:

$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\cdot\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (1)$$

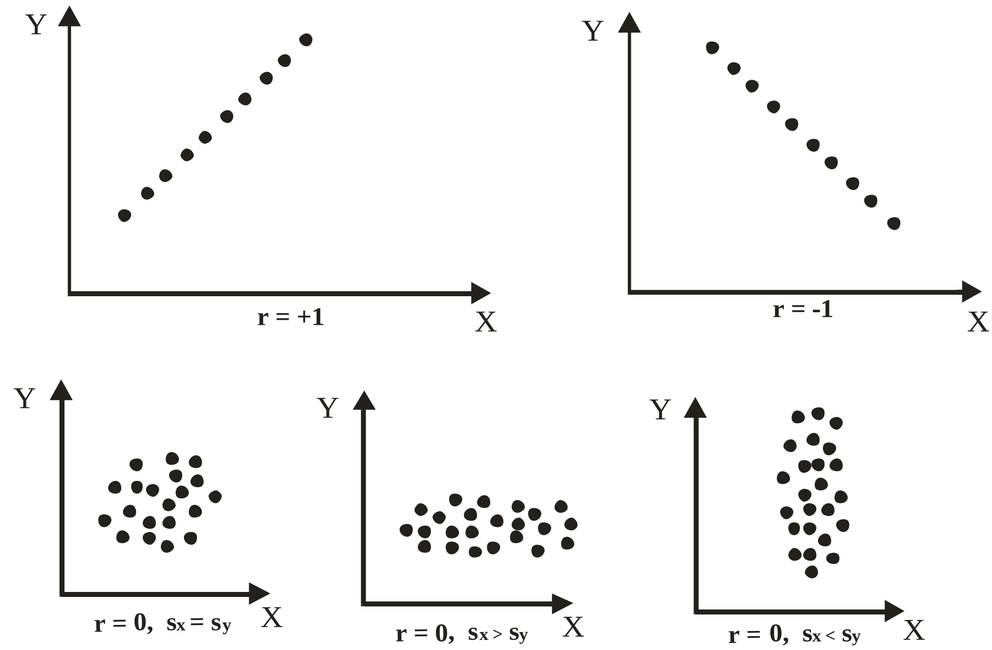
Ο αριθμητής του κλάσματος συμβολίζεται με cov(X,Y) και λέγεται *συνδιακύμανση* των X και Y (*covariance*). Συνεπώς $r = \frac{\mathrm{cov}(X,Y)}{s_x \cdot s_y}$ δηλαδή ο συντελεστής συσχέτισης ισούται με το πηλίκο της συνδιακύμανσης μεταξύ των X και Y προς το γινόμενο των τυπικών αποκλίσεων των X και Y.

Μια άλλη μορφή της (1) πιο κατάλληλη για υπολογισμούς είναι ο επόμενος τύπος (2), στον οποίο μπορούμε εύκολα να αντικαταστήσουμε τις τιμές όλων των αθροισμάτων που υπολογίσαμε στον αρχικό πίνακα, και να υπολογίσουμε έτσι την τιμή της συνδιακύμανσης r.

$$r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\cdot\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\sum_{i=1}^{n} x_i^{2} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}\cdot\sqrt{\sum_{i=1}^{n} y_1^{2} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}} \qquad (2)$$

Αποδεικνύεται ότι ο συντελεστής συσχέτισης παίρνει τιμές στο διάστημα [–1, +1] δηλαδή $-1 \le r \le 1$. Αν $0 < r \le 1$ οι μεταβλητές είναι θετικά συσχετισμένες, ενώ αν

$-1 \le r < 0$ είναι αρνητικά συσχετισμένες. Είναι $r = \pm 1$ όταν όλα τα σημεία βρίσκονται πάνω σε ευθεία. Για να είναι το r κοντά στο 0, πρέπει τα σημεία να είναι το ίδιο κατανεμημένα προς όλες τις διευθύνσεις ($s_x = s_y$), ή να εκτείνονται παράλληλα προς τον οριζόντιο άξονα ($s_x > s_y$) ή να εκτείνονται παράλληλα προς τον κατακόρυφο άξονα ($s_x < s_y$).



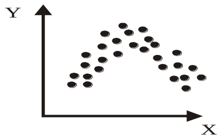
Από τα στοιχεία του πίνακα των X και Y και χρησιμοποιώντας τον τύπο (2) υπολογίζουμε για το παράδειγμά μας την τιμή του συντελεστή συσχέτισης

$$r = \frac{2146 - \frac{1}{15} \cdot 163 \cdot 179}{\sqrt{1959 - \frac{163^2}{15}} \cdot \sqrt{2427 - \frac{179^2}{15}}} = 0,86 .$$

Να σημειωθεί ότι ο συντελεστής συσχέτισης δεν έχει μονάδα μέτρησης, δηλαδή οι τιμές του είναι αριθμοί χωρίς διάσταση. Στο παράδειγμά μας η τιμή r = 0,86 δείχνει μια ισχυρή θετική συσχέτιση μεταξύ των επιδόσεων των μαθητών στα δύο μαθήματα.

Παρατηρήσεις

1) Όσο η μορφή του σμήνους των σημείων αποκλίνει από την γραμμική, τόσο ο συντελεστής συσχέτισης υποεκτιμά το βαθμό συσχέτισης μεταξύ των μεταβλητών



επειδή είναι προορισμένος να μετράει το βαθμό συσχέτισης όταν αυτή είναι γραμμική. Αν π.χ. το σμήνος των σημείων έχει τη μορφή του σχήματος αυτού, ο r υποεκτιμά τον βαθμό συσχέτισης.

2) Η συσχέτιση μεταξύ δύο μεταβλητών δεν πρέπει να μας οδηγεί στο συμπέρασμα ότι μεταξύ των μεταβλητών υπάρχει αιτιατή σχέση, δηλαδή ότι η μία προκαλεί την άλλη.

§ 5. 2. Ευθεία παλινδρόμησης ή ελαχίστων τετραγώνων

Από το διάγραμμα διασποράς του προηγουμένου παραδείγματος φαίνεται ότι υπάρχει μία σχέση ανάμεσα στα μαθήματα X και Y. Τα σημεία (x, y) είναι συγκεντρωμένα περίπου γύρω από μια ευθεία, (την ευθεία που φαίνεται στο διάγραμμα διασποράς), δηλαδή η συσχέτιση των X και Y είναι κατά προσέγγιση γραμμική.

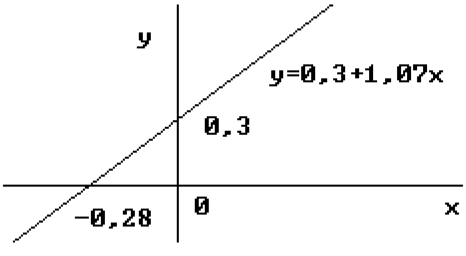
Αν θεωρήσουμε το μάθημα X ως ανεξάρτητη μεταβλητή και το μάθημα Y ως εξαρτημένη, τότε η ευθεία που ‘προσαρμόζεται’ καλύτερα στα σημεία αυτά είναι η <u>ευθεία παλινδρόμησης</u>. Λέγεται επίσης και <u>ευθεία ελαχίστων τετραγώνων</u> διότι έχει την ιδιότητα να ελαχιστοποιείται το άθροισμα των τετραγώνων των κατακόρυφων αποστάσεων των σημείων (x_i, y_i) από την ευθεία αυτή.

Η εξίσωση της ευθείας αυτής όπως γνωρίζουμε είναι της μορφής: $y = a + bx$. Οι παράμετροι a και b που καθορίζουν την εξίσωση της ευθείας παλινδρόμησης λέγονται <u>εκτιμήτριες ελαχίστων τετραγώνων</u> και υπολογίζονται από τις σχέσεις:

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\cdot\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2}, \text{ και } a = \bar{y} - b\bar{x}, \text{ όπου } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,\ \ \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

Έτσι, αν αντικαταστήσουμε στους παραπάνω τύπους τα δεδομένα του πίνακα των μαθημάτων X και Y βρίσκουμε:

$b = \dfrac{15\cdot 2146 - 163\cdot 179}{15\cdot 1959 - 163^2} = 1{,}07$ και $a = \dfrac{179}{15} - 1{,}07\cdot\dfrac{163}{15} = 0{,}3$. Συνεπώς η εξίσωση της ευθείας είναι: $y = 0{,}3 + 1{,}07x$ και η γραφική της παράσταση, όπως προκύπτει με το μαθηματικό πρόγραμμα Derive for Windows, είναι η διπλανή, η οποία βέβαια συμπίπτει με αυτή του διαγράμματος διασποράς που έγινε με το στατιστικό πρόγραμμα SPSS 14.



Ασκήσεις

1) Με τη μέθοδο των ελαχίστων τετραγώνων να βρεθεί η ευθεία γραμμικής παλινδρόμησης της y πάνω στη x για τα παρακάτω δεδομένα ζευγών (x, y) πέντε μαθητών, όπου x εκφράζει την επίδοση στα Μαθηματικά και y την επίδοση στη Φυσική: (12,13), (15,14), (16,18), (18,18) και (18,20). Αν υποθέσουμε ότι χάθηκε ο βαθμός της Φυσικής για τον μαθητή που πήρε 15 στα Μαθηματικά, για να μην υποχρεωθεί ο μαθητής να ξαναδώσει εξετάσεις στη Φυσική, ποιο βαθμό πρέπει να πάρει;

2) Από 8 γάμους που έγιναν σε μια εκκλησία κατά τη διάρκεια ενός μηνός, οι ηλικίες των ανδρογύνων ήταν τα ζεύγη (x, y) όπου x η ηλικία της νύφης και y η ηλικία του γαμπρού: (20,20), (22,20), (24,22), (25,27), (28,24), (30, 25), (33,28), (38,34). α) Να υπολογιστεί με τη μέθοδο των ελαχίστων τετραγώνων η ευθεία γραμμικής παλινδρόμησης της y πάνω στη x. β) Να βρεθεί η αναμενόμενη ηλικία του γαμπρού για μια υποψήφια νύφη 25 ετών. γ) Για κάθε έτος που μια γυναίκα καθυστερεί να παντρευτεί, πόσο αυξάνεται η ηλικία του υποψήφιου γαμπρού; δ) Για τα ίδια δεδομένα, να βρεθεί η ευθεία γραμμικής παλινδρόμησης της x πάνω στην y. ε) την αναμενόμενη ηλικία της νύφης για έναν υποψήφιο γαμπρό 28 ετών. ζ) Για κάθε έτος που ένας άνδρας καθυστερεί να παντρευτεί, πόσο αυξάνεται η ηλικία της υποψήφιας νύφης;

3) Σε μια εξέταση οκτώ μαθητών στα Μαθηματικά από 2 εξεταστές Α και Β η βαθμολογία ήταν ως ακολούθως: (55,54), (62,56), (71,61), (66,66), (63,63), (56,61), (72,73), 51,54). Να εξεταστεί αν υπάρχει γραμμική συσχέτιση μεταξύ της βαθμολογίας των δύο εξεταστών.

4) Ένας μαθητής γνώριζε ότι η σχέση που συνδέει τους βαθμούς Φαρενάιτ (F) με τους βαθμούς Κελσίου (C) είναι γραμμική, δηλαδή $F = \alpha + \beta C$. Επειδή όμως δε θυμόταν τις σταθερές *α, β,* μέτρησε τη θερμοκρασία του δωματίου του σε πέντε διαφορετικές ώρες με δύο θερμόμετρα, με κλίμακα σε F^o και C^o αντίστοιχα, και πήρε τα παρακάτω ζεύγη τιμών (C, F): (15,59), (20,68), (25,77), (30,86) και (35,95). Να βρεθούν τα α και β της σχέσης που συνδέει τις δύο κλίμακες.

5) Να βρεθεί ο συντελεστής συσχέτισης για 4 ζεύγη (x_i, y_i) (i = 1, 2, 3, 4) παρατηρήσεων, αν είναι γνωστό ότι έχουμε: $\bar{x} = 7$, $\bar{y} = 4{,}5$, $\sum x_i^2 = 210$, $\sum y_i^2 = 92$ και $\sum x_i y_i = 138$.

Κεφάλαιο 6.

Έλεγχοι στατιστικών υποθέσεων

§ 6. 1. Γενικές αρχές

Οι έλεγχοι στατιστικών υποθέσεων ή αλλιώς στατιστικές δοκιμασίες έχουν ως σκοπό να ελέγξουν, ξεκινώντας από δειγματικά δεδομένα, την εγκυρότητα (ορθότητα) ορισμένων υποθέσεων που σχετίζονται με ένα ή περισσότερους πληθυσμούς. Στις πιο πολλές περιπτώσεις οι υποθέσεις αφορούν παραμέτρους πληθυσμών.

