ΚΕΦΑΛΑΙΟ 5

Αντικείμενο και σκοπός της στατιστικής

5.1 Τι είναι η Στατιστική

Ο όρος «στατιστική» προέρχεται από τη Λατινική λέξη “status”. Στον πληθυντικό εκφράζει αριθμητικά δεδομένα που αφορούν διάφορα φαινόμενα. Στον ενικό εκφράζει το αντικείμενο της στατιστικής επιστήμης.

Η **στατιστική** είναι ο κλάδος των εφαρμοσμένων μαθηματικών, ο οποίος ασχολείται με τις μεθόδους σχεδιασμού μιας μελέτης, συλλογής , επεξεργασίας, παρουσίασης και ανάλυσης αριθμητικών δεδομένων, με σκοπό την εξαγωγή συμπερασμάτων για την λήψη ορθών αποφάσεων.

Τα βασικά στάδια μιας στατιστικής μελέτης ενός φαινομένου είναι :

1. Ο σχεδιασμός της μελέτης του φαινομένου.
2. Η συγκέντρωση των απαραίτητων πληροφοριών (στοιχείων).
3. Η επεξεργασία και η παρουσίαση των στοιχείων αυτών.
4. Η ανάλυση των στοιχείων με επιστημονικές μεθόδους.
5. Η εξαγωγή χρήσιμων συμπερασμάτων.
6. Η εισήγηση για την λήψη αποφάσεων και μέτρων.

5.2 Πληθυσμός και Δείγμα

Ο όρος πληθυσμός εδώ χρησιμοποιείται με την γενική (καθημερινή) έννοια, και αποτελείται θεωρητικά από το σύνολο των τιμών μιας τυχαίας μεταβλητής X, ή οποία εκφράζει ποσοτικά το υπό μελέτη χαρακτηριστικό. Ο όρος **πληθυσμός** στην στατιστική σημαίνει πάντοτε απαριθμήσεις ή μετρήσεις μιας ιδιότητας ή χαρακτηριστικού ενός συνόλου. Τα άτομα του πληθυσμού στα οποία πραγματοποιείται η παρατήρηση της τυχαίας μεταβλητής X λέγονται απλά στοιχεία (elements) ή μονάδες (units) του πληθυσμού. Ο πληθυσμός προσδιορίζεται και εξαρτάται από το αντικείμενο που μελετάται και αναλύεται. Κάθε πληθυσμός έχει ένα σύνολο ιδιοτήτων και χαρακτηριστικών, τα οποία είναι δυνατό να αποτελέσουν αντικείμενο στατιστικής ανάλυσης. Οι πληθυσμοί διακρίνονται σε **Άπειρους** και σε **Πεπερασμένους.**

Άπειρος ονομάζεται ένας πληθυσμός, όταν το πλήθος των στοιχείων του είναι πολύ μεγάλο, έτσι ώστε στην πράξη να μπορεί να θεωρηθεί ότι είναι

άπειρος.
Πεπερασμένος ονομάζεται ένας πληθυσμός που ο αριθμός των μονάδων που τον αποτελούν είναι καθορισμένος απόλυτα.
Στην πράξη επειδή υπάρχουν διάφορες δυσκολίες στην μελέτη μιας ιδιότητας σε πληθυσμό λόγω υψηλού κόστους ή λόγω έλλειψης χρόνου κλπ συνήθως ο ερευνητής μαζεύει πληροφορίες από κάποια μικρή ομάδα ή υποσύνολο του πληθυσμού κατά τέτοιο τρόπο ώστε τα αποτελέσματα που προκύπτουν από την εξέταση της ομάδας αυτής να είναι αντιπροσωπευτικά για τον μελετώμενο πληθυσμό.
Μια τέτοια ομάδα αποτελούμενη από n άτομα του πληθυσμού λέγεται **δείγμα μεγέθους n.**

5.3 Βασικές αρχές δειγματοληψίας

Είναι φανερό ότι το πρώτο στάδιο μιας στατιστικής μελέτης είναι η προσεκτική συλλογή των δεδομένων. Οι αρχές και οι μέθοδοι για την συλλογή και ανάλυση δεδομένων από πεπερασμένους πληθυσμούς είναι ένας κλάδος της στατιστικής γνωστός σαν «**Μέθοδος Δειγματοληπτικών Ερευνών»** (sample survey Methods) ή γενικότερα σαν Δειγματοληψία (sampling).
Η ανάλυση δειγματοληπτικών ερευνών είναι το κυριότερο εργαλείο έρευνας σε ένα μεγάλο φάσμα εφαρμογών, όπως στην Βιομηχανία, Εκπαίδευση, Ιατρική, Ψυχολογία, Γεωπονική, Αθλητισμό κλπ.
Το πρώτο βήμα σε μια δειγματοληπτική έρευνα είναι ο ακριβής προσδιορισμός του πληθυσμού για τον οποίο ενδιαφερόμαστε και από τον οποίο θα πάρουμε δείγμα. Η **διαδικασία** που θα χρησιμοποιήσουμε για την επιλογή του δείγματος, όπως και το **μέγεθος** δείγματος που θα εκλεγεί είναι σημαντικά στοιχεία για να έχουμε πράγματι ένα αντιπροσωπευτικό δείγμα του υπό εξέταση πληθυσμού. Αυξάνοντας το μέγεθος δείγματος ασφαλώς θα έχουμε και καλύτερη ακρίβεια στην εκτίμηση των χαρακτηριστικών του πληθυσμού. Από την άλλη πλευρά όμως το κόστος για την δειγματοληψία, την επεξεργασία και παρουσίαση των αποτελεσμάτων αυξάνει.
Πολύ μεγάλο δείγμα συνεπάγεται σπατάλη χρόνου, χρήματος και κόπου. Αντίθετα πολύ μικρό δείγμα είναι δυνατό να οδηγήσει σε μεροληπτικές εκτιμήσεις, αν και θα πρέπει να σημειωθεί ότι μια «προσεκτική» επιλογή μικρού δείγματος είναι δυνατό να δώσει πολύ καλύτερα αποτελέσματα από ένα μεγαλύτερο δείγμα που δεν έχει εκλεγεί κατάλληλα.
Το κυριότερο **μειονέκτημα** μιας δειγματοληπτικής μεθόδου έναντι ολικής απογραφής του πληθυσμού είναι ότι οι εκτιμήσεις των χαρακτηριστικών του

πληθυσμού από ένα μέρος αυτού έχουν τα λεγόμενα «**σφάλματα δειγματοληψίας**». Με κατάλληλο όμως μηχανισμό για την επιλογή των μονάδων του δείγματος είναι δυνατός ο περιορισμός των σφαλμάτων αυτών. Για την αξιολόγηση της ακρίβειας μιας δειγματοληπτικής έρευνας και για λόγους ευκολίας τα σφάλματα ταξινομούνται σε δύο κατηγορίες :

i) Δειγματοληπτικά σφάλματα

ii) Μη δειγματοληπτικά σφάλματα

Τα δειγματοληπτικά σφάλματα μπορεί να προκύψουν από τη μη σωστή επιλογή της κατάλληλης μεθόδου δειγματοληψίας, και την επιλογή ακατάλληλου δείγματος, όσον αφορά τη μεταβλητότητα και το μέγεθος του δείγματος.

Μια καλή δειγματοληπτική έρευνα πρέπει να περιλαμβάνει τον υπολογισμό των δειγματοληπτικών σφαλμάτων. Οι πληροφορίες για τα δειγματοληπτικά σφάλματα θα πρέπει να είναι άμεσα διαθέσιμες σε όσους προτίθενται να χρησιμοποιήσουν τα στατιστικά αποτελέσματα. Δηλαδή θα πρέπει να αναφέρεται τόσο η δειγματοληπτική μέθοδος, όσο και το σφάλμα της εκτίμησης μιας παραμέτρου.

Ο όρος "μη δειγματικό σφάλμα" προέρχεται από το γεγονός ότι αυτός ο τύπος σφάλματος μπορεί να συμβεί σε οποιαδήποτε έρευνα, είτε αυτή γίνεται με τη μέθοδο της απογραφής, είτε με τη μέθοδο της δειγματοληψίας.

ΚΕΦΑΛΑΙΟ 6

Οργάνωση και παρουσίασης στατιστικών στοιχείων

6.1 Οργάνωση και Γραφική Παράσταση Δεδομένων

Μεταβλητή (Variable) ενός πληθυσμού είναι το χαρακτηριστικό ως προς το οποίο εξετάζεται ο πληθυσμός. Οι **τυχαίες** μεταβλητές διακρίνονται ανάλογα με το είδος των τιμών που μπορούν να πάρουν σε ποιοτικές και ποσοτικές.

Ποιοτικές μεταβλητές είναι εκείνες που δεν επιδέχονται αριθμητικές μετρήσεις, αλλά περιγράφονται οι κατηγορίες στις οποίες ταξινομούνται οι παρατηρήσεις όπως πχ. το φύλλο, το χρώμα η γεύση, η υγεία (κακή, μέτρια ή καλή) κ.ο.κ. Η απλούστερη μορφή ποιοτικών παρατηρήσεων είναι εκείνη με δύο μόνο κατηγορίες, όπως για το φύλο είναι: άνδρας-γυναίκα, για το κάπνισμα μπορεί να είναι: καπνιστής-μη καπνιστής κλπ. Τα δεδομένα αυτά ονομάζονται **δυαδικά** (binary) ή **διχοτομικά** ή ναι /όχι ή 0-1. Άλλες ποιοτικές μεταβλητές έχουν περισσότερες από δύο κατηγορίες, όπως η ομάδα αίματος: Α, Β, ΑΒ, Ο, η ποιότητα οίνου: άριστη, καλή, μέτρια.

Ποσοτικές μεταβλητές είναι εκείνες οι οποίες επιδέχονται αριθμητικές, μετρήσεις, όπως τα φυσικά μεγέθη (ανάστημα, βάρος), τα βιολογικά μεγέθη (χοληστερόλη, σάκχαρα) κτλ. Θα πρέπει να τονισθεί ότι οι ποσοτικές μεταβλητές μπορούν να μετατραπούν σε διατάξιμες και οι διατάξιμες σε ποιοτικές, ενώ το αντίθετο είναι κατά βάση αδύνατο. Οι ποσοτικές μεταβλητές χωρίζονται σε **συνεχείς** και **ασυνεχείς**.

Συνεχείς ποσοτικές μεταβλητές λέγονται εκείνες που μπορούν να πάρουν θεωρητικά όλες τις τιμές των πραγματικών αριθμών, τουλάχιστον σε ένα διάστημα. **Ασυνεχείς** λέγονται εκείνες που είναι δυνατόν να λάβουν μόνο ορισμένες αριθμητικές τιμές.

Τα δεδομένα μιας στατιστικής έρευνας αποτελούνται συνήθως από ένα μεγάλο πλήθος στοιχείων που αφορούν τον πληθυσμό που μας ενδιαφέρει. Τα στοιχεία αυτά οργανώνονται αρχικά σε μορφή πινάκων με τέτοιο τρόπο ώστε να μπορεί κανείς με μια απλή ανάγνωση να σχηματίσει μια ιδέα για το δείγμα (ή τον πληθυσμό). Στη συνέχεια, για μια πιο αποτελεσματική παρουσίαση, γίνεται χρήση είτε γραφικών είτε αριθμητικών μεθόδων.

Υπάρχουν τρεις μέθοδοι παρουσίασης των στατιστικών δεδομένων α) με την ενσωμάτωση τους σε ειδική παράγραφο του κειμένου, β) με την εμφάνισή τους με τη μορφή πίνακα και γ) με τη μορφή γραφικής απεικόνισης. Η κατάταξη των στατιστικών δεδομένων μπορεί να γίνει είτε αλφαβητικά, είτε γεωγραφικά, είτε κατά μειωμένο ή αυξανόμενο, είτε ιστορικά (κατά χρονολογική σειρά).

Στατιστικοί πίνακες

Τα στατιστικά δεδομένα μπορούν να εξυπηρετήσουν το σκοπό για τον οποίο χρησιμοποιούνται, μόνο όταν η παρουσίασή τους γίνεται με τρόπο απλό και σύμφωνα με ορισμένη λογική τάξη, ώστε να διευκολύνονται οι συγκρίσεις. Ένας τέτοιος τρόπος είναι οι στατιστικοί πίνακες που αποτελούν συστηματικές κατατάξεις αριθμητικών δεδομένων σε στήλες και γραμμές. Η μορφή τους εξαρτάται κάθε φορά από τη χρήση για την οποία προορίζονται, από τη φύση των στοιχείων και από το χώρο που είναι διαθέσιμος για την κατασκευή τους.

Όταν κατασκευάζουμε ένα στατιστικό πίνακα είναι σκόπιμο να ακολουθούμε ορισμένους γενικούς κανόνες, οι οποίοι αποβλέπουν, ανάμεσα σε άλλα και στο να γίνει ο πίνακας αυτοτελής, δηλαδή όλες οι ποσοτικές πληροφορίες που μπορεί να ληφθούν να προκύπτουν από αυτόν εύκολα και με σαφήνεια.

α) Τίτλος. Ένας στατιστικός πίνακας πρέπει στο επάνω μέρος να έχει τίτλο. Ο τίτλος πρέπει να είναι σύντομος και να εκφράζει τη φύση των στοιχείων, τη γεωγραφική περιοχή στην οποία αφορούν και τη χρονική περίοδο που καλύπτουν.

β) Στήλες και γραμμές. Στην κορυφή κάθε στήλης (και στην αρχή κάθε γραμμής) γράφουμε συνοπτικά τη φύση των στοίχων που περιέχονται σ' αυτήν, καθώς και τι μονάδα μετρήσεως στην οποία εκφράζονται.

γ) Στρογγυλοποίηση αριθμών. Για διευκόλυνση των συγκρούσεων και απλούστευση του πίνακα πρέπει ο αριθμός των ψηφίων, από τα οποία αποτελούνται τα στοιχεία του, να είναι όσο το δυνατό μικρότερος. Το πόσο μπορεί, σε μια συγκεκριμένη περίπτωση, να περιορισθεί ο αριθμός των ψηφίων εξαρτάται από τη φύση των στοιχείων και από το σκοπό για τον οποίο πρόκειται να χρησιμοποιηθούν. Ο περιορισμός των ψηφίων γίνεται με την εφαρμογή κανόνων στρογγυλοποίησης, που στηρίζονται στον αλληλοσυμψηφισμό των σφαλμάτων.

δ) Μηδενική τιμή. Στους στατιστικούς πίνακες αποφεύγουμε να γράφουμε τον αριθμό 0, αλλά χρησιμοποιούμε, αντί γι' αυτόν, ένα σύμβολο πχ μια παύλα (−).

ε) Σύνολα. Όταν σε ένα στατιστικό πίνακα θέλουμε να παρουσιάσουμε και τα αθροίσματα των στοιχείων κατά στήλες ή κατά γραμμές, τα σύνολα που προκύπτουν τα γράφουμε στο τέλος (κάτω μέρος) των στηλών και στο τέλος (άκρη δεξιά) των γραμμών.

στ) Υποσημειώσεις. Αμέσως κάτω απ' τον κορμό του πίνακα μπορούμε να γράψουμε υποσημειώσεις, για να επεξηγήσουμε επιμέρους στοιχεία του και γενικά για να δώσουμε διευκρινίσεις, που μπορεί να είναι απαραίτητες για την κατανόηση και την καλή χρήση τους.

ζ) Πηγή. Μετά τις υποσημειώσεις < αν υπάρχουν > και πάντα κάτω από τον κορμό του πίνακα γράφουμε απαραίτητα την πηγή (ή τις πηγές) από την οποία έχουμε λάβει τα στοιχεία, εκτός εάν από το υλικό αυτό είναι πρωτογενές και το παρουσιάζουμε για πρώτη φορά.

Μετά την κατασκευή των πινάκων ακολουθεί η σύνταξη μιας έκθεσης, η οποία αναφέρεται στο αντικείμενο της έρευνας, στις μεθόδους που χρησιμοποιήθηκαν για τη συλλογή των στατιστικών στοιχείων και πληροφοριών, στην ανάλυση τους με τα ευρήματα που προέκυψαν, καθώς και τα συμπεράσματα.

6.2 Κατανομές συχνοτήτων

6.2.1 Έννοια των κατανομών συχνοτήτων

Το πρώτο βήμα της στατιστικής ανάλυσης είναι η οργάνωση και η συστηματοποίηση των δεδομένων, ώστε το έργο της στατιστικής ανάλυσης να διευκολυνθεί χωρίς να χάσει την ακρίβειά του.

Η απλούστερη μορφή συστηματοποίησης μιας ομάδας αριθμητικών δεδομένων είναι η διευθέτηση τους σε **σειρά μεγέθους.** Η συνήθης στατιστική μέθοδος που χρησιμοποιείται για την συστηματοποίηση μιας ομάδας δεδομένων, είναι οι κατανομές συχνότητας.

Οι κατανομές συχνότητας δείχνουν πόσο συχνά απαντά στην ομάδα κάθε τιμή (απλές κατανομές), ή κάθε διάστημα τιμών (ομαδοποιημένες κατανομές). Οι κατανομές συχνότητας συμπτύσσουν και απλοποιούν τα δεδομένα. Επιπλέον, με τις κατανομές συχνοτήτων, διευκολύνεται όπως θα δούμε σε άλλα κεφάλαια, ο υπολογισμός των διαφόρων αριθμητικών στατιστικών δεικτών.

6.2.2 Είδη κατανομών συχνοτήτων

Οι κατανομές συχνότητας των αριθμητικών μονομεταβλητών είναι δύο κατηγορίες απλές και ομαδοποιημένες κατανομές.

1) Απλές κατανομές :

Μια απλή κατανομή δείχνει πόσες φορές απαντά κάθε μια τιμή στην ομάδα. Στην αριστερή στήλη του πίνακα μιας απλής κατανομής καταχωρούνται όλες οι πιθανές τιμές της κλίμακας που περιλαμβάνονται μεταξύ της μικρότερης και της μεγαλύτερης τιμής της ομάδας. Στην δεξιά στήλη του πίνακα καταχωρούνται οι **συχνότητες**, δηλαδή οι αριθμοί που δείχνουν πόσο συχνά απαντά στην ομάδα κάθε τιμή της κλίμακας.

Για παράδειγμα στον πίνακα (2.1) παριστάνεται με μια απλή κατανομή ο αριθμός των αυγών που έκαναν 50 κότες σε περίοδο μιας εβδομάδας.

Α.Α.	Αριθμός αυγών Χ	Αριθμός κοτών συχνότητα ν
1	1	1
2	2	2
3	3	2
4	4	4
5	5	10
6	6	12
7	7	19
Σύνολο		50

πίνακα (6.1)

Η κατανομή συχνότητας δείχνει μια εικόνα του ¨πως¨ κατανέμονται οι τιμές της ομάδας κατά μήκος της κλίμακας. Στο παράδειγμα μας, παρατηρείται ότι οι περισσότερες τιμές συγκεντρώνονται στην κλίμακα μεταξύ 6 και 7 δηλαδή στο τέλος της κλίμακας.

Η πορεία που ακολουθούμε για την οργάνωση μιας ομάδας δεδομένων σε απλή κατανομή είναι η εξής:

α) Εντοπίζουμε την μικρότερη και τη μεγαλύτερη τιμή της ομάδας .

β) Σε μια στήλη **καταχωρίζουμε** όλες τις πιθανές τιμές της κλίμακας που περιλαμβάνονται μεταξύ της μικρότερης και της μεγαλύτερης τιμής της ομάδας.
γ) Για κάθε τιμή της κλίμακας, βρίσκουμε από τις μετρήσεις **πόσες φορές** εμφανίστηκε, και σημειώνουμε τον αριθμό αυτόν απέναντι της στην διπλανή στήλη. Ο αριθμός αυτός αποτελεί την απόλυτη συχνότητα εμφάνισης της τιμής αυτής. Η απόλυτη **συχνότητα** συμβολίζεται με το γράμμα f ή ν.

2) Ομαδοποιημένες κατανομές

Στις ομαδοποιημένες κατανομές το τμήμα της κλίμακας μέτρησης που περιλαμβάνεται μεταξύ της μικρότερης και μεγαλύτερης τιμής της ομάδας χωρίζεται σε **διαστήματα** και δίπλα σε κάθε διάστημα δίνεται η αντίστοιχη συχνότητα όπως φαίνεται στον πίνακα (6.2) στον οποίο φαίνονται ομαδοποιημένες σε οκτώ διαστήματα (κλάσεις) οι συχνότητες των παρατηρήσεων που πάρθηκαν από 200 σταθμούς για το ύψος της ετήσιας βροχόπτωσης (x) μιας χώρας.

	X	$f_i = v_i$
1	20 – 30	11
2	30 – 40	14
3	40 – 50	31
4	50 – 60	41
5	60 – 70	48
6	70 – 80	30
7	80 – 90	15
8	90 – 100	10
ΣΥΝΟΛΟ		200

πίνακα (6.2)

Μια ομαδοποιημένη κατανομή έχει τα εξής βασικά στοιχεία :
α) **Τα όρια του διαστήματος** ή της κλάσης: κάθε διάστημα έχει κατώτερο και ανώτερο όριο. Το κατώτερο όριο ενός διαστήματος είναι η μικρότερη τιμή της κλίμακας που περιλαμβάνει το διάστημα. Το ανώτερο όριο είναι η μεγαλύτερη τιμή της κλίμακας που περιλαμβάνει το διάστημα.
β) **Το εύρος (πλάτος) του διαστήματος**:

Το εύρος ενός διαστήματος είναι η απόσταση μεταξύ των πραγματικών ορίων του διαστήματος. Συνήθως το εύρος των διαστημάτων σε μια κατανομή είναι το ίδιο για όλα τα διαστήματα. Είναι όμως δυνατόν στην ίδια κατανομή το εύρος των διαστημάτων να είναι διαφορετικό. Στην κατανομή με ίσο εύρος σε όλα τα διαστήματα, το εύρος του διαστήματος υπολογίζεται από την διαφορά διαδοχικών κατώτερων ορίων δύο διαστημάτων, ενώ σε κατανομές με διαφορετικό εύρος, το εύρος κάθε διαστήματος υπολογίζεται από την αφαίρεση από το ανώτερο φαινομενικό όριο, το κατώτερο φαινομενικό όριο του διαστήματος.

γ) **Η μέση τιμή του διαστήματος**:

Η μέση τιμή του διαστήματος είναι το μέσο (το κέντρο) του διαστήματος και ισούται με το μισό του αθροίσματος των ορίων του διαστήματος.

δ) **Η συχνότητα του διαστήματος** :

Η συχνότητα ενός διαστήματος είναι ο αριθμός που δείχνει πόσο συχνά απαντούν οι τιμές του διαστήματος στην ομάδα , συμβολίζεται με fi ή ν_i . Η ομαδοποίηση δεδομένων χρειάζεται μεγάλη προσοχή ώστε να διατηρηθεί η αρχική μορφή της κατανομής. Η αυθαίρετη ομαδοποίηση μπορεί να οδηγήσει σε παραπλανητικά συμπεράσματα για τα δεδομένα της κατανομής.

6.3. Ομαδοποίηση δεδομένων

Η γενική πορεία που ακολουθούμε κατά την κατασκευή μιας ομαδοποιημένης κατανομής είναι:

1) **Υπολογίζουμε το εύρος (R)** της κατανομής από την σχέση

R = Μέγιστη τιμή της κατανομής – ελάχιστη τιμή της κατανομής.

2) **Επιλέγουμε τον αριθμό των διαστημάτων** (κλάσεων ή ομάδων) q Ο αριθμός των διαστημάτων συνήθως ορίζεται αυθαίρετα από τον ερευνητή σύμφωνα με την πείρα του. Συνήθως ο αριθμός των διαστημάτων κυμαίνεται μεταξύ 5 και 20 διαστήματα. Υπάρχει όμως και ένας τύπος που χρησιμοποιείται ως οδηγός. Αυτός είναι γνωστός ως τύπος του **Sturges** και ορίζεται ως εξής: $q = 1+3,32 \log N$ όπου q = αριθμός των διαστημάτων (κλάσεων), N = αριθμός των μετρήσεων του δείγματος.

3) **Υπολογίζουμε το εύρος (πλάτος)** του διαστήματος **c**.

Το εύρος ή το πλάτος του διαστήματος συνίσταται να είναι το ίδιο για όλα τα διαστήματα της κατανομής. Συνήθως το πλάτος c υπολογίζεται διαιρώντας το εύρος (R) της κατανομής δια του αριθμού των διαστημάτων της. Δηλαδή $C = R/q$

4) Το τέταρτο βήμα είναι **ο καθορισμός των διαστημάτων**. Το πρώτο διάστημα διαλέγεται έτσι ώστε να περιέχει την μικρότερη παρατήρηση και το τελευταίο να περιέχει τη μεγαλύτερη. Καλό είναι όμως η κατώτερη τιμή του πρώτου διαστήματος να είναι μικρότερη από την ελάχιστη τιμή της κατανομής, και το ανώτερο όριο του τελευταίου διαστήματος να είναι μεγαλύτερη από την μέγιστη τιμή της κατανομής, και καμιά τιμή να μην πέφτει σε άκρο διαστήματος.
5) **Υπολογίζουμε την συχνότητα κάθε διαστήματος** δηλαδή πόσο συχνά απαντούν οι τιμές του διαστήματος.

Παράδειγμα 6.1
Η συγκέντρωση (σε $\mu g/cm^3$) ενός συγκεκριμένου ρύπου σε δείγματα αέρος που πάρθηκαν από 57 περιοχές της Ελλάδας, δίνονται από τον επόμενο πίνακα.(6.3)

70	64	43	28	31	37	29	33	80	28	23	24	25	26	25
67	44	26	75	52	37	43	29	32	29	26	46	13	58	52
13	33	50	39	43	28	32	51	39	22	17	25	70	48	24
23	44	28	50	49	24	13	20	47	31	50	50	30	13	77

Η ομαδοποίηση της παραπάνω κατανομής γίνεται σύμφωνα με τα πέντε στάδια που αναφέρθηκαν παραπάνω:
1) Το εύρος της κατανομής R = Μέγιστη τιμή – Ελάχιστη τιμή
$$R = 80 - 13 = 67$$
2) Ο αριθμός των διαστημάτων $q = 1+3,32 \log N = 1+ 3,32\log 60$
$$q = 6,90 \approx 7$$
3) Το εύρος του διαστήματος $c = R/q = 67/7 = 9,6 \approx 10$
4) Διαλέγουμε το κατώτερο όριο του πρώτου διαστήματος να είναι 11,5 οπότε καμία τιμή δεν πέφτει σε άκρο διαστήματος.
5) μετράμε τις τιμές που περιλαμβάνονται σε κάθε διάστημα οπότε προκύπτει ο παρακάτω πίνακας (6.4)

i	Κάτω όριο – Άνω όριο	Κέντρο x_i	$v_i = f_i$
1	11,5 – 21,5	16,5	6
2	21,5 – 31,5	26,5	21
3	31,5 – 41,5	36,5	9
4	41,5 – 51,5	46,5	14
5	51,5 – 61,5	56,5	3
6	61,5 – 71,5	66,5	4
7	71,5 – 81,5	76,5	3
			60

πίνακας (6.4)

6.4. Γραφικές μέθοδοι παρουσίασης στατιστικών δεδομένων

Η παρουσίαση των στατιστικών δεδομένων (στοιχείων) μπορεί να γίνει με πίνακες ή με διαγράμματα ή γραφικές παραστάσεις (απεικονίσεις). Η γραφική παράσταση της κατανομής συχνοτήτων ενός συνόλου δεδομένων είναι ένα χρήσιμο εργαλείο το οποίο επιτρέπει την άμεση οπτική παρουσίαση των βασικών χαρακτηριστικών τους ακόμη και στον μη ειδικό. Η παραστατικότητα συμβάλλει στην γρήγορη κατανόηση των στοιχείων και στην διατήρηση της εικόνας τους στην μνήμη του αναγνώστη. Οι γραφικές παραστάσεις πρέπει να είναι όσο γίνεται ακριβείς, ώστε οι πληροφορίες που δίνουν να μην απέχουν από την πραγματικότητα.
Ανάλογα με το είδος των δεδομένων που διαθέτουμε υπάρχουν διάφοροι τρόποι παρουσίασης.

Α) Παρουσίαση ποιοτικών δεδομένων

Η γραφική παράσταση ποιοτικών δεδομένων γίνεται ή με **ραβδογράμματα** (bar chart) ή με **κυκλικό διάγραμμα συχνοτήτων** (pie chart). Στο ραβδόγραμμα οι κατηγορίες της τυχαίας μεταβλητής παριστάνονται στον οριζόντιο άξονα με ίσα διαστήματα και συνήθως με κενά ανάμεσά τους, ενώ οι συχνότητες ή οι σχετικές συχνότητες στον κατακόρυφο άξονα. Η κατασκευή ενός ραβδογράμματος θα γίνει κατανοητά με την εφαρμογή του παρακάτω παραδείγματος:

Παράδειγμα (6.2)

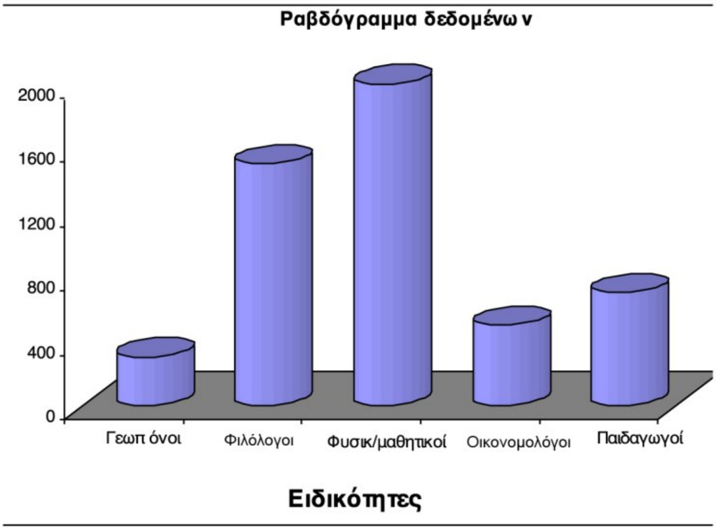
Στον παρακάτω πίνακα (6.5) δίνονται οι συχνότητες ανά ειδικότητα για ένα δείγμα 5000 πτυχιούχων τριτοβάθμιας εκπαίδευσης που συμμετέχουν στον διαγωνισμό του ΑΣΕΠ για το διορισμό στο δημόσιο τομέα.

I	ειδικότητα (y_i)	συχνότητα v_i	συχνότητα f_i	f_i x100
1	Γεωπόνοι	300	0,06	6
2	Φιλόλογοι	1500	0,30	30
3	Φυσικομαθηματικοί	2000	0,40	40
4	Οικονομολόγοι	500	0,10	10
5	Παιδαγωγοί	700	0,14	14

πίνακας 6.5

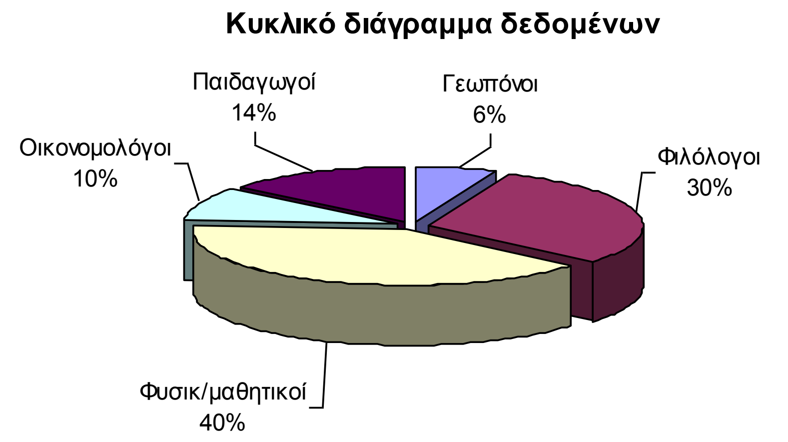
Η κατασκευή του ραβδογράμματος των δεδομένων του πίνακα (6.5) γίνεται ως εξής:

1. Καταγράφουμε τις συχνότητες ή τις σχετικές συχνότητες κατά μήκος του κατακόρυφου άξονα, και τις κατηγορίες της ποιοτικής μεταβλητής κατά μήκος του οριζόντιου άξονα.
2. Κατασκευάζουμε ορθογώνια ή κυλίνδρους πάνω από κάθε κατηγορία της μεταβλητής με ύψος ίσο προς την συχνότητα της κατηγορίας αντίστοιχα.
3. Αφήνουμε ίσα κενά ανάμεσα σε κάθε κατηγορία στον οριζόντιο άξονα για καλύτερη παρουσίαση. Το ραβδόγραμμα των δεδομένων του πίνακα (6.5) φαίνεται στο σχήμα (6.1)



Σχήμα(6.1)
Ραβδόγραμμα των δεδομένων του πίνακα (6.5)

Το **κυκλικό** διάγραμμα χρησιμοποιείται για να δείξει το ποσοστό του συνολικού αριθμού των παρατηρήσεων που περιέχει κάθε κατηγορία της μεταβλητής. Ο κύκλος διαιρείται σε τόσα τμήματα όσες είναι οι κατηγορίες και κάθε τμήμα να αντιστοιχεί σε ένα ποσοστό της κατηγορίας που αντιπροσωπεύει. Το κυκλικό διάγραμμα για τα ποσοστά του πίνακα (6.5) φαίνεται στο σχήμα (6.2)



Σχήμα(6.2) Κυκλικό διάγραμμα των δεδομένων του πίνακα (6.5)

Β) Παρουσίαση ποσοτικών δεδομένων

Η γραφική παράσταση της κατανομής μιας ποσοτικής (αριθμητικής) μεταβλητής μπορεί να γίνει με δυο τρόπους: με το **ιστόγραμμα** συχνοτήτων ή σχετικών συχνοτήτων και με **πολύγωνο** συχνοτήτων**.**

Το πιο συνηθισμένο μέσο περιγραφής ποσοτικών δεδομένων είναι το ιστόγραμμα (histogram) συχνοτήτων ή σχετικών συχνοτήτων. Η κατασκευή τόσο των ιστογραμμάτων όσο και των πολυγώνων συχνοτήτων ακολουθεί την ίδια περίπου διαδικασία:

1. Κατασκευάζουμε δύο καθέτους περίπου άξονες. Στον οριζόντιο άξονα τοποθετούμε τις τιμές της κλίμακας μέτρησης της μεταβλητής ενώ στον κάθετο άξονα την κλίμακα συχνότητας. Η συχνότητα μπορεί να εκφραστεί είτε σε απόλυτους αριθμούς (απόλυτη συχνότητα) είτε σε% ποσοστά (σχετική συχνότητα)

2. Σε κάθε τιμή του οριζόντιου άξονα υψώνουμε ένα ορθογώνιο παραλληλόγραμμο που το ύψος του είναι ισοδύναμο με την συχνότητα της τιμής (αν πρόκειται για ιστόγραμμα συχνοτήτων) ή με την σχετική συχνότητα της τιμής (αν πρόκειται για ιστόγραμμα σχετικών συχνοτήτων). Τα ορθογώνια παραλληλόγραμμα είναι διαδοχικά (χωρίς διαστήματα ανάμεσά τους) και με ίσο πλάτος.

3. Για την κατασκευή του πολυγώνου, σε κάθε τιμή του οριζόντιο άξονα και σε ύψος αντίστοιχο με την συχνότητα (ή την σχετική συχνότητα) της τιμής. Τοποθετούμε ένα σημείο (τελεία ή κουκίδα), στην συνέχεια συνδέουμε διαδοχικά τα σημεία για όλες τις τιμές του οριζόντιου άξονα με τεθλασμένη γραμμή. Η εικόνα που δίνει μια σχηματική παρουσίαση μιας κατανομής πρέπει να είναι πιστή και συγχρόνως να είναι οπτικά και αισθητικά ευχάριστη για το λόγο αυτό πρέπει το μήκος που λαμβάνουν οι τιμές στον κάθετο άξονα να είναι περίπου ¾ του μήκους που καταλαμβάνουν οι τιμές στον οριζόντιο άξονα.

Παράδειγμα (6.3)
Ο αριθμός των φυτών του σπαρτού Carex Flacca που βρέθηκαν σε 300 τετράγωνα του 1 m^2 μιας περιοχής είναι:

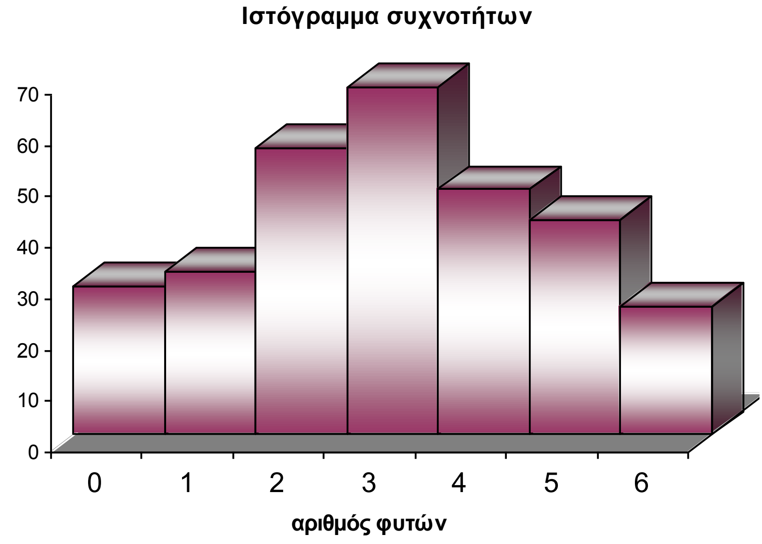
x_i	0	1	2	3	4	5	6
v_i	29	32	56	68	48	42	25

Εάν x_i ο αριθμός των φυτών κατά τετράγωνο και v_i η συχνότητα εμφάνισης, να γίνει γραφική παράσταση της κατανομής της συχνότητας και της σχετικής συχνότητας με ιστόγραμμα και με πολύγωνο.

Λύση:
Το πρώτο βήμα είναι να φτιάξουμε τον πίνακα με τις αθροιστικές συχνότητες N_i , σχετικές συχνότητες f_i , και σχετικές αθροιστικές συχνότητες F_i:

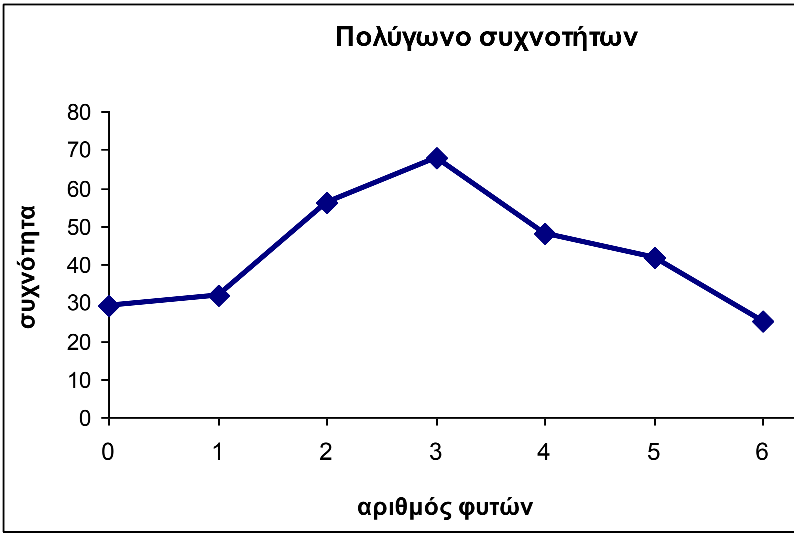
	x_i	v_i	$\%f_i$	N_i	$F_i\%$
1	0	29	9,667	29	9,667
2	1	32	10,667	61	20,333
3	2	56	18,667	117	39,00
4	3	68	22,667	185	61,667
5	4	48	16,00	233	77,667
6	5	42	14,00	275	91,667
7	6	25	8,333	300	100
ΣΥΝΟΛΟ		300	100		

Το δεύτερο βήμα είναι να χαράζουμε τους κάθετους μεταξύ τους άξονες και πάνω στους άξονες να τοποθετούμε τις κλίμακες της μεταβλητής της συχνότητας και της σχετικής συχνότητας.
Το τρίτο βήμα είναι να υψώσουμε τα ορθογώνια παραλληλόγραμμα σε κάθε τιμή του οριζόντιου άξονα σε ύψος ίσο με τη συχνότητα ή την σχετική συχνότητα (για το ιστόγραμμα), ή να σημειώσουμε με κουκίδα ή τελεία (για το πολύγωνο) και ύστερα να συνδέσουμε τα σημεία του πολυγώνου. Το ιστόγραμμα και το πολύγωνο συχνότητας και σχετικής συχνότητας φαίνονται στα παρακάτω σχήματα .