Κατά ένα γενικό τρόπο διατυπώνουμε στην αρχή μια στατιστική υπόθεση (της οποίας θέλουμε να ελέγξουμε την εγκυρότητα) που συμβολίζεται με $\mathbf{H_0}$ και ονομάζουμε μηδενική υπόθεση. Μετρούμε κατόπιν την απόκλιση μεταξύ χαρακτηριστικών του δείγματος και του πληθυσμού ή μεταξύ χαρακτηριστικών των δειγμάτων, και υπολογίζουμε την πιθανότητα, θεωρώντας έγκυρη την H_0, να βρεθεί μια απόκλιση το ίδιο ή πιο σημαντική από την παρατηρούμενη.

Αν αυτή η πιθανότητα είναι μεγαλύτερη από μια πιθανότητα *a* που ορίζουμε από πριν και ονομάζεται επίπεδο σημαντικότητας, δεχόμαστε τη μηδενική υπόθεση ως έγκυρη. Στην περίπτωση αυτή θεωρούμε ότι η παρατηρούμενη απόκλιση οφείλεται στο τυχαίο της δειγματοληψίας και δεν ισχύει σε επίπεδο πληθυσμού.

Αν αντίθετα η πιθανότητα αυτή είναι μικρότερη από το επίπεδο σημαντικότητας *a* απορρίπτουμε την μηδενική υπόθεση και δεχόμαστε την αλήθεια μιας άλλης υπόθεσης που συμβολίζεται με $\mathbf{H_1}$ και ονομάζεται εναλλακτική υπόθεση. Στην τελευταία αυτή περίπτωση θεωρούμε ότι η παρατηρούμενη απόκλιση ελάχιστα οφείλεται στο τυχαίο της δειγματοληψίας και συνεπώς ισχύει σε επίπεδο πληθυσμού.

Το επίπεδο σημαντικότητας των στατιστικών ελέγχων καθορίζεται συνήθως σε $a = 0{,}05$ ή $a = 0{,}01$ ή $a = 0{,}001$. Πρέπει να σημειωθεί ότι οι παρατηρούμενες αποκλίσεις αποτελούν στατιστικά των οποίων οι δειγματοληπτικές κατανομές ακολουθούν γνωστές κατανομές όπως οι κατανομές Z, t, F, X^2 με την προϋπόθεση ότι η μηδενική υπόθεση ισχύει.

Έτσι, το επίπεδο σημαντικότητας *a* ορίζει συγκεκριμένες τιμές των παραπάνω κατανομών που ονομάζονται κρίσιμες τιμές. Οι κρίσιμες τιμές ορίζουν με τη σειρά τους τις περιοχές αποδοχής ή απόρριψης της μηδενικής υπόθεσης. Συνεπώς, δεν απαιτείται να υπολογίζουμε πιθανότητες και να τις συγκρίνουμε με το α για να δούμε αν δεχόμαστε ή όχι την μηδενική υπόθεση, αλλά συγκρίνουμε την τιμή του στατιστικού (που εκφράζει την παρατηρούμενη απόκλιση) με την κρίσιμη τιμή (που ορίζεται από το α). Αν είναι μεγαλύτερη από την κρίσιμη τιμή (κατ' απόλυτη τιμή) απορρίπτουμε την H_0, αν είναι μικρότερη δεχόμαστε την H_0. Στην πρώτη περίπτωση η τιμή του στατιστικού θα βρίσκεται στην περιοχή απόρριψης ενώ στη δεύτερη στην περιοχή αποδοχής.

§ 6. 2. Σφάλματα τύπου Ι και ΙΙ. – Ισχύς ελέγχου

Στους στατιστικούς ελέγχους και λόγω του παράγοντα «τύχη» που επενεργεί κατά τη λήψη των δειγμάτων μπορούν να συμβούν δύο ειδών σφάλματα:

1) Το σφάλμα τύπου Ι (πρώτου είδους) αφορά την απόρριψη της μηδενικής υπόθεσης ενώ αυτή είναι έγκυρη. Η πιθανότητα να συμβεί το σφάλμα αυτό ονομάζεται κίνδυνος σφάλματος τύπου Ι και το ανώτερο όριό του ισούται με το επίπεδο σημαντικότητας α. Αν δηλαδή το $\alpha = 0{,}05$ τόσο είναι το ανώτερο όριο του κινδύνου σφάλματος τύπου Ι.

2) Το σφάλμα τύπου ΙΙ (δευτέρου είδους) αφορά την αποδοχή της μηδενικής υπόθεσης ενώ είναι λανθασμένη. Η πιθανότητα του σφάλματος αυτού ονομάζεται κίνδυνος σφάλματος τύπου ΙΙ και συμβολίζεται με β.

Ας σημειωθεί ότι υπάρχει η πιθανότητα να συμβεί το σφάλμα τύπου Ι μόνο όταν αποφασίσουμε να απορρίψουμε την μηδενική υπόθεση. Αν αντίθετα, η απόφαση οδηγεί στην μη απόρριψη της μηδενικής απόφασης τότε υπάρχει η πιθανότητα να συμβεί το σφάλμα τύπου ΙΙ.

Τα δύο σφάλματα είναι ανταγωνιστικά, δηλαδή όσο μειώνεται το α τόσο αυξάνει το β. Ισχύς ενός ελέγχου είναι η πιθανότητα να απορρίψουμε τη μηδενική υπόθεση H_0 ενώ είναι πράγματι λανθασμένη και ισούται με $1 - \beta$.

Συνήθως καθορίζουμε από πριν σε έναν έλεγχο το α και για να μικρύνουμε το β επιλέγουμε δείγματα μεγάλου μεγέθους. Έτσι, αυξάνει η ισχύς του ελέγχου.

Με το παρακάτω παράδειγμα που αποτελεί μία από τις βασικές στατιστικές δοκιμασίες θα γίνουν κατανοητές οι αρχές των στατιστικών ελέγχων και θα δοθεί η ευκαιρία να γίνει ο διαχωρισμός τους σε μονόπλευρους και δίπλευρους.

§ 6. 3. Έλεγχος συμφωνίας του μέσου όρου

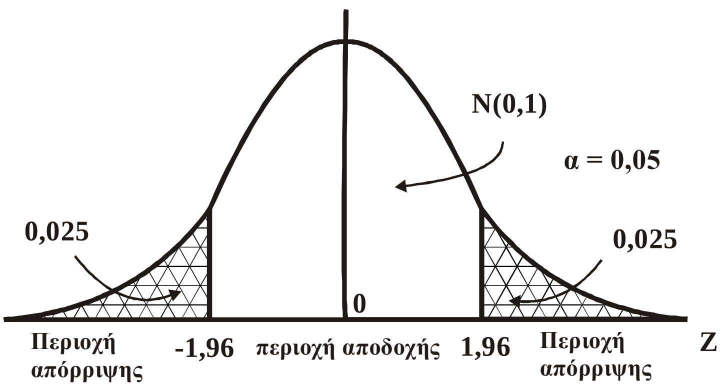
Από ένα πληθυσμό που ακολουθεί κανονική κατανομή ως προς ένα χαρακτηριστικό, παίρνουμε ένα τυχαίο δείγμα μεγέθους *n* και βρίσκουμε ότι ο δειγματικός μέσος όρος είναι $\bar{x}$. Έστω ότι η διακύμανση του πληθυσμού σ^2 είναι γνωστή. Σκοπός του ελέγχου αυτού είναι να ελεγχθεί βάσει της τιμής $\bar{x}$ του δείγματος, αν ο μέσος όρος μ του πληθυσμού (που είναι άγνωστος) ισούται ή όχι με μια τιμή μ_0. Οι δύο υποθέσεις, η μηδενική και η εναλλακτική, είναι οι εξής:

Μηδενική $H_0 : \mu = \mu_0$

Εναλλακτική $H_1 : \mu \neq \mu_0$

Γνωρίζουμε (παρ. 4.1) ότι η τυχαία μεταβλητή $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ ακολουθεί την κατανομή N(0, 1). Θεωρούμε στην αρχή ότι η H_0 είναι έγκυρη. Συνεπώς η τυχαία μεταβλητή $Z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ ακολουθεί την N(0, 1).

Έστω *α* το επίπεδο σημαντικότητας του ελέγχου που αποτελεί και το ανώτερο όριο του κινδύνου σφάλματος τύπου Ι. Αν $|Z| < Z_{1-\alpha/2}$ δεχόμαστε την H_0 ως έγκυρη. Διαφορετικά την απορρίπτουμε δεχόμενοι την αλήθεια της εναλλακτικής υπόθεσης H_1. Στην πρώτη περίπτωση το $|Z| < Z_{1-\alpha/2}$ σημαίνει ότι η τιμή *Z* βρίσκεται μεταξύ των τιμών $Z_{\alpha/2}$ και $Z_{1-\alpha/2}$. Αυτές οι τιμές ορίζουν την περιοχή αποδοχής της H_0 και την περιοχή απόρριψης. Έτσι στη δεύτερη περίπτωση, αν το *Z* είναι μεγαλύτερο του $Z_{1-\alpha/2}$ ή μικρότερο του $Z_{\alpha/2}$, δηλαδή βρίσκεται στην περιοχή απόρριψης, απορρίπτουμε την H_0 δεχόμενοι την H_1. Αν απορρίψουμε την H_0 σημαίνει ότι η απόκλιση μεταξύ $\bar{X}$ και μ_0 λίγο θα οφείλεται στο τυχαίο της δειγματοληψίας αλλά θα είναι συστηματική, συνεπώς θεωρούμε ότι $\mu \neq \mu_0$. Αν δεχθούμε την H_0 σημαίνει ότι η απόκλιση θα οφείλεται στον παράγοντα

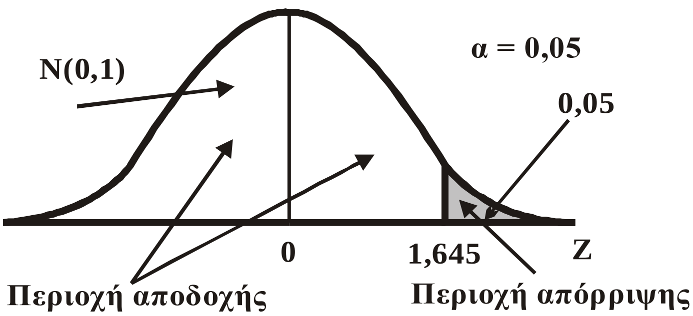


'τύχη', έτσι θεωρούμε ότι $\mu = \mu_0$. Οι τιμές $Z_{1-\alpha/2}$ και $Z_{\alpha/2}$ είναι οι κρίσιμες τιμές του ελέγχου και υπολογίζονται από τον πίνακα της τυποποιημένης κατανομής Z δηλαδή της N(0, 1).

Έτσι, για α = 0,05 είναι $Z_{1-\alpha/2} = Z_{0,975} = 1{,}96$ και $Z_{\alpha/2} = Z_{0,025} = -1{,}96$. Συνεπώς, για α = 0,05 οι κρίσιμες τιμές είναι 1,96 και –1,96. Το παραπάνω σχήμα δείχνει τις κρίσιμες τιμές καθώς και τις περιοχές απόρριψης και αποδοχής της H_0. Δεξιά της τιμής 1,96 η πιθανότητα είναι 0,025 και αριστερά της – 1,96 είναι πάλι 0,025 δηλαδή το μισό του α.

Με το ίδιο σκεπτικό αν α = 0,01, $Z_{1-\alpha/2} = \mathrm{Z}_{0,995} = 2{,}58$, ενώ $Z_{\alpha/2} = Z_{0,005} = -2{,}58$ από τον πίνακα N(0, 1). Συνεπώς οι κρίσιμες τιμές είναι 2,58 και –2,58 για α = 0,01.

Ο έλεγχος που παρουσιάσαμε μέχρι τώρα ονομάζεται δίπλευρος γιατί η εναλλακτική υπόθεση H_1 ήταν $\mu \neq \mu_0$ και ο έλεγχος αφορούσε και τα δύο άκρα της κατανομής (βλέπε παραπάνω σχήμα). Θα μπορούσαμε όμως να έχουμε αντί για $\mu \neq \mu_0$ την $\mu > \mu_0$ ή την $\mu < \mu_0$. Σε αυτή την περίπτωση η πιθανότητα δεν είναι μοιρασμένη σε δύο μέρη αλλά βρίσκεται στο ένα άκρο της κατανομής και ο έλεγχος ονομάζεται μονόπλευρος.



Πρέπει να τονιστεί ότι στους μονόπλευρους ελέγχους οι κρίσιμες τιμές είναι $Z_{1-\alpha}$ και Z_α. Έτσι, αν α = 0,05 τότε $Z_{1-\alpha} = Z_{0,95} = 1{,}64$ (σχήμα), ενώ αν α = 0,01 είναι $Z_{1-\alpha} = Z_{0,99} = 2{,}33$ θεωρώντας πάντοτε το δεξιό άκρο της κατανομής. Αν είμαστε στο αριστερό θεωρούμε τις ίδιες αυτές τιμές αλλά με αρνητικό πρόσημο. Στο δεξιό άκρο θα είμαστε όταν $H_1 : \mu > \mu_0$, ενώ στο αριστερό όταν $H_1 : \mu < \mu_0$.

Γενική παρατήρηση

Όπως στον παραπάνω έλεγχο της συμφωνίας του μέσου όρου, έτσι και σε οποιονδήποτε άλλο έλεγχο όπου το στατιστικό του ελέγχου ακολουθεί Z κατανομή, δηλαδή την N(0, 1), ενεργούμε ως εξής: Υπολογίζουμε την τιμή του στατιστικού Z (π.χ. $Z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$) και παίρνουμε την απόλυτη τιμή της, δηλαδή την θεωρούμε πάντα θετική. Κατόπιν τη συγκρίνουμε με τις κρίσιμες τιμές. Αν είναι μεγαλύτερη απορρίπτουμε την H_0 αν είναι μικρότερη τη δεχόμαστε. Οι κρίσιμες τιμές είναι:

***Για α = 0,05**: Δίπλευρος έλεγχος: 1,96, Μονόπλευρος έλεγχος: 1,64*

***Για α = 0,01**: Δίπλευρος έλεγχος: 2,58, Μονόπλευρος έλεγχος: 2,33*

Παράδειγμα 1

Έστω ότι στο μαθητικό πληθυσμό, η βαθμολογία ενός μαθήματος κατανέμεται κανονικά με τυπική απόκλιση $\sigma = 1{,}2$. Για το μάθημα αυτό υποστηρίχθηκε ότι ο μέσος όρος επίδοσης είναι 6. Για να ελεγχθεί αν ο ισχυρισμός αυτός ευσταθεί, πάρθηκε μέσα από τον πληθυσμό τυχαίο δείγμα μεγέθους $n = 100$ μαθητών και βρέθηκε ότι η μέση επίδοση στο μάθημα ήταν 5,6. Με βάση την τιμή αυτή να ελεγχθεί η αλήθεια του ισχυρισμού.

Λύση

Έστω ότι το επίπεδο σημαντικότητας είναι $\alpha = 0{,}05$. Θεωρούμε ότι ο έλεγχος είναι δίπλευρος. Οι δύο υποθέσεις θα είναι:

$$H_0 : \mu = 6, \quad H_1 : \mu \neq 6$$

Υπολογίζουμε (θεωρώντας κατ' αρχή ότι ισχύει η H_0) την ποσότητα:

$Z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{5{,}6 - 6}{1{,}2/\sqrt{100}} = -3{,}33$. Άρα η απόλυτη τιμή $|Z| = |-3{,}33| = 3{,}33$.

Αφού έχουμε δίπλευρο έλεγχο και $\alpha = 0{,}05$ η κρίσιμη τιμή είναι 1,96. Επειδή 3,33 είναι μεγαλύτερο του 1,96 απορρίπτουμε την μηδενική υπόθεση. Αν θεωρούσαμε ως $\alpha = 0{,}01$ η κρίσιμη τιμή είναι 2,58. Πάλι είναι 3,33 > 2,58 συνεπώς απορρίπτεται και πάλι η H_0. Αυτό σημαίνει ότι ο μέσος όρος επίδοσης δεν είναι 6 και επειδή το $\bar{x} = 5{,}6$ συμπεραίνουμε ότι ο μέσος όρος επίδοσης είναι μικρότερος της τιμής 6.

Το ίδιο αποτέλεσμα (απόρριψη της H_0) συνάγεται επίσης και από ένα μονόπλευρο έλεγχο $(H_1 : \mu < 6)$. Πράγματι, στο μονόπλευρο έλεγχο η κρίσιμη τιμή για $\alpha = 0{,}05$ είναι 1,64 ενώ για $\alpha = 0{,}01$ είναι 2,33. Και για τα δύο επίπεδα σημαντικότητας, η τιμή 3,33 είναι μεγαλύτερη, συνεπώς και στην περίπτωση αυτή απορρίπτουμε την H_0 και δεχόμαστε την $H_1 : \mu < 6$.

Στην ίδια στατιστική δοκιμασία δηλαδή στον έλεγχο συμφωνίας του μέσου όρου, αν η τυπική απόκλισης σ του πληθυσμού είναι άγνωστη μπορούμε να την εκτιμήσουμε από την τυπική απόκλιση s του δείγματος. Σε αυτήν όμως την περίπτωση το στατιστικό $\dfrac{\bar{x}-\mu}{s/\sqrt{n}}$ δεν ακολουθεί Z κατανομή αλλά t - κατανομή (Student) με $n-1$ βαθμούς ελευθερίας. (Β. Ε.) όπου n το μέγεθος του δείγματος.

Στην περίπτωση αυτή ενεργούμε ως εξής: υπολογίζουμε την τιμή $t=\dfrac{\bar{x}-\mu_0}{s/\sqrt{n}}$ και παίρνουμε πάντα την απόλυτη τιμή της δηλαδή το $|t|$. Στον δίπλευρο έλεγχο συγκρίνουμε την τιμή $|t|$ με την κρίσιμη τιμή $t_{\alpha/2}$ για n -1 Β. Ε. Στο μονόπλευρο έλεγχο συγκρίνουμε την $|t|$ με την κρίσιμη τιμή t_α για $n-1$ Β. Ε. Αν $|t|$ είναι μικρότερη από την κρίσιμη τιμή δεχόμαστε την μηδενική υπόθεση, διαφορετικά την απορρίπτουμε. Οι κρίσιμες τιμές βρίσκονται από τον πίνακα της t – κατανομής (σελ. 79).

Παράδειγμα 2

Ας θεωρήσουμε το προηγούμενο παράδειγμα, όπου όμως η τυπική απόκλιση σ είναι άγνωστη. Έτσι έστω ότι πάρθηκε από τον πληθυσμό τυχαίο δείγμα μεγέθους n = 20 και βρέθηκε ότι $\bar{x}=5{,}8$ και s = 1,4. Να ελεγχθεί η υπόθεση ότι για τον πληθυσμό αυτόν ο μέσος όρος επίδοσης στο μάθημα ισούται με την τιμή 6.

Λύση

Έστω ότι α = 0,05 και ο έλεγχος είναι δίπλευρος δηλαδή: H_0 : μ = 6, H_1: $\mu \neq 6$.

Είναι: $t=\dfrac{\bar{x}-\mu_0}{s/\sqrt{n}}=\dfrac{5{,}8-6}{1{,}4/\sqrt{20}}=-0{,}64$. Άρα $|t|=|-0{,}64|=0{,}64$.

Οι βαθμοί ελευθερίας είναι $n-1$ = 20 – 1 = 19. Για 19 βαθμούς ελευθερίας και α = 0,05 είναι $t_{\alpha/2}=t_{0{,}025}=2{,}023$. Αυτή την τιμή διαβάζουμε στον πίνακα της κατανομής t, στη διασταύρωση της γραμμής ν = 19 και της στήλης 0,025. Επειδή το 0,64 είναι μικρότερο από την κρίσιμη τιμή 2,093 δεν έχουμε λόγους να απορρίψουμε τη μηδενική υπόθεση, άρα θεωρούμε ότι πράγματι μ = 6.

Αν ο έλεγχος ήταν μονόπλευρος (H_1: μ < 6) το 0,64 συγκρίνεται με την κρίσιμη τιμή $t_\alpha=t_{0{,}05}=1{,}729$. Η τιμή αυτή βρίσκεται στη διασταύρωση της γραμμής ν = 19 και της στήλης 0,05. Επειδή πάλι 0,64 < 1,729 δεχόμαστε την H_0 ως έγκυρη, θεωρώντας ότι η παρατηρούμενη απόκλιση οφείλεται στο τυχαίο της δειγματοληψίας.

Παρατήρηση

Όταν το μέγεθος του δείγματος είναι μεγάλο και ειδικά όταν $n > 30$ η προϋπόθεση της κανονικότητας του πληθυσμού είναι δευτερεύουσας σημασίας. Αυτή η διαπίστωση απορρέει από το κεντρικό οριακό θεώρημα (βλ. παρ. 4.1). Είναι γνωστό εξ' άλλου (βλ. 3.11), ότι για βαθμούς ελευθερίας μεγαλύτερους από 30, η κατανομή t προσεγγίζει ικανοποιητικά την τυποποιημένη κατανομή *Ζ*.

§ 6. 4. Έλεγχος ισότητας δύο μέσων τιμών σε ανεξάρτητα δείγματα

(A)

Θεωρούμε δύο πληθυσμούς οι οποίοι κατανέμονται κανονικά με διακυμάνσεις $\sigma_1{}^2$ και $\sigma_2{}^2$ που είναι άγνωστες αλλά ίσες $(\sigma_1{}^2 = \sigma_2{}^2)$. Μέσα από κάθε πληθυσμό παίρνουμε από ένα δείγμα μεγέθους n_1 και n_2 αντίστοιχα. Τα δύο αυτά δείγματα λαμβάνονται με τυχαίο τρόπο και ανεξάρτητα το ένα από το άλλο. Ο σκοπός του ελέγχου αυτού είναι να ελέγξει την υπόθεση της ισότητας των άγνωστων μέσων όρων μ_1 και μ_2 των δύο πληθυσμών. Οι δύο υποθέσεις H_0 και H_1 είναι οι εξής:

$H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$ (δίπλευρος έλεγχος).

Με την προϋπόθεση ότι η μηδενική υπόθεση H_0 ισχύει, αποδεικνύεται ότι το στατιστικό:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \cdot \left[\frac{(n_1 - 1)s_1{}^2 + (n_2 - 1)s_2{}^2}{n_1 + n_2 - 2}\right]}} \qquad (1)$$

ακολουθεί t – κατανομή με $n_1 + n_2 - 2$ βαθμούς ελευθερίας. Με $\bar{x}_1, s_1{}^2$ και $\bar{x}_2, s_2{}^2$ συμβολίζονται οι μέσοι όροι και οι διακυμάνσεις αντίστοιχα των δύο δειγμάτων.

Η ποσότητα $\left[\frac{(n_1 - 1)s_1{}^2 + (n_2 - 1)s_2{}^2}{n_1 + n_2 - 2}\right]$ αποτελεί εκτίμηση της άγνωστης κοινής διακύμανσης των δύο πληθυσμών.

Έστω α το επίπεδο σημαντικότητας του ελέγχου. Στο δίπλευρο έλεγχο η κρίσιμη τιμή είναι $t_{\alpha/2}$. Αν $|t| > t_{\alpha/2}$ απορρίπτουμε την H_0 και δεχόμαστε ως έγκυρη την εναλλακτική υπόθεση H_1. Διαφορετικά ($|t| < t_{\alpha/2}$) δεχόμαστε την μηδενική υπόθεση H_0. Αν έχουμε μονόπλευρο έλεγχο δηλαδή αν $H_1 : \mu_1 < \mu_2$ ή $H_1 : \mu_1 > \mu_2$ ακολουθούμε το ίδιο σκεπτικό αλλά τη φορά αυτή συγκρίνουμε την τιμή $|t|$ με την κρίσιμη τιμή t_α.

Παράδειγμα 1

Θέλουμε να συγκρίνουμε δύο διαφορετικές μεθόδους διδασκαλίας. Για το λόγο αυτό παίρνουμε δύο τυχαία και ανεξάρτητα δείγματα μεγέθους το ένα 10 μαθητών και το άλλο 13 μαθητών. Στο πρώτο δείγμα εφαρμόζουμε την πρώτη μέθοδο διδασκαλίας και στο δεύτερο δείγμα τη δεύτερη μέθοδο. Μετά πάροδο ενός εξαμήνου μετρούμε τις επιδόσεις τους και βρίσκουμε ότι στο πρώτο δείγμα ο μέσος όρος βαθμολογίας είναι $\bar{x}_1 = 6{,}8$ με διακύμανση $s_1^2 = 1{,}1$ ενώ στο δεύτερο δείγμα ο μέσος όρος είναι $\bar{x}_2 = 6{,}3$ με διακύμανση $s_2^2 = 1$. Να ελεγχθεί σε επίπεδο σημαντικότητας α = 0,05 αν κατά μέσο όρο διαφέρουν οι επιδόσεις στις δύο μεθόδους διδασκαλίας.

Λύση

Θεωρούμε ότι οι επιδόσεις κάτω από τις δύο μεθόδους διδασκαλίας ακολουθούν σε επίπεδο πληθυσμών κανονική κατανομή με ίσες διακυμάνσεις. Η τελευταία αυτή προϋπόθεση φαίνεται να ισχύει γιατί οι διακυμάνσεις στα δύο δείγματα είναι σχεδόν ίσες. Για το παράδειγμα αυτό έχουμε συνεπώς:

$n_1 = 10$ $n_2 = 13$ $\bar{x}_1 = 6{,}8$ $\bar{x}_2 = 6{,}3$ $s_1^2 = 1{,}1$ $s_2^2 = 1$ και $H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 \neq \mu_2$.