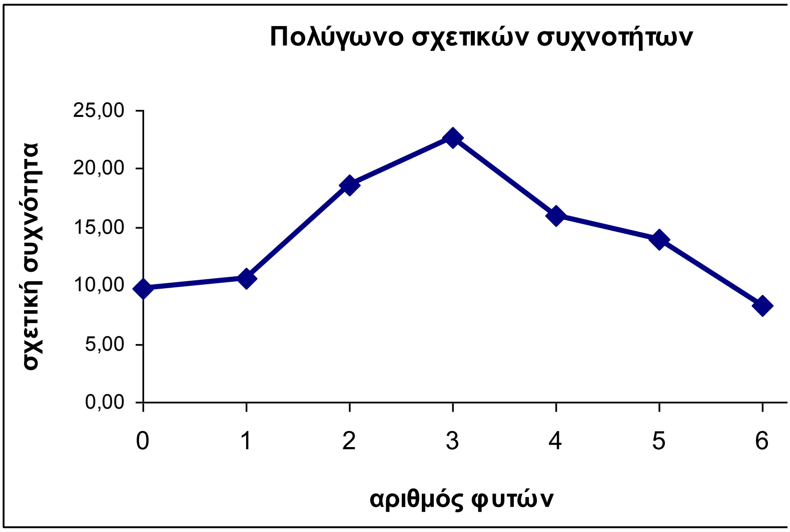
σχήμα (6.3)

α) Ιστόγραμμα συχνότητας και σχετικής συχνότητας



σχήμα (6.4)

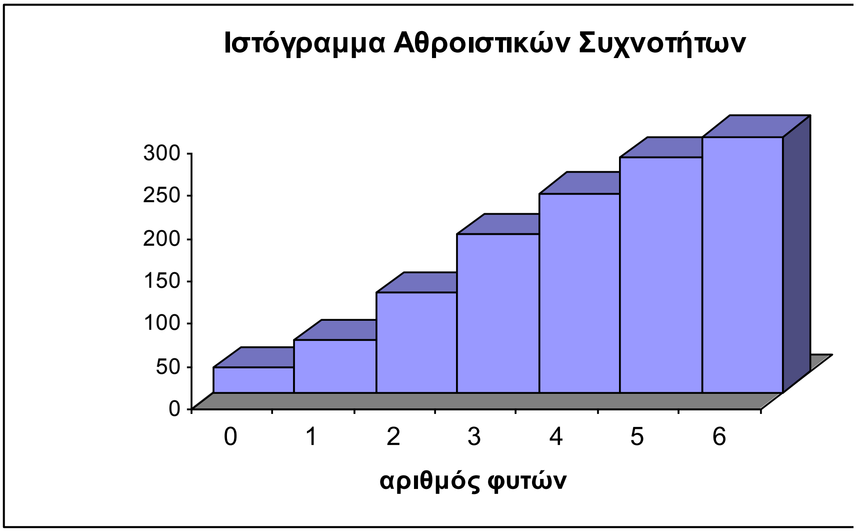
β) Πολύγωνο συχνοτήτων



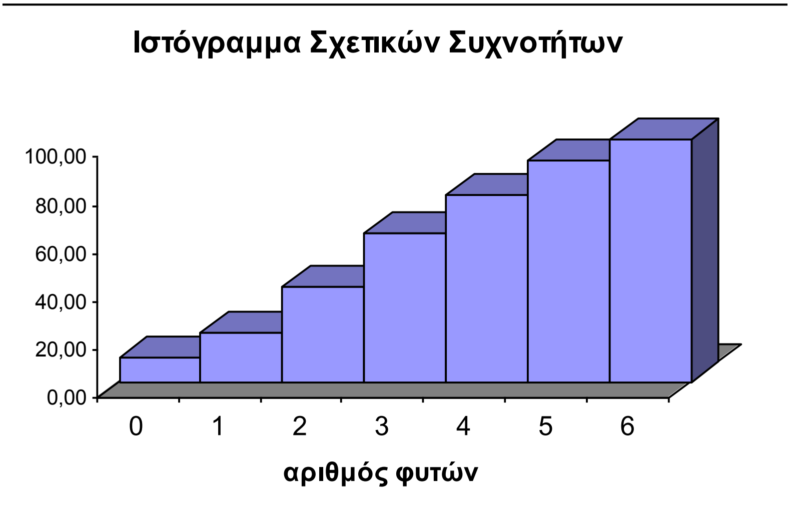
σχήμα (6.5)

γ) Πολύγωνο σχετικών συχνοτήτων

Μπορούμε επίσης να χαράξουμε το ιστόγραμμα αθροιστικών συχνοτήτων και αθροιστικών σχετικών συχνοτήτων όπως φαίνεται στο παρακάτω σχήμα



σχήμα (6.6)
Ιστόγραμμα αθροιστικών συχνοτήτων κατανομής των φυτών του σπαρτού Carex Flacca.



σχήμα (6.7)
Ιστόγραμμα σχετικών αθροιστικών συχνοτήτων κατανομής των φυτών του σπαρτού Carex Flacca.

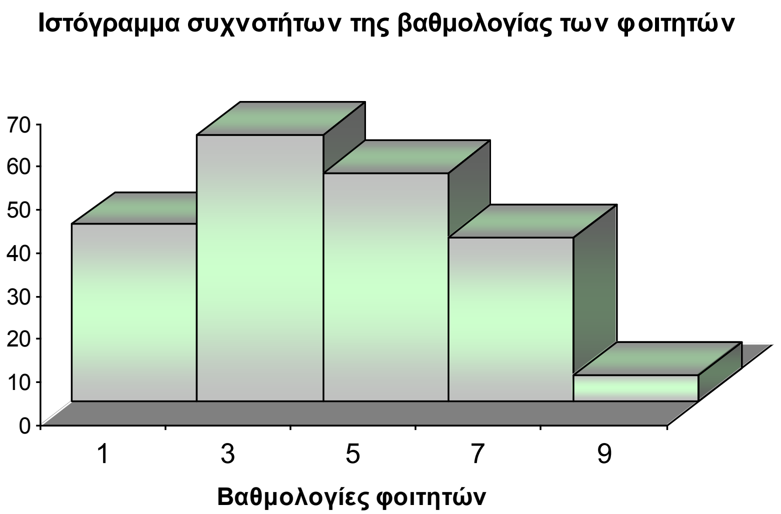
Παράδειγμα (6.4)
Στον παρακάτω πίνακα φαίνεται η καταχώρηση της βαθμολογίας ενός μαθήματος, 200 φοιτητών

Βαθμός	0-2	2-4	4-6	6-8	8-10
Πλήθος φοιτητών	41	62	53	38	6

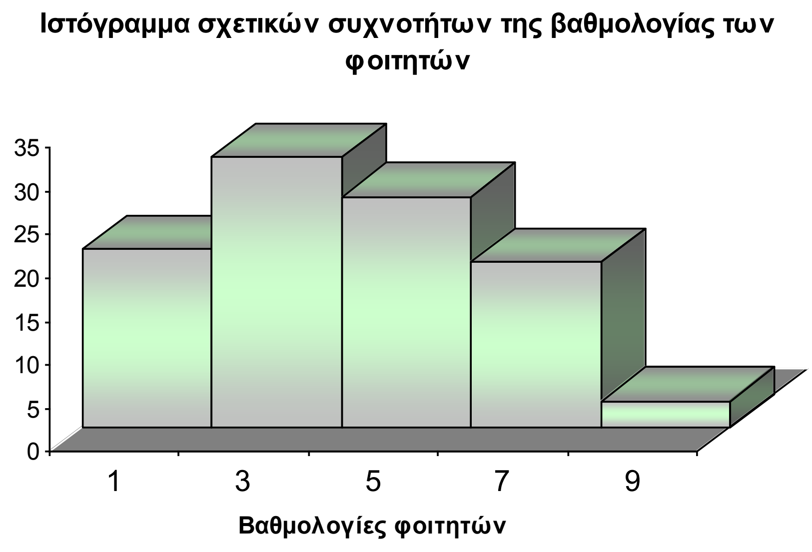
Να κατασκευαστεί ένα ιστόγραμμα συχνοτήτων για τα δεδομένα του παραπάνω πίνακα.
Λύση
Καταγράφουμε τον πίνακα συχνοτήτων και σχετικών συχνοτήτων όπως φαίνεται παρακάτω:

Βαθμός	Συχνότητα v_i	%Σχετική συχνότητα f_i
0-2	41	20,5
2-4	62	31
4-6	53	26,5
6-8	38	19
8-10	6,0	3,0
Σύνολο	200	100

Το ιστόγραμμα συχνότητας και σχετικής συχνότητας παρουσιάζεται στο παρακάτω σχήμα 2.8



σχήμα 6.8



Σχήμα(6.9)
Τα ιστογράμματα των συχνοτήτων και σχετικών συχνοτήτων της βαθμολογίας 200 φοιτητών σε ένα μάθημα.

Κεφάλαιο 7

Αντιπροσωπευτικές τιμές των κατανομών συχνοτήτων ποσοτικών μεταβολών

7.1 Εισαγωγή

Αντιπροσωπευτικές τιμές μιας κατανομής θεωρούνται οι αριθμητικές εκείνες τιμές, οι οποίες υπολογίζονται με βάση τα αναλυτικά στοιχεία κατανομής, και μπορούν να υποδείξουν τα κύρια χαρακτηριστικά της.

Οι αντιπροσωπευτικές τιμές διακρίνονται σε ,(α) εκείνες που προσδιορίζουν τη **θέση** της συνολικής κατανομής στον οριζόντιο άξονα, και ονομάζονται **μέτρα κεντρικής τάσης** ή **μέτρα θέσης** των στατιστικών δεδομένων και β) σε εκείνες που προσδιορίζουν το **βαθμό διασποράς** των παρατηρήσεων στον οριζόντιο άξονα και ονομάζονται **μέτρα διασποράς.**

7.2 Μέτρα κεντρικής τάσης (μέτρα θέσης) (measures of central tendency):

Η κατανομή συχνοτήτων αποτελεί ομαδοποίηση ενός συνόλου στατιστικών δεδομένων και περιέχει περιληπτικά όλες τις πληροφορίες που περιέχονται στα αρχικά δεδομένα. Μια τέτοια ταξινόμηση δεν μας δίνει όμως την δυνατότητα να κάνουμε ποσοτική περιγραφή της κατανομής ή και τη σύγκρισή της με μια άλλη ή περισσότερες άλλες κατανομές. Για την ποσοτική περιγραφή της κατανομής χρησιμοποιούνται τα μέτρα κεντρικής τάσης τα οποία εκφράζουν με σαφή τρόπο τα κύρια γνωρίσματα της κατανομής συχνότητας. Τα μέτρα κεντρικής τάσης είναι:

7.2.1 Μέση τιμή ή μέσος όρος (αριθμητικός μέσος) ή δειγματική μέση τιμή (mean value ή sample mean):

Ο μέσος όρος (μέση τιμή) είναι ο πιο γνωστός δείκτης κεντρικής τάσης και ορίζεται ως αλγεβρικό άθροισμα των τιμών όλων των παρατηρήσεων του δείγματος διαιρεμένο με το πλήθος των μετρήσεων αυτών και συμβολίζεται με $\bar{x}$ και ορίζεται με τη σχέση

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (7.1)$$

Όπου n είναι το πλήθος των μετρήσεων και το σύμβολο $\sum$ δείχνει άθροισμα. Όταν αναφερόμαστε σε ολόκληρο τον στατιστικό πληθυσμό η μέση τιμή του πληθυσμού συμβολίζεται με μ και το πλήθος των τιμών με Ν και τότε:

$$\mu = \frac{1}{N}\sum_{i=1}^{n} x_i \qquad (7.2)$$

Όταν χρησιμοποιούμε πίνακα συχνοτήτων σε μια απλή κατανομή, η μέση τιμή προκύπτει από τις ισοδύναμες εκφράσεις

$$\overline{x} = \frac{\sum_{i=1}^{\kappa} \nu_i x_i}{\sum_{i=1}^{\kappa} \nu_i} = \sum_{i=1}^{\kappa=1} f_i x_i \qquad (7.3) \quad , \quad \text{όπου } f_i = \frac{\nu_i}{\sum \nu_i} \qquad (7.4)$$

και ο υπολογισμός της γίνεται ως εξής:

α) Πολλαπλασιάζουμε κάθε τιμή x_i της κλίμακας με την απόλυτη συχνότητα εμφάνισης της ν_i και τα γινόμενα καταχωρίζονται σε διπλανή στήλη.

β) Αθροίζουμε τα γινόμενα του σταδίου (α) και έτσι βρίσκουμε την ποσότητα $\sum \nu_i x_i$ η οποία αντιστοιχεί στο άθροισμα όλων των τιμών της ομάδας.

γ) Διαιρούμε το άθροισμα ($\sum \nu_i x_i$) του σταδίου (β) με το μέγεθος n το οποίο είναι ίσο με $\sum \nu_i$ δηλαδή με το σύνολο των μετρήσεων.

Παράδειγμα (7.1)

Σε δείγμα 1000 οικογενειών σε μια χώρα της μέσης Ανατολής καταγράφηκε ο αριθμός των παιδιών που έχουν αποκτήσει. Τα αποτελέσματα της καταγραφής φαίνονται στον παρακάτω πίνακα:

Αριθμός παιδιών x_i	Αριθμός οικογενειών ν_i
0	30
1	45
2	80
3	150
4	350
5	235
6	57
7	53
	1000

Να υπολογισθεί ο μέσος όρος των παιδιών ανά οικογένεια.

Λύση
Για τον υπολογισμό της μέσης τιμής του δείγματος ακολουθούμε τα σημεία α), β), και γ) που αναφέρθηκαν παραπάνω φτιάχνοντας τον παρακάτω πίνακα:

x_i	v_i	$x_i v_i$
0	30	0
1	45	45
2	80	160
3	170	510
4	330	1320
5	235	1175
6	57	342
7	53	371
Σύνολο	1000	3923

Από τον πίνακα έχουμε: $\sum \nu_i = n = 100$, $\sum \nu_i x_i = 394$ οπότε

$$\bar{x} = \frac{\sum \nu_i x_i}{\sum \nu_i} = \frac{3923}{1000} = 3{,}923 \approx 3{,}9$$

Άρα ο δειγματικός μέσος όρος των παιδιών ανά οικογένεια είναι 3,9 παιδιά.

Σε ομαδοποιημένες κατανομές ο υπολογισμός του δειγματικού μέσου όρου (μέση τιμή) $\bar{x}$ γίνεται με παρόμοιο τρόπο όπως στην περίπτωση απλής κατανομής με πίνακα συχνοτήτων. Η διαφορά είναι ότι στην ομαδοποιημένη κατανομή θεωρούμε ότι οι τιμές που ανήκουν σε μια κλάση (διάστημα) κατανέμονται ομοιόμορφα σ αυτό, οπότε η κεντρική τιμή του διαστήματος ισούται με τον αριθμητικό μέσο όρο του διαστήματος. Οπότε βρίσκουμε την κεντρική τιμή για κάθε διάστημα και την θέτουμε ως x_i και ακολουθούμε την μεθοδολογία της απλής κατανομής

Παράδειγμα (7.2)

Μετρήσαμε το βάρος 50 ατόμων και πήραμε τον παρακάτω πίνακα συχνοτήτων

Βάρος σε kgr	45-55	55-65	65-75	75-85	85-95
Πλήθος ατόμων	8	12	11	10	9

Να υπολογισθεί η μέση τιμή (μέσος όρος) του βάρους των ατόμων των ομαδοποιημένων μετρήσεων

Λύση:

Το πρώτο που κάνουμε είναι να υπολογίσουμε την κεντρική τιμή (μέσος όρος) κάθε κλάσης (διαστήματος) x_i και ύστερα να βρούμε το γινόμενο $\nu_i x_i$ και να τα τοποθετήσουμε στον πίνακα σε διπλανές στήλες, και να βρούμε το άθροισμα των γινόμενων $\sum \nu_i x_i$

Βάρος σε kgr	Πλήθος ν_i	κ.τ = x_i	$\nu_i x_i$
45-55	8,0	50	400
55-65	12	60	720
65-75	11	70	770
75-85	10	80	800
85-95	9,0	90	810
Σύνολο	50		3500
	$\sum \nu_i = n$		$\sum \nu_i x_i$

Η μέση τιμή $\bar{x}$ βρίσκεται από τη σχέση

$$\bar{x} = \frac{\sum \nu_i x_i}{n} = \frac{3500}{50} = 70kg$$

Ιδιότητες της μέσης τιμής (αριθμητικού μέσου):

Η μέση τιμή προσφέρεται ιδιαίτερα σε μαθηματικούς χειρισμούς εξ αιτίας δύο θεμελιωδών ιδιοτήτων που διαθέτει:

α) Το αλγεβρικό άθροισμα των αποκλίσεων (διαφορών) κάθε μέτρησης από την μέση τιμή ισούται με μηδέν και

β) Το άθροισμα των τετραγώνων των αποκλίσεων των μετρήσεων από την μέση τιμή είναι μικρότερο από το άθροισμα των τετραγώνων των αποκλίσεων των μετρήσεων από οποιαδήποτε άλλη τιμή.

Αξίζει να σημειωθεί ότι ο δειγματικός μέσος όρος (μέση τιμή ή αριθμητικός μέσος) εκτός από τις παραπάνω διαθέτει και άλλες ιδιότητες όπως:

1)Αν σε μια κατανομή αντικαταστήσουμε τις τιμές x_i με λx_i δηλαδή αν $x_i' = \lambda x_i$ τότε η νέα μέση τιμή $\bar{x}_i' = \lambda \bar{x}$. Αναλυτικά έχουμε

$$\bar{x}' = \frac{1}{n}\sum_{i=1}^{n} x_i' = \frac{1}{n}\sum_{i=1}^{n} \lambda x_i = \frac{\lambda}{\nu}\sum_{i=1}^{n} x_i = \lambda \bar{x} \qquad (7.5)$$

Η παραπάνω διαδικασία ονομάζεται **σχηματισμός κλίμακα** (scale transformation)

α)Αν σε μια κατανομή αντικαθιστούμε τις τιμές x_i με $x_i+\alpha$ δηλαδή αν $x_i' = x_i + \alpha$ τότε η νέα μέση τιμή $\bar{x}_i' = \bar{x} + \alpha$.

Αναλυτικά έχουμε

$$\bar{x}' = \frac{1}{n}\sum_{i=1}^{n} x_i' = \frac{1}{n}\sum_{i=1}^{n} (x_i + \alpha) = \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} a = \bar{x} + \alpha \qquad (7.6)$$

Η παραπάνω διαδικασία ονομάζεται **μετασχηματισμός αρχής** (origin transformation).

β)Η μέση τιμή βρίσκεται μέσα στο εύρος των τιμών x_i δηλαδή ισχύει

$$x_{i\,min} \le \bar{x} \le x_{i\,max}$$

γ) Αν $\bar{x}_1$ η μέση τιμή n_1 τιμών και $\bar{x}_2$ η μέση τιμή n_2 τιμών μιας κατανομής τότε η μέση τιμή $\bar{x}$ των n_1+n_2 τιμών ισούται με

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} \qquad (7.7)$$

7.2.2 **Διάμεσος** (median)

Η διάμεσος μιας σειράς μετρήσεων οι οποίες έχουν διαταχθεί κατά αύξουσα ή φθίνουσα τάξη είναι το σημείο της κλίμακας μέτρησης που χωρίζει τη σειρά σε δύο ίσα μέρη. Δηλαδή η διάμεσος είναι η τιμή της σειράς που είναι συγχρόνως μεγαλύτερη από τις μισές μετρήσεις και μικρότερη από τις άλλες μισές.

Η διάμεσος συμβολίζεται με Δμ ή δ η m (συνήθως με δ). Για να βρούμε τη διάμεσο σε μια σειρά μετρήσεων (χωρίς πίνακα συχνοτήτων) κατατάσσουμε πρώτα τις τιμές κατά αύξουσα ή φθίνουσα σειρά. Αν ο αριθμός των τιμών είναι περιττός, τότε η διάμεσος είναι η μεσαία τιμή της σειράς, ενώ αν ο αριθμός των τιμών είναι άρτιος, τότε η διάμεσος είναι ο μέσος όρος των δύο μεσαίων τιμών της σειράς. Η διάμεσος παίρνει την τιμή της παρατήρησης εκείνης που καθορίζεται από τον όρο $\frac{n+1}{2}$, ανεξάρτητα από το αν το πλήθος των παρατηρήσεων (n) είναι άρτιος ή περιττός.

Παράδειγμα (7.3)
Ποια είναι η διάμεσος των παρακάτω 9 παρατηρήσεων
1 8 3 4,5 4 4,8 5,7 5,2 7,5
Λύση:
Πρώτα κατατάσσουμε τις τιμές κατά αύξουσα σειρά:
1 3 4 4,5 4,8 5,2 5,7 7,5 8
Η θέση της διαμέσου είναι $\frac{n+1}{2}=\frac{9+1}{2}=5^{\eta}$
Άρα η διάμεσος είναι δ=4,8

Παράδειγμα (7.4)
Προσδιορίσετε την διάμεσο στις παρακάτω παρατηρήσεις
110 150 130 115 120 118 135 140
Λύση:
Κατατάσσουμε πρώτα τις τιμές σε αύξουσα σειρά
110 115 118 120 130 135 140 150
Η θέση της διαμέσου είναι $\frac{n+1}{2}=\frac{8+1}{2}=4{,}5^{\eta} \Rightarrow \delta=\frac{4^{\eta}+5^{\eta}}{2}$
Άρα $\delta=\frac{120+130}{2}=125$

Όταν οι παρατηρήσεις παρουσιάζονται σε μορφή απλής κατανομής τότε για τον υπολογισμό της διαμέσου ακολουθούμε την εξής πορεία: Τακτοποιούμε τις τιμές της κλίμακας κατά αύξουσα σειρά, βρίσκουμε την αθροιστική συχνότητα για κάθε τιμή και η θέση της διαμέσου θα είναι $\frac{n+1}{2}$.

Παράδειγμα (7.5)
Για τον επόμενο πίνακα παρατηρήσεων

x_i	10	20	30	40	50
v_i	10	35	40	10	5

Να υπολογισθεί η διάμεσος.

Λύση:
Φτιάχνουμε τον πίνακα με τις αθροιστικές συχνότητες

x_i	v_i	N_i
10	10	10
20	35	45
30	40	85
40	10	95
50	5	100
Σύνολο	100	

Επειδή n=100 άρτιος αριθμός η διάμεσος είναι η τιμή που έχει θέση $\frac{100+1}{2} = 50{,}5^{\eta}$, δηλαδή είναι ανάμεσα στην 50^{η} και την 51^{η} ή

$$\delta = \frac{50^{\eta} + 51^{\eta}}{2} = \frac{30+30}{2} = 30$$

Υπολογισμός της διαμέσου σε ομαδοποιημένες κατανομές
Όταν οι παρατηρήσεις παρουσιάζονται σε μορφή ομαδοποιημένης κατανομής τότε ο υπολογισμός της διαμέσου δ γίνεται ως εξής:
α) Υπολογίζουμε την αθροιστική συχνότητα για κάθε κλάση (διάστημα) της κατανομής.
β) Εντοπίζουμε την κλάση (διάστημα) με αθροιστική συχνότητα ίση ή μεγαλύτερη από $\frac{n}{2}$, στην οποία βρίσκεται η διάμεσος .
γ) Η τιμή της διαμέσου δ υπολογίζεται από τον παρακάτω τύπο

$$\delta = L_i + \frac{\frac{n}{2} - N_{i-1}}{\nu_i} \cdot C \qquad (7.8)$$

όπου L_i= το κατώτερο όριο της κλάσης που περιέχει την διάμεσο
n= το σύνολο των παρατηρήσεων
N_{i-1} = η αθροιστική συχνότητα της προηγούμενης κλάσης από την κλάση της διαμέσου
v_i= η απόλυτη συχνότητα της κλάσης που περιέχει τη διάμεσο
C= το εύρος της κλάσης

Παράδειγμα(7.6)
Μετρήθηκε το βάρος 50 ατόμων και πάρθηκε ο παρακάτω πίνακας συχνοτήτων

Βάρος σε kg	45-55	55-65	65-75	75-85	85-95
Πλήθος ατόμων	8	12	11	10	9

Να υπολογισθούν η μέση τιμή και η διάμεσος
Λύση
Για τον υπολογισμό της μέσης τιμής χρειαζόμαστε την κεντρική τιμή x_i για κάθε κλάση, το γινόμενο $v_i x_i$ και το $\sum v_i x_i$, ενώ για τον υπολογισμό της διαμέσου χρειαζόμαστε την αθροιστική συχνότητα κάθε κλάσης της κατανομής. Κατασκευάζουμε πίνακα με τις στήλες που χρειάζονται και υπολογίζουμε τα μεγέθη x_i, $v_i x_i$, N_i και $\sum v_i x_i$ όπως φαίνεται παρακάτω:

Διάστημα βάρους	χ_i	v_i	N_i	$v_i x_i$
45-55	50	8	8	400
55-65	60	12	20	720
65-75	70	11	31	770
75-85	80	10	41	800
85-95	90	9	50	810
Σύνολο		50		3500

Η μέση τιμή $\bar{x}$ υπολογίζεται από τη σχέση

$$\bar{x} = \frac{\sum_{i=1}^{n} v_i x_i}{\sum v_i} = \frac{3500}{50} = 70 kg \,/ \alpha\tau o\mu o$$

Η διάμεσος υπολογίζεται από τον τύπο:

$$\delta = Li + \frac{\frac{n}{2} - N_{i-1}}{v_i} \cdot C$$

Το διάστημα που έχει τη διάμεσο είναι 65-75 διότι έχει αθροιστική συχνότητα 31 (η τιμή με $\frac{51}{2}$ θέση βρίσκεται στο διάστημα αυτό) οπότε

L_i=65 , n=50 , N_{i-1}=20 , v_i=11 , C=10 άρα $\delta = 65 + \dfrac{\frac{50}{2} - 20}{11} \cdot 10 = 69{,}5$

Η διάμεσος έχει συνήθως μεγαλύτερη αξία από την μέση τιμή στην περιγραφική στατιστική, διότι περιγράφει με τρόπο ικανοποιητικό την θέση ενός συνόλου τιμών επάνω στην ευθεία των πραγματικών αριθμών. Έχει όμως ορισμένα μειονεκτήματα έναντι της μέσης τιμής όπως:
i) Δεν είναι μαθηματική ποσότητα που προσφέρεται για περαιτέρω υπολογισμούς. Έτσι αν πχ γνωρίζουμε τις διαμέσους δυο κατανομών δεν μπορούμε να βρούμε απευθείας την διάμεσο της κατανομής που προκύπτει από τον συνδυασμό τους.
ii) Επηρεάζεται από τις διακυμάνσεις της δειγματοληψίας περισσότερο από την μέση τιμή.

7.2.3 Η επικρατούσα τιμή (κορυφή ή δεσπόζουσα τιμή) (median)

Η επικρατούσα τιμή η κορυφή M_o ενός συνόλου παρατηρήσεων ορίζεται η παρατήρηση με την μεγαλύτερη συχνότητα. Στις ομαδοποιημένες κατανομές ο υπολογισμός της κορυφής γίνεται ως εξής:
Εντοπίζουμε την κλάση (το διάστημα) με την μεγαλύτερη συχνότητα (επικρατούσα κλάση) όπου η επικρατούσα τιμή βρίσκεται σε αυτή την κλάση.
Η επικρατούσα τιμή υπολογίζεται από τον τύπο

$$M_o = L_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot C \qquad (7.9)$$

όπου L_i=το κατώτερο όριο της επικρατούσας κλάσης
Δ_1= v_i-v_{i-1} (διαφορά μεταξύ της μεγαλύτερης συχνότητας και της συχνότητας της προηγούμενης κλάσης)
Δ_2= v_i-v_{i+1} (διαφορά μεταξύ της μεγαλύτερης συχνότητας και της συχνότητας της επόμενης κλάσης)
C= το εύρος της κλάσης

7.2. 4 Ποσοστημόρια-εκατοστημόρια

Το P-ποσοστημόριο είναι η παρατήρηση ή το σημείο μιας ομάδας μετρήσεων για την οποία ισχύει: ποσοστό 100 P % των παρατηρήσεων έχουν τιμή μικρότερη από αυτή και ποσοστό 100(1-P)% των παρατηρήσεων έχουν τιμή μεγαλύτερη. Το ποσοστημόριο συμβολίζεται γενικά με P_α.

Στις ομαδοποιημένες κατανομές το ποσοστημόριο υπολογίζεται από τον τύπο:

$$P_\alpha = L_i + \frac{\alpha \cdot n - N_{i-1}}{V_i} C \qquad (7.10)$$

όπου L_i= κατώτερο όριο της κλάσης που έχει το ποσοστημόριο

α= το ποσοστό

N_{i-1} = η αθροιστική συχνότητα της κλάσης που είναι πριν την κλάση που περιέχει το ποσοστημόριο, v_i= η συχνότητα της κλάσης που περιέχει το ποσοστημόριο , C= το εύρος της κλάσης.

Μεγάλη σημασία έχουν τα ποσοστά που αντιστοιχούν σε 25% και 75% (P=0,25 και P=0,75) των παρατηρήσεων τα οποία ονομάζονται πρώτο και τρίτο τεταρτημόριο και συμβολίζονται με Q_1 για το 1^o τεταρτημόριο και Q_3 για το 3^o τεταρτημόριο. Τα Q_1 και Q_3 υπολογίζονται, με βάση τον παραπάνω τύπο, από τους τύπους

$$Q_1 = L_i + \frac{\frac{n}{4} - N_{i-1}}{v_i} \cdot C \qquad (7.11)$$

$$Q_3 = L_i + \frac{\frac{3n}{4} - N_{i-1}}{v_i} \cdot C \qquad (7.12)$$

Παράδειγμα (7.7)

Στον πίνακα που ακολουθεί φαίνονται ομαδοποιημένες σε οκτώ κλάσεις, οι συχνότητες των παρατηρήσεων που πάρθηκαν από 200 σταθμούς της χώρας για το ύψος της ετήσιας βροχόπτωσης x.

x_{cm}	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
v_i	11	14	31	41	48	30	15	10

Να υπολογισθούν:

α) η διάμεσος (δ)

β) το πρώτο και τρίτο τεταρτημόριο (Q_1), (Q_3)

γ) η επικρατούσα κλάση και η κορυφή (M_o).

Λύση

Υπολογίζουμε πρώτα την αθροιστική συχνότητα N_i για κάθε κλάση και τις τοποθετούμε σε αντίστοιχη στήλη στον πίνακα όπως φαίνεται παρακάτω

X	ν_i	N_i
20-30	11	11
30-40	14	25
40-50	31	56
50-60	41	97
60-70	48	145
70-80	30	175
80-90	15	190
90-100	10	200
Σύνολο	200	

α)Για τον υπολογισμό της διαμέσου χρησιμοποιούμε την σχέση (7.8)

$$\delta = L_i + \frac{\frac{n}{2} - N_{i-1}}{\nu_i} \cdot C$$

Η κλάση που περιέχει την διάμεσο είναι αυτή που έχει αθροιστική συχνότητα 145 (περιέχει την 100^{η} και την 101^{η} τιμή), άρα:

L_i=60, $\frac{n}{2} = 100$, N_{i-1}=97, ν_i=48 και c=10

Οπότε $\delta = 60 + \frac{100-97}{48} \cdot 10 = 60,6\ cm$

β) Για το Q_1 χρησιμοποιούμε τον τύπο

$$Q_1 = L_i + \frac{\frac{n}{4} - N_{i-1}}{v_i} \cdot C$$

Η κλάση που περιέχει το 25% των μετρήσεων είναι αυτή που αντιστοιχεί σε $\frac{200}{4} = 50$ δηλαδή η κλάση (40-50), οπότε:

L_i=40, N_{i-1}=25, v_i=31 και c=30, άρα

$$Q_1 = 40 + \frac{50-25}{31} \cdot 10 = 48$$

Για το Q_3 έχουμε την σχέση:

$$Q_3 = L_i + \frac{\frac{3n}{4} - N_{i-1}}{v_i} \cdot c$$

Η κλάση που έχει το Q_3 είναι αυτή που αντιστοιχεί σε αθροιστική συχνότητα $\frac{3 \cdot 200}{4} = 150$ η οποία είναι (70-80) οπότε:

L_i=70, N_{i-1}=145, v_i=30 και c=10, άρα:

$$Q_3 = 70 + \frac{150-145}{30} \cdot 10 = 71,7\ cm$$

γ) Η επικρατούσα κλάση είναι αυτή που έχει την μεγαλύτερη απόλυτη συχνότητα άρα είναι η κλάση (60-70).

Η κορυφή M_o υπολογίζεται από τον τύπο $M_o = L_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot C$

L_i=60, Δ_1=48-41=7, Δ_2=48-30=18 και C=10,
οπότε $M_o = 60 + \frac{7}{7+18} \cdot 10 = 62{,}8cm$

7.2.5 Γεωμετρικός μέσος

Ο Γεωμετρικός μέσος συμβολίζεται με g και χρησιμεύει κυρίως στον υπολογισμό του μέσου ρυθμού της μεταβολής ενός φαινομένου (πχ μέσος ρυθμός αύξησης πωλήσεων σε έξι διαδοχικά χρόνια), στον υπολογισμό του μέσου αναλογιών και στον ανατοκισμό. Ο γεωμετρικός μέσος υπολογίζεται με τον ακόλουθο τύπο

$$g = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot ....x_\nu} \qquad (7.13)$$

Ο υπολογισμός του g διευκολύνεται με την λογαρίθμηση του παραπάνω τύπου δηλαδή

$$\log g = \frac{1}{n}(\log x_1 + \log x_2 + ... + \log x_n) = \frac{1}{n}\sum_{i=1}^{n} \log x_i \qquad (7.14)$$

Παράδειγμα (7.8)
Να υπολογισθεί ο μέσος ρυθμός αυξήσεων στις πωλήσεις μιας εταιρίας που έχει τις ακόλουθες συνεχείς ετήσιες αυξήσεις:
15%, 17%, 27%, 40%, και 70%
Λύση:
Οι ετήσιες μεταβολές (συντελεστές αυξήσεων) είναι: 1,15, 1,17, 1,27, 1,40, 1,70
Ο γεωμετρικός τους μέσος είναι
$g = \sqrt[5]{1{,}15 \cdot 1{,}17 \cdot 1{,}27 \cdot 1{,}40 \cdot 1{,}70} = 1{,}324$
Άρα ο μέσος ρυθμός αύξησης των πωλήσεων είναι 32,4%

7.3 Δείκτες ή μέτρα διασποράς

Δείκτες ή μέτρα διασποράς ονομάζονται τα μέτρα που εκφράζουν τις αποκλίσεις των τιμών μιας μεταβλητής γύρω από τα μέτρα κεντρικής τάσης, δηλαδή είναι στατιστικοί δείκτες που εκφράζουν αριθμητικά την ανομοιογένεια των τιμών. Έτσι, ενώ οι κατανομές συχνότητας δείχνουν πως κατανέμονται οι τιμές της ομάδας κατά μήκος της κλίμακας μέτρησης αναλυτικά, οι δείκτες διασποράς δείχνουν την μεταβλητότητα των τιμών συνοπτικά με ένα αριθμητικό δείκτη.

Οι τιμές των δεδομένων ενός φαινομένου έχουν την ιδιότητα να συσσωρεύονται γύρω από κάποια "κεντρική τιμή" αλλά συγχρόνως να είναι και διασπαρμένες σε ένα διάστημα, μικρό ή μεγάλο, γύρω από αυτήν την τιμή.

Οι κυριότεροι αριθμητικοί δείκτες με τους οποίους εκφράζεται ο βαθμός διασποράς των τιμών ποσοτικών δεδομένων σε μια κατανομή συχνοτήτων είναι:

1.Εύρος – Κύμανση (range)

Το εύρος της κατανομής μιας ομάδας μετρήσεων ορίζεται ως η διαφορά μεταξύ της μεγαλύτερης και μικρότερης τιμής της κατανομής. Συμβολίζεται με R και δίνεται από τον τύπο

$$R = X_{\text{μέγιστη}} - X_{\text{ελάχιστη}} \qquad (7.15)$$

Το εύρος αν και είναι πολύ εύκολο στον υπολογισμό του, δεν θεωρείται αξιόπιστο μέτρο διασποράς διότι βασίζεται μόνο στις δυο ακραίες τιμές και δεν επηρεάζεται καθόλου από την κατανομή των υπόλοιπων τιμών στο ενδιάμεσο διάστημα.

2. Ενδοτεταρτημοριακό και Ημιενδοτεταρτημοριακό εύρος (interquantile range)

Το ενδοτεταρτημοριακό εύρος ορίζεται ως η διαφορά του πρώτου τεταρτημορίου Q_1 από το τρίτο τεταρτημόριο Q_3, $(Q_3 – Q_1)$. Το ημιενδοτεταρτημοριακό εύρος (semi-interquantile range) ορίζεται ως το μισό της διαφοράς $Q_3 – Q_1$ και συμβολίζεται με Q.

$$Q = \frac{Q_3 - Q_1}{2} \qquad (7.16)$$

3.Μέση απόκλιση (mean deviation)

Ως μέση απόκλιση των τιμών μιας σειράς παρατηρήσεων, ορίζεται ο αριθμητικός μέσος των απολύτων τιμών των αποκλίσεων των τιμών της μεταβλητής, από τη μέση τιμή των παρατηρήσεων. Συμβολίζεται με MD και δίνεται από την σχέση

$$MD = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \overline{x}\right| \qquad (7.17)$$

ή για ομαδοποιημένες παρατηρήσεις

$$\mathrm{MD} = \frac{\sum_{i=1}^{n}\nu_i\left|x_i - \overline{x}\right|}{\sum_{i=1}^{n}\nu_i} \qquad (7.18)$$

4. Διασπορά ή Διακύμανση (variance) **ή μεταβλητότητα**

Είναι το μέτρο του βαθμού της διασποράς, συμβολίζεται με σ^2 όταν πρόκειται για το σύνολο του πληθυσμού, και με S^2 όταν πρόκειται για δείγμα του πληθυσμού (δειγματική διασπορά). Η διακύμανση (διασπορά) ενός πληθυσμού σ^2 υπολογίζεται ως

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \mu\right)^2 \qquad (7.19)$$

ενώ η δειγματική διασπορά S^2 υπολογίζεται ως

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 \qquad (7.20)$$

Όταν το μέγεθος του δείγματος είναι μικρό, η εκτίμηση της διασποράς είναι σωστότερη αν ο παρανομαστής n αντικατασταθεί με το (n-1) οπότε η σχέση (7.20) γίνεται

$$S^2 = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n-1} \qquad (7.21)$$

Ο όρος $\sum(x_i - \overline{x})^2$ μπορεί να μετασχηματισθεί αλγεβρικά στο ισοδύναμο $\sum x^2 - \frac{\left(\sum x\right)^2}{n}$ ή σε $\left[\sum x_i^2 - n(\overline{x})^2\right]$,οπότε η σχέση (7.21) γίνεται

$$S^2 = \frac{\sum x_i^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{\left[\sum x_i^2 - n\overline{x}^2\right]}{n-1} \qquad (7.22)$$

Η διασπορά είναι η κυριότερη παράμετρος μεταβλητότητας. Όταν η τιμή της διασποράς είναι μικρή τότε οι τιμές του δείγματος (συνόλου) δεν διαφέρουν πολύ από την μέση τιμή τους, ενώ αντίθετα αν η τιμή της διασποράς είναι μεγάλη τότε οι τιμές διαφέρουν πολύ από την μέση τιμή και είναι σκορπισμένες σε μεγάλη απόσταση από την μέση τιμή.

5.Τυπική απόκλιση ή σταθερή απόκλιση (standard deviation)
Η τυπική (σταθερή) απόκλιση αποτελεί αξιόλογο μέτρο του βαθμού διασποράς των παρατηρήσεων, πολύτιμη έννοια γενικά και ειδικά στη βιοστατιστική. Ορίζεται ως η τετραγωνική ρίζα της διασποράς και συμβολίζεται με σ όταν εκφράζει το σύνολο του πληθυσμού και με s όταν εκφράζει δείγμα του πληθυσμού, οπότε

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x_i - \mu)^2}{n-1}} = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}}$$

και $$S = \sqrt{S^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}} \qquad (7.23)$$

Η σταθερά απόκλισης εκφράζεται στην μονάδα μέτρησης των τιμών x_i, σε αντίθεση με την διασπορά η οποία εκφράζεται στην μονάδα μέτρησης των τιμών υψωμένο στο τετράγωνο.

Όταν οι παρατηρήσεις δίνονται σε πίνακα συχνοτήτων ή όταν είναι ομαδοποιημένες σε πίνακα συχνοτήτων με κλάσεις τότε η τυπική απόκλιση δίνεται από τον ισοδύναμο τύπο

$$S = \sqrt{\frac{\sum \nu_i x_i^2 - \frac{\left(\sum \nu_i x_i\right)^2}{n}}{n-1}} \qquad (7.23)$$

όπου x_i είναι η κεντρική τιμή της κλάσης στις ομαδοποιημένες παρατηρήσεις.

Αξίζει να σημειωθεί ότι η διακύμανση και η τυπική απόκλιση έχουν τις εξής ιδιότητες

α) Αν κάθε τιμή x_i πολλαπλασιασθεί επί μία σταθερά λ τότε η νέα διακύμανση $S'^2=\lambda^2 S^2$ ενώ η τυπική απόκλιση $S'=\lambda S$

β) Αν σε κάθε τιμή x_i προστεθεί μια σταθερά α τότε η διακύμανση και η τυπική απόκλιση δεν μεταβάλλονται $S'^2 = S^2$, $S' = S$

Παράδειγμα (7.9)
Σε μια ομάδα 9 ψαριών χορηγήθηκε για ένα χρονικό διάστημα μια συγκεκριμένη δίαιτα. Μετά το τέλος της δίαιτας, μετρήθηκε η αύξηση του βάρους και είχαμε τα εξής αποτελέσματα:
Αύξηση βάρους X: 60 70 80 65 90 90 50 50 30
Να βρεθούν: α) Η δειγματική μέση τιμή, β) η διασπορά, γ) η τυπική απόκλιση του δείγματος
Λύση
α) Η δειγματική μέση τιμή βρίσκεται από την σχέση

$$\bar{x} = \frac{\sum_{i=1}^{9} x_i}{9} = \frac{60+70+80+65+90+90+50+50+30}{9} = 6$$

β) Η διασπορά S^2 βρίσκεται είτε από την σχέση

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

οπότε χρειαζόμαστε δύο στήλες η μία με την διαφορά $x_i - \bar{x}$ και η άλλη με το τετράγωνο της διαφοράς $(x_i - \bar{x})^2$ είτε από την σχέση 7.22

$S^2 = \frac{1}{n-1}\left[\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right]$ οπότε χρειαζόμαστε στήλη με το x_i^2. Θα λύσουμε την άσκηση και με τους δύο τύπους. Φτιάχνουμε τον πίνακα με τις απαιτούμενες στήλες:

	x_i	x_i^2	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	60	3600	-5	25
2	70	4900	5	25
3	80	6400	15	225
4	65	4225	0	0
5	90	8100	25	625
6	90	8100	25	625
7	50	2500	-15	225
8	50	2500	-15	225
9	30	900	-35	1225
Σύνολο	585	41225	0	3200

$$\bar{x} = \frac{585}{9} = 65$$

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{3200}{8} = 400 \;\; g^2, \text{ οπότε } S = \sqrt{S^2} = \sqrt{400} = 20 \quad ή$$

$$S^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{41225 - \frac{(585)^2}{9}}{8} = 400 g^2 \;\; \Rightarrow S = 20 \; g$$

Παράδειγμα (7.10)
Δίνεται η κατανομή

x_i	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100	Σύνολο
ν_i	11	48	31	41	30	14	15	10	200

Να βρεθούν α) Η μέση τιμή $\bar{x}$, β) Η τυπική απόκλιση S, γ) τα Q_1 και Q_3, δ) το ενδοτεταρτημοριακό εύρος Q

Λύση

Επειδή έχουμε ομαδοποιημένες παρατηρήσεις χρειαζόμαστε την κεντρική τιμή κάθε κλάσης για τον υπολογισμό της μέσης τιμής και της τυπικής απόκλισης. Για τον υπολογισμό της S χρησιμοποιούμε την σχέση 3.23

$$S = \sqrt{\frac{\sum \nu_i x_i^2 - \frac{\left(\sum \nu_i x_i\right)^2}{n}}{n-1}}$$

άρα χρειαζόμαστε να φτιάξουμε πίνακα με πρόσθετες στήλες $x_i, x_i^2, \nu_i x_i, \nu_i x_i^2$ και μια στήλη με αθροιστική συχνότητα για τον υπολογισμό των Q_1 και Q_3. Ο πίνακας με τους υπολογισμούς φαίνεται παρακάτω

Κλάσεις	ν_i	x_i	N_i	$\nu_i x_i$	x_i^2	$\nu_i x_i^2$
20 – 30	11	25	11	275	625	6875
30 – 40	48	35	59	1680	1225	58800
40 – 50	31	45	90	1395	2025	62775
50 – 60	41	55	131	2255	3025	124025
60 – 70	30	65	161	1950	4225	126750
70 – 80	14	75	175	1050	5625	78750
80 – 90	15	85	190	1275	7225	108375
90 – 100	10	95	200	950	9025	90250
Σύνολο	200			10830		656600
	n			$\sum \nu_i x_i$		$\sum \nu_i x_i^2$

α) Η μέση τιμή $\bar{x} = \dfrac{\sum \nu_i x_i}{\sum \nu_i} = \dfrac{10830}{200} = 54{,}15$

β) Η τυπική απόκλιση $S = \sqrt{\dfrac{\sum \nu_i x_i^2 - \dfrac{\left(\sum \nu_i x_i\right)^2}{n}}{n-1}}$

$$S = \sqrt{\frac{656600 - \dfrac{(10830)^2}{200}}{199}} = 18{,}78$$

γ) Για τον υπολογισμό του Q_1 χρησιμοποιούμε τον τύπο

$$Q_1 = L_i + \frac{\dfrac{n}{4} - N_{i-1}}{\nu_i} \cdot C$$

και για το Q_3 τον τύπο

$$Q_3 = L_i + \frac{\dfrac{3n}{4} - N_{i-1}}{\nu_i} \cdot C$$

Η κλάση που περιέχει το Q_1 είναι αυτή που περιέχει την παρατήρηση

$\frac{n}{4} = \frac{200}{4} = 50^n$ δηλαδή η κλάση (30 – 40), ενώ η κλάση που περιέχει το Q_3 είναι αυτή που περιέχει την παρατήρηση με θέση

$\frac{3n}{4} = \frac{3 \cdot 200}{4} = 150^n$ δηλαδή η κλάση (60 – 70) οπότε

$$Q_1 = 30 + \frac{\frac{200}{4} - 11}{48} \cdot 10 = 38,1$$

$$Q_3 = 30 + \frac{150 - 131}{30} \cdot 10 = 66,3$$

δ) $Q = \frac{Q_3 - Q_1}{2} = \frac{66,3 - 38,1}{2} = 14,1$

7.4 Μέτρα σχετικής μεταβλητότητας (διασποράς)

Ως μέτρο της σχετικής μεταβλητότητας ορίζεται ο συντελεστής μεταβλητότητας (coefficient of variation) CV ο οποίος εκφράζεται ως ο λόγος της δειγματικής τυπικής απόκλισης προς την δειγματική μέση τιμή

δηλαδή $CV = \frac{S}{\bar{x}}$ (7.24).

Συνήθως εκφράζεται σαν ποσοστό, δηλαδή

$$CV = \frac{S}{\bar{x}} \cdot 100\% \qquad (7.25)$$

Ο συντελεστής μεταβλητότητας είναι ένα μέτρο σχετικής μεταβλητότητας των τιμών και όχι της απόλυτης μεταβλητότητας. Χρησιμοποιείται για τη σύγκριση της μεταβλητότητας α) ομάδων τιμών, οι οποίες εκφράζονται στην ίδια μονάδα μέτρησης αλλά έχουν εντελώς διαφορετικές μέσες τιμές, β) δύο ή περισσοτέρων διαφορετικών μεγεθών (π.χ. ανάστημα και βάρος, συστολική και διαστολική πίεση).

Με τον συντελεστή μεταβλητότητας μπορεί να εκτιμηθεί η ομοιογένεια ενός δείγματος τιμών μιας μεταβλητής. Γενικά εάν ο συντελεστής μεταβλητότητας CV δεν ξεπερνά το 10% τότε δεχόμαστε ότι το δείγμα είναι ομοιογενές. Δηλαδή όσο πιο μικρό CV έχει ένα δείγμα τόσο πιο ομοιογενές είναι.

Παράδειγμα(7.11)

Σε ένα βιοχημικό εργαστήριο καταγράφηκαν οι τιμές του ουρικού οξέος σε 300 υγιείς άντρες (σε mg/100mL) της ίδιας ηλικίας. Οι τιμές ομαδοποιήθηκαν σε 12 κλάσεις και η κατανομή τους βρέθηκε:

Ουρικό οξύ mg/100mL	Συχνότητα
3,0 – 3,4	3
3,5 – 3,9	17
4,0 – 4,4	35
4,5 – 4,9	45
5,0 – 5,4	58
5,5 – 5,9	52
6,0 – 6,4	41
6,5 – 6,9	20
7,0 – 7,4	18
7,5 – 7,9	5
8,0 – 8,4	2
8,5 – 8,9	4
	300

α) Να υπολογισθούν:

i. Η δειγματική μέση τιμή $\bar{x}$
ii. Η διασπορά και η τυπική απόκλιση
iii. Η διάμεσος, η κορυφή, τα Q_1 και Q_3
iv. Ο συντελεστής μεταβλητότητας

β) Να κατασκευασθούν το ραβδόγραμμα συχνοτήτων και αθροιστικών συχνοτήτων.

Λύση:

Όπως αναφέραμε στη λύση του παραδείγματος (), για να υπολογισθούν η $\bar{x}$, S^2, S και δ χρειάζεται να φτιάξουμε πίνακα με τις εξής πρόσθετες στήλες: (κ.τ = x_i), x_i^2, $v_i\, x_i$, $v_i\, x_i^2$

Επίσης για τον υπολογισμό των Q_1 και Q_3 χρειαζόμαστε στήλη με αθροιστική συχνότητα. Ο πίνακας με τους υπολογισμούς φαίνεται παρακάτω:

Κλάσεις	v_i	**κ.τ=x_i**	N_i	$v_i\ x_i$	x_i^2	$v_i\ x_i^2$
3,0-3,4	3	3,2	3	9,6	10,24	30,72
3,5-3,9	17	3,7	20	62,9	13,69	232,73
4,0-4,4	35	4,2	55	147	17,64	617,4
4,5-4,9	45	4,7	100	211,5	22,09	994,05
5,0-5,4	58	5,2	158	301,6	27,04	1568,32
5,5-5,9	52	5,7	210	296,4	32,49	1689,48
6,0-6,4	41	6,2	251	254,2	38,44	1576,04
6,5-6,9	20	6,7	271	134	44,89	897,8
7,0-7,4	18	7,2	289	129,6	51,84	933,12
7,5-7,9	5	7,7	294	38,5	59,29	296,45
8,0-8,4	2	8,2	296	16,4	67,24	134,48
8,5-8,9	4	8,7	300	34,8	75,69	302,76
	300			**1636,5**		**9273,35**

i) Η δειγματική μέση τιμή $\bar{x} = \dfrac{\sum v_i x_i}{\sum v_i} = \dfrac{1636{,}5}{300} = 5{,}46$

ii) $S^2 = \dfrac{\sum x_i^2 - \dfrac{(\sum x_i)^2}{n}}{n-1} = \dfrac{9273{,}35 - \dfrac{(1636{,}5)^2}{300}}{299} = 1{,}16$

$S = \sqrt{S^2} = 1{,}08$

iii) Για τον υπολογισμό της διαμέσου δ χρησιμοποιούμε τον τύπο

$$\delta = L_i + \frac{\frac{n}{2} - N_{i-1}}{v_i} \cdot C$$

Η κλάση που περιέχει την διάμεσο δ είναι η κλάση που περιέχει την παρατήρηση με την θέση $\dfrac{300}{2} = 150^{\eta}$ δηλαδή το διάστημα (5,0-5,4) άρα

Li=5,0 , v_i=58 , N_{i-1}=100 , C= 0,5 $\Rightarrow$

$$\delta = 5{,}0 + \frac{150-100}{58} \cdot 0{,}5 = 5{,}4\, mg/100\, ml.$$

Για τον υπολογισμό της κορυφής M_o χρησιμοποιούμε τον τύπο

$$M_o = L_i + \frac{\Delta_1}{\Delta_1+\Delta_2}\cdot C = 5,0 + \frac{(58-45)}{(58-45)+(58-52)}\cdot 0,5 = 5,1$$

Για τον υπολογισμό των Q_1 και Q_3 χρησιμοποιούμε τους τύπους

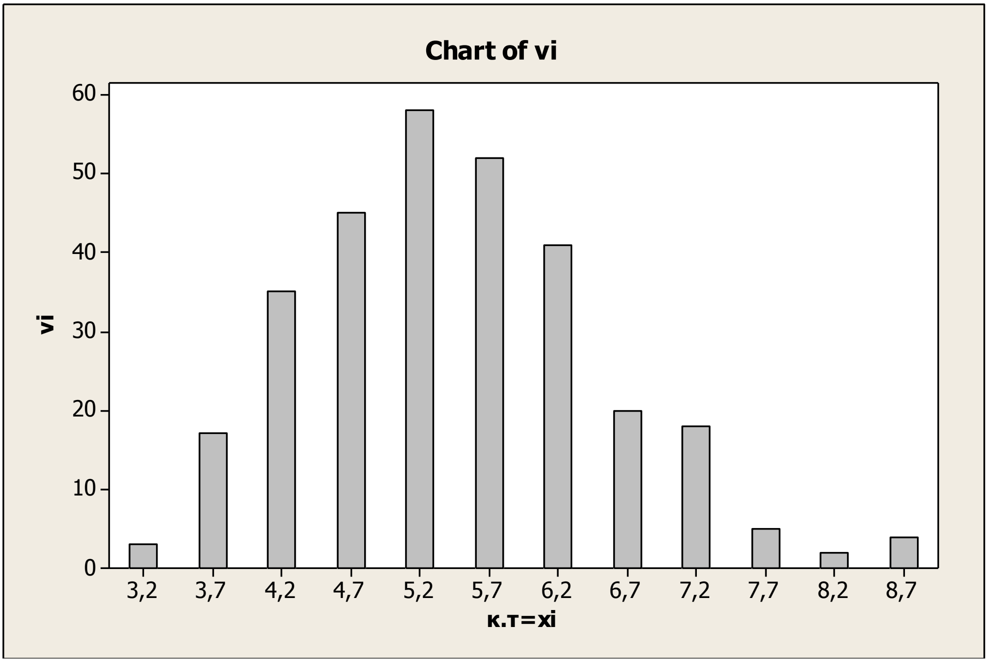
$$Q_1 = L_i + \frac{\frac{n}{4} - N_{i-1}}{\nu_i}\cdot C \quad \text{και} \quad Q_3 = L_i + \frac{\frac{3n}{4} - N_{i-1}}{\nu_i}\cdot C$$

Το διάστημα που περιέχει το Q_1 είναι (4,5-4,9) και το Q_3 είναι (6,0-6,4) οπότε

$$Q_1 = 4,5 + \frac{75-55}{35}\cdot 0,5 = 4,8 \quad , \text{ και } Q_3 = 6,0 + \frac{225-210}{41}\cdot 0,5 = 6,2$$

iv) $CV = \frac{S}{\bar{x}}\cdot 100 = \frac{1,1\cdot 100}{5,45} = 20,2\%$

β) Το ραβδόγραμμα συχνοτήτων και αθροιστικών συχνοτήτων φαίνονται στα παρακάτω σχήματα :



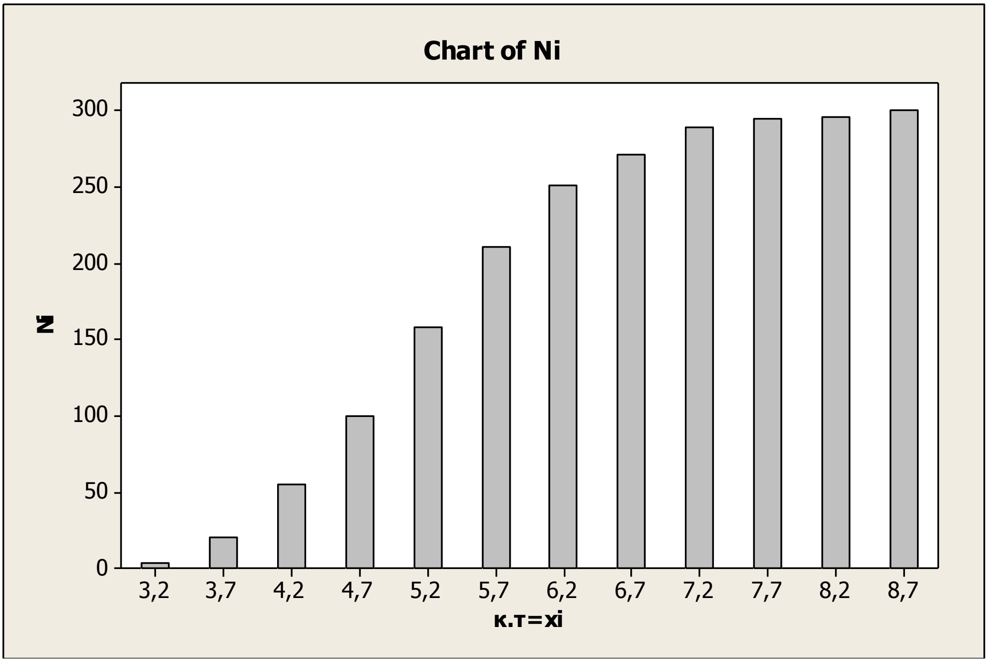


Chart of Ni
Ni
300
250
200
150
100
50
0
3,2
3,7
4,2
4,7
5,2
5,7
6,2
6,7
7,2
7,7
8,2
8,7
к.т=xi

ΑΣΚΗΣΕΙΣ

7.1 Σε ένα δείγμα από 212 κομμάτια σχιστόλιθου από μια περιοχή , βρέθηκε ότι τα ποσοστά του φωσφορίτη $Ca_3\left(PO_4\right)_2$ ήταν :

Ποσοστ%	14-16	16-18	18-20	20-22	22-24	24-26	26-28	28-30
πλήθος	1	1	5	20	44	54	57	30

Να βρεθούν : η μέση τιμή , η διάμεσος , η τυπική απόκλιση, το ενδοτεταρτημοριακό εύρος και το συντελεστή μεταβλητότητας των ποσοστών του φωσφορίτη του δείγματος.

7.2 Η βαθμολογία των 235 φοιτητών/τριών που προσήλθαν στις εξετάσεις σε ένα μάθημα ήταν:

Βαθμός	0-3	3-5	5-7	7-10
Συχνότητα	141	26	47	21

i) Να υπολογισθούν η μέση τιμή , η διάμεσος , η τυπική απόκλιση, το ενδοτεταρτημοριακό εύρος και ο συντελεστής μεταβλητότητας
ii) Να κατασκευαστούν τα ιστογράμματα συχνοτήτων και αθροιστικών σχετικών συχνοτήτων.

7.3 Το ποσοστό του ιδανικού βάρους υπολογίστηκε για 18 τυχαία επιλεγμένους διαβητικούς με εξάρτηση ινσουλίνης. Τα αποτελέσματα παρατίθενται κατωτέρω:

107 119 99 114 120 104 88 114 124
116 101 121 152 100 125 114 95 117 (%)

i) Να υπολογίσετε την διάμεσο, την επικρατούσα τιμή και τον αριθμητικό μέσο.
ii). Να υπολογίσετε την δειγματική τυπική απόκλιση ,το εύρος και τον συντελεστή μεταβλητότητας.

7.4 Η κατανομή της ηλικίας σε τυχαίο δείγμα 40 εργαζομένων κατοίκων της Αθήνας δίνεται στον πίνακα:

Ηλικία	Συχνότητα	Ηλικία	Συχνότητα
15-20	3	40-45	4
20-25	6	45-50	3
25-30	7	50-55	3
30-35	7	55-60	2
35-40	4	60-65	1

(α) Να υπολογιστούν και να ερμηνευτούν ο δειγματικός μέσος, η δειγματική τυπική απόκλιση και η δειγματική διάμεσος.
(β) Να κατασκευαστούν τα ιστογράμματα συχνοτήτων και αθροιστικών συχνοτήτων.

Κεφάλαιο 8

Εκτιμητική – Διάστημα εμπιστοσύνης

8.1 Εισαγωγή

Η εκτιμητική είναι ο κλάδος της στατιστικής συμπερασματολογίας που ασχολείται με την εκτίμηση των παραμέτρων του πληθυσμού όπως η μέση τιμή και η τυπική απόκλιση. Στηρίζεται στις πληροφορίες που παίρνουμε από το δείγμα. Επειδή υπάρχουν πολλές μέθοδοι εκτίμησης, η μέθοδος που θα επιλεγεί πρέπει να δίνει τις περισσότερες πληροφορίες με τη μικρότερη αμφιβολία.

Οι μέθοδοι εκτίμησης ταξινομούνται σε δύο κατηγορίες:

α) εκτίμηση σε σημείο, δηλαδή εκτιμάται μία τιμή της παραμέτρου,

β) εκτίμηση σε διάστημα, δηλαδή εκτιμάται ένα διάστημα στο οποίο βρίσκεται η πραγματική τιμή της παραμέτρου.

Με την εκτίμηση σε σημείο ποτέ δεν απαντάμε "η μέση τιμή μ είναι ίση" αλλά εκτιμάται ίση με τη μέση τιμή $\overline{\psi}$ του δείγματος. Ενώ με την εκτίμηση σε διάστημα έχουμε εμπιστοσύνη 100(1-α)% ότι το διάστημα που προτείνουμε περιέχει την πραγματική μέση τιμή της παραμέτρου.

Η εκτίμηση σε διάστημα γίνεται πάντα με ένα συντελεστή εμπιστοσύνης (με μια πιθανότητα).Το διάστημα αυτό ονομάζεται διάστημα εμπιστοσύνης **(δ.ε)**

8.2 Διάστημα εμπιστοσύνης για τη μέση τιμή μ

Το διάστημα εμπιστοσύνης της μέσης τιμής ισχύει μόνο όταν η κατανομή του δείγματος είναι κανονική.

8.2.1 Δ.ε για μεγάλο αριθμό παρατηρήσεων n (συνήθως $n \geq 30$) και άγνωστη σ^2 του πληθυσμού, ένα προσεγγιστικό διάστημα εμπιστοσύνης

(δ.ε) για το μ με συντελεστή εμπιστοσύνης (σ.ε) 100(1-α)% είναι:

$$\left[\overline{x} - \frac{S}{\sqrt{n}} \cdot Z_{\frac{\alpha}{2}}, \overline{x} + \frac{S}{\sqrt{n}} \cdot Z_{\frac{\alpha}{2}}\right] \qquad (8.1)$$

Για $n \geq 30$ ισχύει $Z_{\frac{\alpha}{2}} = t_{n-1;\frac{\alpha}{2}}$ οπότε το διάστημα εμπιστοσύνης βρίσκεται

από τον τύπο:

$$\delta.\varepsilon = \overline{x} \pm \frac{S}{\sqrt{n}} \cdot t_{n-1;\frac{\alpha}{2}} \qquad (8.2)$$

όπου $t_{n-1;\frac{\alpha}{2}}$ και $Z_{\frac{a}{2}}$ είναι οι τιμές των τυχαίων μεταβλητών t και Z για

επίπεδο σημαντικότητας $\frac{a}{2}$, οι οποίες βρίσκονται από θεωρητικούς πίνακες που βρίσκονται στο τέλος του βιβλίου.

Το παραπάνω διάστημα εμπιστοσύνης μπορεί να παρουσιαστεί ως εξής:

$$\bar{x} - \frac{S}{\sqrt{n}} \cdot z_{\frac{a}{2}} < \mu < \bar{x} + \frac{S}{\sqrt{n}} \cdot z_{\frac{\alpha}{2}} \qquad (8.3)$$

ή το διάστημα εμπιστοσύνης δ.ε είναι: $\bar{x} \pm \frac{S}{\sqrt{n}} \cdot z_{\frac{\alpha}{2}}$ (8.4)

Παράδειγμα(8.1)

Ο πληθυσμός των τιμών των πιέσεων του αίματος τριαντάχρονων ανδρών έχει μέση τιμή μ αλλά άγνωστη τυπική απόκλιση. Ένα τυχαίο δείγμα 81 φαινομενικά κανονικών τριαντάχρονων αντρών έδωσε μέση τιμή 125 και τυπική απόκλιση 15. Να βρεθούν

i. Ένα δ.ε με συντελεστή εμπιστοσύνης 90% για το μ

ii. Ένα δ.ε με συντελεστή εμπιστοσύνης 95% για το μ

Λύση:

Για να κατασκευάσουμε ένα διάστημα εμπιστοσύνης πρώτα κοιτάμε εάν η διασπορά σ^2 είναι γνωστή ή όχι και αν το ν είναι μεγάλο ή μικρό.

Επειδή στο παράδειγμά μας η σ^2 είναι άγνωστη και $n > 30$ τότε το διάστημα εμπιστοσύνης είναι

$$\delta.\ \varepsilon = \bar{x} \pm \frac{S}{\sqrt{n}} \cdot z_{\frac{\alpha}{2}} \qquad (8.4)$$

Έχουμε $\bar{x} = 125$, S=15, n = 81, $z_{\frac{a}{2}} = 1,645$ (για α = 0,10 ή 90% σ.ε)

και $z_{\frac{a}{2}} = 1,96$ (για α = 0,05 ή 95% σ.ε) $\Rightarrow$

i. Για συντελεστή εμπιστοσύνης 90% έχουμε

$$\delta.\varepsilon = 125 \pm \frac{15}{\sqrt{81}} \cdot 1,645 = 125 \pm 2,7 \text{ ή } 122,3<\mu<127,7$$

ii. Για σ.ε 95% έχουμε:

$$\delta.\varepsilon = 125 \pm \frac{15}{\sqrt{81}} \cdot 1,96 = 125 \pm 3,3 \text{ , ή } 121,7<\mu<128,3$$

Παράδειγμα (8.2)
Για να εκτιμηθεί η μέση βαθμολογία των φοιτητών στις εξετάσεις κάποιου μαθήματος, παίρνουμε ένα δείγμα 16 φοιτητών και καταγράφουμε τη βαθμολογία τους:
10, 3, 5, 4, 7, 8, 9, 9, 8, 5, 5, 8, 6, 6, 9, 10.
Να εκτιμηθεί το μ με διάστημα εμπιστοσύνης 95% . Υποθέστε ότι η βαθμολογία στις εξετάσεις ακολουθεί κανονική κατανομή Ν(μ,4), δηλαδή $\sigma^2=4$, γνωστή διασπορά).
Λύση:
Το πρώτο βήμα που πρέπει να κάνουμε είναι να βρούμε τη δειγματική μέση τιμή της βαθμολογίας των 16 φοιτητών $\bar{x}$.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{10+3+5+4+7+8+9+9+8+5+5+8+6+6+9+10}{16} = 7$$

Το δείγμα ανήκει σε έναν πληθυσμό γνωστής τυπικής απόκλισης $\sigma = \sqrt{4} = 2$.
Το διάστημα εμπιστοσύνης της μέσης τιμής δίνεται σε αυτήν την περίπτωση από τον τύπο: $\bar{x} \pm \frac{\sigma}{\sqrt{n}} \cdot z_{\frac{\alpha}{2}}$

έχουμε $\bar{x} = 7$, $\sigma = 2$, $n = 16$ και $Z_{\frac{a}{2}} = 1{,}96$ (για α = 0,05)

Άρα $\delta.\varepsilon = 7 \pm \frac{2}{\sqrt{16}} \cdot 1{,}96 = 7 \pm 1$, ή $6 < \mu < 8$

8.2.2 Διάστημα εμπιστοσύνης για τη μέση τιμή μ για δείγματα με μικρό n ($n \leq 30$) και άγνωστη τυπική απόκλιση του πληθυσμού.

Για μικρό αριθμό παρατηρήσεων ($n \leq 30$), η τυχαία μεταβλητή $Z_{\frac{a}{2}}$ αντικαθίσταται με την τυχαία μεταβλητή t που ακολουθεί την t-κατανομή με n-1 βαθμούς ελευθερίας οπότε το διάστημα εμπιστοσύνης της δειγματικής μέσης τιμής δίνεται από τον τύπο:

$$\bar{x} - \frac{S}{\sqrt{n}} \cdot t_{(n-1),\frac{a}{2}} < \mu < \bar{x} + \frac{S}{\sqrt{n}} \cdot t_{(n-1),\frac{\alpha}{2}} \qquad (8.5)$$

ή $$\delta.\varepsilon = \bar{x} \pm \frac{S}{\sqrt{n}} \cdot t_{(n-1);\frac{a}{2}} \qquad (8.5)$$

Ο παραπάνω τύπος του δ.ε χρησιμοποιείται όταν η τυπική απόκλιση του πληθυσμού είναι άγνωστη άσχετα από τον αριθμό των μετρήσεων. Αυτό γίνεται διότι όταν $n \geq 30$ και σ άγνωστη έχουμε $t_{n-1,\frac{a}{2}} \approx Z_{\frac{a}{2}}$.

Παράδειγμα (8.3)

Ένα τυχαίο δείγμα 18 δεκάχρονων κοριτσιών έδωσε μέσο βάρος 34,5 κιλά και τυπική απόκλιση 4 κιλά. Αν οι τιμές του βάρους ακολουθούν την κανονική κατανομή $N(\mu,\sigma^2)$, να βρεθούν τα δ.ε για το μ με συντελεστή εμπιστοσύνης 90%, 95% και 99%.

Λύση:

Όπως βλέπουμε από τα δεδομένα έχουμε μικρό αριθμό μετρήσεων και άγνωστη τυπική απόκλιση του πληθυσμού, καθώς επίσης έχουμε δείγμα κανονικής κατανομής. Σε τέτοιες περιπτώσεις το διάστημα εμπιστοσύνης βρίσκεται από τον τύπο:

$$\bar{x} \pm t_{(n-1);\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} \quad \text{ή} \quad \bar{x} - t_{(n-1);\frac{a}{2}} \cdot \frac{S}{\sqrt{n}} < \mu < \bar{x} + t_{(n-1)\frac{a}{2}} \cdot \frac{S}{\sqrt{n}}$$

Οπότε για 90% δ.ε έχουμε: $t_{(n-1);\frac{a}{2}} = t_{17;0,05} = 1,740$ και το δ.ε είναι:

$$34,5 - 1,740 \cdot \frac{4}{\sqrt{18}} < \mu < 34,5 + 1,740 \cdot \frac{4}{\sqrt{18}}, \text{ ή } \quad 32,86 < \mu < 36,14$$

Για 95% έχουμε $t_{17;0,025} = 2,110$ οπότε το δ.ε είναι:

$$34,5 - 2,11 \cdot \frac{4}{\sqrt{18}} < \mu < 34,5 + 2,11 \cdot \frac{4}{\sqrt{18}}, \text{ ή } \quad 32,51 < \mu < 36,49$$

Για 99% δ.ε $t_{17;0,005} = 2,576$ οπότε το δ.ε είναι:

$$34,5 - 2,576 \cdot \frac{4}{\sqrt{18}} < \mu < 34,5 + 2,576 \cdot \frac{4}{\sqrt{18}}, \text{ ή } \quad 32,07 < \mu < 36,93$$

8.3 Διάστημα εμπιστοσύνης για τη διαφορά των μέσων τιμών δύο πληθυσμών (μεγάλα δείγματα $n_1, n_2 \geq 30$).

α) Όταν οι διασπορές σ_1^2 και σ_2^2 των δύο πληθυσμών είναι γνωστές τότε το διάστημα εμπιστοσύνης της διαφοράς των δύο μέσων τιμών (μ_1-μ_2) εκφράζεται από τον τύπο:

$$\text{δ.ε: } (\bar{x} - \bar{\psi}) \pm Z_{\frac{a}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad (8.6)$$

ή

$$(\bar{x}-\bar{\psi})-Z_{\frac{a}{2}}\cdot\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\quad,\ (\bar{x}-\bar{\psi})+Z_{\frac{a}{2}}\cdot\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\qquad(8.6)$$

β) Όταν οι διασπορές σ_1^2 και σ_2^2 των δύο πληθυσμών είναι άγνωστες τότε το διάστημα εμπιστοσύνης της διαφορά των δύο μέσων τιμών (μ_1-μ_2) εκφράζεται από τον τύπο:

$$(\bar{x}-\bar{\psi})\pm Z_{\frac{a}{2}}\cdot\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\qquad(8.7)$$

ή

$$(\bar{x}-\bar{\psi})-Z_{\frac{a}{2}}\cdot\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\quad,\ (\bar{x}-\bar{\psi})+Z_{\frac{a}{2}}\cdot\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\qquad(8.7)$$

Παράδειγμα (8.4)

Σε μια έρευνα μετρήθηκε το ύψος 169 ατόμων μιας ευρωπαϊκής χώρας και έδωσε μέση τιμή $\bar{x}_1=174$ cm με τυπική απόκλιση $S_1=9$ cm. Στην ίδια έρευνα μετρήθηκε επίσης το ύψος 256 ατόμων μιας χώρας της μέσης ανατολής και η μέση τιμή $\bar{x}_2$ βρέθηκε ίση με 171 cm με τυπική απόκλιση $S_2=8$ cm. Βρείτε 95% διάστημα εμπιστοσύνης για τη διαφορά μ_1-μ_2.

Λύση:

Έχουμε δύο ομάδες μετρήσεων με n_1και n_2 >30 που προέρχονται από δύο πληθυσμούς με άγνωστες σ_1^2 και σ_2^2, οπότε το διάστημα εμπιστοσύνης δίνεται από τη σχέση: $(\bar{x}_1-\bar{x}_2)\pm z_{\frac{a}{2}}\cdot\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$

Από τα δεδομένα έχουμε: $\bar{x}_1=174$ cm, $n_1=169$, $S_1=9$ cm, $\bar{x}_2=171$ cm, $n_2=256$, $S_2=8$ cm και από τους πίνακες για α = 0,05 η τιμή $Z_{\frac{a}{2}}=1.96$

οπότε το δ.ε είναι: $$(174-171)\pm1,96\cdot\sqrt{\frac{9^2}{169}+\frac{8^2}{256}}=3\pm1,7$$

Δηλαδή το δ.ε είναι (3-1,7) ,(3+1,7) ή (1.3 , 4.7)

Παράδειγμα (8.5)

Η ποσότητα οινοπνεύματος στο αίμα (σε mg/L) 120 τυχαία επιλεγμένων οδηγών της Αθήνας και 80 τυχαία επιλεγμένων οδηγών της Θεσσαλονίκης βρέθηκε $x_1, x_2.....x_{120}$ και $\psi_1, \psi_2...\psi_{80}$. Μετά από επεξεργασία των δεδομένων προέκυψε ότι:

$$\sum_{i=1}^{120} x_i = 120 \quad , \sum_{i=1}^{120} x_i^2 = 300, \quad \sum_{i=1}^{80} \psi_i = 100 \quad , \quad \sum_{i=1}^{80} \psi_i^2 = 600$$

α) Να υπολογιστούν οι δειγματικοί μέσοι $\overline{x}, \overline{\psi}$ καθώς και οι δειγματικές τυπικές αποκλίσεις S_1,S_2 των δύο δειγμάτων.

β) έστω μ_1 και μ_2 οι θεωρητικοί μέσοι στις δύο πόλεις. Να κατασκευαστεί ένα 95% δ.ε για το μ_1, για το μ_2 και για τη διαφορά μ_1-μ_2.

Λύση:

α) Ο δειγματικός μέσος ή η μέση τιμή υπολογίζεται από τη σχέση:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{, οπότε} \quad \overline{x} = \frac{\sum_{i=1}^{120} x_i}{120} = \frac{120}{120} = 1 \quad \text{mg/L}$$

$$\overline{\psi} = \frac{\sum_{i=1}^{80} \psi_i}{80} = \frac{100}{80} = 1,25 \quad \text{mg/L.}$$

Η τυπική απόκλιση δείγματος S υπολογίζεται από τη σχέση:

$$S = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n-1}} \quad \text{οπότε}$$

$$S_1 = \sqrt{\frac{300 - \frac{(120)^2}{120}}{119}} = 1,23 \text{ mg/L}$$

$$S_2 = \sqrt{\frac{600 - \frac{(100)^2}{80}}{79}} = 2,45 \text{mg/L}$$

β) Όπως προκύπτει από τα δεδομένα πρόκειται για δείγματα με μεγάλο αριθμό παρατηρήσεων n_1>30, n_2>30 με άγνωστες τυπικές αποκλίσεις των πληθυσμών τους. Το διάστημα εμπιστοσύνης δίνεται από τη σχέση:

$$\bar{x} \pm \frac{S}{\sqrt{n}} \cdot t_{(n-1);\frac{a}{2}}$$

Άρα για το μ₁ το δ.ε είναι: $$1 \pm \frac{1,23}{\sqrt{120}} \cdot 1,96 = 1 \pm 0,22$$

Για το μ_2 το δ.ε είναι:

$$1,25 \pm \frac{2,45}{\sqrt{80}} \cdot 1,96 = 1,25 \pm 0,54$$

Το διάστημα εμπιστοσύνης για τη διαφορά των μέσων τιμών μ_1-μ_2 δίνεται από τη σχέση:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\frac{a}{2}} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

οπότε το δ.ε είναι:

$$(1,23 - 1) \pm 1,96 \cdot \sqrt{\frac{1,23^2}{120} + \frac{2,45^2}{80}} = 0,23 \pm 0,58$$

8.4 Διάστημα εμπιστοσύνης για τη διαφορά δύο μέσων τιμών (μικρά δείγματα)

α) Για δείγματα παρμένα από κανονικούς πληθυσμούς των οποίων η διασπορές σ_1^2 και σ_2^2 είναι ίσες και άγνωστες.

Αν οι δύο πληθυσμοί από τους οποίους είναι παρμένα τα δείγματα είναι κανονικοί και ισχύει $\sigma_1^2 = \sigma_2^2$ τότε η κοινή τους διασπορά είναι σ^2.

Η κοινή διασπορά των δύο δειγμάτων συμβολίζεται με s^2 και δίνεται από τη σχέση:

$$S^2 = \frac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{n_1 + n_2 - 2} \qquad (8.8)$$

Όταν ο αριθμός των παρατηρήσεων για κάθε δείγμα είναι μικρότερος από το 30 τότε το διάστημα εμπιστοσύνης της διαφοράς των δύο μέσων τιμών μ_1-μ_2 δίνεται από τον τύπο:

$$(\bar{x} - \bar{\psi}) \pm t_{(n_1+n_2-2);\frac{a}{2}} \cdot S \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad (8.9)$$

όπου $t_{(n_1+n_2-2);\frac{a}{2}}$ η τιμή της t-κατανομής για $n_1 + n_2 - 2$ βαθμούς ελευθερίας και πιθανότητα α/2.

Παράδειγμα (8.6)
Δύο μηχανές κατασκευάζουν πυκνωτές. Από τυχαία δείγματα μεγέθους n_1=16 και n_2=13 από τις δύο μηχανές αντίστοιχα προέκυψαν:

$\overline{x}_1 = 52$ μF, $\overline{x}_2 = 48$ μF, $S_1 = 4$ μF, $S_2 = 6$ μF.

Να δοθεί ένα διάστημα εμπιστοσύνης της διαφοράς των δύο μέσων τιμών, με δεδομένη την ισότητα των διασπορών.

Λύση:
Έχουμε δύο δείγματα από πληθυσμούς ίσων διασπορών και $n_1, n_2 < 30$. Σε αυτή την περίπτωση το διάστημα εμπιστοσύνης της διαφοράς μ_1-μ_2 δίνεται από τον παρακάτω τύπο:

$$(\overline{x}_1 - \overline{x}_2) \pm t_{(n_1+n_2-2);\frac{a}{2}} \cdot S \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

όπου S είναι η κοινή τυπική απόκλιση η οποία δίνεται από τη σχέση:

$$S = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

Από τα δεδομένα έχουμε: $\overline{x}_1 = 52$ μF, $\overline{x}_2 = 48$ μF, $s_1 = 4$ μF, $s_2 = 6$ μF, n_1=16 και n_2=13, οπότε:

$$S = \sqrt{\frac{(16-1)\cdot 4^2 + (13-1)\cdot 6^2}{16+13-2}} = 4,99 \text{ μF}.$$

Επίσης για 27 β.ε η $t_{\frac{a}{2}} = 2,052$

Το δ.ε της διαφοράς δύο μέσων τιμών είναι:

$$(52-48) \pm 2,052 \cdot 4,99 \cdot \sqrt{\frac{1}{16} + \frac{1}{13}} = 4 \pm 3,82 \text{ ή } 0,18 < \mu_1\text{-}\mu_2 < 7,82$$

β) Για δείγματα παρμένα από κανονικούς πληθυσμούς με διασπορές σ_1^2, σ_2^2 άγνωστες και άνισες.

Όταν οι δύο πληθυσμοί από τους οποίους είναι παρμένα τα δείγματα είναι κανονικοί με $\sigma_1^2 \neq \sigma_2^2$ και n_1, n_2 μικρότερα του 30 τότε το διάστημα εμπιστοσύνης της διαφοράς μ_1-μ_2 θα είναι:

$$(\overline{x} - \overline{\psi}) \pm t_{(n_1+n_2-2);\frac{a}{2}} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \qquad (8.10)$$

$$(\overline{x} - \overline{\psi}) - t_{(n_1+n_2-2);\frac{a}{2}} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, \quad (\overline{x} - \overline{\psi}) + t_{(n_1+n_2-2);\frac{a}{2}} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Αξίζει να σημειωθεί ότι το μέγεθος $\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$ εκφράζει το πιθανό σφάλμα της διαφοράς των δύο μέσων τιμών και συμβολίζεται με SE_Δ. Το πιθανό σφάλμα κάθε μέσης τιμής SE δίνεται από τη σχέση $SE=\frac{S}{\sqrt{n}}$ και ισχύει:

$$SE_\Delta=\sqrt{SE_1^2+SE_2^2}=\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}} \qquad (8.11)$$

Παράδειγμα (8.7)
Μετρήθηκε το κάλιο του ορού σε 9 υγιή άτομα και σε 4 άτομα που έπασχαν από μία νόσο. Στα υγιή άτομα βρέθηκε μέση τιμή 4 mEq/L και τυπική απόκλιση 0,8 mEq/L, ενώ στους ασθενείς βρέθηκε μέση τιμή 5 mEq/L και τυπική απόκλιση 0,8 mEq/L. Να υπολογιστεί το 95% διάστημα εμπιστοσύνης της διαφοράς των δύο μέσων τιμών.
Λύση:
Επειδή τα δείγματα προέρχονται από δύο πληθυσμούς με άγνωστες και άνισες διασπορές $\sigma_1^2 \neq \sigma_2^2$ και $n_1<30$, $n_2<30$ το διάστημα εμπιστοσύνης βρίσκεται με βάση τον τύπο:

$$(\bar{x}_1-\bar{x}_2)\pm t_{(n_1+n_2-2);\frac{a}{2}}\cdot\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

έχουμε $\bar{x}_1=5$, $\bar{x}_2=4$, $S_1=0,8$, $S_2=0,9$, $n_1=9$, $n_2=4$
και $t_{7;0,025}=2,365$ οπότε το δ.ε θα είναι:

$$(5-4)\pm 2,365\cdot\sqrt{\frac{0,8^2}{9}+\frac{0,9^2}{4}}=1\pm 1,24 \quad \text{ή} \quad -0,24<\mu_1-\mu_2<2,24$$

8.5 Διάστημα εμπιστοσύνης της διαφοράς μ_1-μ_2 συσχετισμένων πληθυσμών

Πολλές φορές οι παρατηρήσεις που προέρχονται από δύο πληθυσμούς εμφανίζονται κατά ζεύγη (χ_i, ψ_i) και δεν είναι ανεξάρτητες μεταξύ τους. Τέτοιες παρατηρήσεις είναι η μέτρηση της θερμοκρασίας του σώματος πρωί και απόγευμα, ή η αντίδραση του οργανισμού ενός ασθενούς κατά τη χορήγηση σε διαφορετικές περιόδους δύο φαρμάκων Α και Β ή κατά τη μέτρηση ενός βιοχημικού μεγέθους πριν και μετά τη χορήγηση ενός φαρμάκου σε μια ομάδα ασθενών κλπ.