Υπολογίζουμε την ποσότητα t αντικαθιστώντας τα παραπάνω δεδομένα στον (1):

$$t = \frac{6{,}8 - 6{,}3}{\sqrt{\left(\frac{1}{10} + \frac{1}{13}\right) \cdot \left[\frac{(10-1)\cdot 1{,}1 + (13-1)\cdot 1}{10+13-2}\right]}} = \frac{0{,}5}{0{,}43} = 1{,}163 .$$

Οι βαθμοί ελευθερίας είναι $n_1 + n_2 - 2 = 21$. Επειδή ο έλεγχος είναι δίπλευρος η κρίσιμη τιμή είναι $t_{\alpha/2}$. Για επίπεδο σημαντικότητας α = 0,05 και 21 βαθμούς ελευθερίας από τον πίνακα κατανομής t έχουμε $t_{\alpha/2} = t_{0{,}025} = 2{,}08$.

Επειδή 1,163 < 2,08 δεν έχουμε λόγους να απορρίψουμε τη μηδενική υπόθεση H_0. Αν ο έλεγχος ήταν μονόπλευρος ($H_1 : \mu_1 > \mu_2$), τότε η κρίσιμη τιμή θα ήταν $t_\alpha = t_{0,05} = 1{,}721$, οπότε πάλι 1,163 < 1,721. Συνεπώς η μηδενική υπόθεση της ισότητας των δύο μέσων τιμών είναι έγκυρη. Μπορούμε λοιπόν να συμπεράνουμε ότι δεν υπάρχει στατιστικά σημαντική διαφορά στις μέσες επιδόσεις των μαθητών με τις δύο μεθόδους διδασκαλίας. Η παρατηρούμενη απόκλιση μεταξύ των μέσων όρων $\bar{x}_1$ και $\bar{x}_2$ μπορεί να αποδοθεί στο τυχαίο της δειγματοληψίας και δεν είναι συστηματική.

Παρατήρηση

Έχει αποδειχθεί ότι ο στατιστικός αυτός έλεγχος που είναι γνωστός και ως t – test, παραμένει έγκυρος ακόμα και αν οι δύο πληθυσμοί δεν ακολουθούν κανονική κατανομή, αλλά οποιαδήποτε κατανομή περίπου συμμετρική. Αυτό αληθεύει ιδίως όταν τα δείγματά μας είναι μεγάλα δηλ. n_1, n_2 > 30. Επίσης, παραμένει έγκυρος ακόμα και όταν η δεύτερη προϋπόθεση της ισότητας των δύο διακυμάνσεων $(\sigma_1^2 = \sigma_2^2)$ δεν ισχύει ακριβώς. Στην περίπτωση όμως αυτή πρέπει τα μεγέθη των δύο δειγμάτων να είναι ίσα δηλαδή $n_1 = n_2$.

(Β)

Ας διατηρήσουμε τώρα τις προϋποθέσεις της περίπτωσης (Α) θεωρώντας όμως ότι οι άγνωστες διακυμάνσεις σ_1^2 και σ_2^2 δεν είναι ίσες δηλαδή $\sigma_1^2 \neq \sigma_2^2$.

Στην περίπτωση που τα μεγέθη των δειγμάτων είναι μεγάλα δηλαδή n_1, n_2 > 30 και θεωρώντας τη μηδενική υπόθεση αληθινή ($\mu_1 = \mu_2$) αποδεικνύεται ότι η δειγματοληπτική κατανομή του στατιστικού:

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \qquad (2)$$

είναι προσεγγιστικά η τυποποιημένη κανονική κατανομή N(0, 1). Γνωρίζουμε τις κρίσιμες τιμές (βλ. γενική παρατήρηση παράγρ. 6.3) για α = 0,05 α = 0,01 για δίπλευρο ή μονόπλευρο έλεγχο. Αν $|Z|$ είναι μεγαλύτερη από την κρίσιμη τιμή, απορρίπτουμε τη μηδενική υπόθεση θεωρώντας έγκυρη την εναλλακτική. Αν $|Z|$ είναι μικρότερη, δεν έχουμε λόγους να μην θεωρήσουμε τη μηδενική υπόθεση έγκυρη.

Παράδειγμα 2

Σε δύο τυχαία δείγματα, το πρώτο μεγέθους n_1 = 40 και το δεύτερο n_2 = 60 ατόμων, μετρήθηκε ο χρόνος αντίδρασης σε δύο διαφορετικά ερεθίσματα Α και Β. Στο πρώτο δείγμα ο μέσος χρόνος αντίδρασης στο ερέθισμα Α ήταν $\bar{x}_1 = 2{,}3$ sec με διακύμανση $s_1^2 = 0{,}8$ ενώ στο δεύτερο ο μέσος χρόνος αντίδρασης στο ερέθισμα Β ήταν $\bar{x}_2 = 3{,}5$ sec με διακύμανση $s_2^2 = 1{,}9$. Θέλουμε να ελέγξουμε αν οι μέσοι χρόνοι στα δύο ερεθίσματα διαφέρουν στατιστικά σε επίπεδο σημαντικότητας α = 0,01.

Λύση

Στο παράδειγμα αυτό τα μεγέθη των δύο δειγμάτων διαφέρουν αρκετά. Εδώ είναι δύσκολη η παραδοχή της ισότητας των διακυμάνσεων σ_1^2 και σ_2^2. Συνεπώς δεν φαίνεται να πληρούνται οι προϋποθέσεις της παραγράφου (Α). Επειδή τα δύο δείγματα έχουν μέγεθος μεγαλύτερο του 30, ο έλεγχος μπορεί να γίνει μέσω του στατιστικού (2) της παραγράφου (Β). Είναι:

$H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$ (δίπλευρος έλεγχος)

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{2{,}3 - 3{,}5}{\sqrt{\frac{0{,}8}{40} + \frac{1{,}9}{60}}} = \frac{-1{,}2}{0{,}227} = -5{,}28 .$$

Για α = 0,01 και δίπλευρο έλεγχο η κρίσιμη τιμή είναι 2,58.

Επειδή $|Z| = |-5{,}28| = 5{,}28 > 2{,}58$ απορρίπτουμε την μηδενική υπόθεση δεχόμενοι ότι οι μέσοι χρόνοι αντίδρασης στα δύο ερεθίσματα διαφέρουν. Η αντίδραση στο ερέθισμα Α είναι πιο γρήγορη από αυτήν του ερεθίσματος Β.

Παρατήρηση

Η παραπάνω μεθοδολογία στηρίχθηκε στο γεγονός ότι τα μεγέθη των δειγμάτων ήταν μεγάλα δηλαδή n_1 και $n_2 > 30$. Στην περίπτωση που τα μεγέθη είναι μικρά, το στατιστικό (2) της παραγράφου (Β) μπορεί να θεωρηθεί ότι ακολουθεί προσεγγιστικά την t – κατανομή με βαθμούς ελευθερίας που δίνονται επίσης προσεγγιστικά από τον πλησιέστερο ακέραιο αριθμό της τιμής V όπου

$$V = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} .$$

Παράδειγμα 3

Ας θεωρήσουμε και πάλι το προηγούμενο παράδειγμα, αλλά, έστω ότι τα μεγέθη των δύο δειγμάτων είναι $n_1 = 10$ και $n_2 = 15$.

Λύση

Είναι $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^{\ 2}}{n_1} + \dfrac{s_2^{\ 2}}{n_2}}} = \dfrac{2{,}3 - 3{,}5}{\sqrt{\dfrac{0{,}8}{10} + \dfrac{1{,}9}{15}}} = \dfrac{-1{,}2}{0{,}455} = -2{,}637$. Οι βαθμοί ελευθερίας είναι:

$$V = \frac{\left(\dfrac{s_1^{\ 2}}{n_1} + \dfrac{s_2^{\ 2}}{n_2}\right)^2}{\dfrac{1}{n_1 - 1}\left(\dfrac{s_1^{\ 2}}{n_1}\right)^2 + \dfrac{1}{n_2 - 1}\left(\dfrac{s_2^{\ 2}}{n_2}\right)^2} = \frac{\left(\dfrac{0{,}8}{10} + \dfrac{1{,}9}{15}\right)^2}{\dfrac{1}{9}\cdot\left(\dfrac{0{,}8}{10}\right)^2 + \dfrac{1}{14}\cdot\left(\dfrac{1{,}9}{15}\right)^2} = 22{,}99$$

Ο πλησιέστερος ακέραιος του V είναι το 23, άρα οι βαθμοί ελευθερίας είναι 23.

Για $\alpha = 0{,}01$ και 23 βαθμούς ελευθερίας, λόγω του ότι ο έλεγχος είναι δίπλευρος, η κρίσιμη τιμή που βρίσκουμε από τον πίνακα της t – κατανομής είναι 2,807. Επειδή $|t| = |-2{,}637| = 2{,}637 < 2{,}807$ δεχόμαστε την μηδενική υπόθεση ως έγκυρη δηλαδή οι μέσοι χρόνοι αντίδρασης δεν διαφέρουν.

Για $\alpha = 0{,}05$ όμως, η κρίσιμη τιμή είναι 2,069. Στην περίπτωση αυτή και επειδή 2,637 > 2,069 μπορούμε να απορρίψουμε την υπόθεση της ισότητας και να δεχθούμε την εναλλακτική δηλαδή ότι $\mu_1 \neq \mu_2$.

§ 6. 5. Έλεγχος ισότητας δύο διακυμάνσεων

Παίρνουμε δύο ανεξάρτητα και τυχαία δείγματα από δύο πληθυσμούς κανονικά κατανεμημένους. Έστω n_1, n_2 τα μεγέθη των δύο δειγμάτων και $s_1^{\ 2}, s_2^{\ 2}$ οι παρατηρούμενες διακυμάνσεις στα δύο δείγματα. Ο έλεγχος αυτός έχει σκοπό να διαπιστώσει αν οι διακυμάνσεις των δύο πληθυσμών $\sigma_1^{\ 2}$ και $\sigma_2^{\ 2}$ είναι ίσες.

Συνεπώς η μηδενική υπόθεση είναι: $H_0 : \sigma_1^{\ 2} = \sigma_2^{\ 2}$

Με την προϋπόθεση ότι η H_0 ισχύει, το στατιστικό $s_1^{\ 2} / s_2^{\ 2}$ ακολουθεί την F κατανομή με $(n_1 - 1)$ και $(n_2 - 1)$ βαθμούς ελευθερίας. Αν οι διακυμάνσεις των δύο πληθυσμών είναι ίσες, ο λόγος $s_1^{\ 2} / s_2^{\ 2}$ δεν θα διαφέρει πολύ από την τιμή 1.

Στην πράξη σχηματίζουμε πάντοτε το λόγο s_A^2 / s_B^2 όπου s_A^2 είναι η μεγαλύτερη τιμή από τις s_1^2 και s_2^2, ενώ s_B^2 είναι η μικρότερη. Συνεπώς, θα είναι πάντα $F \geq 1$. Έστω α το επίπεδο σημαντικότητας του ελέγχου. Αν $F > F_\alpha$ απορρίπτουμε την H_0, διαφορετικά τη δεχόμαστε. Η κρίσιμη τιμή F_α προσδιορίζεται από τον πίνακα της F κατανομής με $n_A - 1$ και $n_B - 1$ βαθμούς ελευθερίας, όπου n_A το μέγεθος του δείγματος που αντιστοιχεί στη διακύμανση s_A^2 ενώ n_B το μέγεθος του δείγματος που αντιστοιχεί στη διακύμανση s_B^2. Ας σημειωθεί ότι ο έλεγχος είναι μονόπλευρος.

Παρατήρηση

Για την εγκυρότητα του ελέγχου αυτού είναι απαραίτητη η κανονικότητα των δύο πληθυσμών. Τονίζεται επίσης ότι τα δείγματα πρέπει να είναι ανεξάρτητα.

Παράδειγμα

Θέλουμε να συγκρίνουμε τις διακυμάνσεις των επιδόσεων σε δύο μαθήματα. Θεωρούμε ότι οι επιδόσεις στα δύο αυτά μαθήματα ακολουθούν κανονική κατανομή. Δύο δείγματα μεγέθους 10 και 12 μαθητών παίρνονται τυχαία και έστω 1,3 και 2,1 οι αντίστοιχες διακυμάνσεις των επιδόσεων στα δύο αυτά μαθήματα. Να χρησιμοποιηθεί ως επίπεδο σημαντικότητας του ελέγχου το $\alpha = 0{,}05$.

Λύση

Η μηδενική υπόθεση είναι H_0 : $\sigma_1^2 = \sigma_2^2$. Είναι $F = s_A^2 / s_B^2 = 2{,}1/1{,}3 = 1{,}615$.

Για $\alpha = 0{,}05$ με 9 και 11 βαθμούς ελευθερίας η κρίσιμη τιμή $F_{0,05}$ από τον πίνακα κατανομής F είναι 3,10. Επειδή $1{,}615 < 3{,}1$ δεν έχουμε λόγους να απορρίψουμε την H_0. Οι διακυμάνσεις των επιδόσεων στα δύο μαθήματα φαίνεται ότι είναι ίσες.