Σε τέτοιες περιπτώσεις, η κατασκευή ενός διαστήματος εμπιστοσύνης της διαφοράς των μέσων τιμών μ_1-μ_2 των δύο πληθυσμών δίνεται από τον τύπο:

$$\left[(\bar{x}-\bar{\psi})=\bar{D}\right]\pm t_{(n-1);\frac{\alpha}{2}}\cdot\frac{S_D}{\sqrt{n}} \quad \text{ή} \quad \bar{D}\pm t_{(n-1);\frac{\alpha}{2}}\cdot SE \qquad (8.11) \qquad \text{όπου:}$$

$$S_D=\sqrt{\frac{\sum(D_i-\bar{D})^2}{n-1}}$$

$$\bar{D}=\bar{x}-\bar{\psi}=\frac{\sum_{i=1}^{n}D_i}{n} \qquad (8.12) \quad ,SE=\frac{S_D}{\sqrt{n}}$$

, και n ο αριθμός των ζευγών.

Παράδειγμα (8.8)

Δεκατρία άτομα παρακολούθησαν ένα πρόγραμμα βελτίωσης και ικανότητας απομνημόνευσης. Με ειδικό τεστ προσδιορίστηκε για κάθε άτομο ο βαθμός απομνημόνευσης πριν και μετά το πρόγραμμα. Τα αποτελέσματα δίνονται στον παρακάτω πίνακα:

Πριν	64	58	76	89	84	70	78	67	75	83	85	77	66
Μετά	70	65	81	93	88	76	85	74	80	88	91	83	71

Να κατασκευαστεί 95% (0,95) διάστημα εμπιστοσύνης της διαφοράς των μέσων βαθμών απομνημόνευσης πριν και μετά από το πρόγραμμα.

Λύση:

Όπως βλέπουμε από τα δεδομένα κάθε άτομο υποβάλλεται στο τεστ δυο φορές πριν και μετά το πρόγραμμα, άρα τα δύο τεστ ανά άτομο θα δώσουν συσχετισμένο ζεύγος (χ_i, ψ_i) (i= 1,2,…..,n). Συνεπώς για την κατασκευή του διαστήματος εμπιστοσύνης της διαφοράς $\mu_{μετά}$-$\mu_{πριν}$ χρησιμοποιούμε τον τύπο:

$$\bar{D}\pm t_{(n-1);\frac{\alpha}{2}}\cdot\frac{S_D}{\sqrt{n}} \quad \text{ή} \quad \bar{D}\pm t_{(n-1);\frac{\alpha}{2}}\cdot SE \;, \quad \text{όπου } \bar{D}=\bar{x}-\bar{\psi} \quad \text{και}$$

$$S_D=\sqrt{\frac{\sum(D_i-\bar{D})^2}{n-1}}=\sqrt{\frac{\sum_{i=1}^{n}D_i^2-\frac{(\sum D_i)^2}{n}}{n-1}}$$

Για να υπολογίσουμε την $\bar{D}$ και S_D κατασκευάζουμε τον παρακάτω πίνακα:

	χ (πριν)	ψ (μετά)	Διαφορά D	$D-\overline{D}$	$\left(D-\overline{D}\right)^2$	D^2
1	64	70	6	0,4	0,16	36
2	58	65	7	1,4	1,96	49
3	76	81	5	-0,6	0,36	25
4	89	93	4	-1,6	2,56	16
5	84	88	4	-1,6	2,56	16
6	70	76	6	0,4	0,16	36
7	78	85	7	1,4	1,96	49
8	67	74	7	1,4	1,96	49
9	75	80	5	-0,6	0,36	25
10	83	88	5	-0,6	0,36	25
11	85	91	6	0,4	0,16	36
12	77	83	6	0,4	0,16	36
13	66	71	5	-0,6	0,36	25
Σύνολ	972	1045	73		13,08	423

$$\overline{D}=\frac{73}{13}=5,6 \quad , \quad \overline{x}=\frac{972}{13}=74,8 \quad , \quad \overline{\psi}=\frac{1045}{13}=80,4\,, \quad S_D=\sqrt{\frac{13,08}{12}}=1,04$$

Από τους πίνακες $t_{12;0,025}=2,179$

Το διάστημα εμπιστοσύνης της μέσης διαφοράς (ή των μέσων τιμών) πριν και μετά το πρόγραμμα είναι:

$$5,6\pm 2,179\cdot 1,04=5,6\pm 2,27 \text{ ,ή} \quad 3,33<\overline{D}<7,87$$

8.6 Διάστημα εμπιστοσύνης ή όρια αξιοπιστίας για την πιθανότητα P (αναλογία ή ποσοστό) της διωνυμικής κατανομής

Όπως οι μέσες τιμές των ποσοτικών χαρακτηριστικών, έτσι και οι αναλογίες (ποσοστά ή πιθανότητες) επηρεάζονται από την τυχαία (δειγματοληπτική) διακύμανση. Συνεπώς η πιθανότητα που βρέθηκε σε κάποιο τυχαίο ή αντιπροσωπευτικό δείγμα ενδέχεται να είναι, εξαιτίας της τυχαίας διακύμανσης, μεγαλύτερη ή μικρότερη από την πραγματική πιθανότητα στο γενικότερο σύνολο από το οποίο προήλθε το δείγμα. Ο προσδιορισμός των ορίων αξιοπιστίας (διάστημα εμπιστοσύνης) μιας πιθανότητας (αναλογίας) γίνεται ανάλογα με τον αριθμό των παρατηρήσεων και την τιμή της πιθανότητας P όπως φαίνεται παρακάτω.

α) Ο αριθμός των παρατηρήσεων είναι αρκετά μεγάλος (n>40) και η τιμή της πιθανότητας που υπολογίστηκε είναι μεταξύ 0,10 και 0,90 (0,10< P< 0,90).

Στις περιπτώσεις αυτές το διάστημα εμπιστοσύνης της αναλογίας (πιθανότητας) για ένα επίπεδο σημαντικότητας $(1-\alpha)\cdot 100$ υπολογίζεται με την σχέση:

$$P \pm Z_{\frac{a}{2}} \cdot SE_P = P \pm Z_{\frac{a}{2}} \cdot \sqrt{\frac{p \cdot q}{n}} \qquad (8.13)$$

όπου SE_p= πιθανό σφάλμα της πιθανότητας = $\sqrt{\frac{P \cdot q}{n}} = \sqrt{\frac{P(1-P)}{q}}$ και

$q = 1 - P$ ή

$$P - z_{\frac{a}{2}} \cdot \sqrt{\frac{P(1-P)}{n}} < P < P + z_{\frac{a}{2}} \cdot \sqrt{\frac{P(1-P)}{n}} \qquad (8.13)$$

όπου P είναι η πιθανότητα που υπολογίζεται από τα πειραματικά δεδομένα.

Παράδειγμα (8.9)

Μετά από μια διαφημιστική καμπάνια ρωτήθηκαν 150 καπνιστές και βρέθηκαν 20 να καπνίζουν τα τσιγάρα μάρκας Α. Να βρεθεί 95% διάστημα εμπιστοσύνης για το ποσοστό των καπνιστών που καπνίζουν τα τσιγάρα μάρκας Α μετά τη διαφήμιση.

Λύση:

Το ποσοστό των καπνιστών που καπνίζουν τσιγάρα μάρκας Α είναι

$$P = \frac{20}{150} = 0{,}133$$

Το μέγεθος του δείγματος είναι μεγάλο n> 40 και ισχύει:

0,10< P< 0,90 συνεπώς το διάστημα εμπιστοσύνης του ποσοστού αυτού δίνεται από τον τύπο:

$$P \pm Z_{\frac{a}{2}} \cdot \sqrt{\frac{p \cdot q}{n}} = P \pm Z_{\frac{a}{2}} \sqrt{\frac{P(1-P)}{n}}$$

έχουμε P=0,133 , q = 1-P=1-0,133=0,867 και $Z_{0,025} = 1{,}96$ οπότε το 95% δ.ε είναι:

$$0{,}133 \pm 1{,}96 \cdot \sqrt{\frac{0{,}133 \cdot 0{,}867}{150}} = 0{,}133 \pm 0{,}054, \text{ ή } \ 0{,}079 < P < 0{,}187$$

Παράδειγμα (8.10)
Έγινε η εξής ερώτηση σε 800 φοιτητές 'Αγοράσατε τουλάχιστον ένα ζευγάρι παπούτσια στο χρονικό διάστημα μεταξύ των μηνών Σεπτεμβρίου και Ιουνίου;'. Ο αριθμός των θετικών απαντήσεων ήταν 100. Να βρεθεί 95% διάστημα εμπιστοσύνης για το ποσοστό των φοιτητών που αγόρασαν τουλάχιστον ένα ζευγάρι παπούτσια, το παραπάνω χρονικό διάστημα.
Λύση:
Το ποσοστό των φοιτητών που αγόρασαν τουλάχιστον ένα ζευγάρι παπούτσια την περίοδο Σεπτεμβρίου-Ιουνίου είναι $P=\frac{100}{800}=0,125$

$$q=1-P=1-0,125=0,875$$

$$SE_P=\sqrt{\frac{P\cdot q}{n}}=\sqrt{\frac{P(1-P)}{q}}=\sqrt{\frac{0,125\cdot 0,875}{800}}=0,012$$

Λόγω του μεγάλου αριθμού παρατηρήσεων n> 40 και 0,10 < P < 0,90 το δ.ε είναι:

$$P\pm Z_{\frac{a}{2}}\sqrt{\frac{P(1-P)}{n}}=0,125\pm 1,96\cdot 0,012=0,125\pm 0,022$$

Παράδειγμα (8.11)
Σε μια ομάδα 200 ανδρών ηλικίας 70 ετών επιβιώνουν για μια πενταετία οι 170. Ποια τα 95% όρια αξιοπιστίας του αριθμού 170;
Λύση:
Το πρώτο που κάνουμε σε τέτοια περίπτωση είναι να μετατρέψουμε τον αριθμό 170 σε ποσοστό:

$$P=\frac{170}{200}=0,85$$

Το δεύτερο στάδιο είναι να βρούμε τα όρια αξιοπιστίας (διάστημα εμπιστοσύνης) του ποσοστού 0,85.

$$\text{δ.ε} = P\pm Z_{\frac{a}{2}}\sqrt{\frac{P(1-P)}{n}}=0,85\pm 1,96\cdot\sqrt{\frac{0,85\cdot 0,15}{200}}=0,85\pm 0,05$$

Το τρίτο στάδιο είναι να μετατρέψουμε τα ποσοστά σε αριθμούς οπότε τα όρια αξιοπιστίας του αριθμού 170 είναι:

$$0,85\cdot 200\pm 0,05\cdot 200=170\pm 10$$

β) Ο αριθμός των παρατηρήσεων είναι αρκετά μεγάλος n>40 και η πιθανότητα (αναλογία ή ποσοστό) που υπολογίστηκε είναι πολύ μικρή (P< 0,10) ή πολύ μεγάλη (P> 0,90).
Στην περίπτωση που ισχύει $P \le 0{,}10$ ή $P \ge 0{,}90$ και n> 40 το διάστημα εμπιστοσύνης μπορεί να υπολογιστεί μόνο αν το γινόμενο της μικρότερης αναλογίας (P ή q) επί τον αριθμό των παρατηρήσεων, δηλαδή $n \cdot P$ ή $n \cdot q$ είναι μεγαλύτερο του 5. Αν συμβαίνει αυτό τότε το δ.ε υπολογίζεται με βάση τον τύπο:

$$P \pm Z_{\frac{a}{2}} \cdot SE_P = P \pm Z_{\frac{a}{2}} \cdot \sqrt{\frac{p \cdot q}{n}} \qquad (8.13)$$

Αν η παραπάνω προϋπόθεση δεν ισχύει τότε το διάστημα εμπιστοσύνης απέχει ανισόμετρα από την P και ο υπολογισμός του προϋποθέτει εκτέλεση πολύπλοκων υπολογισμών ή χρήση έτοιμων πινάκων.

Παράδειγμα (8.12)
Αν η πιθανότητα να πεθάνει ένα άτομο από μια συγκεκριμένη νόσο είναι 0,009 και προσβληθούν από αυτή 860 άτομα, ποιος ο αριθμός των θανάτων που προβλέπεται και ποιο είναι το 95% διάστημα εμπιστοσύνης του αριθμού αυτού;
Λύση:
Από τα δεδομένα της άσκησης έχουμε $P = 0{,}009 < 0{,}10$ άρα $q = 1 - 0{,}009 = 0{,}991$.
Επειδή $P < 0{,}10$ για να βρούμε το δ.ε πρέπει να εξετάσουμε αν $n \cdot P > 5$,οπότε:
$860 \cdot 0{,}009 = 7{,}75 > 5$, επομένως το διάστημα εμπιστοσύνης υπολογίζεται από τον τύπο:

$$P \pm Z_{\frac{a}{2}} \cdot SE_P = P \pm Z_{\frac{a}{2}} \cdot \sqrt{\frac{p \cdot q}{n}} = 0{,}009 \pm 1{,}96 \cdot \sqrt{\frac{0{,}009 \cdot 0{,}991}{860}} = 0{,}009 \pm 0{,}0$$

Ο αριθμός των θανάτων που προβλέπεται είναι $n \cdot P = 7{,}45$.
Τα όρια αξιοπιστίας (δ.ε) του αριθμού 7,45 είναι:

$$7{,}45 \pm 0{,}006 \cdot 860 = 7{,}45 \pm 5{,}16$$

ή τα όρια αξιοπιστίας του αριθμού 7,45 είναι:
7,45 + 5,16= 12,61και 7,45 – 5,16= 2,29

γ) Ο αριθμός των παρατηρήσεων είναι μικρός.
Στις περιπτώσεις που ο αριθμός των παρατηρήσεων n< 40 και P< 0,30 ή P> 0,70, το πιθανό σφάλμα που υπολογίζεται από τον τύπο $\sqrt{\frac{P\cdot q}{n}}$ είναι σχετικά μεγάλο. Επομένως το διάστημα εμπιστοσύνης που υπολογίζεται με τη συμμετρική εφαρμογή του πιθανού σφάλματος πρέπει να αποφεύγεται. Η εκτίμηση των ορίων αξιοπιστίας σε επίπεδο σημαντικότητας 95% ή 99% γίνεται με βάση θεωρητικών πινάκων όπου δίνονται τα ακριβή όρια αξιοπιστίας για όλες τις αναλογίες που είναι δυνατό να προκύψουν σε δείγματα με αριθμό μετρήσεων από 2 μέχρι 40.

8.7 Διάστημα εμπιστοσύνης της διαφοράς δύο αναλογιών P_1 και P_2.

Το διάστημα εμπιστοσύνης της διαφοράς P_1-P_2 δύο ανεξάρτητων διωνυμικών πληθυσμών με $n_1 \geq 30$ και $n_2 \geq 30$ δίνεται από τον παρακάτω τύπο:

δ.ε: $P_1 - P_2 \pm Z_{\frac{a}{2}} \cdot SE_{\Delta}$ (8.14)

όπου $P_1 = \frac{x_1}{n_1}$, $P_2 = \frac{x_2}{n_2}$, x_1 = αριθμός επιτυχιών στον πληθυσμό 1 σε n_1 ανεξάρτητες προσπάθειες, x_2 = αριθμός επιτυχιών στον πληθυσμό 2 σε n_2 ανεξάρτητες προσπάθειες.

$SE_{\Delta} = \sqrt{SE_1^2 + SE_2^2}$ όπου:

$SE_1 = \sqrt{\frac{P_1(1-P_1)}{n_1}}$ και $SE_2 = \sqrt{\frac{P_2(1-P_2)}{n_2}}$.

Άρα το διάστημα εμπιστοσύνης της διαφοράς $P_1 - P_2$ μπορεί να γραφεί ως:

$$P_1 - P_2 \pm Z_{\frac{a}{2}} \cdot \sqrt{\frac{P_1\cdot(1-P_1)}{n_1} + \frac{P_2\cdot(1-P_2)}{n_2}} \qquad (8.15)$$

Παράδειγμα (8.13)
Από τις 450 μποτίλιες που παίρνουμε στην τύχη από μηχανή που φτιάχνει μποτίλιες, οι 18 είναι ελαττωματικές, ενώ από τις 350 μιας άλλης μηχανής, οι 28 βρέθηκαν ελαττωματικές. Να βρεθεί το 95% διάστημα εμπιστοσύνης

για τη διαφορά των αναλογιών των ελαττωματικών μποτιλιών, ανάμεσα στις δύο μηχανές.

Λύση:

Από τα δεδομένα έχουμε δύο πληθυσμούς ανεξάρτητους μεταξύ τους με $n_1 = 350$, $x_1 = 28$, $n_2 = 450$, $x_2 = 18$ οπότε , $P_1 = \frac{28}{350} = 0,08$ και $P_2 = \frac{18}{450} = 0,04$

Επειδή n_1 και $n_2 > 30$ το διάστημα εμπιστοσύνης της διαφοράς $P_1 - P_2$ δίνεται από τον τύπο:

$$P_1 - P_2 \pm Z_{\frac{a}{2}} \cdot \sqrt{\frac{P_1 \cdot (1-P_1)}{n_1} + \frac{P_2 \cdot (1-P_2)}{n_2}} \Rightarrow$$

$$\delta.\varepsilon = 0,08 - 0,04 \pm 1,96 \cdot \sqrt{\frac{0,08 \cdot 0,92}{350} + \frac{0,04 \cdot 0,96}{450}} \quad = 0,04 \pm 0,034$$

ή $0,006 < P_1 - P_2 < 0,074$

Παράδειγμα (8.14)

Σε μια σκοποβολή ένας σκοπευτής έκανε συνολικά 250 βολές από τις οποίες οι 90 ήταν πετυχημένες. Ένας άλλος, από τις 331 βολές που έκανε πέτυχε τις 96 βολές. Βρείτε ένα 99% διάστημα εμπιστοσύνης για τη διαφορά των αναλογιών των επιτυχημένων βολών.

Λύση:

Για να βρούμε το διάστημα εμπιστοσύνης της διαφοράς των δύο αναλογιών πρέπει πρώτα να βρούμε τις αναλογίες P_1 και P_2.

$$P_1 = \frac{x_1}{n_1} = \frac{90}{250} = 0,36 \,,\; P_2 = \frac{x_2}{n_2} = \frac{96}{331} = 0,29$$

Για συντελεστή σημαντικότητας 99% η τιμή $z_{\frac{a}{2}} = 2,576$

Το δ.ε της διαφοράς $P_1 - P_2$ των επιτυχημένων βολών δίνεται ως:

$$(P_1-P_2)\pm Z_{\frac{a}{2}}\cdot\sqrt{\frac{P_1\cdot(1-P_1)}{n_1}+\frac{P_2\cdot(1-P_2)}{n_2}}\Rightarrow$$

$$\delta.\varepsilon=0,36-0,29\pm 2,576\cdot\sqrt{\frac{0,36\cdot(1-0,36)}{250}+\frac{0,29\cdot(1-0,29)}{331}}=0,07\pm 0,10 \text{ ή}$$

$$-0,03<P_1-P_2<0,17$$

8.8 Διάστημα εμπιστοσύνης για τη διασπορά ενός πληθυσμού σ²

Άσχετα από το μέγεθος του δείγματος (μικρό ή μεγάλο) το διάστημα εμπιστοσύνης της δειγματικής διασποράς ενός πληθυσμού δίνεται ως:

$$\frac{(n-1)\cdot S^2}{X^2_{(n-1);\frac{a}{2}}}<\sigma^2<\frac{(n-1)\cdot S^2}{X^2_{(n-1);\left(1-\frac{a}{2}\right)}} \qquad (8.16)$$

όπου X^2 είναι μια X^2-κατανομή με n-1 βαθμούς ελευθερίας. Οι τιμές $X^2_{(n-1);\frac{a}{2}}$ και $X^2_{(n-1);\left(1-\frac{a}{2}\right)}$ βρίσκονται από τον πίνακα (Π…) στο τέλος του βιβλίου.

Το διάστημα εμπιστοσύνης για την τυπική απόκλιση του πληθυσμού είναι:

$$\sqrt{\frac{(n-1)\cdot S^2}{X^2_{n-1;\frac{\alpha}{2}}}}<\sigma<\sqrt{\frac{(n-1)\cdot S^2}{X^2_{n-1;(1-\frac{\alpha}{2})}}} \qquad (8.17)$$

Παράδειγμα (8.15)

Μετρήσεις του δείκτη διάθλασης 25 τεμαχίων διαφανούς ύαλου έδωσαν τυπική απόκλιση 0,012. Να κατασκευασθεί ένα διάστημα εμπιστοσύνης της διασποράς του δείκτη διάθλασης με βαθμό (συντελεστή) εμπιστοσύνης 0,90.

Λύση:

Το διάστημα εμπιστοσύνης της διασποράς του δείκτη διάθλασης σ^2 είναι:

$$\frac{(n-1)\cdot S^2}{X^2_{(n-1);\frac{\alpha}{2}}}<\sigma^2<\frac{(n-1)\cdot S^2}{X^2_{(n-1);(1-\frac{\alpha}{2})}}$$

Από τα δεδομένα έχουμε : S= 0,012 άρα $S^2=1,44\cdot 10^{-4}$, a = 0,10 , $\frac{a}{2}$ = 0,05 , n = 25 οπότε οι βαθμοί ελευθερίας n-1 = 25-1 = 24 .

Από τον πίνακα (3) στο παράρτημα βρίσκουμε ότι $X^2_{24;0,05}=36,4151$ και $X^2_{24;0,95}=13,8484$, συνεπώς το 0,90 Δ.Ε είναι

$$\frac{24\cdot 1,44\cdot 10^{-4}}{36,4151}<\sigma^2<\frac{24\cdot 1,44\cdot 10^{-4}}{13,8484}\Rightarrow$$

$$9,49\cdot 10^{-5}<\sigma^2<2,50\cdot 10^{-4}$$

Αν θέλαμε το δ.ε της τυπικής απόκλισης τότε θα είχαμε:

$$\sqrt{9,49\cdot 10^{-5}}<\sigma<\sqrt{2,5\cdot 10^{-5}}\Rightarrow \quad 9,74\cdot 10^{-3}<\sigma<0,0158$$

Παράδειγμα (8.16)

Ένας ζωολόγος θέλει να εκτιμήσει την περιεκτικότητα ζάχαρης στο αίμα ορισμένου είδους ζώων, όταν τους γίνει ένεση αδρεναλίνης. Ένα δείγμα 55 ζώων, ίδιας γενιάς, που του γίνεται τέτοια ένεση, δίνει μέση τιμή και τυπική απόκλιση αντίστοιχα 126,9 και 10,5 mg/100 mL αίματος. Να ορισθεί ένα 90% διάστημα εμπιστοσύνης για τη μέση τιμή και τη διασπορά του πληθυσμού.

Λύση:

Από τα δεδομένα της άσκησης έχουμε ότι n = 55, $\overline{x}=126,9$ mg/100mL, S= 10,5 mg/100mL, σ^2 = άγνωστη, n > 30, οπότε το δ.ε της μέσης τιμής δίνεται από τη σχέση: $\overline{x}\pm\frac{S}{\sqrt{n}}\cdot z_{\frac{a}{2}}$ και $Z_{0,05}=1,645$ οπότε

$$\overline{x}\pm\frac{S}{\sqrt{n}}\cdot z_{\frac{a}{2}}=126,9\pm\frac{10,5}{\sqrt{55}}\cdot 1,645=126,9\pm 2,3$$

Το δ.ε της διασποράς είναι:

$$\frac{(n-1)\cdot S^2}{X^2_{(n-1);\frac{\alpha}{2}}}<\sigma^2<\frac{(n-1)\cdot S^2}{X^2_{(n-1);(1-\frac{\alpha}{2})}}$$

Για 90% δ.ε με βαθμούς ελευθερίας 54 έχουμε :

$X^2_{54;0,05}=73,29$ και $X^2_{54;0,95}=38,98$ συνεπώς το δ.ε είναι:

$$\sqrt{\frac{54\cdot(10,5)^2}{73,29}}<\sigma<\sqrt{\frac{54\cdot(10,5)^2}{38,98}}$$

$$9,01<\sigma<12,36$$

8.9 Διάστημα εμπιστοσύνης για το λόγο των διασπορών δύο κανονικών πληθυσμών

Στην περίπτωση δύο κανονικών πληθυσμών με n_1, σ_1^2 για τον πρώτο πληθυσμό και n_2, σ_2^2 για το δεύτερο πληθυσμό, το διάστημα εμπιστοσύνης του λόγου των δειγματικών διασπορών είναι:

$$\frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{n_1-1, n_2-1, \frac{a}{2}}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} \cdot F_{n_2-1, n_1-1; \frac{a}{2}} \qquad (8.18)$$

$F_{n-1;n-1;\frac{\alpha}{2}}$ είναι η τυχαία μεταβλητή F της F- κατανομής με βαθμούς ελευθερίας $n_1 - 1, n_2 - 1$ στο επίπεδο σημαντικότητας $\frac{a}{2}$, της οποίας η τιμή βρίσκεται από τον πίνακα (4) στο τέλος του βιβλίου.

Παράδειγμα (8.17)

Δύο ανεξάρτητα τυχαία δείγματα $\chi_1, \chi_2, \ldots \chi_8$ και $\psi_1, \psi_2 \ldots \psi_5$ με κανονικές κατανομές $N(\mu_1, \sigma_1^2)$ και $N(\mu_2, \sigma_2^2)$ αντίστοιχα έδωσαν $\sum_{i=1}^{8} x_i = 96$ και $\sum_{i=1}^{8} x_i^2 = 1194$ και ομοίως $\sum_{j=1}^{5} \psi_j = 50$ και $\sum_{j=1}^{5} \psi_j^2 = 504$. Να κατασκευαστεί 90% δ.ε για το πηλίκο $\frac{\sigma_1^2}{\sigma_2^2}$.

Λύση:

Για να βρούμε το διάστημα εμπιστοσύνης για το πηλίκο των διασπορών πρέπει πρώτα να βρούμε τις διασπορές σ_1^2 και σ_2^2 των δύο πληθυσμών αντίστοιχα. Η δειγματική διασπορά s^2 δίνεται από τη σχέση:

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

Για τον πληθυσμό 1 έχουμε $n_1 = 8$, $\sum_{i=1}^{8} x_i^2 = 1194$ και $\sum_{i=1}^{8} x_i = 96$. Οπότε:

$$S_1^2 = \frac{1}{7}(1194 - \frac{(96)^2}{8}) = 6$$

Για τον πληθυσμό 2 έχουμε n_2= 5, $\sum_{j=1}^{5}\psi_j^2 = 504$ και $\sum_{j=1}^{5}\psi_j = 50$ άρα:

$$S_2^2 = \frac{1}{4}(504 - \frac{(50)^2}{5} = 1$$

Το διάστημα εμπιστοσύνης για το λόγο $\frac{\sigma_1^2}{\sigma_2^2}$ είναι:

$$\frac{S_1^2}{S_2^2}\cdot\frac{1}{F_{n_1-1;n_2-2;\frac{\alpha}{2}}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\cdot F_{n_2-1;n_1-1;\frac{\alpha}{2}}$$

Από τον πίνακα (….) έχουμε $F_{7;5;0,05} = 4,88$ και $F_{5;7;0,05} = 3,97$

Οπότε το δ.ε είναι:

$$\frac{6}{1}\cdot\frac{1}{4,88} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{6}{1}\cdot 3,97 \Rightarrow$$

$$1,23 < \frac{\sigma_1^2}{\sigma_2^2} < 23,82$$

Παράδειγμα (8.18)

Δύο μηχανές κατασκευάζουν πυκνωτές. Από τυχαία δείγματα μεγέθους n_1=12 και n_2=10 από τις δύο μηχανές αντίστοιχα προέκυψαν: $\overline{x}$ = 46 μF, $\overline{\chi_2}$ = 42 μF, $S_1 = 5$ μF και $S_2 = 7$ μF. Να δοθεί ένα 98% διάστημα εμπιστοσύνης του λόγου $\frac{\sigma_1^2}{\sigma_2^2}$.

Λύση:

Από τα δεδομένα έχουμε $S_1^2 = 25, n_1 = 12, S_2^2 = 49, n_2 = 10$.Το διάστημα εμπιστοσύνης του λόγου $\frac{\sigma_1^2}{\sigma_2^2}$ είναι: $\frac{S_1^2}{S_2^2}\cdot\frac{1}{F_{n_1-1;n_2-2;\frac{\alpha}{2}}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\cdot F_{n_2-1;n_1-1;\frac{\alpha}{2}}$

Η τιμή της $F_{11;9;0,01} = 3,07$ και $F_{9;11;0,01} = 2,90 \Rightarrow$

$$\frac{25}{49}\cdot\frac{1}{3,07} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{25}{49}\cdot 2,90 \Rightarrow \quad 0,166 < \frac{\sigma_1^2}{\sigma_2^2} < 1,48$$

ΑΣΚΗΣΕΙΣ

8.1 Μια μηχανή κατασκευάζει μπίλιες για ρουλεμάν. Οι 200 μπίλιες ενός τυχαίου δείγματος από την παραγωγή μιας εβδομάδας έχουν διαμέτρους με μέση τιμή 0.824 cm και τυπική απόκλιση0.042 cm. Υπολογίστε (α) 95% και (β) 99% διαστήματα εμπιστοσύνης για τη μέση τιμή της διαμέτρου του πληθυσμού.

8.2 Σε ένα τυχαίο δείγμα από 50 άνδρες το μέσο βάρος ήταν 75 κιλά με τυπική απόκλιση 10. (α) Υπολογίστε 95% διάστημα εμπιστοσύνης για το μέσο βάρος του πληθυσμού από τον οποίο προέρχεται το δείγμα. (β) Με τι εμπιστοσύνη μπορεί κανείς να πει ότι τι μέσο βάρος του πληθυσμού είναι μεταξύ 75±1.

8.3 Δέκα μετρήσεις της διαμέτρου μιας σφαίρας έδωσαν μέση τιμή $\overline{X}=4.38$ cm και τυπική απόκλιση $S = 0.06$ cm. Υπολογίστε (α) 95% και (β) 99% διαστήματα εμπιστοσύνης για την πραγματική διάμετρο της σφαίρας. (Είναι γνωστό ότι τυχαία σφάλματα σε μετρήσεις ακολουθούν κανονική κατανομή και άρα μπορούμε εδώ να θεωρήσουμε ότι οι δέκα μετρήσεις αποτελούν τυχαίο αποτελούν τυχαίο δείγμα από πληθυσμό με κανονική κατανομή).

8.4 Ένα δείγμα 150 λαμπτήρων τύπου Α έδωσε μέση τιμή ζωής 1400 ώρες για κάθε λαμπτήρα με τυπική απόκλιση 120 ώρες. Ένα άλλο δείγμα 200 λαμπτήρων τύπου Β έδωσε μέση ζωή 1200 ώρες για κάθε λαμπτήρα με τυπική απόκλιση 80 ώρες. Υπολογίστε (α) 95% και (β) 99% διαστήματα εμπιστοσύνης για τη διαφορά της μέσης ζωής των λαμπτήρων τύπου Α μείον τη μέση ζωή των λαμπτήρων τύπου Β.

8.5 Η τυπική απόκλιση της ζωής ενός τύπου λυχνιών είναι 100 ώρες. Πόσο μεγάλο πρέπει να είναι ένα δείγμα ώστε να είμαστε (α) 95%, (β) 90% και (γ) 99.73% βέβαιοι ότι το σφάλμα στην εκτίμηση της μέσης ζωής δεν υπερβαίνει τις 20 ώρες.

8.6 Ένας άνθρωπος αντέδρασε σε 5 εξωτερικούς ερεθισμούς σε χρόνο 0.28 0.30, 0.27 , 0.33 , 0,31 sec αντίστοιχα . Υπολογίστε (α) 95% και (β) 99% διαστήματα εμπιστοσύνης για την πραγματική μέση τιμή του χρόνου αντίδρασης του ανθρώπου αυτού (υποθέτοντας ότι ο χρόνος αντίδρασης ακολουθεί κανονική κατανομή).

8.7 Τα παρακάτω δεδομένα δίνουν το ποσοστό επιμήκυνσης (κάτω από ορισμένο βάρος) για δύο δείγματα από κομμάτια ενός συγκεκριμένου τύπου νήματος όπου το πρώτο δείγμα ελήφθη πριν πλυθεί το νήμα και το δεύτερο αφού το νήμα πλύθηκε 6 φορές.

Χωρίς πλύση:	12.3	13.7	10.4	11.4	14.9	12.6	
Μετά 6 πλύσεις:	15.7	10.3	12.6	14.5	12.6	13.8	11.9

Να δοθούν (α) 95% και (β) 99% διαστήματα εμπιστοσύνης για την διαφορά στο ποσοστό επιμήκυνσης του συγκεκριμένου νήματος πριν και μετά την πλύση. (Μπορείτε να υποθέσετε ότι τα ποσοστά επιμήκυνσης του συγκεκριμένου νήματος πριν και μετά την πλύση, ακολουθούν κανονικές κατανομές με ίσες διασπορές.)

8.8 Σε ένα άλλο πείραμα με τον ίδιο τύπο νήματος πάρθηκαν 6 κομμάτια τυχαίου μήκους από το νήμα και κάθε κομμάτι κόπηκε στα δύο. Από τα δύο κομμάτια, το ένα υποβλήθηκε σε έξι πλύσεις και το άλλο δεν πλύθηκε καθόλου. Τα ποσοστά επιμήκυνσης ήταν:

Χωρίς πλύση:	13.9	12.5	11.0	11.8	10.8	14.6
Μετά 6 πλύσεις:	14.7	12.1	13.2	13.6	11.5	15.4

Να δοθούν (α) 95% και (β) 90% διαστήματα εμπιστοσύνης για την διαφορά στο ποσοστό επιμήκυνσης του συγκεκριμένου νήματος πριν και μετά την πλύση. (Μπορείτε να υποθέσετε ότι οι διαφορές στα ποσοστά επιμήκυνσης μεταξύ των κομματιών που δεν πλύθηκαν και των αντίστοιχων που πλύθηκαν, ακολουθούν κανονική κατανομή.

8.9 Τα παρακάτω δεδομένα προέρχονται από πληθυσμό με κανονική κατανομή, με μέση τιμή μ και διασπορά σ^2, άγνωστα και τα δύο: 6.82, 6.07, 6.87, 5.92. Για αυτά τα δεδομένα, $\Sigma_i x_i$=29.42 και $\Sigma_i x_i^2$=179.5882.
(α) Δώστε 95% διάστημα εμπιστοσύνης για τον μέσο μ του πληθυσμού και συγκεκριμένα δείξτε ότι αυτό έχει μήκος (περίπου) 3.165.
(β) Έστω ότι μαθαίνουμε ότι η διασπορά σ^2 του πληθυσμού είναι 1.Δώστε 95% διάστημα εμπιστοσύνης για τον μέσο μ με τα νέα δεδομένα και συγκεκριμένα δείξτε ότι αυτό το νέο διάστημα έχει μήκος (περίπου) 1.76.

8.10 Η μέση τιμή και η τυπική απόκλιση των μέγιστων φορτίων που άντεξαν 30 συρματόσχοινα ενός συγκεκριμένου τύπου ήταν 11.09 και 0.73 τόνοι αντίστοιχα. Είναι επίσης γνωστό ότι η αντοχή του συγκεκριμένου τύπου συρματόσχοινου ακολουθεί κανονική κατανομή. Δώστε 95% και 99%διαστήματα εμπιστοσύνης για την μέση αντοχή του συγκεκριμένου τύπου συρματόσχοινου, χρησιμοποιώντας
(α) τον τύπο που στηρίζεται στην τυπική κανονική κατανομή
(β) τον τύπο που στηρίζεται στην κατανομή t του Student
και συγκρίνετε τα αντίστοιχα διαστήματα των (α) και (β).
(Εδώ το μέγεθος του δείγματος είναι στην οριακή περίπτωση n=30.)

8.11 Σε μία νέα μέθοδο καθαρισμού ενός μεταλλεύματος, βρέθηκε ότι ο χρόνος, σε min που χρειάζεται για τον καθαρισμό ενός κιλού μεταλλεύματος σε 5 διαφορετικά δείγματα ήταν: 30, 36, 35, 28, 41. Μια άλλη μέθοδος καθαρισμού χρησιμοποιήθηκε σε 5 άλλα δείγματα του ενός κιλού και χρειάστηκε χρόνο 22, 30, 27, 25, 31. Με την υπόθεση ότι τα δύο δείγματα προέρχονται από κανονικούς πληθυσμούς με ίσες διασπορές, δώστε 99% διάστημα εμπιστοσύνης για την διαφορά στον χρόνο καθαρισμού ενός κιλού του συγκεκριμένου μεταλλεύματος μεταξύ των δύο μεθόδων.

8.12 Δύο διαφορετικά συστήματα πάχυνσης αγελάδων εφαρμόστηκαν σε 5 ζευγάρια δίδυμων αγελάδων. Το σύστημα Α εφαρμόστηκε σε μία από τις δύο δίδυμες αγελάδες και το σύστημα Β στην άλλη. Το βάρος που κέρδισαν σε κιλά ήταν:

Σύστημα Α: 40 39 39 42 46
Σύστημα Β: 37 36 38 43 42

(α) Δώστε 95% διάστημα εμπιστοσύνης για την διαφορά στην αύξηση βάρους μεταξύ των δύο μεθόδων.
(β) Δώστε 95% διάστημα εμπιστοσύνης για την αύξηση βάρους με την μέθοδο Α.
(Μπορείτε να υποθέσετε ότι οι αυξήσεις βάρους με κάθε σύστημα, καθώς επίσης και διάφορές στις αυξήσεις βάρους μεταξύ των δύο συστημάτων Α και Β, ακολουθούν κανονικές κατανομές.)

ΚΕΦΑΛΑΙΟ 9

Δοκιμασία ή έλεγχος Στατιστικών Υποθέσεων

9.1 Εισαγωγή

Ο έλεγχος στατιστικών υποθέσεων είναι η διαδικασία με την οποία, με βάση τα δεδομένα ενός δείγματος ελέγχουμε την **ορθότητα** (αποφαινόμαστε κατά πόσο είναι εσφαλμένη ή όχι) μιας υπόθεσης που έχουμε διατυπώσει για κάποιο χαρακτηριστικό του πληθυσμού.

Παραμετρικοί έλεγχοι (parametric tests): ονομάζονται οι έλεγχοι των υποθέσεων για μία ή περισσότερες παραμέτρους πληθυσμού οι οποίοι γίνονται κάτω από την υπόθεση ότι η κατανομή του πληθυσμού είναι γνωστή, συνήθως κανονική ή διωνυμική. Ορισμένοι από τους ελέγχους αυτούς μπορούν να εφαρμοστούν και στην περίπτωση που αγνοούμε την κατανομή πληθυσμού, αρκεί να έχουμε μεγάλο μέγεθος δείγματος.

Για να κάνουμε ένα test (έλεγχο) σημαντικότητας πρέπει να ακολουθήσουμε τα εξής βήματα:

1. Ορίζουμε τη μηδενική υπόθεση, $\mathbf{H_0}$.

2. Ορίζουμε την εναλλακτική υπόθεση, $\mathbf{H_1}$.

3. Ορίζουμε το κριτήριο απόφασης για την αποδοχή ή απόρριψη της μηδενικής υπόθεσης. Για το σκοπό αυτό επιλέγουμε το στατιστικό έλεγχο που θα κάνουμε και το επίπεδο σημαντικότητας του ελέγχου το οποίο ορίζεται συνήθως στις τιμές 0,10 ή 0,05 ή 0,01. Η τιμή, ή οι τιμές, του στατιστικού ελέγχου οι οποίες αντιστοιχούν στο επίπεδο σημαντικότητας α ονομάζονται **κριτικές τιμές (critical values)** και προσδιορίζουν την περιοχή αποδοχής και την περιοχή απόρριψης της μηδενικής υπόθεσης.

4. Βγάζουμε τα συμπεράσματα και παίρνουμε την απόφαση ως εξής:
Αν η τιμή του στατιστικού ελέγχου βρίσκεται στην περιοχή αποδοχής της H_0 την δεχόμαστε, διαφορετικά την απορρίπτουμε.

Μηδενική υπόθεση λέγεται η υπόθεση που κάνουμε για μια συγκεκριμένη τιμή της παραμέτρου και η οποία είναι η σοβαρότερη υπόθεση στο test, και συμβολίζεται με H_0.

Εναλλακτική υπόθεση λέγεται η άλλη υπόθεση, ως προς την οποία ελέγχεται η H_0, συμβολίζεται με H_1.

9.2 Σφάλμα τύπου 1 και σφάλμα τύπου 2

Κατά την εφαρμογή ενός στατιστικού ελέγχου επιδιώκεται ουσιαστικά να ελεγχθεί αν ισχύει η μηδενική υπόθεση (δηλαδή δεν υπάρχει οποιαδήποτε διαφορά) ή αν αυτή δεν ισχύει (οπότε υπάρχει πράγματι διαφορά)

Τα αποτελέσματα μιας στατιστικής δοκιμασίας ενδέχεται να οδηγήσουν σε δύο είδη σφαλμάτων:

α).**Σφάλμα τύπου 1** (ή σφάλμα πρώτου βαθμού).

Είναι το σφάλμα που κάνουμε όταν απορρίπτουμε τη μηδενική υπόθεση H_0 ενώ αυτή είναι σωστή. Δηλαδή βγάζουμε συμπέρασμα ότι σε ένα επίπεδο σημαντικότητας η μηδενική υπόθεση δεν ισχύει ενώ στην πραγματικότητα αυτή ισχύει. Η πιθανότητα του σφάλματος αυτού συμβολίζεται με α. Το α λέγεται στάθμη σημαντικότητας του ελέγχου (test).

β) Σφάλμα τύπου 2 (ή σφάλμα δεύτερου βαθμού)

Είναι το σφάλμα που κάνουμε όταν δεν απορρίπτουμε (δηλαδή δεχόμαστε) την υπόθεση H_0, ενώ αυτή δεν ισχύει, δηλαδή το σφάλμα τύπου 2 αφορά στη διαπίστωση ότι δεν υπάρχει διαφορά, ενώ στην πραγματικότητα αυτή υπάρχει.

Η **ισχύς** (power) μιας στατιστικής διαδικασίας ορίζεται με βάση την πιθανότητα σφάλματος τύπου 2 σύμφωνα με τον τύπο:

Ισχύς = 1- (P σφάλματος τύπου 2)

Γενικά όσο μικρότερο είναι ένα λάθος τύπου 1, τόσο συνεπέστερο είναι το test που κάνουμε.