§ 6. 6. Έλεγχος ισότητας δύο μέσων τιμών σε συζευγμένα δείγματα

Στην προηγούμενη παράγραφο θεωρήσαμε ότι τα δύο δείγματα λαμβάνονται κατά τέτοιο τρόπο που να μην εμφανίζουν καμία σύνδεση ή συσχέτιση μεταξύ τους. Σε πολλές όμως περιπτώσεις είναι χρήσιμη η επιλογή δειγμάτων, έτσι ώστε κάθε άτομο του ενός δείγματος να είναι συσχετισμένο με κάποιο τρόπο με ένα άτομο του άλλου δείγματος. Συνήθως διακρίνουμε δύο περιπτώσεις συσχέτισης:

α) Η πρώτη συσχέτιση αφορά ορισμένους παράγοντες όπως το φύλο, η ηλικία, το I.Q., το μορφωτικό επίπεδο κ.τ.λ. Συλλέγουμε κατ' αρχήν ένα τυχαίο δείγμα μεγέθους *n* και κατόπιν, σε κάθε άτομο του δείγματος αντιστοιχούμε ένα όμοιό του ως προς ορισμένους παράγοντες (χαρακτηριστικά). Ο τρόπος αυτός συσχέτισης ονομάζεται εξίσωση κατά ζεύγη.

β) Η δεύτερη αναφέρεται στις επαναληπτικές μετρήσεις στα ίδια άτομα του δείγματος (μεγέθους *n*) όπου γίνονται δύο μετρήσεις. Στην περίπτωση αυτή, όπως φαίνεται, χρησιμοποιούμε ένα μόνο δείγμα.

Ο σκοπός των συζευγμένων δειγμάτων και στις δύο παραπάνω περιπτώσεις είναι η εξουδετέρωση ορισμένων εξωτερικών παραγόντων που είναι συσχετισμένοι με την εξαρτημένη μεταβλητή που μελετούμε, έτσι ώστε οι διαφορές που προκύπτουν να είναι αποτέλεσμα μόνο της ανεξάρτητης μεταβλητής. Έτσι, στο παράδειγμα 1 (παρ. 6.4) θα ήταν καλό να γινόταν εξίσωση κατά ζεύγη των μαθητών ως προς τις I.Q. επιδόσεις τους στο μάθημα που εξετάζονται. Συνεπώς, αν υπήρχε διαφορά μεταξύ των δύο μεθόδων διδασκαλίας (που είναι ο παράγοντας που μελετούμε) αυτή θα οφείλονταν κυρίως στον παράγοντα αυτόν. Φυσικά, όσο πιο πολλούς εξωτερικούς παράγοντες εξουδετερώνουμε, τόσο είμαστε σίγουροι ότι τα αποτελέσματά μας οφείλονται στον παράγοντα που μελετούμε.

Αν (x_i, y_i) είναι τα ζεύγη των τιμών (i = 1, 2, 3, … , n) θεωρούμε τις διαφορές $x_i - y_i = d_i$. Έστω $\bar{x}$ και $\bar{y}$ οι μέσοι όροι των x_i και y_i αντίστοιχα και $\bar{d}$ ο μέσος όρος των d_i. Θα είναι: $\bar{d} = \bar{x} - \bar{y}$. Αν μ_1 και μ_2 είναι οι μέσοι όροι των δύο μεταβλητών και δ ο μέσος όρος της μεταβλητής d (σε επίπεδο πληθυσμού) θα είναι επίσης: $\delta = \mu_1 - \mu_2$. Έστω ότι η μεταβλητή d ακολουθεί κανονική κατανομή με μέσο όρο, όπως είπαμε, δ και άγνωστη τυπική απόκλιση. Μπορούμε να εκτιμήσουμε αυτή την τυπική απόκλιση με την τυπική απόκλιση s_d των n τιμών d_i.

Η μηδενική υπόθεση θα είναι $H_0 : \mu_1 = \mu_2$, που σημαίνει $\mu_1 - \mu_2 = 0$. Δηλαδή τελικά

$H_0 : \delta = 0$.

Φαίνεται λοιπόν καθαρά ότι ο στατιστικός αυτός έλεγχος συμπίπτει με τον έλεγχο συμφωνίας του μέσου όρου (παράγραφος 6.3) για την μεταβλητή d.

Συνεπώς, με την προϋπόθεση ότι η H_0 ισχύει, το στατιστικό $t = \dfrac{\bar{d}}{s_d / \sqrt{n}}$ ακολουθεί *t* κατανομή με *n* – 1 βαθμούς ελευθερίας. Έστω *α* το επίπεδο σημαντικότητας. Στο δίπλευρο έλεγχο ($H_1 : \delta \neq 0$) η κρίσιμη τιμή είναι $t_{\alpha/2}$. Αν $|t| < t_{\alpha/2}$ δεχόμαστε την H_0 ως έγκυρη. Αν $|t| > t_{\alpha/2}$ απορρίπτουμε την H_0 αποδεχόμενοι την H_1. Στην περίπτωση μονόπλευρου ελέγχου ($H_1 : \delta > 0$ ή $H_1 : \delta < 0$) η κρίσιμη τιμή είναι t_α.

Παράδειγμα

Θέλουμε να διαπιστώσουμε αν η χορήγηση μιας ουσίας μειώνει το χρόνο αντίδρασης σε κάποιο ερέθισμα. Παίρνουμε ένα τυχαίο δείγμα 7 ατόμων και μετρούμε το χρόνο αντίδρασης στο ερέθισμα. Κατόπιν χορηγείται στα άτομα η υπό μελέτη ουσία και μετρείται πάλι ο χρόνος αντίδρασής τους. Οι χρόνοι αντίδρασης πριν και μετά τη χορήγηση είναι σε δέκατα δευτερολέπτου:

Πριν (*x*)	*Μετά* (*y*)	*Διαφορά* (*d*)
12	11	1
7	5	2
11	9	2
14	10	4
9	9	0
10	9	1
8	6	2

Λύση

Για τις 7 τιμές d_i θα είναι, μέσος όρος:

$\bar{d} = \frac{1}{7}\sum_{i=1}^{7} d_i = 1{,}714$ και διακύμανση:

$$s_d^{\,2} = \frac{1}{7-1}\left(\sum_{i=1}^{7} d_i^{\,2} - \frac{\left(\sum_{i=1}^{7} d_i\right)^2}{7} \right) = 1{,}57 .$$

Άρα η τυπική απόκλιση θα είναι $s_d = \sqrt{1{,}57} = 1{,}253$. Συνεπώς,

Η μηδενική υπόθεση $\quad H_0 : \delta = 0$

Η εναλλακτική υπόθεση $\quad H_1 : \delta > 0$

Στην H_1 είναι $\delta > 0$ διότι ελέγχουμε μείωση του χρόνου αντίδρασης, δηλαδή ο έλεγχος είναι μονόπλευρος. Αφού όπως είδαμε $\delta = \mu_1 - \mu_2$ αυτό σημαίνει ότι $\mu_1 > \mu_2$, δηλαδή μετά τη χορήγηση έχουμε κατά μέσο όρο μείωση του χρόνου αντίδρασης.

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} = \frac{1{,}714}{1{,}253/\sqrt{7}} = 3{,}6$$

Για επίπεδο σημαντικότητας *α* = 0,01 και *n* – 1 = 7 – 1 = 6 βαθμούς ελευθερίας, στο μονόπλευρο έλεγχο η κρίσιμη τιμή είναι $t_\alpha = t_{0{,}01} = 3{,}143$ (πίνακα κατανομής *t* σελ.79). Επειδή 3,6 > 3,143 η μηδενική υπόθεση απορρίπτεται. Συνεπώς, δεχόμαστε ότι $\delta > 0$ δηλαδή $\mu_1 > \mu_2$ που σημαίνει ότι κατά μέσο όρο ο χρόνος αντίδρασης πριν τη χορήγηση της ουσίας ήταν μεγαλύτερος από αυτόν μετά τη χορήγηση. Ας σημειωθεί ότι στο παραπάνω παράδειγμα είχαμε επαναληπτικές μετρήσεις στα ίδια άτομα.

§ 6. 7. Έλεγχος συμφωνίας μιας αναλογίας

Έστω ένας πληθυσμός τον οποίο μελετούμε ως προς ένα χαρακτηριστικό. Το χαρακτηριστικό αυτό χωρίζει τον πληθυσμό σε δύο τάξεις. Στην πρώτη ανήκουν τα άτομα του πληθυσμού που κατέχουν το χαρακτηριστικό και στη δεύτερη αυτά που δεν το κατέχουν. Πρόκειται συνεπώς για ένα διχοτομικό πληθυσμό. Παίρνουμε ένα τυχαίο δείγμα από τον πληθυσμό μεγέθους n και έστω x ο αριθμός των ατόμων που κατέχουν το υπό μελέτη χαρακτηριστικό. Η στατιστική αυτή δοκιμασία έχει σκοπό να ελέγξει αν η αναλογία p των ατόμων που κατέχουν το χαρακτηριστικό στον πληθυσμό συμφωνεί με μία τιμή p_0. Η μηδενική υπόθεση είναι $H_0 : p = p_0$.

Με την υπόθεση ότι η H_0 ισχύει, αποδεικνύεται ότι το στατιστικό

$$Z = \frac{(x \pm 0{,}5) - np_0}{\sqrt{np_0 \cdot (1 - p_0)}} \qquad (1)$$

ακολουθεί την τυποποιημένη κανονική κατανομή N(0, 1). Στον παραπάνω τύπο έχουμε $(x + 0{,}5)$ όταν $x < n \cdot p_0$ ενώ $(x - 0{,}5)$ όταν $x > n \cdot p_0$. Προϋπόθεση εφαρμογής του ελέγχου είναι όταν το p_0 είναι κοντά στο 0,5 να είναι $n > 25$. Όταν το p_0 είναι κοντά στο 0 ή στο 1 θα πρέπει $np_0 \cdot (1 - p_0) > 9$. Αν $|Z|$ μεγαλύτερο από την κρίσιμη τιμή (που είναι ανάλογη του επιπέδου σημαντικότητας και του αν ο έλεγχος είναι μονόπλευρος ή δίπλευρος), απορρίπτουμε την H_0, διαφορετικά τη δεχόμαστε ως έγκυρη.

Παράδειγμα

Σε δείγμα 1000 μαθητών τέθηκε η ερώτηση αν είναι ικανοποιημένοι ή όχι από τις συνθήκες που επικρατούν στο σχολείο τους. Από τους 1000 μαθητές οι 460 δήλωσαν ικανοποιημένοι. Να ελεγχθεί σε επίπεδο σημαντικότητας $\alpha = 0{,}05$ αν η αναλογία των ικανοποιημένων διαφέρει στατιστικά από αυτή των όχι ικανοποιημένων.

Λύση

Έστω p η αναλογία που δήλωσαν ικανοποιημένοι σε επίπεδο πληθυσμού. Τότε η αναλογία των μη ικανοποιημένων θα είναι $1 - p$. Αν οι δύο αναλογίες δεν διαφέρουν, θα πρέπει $p = 0{,}5$ οπότε $1 - p = 0{,}5$. Συνεπώς οι υποθέσεις είναι:

$H_0: \ p = 0{,}5$

$H_1: \ p \neq 0{,}5$

Βάσει του τύπου (1) και θεωρώντας την H_0 ως αληθινή έχουμε:

$$Z = \frac{(460 + 0{,}5) - 1000 \cdot 0{,}5}{\sqrt{1000 \cdot 0{,}5 \cdot 0{,}5}} = -2{,}5$$

Ο έλεγχος είναι δίπλευρος και α = 0,05 συνεπώς η κρίσιμη τιμή είναι 1,96. Επειδή $|Z| = |-2{,}5| = 2{,}5 > 1{,}96$ απορρίπτεται η μηδενική υπόθεση. Η αναλογία των ικανοποιημένων διαφέρει στατιστικά (για την ακρίβεια είναι μικρότερη) από αυτή των μη ικανοποιημένων μαθητών. Στον παραπάνω τύπο βάλαμε x + 0,5 μέσα στην παρένθεση γιατί $x < np_0$. Πράγματι $x = 460$ ενώ $np_0 = 1000 \cdot 0{,}5 = 500$.

Παρατήρηση

Η πρόσθεση (ή η αφαίρεση) του αριθμού 0,5 στο x γίνεται πάντοτε για να αυξηθεί η ακρίβεια του ελέγχου.

§ 6. 8. Έλεγχος ανεξαρτησίας δύο ποιοτικών μεταβλητών – δοκιμασία Χ2

Με τη στατιστική αυτή δοκιμασία ελέγχουμε αν δύο ποιοτικές μεταβλητές (χαρακτηριστικά) είναι στατιστικά ανεξάρτητες. Έστω ότι η πρώτη μεταβλητή έχει k κατηγορίες και η άλλη l κατηγορίες. Οι δύο αυτές μεταβλητές δημιουργούν μια ταξινόμηση του πληθυσμού (στον οποίο εφαρμόζονται) σε $k{\cdot}l$ κλάσεις.