9.3 Έλεγχος υποθέσεων για τη μέση τιμή του πληθυσμού μ

9.3 1 Έλεγχος για τη μέση τιμή μ πληθυσμού με γνωστή διασπορά σ^2

Ο έλεγχος για τη μέση τιμή πληθυσμού με γνωστή διασπορά ανεξάρτητα από τον αριθμό των παρατηρήσεων του δείγματος γίνεται με το κριτήριο απόφασης:

$$Z = \frac{\bar{x} - \mu_0}{SE} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{(\bar{x} - \mu_0) \cdot \sqrt{n}}{\sigma} \qquad (9.1)$$

Η μηδενική υπόθεση H_0: $\mu = \mu_0$

Η εναλλακτική υπόθεση H_1: 1) $\mu \neq \mu_0$

2) $\mu > \mu_0$

3) $\mu < \mu_0$

Η διαδικασία ελέγχου γίνεται ως εξής:
1.Υπολογίζουμε τη δειγματική μέση τιμή $\overline{x}$.
2.Υπολογίζουμε την τιμή του κριτηρίου Z από τη σχέση:

$$Z = \frac{(\overline{x} - \mu_0) \cdot \sqrt{n}}{\sigma}$$

3) i) Για τον έλεγχο της H_0 έναντι της H_1: $\mu \neq \mu_0$, αν $|Z_\pi| > Z_{\frac{\alpha}{2}}$ τότε απορρίπτεται η H_0 και δεχόμαστε ότι $\mu \neq \mu_0$

ii) Για τον έλεγχο της H_0 έναντι της H_1: $\mu > \mu_0$, αν $Z_\pi > Z_a$ τότε απορρίπτεται η H_0 και δεχόμαστε ότι $\mu > \mu_0$

iii) Για τον έλεγχο της H_0 έναντι της H_1: $\mu < \mu_0$, αν $Z_\pi < -Z_a$ τότε απορρίπτεται η H_0 και δεχόμαστε ότι $\mu < \mu_0$.

Παράδειγμα (9.1)

Το βάρος Χ των φοιτητών, σε κιλά, ακολουθεί την κανονική κατανομή και το 1990 ήταν N(68, 150). Το 2000 πήραμε τυχαία 10 φοιτητές και τα βάρη σε κιλά ήταν: 53, 69, 62, 78, 81, 55, 66, 62, 74, 60. Μπορούμε να ισχυριστούμε με συντελεστή σημαντικότητας α = 0,05, ότι μειώθηκε ο μέσος όρος του πληθυσμού το 2000;

Λύση:

Έχουμε κανονική κατανομή N (68, 150), δηλαδή η μέση τιμή του πληθυσμού μ = 68 και διασπορά σ^2= 150. Έχουμε επίσης ένα δείγμα 10 μετρήσεων και θέλουμε να ελέγξουμε αν η μέση τιμή του δείγματος είναι μικρότερη από τη μέση τιμή του πληθυσμού δηλαδή αν $\mu<\mu_0$.

Ο έλεγχος γίνεται με το κριτήριο:

$$Z_\pi = \frac{(\overline{x} - \mu_0) \cdot \sqrt{n}}{\sigma}$$

Θέτουμε H_0 : $\overline{x} = \mu_0 = 68$ και H_1: $\overline{x} < 68$

Για να πραγματοποιήσουμε τον έλεγχο υπολογίζουμε πρώτα τη μέση τιμή $\overline{x}$.

$$\overline{x} = \sum_{i=1}^{10} x_i = \frac{53+69+62+78+81+55+66+62+74+60}{10} = 66$$

Η τυπική απόκλιση $\sigma = \sqrt{150} = 12,25$

Υπολογίζουμε το Z_π:

$$Z = \frac{(\bar{x} - \mu_0) \cdot \sqrt{n}}{\sigma} = \frac{(66 - 68) \cdot \sqrt{10}}{12,25} = -0,516$$

Για α = 0,05, η τιμή Z= 1,645

Επειδή Z_π= -0,516 > -1,645 δεχόμαστε την H_0 που ισχυρίζεται ότι μ= μ_0 και απορρίπτουμε την H_1.

Το συμπέρασμα του ελέγχου είναι ότι σε επίπεδο σημαντικότητας 0,05 ο μέσος όρος του πληθυσμού δεν μειώθηκε.

Παράδειγμα (9.2)

Η μέτρηση του ύψους σε ένα δείγμα 20 ατόμων από ένα πληθυσμό έδωσε τα παρακάτω αποτελέσματα σε εκατοστά:

173, 166, 168, 166, 169, 166, 173, 170, 170, 173,166, 161, 166, 170, 168, 158, 173, 166, 165, 165

α) Αν είναι γνωστό ότι τα ύψη ακολουθούν κανονική κατανομή N(μ, σ^2) με τυπική απόκλιση σ = 5 cm, να ελεγχθεί σε επίπεδο σημαντικότητας 5% αν το μέσο ύψος του πληθυσμού είναι μεγαλύτερο των 165 cm.

Λύση:

Από τα δεδομένα έχουμε κανονική κατανομή με γνωστή διασπορά σ= 5 cm και πραγματική μέση τιμή μ_0= 165 cm.

Το ζητούμενο είναι να ελέγξουμε αν $\mu > \mu_0$.

Θέτουμε ως μηδενική υπόθεση H_0: μ = μ_0 ,και H_1: μ> μ_0.

Ο έλεγχος γίνεται με το κριτήριο:

$$Z = \frac{(\bar{x} - \mu_0) \cdot \sqrt{n}}{\sigma}$$

Για να πραγματοποιηθεί ο έλεγχος πρέπει να έχουμε τη δειγματική μέση τιμή η οποία υπολογίζεται από τη σχέση:

$$\bar{x} = \sum_{i=1}^{20} x_i = \frac{x_1 + x_2 + x_3 + ... + x_{20}}{20} = \frac{3352}{20} = 167,6cm$$

Βρίσκουμε την τιμή του Z από τα πειραματικά δεδομένα:

$$Z_\pi = \frac{(\bar{x} - \mu_0) \cdot \sqrt{n}}{\sigma} = \frac{(167,6 - 165) \cdot \sqrt{20}}{5} = 2,325$$

$Z_{0,05} = 1,645$.

Επειδή $Z_\pi = 2,325 > (Z_{0,05} = 1,645)$ απορρίπτουμε την H_0 και δεχόμαστε την H_1, δηλαδή δεχόμαστε με πιθανότητα σφάλματος 5% (επίπεδο σημαντικότητας 0,05) ότι η μέση τιμή του ύψους του πληθυσμού είναι μεγαλύτερη από 165.

Παράδειγμα (9.3)
Για να εκτιμηθεί η μέση βαθμολογία των φοιτητών στις εξετάσεις κάποιου μαθήματος παίρνουμε ένα δείγμα 16 φοιτητών και καταγράφουμε την βαθμολογία τους:
10, 3, 5, 4, 7, 8, 9, 9, 8, 5, 5, 8, 6, 6, 9,10.
Υποθέστε ότι η βαθμολογία στις εξετάσεις ακολουθεί κανονική κατανομή N(μ, 4), (δηλαδή $\sigma^2 = 4$). Να ελεγχθεί σε επίπεδο σημαντικότητας 5%, ο ισχυρισμός ότι η μέση βαθμολογία στις εξετάσεις είναι διαφορετική του 6,5.
Λύση:
Όπως στο προηγούμενο παράδειγμα το πρώτο που κάνουμε είναι να υπολογίσουμε τη μέση τιμή $\bar{x}$ και την τυπική απόκλιση σ.

$$\bar{x} = \frac{10+3+5+4+7+8+9+9+8+5+5+8+6+6+9+10}{16} = 7$$

και $\sigma = \sqrt{4} = 2$
Θέτουμε H_0: $\mu = \mu_0$ και H_1: $\mu \neq \mu_0$.
Υπολογίζουμε το z_π από τη σχέση:

$$Z_\pi = \frac{(\bar{x} - \mu_0) \cdot \sqrt{n}}{\sigma}$$

$$Z_\pi = \frac{(7 - 6,5) \cdot \sqrt{16}}{2} = 1$$

Για να απορρίψουμε τη μηδενική υπόθεση πρέπει $|Z_\pi| > Z_{\frac{\alpha}{2}}$

Από τους πίνακες έχουμε $z_{0,025} = 1,96$
Επειδή $Z_\pi = 1 < 1,96$ δεν μπορούμε να απορρίψουμε την H_0.
Το συμπέρασμα είναι ότι δεχόμαστε ότι $\mu = \mu_0$ δηλαδή η μέση τιμή της βαθμολογίας των φοιτητών δεν είναι διαφορετική από την τιμή 6,5.

9.3.2 Έλεγχος για τη μέση τιμή μ πληθυσμού με άγνωστη διασπορά σ^2.
Όταν έχουμε ένα δείγμα από έναν κανονικό πληθυσμό με μέση τιμή μ και αγνοούμε τη διασπορά σ^2, την εκτιμούμε με τη δειγματική διασπορά S^2. Αν $\bar{x}$ και S^2 είναι η μέση τιμή και η διασπορά αντίστοιχα του δείγματος μεγέθους n, τότε ο έλεγχος για τη μέση τιμή γίνεται ως εξής:

i. Για $n \geq 30$ το κριτήριο απόφασης είναι

$$Z = \frac{\bar{x} - \mu_0}{SE} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{(\bar{x} - \mu_0) \cdot \sqrt{n}}{\sigma} = \frac{(\bar{x} - \mu_0) \cdot \sqrt{n}}{S} \qquad (9.2)$$

ii. Για $n \leq 30$ ο έλεγχος γίνεται με t-test και το κριτήριο απόφασης είναι

$$t = \frac{\bar{x} - \mu_0}{SE} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{(\bar{x} - \mu_0) \cdot \sqrt{n}}{\sigma} = \frac{(\bar{x} - \mu_0) \cdot \sqrt{n}}{S} \qquad (9.3)$$

Αξίζει να σημειωθεί ότι για $n \geq 30$ δηλαδή για μεγάλα δείγματα η τιμή της μεταβλητής z ισούται περίπου με την τιμή της μεταβλητής t της κατανομής t-student, έτσι ο έλεγχος της μέσης τιμής μ του πληθυσμού με άγνωστη διασπορά σ^2 άσχετα από τον αριθμό μετρήσεων γίνεται με t-test και κριτήριο απόφασης

$$t = \frac{(\bar{x} - \mu_0) \cdot \sqrt{n}}{\sigma} = \frac{(\bar{x} - \mu_0) \cdot \sqrt{n}}{S} \qquad (9.3)$$

Η μηδενική υπόθεση είναι πάντα H_0: $\mu = \mu_0$

Η εναλλακτική υπόθεση H_1 μπορεί να είναι μία από τις εξής τρεις υποθέσεις:

i. $\mu \neq \mu_0$

ii. $\mu > \mu_0$

iii. $\mu < \mu_0$

Ο έλεγχος γίνεται ως εξής:

1) Θέτουμε τη μηδενική υπόθεση H_0 και την εναλλακτική υπόθεση H_1.

2) Υπολογίζουμε την τιμή του κριτηρίου απόφασης t, από τη σχέση:

$$t = \frac{(\bar{x} - \mu_0) \cdot \sqrt{n}}{\sigma} = \frac{(\bar{x} - \mu_0) \cdot \sqrt{n}}{S}$$

3) Υπολογίζουμε τους βαθμούς ελευθερίας: Β.Ε= n-1

4) Συγκρίνουμε την τιμή t_π που υπολογίζεται από τα πειραματικά δεδομένα με την τιμή του t-κριτηρίου για τους βαθμούς ελευθερίας n-1 και στο επίπεδο σημαντικότητας α ή α/2 ανάλογα με την εναλλακτική υπόθεση, που βρίσκεται από πίνακες πίσω στο βιβλίο.

5) i) Για H_1: $\mu \neq \mu_0$, απορρίπτουμε την H_0 αν $|t_\pi| > t_{n-1,\frac{a}{2}}$

ii) Για H_1: $\mu > \mu_0$, απορρίπτουμε τη H_0 αν $t_\pi > t_{n-1,\alpha}$

iii) Για H$_1$: $\mu<\mu_0$, απορρίπτουμε τη H$_0$ αν $t_\pi<-t_{n-1,\alpha}$

Αξίζει να σημειωθεί ότι για τη βιοστατιστική οι επιστήμονες που ασχολούνται με την αξιολόγηση των εργαστηριακών αποτελεσμάτων υποστηρίζουν ότι η μηδενική υπόθεση H$_0$ απορρίπτεται έναντι της εναλλακτικής H$_1$: $\mu\neq\mu_0$ όταν $|t_\pi|>t_{n-1,\alpha}$.

Παράδειγμα (9.4)

Δίνονται τα ύψη 725 στρατιωτών με $\bar{x}$ = 171cm και S=5cm

Αν γνωρίζουμε ότι το πραγματικό μέσο ύψος των στρατιωτών, πριν 10 χρόνια ήταν 170 cm, μπορούμε να δεχθούμε τον ισχυρισμό ότι τώρα οι στρατιώτες είναι κατά μέσο όρο ψηλότεροι; (α=0,01)

Λύση:

Όπως βλέπουμε από τα δεδομένα έχουμε δείγμα από πληθυσμό με άγνωστη διασπορά, γνωστή πραγματική μέση τιμή μ=170 cm και μεγάλο αριθμό μετρήσεων n>30. Θέλουμε να ελέγξουμε αν η μέση τιμή του δείγματος είναι μεγαλύτερη από την πραγματική μέση τιμή. Ο έλεγχος γίνεται με το t-test και το κριτήριο απόφασης είναι:

$$t_\pi=\frac{\bar{x}-\mu}{SE}=\frac{(\bar{x}-\mu)}{S}\cdot\sqrt{n}$$

Θέτουμε πρώτα την μηδενική υπόθεση και την εναλλακτική

$$H_0:\bar{x}=\mu$$

$$H_1:\bar{x}>\mu$$

Υπολογίζουμε την τιμή του κριτηρίου t_π:

$$t_\pi=\frac{(\bar{x}-\mu)\cdot\sqrt{n}}{S}=\frac{(171-170)}{5}\cdot\sqrt{725}=5.38$$

Υπολογίζουμε τους βαθμούς ελευθερίας π-1=725-1=724

Συγκρίνουμε την τιμή t_π=5,38 με την θεωρητική τιμή του t κριτηρίου για βαθμούς ελευθερίας 724 και σε επίπεδο συμβατικότητας α=0,01:

Για 724 βαθμούς ελευθερίας $t_{0.01}$=2.33

t_π=5,38> $t_{725,0.01}$ =2.33

Επειδή t_π >$t_{n-1,\alpha}$ απορρίπτουμε την μηδενική υπόθεση και δεχόμαστε την εναλλακτική, δηλαδή δεχόμαστε τον ισχυρισμό ότι οι στρατιώτες τώρα είναι κατά μέσο όρο υψηλότεροι από ότι ήταν πριν 10 χρόνια σε επίπεδο σημαντικότητας 0,01.

Παράδειγμα (9.5)
Σε τυχαίο δείγμα 20 εργαζομένων κατοίκων μιας πόλης, βρέθηκε ότι ο χρόνος μετάβασης τους στην εργασία τους (σε λεπτά της ώρας) ήταν :

Λεπτά	**Συχνότητα**	**Λεπτά**	**Συχνότητα**
00-10	1	50-60	2
10-20	2	60-70	2
20-30	3	70-80	2
30-40	4	80-90	1
40-50	2	90-100	1

A) Να υπολογισθεί ο δειγματικός μέσος και η δειγματική διασπορά από τα παραπάνω ομαδοποιημένα δεδομένα.
B) Αν υποθέσουμε ότι ο χρόνος μετάβασης ενός εργαζόμενου κατοίκου της πόλης στην εργασία του ακολουθεί κανονική κατανομή, να εξεταστεί σε επίπεδο σημαντικότητας 5% αν ο μέσος χρόνος μετάβασης ενός εργαζόμενου κατοίκου της πόλης στην εργασία του υπερβαίνει τα 35 λεπτά.

Λύση:
A) Σε ομαδοποιημένες κατανομές ο δειγματικός μέσος υπολογίζεται από την σχέση :

$$\bar{x} = \frac{\sum v_i \kappa.\tau}{\sum v_i}$$

Άρα χρειαζόμαστε την κεντρική τιμή (κ.τ) κάθε κλάσης και το γινόμενο v_i.κτ . Η δειγματική διασπορά υπολογίζεται από την σχέση :

$$S^2 = \frac{\sum v_i(\kappa.\tau)^2 - \frac{\left(\sum v_i \kappa.\tau\right)^2}{N}}{N-1}$$

Άρα χρειαζόμαστε τα γινόμενα $v_i \cdot \kappa\tau^2 = v_i \cdot x_i^2$ και $v_i \cdot \kappa\tau = v_i \cdot x_i$

Φτιάχνουμε πίνακα με τις εξής πρόσθετες στα δεδομένα στήλες $\kappa\tau = x_i$, x_i^2 , $v_\iota \cdot x_i^2$

Λεπτά	**V_i**	**κτ= x_i**	x_i^2	$v_i \cdot x_i$	$v_i \cdot x_i^2$
00-10	1	5	25	5	25
10-20	2	15	225	30	450
20-30	3	25	625	75	1875
30-40	4	35	1225	140	4900
40-50	2	45	2025	90	4050
50-60	2	55	3025	110	6050
60-70	2	65	4225	130	8450
70-80	2	75	5625	150	11250
80-90	1	85	7225	85	7225
90-100	1	95	9025	95	9025
Σύνολο	20			910	53300

$$\overline{x} = \frac{\sum v_i x_i}{\sum v_i} = \frac{910}{20} = 45.5 \text{ λεπτά}$$

$$S^2 = \frac{\sum v_i x_i^2 - \frac{\left(\sum v_i x_i\right)^2}{N}}{N-1} = \frac{53300 - \frac{(910)^2}{20}}{19} = 626.05\ (\lambda\varepsilon\pi\tau\acute{\alpha})^2$$

και $S = \sqrt{626.05} = 25.02$ λεπτά

Β) Επειδή έχουμε κανονική κατανομή με άγνωστη διασπορά σ^2 και θέλουμε να ελέγξουμε αν η δειγματική μέση τιμή x είναι μεγαλύτερη από 35 λεπτά (σταθερή υποθετική πραγματική τιμή) πρέπει να υπολογίζουμε την τιμή t_Π του κριτηρίου t και να συγκρίνουμε με την θεωρητική τιμή $t_{n-1.\alpha}$ (πίνακας 1 του παραρτήματος)

$$t_\pi = \frac{\overline{x} - \mu}{S} \cdot \sqrt{n} = \frac{(45.5 - 35)}{23} \cdot \sqrt{20} = 1.877$$

Θέτουμε Ho: $\overline{x} = \mu$ δηλαδή 46.5 = 35

H_1: $\overline{x} > \mu$ δηλαδή 46.5 > 35

Οι βαθμοί ελευθερίας είναι 20-1=19 , $t_{19,0.05}$ =1,73 $\Rightarrow$

t_π =1,877 > $t_{19,0.05}$ =1,73

Απόφαση : απορρίπτουμε την μηδενική υπόθεση Ho και δεχόμαστε την εναλλακτική υπόθεση H_1, δηλαδή δεχόμαστε ότι $45{,}5 > 35$ σε επίπεδο 5%

Άρα ο χρόνος μετάβασης ενός εργαζόμενου κατοίκου της πόλης στην εργασία του υπερβαίνει τα 35 λεπτά.

9.4 Έλεγχος υποθέσεων για την Διαφορά των μέσων τιμών δυο δειγμάτων (ή Test σημαντικότητας για την διαφορά δυο μέσων τιμών)

Εφαρμόζεται όταν πρόκειται να συγκρίνουμε δύο ανεξάρτητους μεταξύ τους πληθυσμούς Π_1, Π_2 με βάση δυο αντίστοιχα ανεξάρτητα δείγματα, πχ όταν θέλουμε να συγκρίνουμε δυο νέα φάρμακα, δύο νέες διδακτικές μεθόδους, την απόδοση δυο φυλών αγελάδων κλπ, και πρέπει να γίνει προτίμηση ένα από τα δύο. Το test για τη διαφορά δύο μέσων τιμών απαντά στην ερώτηση: Είναι η διαφορά ανάμεσα στις δύο μέσες τιμές σημαντική ή όχι;

9.4.1 Για μεγάλα δείγματα με $n_1, n_2 \geq 30$

Για την σύγκριση των μέσων τιμών $\bar{x}_1$ και $\bar{x}_2$ δύο ανεξάρτητων δειγμάτων με n_1 και n_2 αντίστοιχα, με άγνωστη διασπορά σ^2. Το κριτήριο της απόφασης είναι:

$$Z = \frac{|\bar{x}_1 - \bar{x}_2|}{SE_\Delta} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \qquad (9.4)$$

Η μηδενική υπόθεση είναι $H_0 : \mu_1 = \mu_2$

Η εναλλακτική υπόθεση είναι: 1) $H_1 : \mu_1 \neq \mu_2$

2) $\mu_1 > \mu_2$

3) $\mu_1 < \mu_2$

Απόφαση: απορρίπτουμε την H_0 εάν

1) $|Z| > Z_{\frac{a}{2}} \Rightarrow \mu_1 \neq \mu_2$

2) $Z > Z_a \Rightarrow \mu_1 > \mu_2$

3) $Z < -Z_a \Rightarrow \mu_1 < \mu_2$

Για $n \geq 30$ η τιμή της μεταβλητής Z ισούται περίπου με την τιμή της μεταβλητής t της κατανομής t-student. Έτσι το παραπάνω κριτήριο απόφασης γίνεται

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{SE_\Delta} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \qquad (9.5)$$

Οι βαθμοί ελευθερίας είναι $= n_1 + n_2 - 2$

9.4.2 Για μικρά δείγματα με n$_1$, n$_2$ < 30

1) Στην περίπτωση δυο δειγμάτων με n_1, $n_2 < 30$ και μέσες τιμές $\overline{x}_1, \overline{x}_2$ αντίστοιχα για τα οποία υποθέτουμε ότι προέρχονται από δυο κανονικούς πληθυσμούς με την ίδια διασπορά σ^2, αλλά διαφορετικές δειγματικές διασπορές S_1^2, S_2^2, ο έλεγχος γίνεται με τη μεταβλητή t.

Θέτουμε την $H_0 : \mu_1 = \mu_2$

$H_1 : 1) \mu_1 > \mu_2$

$2) \mu_1 < \mu_2$

$3) \mu_1 \neq \mu_2$

Το κριτήριο απόφασης είναι: $t = \dfrac{\overline{x}_1 - \overline{x}_2}{S \cdot \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ (9.6)

όπου $S = \sqrt{\dfrac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{n_1 + n_2 - 2}}$ (9.7)

Απορρίπτουμε την H_0 αν 1) $t > t_{(n_1+n_2-2);a} \Rightarrow \mu_1 > \mu_2$

$2) t > -t_{(n_1+n_2-2);a} \Rightarrow \mu_1 < \mu_2$

3) $|t| > t_{(n_1+n_2-2);\frac{a}{2}} \Rightarrow \mu_1 \neq \mu_2$

2) Στην περίπτωση που τα δύο δείγματα ανήκουν σε κανονικούς πληθυσμούς με διαφορετική μέση τιμή και άγνωστες και διαφορετικές διασπορές τότε η τυχαία μεταβλητή t δίνεται ως:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \qquad (9.8)$$

και $H_0 : \mu_1 = \mu_2$

$H_1 : 1) \mu_1 > \mu_2$

$2) \mu_1 < \mu_2$

$3) \mu_1 \neq \mu_2$

Οι βαθμοί ελευθερίας = $n_1 + n_2 - 2$

Απορρίπτουμε την H_0 αν 1) $t > t_{(n_1+n_2-2);a}$

2) $t > -t_{(n_1+n_2-2);a}$

3) $|t| > t_{(n_1+n_2-2);a}$

9.5 Test σημαντικότητας για την σύγκριση μέσων τιμών ζευγαρωτών παρατηρήσεων (t – tests για παρατηρήσεις κατά ζεύγη)

Στα προηγούμενα μιλήσαμε για σύγκριση μέσων τιμών δύο κανονικών πληθυσμών , στηριζόμενοι στα στοιχεία που έχουμε από δύο δείγματα που είναι μεταξύ τους ανεξάρτητα , και δεν έχουν ατομική αντιστοιχία . Όταν οι παρατηρήσεις των δύο συγκεκριμένων ομάδων εμφανίζουν ατομική αντιστοιχία οι συγκρίσεις μπορούν να γίνουν ανά ζεύγη (paired comparisons) . Για παράδειγμα αν θέλουμε να συγκρίνουμε δύο αναλγητικά φάρμακα ως προς την αποτελεσματικότητά τους , είναι σκόπιμο να χορηγούνται τα δύο αναλγητικά στους ίδιους ασθενείς , σε διαφορετικό χρόνο και να εξετασθεί η διαφορά αποτελεσματικότητας των φαρμάκων . Μερικές φορές οι αντιστοιχούσες μετρήσεις δεν γίνονται στα ίδια άτομα ή άλλες ερευνητικές ομάδες , αλλά σε ομοιόμορφες ερευνητικές ομάδες η σε εξομοιωμένα άτομα . Έτσι είναι δυνατόν να δοθούν τα δύο φάρμακα σε δίδυμα αδέλφια , ή σε άτομα του ίδιου φύλλου και ηλικίας κ.ο.κ . Το κατά ζεύγη t – test (paired t – test) σημαντικότητας , είναι γενικώς περισσότερο ισχυρό από το συνηθισμένο t – test επειδή τεκμηριώνει κατά κανόνα και με μικρότερο αριθμό παρατηρήσεων την ενδεχόμενη στατιστική σημαντικότητας μιας πραγματικής διαφοράς .

Για την εφαρμογή της δοκιμασίας t – test κατά ζεύγη , υπολογίζονται οι διαφορές δ των αντίστοιχων μετρήσεων και τα τετράγωνα δ^2 των διαφορών αυτών . Η διαφορά δ μπορεί τώρα να θεωρηθεί ως αυθυπόστατο μέγεθος με μέση τιμή

$\bar{\delta} = \frac{\Sigma \delta i}{n}$ και σταθερή απόκλιση

SD ή $S_\delta = \sqrt{\dfrac{\Sigma\delta_i^2 - \dfrac{(\Sigma\delta_i)^2}{n}}{n-1}}$ (9.9)

και πιθανό σφάλμα $SE_\delta = \frac{S_\delta}{\sqrt{n}}$ (9.10) .

Το κριτήριο απόφασης είναι :

$$t_{\pi} = \frac{\overline{\delta}}{SE_{\delta}} = \frac{\overline{\delta}}{\frac{S_{\delta}}{\sqrt{n}}} = \overline{\delta} \cdot \frac{\sqrt{n}}{S_{\delta}} \qquad (9.11)$$

Θέτουμε H_0 : $\overline{\delta} = 0$ δηλαδή $\mu_1 = \mu_2$

H_1 : 1) $\overline{\delta} > 0$ δηλαδή $\mu_1 > \mu_2$

2) $\overline{\delta} < 0$ δηλαδή $\mu_1 < \mu_2$

3) $\overline{\delta} \neq 0$ δηλαδή $\mu_1 \neq \mu_2$

Απόφαση :

1) Αν το κριτήριο απόφασης $t_{\pi} > t_{n-1,a}$ τότε απορρίπτουμε των H_0 και δεχόμαστε ότι $\overline{\delta} > 0$ δηλαδή $\mu_1 > \mu_2$

2) Αν $t_{\pi} < -t_{n-1,a}$ τότε απορρίπτουμε την H_0 και δεχόμαστε ότι $\overline{\delta} < 0$ δηλαδή $\mu_1 < \mu_2$

3) Αν $|t_{\pi}| > t_{n-1,\frac{a}{2}}$ απορρίπτουμε την H_o και δεχόμαστε ότι $\overline{\delta} \neq 0$ δηλαδή $\mu_1 \neq \mu_2$ και πάντα σε επίπεδο σημαντικότητας α .

Παράδειγμα (9.6)

Δύο εργοστάσια Α και Β παράγουν λάστιχα αυτοκινήτων μας πληροφορούν ότι στα 20 λάστιχα του Α η διάρκεια ζωής είχε μέση τιμή $\overline{x}_1 = 27000$ km και $S_1 = 4300$ km ενώ σε 20 λάστιχα του Β η διάρκεια ζωής βρέθηκε ίση με $\overline{x}_2 = 26000$ km και $S_2 = 2900$ km . Μπορούμε να ισχυριστούμε ότι η μέση διάρκεια ζωής είναι η ίδια για τα λάστιχα Α και Β ; (α = 0,05) .

Λύση

Από τα δεδομένα διαπιστώνουμε ότι τα δύο δείγματα ανήκουν σε δυο κανονικούς πληθυσμούς με άγνωστες και διαφορετικές διασπορές . Οπότε για τον στατιστικό έλεγχο της διαφοράς των δύων μέσων τιμών $\overline{x}_1$ και $\overline{x}_2$ χρησιμοποιούμε την τ.μ t της t- κατανομής με κριτήριο απόφασης :

$$t_{\Pi} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

έχουμε $n_1 = 20$, $\overline{x}_1 = 27000$ km , $S_1 = 4300$ km

$n_2 = 20$, $\overline{x}_2 = 26000$ km , $S_2 = 2900$ km

θέτουμε $H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

Υπολογίζουμε την τιμή t_Π

$$t_\Pi = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{27000 - 26000}{\sqrt{\frac{(4300)^2}{20} + \frac{(2900)^2}{20}}} = \frac{1000}{1159,74} = 0,862$$

Οι βαθμοί ελευθερίας είναι 20 + 20 – 2 = 38 , $t_{38,0.05} = 1,645$, $t_{38,0.025} = 1,96$

Απόφαση : $t_\pi = 0,862 < 1,96 = t_{38,0.025}$

Άρα δεχόμαστε την μηδενική υπόθεση H_0 δηλαδή $\mu_1 = \mu_2$, άρα μπορούμε να ισχυριστούμε σε επίπεδο σημαντικότητας 0,05 ότι η μέση διάρκεια ζωής είναι η ίδια για τα λάστιχα Α και Β .

Παράδειγμα (9.7)

Η μέση τιμή του αναστήματος μιας ομάδας Α 64 ατόμων είναι 166,2 cm και η σταθερή απόκλισης 3,2 cm . Η μέση τιμή του αναστήματος μιας άλλης ομάδας Β 100 ατόμων είναι 167,1 cm και η σταθερή απόκλισης 3,0 cm .

Α) Να αξιολογηθεί στατιστικά η διαφορά των δύο μέσων τιμών .

Β) Να υπολογισθούν τα 95 % όρια αξιοπιστίας της διαφοράς αυτής .

Λύση

Από τα δεδομένα βλέπουμε ότι έχουμε ποσοτικά δεδομένα από τα δύο δείγματα με n_1 και $n_2 > 30$ και άγνωστες και διαφορετικές διασπορές των πληθυσμών τους . Η στατιστική αξιολόγηση γίνεται με t – test και το κριτήριο απόφασης είναι :

$$t_\Pi = \frac{\bar{x}_1 - \bar{x}_2}{SE_\Delta} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{SE_1^2 + SE_2^2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Έχουμε για την ομάδα Α : $n_1 = 64$, $\bar{x}_1 = 166,2$, $S_1 = 3,2$ cm

για την ομάδα Β : $n_2 = 100$, $\bar{x}_2 = 167,1$, $S_2 = 3,0$ cm

Θέτουμε ως μηδενική υπόθεση H_0 : $\mu_1 = \mu_2$

H_1 : $\mu_1 \neq \mu_2$

Υπολογίζουμε την t_Π και αν $|t_\Pi| > t_{(n_1+n_2-2),\frac{\alpha}{2}}$ τότε απορρίπτουμε την H_0 και δεχόμαστε την H_1

$$t_{\Pi}=\frac{166,2-167,1}{\sqrt{\frac{(3,2)^2}{64}+\frac{(3,0)^2}{100}}}=\frac{-0,9}{0,5}=-1,8$$

Οι βαθμοί ελευθερίας βε = 100 + 64 -2 = 162

$$t_{\Pi}<t_{162,0.025}=1,96$$

Άρα σε επίπεδο σημαντικότητας (α = 0,05) δεχόμαστε την μηδενική υπόθεση ότι $\mu_1=\mu_2$ και απορρίπτουμε την εναλλακτική H_1 : $\mu_1\neq\mu_2$.
Αυτό σημαίνει ότι οι δύο μέσες τιμές δεν διαφέρουν μεταξύ τους σε επίπεδο σημαντικότητας 0,05 .
Το διάστημα εμπιστοσύνης (όρια αξιοπιστίας) της διαφοράς των μέσων τιμών δίνεται από τον τύπο :

$$(\bar{x}_1-\bar{x}_2)\pm t_{(n_1+n_2-2),\frac{\alpha}{2}}\cdot\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}=0,9\pm1,96\cdot\sqrt{\frac{(3,2)^2}{64}+\frac{(3,0)^2}{100}}=$$

$=0,9\pm1,96\cdot0,5=0,9\pm0,98$ ή δ.ε είναι $-0,08<\bar{\Delta}<1,88$

Παράδειγμα (9.8)
Για να δεχτούμε ότι η περιεκτικότητα σε σίδηρο δύο περιοχών Α και Β είναι η ίδια , πήραμε 10 δείγματα από κάθε περιοχή και μετρήσαμε την περιεκτικότητα σε g/m^3 . Βρήκαμε ότι :

$\bar{x}_1=80,4\ \text{g/m}^3$ $\qquad$ $\bar{x}_2=75,2\ \text{g/m}^3$

S_1 = 3,1 g/m^3 $\qquad$ S_2 = 1,9 g/m^3

Αν δεχτούμε ότι η περιεκτικότητα σε σίδηρο ακολουθεί κανονική κατανομή τότε σε στάθμη σημαντικότητας 5% κάνετε δοκιμασία των υποθέσεων : H_0 : $\mu_1=\mu_2$, H_1 : $\mu_1\neq\mu_2$

Λύση
Τα δεδομένα δείχνουν κανονική κατανομή δύο πληθυσμών με άγνωστες και διαφορετικές διασπορές .

Έχουμε : n_1 = 10 $\quad$ $\bar{x}_1$ = 80,4 g/m^3 $\quad$ S_1 = 3,1 g/m^3

n_2 = 10 $\quad$ $\bar{x}_2$ = 75,2 g/m^3 $\quad$ S_2 = 1,9 g/m^3

Το κριτήριο απόφασης είναι η δοκιμασία t – test με τ.μ. $t_{\Pi}=\dfrac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$.

Έχουμε H_0 : $\mu_1=\mu_2$,

$H_1 : \mu_1 \neq \mu_2$

Υπολογίζουμε την τιμή t_{Π} από τα δεδομένα :

$t_{\Pi} = \dfrac{80,4-75,2}{\sqrt{\dfrac{(3,1)^2}{10}+\dfrac{(1,9)^2}{10}}} = \dfrac{5,2}{1,15} = 4,52$ Συγκρίνουμε t_{π} και $t_{n_1+n_2-2,\frac{\alpha}{2}}$. Από τον πίνακα (2) βρίσκουμε

$t_{18,0.025} = 2,101$ οπότε

$t_{\Pi} = 4,52 > t_{18,0.025} = 2,101$

Απόφαση: απορρίπτουμε την H_0 και δεχόμαστε την H_1 δηλαδή δεχόμαστε ότι $\mu_1 \neq \mu_2$ σε επίπεδο σημαντικότητας 0,05 .

Παράδειγμα (9.10)

Σε τυχαίο δείγμα 500 οικογενειών από αγροτικές περιοχές υπολογίσαμε μέσο αριθμό παιδιών ανά οικογένεια $\bar{x}_1$ = 4,26 και τυπική απόκλιση S_1 = 3,85 , ενώ τυχαίο δείγμα 420 οικογενειών αστικών περιοχών υπολογίσαμε $\bar{x}_2$ = 3,2 και S_2 = 4,12 . Να ελεγχθεί η υπόθεση ότι η διαφορά στον μέσο αριθμό ανά οικογένεια στις δύο περιοχές είναι στατιστικά σημαντική σε επίπεδο σημαντικότητας α = 0,05 .

Λύση

Πρόκειται για δύο δείγματα με n_1 και n_2 > 30 τα οποία προέρχονται από δυο διαφορετικούς πληθυσμούς με άγνωστες διασπορές οπότε το κριτήριο ελέγχου και απόφασης είναι :

$$Z_{\Pi} \cong t_{\Pi} = \frac{\bar{x}_1 - \bar{x}_2}{SE_{\Delta}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}}}$$

Έχουμε $n_1 = 500$ $\bar{x}_1 = 4,26$ $S_1 = 3,85$

$n_2 = 420$ $\bar{x}_2 = 3,2$ $S_2 = 4,12$

Θέτουμε $H_0 : \mu_1 = \mu_2$ δηλαδή $\bar{x}_1 = \bar{x}_2$

$H_1 : \mu_1 \neq \mu_2$

Υπολογίζουμε την τ.μ t_{Π} , και αν $|t_{\Pi}| > t_{n1+n2-2,\frac{\alpha}{2}}$ τότε απορρίπτουμε την H_0 και δεχόμαστε την H_1 .

$$Z_{\Pi} = t_{\Pi} = \frac{4,26-3,2}{\sqrt{\frac{(3,85)^2}{500}+\frac{(4,12)^2}{420}}} = \frac{1,06}{0,265} = 4$$

Η συγκριτική τιμή $t_{n_1+n_2-2,\frac{\alpha}{2}} = t_{918,0.025} = 1,96$

Επειδή

$t_{\Pi} = 4 > t_{918,0.025} = 1,96$ απορρίπτουμε την μηδενική υπόθεση H_0 : $\mu_1=\mu_2$ και δεχόμαστε την εναλλακτική H_1 : $\mu_1 \neq \mu_2$, σε επίπεδο σημαντικότητας α = 0,05 . Με άλλα λόγια υπάρχει διαφορά στον μέσο αριθμό παιδιών ανά οικογένεια στις δύο περιοχές σε επίπεδο στατιστικά σημαντικό $p < 0,05$.

Παράδειγμα (9.11)

Σε δύο ομάδες ποντικιών (ίδιας ηλικίας και φύλλου) δόθηκε , μετά από τυχαιοποίηση , για δύο μήνες διαφορετικό σε περιεκτικότητα πρωτεϊνών είδος διατροφής . Η ομάδα Α είχε διατροφή υψηλής περιεκτικότητας σε πρωτεΐνες , ενώ η Β χαμηλής . Το βάρος (σε γραμμάρια) που κέρδισαν τα ποντίκια στο διάστημα αυτό ήταν :

Ομάδα Α : 136 144 100 123 128 150 107 90 110 125 92 130 Ομάδα Β : 60 122 102 102 89 105 140 95 .

Επιδρά το διαφορετικό είδος διατροφής στο βάρος των ποντικιών αυτών ;

Λύση :

Εδώ έχουμε δύο δείγματα με $n_1 < 30$, $n_2 < 30$ από δύο πληθυσμούς με άγνωστες διασπορές . Τα δεδομένα είναι ποσοτικά και το ζητούμενο είναι να αξιολογήσουμε στατιστικά αν επιδρά το είδος της διατροφής στο βάρος των ποντικιών , δηλαδή να αξιολογήσουμε στατιστικά αν υπάρχει διαφορά στην μέση τιμή της αύξησης του βάρους των δύο ομάδων . Η αξιολόγηση γίνεται με την δοκιμασία t – test και το κριτήριο απόφασης είναι :

$$t_{\Pi} = \frac{\bar{x}_1-\bar{x}_2}{SE_{\Delta}} = \frac{\bar{x}_1-\bar{x}_2}{\sqrt{SE_1^2+SE_2^2}} = \frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$$

Θέτουμε ως H_0 : $\mu_1 = \mu_2$ δηλαδή $\bar{x}_1 = \bar{x}_2$

H_1 : $\mu_1 \neq \mu_2$ δηλαδή $\bar{x}_1 \neq \bar{x}_2$

Για να υπολογίσουμε την τιμή t_{Π} χρειαζόμαστε την μέση τιμή κάθε ομάδας η οποία υπολογίζεται από τον τύπο $\overline{x} = \dfrac{\Sigma x_1}{n_1}$,και την τυπική απόκλιση από τον τύπο

$$S_1 = \sqrt{\frac{\Sigma x_1^2 - \frac{(\Sigma x_1)^2}{n_1}}{n_1 - 1}}$$

Αυτό σημαίνει ότι χρειαζόμαστε πίνακα με τις στήλες x_1, x_2, x_1^2, x_2^2 ο οποίος φαίνεται παρακάτω

	x_1	x_2	x_1^2	x_2^2
1	136	60	18496	3600
2	144	128	20736	16384
3	100	102	10000	10404
4	123	102	15129	10404
5	128	89	16384	7921
6	150	105	22500	11025
7	107	140	11449	19600
8	90	90	8100	8100
9	110		12100	
10	125		15625	
11	92		8464	
12	130		16900	
Συν.	Σx_1=1435	Σx_2=816	Σx_1^2=175883	Σx_2^2=87438

Από τον πίνακα υπολογίζουμε τα $\overline{x_1}$, $\overline{x_2}$, S_1 , S_2 .

$$\overline{x}_1 = \frac{\Sigma x_1}{n_1} = \frac{1435}{12} = 119,58 \quad , \; \overline{x}_2 = \frac{\Sigma x_2}{n_2} = \frac{816}{8} = 102$$

$$S_1 = \sqrt{\frac{\Sigma x_1^2 - \frac{(\Sigma x_1)^2}{n_1}}{n_1 - 1}} = \sqrt{\frac{175883 - \frac{(1435)^2}{12}}{11}} = 19,73$$

$$S_2 = \sqrt{\frac{\Sigma x_2^2 - \frac{(\Sigma x_2)^2}{n_2}}{n_2 - 1}} = \sqrt{\frac{87438 - \frac{(816)^2}{8}}{7}} = 24,51$$

Και τώρα υπολογίζουμε το t_Π .

$$t_\Pi = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{119,58 - 102}{\sqrt{\frac{(19,73)^2}{12} + \frac{(24,51)^2}{8}}} = \frac{17,58}{10,37} = 1,69$$

Από τον πίνακα (2) και για βαθμούς ελευθερίας 12 + 8 -2 = 18 βρίσκουμε ότι $t_{18,0.05} = 1,734$.