Σε ένα τυχαίο δείγμα μεγέθους n παρμένο από τον πληθυσμό, τα δεδομένα παρουσιάζονται υπό μορφή ενός <u>πίνακα συνάφειας</u>. Σε έναν τέτοιο πίνακα οι k γραμμές είναι οι κατηγορίες της μίας μεταβλητής, ενώ οι l στήλες οι κατηγορίες της άλλης. Στη διασταύρωση μιας γραμμής i (i = 1, 2, 3, …, k) και μιας στήλης j (j = 1, 2, 3,…, l) βρίσκεται ο αριθμός των ατόμων ε_{ij} του δείγματος, που ανήκουν ταυτόχρονα στις κατηγορίες i και j. Συνεπώς ο πίνακας συνάφειας είναι ένας πίνακας συχνοτήτων ε_{ij} και $n = \sum_{ij} \varepsilon_{ij}$.

Στον παρακάτω πίνακα συνάφειας συμβολίζεται με Κ(i) το άθροισμα των συχνοτήτων της κατηγορίας (γραμμής) i ενώ με Κ(j) το άθροισμα των συχνοτήτων της κατηγορίας (στήλης) j.

		2η μεταβλητή						
		1	2	…	j	…	l	
1η μεταβλητή	1							K(1)
	2							K(2)
	...							…
	i				ε_{ij}			K(i)
	…							…
	k							K(k)
		K(1)	K(2)	…	K(j)	…	K(l)	n

Είναι $n=\sum_{i=1}^{k} K(i)$ και $n=\sum_{j=1}^{l} K(j)$. Το άθροισμα K(i) ισούται με τον αριθμό των ατόμων που ανήκουν στην κατηγορία i ενώ K(j) είναι ο αριθμός των ατόμων που ανήκουν στην κατηγορία j. Οι συχνότητες ε_{ij} ονομάζονται <u>εμπειρικές συχνότητες</u> ή και <u>παρατηρούμενες.</u>

Έστω τώρα p_{ij} η πιθανότητα να ανήκει κάποιο άτομο του πληθυσμού ταυτόχρονα στις κατηγορίες i και j. Έστω p_i η πιθανότητα να ανήκει στην κατηγορία i ενώ p_j η πιθανότητα να ανήκει στην κατηγορία j. Αποδεικνύεται ότι αν οι δύο ποιοτικές μεταβλητές είναι ανεξάρτητες, τότε ισχύει η εξής σχέση:

$$p_{ij} = p_i \cdot p_j$$

Οι παραπάνω πιθανότητες είναι φυσικά άγνωστες. Η πιθανότητα p_i μπορεί να εκτιμηθεί από την ποσότητα $K(i)/n$ ενώ η πιθανότητα p_j από την $K(j)/n$ όπου οι δύο αυτές ποσότητες είναι σχετικές συχνότητες.

Τελικά, η <u>αναμενόμενη συχνότητα</u> θ_{ij} δηλαδή ο αριθμός των ατόμων που θα έπρεπε να ανήκουν ταυτόχρονα στις κατηγορίες i και j αν οι δύο μεταβλητές ήταν ανεξάρτητες, μπορεί να εκτιμηθεί (βάσει των παραπάνω) από την ποσότητα

$$\theta_{ij} = \frac{K(i)\cdot K(j)}{n}. \qquad (1)$$

Συνεπώς, αν οι δύο υπό μελέτη μεταβλητές είναι ανεξάρτητες, οι εμπειρικές συχνότητες ε_{ij} λίγο θα διαφέρουν από τις αναμενόμενες θ_{ij}, ενώ αν είναι εξαρτημένες η απόκλιση θα είναι μεγάλη.

Η απόκλιση μετριέται από την ποσότητα:

$$X^2 = \sum_{i=1,j=1}^{k,l} \frac{(\varepsilon_{ij} - \theta_{ij})^2}{\theta_{ij}} \tag{2}$$

Στον στατιστικό αυτό έλεγχο θεωρούμε πάντοτε ότι η μηδενική υπόθεση H_0 είναι η ανεξαρτησία των δύο μεταβλητών, ενώ η εναλλακτική H_1 η εξάρτηση αυτών. Αποδεικνύεται ότι με την προϋπόθεση ότι ισχύει η μηδενική υπόθεση H_0 (άρα ισχύει η σχέση (1)), το στατιστικό (2) ακολουθεί προσεγγιστικά την κατανομή X^2 με:

$$(k-1)\cdot(l-1)$$

βαθμούς ελευθερίας όπου *k* και *l* είναι όπως είδαμε ο αριθμός των γραμμών και στηλών του πίνακα συνάφειας. Έστω *α* το επίπεδο σημαντικότητας του ελέγχου. Τότε η κρίσιμη τιμή είναι $x_\alpha^{\ 2}$. Για *α* = 0,05 ή *α* = 0,01 οι κρίσιμες τιμές δίνονται από τον πίνακα της κατανομής X^2. Αν η τιμή της ποσότητας (2) είναι μεγαλύτερη από $x_\alpha^{\ 2}$, απορρίπτουμε την μηδενική υπόθεση H_0 και δεχόμαστε την H_1. Διαφορετικά θεωρούμε την H_0 ως έγκυρη.

Παράδειγμα

		'Τύπος Σχολής'		
		Α	*Β*	*Γ*
'ΚΟΕ'	*Ι*	50	53	49
	ΙΙ	90	91	89
	ΙΙΙ	17	21	40

Τυχαίο δείγμα 500 μαθητών της Γ΄ τάξης Λυκείου ταξινομήθηκε ως προς δύο χαρακτηριστικά: α) το κοινωνικοοικονομικό επίπεδο των μαθητών (ΚΟΕ) και β) την επιθυμία τους να εισαχθούν σε διαφόρου τύπου πανεπιστημιακές σχολές. Η μεταβλητή '(ΚΟΕ)' περιλάμβανε τρεις κατηγορίες Ι, ΙΙ, ΙΙΙ, ενώ η μεταβλητή 'τύπος σχολής' επίσης τρεις κατηγορίες σχολών Α, Β, Γ. Τα δεδομένα συγκεντρώθηκαν στον διπλανό πίνακα συνάφειας.

Να ελεγχθεί σε επίπεδο σημαντικότητας *α* = 0,05 αν ο τύπος σχολής που επιθυμούν να εισαχθούν οι μαθητές είναι ανεξάρτητος ή όχι από το κοινωνικοοικονομικό τους επίπεδο.

Λύση

Από τον παραπάνω πίνακα υπολογίζουμε τα αθροίσματα των γραμμών και των στηλών. Κατόπιν, για κάθε εμπειρική συχνότητα υπολογίζουμε (εκτιμούμε) την αντίστοιχη αναμενόμενη από τον τύπο (1) αφού βέβαια υποθέσουμε κατ' αρχήν ότι οι μεταβλητές μας είναι ανεξάρτητες.

		'Τύπος Σχολής'			*Σύνολο*
		Α	*Β*	*Γ*	
'ΚΟΕ'	*I*	50 **(47,73)**	53 **(50,16)**	49 **(54,11)**	**152**
	II	90 **(84,78)**	91 **(89,10)**	89 **(96,12)**	**270**
	III	17 **(24,50)**	21 **(25,74)**	40 **(27,77)**	**78**
Σύνολο		**157**	**165**	**178**	**500**

Οι αναμενόμενες συχνότητες είναι μέσα σε παρενθέσεις και έντονες. Π.χ. η αναμενόμενη συχνότητα *Α-I* = 47,73 προκύπτει απ' τη σχέση (1): $\frac{152 \cdot 157}{500} = 47{,}73$.

Η συχνότητα *Β-I* = 50,16 απ' τη σχέση (1): $\frac{152 \cdot 165}{500} = 50{,}16$ κ.ο.κ.

Οι υποθέσεις είναι:

Μηδενική H_0 : ανεξαρτησία (Σχολή ανεξάρτητη από (ΚΟΕ))

Εναλλακτική H_1 : εξάρτηση (Σχολή εξαρτημένη από (ΚΟΕ))

Βάσει του τύπου (2) έχουμε:

$$X^2 = \sum_{i=1,j=1}^{k,l} \frac{(\varepsilon_{ij} - \theta_{ij})^2}{\theta_{ij}} = \frac{(50-47{,}73)^2}{47{,}73} + \frac{(53-50{,}16)^2}{50{,}16} + \ldots + \frac{(40-27{,}77)^2}{27{,}77} = 10{,}19$$

Για $\alpha = 0{,}05$ και για βαθμούς ελευθερίας $(k-1)\cdot(l-1)$ = (3 – 1)·(3 – 1) = 4 ο πίνακας των X^2 – κατανομών δίνει ως κρίσιμη τιμή $x^2_{0,05} = 9{,}488$. Επειδή 10,19 > 9,488 απορρίπτουμε την μηδενική της ανεξαρτησίας και δεχόμαστε ότι υπάρχει εξάρτηση μεταξύ των δύο ποιοτικών μεταβλητών. Αυτό σημαίνει ότι οι προτιμήσεις των μαθητών από διαφορετικά κοινωνικοοικονομικά επίπεδα δεν είναι όμοιες. Φαίνεται ότι οι μαθητές του επιπέδου ΙΙΙ (αντίθετα από τους άλλους) προτιμούν να εισαχθούν σε σχολή τύπου Γ.

Παρατήρηση

Για να είναι έγκυρη η δοκιμασία ανεξαρτησίας του X^2 πρέπει οι αναμενόμενες συχνότητες να είναι μεγαλύτερες ή ίσες του 5. Αν αυτό δεν ισχύει θα πρέπει οι αναμενόμενες συχνότητες να είναι μικρότερες από 5 σε ποσοστό μικρότερο από 20% και καμία να μην είναι μικρότερη από 1.

§ 6. 9. Έλεγχος ισότητας δύο αναλογιών (ποσοστών)

Έστω δύο ανεξάρτητα και τυχαία δείγματα n_1 και n_2 και έστω f_1 και f_2 είναι οι αναλογίες ενός ποιοτικού χαρακτηριστικού στα δύο δείγματα. Αν p_1 και p_2 είναι οι αναλογίες του χαρακτηριστικού στους δύο πληθυσμούς από τους οποίους προέρχονται τα δείγματα, θέλουμε να ελέγξουμε αν αυτές οι αναλογίες μπορούν να θεωρηθούν ίσες. Οι δύο υποθέσεις είναι:

Η μηδενική $H_0 : p_1 = p_2$

Η εναλλακτική $H_1 : p_1 \neq p_2$ (δίπλευρος έλεγχος)

Έστω ότι κατ' αρχήν η H_0 είναι αληθινή. Αποδεικνύεται ότι η ποσότητα:

$$Z = \frac{f_1 - f_2}{\sqrt{p \cdot (1-p) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

είναι τιμή της τυποποιημένης κατανομής N(0, 1) όπου $p = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$ είναι εκτίμηση της άγνωστης κοινής αναλογίας $(p_1 = p_2 = p)$ του χαρακτηριστικού στον πληθυσμό.

Έστω α το επίπεδο σημαντικότητας του ελέγχου. Γνωρίζουμε (βλέπε γενική παρατήρηση παράγρ. 6.3) ποιες είναι οι κρίσιμες τιμές για $\alpha = 0,05$ $\alpha = 0,01$ και για μονόπλευρο ή δίπλευρο έλεγχο. Αν $|Z|$ είναι μεγαλύτερο από την κρίσιμη τιμή απορρίπτουμε την H_0, διαφορετικά τη δεχόμαστε.

Παρατήρηση

Ο παραπάνω έλεγχος ισχύει για μεγάλα δείγματα και συγκεκριμένα όταν:

$n_1 p \geq 5, \quad n_2 p \geq 5, \quad n_1(1-p) \geq 5, \quad n_2(1-p) \geq 5$.

Παράδειγμα

Σε μια μελέτη για τη διαπίστωση της επιθυμίας των ατόμων να γίνουν δωρητές οργάνων του σώματός τους μετά θάνατο, πάρθηκαν δύο τυχαία δείγματα, το ένα από 200 άνδρες και το άλλο από 300 γυναίκες και κατόπιν καταγράφηκε η επιθυμία τους. Βρέθηκε ότι 20 άνδρες και 37 γυναίκες εξέφρασαν την επιθυμία να γίνουν δωρητές. Να ελεγχθεί για $\alpha = 0{,}05$ αν η αναλογία αυτών που θέλουν να γίνουν δωρητές είναι ίδια στους άνδρες και στις γυναίκες.