$t_\Pi = 1,69 < t_{18,0.05} = 1,734$.

Η απόφαση είναι ότι δεχόμαστε την H_0 και απορρίπτουμε την H_1 δηλαδή το είδος της διατροφής δεν επιδρά στο βάρος των ποντικιών των δύο ομάδων .

Παράδειγμα (9.12)

Μετρήθηκε η συστολική αρτηριακή πίεση 5 ατόμων και βρέθηκε ίση με 16 , 15 , 18 , 22 και 23 cm Hg. Η πίεση τους μετρήθηκε ξανά 6 ώρες μετά την χορήγηση του φαρμάκου Α και βρέθηκε ίση με 14 ,12 12 , 18 και 20 αντίστοιχα . Άλλα 4 άτομα (ίδιου φύλλου και ηλικίας με τα προηγούμενα) είχαν συστολική αρτηριακή πίεση 17 , 18 , 22 και 18 cm Hg πριν την χορήγηση του φαρμάκου Β , ενώ 6 ώρες μετά η πίεση τους ήταν 15, 15, 18 και 17 cm Hg αντίστοιχα . Διαφέρει η ενδεχόμενη αντιυπερτασική δράση των φαρμάκων Α και Β .

Λύση

Όπως βλέπουμε έχουμε δύο ομάδες από διαφορετικούς πληθυσμούς με άγνωστες διασπορές . Οι δύο ομάδες παίρνουν διαφορετικό αντιυπερτασικό φάρμακο και ζητείται η στατιστική αξιολόγηση της διαφοράς των μέσων ελάττωσης της αρτηριακής πίεσης που προκαλούν τα φάρμακα αυτά . Αυτό απαιτεί πρώτα να υπολογίσουμε την μέση μεταβολή για κάθε ομάδα . Η αξιολόγηση της διαφοράς των μέσων τιμών X_A και X_B γίνεται με t-test .

Το κριτήριο απόφασης είναι :

$$t_\Pi = \frac{\overline{x}_A - \overline{x}_B}{SE_\Delta} = \frac{\overline{x}_A - \overline{x}_B}{\sqrt{SE_A^2 + SE_B^2}} = \frac{\overline{x}_A - \overline{x}_B}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}}$$

Θέτουμε ως H_0 : $\mu_1 = \mu_2$ δηλαδή $\overline{x}_A = \overline{x}_B$

$H_1 : \mu_1 \neq \mu_2$ δηλαδή $\bar{x}_A \neq \bar{x}_B$

Για τον υπολογισμό της τιμής t_π χρειαζόμαστε τις μεταβολές X_A και X_B και X_A^2, X_B^2 για τον υπολογισμό των τυπικών αποκλίσεων S_A και S_B. Τα μεγέθη που χρειάζονται φαίνονται στο παρακάτω πίνακα .

A				B				
	πριν	Μετά	$X_A=\delta_A$	πριν	μετά	$X_B=\delta_B$	X_A^2	X_B^2
1	16	14	2	17	15	2	4	4
2	15	12	3	18	15	3	9	9
3	18	12	6	22	18	4	36	16
4	22	18	4	18	17	1	16	1
5	23	20	3				9	
Σύνολο			18			10	74	30
			$\Sigma\chi_A$			ΣX_B	ΣX_A^2	ΣX_B^2

$$\bar{x}_A = \frac{\Sigma x_A}{n_A} = \frac{18}{5} = 3,6 \qquad \bar{x}_B = \frac{\Sigma x_B}{n_B} = \frac{10}{4} = 2,5 \quad ,$$

$$S_A = \sqrt{\frac{\Sigma x_A^2 - \frac{(\Sigma x_A)^2}{n_A}}{n_A - 1}} = \sqrt{\frac{74 - \frac{(18)^2}{5}}{4}} = 1,52$$

$$S_B = \sqrt{\frac{\Sigma x_B^2 - \frac{(\Sigma x_B)^2}{n_B}}{n_B - 1}} = \sqrt{\frac{30 - \frac{(10)^2}{4}}{3}} = 1,29$$

$$t_\Pi = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}} = \frac{3,6 - 2,5}{\sqrt{\frac{1,52^2}{5} + \frac{1,29^2}{4}}} = 1,174$$

Από τον πίνακα (2) βρίσκουμε ότι $t_{9,0.05} = 1,895$ Επειδή

$t_\Pi = 1,174 < t_{9,0.05} = 1,895$ δεχόμαστε την $H_0 : \mu_1 = \mu_2$ δηλαδή $\bar{x}_A = \bar{x}_B$

Απόφαση : Δεν διαφέρει η ενδεχόμενη αντιυπερτασική δράση των φαρμάκων Α και Β σε επίπεδο σημαντικότητας 0.05.

Πρόβλημα (9.13)
Σε 10 ασθενείς που υπέφεραν από φοβία χορηγήθηκε για 2 μήνες ένα νέο φάρμακο . Μετά από διακοπή ενός μηνός χορηγήθηκε για δύο μήνες ένα "αδρανές" φάρμακο (placebo) . Η απάντηση αξιολογήθηκε με μία κλίμακα (score) και ήταν η εξής.

Φάρμακο	1ος	2ος	3ος	4ος	5ος	6ος	7ος	8ος	9ος	10ος
ΝΕΟ	21	13	16	19	25	13	17	21	13	20
Αδρανές	24	20	19	21	24	14	16	13	21	19

Υπάρχει διαφορά στην απάντηση φοβίας μεταξύ του νέου και του "αδρανούς" φάρμακου ;

Λύση
Σε αυτό το παράδειγμα όπως βλέπουμε οι μετρήσεις γίνανε στα ίδια άτομα για τα δύο φάρμακα , δηλαδή υπάρχει ατομική αντιστοιχία . Άρα η σύγκριση μπορεί να γίνει ανά ζεύγη με test σημαντικότητας t – test κατά ζεύγη (ζευγαρωτών παρατηρήσεων) .

Για την εφαρμογή της δοκιμασίας t – test κατά ζεύγη υπολογίζουμε τις διαφορές δ των αντίστοιχων μετρήσεων και το δ^2 όπως φαίνεται στο παρακάτω πίνακα :

	ΝΕΟ	Αδρανές	δ	δ^2
1	21	24	3	9
2	13	20	7	49
3	16	19	3	9
4	19	21	2	4
5	25	24	-1	1
6	13	14	1	1
7	17	16	1	1
8	21	13	-8	64
9	13	21	8	64
10	20	19	-1	1
Σύνολο			15	203

Το κριτήριο απόφασης είναι $t_{\Pi} = \dfrac{\overline{\delta}}{SE_{\delta}} = \overline{\delta} \cdot \dfrac{\sqrt{n}}{S_{\delta}}$

Άρα χρειαζόμαστε το $\overline{\delta}$ και S_{δ} . Η $\overline{\delta}$ υπολογίζεται από την σχέση :

$\bar{\delta} = \frac{\Sigma\delta_i}{n} = \frac{15}{10} = 1,5$ και η S_δ από την σχέση

$$S_\delta = \frac{\Sigma\delta_i^2 - \frac{(\Sigma\delta_i^2)}{n}}{n-1} = \sqrt{\frac{203 - \frac{(15)^2}{10}}{9}} = 4,48$$

Θέτουμε H_0 : δ=0 δηλαδή $\mu_1 = \mu_2$

H_1 : $\delta \neq 0$ δηλαδή $\mu_1 \neq \mu_2$

Υπολογίζουμε την τιμή t_π .

$$t_\Pi = \frac{\bar{\delta}\cdot\sqrt{n}}{S_\delta} = \frac{1,5\cdot\sqrt{10}}{4,48} = 1,06$$

Συγκρίνουμε την $|t_\Pi|$ με την τιμή πινάκων $t_{n-1,\frac{\alpha}{2}}$

$|t_\Pi| = 1,06 < 1,372$ $(t_{\alpha,0,1})$

Επειδή $|t_\Pi| < t_{\alpha,0,1}$ τότε δεχόμαστε την μηδενική υπόθεση δηλαδή δ= 0 ή $\mu_1=\mu_2$ οπότε δεν υπάρχει διαφορά στην απάντηση φοβίας μεταξύ του νέου και του αδρανούς φαρμάκου . (P>0,1)

Παράδειγμα (9.14)

Σε 5 ζεύγη μονογενών διδύμων έγινε ένα ψυχολογικό τεστ για να διερευνηθεί αν το πρωτότοκο παιδί είναι πιο επιθετικό από το δευτερότοκο (μεγαλύτερη τιμή σημαίνει μεγαλύτερη επιθετικότητα) . Τα αποτελέσματα ήταν :

Ζεύγος	Πρωτότοκο	Δευτερότοκο
1	86	88
2	71	77
3	77	76
4	68	64
5	91	90

Λύση

Όπως βλέπουμε έχουμε ζεύγη διδύμων που γίνεται η ίδια μέτρηση στα δυο άτομα (πρωτότοκο και δευτερότοκο) , δηλαδή υπάρχει ατομική αντιστοιχία

οπότε η σύγκριση των μέσων τιμών γίνεται με t – test κατά ζεύγη (ζευγαρωτές παρατηρήσεις) .
Το κριτήριο απόφασης είναι :

$$t_{\Pi} = \frac{\overline{\delta}}{SE_{\delta}} = \frac{\overline{\delta}}{S_{\delta}} \cdot \sqrt{n}$$

όπου $\overline{\delta}$ η μέση διαφορά και δίνεται από την σχέση $\overline{\delta} = \frac{\Sigma\delta_i}{n}$, και S_{δ} είναι η σταθερά απόκλισης της διαφοράς και δίνεται από την σχέση :

$$S_{\delta} = \sqrt{\frac{\Sigma\delta_i^2 - \frac{(\Sigma\delta_i)^2}{n}}{n-1}}$$

Από τις παρακάτω σχέσεις βλέπουμε ότι χρειαζόμαστε να υπολογίσουμε τις διαφορές δ_i και τα τετράγωνα τους δ_i^2 . Τα δεδομένα και οι υπολογισμοί φαίνονται στο παρακάτω πίνακα .

Ζεύγος	Πρωτότοκος	Δευτερότοκος	δ	δ^2
1	86	88	-2	4
2	71	77	-6	36
3	77	76	+1	1
4	68	64	+4	16
5	91	90	+1	1
Σύνολο			-2	58
			$\Sigma\delta_i$	$\Sigma\delta_i^2$

Επειδή διερευνούμε αν το πρωτότοκο είναι πιο επιθετικό από το δευτερότοκο δηλαδή αν $\mu_1 > \mu_2$, θέτουμε :

H_0 : $\mu_1=\mu_2$ δηλαδή $\overline{\delta} = 0$

H_1 : $\mu_1 > \mu_2$ δηλαδή $\overline{\delta} > 0$

Υπολογίζουμε τις τιμές $\overline{\delta}$, S_{δ} , t_{Π} :

$$\overline{\delta} = \frac{\Sigma\delta_i}{n} = \frac{-2}{5} = -0,4$$

$$S_{\delta}=\sqrt{\frac{\Sigma\delta_i^2-\frac{(\Sigma\delta_i)^2}{n}}{n-1}}=\sqrt{\frac{58-\frac{(-2)^2}{5}}{4}}=3,78$$

$$t_{\Pi}=\frac{\overline{\delta}}{S_{\delta}}\cdot\sqrt{n}=\frac{-0,4}{3,78}\cdot\sqrt{5}=-0,24$$

Βαθμοί ελευθερίας = n-1= 5-1= 4

$t_{\Pi}=-0,24<t_{4,0,025}$ = 2,776 p > 0,05

Επίσης $t_{\Pi}=-0,24<t_{4,0,05}$ = 2,131 p > 0,10

Άρα δεχόμαστε την μηδενική υπόθεση H_0 και απορρίπτουμε την H_1 , δηλαδή τα πρωτότοκα δεν είναι πιο επιθετικά από τα δευτερότοκα σε επίπεδο σημαντικότητας α=0,05 αλλά ούτε σε επίπεδο σημαντικότητας α = 0,10 .

Παράδειγμα (9.15)

Σε 10 ασθενείς χορηγήθηκαν για την καταπολέμηση της αϋπνίας τους δυο διαφορετικά φάρμακα Α και Β . Οι επιπλέων ώρες ύπνου των 10 αυτών ασθενών , οι οποίες οφείλονται στην λύση του φαρμάκου Α έναντι του Β , ήταν : 1,2 2,4 1,3 1,0 1,8 0,8 4,6 1,4 1,3

Α) Να εξεταστεί αν η υπνωτική επίδραση των δυο αυτών φαρμάκων διαφέρει

Β)Να υπολογιστούν τα 95% όρια αξιοπιστίας της μέσης αύξησης των ωρών ύπνου η οποία οφείλεται στην λήψη του φαρμάκου Α αντί Β .

Λύση

Από τις πληροφορίες της άσκησης διαπιστώνουμε ότι έχουμε ποσοτικές μετρήσεις , και έχουμε ατομική αντιστοίχηση διότι τα δύο φάρμακα δόθηκαν στους ίδιους ασθενείς . Οπότε η αξιολόγηση της διαφοράς της υπνωτικής επίδρασης των δυο φαρμάκων γίνεται με t – test κατά ζεύγη . Στην άσκηση δόθηκε η διαφορά στις ώρες ύπνου των δυο φαρμάκων μ_A-μ_B δηλαδή το δ .

Το κριτήριο απόφασης είναι $t_{\Pi}=\frac{\overline{\delta}}{S_{\delta}}\cdot\sqrt{n}$

Άρα χρειαζόμαστε $\overline{\delta}$ και S_{δ} .

$$\overline{\delta}=\frac{\Sigma\delta_i}{n} \text{ και } S_{\delta}=\sqrt{\frac{\Sigma\delta_i^2-\frac{(\Sigma\delta_i)^2}{n}}{n-1}}$$

Όπως βλέπουμε χρειαζόμαστε εκτός από το δ και το δ^2 οπότε έχουμε :

Ζεύγος	δ	δ^2
1	1.2	1.4
2	2.4	5.76
3	1.3	1.69
4	0.0	0.0
5	1.0	1.0
6	1.8	3.24
7	0.8	0.64
8	4.6	21.16
9	1.4	1.96
10	1.3	1.69
Σύνολο	15.8	38.58
	$\Sigma\delta_i$	$\Sigma\delta_i^2$

Υπολογίζουμε τα μεγέθη $\overline{\delta}$ και S_δ .

$$\overline{\delta} = \frac{15,8}{10} = 1,58 \text{ και } S_\delta = \sqrt{\frac{38,58 - \frac{(15,8)^2}{10}}{9}} = 1,23 \ .$$

Θέτουμε $H_0 : \delta = 0$ δηλαδή $\mu_A = \mu_B$

$H_1 : \delta \neq 0$ δηλαδή $\mu_A \neq \mu_B$

Υπολογίζουμε την τιμή t_Π : $t_\Pi = \frac{\overline{\delta}}{S_\delta} \cdot \sqrt{n} = \frac{1,58 \cdot \sqrt{10}}{1,23} = 4,06$

Συγκρίνουμε την t_Π με την τιμή t πινάκων και έχουμε

$t_\Pi = 4,06 > 3,250 = t_{9,0,005}$

Άρα σε στατιστικό επίπεδο 0,01 (P<0,01) απορρίπτουμε την μηδενική υπόθεση και δεχόμαστε την H_1 δηλαδή δεχόμαστε ότι $\delta \neq 0$ ή $\mu_A \neq \mu_B$, με άλλα λόγια δεχόμαστε ότι η υπνωτική επίδραση των δυο φαρμάκων Α και Β διαφέρουν σε επίπεδο στατιστικά πολύ σημαντικό (α=0,01) με P<0,01 .

Β) Το διάστημα εμπιστοσύνης της διαφοράς ζευγαρωτών παρατηρήσεων δίνεται από τον τύπο $\delta \cdot \varepsilon = \overline{\delta} \pm t_{\frac{a}{2}, n-1} \cdot \frac{S_\delta}{\sqrt{n}} = 1.58 \pm 2,26 \cdot \frac{1,23}{\sqrt{10}} = 1,58 \pm 0,88$

Δηλαδή το 95% είναι 1,58-0,88 , 1,58+0,88 ή $\delta \cdot \varepsilon = 0,70/2,46$.

Παράδειγμα (9.16)
Με μια νέα μέθοδο καθαρισμού ενός μεταλλεύματος , βρέθηκε ότι ο χρόνος που χρειάζεται για καθαρισμό 1 κιλού ήταν σε 9 διαφορετικά δείγματα σε min
35 31 29 25 34 40 27 32 31 .
Μία άλλη μέθοδος καθαρισμού χρησιμοποιήθηκε σε 7 δείγματα του 1 κιλού και χρειάστηκαν σε min
32 37 35 28 41 44 35.
i) Μπορούμε να ισχυριστούμε ότι οι δύο μέθοδοι καθορισμού είναι διαφορετικές (α = 0,05)
ii) Δώστε 95% δ.ε για την διαφορά των μέσων τιμών του χρόνου καθαρισμού 1 κιλού μεταλλεύματος , με τις δύο μεθόδους .
Λύση
Από τα δεδομένα της άσκησης βλέπουμε ότι έχουμε δείγματα με μικρό αριθμό μετρήσεων με άγνωστες και διαφορετικές διασπορές θεωρούμε ότι τα δείγματα ανήκουν σε κανονικούς πληθυσμούς .
Τα δεδομένα είναι ποσοτικά και το ζητούμενο είναι να εξετάσουμε αν $\mu_1 \neq \mu_2$. Σε τέτοιες περιπτώσεις το test που εφαρμόζεται είναι t – test με τ.μ :

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Για τον υπολογισμό της τιμής t χρειαζόμαστε $\overline{x}_1$, $\overline{x}_2$, S_1 , S_2 . Για τον σκοπό αυτόν φτιάχνουμε τον παρακάτω πίνακα με τις στήλες x_1, x_2, x_1^2 , x_2^2 .

	X_1	X_1^2	X_2	X_2^2
1	35	1225	32	1024
2	31	961	37	1369
3	29	841	35	1225
4	25	625	28	784
5	34	1156	41	1681
6	40	1600	44	1936
7	27	729	35	1225
8	32	1024	-	-
9	31	961	-	-
	Σx_1 =284	Σx_1^2 =9122	Σx_2 =252	Σx_2^2 =9244

Από τον πίνακα υπολογίζουμε :

$$\bar{x}_1 = \frac{\Sigma x_1}{n_1} = \frac{284}{9} = 31,56$$

$$\bar{x}_2 = \frac{\Sigma x_2}{n_2} = \frac{252}{7} = 36$$

$$S_1 = \sqrt{\frac{\Sigma x_1^2 - \frac{(\Sigma x_1)^2}{n_1}}{n_1}} = \sqrt{\frac{9122 - \frac{(284)^2}{9}}{8}} = 4,475$$

$$S_2 = \sqrt{\frac{\Sigma x_2^2 - \frac{(\Sigma x_2)^2}{n_2}}{n_2 - 1}} = \sqrt{\frac{9244 - \frac{(252)^2}{7}}{6}} = 5,354$$

Ορίζουμε H$_0$: $\mu_1 = \mu_2$ και H$_1$: $\mu_1 \neq \mu_2$

Υπολογίζουμε την t_Π : $t_\Pi = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} = \dfrac{31,56 - 36}{\sqrt{\dfrac{(4,475)^2}{9} + \dfrac{(5,354)^2}{7}}} = -1,766$

Οι βαθμοί ελευθερίας είναι $n_1 + n_2 - 2 = 9 + 7 - 2 = 14$

Από τους πίνακες η τιμή $t_{n_1+n_2-2,\frac{a}{2}} = 2,145$

Συγκρίνουμε την $|t_\Pi|$ με $t-$πινάκων οπότε $|t_\Pi| = 1,766 < 2,145$

Συμπέρασμα : δεχόμαστε την H$_0$ και απορρίπτουμε την H$_1$.
Δηλαδή μπορούμε σε α=0,05 να ισχυριστούμε ότι οι δύο μέθοδοι καθαρισμού **Δεν** είναι διαφορετικές .

9.6 Έλεγχος υποθέσεων (test σημαντικότητας) για την P της διωνυμικής κατανομής (έλεγχος για αναλογία)

Ο έλεγχος σημαντικότητας για την αναλογία P ενός πληθυσμού γίνεται με βάση τη δειγματική αναλογία $p = \dfrac{X}{n}$ (κατανομή Bernoulli) όπου X ο αριθμός των επιτυχιών και n το σύνολο των επαναλήψεων (προσπαθειών). Η αναλογία αυτή ακολουθεί κανονική κατανομή, ή η διωνυμική κατανομή στην οποία ανήκει η αναλογία αυτή προσεγγίζει την κανονική κατανομή όταν

i) $n.p \geq 5$ ii) $[n.q = n(1-p)] \geq 5$.

Η τυπική απόκλιση S της αναλογίας p βρίσκεται από την σχέση

$$S = \sqrt{\frac{p.q}{n}} = \sqrt{\frac{p(1-p)}{n}} \qquad (9.12)\,.$$

Όταν ισχύουν οι παραπάνω προϋποθέσεις, δηλαδή όταν η αναλογία ακολουθεί ή προσεγγιστικά ακολουθεί κανονική κατανομή τότε ο στατιστικός έλεγχος για την αναλογία p θα γίνει με τρόπο ανάλογο με τον έλεγχο για τη μέση τιμή, ως εξής:

Θέτουμε την μηδενική και την εναλλακτική υπόθεση

H_o: $P=P_o$ H_1: i) $P \neq P_o$ για δίπλευρο test

ii) $P>P_o$ για μονόπλευρο δεξί test, iii) $P<P_o$ για μονόπλευρο αριστερό test

Το κριτήριο απόφασης είναι: $$Z_{\Pi} = \frac{P-P_o}{\sqrt{\dfrac{P_o(1-P_o)}{n}}} \qquad (9.13)$$

υπολογίζεται το Z_π από την παραπάνω σχέση και απορρίπτουμε την μηδενική υπόθεση σε επίπεδο σημαντικότητας α, για τα τρία είδη ελέγχου αντίστοιχα αν:

i) $|Z_\pi| \geq Z_{\frac{\alpha}{2}}$ Τότε δεχόμαστε ότι $p \neq p_0$

ii) $Z_\pi > Z_\alpha$ Τότε δεχόμαστε ότι P>Po

iii) $Z_\pi < -Z_\alpha$ Τότε δεχόμαστε ότι P<Po

Παράδειγμα (9.17)

Είναι γνωστό ότι η πιθανότητα να πάσχει ένα άτομο από αχρωματοψία είναι 0.08. Σ' ένα δείγμα 150 ατόμων είναι πιθανό να βρεθούν 20 άτομα , που να πάσχουν από αχρωματοψία;

Λύση:

Από τα δεδομένα έχουμε $P_0 = 0.08$.
Η πιθανότητα να βρεθούν 20 άτομα με αχρωματοψία είναι :

$$P = \frac{20}{150} = 0{,}133$$

Θέτουμε την μηδενική και την εναλλακτική υπόθεση
Ho: P = Po . H1: P≠Po για δίπλευρο test

Το κριτήριο απόφασης είναι: $Z_{\Pi} = \dfrac{P - P_o}{\sqrt{\dfrac{P_o(1-P_o)}{n}}}$

Υπολογίζεται το Z_{Π} από την παραπάνω σχέση και απορρίπτουμε την μηδενική υπόθεση σε επίπεδο σημαντικότητας α, αν:

$|Z_{\pi}| \geq Z_{\frac{\alpha}{2}}$ Τότε δεχόμαστε ότι P≠Po

$$Z_{\Pi} = \frac{P - P_o}{\sqrt{\dfrac{P_o(1-P_o)}{n}}} = \frac{0.133 - 0.08}{\sqrt{\dfrac{0.08(1-0.08)}{150}}} = 2.26$$

$Z_{\pi} = 2{,}26 > Z_{0{,}025} = 1{,}96$

Συμπέρασμα: απορρίπτουμε την H_0 σε επίπεδο σημαντικότητας α=0,05 δηλαδή η πιθανότητα να βρεθούν 20 άτομα με αχρωματοψία είναι $\langle$ 0,05

Αξίζει να σημειωθεί ότι η αξιολόγηση μπορεί να γίνει με t-test:

Το κριτήριο απόφασης είναι $t_{\Pi} = \dfrac{P - P_0}{\sqrt{\dfrac{P_0(1-P_0)}{n}}}$

Υπολογίζουμε την τιμή t_{π} και απορρίπτουμε την H_0 αν $t_{\pi} > t_{\alpha}$

$$t_{\Pi} = \frac{P - P_0}{\sqrt{\dfrac{P_0(1-P_0)}{n}}} = 2.26 = Z_{\pi}\ , \quad β.ε = 150-1 = 149$$

$t_{\pi} = 2{,}26 > t_{0{,}05} = 1{,}96$

Συμπέρασμα: απορρίπτουμε την H_0 σε επίπεδο σημαντικότητας α=0,05 δηλαδή η πιθανότητα να βρεθούν 20 άτομα με αχρωματοψία είναι $\langle$ 0,05 .

9.7 Σύγκριση δύο αναλογιών P_1 και P_2

Ο έλεγχος της σημαντικότητας διαφοράς δύο αναλογιών (ποσοστών) P_1 και P_2 από τους πληθυσμούς 1 και 2 αντίστοιχα γίνεται ως εξής:

1. **Βρίσκουμε** τις πιθανότητες επιτυχίας (ποσοστά) P_1 και P_2 από τους τύπους $P_1 = \frac{x_1}{n_1}$, όπου x_1 = επιτυχίες από πληθυσμό 1, n_1 = τυχαίες προσπάθειες από τον πληθυσμό 1.

$P_2 = \frac{x_2}{n_2}$, όπου x_2 = επιτυχίες σε n_2 τυχαίες προσπάθειες από το πληθυσμό 2.

2.**Υπολογίζουμε** την παράμετρο P από τα στοιχεία των δύο δειγμάτων με βάση τη σχέση : $P = \frac{x_1 + x_2}{n_1 + n_2}$ (9.14)

3.**Υπολογίζουμε** την παράμετρο Z_π (κριτήριο απόφασης) από την σχέση

$$Z_\pi = \frac{P_1 - P_2}{\sqrt{P(1-P)(\frac{1}{n_1} + \frac{1}{n_2})}} \qquad (9.15)$$

4.**Θέτουμε** την μηδενική και την εναλλακτική υπόθεση

H_o: $P_1 = P_2$

H_1: i) $P_1 \neq P_2$ ii) $P_1 > P_2$ iii) $P_1 < P_2$

5.**Απορρίπτουμε** την H_o και δεχόμαστε την H_1 αντίστοιχα αν

i) $|Z_\Pi| > Z_{\alpha/2}$ ii) $Z_\pi > Z_\alpha$ iii) $Z_\pi < - Z_\alpha$

όπου Z_α, $Z_{\alpha/2}$ είναι οι γνωστές τιμές της τυπικής κατανομής.

Παράδειγμα (9.18)

Στο μάθημα της στατιστικής και τις δύο περιόδους έδωσαν 90 και 118 φοιτητές και πέρασαν το μάθημα 63 και 85 αντίστοιχα. Να εξεταστεί αν η πιθανότητα επιτυχίας και στις δύο περιόδους είναι η ίδια (α=0,05).

Λύση

Η στατιστική αξιολόγηση διαφοράς δύο πιθανοτήτων P_1 και P_2 γίνεται ως εξής:

1.**Υπολογίζουμε** τις P_1 και P_2

$$P_1 = \frac{63}{90} = 0{,}70 \ , \qquad P_2 = \frac{85}{118} = 0{,}72$$

2.Υπολογίζουμε τη μέση πιθανότητα P

$$P = \frac{63+85}{90+118} = 0{,}71$$

3.Υπολογίζουμε την παράμετρο Z_π:

$$Z_\pi = \frac{P_1 - P_2}{\sqrt{P(1-P)(\frac{1}{n_1}+\frac{1}{n_2})}} = \frac{0,70-0,72}{\sqrt{0,71.0,29(\frac{1}{90}+\frac{1}{118})}} = \frac{-0,02}{0,0635} = -0,315$$

4.Θέτουμε

H_o: $P_1=P_2$, H_1: $P_1 \neq P_2$

5.Συγκρίνουμε την $|Z_\pi| = 0,315$ με $Z_{\alpha/2} = Z_{0,025} = 1,96$

$|Z_\Pi| < Z_{\alpha/2}$ συνεπώς δεχόμαστε την H_0 δηλαδή η πιθανότητα επιτυχίας και στις δυο περιόδους είναι η ίδια .

ΑΣΚΗΣΕΙΣ

9.1.Για να ελέγξουμε την αποτελεσματικότητα 2 σιτηρεσίων Α και Β στη γαλακτοπαραγωγή των προβατίνων λαμβάνουμε 5 προβατίνες και εφαρμόζουμε το σιτηρέσιο Α και σε 5 άλλες προβατίνες εφαρμόζουμε το σιτηρέσιο Β. Η ημερήσια γαλακτοπαραγωγή σε λίτρα φαίνεται στον επόμενο πίνακα :

Σιτηρέσιο Α	9	8	9	7	7
Σιτηρέσιο Β	10	9	8	10	8

Να ελεγχθεί σε επίπεδο σημαντικότητας 5% αν το σιτηρέσιο Β είναι καλύτερο από το Α.

9.2.Για την αξιολόγηση της γεύσης ενός γιαουρτιού με 0% λιποπεριεκτικότητα μιας εταιρείας χρησιμοποιήθηκαν 6 καταναλωτές με κλίμακα βαθμολόγησης 1-10. Άλλοι 6 καταναλωτές βαθμολόγησαν τη γεύση του παλιού γιαουρτιού της εταιρείας με 0% λιποπεριεκτικότητα. Τα αποτελέσματα ήταν τα εξής:

ΝΕΟ	8	7	8	7	6	6
ΠΑΛΙΟ	7	5	7	6	4	7

α) Να εξεταστεί σε επίπεδο σημαντικότητας 5% αν η γεύση του νέου γιαουρτιού βελτιώθηκε

β) Να βρεθεί το 95% διάστημα για τη μέση βαθμολογία του νέου γιαουρτιού.

9.3. Το ποσοστό του ιδανικού βάρους υπολογίστηκε για 18 τυχαία επιλεγμένους διαβητικούς με εξάρτηση ινσουλίνης. Τα αποτελέσματα παρατίθενται κατωτέρω:

107 119 99 114 120 104 88 114 124

116 101 121 152 100 125 114 95 117 (%)

α. Να υπολογίσετε την διάμεσο, την επικρατούσα τιμή και τον αριθμητικό μέσο.

β. Η δειγματική τυπική απόκλιση έχει υπολογιστεί και είναι s=14,42. Να υπολογίσετε το εύρος και τον συντελεστή μεταβλητότητας.

γ. Θέλετε να ελέγξετε εάν οι διαβητικοί που εξαρτώνται από την ινσουλίνη έχουν **υψηλότερο** μέσο ιδανικό βάρος από 100%. Να χρησιμοποιήσετε το δείγμα για να διεξαχθεί ο κατάλληλος στατιστικός έλεγχος. Να παρουσιάσετε μηδενική και εναλλακτική υπόθεση, στατιστική συνάρτηση ελέγχου, υποθέσεις που πρέπει να πληρούνται, παρατηρούμενο επίπεδο σημαντικότητας (p-value) και συμπέρασμα.

9.4. Υποθέστε ότι επιθυμείτε να συγκρίνετε τα χαρακτηριστικά της φυματικής μηνιγγίτιδας σε ασθενείς μολυσμένους με τον ιό της ανοσοανεπάρκειας του ανθρώπου (HIV) και σε ασθενείς που δεν έχουν μόλυνση. Ειδικότερα, θα θέλατε να καθορίσετε εάν οι δύο πληθυσμοί έχουν την ίδια μέση ηλικία. Ένα δείγμα 37 μολυσμένων ασθενών έχει μέση ηλικία $\bar{x}_1$ = 27,9 έτη και τυπική απόκλιση S_1 = 5,6 έτη. Ένα δείγμα 19 ασθενών που δεν έχουν μολυνθεί έχει μέση ηλικία $\bar{x}_2$ = 38,8 έτη και τυπική απόκλιση S_2 = 5,7 έτη. Να θεωρήσετε ότι οι_διασπορές στους πληθυσμούς είναι ίσες, και ότι ισχύει το κεντρικό οριακό θεώρημα.

α. Να ελέγξετε τη μηδενική υπόθεση ότι οι δύο πληθυσμοί ασθενών έχουν την ίδια μέση ηλικία στο επίπεδο σημαντικότητας α = 0,05.

β. Να κατασκευάσετε ένα 95% διάστημα εμπιστοσύνης για την αληθινή διαφορά στους μέσους των πληθυσμών.

9.5. Τα βάρη (σε γρ) 13 ποντικιών, τα οποία υπεβλήθησαν σε καθημερινές μεταγγίσεις τεχνητού αίματος επί ένα μήνα, βρέθηκαν μετά τις μεταγγίσεις ως εξής:
75, 78, 95, 71, 80, 78, 69, 77, 77, 80, 65, 90, 79.
(α) Υπολογίστε το δειγματικό μέσο και τη δειγματική διασπορά του βάρους των ποντικιών,
(β) Υπολογίστε την δειγματική διάμεσο.
(γ) Στα συνήθη ποντίκια (που δεν υποβάλλονται σε μεταγγίσεις), το μέσο βάρος είναι $\mu_0 = 90$ γρ. Υποθέτοντας ότι τα βάρη των συγκεκριμένων ποντικιών (του δείγματος) προέρχονται από κάποια κανονική κατανομή Ν (μ, σ2) με μ, σ2 άγνωστα, μπορούμε να συμπεράνουμε σε επίπεδο σημαντικότητας (ε.σ.) α = 5% ότι οι μεταγγίσεις προκαλούν ελάττωση βάρους;

9.6 Οι τιμές της αμόλυβδης βενζίνης (σε λεπτά του Ευρώ ανά λίτρο) σε τυχαίο δείγμα $X_1, X_2....X_{10}$ μεγέθους 10 από την Αθήνα βρέθηκαν
71, 67, 75, 71, 70, 69, 70, 70, 71, 71.
Ομοίως, οι τιμές ενός δεύτερου δείγματος $Y_1, Y_2...Y_{10}$ μεγέθους 10 από την Θεσσαλονίκη ήταν
70, 73, 75, 74, 75, 77, 75, 73, 67, 75.

Υποθέτοντας ότι οι τιμές προέρχονται από κανονικές κατανομές $N\left(\mu_1,\sigma^2\right)$ και $N\left(\mu_2,\sigma^2\right)$, για Αθήνα και Θεσσαλονίκη αντίστοιχα, με κοινή (άγνωστη) διασπορά σ^2,
(α) κατασκευάστε 95% διάστημα εμπιστοσύνης για τη μέση τιμή πώλησης της αμόλυβδης βενζίνης σε Αθήνα και Θεσσαλονίκη,
(β) κατασκευάστε 95% διάστημα εμπιστοσύνης για τη διαφορά των μέσων τιμών πώλησης σε Θεσσαλονίκη και Αθήνα, και (γ) να ελεγχθεί η υπόθεση
$H_0 : \mu_1 > \mu_2$ έναντι της $H_1 : \mu_1 < \mu_2$
σε επίπεδο σημαντικότητας α = 10%, και να αναπτυχθεί το συμπέρασμα που προκύπτει.

9.7 Ο ιός της ανεμοβλογιάς σε δυο μεγάλες πόλεις Α και Β συναντάται σε (άγνωστα) ποσοστά Pα, Pβ, αντίστοιχα. Σε τυχαία δείγματα ηα = 100 και ub = 200 ατόμων από τις δυο πόλεις, βρέθηκαν 20 και 30 ασθενείς, αντίστοιχα. Από τα παραπάνω δεδομένα, έχουμε σημαντικά στοιχεία ώστε να αποφανθούμε ότι στην πόλη Α είναι πιο διαδεδομένη η ανεμοβλογιά από ότι στην πόλη Β; (χρησιμοποιείστε ε.σ. α = 5%).

9.8 Οι παρακάτω τιμές δίνουν τον αριθμό των τσιγάρων που καπνίζει ημερησίως τυχαίο
δείγμα ν = 82 καπνιστών.

Πλήθ. Τσιγά.	10-12	13-15	16-18	19-21	22-24	25-27	28-30
Συχνότ	2	9	18	24	18	9	2

(α) Να υπολογιστεί ο δειγματικός μέσος και η δειγματική τυπική απόκλιση,
(β) Να υπολογιστεί προσεγγιστικά ένα 95% διάστημα εμπιστοσύνης για τον μέσο αριθμό τσιγάρων μ που καπνίζουν ημερησίως οι καπνιστές.
(γ) Καπνοβιομηχανία ισχυρίζεται ότι ο μέσος αριθμός τσιγάρων, μ, που καπνίζουν ημερησίως οι καπνιστές, δεν υπερβαίνει τον αριθμό 19. Τι συμπεραίνετε σε επίπεδο σημαντικότητας α = 5% και τι σε α = 1%, σχετικά με τον ισχυρισμό της καπνοβιομηχανίας;

9.9. Σε δύο τυχαία δείγματα μεγέθους $n_1 = 110$ και $n_2 = 90$ από τυχαία επιλεγμένους μισθωτούς, κατοίκους της Ελλάδας (X) και της Πορτογαλίας (Y), προέκυψαν οι παρακάτω τιμές:

Μισθός	Συχνότητα (X)	Συχνότητα(Y)
600 - 800	20	10
800- 1000	30	20
1000-1200	40	20
1200-1400	10	20
1400-1600	10	20
Σύνολο	$n_1 = 110$	$n_2 = 90$

(α) Να υπολογίσετε τους δειγματικούς μέσους X, Y, και τις δειγματικές διασπορές Sf και S% των παραπάνω (ομαδοποιημένων) δειγμάτων, και να εξηγήσετε ποιες άγνωστες ποσότητες εκτιμούν.
(β) Να κατασκευάσετε ένα 90% διάστημα εμπιστοσύνης για τη διαφορά (μι – μι) των μέσω μισθών σε Ελλάδα και Πορτογαλία. Επίσης να εξηγήσετε τη σημασία του διαστήματος που βρήκατε.
(γ) Να ελέγξετε σε επίπεδο σημαντικότητας α = 10% τη μηδενική υπόθεση ότι οι μέσοι μισθοί είναι ίσοι στις δύο χώρες, και να αναλύσετε το συμπέρασμα που προκύπτει από τον έλεγχο.

9.10 Τα παρακάτω δεδομένα δίνουν το ποσοστό επιμήκυνσης (κάτω από ορισμένο βάρος) για δύο δείγματα από κομμάτια ενός συγκεκριμένου τύπου νήματος όπου το πρώτο δείγμα ελήφθη πριν πλυθεί το νήμα και το δεύτερο αφού το νήμα πλύθηκε 6 φορές.

Χωρίς πλύση:	13.3	14.7	11.4	12.4	15.9	13.6	
Μετά 6 πλύσεις:	16.7	11.3	13.6	15.5	13.6	14.8	12.9

Με αυτά τα δεδομένα και σε επίπεδο σημαντικότητας α = 0.05, μπορούμε να συμπεραίνουμε ότι το πλύσιμο επηρεάζει το ποσοστό επιμήκυνσης του συγκεκριμένου νήματος.
(Μπορείτε να υποθέσετε ότι τα ποσοστά επιμήκυνσης του συγκεκριμένου νήματος πριν και μετά την πλύση, ακολουθούν κανονικές κατανομές με ίσες διασπορές.)

9.11 Σε μια απογραφή που έγινε πριν από πέντε χρόνια βρέθηκε ότι το 10% των προβάτων της χώρας πάσχουν από κάποια ασθένεια. Για να ελεγχθεί αν το ποσοστό άλλαξε από τότε , πάρθηκε δείγμα 500 προβάτων και σε 47 από αυτά παρατηρήθηκε η συγκεκριμένη ασθένεια .
Διαφέρει το ποσοστό των άρρωστων ζώων σήμερα από αυτό που βρέθηκε στην απογραφή πριν από πέντε χρόνια ;($\alpha = 0.05$)

9.12 Δύο χημικά διαλύματα Α και Β εξετάστηκαν για το pH τους. Ανάλυση δειγμάτων του Α έδωσε μέσο pH 7.52 με τυπική απόκλιση 0.024. Ανάλυση δειγμάτων του Β έδωσε μέσο pH 7.49 με τυπική απόκλιση 0.032. Σε επίπεδο σημαντικότητας 0.05 εξετάστε εάν τα δύο διαλύματα έχουν διαφορετικό pH.

9.13 Τα όρια αντοχής ενός τύπου καλωδίων έχει μέση τιμή 1800 κιλά και τυπική απόκλιση 100 κιλά .Η εταιρία που φτιάχνει τα καλώδια ισχυρίζεται ότι μια βελτίωση στη μέθοδο κατασκευής αύξησε το όριο αντοχής. Για να το επαληθεύσουμε δοκιμάζουμε 50 νέα καλώδια . Εάν το μέσο όριο αντοχής τους βρέθηκε 1850 κιλά , είναι σωστός ο ισχυρισμός της εταιρίας σε επίπεδο σημαντικότητας 0.01 ;

9.14 Τέσσερα υγιή άτομα θερμομετρήθηκαν το πρωί και το βράδυ της ίδιας ημέρας . Οι τιμές της θερμοκρασίας $\left({}^{0}C\right)$ ήταν

θερμοκρασία		
Άτομο	Πρωινή	βραδινή
1ο	36.1	36.4
2ο	36.4	36.8
3ο	36.0	36.6
4ο	36.5	36.8

Υπάρχει διαφορά μεταξύ της πρωινής και βραδυνής θερμοκρασίας ;

9.15 Μια εταιρεία παραγωγής γάλακτος, προκειμένου να διαλέξει ανάμεσα σε δύο μάρκες Α και Β για την αντικατάσταση κάποιων μηχανημάτων της, έκανε τον εξής έλεγχο. Πήρε ένα δείγμα από 10 εργάτες, οι οποίοι την πρώτη μέρα χειρίσθηκαν τα μηχανήματα της μάρκας Α και τη δεύτερη μέρα τα μηχανήματα της μάρκας Β, και κατέγραψε τις αποδόσεις των μηχανημάτων. Τα αποτελέσματα ήταν ως εξής:

Εργάτης	1	2	3	4	5	6	7	8	9	10
Μηχάνημα τύπου Α	80	92	102	76	105	98	110	84	92	91
Μηχάνημα τύπου Β	86	94	104	78	103	97	115	91	98	94

α) Ελέγξτε εάν οι αποδόσεις των μηχανημάτων της μάρκας Β είναι μεγαλύτερες από τις αποδόσεις των μηχανημάτων της μάρκας Α σε επίπεδο σημαντικότητας 5%. Να διατυπώσετε με ακρίβεια τις υποθέσεις που ελέγχετε .
β) Δώστε ένα 98% διάστημα εμπιστοσύνης για τη μέση απόδοση μηχανημάτων τύπου Α.

9.16 Από τα στοιχεία ενός εργαστηρίου φαίνονται τα ακόλουθα: 60 transistors τύπου Α σε ένα δείγμα χιλίων transistors βρέθηκαν ελαττωματικά και 40 transistors τύπου Β σε ένα άλλο δείγμα επίσης χιλίων transistors βρέθηκαν ελαττωματικά. Τα παραπάνω δεδομένα δίνουν αρκετές ενδείξεις που να φαίνεται ότι υπάρχει διαφορά στην ποιότητα μεταξύ των transistors τύπου Α και τύπου Β ($\alpha=0.05$);

9.17 Για να μελετηθεί η μόλυνση της ατμόσφαιρας σε μία βιομηχανική περιοχή, μετρήθηκε ο αριθμός των αιωρούμενων σωματιδίων (σκόνης) σε 1lt αέρα σε 14 διαφορετικά σημεία της περιοχής. Οι 14 αυτές μετρήσεις αυτές είχαν μέσο όρο ίσο με 102.6 με τυπική απόκλιση 27. Έγιναν επίσης παρόμοιες μετρήσεις σε 15 σημεία μιας αγροτικής περιοχής, οι οποίες έδωσαν μέσο όρο 47 με τυπική απόκλιση 17.
(α) Σε επίπεδο σημαντικότητας 1%, μπούμε να ισχυριστούμε ότι ο αριθμός σωματιδίων ανά λίτρο αέρα παρουσιάζεται αυξημένος στην βιομηχανική περιοχή έναντι της αγροτικής;
(β) Ας υποθέσουμε ότι το όριο ασφαλείας για την βιομηχανική περιοχή είναι 100 σωματίδια ανά λίτρο αέρα. Με αυτές τις μετρήσεις, μπορούμε να ισχυριστούμε ότι η μόλυνση στη βιομηχανική περιοχή ξεπερνά τα όρια ασφαλείας; ($\alpha = 5\%$).
(γ) Δώστε 98% διάστημα εμπιστοσύνης για το μέσο αριθμό σωματιδίων ανά λίτρο αέρα για την αγροτική περιοχή.

Κεφάλαιο 10

Δοκιμασία ή έλεγχος X^2

10.1 Εισαγωγή

Όταν οι παρατηρήσεις έχουν ποιοτικό χαρακτήρα, ο μόνος τρόπος για να συγκεντρώσει κανείς πληροφορίες είναι να χωρίσει τις παρατηρήσεις σε **κατηγορίες**. Οπότε τελικά τα δεδομένα θα είναι οι μετρήσεις συχνοτήτων σε κάθε κατηγορία. Στις περιπτώσεις αυτές ο στατιστικός έλεγχος γίνεται με σύγκριση των συχνοτήτων των παρατηρήσεων στις διάφορες ταξινομητικές κατηγορίες, και έτσι τα κριτήρια t-test και Z δεν είναι εφαρμόσιμα.
Ο στατιστικός έλεγχος αυτός γίνεται με την στατιστική δοκιμασία (έλεγχος) $\mathbf{X^2}$.

10.2 Η δοκιμασία X^2 σαν test προσαρμογής (ή καλής εφαρμογής)

Όταν οι παρατηρήσεις αναφέρονται σε μία μόνο κατανομή το πρόβλημα που προκύπτει είναι αν η κατανομή αυτή είναι συμβατή (εφαρμόζει καλά ή προσαρμόζεται καλά) με μία ορισμένη θεωρητική, ή με την κατανομή η οποία θα αναμενόταν με βάση μιας ορισμένης θεωρίας ή υπόθεσης. Το κριτήριο (δοκιμασία) X^2 στις περιπτώσεις αυτές αποτελεί δοκιμασία που ελέγχει την προσαρμογή ή καλή εφαρμογή των δεδομένων σε μία θεωρητικά αναμενόμενη κατανομή.
Το κριτήριο ελέγχου που μετρά την συνολική διαφορά ανάμεσα στα παρατηρούμενα και στα θεωρητικά μεγέθη είναι:

$$X^2 = \sum_{i=1}^{\kappa} \frac{(\Pi_i - \theta_i)^2}{\theta_i} = \sum \frac{(O-E)^2}{E} \qquad (10.1)$$

όπου $\Pi_i = O =$ τα παρατηρούμενα μεγέθη (πειραματικά δεδομένα)
$\Theta_i = E =$ τα θεωρητικά ή αναμενόμενα θεωρητικά μεγέθη.
Η θεωρητική ή αναμενόμενη τιμή υπολογίζεται από τον τύπο

$$\theta_i = E = n.P_i \qquad (10.2)$$

όπου n = το πλήθος των παρατηρήσεων
P_i= η πιθανότητα που αντιστοιχεί στην κατηγορία.
Το test που εφαρμόζουμε συνοψίζεται στα εξής βήματα:
1) Ορίζουμε την H_o: δηλαδή δεχόμαστε ότι τα πειραματικά δεδομένα προσαρμόζονται με τα θεωρητικά πχ. $P_1 = P_{1,0}, P_2 = P_{2,0}, \ \ldots, \ P_k = P_{k,0}$ όπου $P_1, P_2 \ldots P_\kappa$ είναι τα παρατηρούμενα και P_1,ο , P_2,ο … P_κ,ο είναι τα θεωρητικά αναμενόμενα.

2)Υπολογίζουμε τα θεωρητικά μεγέθη.
3)Υπολογίζουμε κριτήριο ελέγχου X^2 από την σχέση

$$X^2 = \sum_{i=1}^{\kappa} \frac{(\Pi i - \theta i)^2}{\theta i}$$

4)Υπολογίζουμε τους βαθμούς ελευθερίας από την σχέση: β. ε =Κ-1 ,όπου Κ είναι ο αριθμός των κελιών (κατηγοριών).
5) Συγκρίνουμε την τιμή X^2 που υπολογίσαμε από τα πειραματικά δεδομένα με την τιμή $X^2_{\kappa-1,\alpha}$ που βρίσκεται από τους στατιστικούς πίνακες .
6)Απορρίπτουμε την Ho όταν $X^2 \geq X^2_{\kappa-1,\alpha}$.
Παρατήρηση: Αξίζει να σημειωθεί ότι ο αριθμός των μετρήσεων n πρέπει να είναι μεγάλος και όλα τα θεωρητικά υπολογιζόμενα μεγέθη πρέπει να είναι μεγαλύτερα ή ίσα με το 5 ($\theta_i \geq 5$), ώστε να εφαρμοσθεί το κριτήριο X^2. Σε περίπτωση που $\theta_i < 5$ κάνουμε σύμπηξη τάξεων.

Παράδειγμα (10.1)
Η θεωρία του Mendel λέει ότι τα μπιζέλια των τύπων Α, Β, Γ, Δ εμφανίζονται στις αναλογίες 9:3:3:1. Αν σε 100 μπιζέλια βρέθηκαν 56 τύπου Α , 19 τύπου Β,17 τύπου Γ και 8 τύπου Δ μπορούμε να δεχθούμε ότι συμφωνούν με τη θεωρία (α=0.05);
Λύση:
Οι παρατηρήσεις που δίνονται είναι ποιοτικές , αναφέρονται σε μία μόνο κατανομή και ζητείται αν η κατανομή αυτή είναι συμβατή (εφαρμόζει καλά ή προσαρμόζεται καλά) με μία ορισμένη θεωρητική. Ο στατιστικός έλεγχος γίνεται με την δοκιμασία X^2.Το κριτήριο (δοκιμασία) X^2 στις περιπτώσεις αυτές αποτελεί δοκιμασία που ελέγχει την προσαρμογή ή καλή εφαρμογή των δεδομένων σε μία θεωρητικά αναμενόμενη κατανομή.
Το κριτήριο ελέγχου που μετρά την συνολική διαφορά ανάμεσα στα παρατηρούμενα και στα θεωρητικά μεγέθη είναι:

$$X^2 = \sum_{i=1}^{\kappa} \frac{(\Pi i - \theta i)^2}{\theta i} = \sum \frac{(O-E)^2}{E}$$

Ορίζουμε :
H_o: δεχόμαστε ότι τα πειραματικά δεδομένα προσαρμόζονται με τα θεωρητικά
H_1: δεχόμαστε ότι τα πειραματικά δεδομένα δεν προσαρμόζονται με τα θεωρητικά

Υπολογίζουμε τα θεωρητικά μεγέθη:

$\Theta_1 = \frac{9}{16}\times 100 = 56.25$,$\theta_2 = \frac{3}{16}\times 100 = 18.75$,$\theta_3 = \theta_2 = 18.75$

,$\theta_4 = \frac{1}{16}\times 100 = 6.25$

Υπολογίζουμε την ποσότητα X^2 από τον τύπο:

$$X^2 = \sum_{i=1}^{\kappa} \frac{(\Pi i - \theta i)^2}{\theta i} = \frac{(56-56.25)^2}{56.25} + \frac{(19-18.75)^2}{18.75} + \frac{(17-18.75)^2}{18.75} +$$

$$+ \frac{(8-6.25)^2}{6.75} = 0.001+0.003+0.163+0.454=0.620$$

Συγκρίνουμε την τιμή X^2 που υπολογίσαμε από τα πειραματικά δεδομένα με την τιμή $X^2_{\kappa-1,\alpha}$ που βρίσκεται από τους στατιστικούς πίνακες(3)

Οι βαθμοί ελευθερίας β.ε =K-1= 4-1= 3

Για 3 βαθμούς ελευθερίας η θεωρητική τιμή $X^2_{3,0.05}=7.82$

$X^2 = 0.620 \langle X^2_{3,0.05} = 7.82$

Συμπέρασμα : δεχόμαστε την H_0, δηλαδή δεχόμαστε ότι τα πειραματικά δεδομένα προσαρμόζονται με την θεωρία του Mendel .

Παράδειγμα (10.2)

Σε ένα εργοστάσιο παραγωγής ορισμένου είδους υφάσματος παίρνουμε σε κανονικά χρονικά διαστήματα δείγματα μήκους 10 μέτρων και βρίσκουμε τον αριθμό των ελαττωμάτων στα δείγματα αυτά. Σε 400 τέτοια δείγματα ο αριθμός των ελαττωμάτων και οι αντίστοιχες συχνότητες αυτών δίνονται στον ακόλουθο πίνακα:

Πλήθος ελαττωμάτων	0	1	2	3	4	5	6	7
Πλήθος δειγμάτων.	89	143	94	42	20	8	3	1

Να εξετάσετε αν η κατανομή αυτή ακολουθεί την κατανομή Poisson σε επίπεδο σημαντικότητας α = 0.05. ($e^{-1.5}=0.223$).

Λύση:

Για τα δεδομένα της άσκησης η μηδενική υπόθεση είναι :

H_0 : δεχόμαστε ότι τα δεδομένα ακολουθούν την κατανομή poisson.

H_1: τα δεδομένα δεν ακολουθούν την κατανομή poisson.

Άρα ο έλεγχος θα γίνει με X^2 σαν test προσαρμογής :

$$\mathbf{X^2}= \sum_{i=1}^{\kappa}\frac{(\Pi i-\theta i)^2}{\theta i}=\sum\frac{(O-E)^2}{E}$$

όπου Π_i = O = τα παρατηρούμενα μεγέθη

Θ_i = E= τα θεωρητικά ή αναμενόμενα θεωρητικά μεγέθη. Η θεωρητική ή αναμενόμενη τιμή υπολογίζεται από τον τύπο

$\theta_i = E = nP_i$, όπου n=το πλήθος των παρατηρήσεων .

Χρειαζόμαστε την παράμετρο λ της κατανομής Poisson η οποία είναι άγνωστη, για να υπολογίσουμε τα P . Η παράμετρος λ είναι η μέση τιμή της κατανομής και υπολογίζεται από την σχέση :

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

$$\bar{\mathrm{X}}=\frac{0\times 89+1\times 143+2\times 94+3\times 42+4\times 20+5\times 8+6\times 3+7\times 1}{400}=1.5$$

Οι πιθανότητες p_i για κάθε κατηγορία υπολογίζονται με βάση τον τύπο :

$$P\left(X=x\right)=e^{-\lambda}.\frac{\lambda^x}{x!}$$

$P_{(0)}=e^{-1.5}\frac{1.5^0}{0!}$=0.2231 , $P_{(1)}= e^{-1.5}\frac{1.5^1}{1!}=0.335$,

$P_{(2)}$=0.251 , $P_{(3)}$=0.1255 , $P_{(4)}$=0.0471

, $P_{(5)}$=0.0141 , $P_{(6)}$=0.00353, $P_{(7)}$=0.00076

Υπολογίζουμε τα θεωρητικά μεγέθη $\theta_i = E = nP_i$, και τα τοποθετούμε σε πίνακα με τα πειραματικά δεδομένα .

$\Theta_0=0.2231\times400=89.3$, $\theta_1=0.335\times400=134$,

$\theta_2=0251\times400=100.4$, $\theta_3=0.1255\times400=50.2$

$\theta_4=0.0471\times400=18.8$, $\theta_5= 0.0141\times400=5.6$,

$\theta_6=0.00353\times400=1.4$, $\theta_7=0.00076\times400=0.3$

Κατηγορία	Π_i	Θi	$\frac{(\pi_i - \theta_i)^2}{\theta_i}$
0	89	89.3	0.001
1	143	134	0.604
2	94	100.4	0.408
3	42	50.2	1.339
4	20	18.8	0.077
5	8	5.6	1.029
6	3	1.4	1.829
7	1	0.3	1.633
Σύνολο	400	400	X^2=6.92

Παρατηρούμε ότι τα θεωρητικά αλλά και τα πειραματικά μεγέθη στις τελευταίες δυο κατηγορίες (6 και 7) είναι μικρότερες του πέντε (5) γι' αυτό συμπτύσσουμε τις τρεις τελευταίες τάξεις σε μία με Π_5 =12 και θ_5 =7.3 και συνεπώς $\frac{(\pi_i - \theta_i)^2}{\theta_i} = 3.026$

Η τιμή του X^2 διαμορφώνεται σε 5.455 αντί 6.92 . Οι βαθμοί ελευθερίας = K-1-m όπου K το πλήθος των κατηγοριών και m το πλήθος των παραμέτρων που εκτιμήσαμε από το δείγμα. Δηλαδή β.ε =6-1-1= 4

Συγκρίνουμε την τιμή του X^2 των πειραματικών δεομένων με την θεωρητική .

$$X^2 = 5.455 < X^2_{4,0.05} = 9.488$$

Συμπέρασμα : δεχόμαστε την H_0 δηλαδή τα δεδομένα ακολουθούν την κατανομή poisson.

10.3.Η δοκιμασία X^2 σαν test ανεξαρτησίας ή αλλιώς ως κριτήριο συσχέτισης ποιοτικών χαρακτηριστικών.

Όταν θέλουμε να ελέγξουμε αν τα χαρακτηριστικά Α και Β είναι εξαρτημένα ή ανεξάρτητα ή με άλλα λόγια αν συσχετίζονται μεταξύ τους, εφαρμόζουμε το X^2 -test δηλαδή **βρίσκουμε** την τιμή X^2 για όλα τα πειραματικά δεδομένα από τον τύπο:

$$X^2 = \Sigma \frac{(\Pi i - \theta i)^2}{\theta i} = \Sigma \frac{(O-E)^2}{E} \qquad (10.1)$$

Όπου θi=E= αναμενόμενες θεωρητικά παρατηρήσεις οι οποίες βρίσκονται από την σχέση :

$$E = \theta_i = \frac{(\acute{\alpha}\theta\rho o\iota\sigma\mu\alpha\; \tau\eta\varsigma\; i\, \gamma\rho\alpha\mu\mu\acute{\eta}\varsigma)\times(\acute{\alpha}\theta\rho o\iota\sigma\mu\alpha\; \tau\eta\varsigma\; j\, \sigma\tau\acute{\eta}\lambda\eta\varsigma)}{\gamma\varepsilon\nu\iota\kappa\acute{o} \quad \acute{\alpha}\theta\rho o\iota\sigma\mu\alpha} \qquad (10.3) \quad \text{ή}$$

$$\mathrm{E} = \Theta_i = \frac{\kappa\acute{\alpha}\theta\varepsilon\tau o \text{ σύνολο} \times \text{οριζόντιο σύνολο}}{\Gamma\varepsilon\nu\iota\kappa\acute{o} \text{ σύνολο}} \qquad (10.3)$$

Οι θεωρητικές τιμές τοποθετούνται σε παρένθεση δίπλα στις αντίστοιχες τιμές, και πρέπει όλες να είναι μεγαλύτερες ή τουλάχιστον ίσες με 5.

Οι βαθμοί ελευθερίας υπολογίζονται από την σχέση:
β.ε = (K-1) .(L-1) = (K-1).(S-1) . (10.4)
όπου K= κάθετες στήλες , S = L = οριζόντιες σειρές. Η στήλη και η γραμμή των συνόλων δεν υπολογίζονται.
Θέτουμε :
Ho: τα χαρακτηριστικά Α και Β είναι ανεξάρτητα δηλαδή δεν συσχετίζονται (δεν υπάρχει σχέση μεταξύ τους) και ,
H_1: τα χαρακτηριστικά Α και Β είναι εξαρτημένα δηλαδή συσχετίζονται μεταξύ τους.
Απορρίπτουμε την Ho μόνο εάν $X^2 > X^2_{(s-1)(\kappa-1),\alpha}$.

Η τιμή του κριτηρίου $X^2_{(s-1)(\kappa-1),\alpha}$ βρίσκεται από θεωρητικούς πίνακες που βρίσκονται συνήθως στο τέλος των στατιστικών βιβλίων.

Παράδειγμα (10.3)
Σ’ ένα δείγμα 100 ατόμων, που έπασχαν από καρκίνο του πνεύμονα, η κατανομή των ομάδων αίματος ήταν: O = 45, A = 45, B =8, AB =2. Σε ένα

άλλο δείγμα 900 «υγιών» ατόμων (ίδιας φυλής) η κατανομή των ομάδων αίματος ήταν: Ο =455, Α= 355, Β=77, ΑΒ=13.
Υπάρχει σχέση μεταξύ ομάδας αίματος και καρκίνο του πνεύμονα;
Λύση:
Οι κατηγορίες ομάδα αίματος, καρκίνος κλπ είναι ποιοτικές, και ζητείται να ελέγξουμε αν συσχετίζονται μεταξύ τους. Άρα εφαρμόζουμε την δοκιμασία X^2 ως κριτήριο συσχέτισης ποιοτικών χαρακτηριστικών.
Θέτουμε :
Ho: δεν υπάρχει σχέση μεταξύ ομάδας αίματος και καρκίνο του πνεύμονα .
H_1: υπάρχει σχέση μεταξύ ομάδας αίματος και καρκίνο του πνεύμονα .
Τοποθετούμε τις κατηγορίες με τις αντίστοιχες συχνότητες σε πίνακα δεδομένων.

	O	A	B	AB	Σύνολο
Με καρκίνο	45 (50)	45 (40)	8 (8.5)	2 (1.5)	100
Υγιής	455 (450)	355(360)	77(76.5)	13(13.5)	900
Σύνολο	500	400	85	15	1000

Υπολογίζουμε τα θεωρητικά αναμενόμενα από τον τύπο

$$E = \Theta_i = \frac{(\textit{κάθετο}\ \text{σύνολο}) \cdot (\text{οριζόντιο σύνολο})}{\textit{Γενικό}\ \text{σύνολο}}$$

και τα τοποθετούμε σε ξεχωριστώ πίνακα ή στον ίδιο πίνακα σε παρένθεση.
Υπολογίζουμε την πειραματική τιμή του $\mathbf{X^2}$

$$X^2 = \Sigma\ \frac{(\Pi i - \theta i)^2}{\theta i} = \Sigma \frac{(O-E)^2}{E} = \frac{(45-50)^2}{50} + \frac{(45-40)^2}{40} +$$

$$+\frac{(8-8.5)^2}{8.5} + \frac{(2-1.5)^2}{1.5} + \frac{(455-450)^2}{450} + \frac{(355-360)^2}{360} + \quad + \frac{(77-76.5)^2}{76.5}$$

$$+\frac{(13-13.5)^2}{13.5} = 1.467$$

Οι βαθμοί ελευθερίας υπολογίζονται από την σχέση:
β.ε =(K-1)(L-1)=(S-1).(K-1) = (4-1).(2-1) = 3 .
Συγκρίνουμε την τιμή X^2 που υπολογίσαμε από τα πειραματικά δεδομένα με την τιμές του X^2 που βρίσκονται από τους στατιστικούς πίνακες :
$X^2_{3,0.05}$ =7.81, $X^2_{3,0.01}$=11.35 , $X^2_{3,0.001}$=16.27.

$X^2=1.467 < X^2_{3,0.05}=7.81$
Συμπέρασμα : Δεχόμαστε την Ho: δηλαδή δεν υπάρχει σχέση μεταξύ ομάδας αίματος και καρκίνο του πνεύμονα .

10.4. Η δοκιμασία X^2 σαν test ομογένειας ή κριτήριο ετερογένειας.

Όταν η δειγματοληψία περιλαμβάνει μια διαίρεση του πληθυσμού σε «υποπληθυσμούς» ανάλογα με τις κατηγορίες ενός χαρακτηριστικού, δηλαδή
όταν έχουμε έναν αρχικό πληθυσμό χωρισμένο σε υποπληθυσμούς, και από κάθε υποπληθυσμό σχηματίζουμε ένα τυχαίο δείγμα που το χωρίζουμε σε κατηγορίες ως προς ένα χαρακτηριστικό. Ο πίνακας δεδομένων που σχηματίζεται έχει στήλες μεγαλύτερες ή ίσες με 2 και αριθμό σειρών επίσης μεγαλύτερο ή ίσο με 2.
Η τιμή του x^2 βρίσκεται από τον τύπο :

$$X^2 = \Sigma\ \frac{(\Pi i - \theta i)^2}{\theta i} = \Sigma \frac{(O-E)^2}{E} \qquad (10.1) \qquad \text{για όλα τα κελιά.}$$

Οι βαθμοί ελευθερίας βρίσκονται όπως πριν από τον τύπο :
Β.ε = (S-1)(K-1) = (L-1)(K-1)
Οι θεωρητικές τιμές βρίσκονται επίσης από την σχέση

$$E=\theta_i = \frac{\left(\acute{\alpha}\theta\rho o\iota\sigma\mu\alpha\; \tau\eta\varsigma\; i\, \gamma\rho\alpha\mu\mu\acute{\eta}\varsigma\right)\times\left(\acute{\alpha}\theta\rho o\iota\sigma\mu\alpha\; \tau\eta\varsigma\; j\, \sigma\tau\acute{\eta}\lambda\eta\varsigma\right)}{\gamma\varepsilon\nu\iota\kappa\acute{o}\;\; \acute{\alpha}\theta\rho o\iota\sigma\mu\alpha} \qquad (10.3)$$

ή $$E=\theta_i = \frac{(\kappa\acute{\alpha}\vartheta\varepsilon\tau o\ \text{σύνολο})\ \cdot\ (\text{οριζόντιο σύνολο})}{\Gamma\varepsilon\nu\iota\kappa\acute{o}\ \text{σύνολο}}$$

Απόφαση: Αν $X^2 > X^2_{(s-1)(\kappa-1),\alpha}$ τότε απορρίπτουμε την H_o.
H_o: Υπάρχει ομογένεια.
H_1: Υπάρχει ετερογένεια.

Παράδειγμα (10.4)
Για να διαπιστωθεί η ισχύς τριών εμβολίων παρωτίτιδας (Α,Β,Γ) ,εμβολιάσθηκαν σε 1 σχολείο 200 παιδιά. Ο αριθμός των παιδιών που δεν αρρώστησαν, αρρώστησαν ελαφρά , αρρώστησαν σοβαρά από παρωτίτιδα στα επόμενα 2 χρόνια ήταν :

	Α	Β	Γ
Δεν αρρώστησαν	30	15	15
Αρρώστησαν ελαφρά	40	50	10
Αρρώστησαν σοβαρά	10	25	5

Σε επίπεδο σημαντικότητας 5% να εξετάσετε αν τα εμβόλια Α,Β,Γ έχουν την ίδια αποτελεσματικότητα.

Λύση :

Όπως βλέπουμε οι παρατηρήσεις είναι ποιοτικές, οπότε ο στατιστικός έλεγχος γίνεται με δοκιμασία X^2 ως κριτήριο ετερογένειας .

Η τιμή του X^2 βρίσκεται από τον τύπο :

$X^2=\Sigma \dfrac{(\Pi i - \theta i)^2}{\theta i} = \Sigma \dfrac{(O-E)^2}{E}$ για όλα τα κελιά, όπου

$\Pi_i = O =$ τα παρατηρούμενα μεγέθη , $\Theta_i = E =$ τα θεωρητικά ή αναμενόμενα θεωρητικά μεγέθη.

Ορίζουμε H_o: δεχόμαστε ότι τα εμβόλια Α,Β,Γ έχουν την ίδια αποτελεσματικότητα.

H_1: τα εμβόλια Α,Β,Γ δεν έχουν την ίδια αποτελεσματικότητα.

Φτιάχνουμε δύο πίνακες ,ο ένας με τα πειραματικά δεδομένα

	Α	Β	Γ	Σύνολο
Δεν αρρώστησαν	30	15	15	60
Αρρώστησαν ελαφρά	40	50	10	100
Αρρώστησαν σοβαρά	10	25	5	40
Σύνολο	80	90	30	200

Και ο άλλος με τα θεωρητικά αναμενόμενα .Τα θεωρητικά αναμενόμενα υπολογίζονται από τον τύπο

$$E = \Theta_i = \frac{\textit{κάθετο}\ \text{σύνολο x οριζόντιο σύνολο}}{\Gamma\textit{ενικό}\ \text{σύνολο}}$$

	Α	Β	Γ	Σύνολο
Δεν αρρώστησαν	$\frac{80.60}{200}=24$	$\frac{90.60}{200}=27$	$\frac{30.60}{200}=9$	60
Αρρώστησαν ελαφρά	$\frac{80.100}{200}=40$	$\frac{90.100}{200}=45$	$\frac{30.100}{200}=15$	100
Αρρώστησαν σοβαρά	$\frac{80.40}{200}=16$	$\frac{90.40}{200}=18$	$\frac{30.40}{200}=6$	40
Σύνολο	80	90	30	200

Υπολογίζουμε την πειραματική τιμή του X^2 :

$$X^2=\Sigma\frac{(\Pi i-\theta i)^2}{\theta i}=\frac{(30-24)^2}{24}+\frac{(15-27)^2}{27}+\frac{(15-9)^2}{9}+\frac{(40-40)^2}{40}+$$
$$+\frac{(50-45)^2}{45}+\frac{(10-15)^2}{15}+\frac{(10-16)^2}{16}++\frac{(25-18)^2}{18}+\frac{(5-6)^2}{6}=$$
$=1.5$ +5,333 + 4+ 0+ 0,556+1,667 +2,25 +2,722 +0,167= 18,195

Υπολογίζουμε τους βαθμούς ελευθερίας από την σχέση :

Β.ε = (Κ-1)(L-1) = (S-1).(κ-1) = (3-1).(3-1) = 4

Συγκρίνουμε την τιμή X^2που υπολογίσαμε από τα πειραματικά δεδομένα με την τιμή $X^2_{(s-1)(k-1),a}$ που βρίσκεται από τους στατιστικούς πίνακες . $X^2_{(s-1)(k-1),a}=X^2_{2,2,0,05}=9{,}4877$

$X^2=18{,}195\rangle\ X^2_{2,2,0,05}=9{,}4877$

Συμπέρασμα : απορρίπτουμε την H_o και δεχόμαστε ότι τα εμβόλια Α,Β,Γ δεν έχουν την ίδια αποτελεσματικότητα .

10.5 Η δοκιμασία X^2σε δύο δυαδικές ταξινομήσεις. τετράπτυχος πίνακας

Όταν η κατανομή συχνοτήτων ποιοτικών παρατηρήσεων αναφέρεται σε δύο μόνο οριζόντιες σειρές και σε δύο μόνο κάθετες στήλες (K=2, L=2),τότε σχηματίζεται ένας τετράπτυχος πίνακας ο οποίος έχει 1 βαθμό ελευθερίας (K-1)(L-1) = (2-1)(2-1) = 1 , όπως φαίνεται στο παρακάτω πίνακα

	X	Y	Σύνολο
A	a	b	a+b
B	c	d	c+d
Σύνολο	a+c	b+d	a+b+c+d = n

Η τιμή του X^2 βρίσκεται και σε αυτή την περίπτωση από τον γενικό τύπο :

$$X^2=\Sigma\ \frac{(\Pi i-\theta i)^2}{\theta i}=\Sigma\frac{(O-E)^2}{E}$$

αλλά και από το ισοδύναμο και περισσότερο εύχρηστο με βάση του παραπάνω πίνακα τύπο :

$$X^2=\frac{(ad-bc)^2\cdot n}{(a+b)(c+d)(a+c)(b+d)} \qquad (10.4)$$

Κατά ορισμένους στατιστικολόγους η τιμή του X^2 στους τετράπτυχους πίνακες υπολογίζεται σωστότερα με την διόρθωση κατά Yates ,σύμφωνα με την οποία γίνεται μείωση της απόλυτης τιμής (ad-bc) κατά n/2 ,οπότε ο παραπάνω τύπος διαμορφώνεται ως εξής :

$$X^2=\frac{\left(\left|ad-bc\right|-\frac{n}{2}\right)^2\cdot n}{(a+b)(c+d)(a+c)(b+d)} \qquad (10.5)$$

Η διόρθωση αυτή έχει ως συνέπεια την μείωση της τιμής του X^2.

Παράδειγμα (10.5)

Για την θεραπεία του σακχαρώδη διαβήτη χορηγήθηκαν, μετά από τυχαιοποίηση για δύο έτη σε 200 ασθενείς το φάρμακο A, ενώ σε 160 το φάρμακο B. Απ'αυτούς που έπαιρναν το φάρμακο A, 25 πέθαναν στο διάστημα αυτό, ενώ στο ίδιο διάστημα πέθαναν 5 από εκείνους που έπαιρναν το B. Υπάρχει διαφορά στον αριθμό των θανάτων μεταξύ των δύο ομάδων ασθενών με διαφορετικό φάρμακο θεραπείας

Λύση :

Τα πειραματικά δεδομένα είναι ποιοτικά , και ο στατιστικός έλεγχος γίνεται με την δοκιμασία X^2 .

Τοποθετούμε τις κατηγορίες με τις αντίστοιχες συχνότητες σε πίνακα δεδομένων.

	Πέθαναν	Έζησαν	Σύνολο
Φάρμακο Α	25	175	200
Φάρμακο Β	5	155	160
Σύνολο	30	330	360

Όπως βλέπουμε έχουμε πίνακα 2 επί 2 (τετράπτυχος πίνακας) . Το X^2 μπορεί να θεωρηθεί ως στατιστικός έλεγχος της διαφοράς δύο αναλογιών (ποσοστών).
Η μηδενική υπόθεση H_o : $P_1 = P_2$ (δεν υπάρχει διαφορά στον αριθμό των θανάτων μεταξύ των δύο ομάδων).Ενώ η εναλλακτική :
H_1 $P_1 \neq P_2$ (υπάρχει διαφορά στον αριθμό των θανάτων μεταξύ των δύο ομάδων)
Η τιμή του X^2 υπολογίζεται με βάση του διορθωμένου κατά Yates τύπο :

$$X^2 = \frac{\left(\left|ad - bc\right| - \frac{n}{2}\right)^2 n}{(a+b)(c+d)(a+c)(b+d)} = \frac{\left(\left|25 \times 155 - 175 \times 5\right| - \frac{360}{2}\right)^2 \times 360}{200 \times 160 \times 30 \times 330} = 9.04$$

Για 1 βαθμό ελευθερίας η θεωρητικές τιμές είναι $X^2_{0,05}=3.84$, $X^2_{0,01}=6.64$,$X^2_{0,001}=10.83$.Συγκρίνουμε την πειραματική τιμή X^2 με τις θεωρητικές:

$$X^2_{0,01}=6.64 \;\langle\; X^2=9.04 \;\langle\; X^2_{0,001}=10.83$$

Συνεπώς απορρίπτουμε την H_o και δεχόμαστε ότι υπάρχει διαφορά στον αριθμό των θανάτων μεταξύ των δύο ομάδων ασθενών με διαφορετικό φάρμακο θεραπείας σε επίπεδο στατιστικά πολύ σημαντικό ($0.001 < P < 0.01$).

10.6 Η δοκιμασία X^2 για σύγκριση ποιοτικών παρατηρήσεων κατά ζεύγη

Όταν οι ποιοτικές παρατηρήσεις δυο συγκρινόμενων ομάδων εμφανίζουν ατομική κατά ζεύγη αντιστοιχία η δοκιμασία X^2 που εφαρμόζεται είναι η κατά ζεύγη με 1 βαθμό ελευθερίας , και είναι για τις ποιοτικές παρατηρήσεις το ισοδύναμο της δοκιμασίας κατά ζεύγη t –test σε ποσοτικές παρατηρήσεις .Οι ποιοτικές παρατηρήσεις των δυο συγκρινόμενων ομάδων τοποθετούνται σε πίνακα της μορφής :

Ναι στο Α, ναι στο Β	α
Όχι στο Α, όχι στο Β	β
Ναι στο Α, όχι στο Β	ε
Όχι στο Α, ναι στο Β	ζ
Σύνολο	α+β+ε+ζ

Το κριτήριο X^2 -κατά ζεύγη υπολογίζεται με βάση τον τύπο:

$$X_m^2 = \frac{(\varepsilon - \zeta)^2}{(\varepsilon + \zeta)} \qquad (10.6)$$

Το κατά ζεύγη X^2 είναι κατά κανόνα ισχυρότερο ,και δίνει αποτελέσματα περισσότερο σημαντικά απ' ότι το συνηθισμένο X^2 για τα ίδια δεδομένα. Η διόρθωση κατά Yates εφαρμόζεται επίσης στο κατά ζεύγη X^2 ,έτσι ο παραπάνω τύπος γίνεται :

$$X_m^2 = \frac{\left(\left|\varepsilon - \zeta\right| - 1\right)^2}{(\varepsilon + \zeta)} \qquad (10.7)$$

Παράδειγμα (10.6)
Εκατό φοιτητές εξετάστηκαν σε δύο μαθήματα Α και Β. Απ' αυτούς 20 πέτυχαν και στα δύο , 40 σε κανένα απ' αυτά , 10 μόνο στο Α και 30 μόνο στο Β. Να ελεγχθεί :
α) αν υπάρχει διαφορά στην πιθανότητα επιτυχίας στα δύο αυτά μαθήματα
β) αν άτομα, που πέτυχαν στο μάθημα Α, έχουν μεγαλύτερη πιθανότητα επιτυχίας στο μάθημα Β, σε σύγκριση με άτομα , που δεν πέτυχαν στο μάθημα Α.
Λύση:
Τα δεδομένα δείχνουν σύγκριση ανάμεσα σε δύο μαθήματα με ταυτόχρονη ατομική αντιστοιχία, οπότε ανάλογα με την διατύπωση της ερώτησης, οι παρατηρήσεις τοποθετούνται σε κατάλληλο πίνακα και εφαρμόζεται ο κατάλληλος τύπος του χ^2.
α) Η διαφορά στην πιθανότητα επιτυχίας στα δύο μαθήματα απαιτεί δοκιμασία X^2 κατά ζεύγη, οπότε ο πίνακας των δεδομένων είναι:

Πέτυχαν και στα δύο μαθήματα	20
Δεν πέτυχαν στο Α ούτε στο Β	40
Πέτυχαν στο Α, όχι στο Β	10 (ε)
Απέτυχαν στο Α, πέτυχαν στο Β	30(ζ)
Σύνολο	100

Θέτουμε:
H_0 : $P_1 = P_2$, δεν υπάρχει διαφορά στην πιθανότητα επιτυχίας στα δύο μαθήματα
H_1: $P_1 \neq P_2$ υπάρχει διαφορά στην πιθανότητα επιτυχίας στα δύο μαθήματα
Υπολογίζουμε το X^2 με διόρθωση κατά Yates

$$X_m^2 = \frac{(|\varepsilon - \zeta| - 1)^2}{(\varepsilon + \zeta)} = \frac{(|10 - 30| - 1)^2}{(10 + 30)} = 9.03$$

Θα μπορούσαμε να τοποθετήσουμε τις παρατηρήσεις σε τετράπτυχο πίνακα

	Πέτυχαν	Απέτυχαν	Σύνολο
Μάθημα Α	30	70	100
Μάθημα Β	50	50	100
Σύνολο	80	120	200

Το X^2 υπολογίζεται από τον τύπο:

$$X^2 = \frac{(ad - bc)^2 n}{(a+b)(c+d)(a+c)(b+d)} = \frac{(30 \times 50 - 70 \times 50)^2}{80 \times 120 \times 100 \times 100} \times 200 = 8.33$$

Για 1 βαθμό ελευθερίας οι θεωρητικές τιμές του X^2 είναι :
$X^2_{0.05} = 3.84$, $X^2_{0.01} = 6.64$, $X^2_{0.001} = 10.83$
Συγκρίνοντας την πειραματική X^2 έχουμε:
$X^2_{0.01} = 6.64 < X^2_m = 9.03 < X^2_{0.001} = 10.83$
Ομοίως $X^2_{0.01} = 6.64 < X^2 = 8.33 < X^2_{0.001} = 10.83$
$0.001 \langle P \langle 0.01$

Συμπέρασμα: απορρίπτουμε την H_0,δηλαδή υπάρχει διαφορά στην πιθανότητα επιτυχίας στα δύο μαθήματα και μάλιστα σε επίπεδο στατιστικά πιο σημαντικό.

β) Για το ερώτημα αυτό ο πίνακας των πειραματικών δεδομένων είναι τετράπτυχος και διαμορφώνεται ως εξής

B \ A	Πέτυχαν	Απέτυχαν	Σύνολο
Πέτυχαν	20	10	30
Απέτυχαν	30	40	70
Σύνολο	50	50	100

Η τιμή του X^2 υπολογίζεται με βάση του διορθωμένο κατά Yates τύπο

$$X^2=\frac{\left(\left|ad-bc\right|-\frac{n}{2}\right)^2 n}{(a+b)(c+d)(a+c)(b+d)}=\frac{\left(\left|20\times 40-10\times 30\right|-50\right)^2 100}{30\times 70\times 50\times 50}=3.86$$

Συγκρίνοντας την πειραματική X^2 έχουμε:

$X^2_{0.05}=3.84 < X^2= 3.86 < X^2_{0.01}=6.64$

$0.01 < P < 0.05$

Συμπέρασμα : άτομα, που πέτυχαν στο μάθημα Α, έχουν μεγαλύτερη πιθανότητα επιτυχίας στο μάθημα Β, σε σύγκριση με άτομα , που δεν πέτυχαν στο μάθημα Α σε επίπεδο στατιστικά σημαντικό με $0.01 < P < 0.05$.

Παράδειγμα (10.7)

Καθένας από τους 50 ασθενείς με μια ορισμένη νόσο πήρε σε διαφορετικό χρόνο τα αναλγητικά φάρμακα Α και Β. Απ’ αυτούς 26 εξέφρασαν προτίμηση στο αναλγητικό Α, 12 στο αναλγητικό Β και 12 δεν εξέφρασαν κάποια προτίμηση. Υπερέχει το αναλγητικό Α του αναλγητικού Β;

Λύση:

Η προτίμηση σε ένα αναλγητικό είναι ποιοτικό μέγεθος, και η αξιολόγηση γίνεται με δοκιμασία X^2. Το θέμα είναι με ποιο τύπο του X^2; Τα δεδομένα δείχνουν σύγκριση ανάμεσα σε δύο αναλγητικά φάρμακα με ταυτόχρονη ατομική αντιστοιχία, οπότε ανάλογα με την διατύπωση της ερώτησης, οι παρατηρήσεις τοποθετούνται σε κατάλληλο πίνακα και εφαρμόζεται ο κατάλληλος τύπος του χ^2. Η ερώτηση αν υπερέχει το αναλγητικό Α του αναλγητικού Β, απαιτεί δοκιμασία X^2 κατά ζεύγη.

Θέτουμε H_o: το αναλγητικό Α δεν υπερέχει του αναλγητικού Β ($P_A=P_B$).

H_1: το αναλγητικό Α διαφέρει από το αναλγητικού Β

Κατασκευάζουμε τον πίνακα δεδομένων , και υπολογίζουμε το X^2

Ναι στο Α, ναι στο Β	0
Όχι στο Α, όχι στο Β	12
Ναι στο Α, όχι στο Β	26(ε)
Όχι στο Α, ναι στο Β	12(ζ)
Σύνολο	50

$$X_m^2 = \frac{\left(\left|\varepsilon - \zeta\right| - 1\right)^2}{(\varepsilon + \zeta)} = \frac{\left(\left|26 - 12\right| - 1\right)^2}{(26 + 12)} = 4.447$$

Για 1 βαθμό ελευθερίας οι θεωρητικές τιμές του X^2 είναι $X^2_{0.05}=3.84$, $X^2_{0.01}=6.64$

Συγκρίνοντας την πειραματική X^2 έχουμε:

$X^2_{0.05}=3.84 \langle X^2= 4.447 \langle X^2_{0.01}=6.64$

$0.01 \langle P \langle 0.05$

Συμπέρασμα: απορρίπτουμε την H_0,δηλαδή υπάρχει διαφορά στην προτίμηση στα δύο αναλγητικά φάρμακα και μάλιστα σε επίπεδο στατιστικά πιο σημαντικό. Επειδή $P_A \rangle P_B$ συνεπώς το αναλγητικό Α υπερέχει του αναλγητικού Β.