Λύση

Είναι: $H_0 : p_1 = p_2$, $H_1 : p_1 \neq p_2$ (δίπλευρος έλεγχος). Επίσης:

$f_1 = 20/200 = 0{,}10$ $f_2 = 37/300 = 0{,}123$ είναι οι δύο αναλογίες και

$p = \dfrac{20+37}{200+300} = 0{,}114$ και $(1-p) = 0{,}886$. Συνεπώς

$$Z = \frac{0{,}10 - 0{,}123}{\sqrt{0{,}114 \cdot 0{,}886\left(\dfrac{1}{200} + \dfrac{1}{300}\right)}} = -0{,}79 .$$

Ο έλεγχος είναι δίπλευρος και $\alpha = 0{,}05$. Άρα η κρίσιμη τιμή είναι 1,96. Επειδή

$|Z| = |-0{,}79| = 0{,}79 < 1{,}96$ θεωρούμε την H_0 έγκυρη. Οι δύο αναλογίες δεν διαφέρουν στατιστικά.

Παρατήρηση

Ο έλεγχος ισότητας δύο αναλογιών μπορεί να γίνει προσεγγιστικά και με τον έλεγχο X^2. Θεωρώντας τα δεδομένα του προηγουμένου παραδείγματος μπορούμε να κατασκευάσουμε τον εξής πίνακα συνάφειας με δύο γραμμές και δύο στήλες:

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="2">Φύλο</td><td rowspan="2">Σύνολο</td></tr>
<tr><td>Άνδρες</td><td>Γυναίκες</td></tr>
<tr><td rowspan="2">Δωρεά οργά-νου</td><td>Ναι</td><td>20</td><td>37</td><td>57</td></tr>
<tr><td>Όχι</td><td>180</td><td>263</td><td>443</td></tr>
<tr><td colspan="2">Σύνολο</td><td>200</td><td>300</td><td>500</td></tr>
</table>

Αν η επιθυμία για δωρεά οργάνου είναι ανεξάρτητη από το φύλο, οι δύο αναλογίες δεν θα διαφέρουν στατιστικά. Αν όμως διαπιστωθεί εξάρτηση, τότε θα διαφέρουν. Για οποιονδήποτε 2x2 πίνακα συνάφειας, όπως ο παρακάτω, η ποσότητα (2) απλοποιείται ως εξής:

$$X^2 = \frac{n\left(\left|\varepsilon_1\varepsilon_4 - \varepsilon_2\varepsilon_3\right| - 0{,}5n\right)^2}{(\varepsilon_1 + \varepsilon_2)\cdot(\varepsilon_3 + \varepsilon_4)\cdot(\varepsilon_1 + \varepsilon_3)\cdot(\varepsilon_2 + \varepsilon_4)},$$

όπου η αφαίρεση του 0,5 ονομάζεται διόρθωση του Yates.

ε_1	ε_2	$\varepsilon_1 + \varepsilon_2$
ε_3	ε_4	$\varepsilon_3 + \varepsilon_4$
$\varepsilon_1 + \varepsilon_3$	$\varepsilon_2 + \varepsilon_4$	$\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4 = n$

Έτσι, με αντικατάσταση θα έχουμε:

$$X^2 = \frac{500\left(\left|20\cdot 263 - 37\cdot 180\right| - 0{,}5\cdot 500\right)^2}{57\cdot 443\cdot 200\cdot 300} = 0{,}436$$

Για $(k\ \text{-}1)\cdot(l - 1) = (2 - 1)\cdot(2 - 1) = 1$ βαθμό ελευθερίας και $\alpha = 0{,}05$ η κρίσιμη τιμή από τον πίνακα κατανομής X^2 είναι 3,841. Επειδή 0,436 < 3,841 δεχόμαστε την ανεξαρτησία των δύο μεταβλητών. Συνεπώς οι δύο αναλογίες δεν διαφέρουν.

Ασκήσεις

1) Το όριο αντοχής ενός τύπου καλωδίων έχει μέση τιμή 1800 κιλά και τυπική απόκλιση 100 κιλά. Η εταιρεία που φτιάχνει τα καλώδια ισχυρίζεται ότι μια βελτίωση στη μέθοδο κατασκευής αύξησε το όριο αντοχής. Για να το επαληθεύσουμε, δοκιμάζουμε 50 νέα καλώδια. Αν το μέσο όριο αντοχής τους βρέθηκε 1850 κιλά, είναι σωστός ο ισχυρισμός της εταιρείας σε επίπεδο σημαντικότητας 0,01;

2) Η μέση ζωή 100 λαμπτήρων φθορισμού από την παραγωγή μιας εταιρείας βρέθηκε να είναι 1570 ώρες με τυπική απόκλιση 120 ώρες. Αν μ είναι η μέση ζωή των λαμπτήρων φθορισμού της εταιρείας, ελέγξτε την υπόθεση $\mu = 1600$ με εναλλακτική υπόθεση την $\mu \neq 1600$ σε επίπεδο σημαντικότητας 0,05 και 0,01. Τι αποτελέσματα προκύπτουν αν η εναλλακτική υπόθεση είναι $\mu < 1600$ χρησιμοποιώντας πάλι επίπεδο σημαντικότητας 0,05 και 0,01;

3) Το μέσο βεληνεκές ενός πυραύλου είναι 200 μίλια. Μετά ένα έτος αποθηκεύσεως των πυραύλων, τυχαίο δείγμα 100 πυραύλων έδωσε μέσο βεληνεκές 175 μίλια και μέση απόκλιση 4 μίλια. Να ελεγχθεί αν η αποθήκευση των πυραύλων επί ένα έτος ελάττωσε το μέσο βεληνεκές των πυραύλων ($\alpha = 0{,}05$).

4) Οι εβδομαδιαίες δαπάνες των τριμελών νοικοκυριών μιας πόλης ακολουθούν την κανονική κατανομή με μέση τιμή $\mu = 300$ ευρώ. Δείγμα $n = 25$ νοικοκυριών έδωσε μέση τιμή $\bar{x} = 312$ ευρώ και $s = 32{,}5$ ευρώ. α) Να ελεγχθεί η υπόθεση H_0: $\mu = 300$ έναντι της H_1: $\mu \neq 300$ σε επίπεδο σημαντικότητας $\alpha = 0{,}05$. β) Να ελεγχθεί η υπόθεση H_0: $\mu = 300$ έναντι της H_1: $\mu > 300$ σε επίπεδο σημαντικότητας $\alpha = 0{,}05$. (Δίνεται $t_{24,\,0{,}025} = 2{,}064$ και $t_{24,\,0{,}05} = 1{,}711$).

5) Δοκιμάζεται ένας νέος τύπος πυραύλων για να διαπιστωθεί αν είναι καλύτερος από έναν παλαιό τύπο. Το μέσο βεληνεκές του παλαιού τύπου είναι 340 μίλια. Παίρνουμε τυχαίο δείγμα 10 πυραύλων νέου τύπου και βρίσκουμε μέσο βεληνεκές 360 μίλια και μέση απόκλιση 20 μίλια. Αν η κατανομή των βεληνεκών των πυραύλων ακολουθεί την κανονική κατανομή, να ελεγχθεί αν ο νέος τύπος πυραύλων είναι καλύτερος ή χειρότερος από τον παλαιό τύπο ($\alpha = 0{,}05$).

6) Από έναν κανονικό πληθυσμό πήραμε τυχαίο δείγμα 16 παρατηρήσεων και βρήκαμε μέση τιμή 41,5 και άθροισμα τετραγώνων αποκλίσεων από την μέση τιμή 135 ($\sum_{i=1}^{16}(x_i - \bar{x})^2 = 135$). Ερωτάται: είναι δυνατόν η μέση τιμή του πληθυσμού να είναι: α) 43,5 και β) μικρότερη; ($\alpha = 0{,}05$). Σε ποια όρια θα περιλαμβάνεται η μέση τιμή του πληθυσμού με πιθανότητα 95%;

7) Μία καπνοβιομηχανία ισχυρίζεται ότι το 62% των καπνιστών μιας πόλης καπνίζουν τσιγάρα τύπου Α που παράγει η καπνοβιομηχανία. Για να θεμελιώσει τον ισχυρισμό της διενεργεί δειγματοληπτική έρευνα ανάμεσα στους καπνιστές της πόλης. Σε τυχαίο δείγμα 2000 καπνιστών, οι 1280 κάπνιζαν τσιγάρα τύπου Α της καπνοβιομηχανίας. Ερωτάται: Ευσταθεί ο ισχυρισμός της καπνοβιομηχανίας; ($\alpha = 0{,}05$). Εντός ποιών ορίων περιλαμβάνεται η πραγματική αναλογία (ποσοστό) των καπνιστών τσιγάρων τύπου Α σ' όλους τους καπνιστές της πόλης με πιθανότητα 95%;

8) Σε τυχαίο δείγμα 100 ατόμων βρέθηκαν 40 καπνιστές. Να ελεγχθεί η μηδενική υπόθεση H_0: $p = 0{,}3$ έναντι της εναλλακτικής H_1: $p \neq 0{,}3$ α) σε επίπεδο $\alpha = 0{,}05$ και β) σε επίπεδο $\alpha = 0{,}01$. (Δίνεται $z_{0{,}025} = 1{,}96$ και $z_{0{,}005} = 2{,}58$).

9) Οι μισθοί σε δύο οργανισμούς Α και Β ακολουθούν κανονικές κατανομές. Δείγμα $n_1 = 11$ εργαζομένων από τον Α έδωσε $\bar{x}_1 = 1000$ ευρώ με $s_1 = 54$ και δείγμα

$n_2 = 16$ εργαζομένων από τον B έδωσε $\bar{x}_2 = 1100$ με $s_2 = 45$. Να ελεγχθεί η μηδενική υπόθεση H₀: $\sigma_1^2 = \sigma_2^2$ έναντι της εναλλακτικής H₁: $\sigma_1^2 \neq \sigma_2^2$ (α = 0,05).

10) Εργοστάσιο παράγει δύο τύπους ηλεκτρικών λαμπτήρων. Ο τύπος A έχει μέση απόκλιση 80 ώρες και ο τύπος B 94 ώρες. Πήραμε 50 λαμπτήρες τύπου A και 50 τύπου B. Από τον έλεγχο διαπιστώθηκε ότι ο τύπος A είχε μέση διάρκεια ζωής 1282 ώρες και ο τύπος B 1208. Ερωτάται: Διαφέρουν οι δύο τύποι λαμπτήρων σχετικά με τη μέση διάρκεια ζωής τους ($\mu_1 \neq \mu_2$); Ποιος τύπος είναι καλύτερος; (α =0,05).

11) Συγκρίνονται δύο διαφορετικές μέθοδοι καλλιέργειας, σχετικά με την απόδοση ενός γεωργικού προϊόντος. Σε δείγμα 100 πειραματικών αγρών εφαρμόζεται η μέθοδος A και δίνει $\bar{x}_1 = 55$ τόνους και $s_1^2 = 4$. Σε άλλο ισοπληθές δείγμα πειραματικών αγρών εφαρμόζεται η μέθοδος B η οποία δίνει $\bar{x}_2 = 50$ και $s_2^2 = 9$. Ερωτάται: Διαφέρουν οι δύο μέθοδοι καλλιέργειας από άποψη αποδόσεως; (α = 0,01). Ποια μέθοδος είναι καλύτερη;

12) Ένας δάσκαλος κάνει το ίδιο μάθημα σε δύο τάξεις A και B. Η τάξη A έχει 16 μαθητές ενώ η τάξη B έχει 25 μαθητές. Σε μια εξέταση, αν και δεν παρουσιάστηκε σημαντική διαφορά στο μέσο βαθμό των δύο τάξεων, οι βαθμοί της A είχαν τυπική απόκλιση 9 μονάδων, ενώ οι βαθμοί της B 12 μονάδες. Μπορούμε να συμπεράνουμε σε επίπεδο σημαντικότητας 0,01 και 0,05, ότι η διακύμανση της B είναι μεγαλύτερη από την διακύμανση της A;

13) Για την απόδοση σιταριού έγινε ένα πείραμα σε 24 πειραματικούς αγρούς. Στους 12 αγρούς χρησιμοποιήθηκε για λίπασμα φωσφόρος, ενώ στους άλλους 12 δεν χρησιμοποιήθηκε κανένα λίπασμα. Βρέθηκαν οι παρακάτω αποδόσεις:

Με φωσφόρο x_i	65	40	63	78	67	34	76	57	75	88	77	75
Χωρίς φωσφόρο y_i	60	42	65	71	62	35	74	54	71	82	77	67

Να ελεγχθεί αν η προσθήκη φωσφόρου για λίπασμα ασκεί επίδραση ή όχι στην αύξηση της απόδοσης σιταριού. Υποθέτουμε ότι παράγοντες όπως: είδος εδάφους, ηλιοφάνεια, κ.λ.π., οι οποίοι επιδρούν στην απόδοση του σιταριού είναι κοινοί (α = 0,05).

14) Σε τυχαίο δείγμα 400 τηλεθεατών οι 100 δήλωσαν ότι παρακολουθούν μια ορισμένη τηλεοπτική σειρά. Μπορούμε να δεχθούμε ότι εκείνοι που παρακολουθούν την τηλεοπτική σειρά στο σύνολο των τηλεθεατών είναι 30% ή μικρότερο; (α = 0,05).

15) Στον παρακάτω πίνακα δίνεται ο αριθμός των παιδιών κατά οικογένεια και το αντίστοιχο ετήσιο οικογενειακό εισόδημα.

Εισόδημα	***Αριθμός παιδιών***				***Σύνολο***
	0	**1**	**2**	**3** και άνω	
κάτω των 10.000 €	15	27	50	43	***135***
10.000 – 30.000 €	25	37	12	8	***82***
30.000 και πάνω	8	13	9	10	***40***
Σύνολο	***48***	***77***	***71***	***61***	***N = 257***

Ερωτάται: Υπάρχει σχέση εξαρτήσεως μεταξύ του ύψους του εισοδήματος και του αριθμού των παιδιών που αποκτά μια οικογένεια; (α = 0,01).

16) Στο μάθημα της Στατιστικής πήραν μέρος 1030 σπουδαστές μιας σχολής. Απ’ αυτούς, 798 ήταν αγόρια και 232 κορίτσια. Απ’ τους επιτυχόντες 262 ήταν αγόρια και 74 κορίτσια. Ερωτάται: Το φύλο ασκεί επίδραση στην επίδοση στη Στατιστική; (α = 0,05) (πίνακας συνάφειας με δύο γραμμές και δύο στήλες).

Στατιστικοί Πίνακες

Πίνακας 1: Τυποποιημένη κανονική κατανομή Ζ

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
-3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002
-3,3	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003
-3,2	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
-3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
-3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
-2,9	0,0019	0,0018	0,0017	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
-2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
-2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
-2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
-2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
-2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
-2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
-2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
-2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
-2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
-1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
-1,8	0,0359	0,0352	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
-1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
-1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
-1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
-1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0722	0,0708	0,0694	0,0681
-1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
-1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
-1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
-1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
-0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
-0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
-0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
-0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
-0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2276
-0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
-0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
-0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
-0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
-0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9278	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767

Συνέχεια της τυποποιημένης κατανομής Z …

z	**0,00**	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

Πίνακας 2: Κατανομή t του Student

Βαθμοί Ελευθερίας	***Κρίσιμη τιμή t***				
n	**0,10**	**0,05**	**0,025**	**0,01**	**0,005**
1	3,078	6,314	12,706	31,821	63,657
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750
40	1,303	1,684	2,021	2,423	2,704
60	1,296	1,671	2,000	2,390	2,660
120	1,289	1,658	1,980	2,358	2,617
∞	1,282	1,645	1,960	2,326	2,576

Πίνακας 3: Χ2 – Κατανομή

ν	**Κρίσιμη τιμή *α***							
	0,995	**0,99**	**0,975**	**0,95**	**0,05**	**0,025**	**0,01**	**0,005**
1	0,00004	0,00016	0,00098	0,00393	3,841	5,024	6,635	7,879
2	0,0100	0,0201	0,0506	0,103	5,991	7,378	9,210	10,597
3	0,0717	0,115	0,216	0,352	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	11,070	12,832	15,086	16,750
6	0,676	0,872	1,237	1,635	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	14,067	16,013	18,475	20,278
8	1,344	1,646	2,180	2,733	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	19,675	21,920	24725	26,757
12	3,074	3,571	4,404	5,226	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	36,415	39,364	42,980	45,558
25	10,520	11,524	13,120	14,611	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	38,885	41,923	45,642	48,290
27	11,808	12,879	14,573	16,151	40,113	43,194	46,963	49,645
28	12,461	13,565	15,308	16,928	41,337	44,461	48,278	50,993
29	13,121	14,256	16,047	17,708	42,557	45,722	49,588	52,336
30	13,787	14,953	16,791	18,493	43,773	46,979	50,892	53,672

Πίνακας 4: Κατανομή F (α = 0,05)

ν_1 : βαθμοί ελευθερίας αριθμητή, ν_2 : βαθμοί ελευθερίας παρονομαστή

ν_1 →	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	12	∞
1	16	20	21	22	23	23	23	23	24	24	24	24	24	24	25	25	25	25	25
2	18,	19,	19,	19,	19,	19,	19,	19,	19,	19,	19,	19,	19,	19,	19,	19,	19,	19,	19,
3	10,	9,5	9,2	9,1	9,0	8,9	8,8	8,8	8,8	8,7	8,7	8,7	8,6	8,6	8,6	8,5	8,5	8,5	8,5
4	7,7	6,9	6,5	6,3	6,2	6,1	6,0	6,0	6,0	5,9	5,9	5,8	5,8	5,7	5,7	5,7	5,6	5,6	5,6
5	6,6	5,7	5,4	5,1	5,0	4,9	4,8	4,8	4,7	4,7	4,6	4,6	4,5	4,5	4,5	4,4	4,4	4,4	4,3
6	5,9	5,1	4,7	4,5	4,3	4,2	4,2	4,1	4,1	4,0	4,0	3,9	3,8	3,8	3,8	3,7	3,7	3,7	3,6
7	5,5	4,7	4,3	4,1	3,9	3,8	3,7	3,7	3,6	3,6	3,5	3,5	3,4	3,4	3,3	3,3	3,3	3,2	3,2
8	5,3	4,4	4,0	3,8	3,6	3,5	3,5	3,4	3,3	3,3	3,2	3,2	3,1	3,1	3,0	3,0	3,0	2,9	2,9
9	5,1	4,2	3,8	3,6	3,4	3,3	3,2	3,2	3,1	3,1	3,0	3,0	2,9	2,9	2,8	2,8	2,7	2,7	2,7
10	4,9	4,1	3,7	3,4	3,3	3,2	3,1	3,0	3,0	2,9	2,9	2,8	2,7	2,7	2,7	2,6	2,6	2,5	2,5
11	4,8	3,9	3,5	3,3	3,2	3,0	3,0	2,9	2,9	2,8	2,7	2,7	2,6	2,6	2,5	2,5	2,4	2,4	2,4
12	4,7	3,8	3,4	3,2	3,1	3,0	2,9	2,8	2,8	2,7	2,6	2,6	2,5	2,5	2,4	2,4	2,3	2,3	2,3
13	4,6	3,8	3,4	3,1	3,0	2,9	2,8	2,7	2,7	2,6	2,6	2,5	2,4	2,4	2,3	2,3	2,3	2,2	2,2
14	4,6	3,7	3,3	3,1	2,9	2,8	2,7	2,7	2,6	2,6	2,5	2,4	2,3	2,3	2,3	2,2	2,2	2,1	2,1
15	4,5	3,6	3,2	3,0	2,9	2,7	2,7	2,6	2,5	2,5	2,4	2,4	2,3	2,2	2,2	2,2	2,1	2,1	2,0
16	4,4	3,6	3,2	3,0	2,8	2,7	2,6	2,5	2,5	2,4	2,4	2,3	2,2	2,2	2,1	2,1	2,1	2,0	2,0
17	4,4	3,5	3,2	2,9	2,8	2,7	2,6	2,5	2,4	2,4	2,3	2,3	2,2	2,1	2,1	2,1	2,0	2,0	1,9
18	4,4	3,5	3,1	2,9	2,7	2,6	2,5	2,5	2,4	2,4	2,3	2,2	2,1	2,1	2,1	2,0	2,0	1,9	1,9
19	4,3	3,5	3,1	2,9	2,7	2,6	2,5	2,4	2,4	2,3	2,3	2,2	2,1	2,1	2,0	2,0	1,9	1,9	1,8
20	4,3	3,4	3,1	2,8	2,7	2,6	2,5	2,4	2,3	2,3	2,2	2,2	2,1	2,0	2,0	1,9	1,9	1,9	1,8
21	4,3	3,4	3,0	2,8	2,6	2,5	2,4	2,4	2,3	2,3	2,2	2,1	2,1	2,0	2,0	1,9	1,9	1,8	1,8
22	4,3	3,4	3,0	2,8	2,6	2,5	2,4	2,4	2,3	2,3	2,2	2,1	2,0	2,0	1,9	1,9	1,8	1,8	1,7
23	4,2	3,4	3,0	2,8	2,6	2,5	2,4	2,3	2,3	2,2	2,2	2,1	2,0	2,0	1,9	1,9	1,8	1,8	1,7
24	4,2	3,4	3,0	2,7	2,6	2,5	2,4	2,3	2,3	2,2	2,1	2,1	2,0	1,9	1,9	1,8	1,8	1,7	1,7
25	4,2	3,3	2,9	2,7	2,6	2,4	2,4	2,3	2,2	2,2	2,1	2,0	2,0	1,9	1,9	1,8	1,8	1,7	1,7
26	4,2	3,3	2,9	2,7	2,5	2,4	2,3	2,3	2,2	2,2	2,1	2,0	1,9	1,9	1,9	1,8	1,8	1,7	1,6
27	4,2	3,3	2,9	2,7	2,5	2,4	2,3	2,3	2,2	2,2	2,1	2,0	1,9	1,9	1,8	1,8	1,7	1,7	1,6
28	4,2	3,3	2,9	2,7	2,5	2,4	2,3	2,2	2,2	2,1	2,1	2,0	1,9	1,9	1,8	1,8	1,7	1,7	1,6
29	4,1	3,3	2,9	2,7	2,5	2,4	2,3	2,2	2,2	2,1	2,1	2,0	1,9	1,9	1,8	1,8	1,7	1,7	1,6
30	4,1	3,3	2,9	2,6	2,5	2,4	2,3	2,2	2,2	2,1	2,0	2,0	1,9	1,8	1,8	1,7	1,7	1,6	1,6
40	4,0	3,2	2,8	2,6	2,4	2,3	2,2	2,1	2,1	2,0	2,0	1,9	1,8	1,7	1,7	1,6	1,6	1,5	1,5
60	4,0	3,1	2,7	2,5	2,3	2,2	2,1	2,1	2,0	1,9	1,9	1,8	1,7	1,7	1,6	1,5	1,5	1,4	1,3
12	3,9	3,0	2,6	2,4	2,2	2,1	2,0	2,0	1,9	1,9	1,8	1,7	1,6	1,6	1,5	1,5	1,4	1,3	1,2
∞	3,8	3,0	2,6	2,3	2,2	2,1	2,0	1,9	1,8	1,8	1,7	1,6	1,5	1,5	1,4	1,3	1,3	1,2	1,0

Συνέχεια πίνακα 4 Κατανομής F για α = 0,01…

Πίνακας 4: Κατανομή F (α = 0,01)

ν_1 : βαθμοί ελευθερίας αριθμητή, ν_2 : βαθμοί ελευθερίας παρονομαστή

$\nu_1 \rightarrow$ $\nu_2 \downarrow$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	∞
1	4052	5000	5403	5625	5764	5859	5928	5981	6023	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98,5	99,0	99,2	99,2	99,3	99,3	99,4	99,4	99,4	99,4	99,4	99,4	99,4	99,5	99,5	99,5	99,5	99,5	99,5
3	34,1	30,8	29,5	28,7	28,2	27,9	27,7	27,5	27,3	27,2	27,1	26,9	26,7	26,6	26,5	26,4	26,3	26,2	26,1
4	21,2	18,0	16,7	16,0	15,5	15,2	15,0	14,8	14,7	14,5	14,4	14,2	14,0	13,9	13,8	13,7	13,7	13,6	13,5
5	16,3	13,3	12,1	11,4	11,0	10,7	10,5	10,3	10,2	10,1	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	13,7	10,9	9,78	9,15	8,75	8,47	8,26	8,10	9,98	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	12,2	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	11,3	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	10,6	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	10,0	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	9,7	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	8,86	6,51	5,56	5,04	4,70	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	8,40	6,11	5,19	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	7,77	5,57	4,68	4,18	3,86	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	2,96	2,82	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

Βιβλιογραφία

Α. Ξένη

[1] Kreyszig Erwin «*Statistische Methoden und ihre Anwendungen*»

[2] Lindgren B., και Berry D. «*Elementary Statistics*»

[3] Spiegel Murray «*Theory and Problems of Probability and Statistics*»

Β. Ελληνική

[4] Αγγέλης Β. Δημάκη Κ. «*Στατιστική*»

[5] Αποστολόπουλος Θ. και Κ.

[6] Ζαΐρης Ποσειδών «*Στατιστική Μεθοδολογία*»

[7] Ζαχαροπούλου Χρυσούλα Α, Β τόμοι «*Στατιστική μέθοδοι-εφαρμογές»*

[8] Κιντής Ανδρέας «*Στατιστικές και Οικονομετρικές Μέθοδοι*»

[9] Κιοσέογλου Γρ. «*Σημειώσεις Στατιστικής*»

[10] Πέκος Γεώργιος «*Ασκήσεις Στατιστικής*»

[11] Σταυριανός Β. και Παναγιωτάκος Δ. «*Βιοστατιστική*»

[12] Χαλικιάς Ιωάννης «*Στατιστική*»