Παράδειγμα (10.8)

Για να διαπιστωθεί η ισχύς ενός αντιγριπικού εμβολίου, εμβολιάσθηκαν σε ένα σχολείο 150 παιδιά ενώ σε 100 παιδιά ενός γειτονικού σχολείου αντί για εμβόλιο χορηγήθηκε απλά φυσιολογικός ορός .Τα παιδιά και των 2 σχολείων χωρίσθηκαν σε 4 κατηγορίες ανάλογα με την βαρύτητα της γρίπης. Τα αποτελέσματα που πήραμε φαίνονται στον πίνακα :

	Καθόλου	Ελαφρά	Μέτρια	Βαριά
Εμβόλιο	60	40	28	22
Ορός	40	30	12	18

Μπορούμε να ισχυρισθούμε ότι το εμβόλιο είναι αποτελεσματικό;(α=0.05).

Λύση :

Τα δεδομένα της άσκησης είναι ποιοτικά , αυτό σημαίνει ότι ο έλεγχος θα γίνει με το κριτήριο X^2 .Η τιμή του X^2 βρίσκεται από τον τύπο

$X^2 = \Sigma \dfrac{(\Pi i - \theta i)^2}{\theta i} = \Sigma \dfrac{(O - E)^2}{E}$, όπου Π_i = Ο = τα παρατηρούμενα μεγέθη
Θ_i = Ε= τα θεωρητικά ή αναμενόμενα θεωρητικά μεγέθη.
Ορίζουμε H_o: το εμβόλιο δεν είναι αποτελεσματικό .
H_1 : το εμβόλιο είναι αποτελεσματικό .
Κατασκευάζουμε τον πίνακα των πειραματικών δεδομένων με τα σύνολα, και με βάση αυτών και του παρακάτω τύπου

$$E = \theta_i = \frac{(\textit{κάθετο}\ \text{σύνολο}) \cdot (\text{οριζόντιο σύνολο})}{\textit{Γενικό}\ \text{σύνολο}}$$

υπολογίζοντα τα θεωρητικά αναμενόμενα μεγέθη ,και τα αποτελέσματα τοποθετούνται σε ξεχωριστό πίνακα όπως φαίνεται παρακάτω
Τα πειραματικά δεδομένα :

	Καθόλου	Ελαφρά	Μέτρια	Βαριά	Σύνολο
Εμβόλιο	60	40	28	22	150
Ορός	40	30	12	18	100
Σύνολο	100	70	40	40	250

Τα θεωρητικά αναμενόμενα :

	Καθόλου	Ελαφρά	Μέτρια	Βαριά	Σύνολο
Εμβόλιο	60	42	24	24	150
Ορός	40	28	16	16	100
Σύνολο	100	70	40	40	250

Υπολογίζουμε την πειραματική τιμή του X^2 :

$$X^2 = \Sigma \frac{(\Pi i - \theta i)^2}{\theta i} = \frac{(60-60)^2}{60} + \frac{(40-42)^2}{42} + \frac{(28-24)^2}{24} + \frac{(22-24)^2}{24} +$$

$$+ \frac{(40-40)^2}{40} + \frac{(30-28)^2}{28} + \frac{(12-16)^2}{16} + \frac{(18-16)^2}{16} = 2{,}32$$

Υπολογίζουμε τους βαθμούς ελευθερίας από την σχέση :

Β.ε=(κ-1)(L-1)=(S-1).(κ-1) =(4-1).(2-1) =3
Συγκρίνουμε την τιμή X^2 που υπολογίσαμε από τα πειραματικά δεδομένα με την τιμή $X^2_{(s-1)(k-1),a}$ που βρίσκεται από τους στατιστικούς πίνακες . $X^2_{(s-1)(k-1),a} = X_{3,0.05}$ =7,

$X^2 =2,32 < X_{3,0.05} =7,81$

Συμπέρασμα : δεχόμαστε την H_o δεν υπάρχει διαφορά στην αποτελεσματικότητα μεταξύ του εμβολίου και του ορού, άρα το εμβόλιο δεν είναι αποτελεσματικό .

ΑΣΚΗΣΕΙΣ

10.1.Ο παρακάτω πίνακας δίνει τον αριθμό των υγιών και ασθενών ζώων που βρέθηκαν κατά την εξέταση ενός τυχαίου δείγματος σε κάθε μία από τις 4 κτηνοτροφικές μονάδες μιας περιοχής. Να ελεγχθεί σε επίπεδο σημαντικότητας α=5% η υπόθεση ότι όλες οι μονάδες έχουν το ίδιο ποσοστό ασθενών ζώων. Ποιο είναι το ποσοστό των υγειών ζώων της περιοχής;

	1	2	3	4
Υγιή	12	16	16	6
Ασθενή	20	4	14	12

10.2.1000 φοιτητές και 1000 φοιτήτριες ρωτήθηκαν πως πάνε συνήθως από το σπίτι στο Πανεπιστήμιο. Οι απαντήσεις καταχωρήθηκαν ως εξής:

	Φοιτητές	Φοιτήτριες
Με τα πόδια	250	150
Ποδήλατο	100	50
Αυτοκίνητο Ι.Χ.	200	250
Λεωφορείο	450	550

Μπορούμε να ισχυρισθούμε ότι ο τρόπος μετάβασης στο Πανεπιστήμιο είναι ο ίδιος στους φοιτητές και στις φοιτήτριες (α=0.05).

10.3.Οι εξεταστές Α,Β,Γ,Δ διόρθωσαν συνολικά 200 γραπτά. Ο επόμενος πίνακας δίνει τον αριθμό των γραπτών που βαθμολογήθηκαν πάνω και κάτω από την βάση από κάθε εξεταστή.

Εξεταστής	Α	Β	Γ	Δ
Πάνω από τη βάση	30	40	25	25
Κάτω από τη βάση	20	20	15	25

Να εξεταστεί σε επίπεδο σημαντικότητας 5% η υπόθεση ότι και οι 4 εξεταστές βαθμολόγησαν κάτω από τη βάση το ίδιο ποσοστό εξεταζόμενων.

10.4. Δύο εμβόλια (Α,Β) συγκρίθηκαν με ένα εμβόλιο που περιείχε αποσταγμένο νερό (Γ). Ο αριθμός των ζώων που δεν αρρώστησαν, αρρώστησαν ελαφρά, αρρώστησαν σοβαρά τα επόμενα 2 χρόνια ήταν:

Εμβόλια	Α	Β	Γ
Δεν αρρώστησαν	25	25	10
Αρρώστησαν ελαφρά	40	50	10
Αρρώστησαν σοβαρά	10	5	25

Να εξετάσετε εάν τα εμβόλια Α,Β,Γ έχουν την ίδια αποτελεσματικότητα (α=0.05).

10.5. 300 φοιτητές και φοιτήτριες ρωτήθηκαν αν τους αρέσει η κολύμβηση και οι απαντήσεις ήταν:

	Ναι	Όχι	Αδιάφοροι
Φοιτητές	115	50	15
Φοιτήτριες	85	30	5

Μπορούμε να ισχυριστούμε ότι η απάντηση δεν σχετίζεται με το φύλο (α=0.05);

10.6 Σε μία μελέτη της μόλυνσης με τον HIV ιό μεταξύ γυναικών που εισάγονται στο σωφρονιστικό σύστημα της Πολιτείας της Νέας Υόρκης, 457 φυλακισμένες ταξινομήθηκαν σε σχέση με το αν ο όρος τους ήταν θετικός για τον HIV ιό και το ιστορικό τους όσον αφορά ενδοφλέβια χρήση ναρκωτικών.

	Ιστορικό ενδοφλέβιας χρήσης ναρκωτικών		
	Ναι	Όχι	Σύνολο
Θετικός στον HIV ιό	61	27	88
Αρνητικός στον HIV ιό	75	312	387
Σύνολο	136	339	475

α. Να ελέγξετε τη μηδενική υπόθεση ότι δεν υπάρχει σχέση μεταξύ του ιστορικού ενδοφλέβιας χρήσης ναρκωτικών και του να είναι ο ορός θετικός για τον HIV ιό.

β. Να εκτιμήσετε τον σχετικό λόγο συμπληρωματικών πιθανοτήτων του να είναι κανείς θετικός για τον HIV ιό για τις γυναίκες που είχαν κάνει ενδοφλέβια χρήση ναρκωτικών έναντι αυτών που δεν είχαν κάνει.

10.7 Σε μια αποικία ποντικών βρέθηκαν 120 λευκοί 200 μαύροι και 180 ασπρόμαυροι .

α) Μπορούμε να ισχυριστούμε ότι οι λευκοί , μαύροι , ασπρόμαυροι ποντικοί βρίσκονται στην ίδια αναλογία ; (α =0.05). β) Μπορούμε να ισχυριστούμε ότι τα πραγματικά ποσοστά των λευκών, μαύρων, ασπρόμαυρων ποντικών είναι 28%, 37%, και 35% αντίστοιχα ; (α = 0.05).

10.8 Σύμφωνα με ένα εμπειρικό μοντέλο, όταν ο σύζυγος έχει καστανά μάτια και η σύζυγος γαλανά , τα παιδιά έχουν μάτια με χρώμα : γαλανά ,καστανά ,πράσινα , σε αναλογία 8:7:3. Σε 1000 ζευγάρια που ο πατέρας είχε καστανά μάτια και η μητέρα γαλανά βρέθηκε ότι :

Χρώμα ματιών	Γαλανά	Καστανά	Πράσινα
Αριθμός παιδιών	460	370	170

Σε επίπεδο σημαντικότητας $a = 0.05$ μπορούμε να ισχυριστούμε ότι τα γαλανά , καστανά , πράσινα μάτια είναι πράγματι σε αναλογία 8:7:3 ;

10.9 Σε γεωτρήσεις μετρήθηκε ο αριθμός των βλαβών που παθαίνει το γεωτρύπανο κάθε εβδομάδα . Σε 50 εβδομάδες εργασίας οι βλάβες ήταν :

Αριθμός βλαβών	0	1	2	3	≥ 4
Αριθμός εβδομάδων	5	12	16	12	5

Μπορούμε να ισχυριστούμε σε επίπεδο $a = 0.05$ ότι ο αριθμός των βλαβών ακολουθεί κατανομή Poisson ; (Δίνεται $e^{-2} = 0.135$)

10.10 Σε μια έρευνα που έγινε για να διαπιστωθεί αν το μέγεθος ενός είδους φυτού σχετίζεται με την ανθεκτικότητα του φυτού σε κάποια ασθένεια, μελετήθηκε ένα δείγμα 100 φυτών, τα οποία κατατάχθηκαν σε ‘ψηλά’, ‘μέτρια’ και ‘κοντά’ και εξετάστηκε αν προσβλήθηκαν από την ασθένεια ή όχι. Το πλήθος των φυτών σε κάθε κατηγορία δίνεται στον παρακάτω πίνακα.

	Ψηλά	Μέτρια	Κοντά
Προσεβλήθησαν	13	12	15
Δεν προσεβλήθησαν	22	18	20

Ελέγξτε εάν εξαρτάται η ανθεκτικότητα του φυτού στην ασθένεια, από το μέγεθος του φυτού (α = 0.05).

10.11 Από σύνολο 100 ασθενών που πάσχουν από κάποια ασθένεια σχηματίζουμε τυχαία δύο ομάδες Α και Β με 50 ασθενείς η κάθε μία. Στους ασθενείς της Α ομάδας μόνο δίνουμε ένα νέο φάρμακο. Η ομάδα Β χρησιμοποιείται σαν μάρτυρας. Στον επόμενο πίνακα καταγράφονται οι συχνότητες σε κάθε κατηγορία.

	Έγιναν καλά	Δεν έγιναν καλά
Ομάδα Α (με φάρμακο)	40	10
Ομάδα Β (χωρίς φάρμακο)	35	15

Ελέγξτε εάν το νέο φάρμακο βοηθάει στην καταπολέμηση της ασθένειας (α = 0.05).

10.12 Σύμφωνα με τη θεωρία του Mendel, αν διασταυρωθούν φυτά μπιζελιών με στρογγυλούς-κίτρινους σπόρους, με φυτά μπιζελιών με ρυτιδωμένους πράσινους σπόρους, θα δώσουν σπόρους στρογγυλούς-κίτρινους, ρυτιδωμένους- κίτρινους, στρογγυλούς-πράσινους και ρυτιδωμένους-πράσινους σε αναλογία 9:3:3:1. Σε ένα πείραμα παρατηρήθηκαν τα παρακάτω αποτελέσματα:
i) Στρογγυλοί-κίτρινοι 44 **ii)** Ρυτιδωμένοι –κίτρινοι 17
iii) Στρογγυλοί-πράσινοι 13 **iv)** Ρυτιδωμένοι-πράσινοι 6.
Συμφωνούν οι παρατηρήσεις αυτές με τη θεωρία του Mendel; (α=0.05).

Κεφάλαιο 11

Γραμμική παλινδρόμηση – Συσχέτιση

11.1 Ανάλυση Παλινδρόμησης (Regression Analysis): είναι ο κλάδος της στατιστικής ο οποίος εξετάζει τη σχέση μεταξύ δύο ή περισσοτέρων μεταβλητών, ώστε να είναι δυνατή η πρόβλεψη της μιας από τις υπόλοιπες.

11.2 Η έννοια της στατιστικής συσχέτισης και της στατιστικής εξάρτησης.

Το πρόβλημα της συσχέτισης και εξάρτησης ποσοτικών χαρακτηριστικών υπάρχει, όταν οι εξεταζόμενες μεταβλητές X,Y(χαρακτηριστικά – μεγέθη) μετριούνται στην ίδια ερευνητική μονάδα, δηλαδή στο άτομο ή στην οικογένεια ή μια ευρύτερη ομάδα. Η **συσχέτιση** δύο μεγεθών X,Y (και η εξάρτηση) λέγεται θετική όταν σε υψηλές τιμές του X αντιστοιχούν συνήθως υψηλές τιμές του Y, και αρνητική όταν σε υψηλές τιμές του X αντιστοιχούν συνήθως χαμηλές τιμές του Y.
Η διάκριση ανάμεσα στην συσχέτιση και την εξάρτηση είναι περισσότερο εννοιολογική παρά μαθηματική. Η σχέση των δύο μεταβλητών X και Y είναι δυνατό να μελετηθεί με τρείς τρόπους:
α) να θεωρηθεί η X ως ανεξάρτητη μεταβλητή (independent), και να διερευνηθεί πως εξαρτάται η Y (εξαρτημένη μεταβλητή dependent variable) – **πρόβλημα εξάρτησης**.
β) να θεωρηθεί η Y ως ανεξάρτητη μεταβλητή, και να διερευνηθεί πως εξαρτάται η X (εξαρτημένη μεταβλητή) – **πρόβλημα εξάρτησης.**
γ) να μη θεωρηθεί οποιασδήποτε από τις δύο μεταβλητές ως ανεξάρτητη και να ερευνηθεί απλά ο βαθμός της συμμεταβολής τους – **πρόβλημα συσχέτισης**.
Όταν οι τιμές των δύο μεταβλητών λαμβάνονται τυχαία είναι δυνατό να εφαρμοσθούν τόσο οι μέθοδοι της στατιστικής συσχέτισης όσο και οι μέθοδοι της στατιστικής εξάρτησης. Η καταλληλότερη μέθοδος επιλέγεται ανάλογα με τις ιδιαίτερες συνθήκες του κάθε προβλήματος. Αν η μία από τις μεταβλητές θεωρείται αίτιο της άλλης, συνήθως αυτή λαμβάνεται ως ανεξάρτητη. Αν το πρόβλημα αναφέρεται στην ύπαρξη ή μη συσχέτισης

ανάμεσα σε δύο μεταβλητές και δεν υπάρχει δυνατότητα ή ενδιαφέρον για ενδεχόμενη αιτιολογική σχέση ,τότε οι μέθοδοι συσχέτισης είναι κατάλληλες.
Όταν οι δύο μεταβλητές X και Y που εκφράζουν τις ιδιότητες σχετίζονται με τέτοιο τρόπο ώστε η μία –έστω η Y-να μπορεί να προβλεφθεί καλύτερα όταν η άλλη-η X- είναι γνωστή, τότε λέμε ότι τα Xκαι Y σχετίζονται (θετικά ή αρνητικά).

11.3 Απλή γραμμική παλινδρόμηση

Η ευθεία Y= α + βX λέγεται ευθεία γραμμικής παλινδρόμησης και είναι ο απλούστερος τύπος μοντέλου που συνδέει τα X και Y. Ο συντελεστής **α** είναι η τιμή του Y για X=0 (intercept) και το **β** είναι η κλίση (slope) της ευθείας.
Ένα μοντέλο που δίνει την δυνατότητα στο Y να μη βρίσκεται στην ευθεία

Y = α + βX (11.1) είναι το:

Y= α + βX + ε (11.2)

όπου **ε** είναι ένα τυχαίο σφάλμα και παριστάνει την διαφορά της παρατηρούμενης τιμής y, για δοσμένο X , από την θεωρητική τιμή α + βX .Τα μοντέλα αυτού του είδους ονομάζονται στοχαστικά. Δεδομένης της υπόθεσης περί κανονικότητας των τιμών της Y για ένα δοθέν X, η μέση τιμή της Y συνδέεται γραμμικά με την X. Επομένως , αν υπάρχει γραμμικότητα , τότε η σχέση που συνδέει την Y με την X στον πληθυσμό είναι της μορφής:

$$E\left(y_i / x_i \right) = a + \beta \chi_\iota \qquad (11.3) \quad , \text{ με } E(\varepsilon_{\rm i}) = 0$$

Η σημειακή εκτίμηση της παραπάνω σχέσης θα είναι:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}\, x_i \qquad (11.4)$$

Οι εκτιμήτριες των παραμέτρων $\hat{a}$ και $\hat{\beta}$ του παραπάνω μοντέλου παλινδρόμησης προκύπτουν με τη μέθοδο των ελαχίστων τετραγώνων και είναι:

$$\hat{\beta} = \frac{\mathrm{S_{xy}}}{\mathrm{S_{xx}}} = \frac{\Sigma(\mathrm{x_i} - \bar{\mathrm{x}})(\mathrm{y_i} - \bar{y})}{\Sigma(\mathrm{x_i} - \bar{x})^2} \qquad (11.6)$$

όπου : $S_{xy} = \sum x_{\rm i} y_i - n\bar{x}\,\bar{y}$ (11.7) , $S_{xx} = \sum x_i^2 - n\bar{x}^2$ (11.8)

Η σχέση (11.6) είναι ισοδύναμος με :

$$\hat{\beta} = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\Sigma \mathrm{x_i})^2} \qquad (11.9)$$

Ο συντελεστής α υπολογίζεται από την σχέση :

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\ .\overline{X} \qquad (11.5)$$

11.3.1 Στατιστικός έλεγχος του συντελεστή εξάρτησης β

Ο Στατιστικός έλεγχος του συντελεστή εξάρτησης β γίνεται με την κατασκευή διαστήματος εμπιστοσύνης. Συνήθως μας ενδιαφέρει να ελέγξουμε την υπόθεση **H_0:β = 0** διότι αν ισχύει αυτή η υπόθεση τότε το μοντέλο , $\hat{Y}_i = \hat{\alpha} + \hat{\beta}\, x_i$ γίνεται $\hat{Y}_i = \hat{\alpha}$ που σημαίνει ότι η τιμή της μεταβλητής y είναι ανεξάρτητη από την τιμή της x.

Το 100(1-α)% δ.ε του β δίνεται από την σχέση:

$$\hat{\beta} \pm \mathrm{S_\beta}\,.\,t_{n-2;\frac{a}{2}} \qquad (11.10)\ , \quad \text{όπου} \quad S_\beta = \sqrt{\frac{\mathrm{S}^2}{(\text{n-1})\,S_x^2}} \qquad (11.11) \qquad \text{και}$$

$$S^2 = \frac{n-1}{n-2}\left(S_y^2 - \hat{\beta}^2.\mathrm{S_x^2}\right) \qquad (11.12)$$

Αν στο διάστημα εμπιστοσύνης (δ.ε) δεν περιλαμβάνεται η τιμή μηδέν ,τότε η H_o απορρίπτεται .Το πρόσημο του συντελεστή εξάρτησης (κλίση) $\hat{\beta}$ φανερώνει και την μορφή της γραμμικής σχέσης (θετική ή αρνητική) μεταξύ των δύο μεταβλητών.

11.3.2 Στατιστικός έλεγχος της προβλεπόμενης τιμής ,διάστημα εμπιστοσύνης.

Το δ.ε της προβλεπόμενης τιμής y δίνεται από την σχέση:

$$\hat{y}_{n-1} \pm S.t_{n-1,a/2}\sqrt{\left(\frac{1}{n}\right) + \frac{(x-\overline{x})^2}{(n-1)\,S_x^2}} \qquad (11.13) \qquad \text{όπου}$$

$$S^2 = \frac{n-1}{n-2}\left(S_y^2 - \frac{S_{xy}^2}{S_x^2}\right) = \frac{n-1}{n-2}\left(S_y^2 - \hat{\beta}^2 S_x^2\right) \qquad (11.14) \quad \text{και}$$

$$S_x^2 = \frac{1}{n-1}\left(\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right), \qquad S_y^2 = \frac{1}{n-1}\left(\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right)$$

11.4 Μέθοδος διερεύνησης της στατιστικής συσχέτισης δύο ποσοτικών χαρακτηριστικών

Η διερεύνησης της στατιστικής συσχέτισης δύο ποσοτικών χαρακτηριστικών (μεταβλητών) γίνεται συνήθως με τον προσδιορισμό του παραμετρικού **συντελεστή συσχέτισης κατά Pearson.** Ο συντελεστή συσχέτισης συμβολίζεται με το σύμβολο r και αποτελεί αξιόλογο μέτρο του βαθμού της ευθύγραμμης συσχέτισης των δύο τυχαία επιλεγμένων μεταβλητών, όταν και οι δύο κατανομές είναι κανονικές.
Ο συντελεστή συσχέτισης είναι **καθαρός αριθμός**, δηλαδή στερείται φυσικών διαστάσεων και μονάδων, και μπορεί να πάρει τιμές από -1 μέχρι +1 . Η τιμή -1 δηλώνει ότι υπάρχει απόλυτα αρνητική ευθύγραμμη συσχέτιση, η τιμή +1, ότι υπάρχει απόλυτα θετική ευθύγραμμη συσχέτιση, και η τιμή 0, ότι δεν υπάρχει συσχέτιση (ευθύγραμμη). Ο συντελεστή συσχέτισης υπόκειται σε μεγάλη τυχαία διακύμανση , κατά συνέπεια, χρειάζεται μεγάλος αριθμός παρατηρήσεων για τον περιορισμό της διακύμανσης σε αποδεκτά επίπεδα.

11.4.1 Ο (παραμετρικός) συντελεστής συσχέτισης r (Pearson correlation coefficient) υπολογίζεται από τον τύπο:

$$r = \frac{S_{xy}}{S_x S_y} = \hat{\rho} = \frac{\sum (x_i - \bar{x})(y - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} =$$

$$= \frac{\sum x_i y_i - n\bar{x}\,\bar{y}}{\sqrt{\sum x_i^2 - n\bar{x}_i^2}.\sqrt{\sum y_i^2 - n\bar{y}^2}} \qquad (11.15)$$

Ο παραπάνω τύπος μετατρέπεται στον ισοδύναμο τύπο :

$$r = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\left(\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right)\left(\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right)}} \qquad (11.16)$$

Ο οποίος θεωρείται περισσότερο εύχρηστος.
Ο έλεγχος της στατιστικής σημαντικότητας γίνεται με μηδενική υπόθεση:
$H_0 : r = \rho = 0$ (δεν υπάρχει γραμμική σχέση) , και εναλλακτική
$H_1 : r = \rho \neq 0$ (υπάρχει γραμμική σχέση)
Το κριτήριο ελέγχου ακολουθεί την κατανομή t student με n-2 βαθμούς ελευθερίας, και είναι:

$$t = r.\sqrt{\frac{n-2}{1-r^2}} \qquad (11.17)$$

Η H_0 απορρίπτεται σε επίπεδο σημαντικότητας α αν: $|t| > t_{n-2\,;\,\alpha/2}$

Παράδειγμα (11.1)
Κατά την διεξαγωγή μιας άντλησης σε μια γεώτρηση που έγινε σε χαλαρά πετρώματα, μετρήθηκαν η πτώση στάθμης y σε m, σε σχέση με το χρόνο άντλησης t σε min. Τα δεδομένα είναι

Χρόνος άντλησης t	1	2	3	4	5	6	7
Πτώση στάθμης y	0	1	2	3,5	4,5	5	5

i)Να υπολογιστεί η ευθεία παλινδρόμησης y = α + β.t
ii)Ποια θα είναι η πτώση στάθμης του νερού για το χρόνο άντλησης 3.5 min;
Λύση:
Για να υπολογιστεί η ευθεία **Y = α + βX**
πρέπει να υπολογίσουμε πρώτα την κλίση της ευθείας **β** και ύστερα την τιμή του **α.** Η κλίση της ευθείας υπολογίζεται από τον τύπο:

$$\hat{\beta} = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\Sigma \mathrm{x_i})^2}$$, και το α από τον τύπο: $\hat{\alpha} = \bar{y} - \hat{\beta}.\bar{x}$

Όπως βλέπουμε για να υπολογίσουμε το β χρειαζόμαστε τα μεγέθη,
$\sum x_i y_i, \sum x_i^2, \sum x_i, \sum y_i$
Κατασκευάζουμε πίνακα με τα παραπάνω μεγέθη

X,(t)	Y	X.Y	X^2
1	0	0	1
2	1	2	4
3	2	6	9
4	3.5	14	16
5	4.5	22.5	25
6	5	30	36
7	5	35	49
$\sum X = 28$	$\sum Y = 21$	$\sum x_i y_i$ =109.5	$\sum x_i^2$ =140

Υπολογίζουμε το β από τον τύπο:

$$\hat{\beta}=\frac{n\sum x_i y_i-(\sum x_i)(\sum y_i)}{n\sum x_i^2-(\Sigma \mathrm{x_i})^2}=\frac{7\times 109{,}5-28\times 21}{7\times 140-28^2}=0{,}91$$

$$\bar{y}=\frac{21}{7}=3\ ,\bar{x}=\frac{28}{7}=4$$

Υπολογίζουμε το α από τον τύπο:

$$\hat{\alpha}=\bar{y}-\hat{\beta}.\bar{x}=3-0{,}91.4=-0{,}64$$

Συνεπώς i) η ευθεία παλινδρόμησης είναι: Y= -0,64 + 0,91t

ii) για χρόνο άντλησης 3,5 min προβλέπεται να είναι :

Y= -0,64 +0,91.3,5 = 2,55 min

Παράδειγμα (11.2)

Μελετήθηκε η σχέση ανάμεσα στο άγχος X, που μετρήθηκε με ένα τεστ, και στη συχνότητα καρδιακών παλμώνY ανά λεπτό σε 12 ενήλικες. Τα δεδομένα δίνονται στον πίνακα:

Y	95	90	85	75	80	70	70	60	60	55	55	70
X	48	41	45	41	42	36	38	36	30	32	34	35

Υπάρχει σχέση ανάμεσα στην καρδιακή συχνότητα και στο άγχος;

Λύση:

Η σχέση ανάμεσα στην συχνότητα καρδιακών παλμών και στο άγχος είναι πρόβλημα εξάρτισης , με ανεξάρτητη μεταβλητή X το άγχος, και εξαρτημένη Y η συχνότητα των καρδιακών παλμών.

Υπολογίζουμε τον συντελεστή εξάρτησης β από τον τύπο:

$$\hat{\beta}=\frac{n\sum x_i y_i-(\sum x_i)(\sum y_i)}{n\sum x_i^2-(\Sigma \mathrm{x_i})^2}$$

Κατασκευάζουμε πίνακα δεδομένων με τις στήλες που αντιστοιχούν στα μεγέθη $\sum x_i y_i , \sum x_i^{\,2} , \sum x_i , \sum y_i$.

Αριθμός	X	Y	X.Y	X^2	Y^2
1	48	95	4560	2304	9025
2	41	90	3690	1681	8100
3	45	85	3825	2025	7225
4	41	75	3075	1681	5625
5	42	80	3360	1764	6400
6	36	70	2520	1296	4900
7	38	70	2660	1444	4900
8	36	60	2160	1296	3600
9	30	60	1800	900	3600
10	32	55	1760	1024	3025
11	34	55	1870	1156	3025
12	35	70	2450	1225	4900
Σύνολο	458	865	33730	17796	64325

$$\hat{\beta} = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\Sigma \mathrm{x_i})^2} = \frac{12\times 33730 - 458\times 865}{12\times 17796 - 458^2} = 2{,}268$$

Ο έλεγχος της στατιστικής σημαντικότητας του β γίνεται με την κατασκευή του δ.ε :

$$\hat{\beta} \pm \mathrm{S_\beta} \,.\, t_{n-2;\frac{a}{2}} \quad \text{όπου} \quad S_\beta = \sqrt{\frac{\mathrm{S}^2}{(\mathrm{n}-1)S_x^2}} \quad ,\text{και} \quad S^2 = \frac{n-1}{n-2}\left(S_y^2 - \hat{\beta}^2.\mathrm{S_x^2}\right)$$

$$S_y^2 = \frac{\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}}{n-1} = \frac{64325 - \frac{865^2}{12}}{11} = 179{,}36$$

$$S_x^2=\frac{\sum x_i^2-\frac{\left(\sum x_i\right)^2}{n}}{n-1}=\frac{17796-\frac{458^2}{12}}{11}=28,70$$

$$S^2=\frac{n-1}{n-2}\left(S_y^2-\hat{\beta}^2.\mathrm{S}_\mathrm{x}^2\right)=\frac{11}{12}\left(179,36-2,268^2.28,70\right)=29,09$$

$$S_\beta=\sqrt{\frac{\mathrm{S}^2}{(\mathrm{n}-1)S_x^2}}=\sqrt{\frac{29,09}{11.28,70}}=0,303$$

$$\hat{\beta}\pm \mathrm{S}_\beta\,.\,t_{10;0,005}=2,268\pm 0,303.3,169=2,268\pm 0,687$$

$$\hat{\beta}\pm \mathrm{S}_\beta\,.\,t_{10;0,0005}=2,268\pm 0,303.4.59=2,268\pm 1,391$$

Άρα $0,877\leq \hat{\beta}\leq 3,659$

Συμπέρασμα: επειδή το μηδέν δεν εμπεριέχεται στο δ.ε απορρίπτουμε την H_0. Άρα υπάρχει σχέση ανάμεσα στην καρδιακή συχνότητα και στο άγχος με συντελεστή εξάρτησης β=b=2,268 καρδιακοί παλμοί ανά μονάδα άγχους σε επίπεδο σημαντικότητας 0,001. ($p<0,001$).

Παράδειγμα (11.3)

Σε 10 ασθενείς με χρόνια βαριά αναιμία μετρήθηκαν το επίπεδο της αιμοσφαιρίνης (g/100 mL) στο αίμα και η στεφανιαία ροή αίματος (mL/100 g LV/min) και βρέθηκαν οι παρακάτω τιμές:

Άτομο	1_0	2_0	3_0	4_0	5_0	6_0	7_0	8_0	9_0	10_0
Αιμοσφ	1,6	2,4	3,5	3,5	4,7	4,8	5,1	6,0	6,1	6,4
τεφ.ροή	222	198	193	208	155	139	177	121	140	122

Υπάρχει συσχέτιση ανάμεσα στις τιμές των δύο μεγεθών;

Λύση:

Το πρόβλημα αναφέρεται στην ύπαρξη ή μη συσχέτισης ανάμεσα σε δύο μεταβλητές και δεν υπάρχει δυνατότητα ή ενδιαφέρον για ενδεχόμενη αιτιολογική σχέση ,άρα είναι πρόβλημα συσχέτισης .Η διερεύνησης της στατιστικής συσχέτισης δύο ποσοτικών χαρακτηριστικών (μεταβλητών) γίνεται συνήθως με τον προσδιορισμό του παραμετρικού συντελεστή συσχέτισης κατά Pearson r.O (παραμετρικός) συντελεστής συσχέτισης r (Pearson correlation coefficient) υπολογίζεται από τον τύπο:

$$r = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\left(\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right)\left(\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right)}}$$

Από τον παραπάνω τύπο βλέπουμε ότι χρειαζόμαστε τα μεγέθη
$\sum x_i y_i$, $\sum x_i^2$
$\sum x_i$, $\sum y_i$. $\sum y_i^2$,οπότε κατασκευάζουμε πίνακα δεδομένων με τις αντίστοιχες στήλες:

Άτομο	Αιμοσφ(X)	Στεφ.ροή(Y)	X.Y	X^2	Y^2
1	1,6	222	355,2	2,56	49284
2	2,4	198	475,2	5,76	39204
3	3,5	193	675,5	12,25	37249
4	3,5	208	728	12,25	43264
5	4,7	155	728,5	22,09	24025
6	4,8	139	667,2	23,04	19321
7	5,1	177	902,7	26,01	31329
8	6,0	121	726	36	14641
9	6,1	140	854	37,21	19600
10	6,4	122	780,8	40,96	14884
Σύνολο	44,1	1675	6893,1	218,13	292801

Υπολογίζουμε το r από την σχέση **:**

$$r = \frac{\sum x_i y_i - \dfrac{\sum x_i \sum y_i}{n}}{\sqrt{\left(\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}\right)\left(\sum y_i^{\ 2} - \dfrac{\left(\sum y_i\right)^2}{n}\right)}}$$

$$r = \frac{6893{,}1 - \dfrac{44{,}1 \times 1675}{10}}{\sqrt{\left(218{,}13 - \dfrac{44{,}1^2}{10}\right)\left(292801 - \dfrac{1675^2}{10}\right)}} = -0.918$$

Ο έλεγχος της στατιστικής σημαντικότητας του r γίνεται με t student με n-2 βαθμούς ελευθερίας.
H_0 : $r =$ ρ = 0 (δεν υπάρχει γραμμική σχέση) , και εναλλακτική
H_1 : $r =$ ρ $\neq$ 0 (υπάρχει γραμμική σχέση)
Το κριτήριο ελέγχου είναι:

$$t = r.\sqrt{\frac{n-2}{1-r^2}} = -0{,}918 \times \sqrt{\frac{10-2}{1-(-0{,}918)^2}} = -\,6{,}547$$

Η H_0 απορρίπτεται σε επίπεδο σημαντικότητας α αν: $|t| \rangle\ t_{n-2;\alpha/2}$

Οι θεωρητικές τιμές t από τους στατιστικούς πίνακες είναι :
$t_{8;0,005} = 3{,}355$, $t_{8;0,025} = 2{,}306$

$|\,t\,| = 6{,}547 \rangle\ t_{8;0,005} = 3{,}355$

Συμπέρασμα: Υπάρχει αρνητική γραμμική συσχέτιση ανάμεσα στις τιμές των δύο μεγεθών σε επίπεδο σημαντικότητας α = 0,01με συντελεστή συσχέτισης
r = - 0,918.

Παράδειγμα (11.4)
Ένας γεωπόνος μελετάει την σχέση μεταξύ ποσότητας ενός λιπάσματος (x) και της απόδοσης (y) μιας καλλιέργειας σιταριού. Η μελέτη έδωσε τα ακόλουθα αποτελέσματα:

x	2	4	5	7	8	10	12	15	17	20
y	50	55	60	65	70	75	80	80	90	100

Δίνονται: $\sum_{i=1}^{10} y_i = 725$, $\sum_{i=1}^{10} x_i^2 = 1316$, $\sum_{i=1}^{10} y_i^2 = 54775$, $\sum_{i=1}^{10} x_i y_i = 8077$.

$\sum_{i=1}^{10} x_i = 100$

α) Βρείτε την ευθεία ελάχιστων τετραγώνων για τα δεδομένα αυτά.
β) Υπολογίστε το συντελεστή συσχέτισης. Τι συμπεραίνετε για το μοντέλο;
γ) Για ποσότητα λιπάσματος 11 κιλά , ποια απόδοση αναμένετε σύμφωνα με το μοντέλο σας στο (α) ; Δώστε 95% διάστημα εμπιστοσύνης για την μέση απόδοση του σιταριού, όταν χρησιμοποιούνται 11 κιλά λιπάσματος.

Λύση:

α) Η ευθεία ελάχιστων τετραγώνων έχει την μορφή:

$\hat{Y}_i = \hat{\alpha} + \hat{\beta}.\, x_i$ (Y= α + β.χ) όπου

$$\hat{\beta} = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\Sigma \mathrm{x_i})^2} \quad ,\text{και} \quad \hat{\alpha} = \bar{y} - \hat{\beta}\,.\bar{x} \ .$$

Άρα υπολογίζουμε πρώτα το $\hat{\beta}$ ύστερα το $\hat{\alpha}$ και αντικαθιστούμε στην παραπάνω εξίσωση.

$$\hat{\beta} = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\Sigma \mathrm{x_i})^2} = \frac{10\times 8075 - 100\times 725}{10\times 1316 - 100^2} = \frac{8250}{3160} = 2,61$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{100}{10} = 10 \quad , \quad \bar{y} = \frac{\sum y_i}{n} = \frac{725}{10} = 72,5$$

$\hat{\alpha} = \bar{y} - \hat{\beta}\,.\bar{x} = 72,5 - 2,61\times 10 = 46,4$, συνεπώς η ευθεία ελάχιστων τετραγώνων έχει την μορφή:

$$\hat{Y}_i = 46,4 + 2,61\ \hat{\mathrm{x}}_i$$

β) Ο συντελεστή συσχέτισης υπολογίζεται από τον παρακάτω τύπο:

$$r = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\left(\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right)\left(\sum y_i^{\,2} - \frac{\left(\sum y_i\right)^2}{n}\right)}}$$

$$= \frac{8075 - \frac{100 \times 725}{10}}{\sqrt{\left(1316 - \frac{100^2}{10}\right)\left(54775 - \frac{725^2}{10}\right)}} = 0{,}987$$

Ο έλεγχος της στατιστικής σημαντικότητας του r γίνεται με t student με n-2 βαθμούς ελευθερίας.

H_0 : $r = \rho = 0$ (δεν υπάρχει γραμμική σχέση) , και εναλλακτική

H_1 : $r = \rho \neq 0$ (υπάρχει γραμμική σχέση)

Το κριτήριο ελέγχου είναι:

$$t = r.\sqrt{\frac{n-2}{1-r^2}} = 0{,}987\sqrt{\frac{8}{1-0{,}987^2}} = 17{,}37$$

Οι θεωρητικές τιμές t από τους στατιστικούς πίνακες είναι :

$t_{8;0,005} = 3{,}355$, $t_{8;0,025} = 2{,}306$

$|\, t \,| = 17{,}37 \rangle\ t_{8;0,005} = 3{,}355$

Συμπέρασμα: Υπάρχει θετική γραμμική συσχέτιση ανάμεσα στις τιμές των δύο μεγεθών του μοντέλου σε επίπεδο σημαντικότητας p = 0,01με συντελεστή συσχέτισης r =0,987.

γ)Για ποσότητα λιπάσματος 11κιλά η απόδοση εκτιμάται κατά μέσο όρο, ίση με

$\hat{Y}_i = 46{,}4 + 2{,}61\, x_i$ =46,4+2,61.11=75,11 κιλά

Το 95% διάστημα εμπιστοσύνης για την μέση απόδοση του σιταριού, όταν χρησιμοποιούνται 11 κιλά λιπάσματος δίνεται από την σχέση:

$$\hat{y}_{n-1} \pm S.t_{n-1,\alpha/2}\sqrt{\left(\frac{1}{n}\right) + \frac{(x-\bar{x})^2}{(n-1)\,S_x^2}} \quad \text{όπου } s^2 = \frac{n-1}{n-2}\left(S_y^2 - \hat{\beta}^2 S_x^2\right)$$

και $S_x^2 = \frac{1}{n-1}\left(\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right)$, $S_y^2 = \frac{1}{n-1}\left(\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right)$,άρα

$$S_x^2 = \frac{1}{10-1}\left(1316 - \frac{100^2}{10}\right) = 35,11 \quad , S_y^2 = \frac{1}{10-1}\left(54775 - \frac{725^2}{10}\right) = 245,83$$

συνεπώς

$$S^2 = \frac{9}{8}\left(245,83 - 2,61^2.35,11\right) = 7,489 \quad \Rightarrow s = 2,737$$

$$\hat{y}_{n-1} \pm S.t_{n-1,a/2}\sqrt{\left(\frac{1}{n}\right) + \frac{(x-\overline{x})^2}{(n-1)\,S_x^2}}$$

$$= 75,11 \pm 2,737.2,262\sqrt{\frac{1}{10} + \frac{\left(11-10\right)^2}{9.35,11}} = 75,11 \pm 1,99$$

Άρα δ. ε $=75,11 \pm 1,99$, δηλαδή ($73,12 \langle \ \hat{y} \ \langle \ 77,1$)

Ασκήσεις

11.1 Σ’ ένα τυχαίο δείγμα από 5 καρπούς , ζυγίσαμε το βάρος τους x με το φλοιό και το βάρος y αφού αφαιρέσαμε το φλοιό

X	4	3.5	5	3	4.5
Y	3	2.5	3.5	2.5	3.5

i)Να υπολογιστεί η ευθεία παλινδρόμησης y=α+βχ

ii)Ποιό θα είναι το βάρος ενός καρπού χωρίς φλοιό, όταν το βάρος του καρπού μαζί με το φλοιό είναι x=4;

Δίνονται : Σx =20 , Σy =15 , Σx =82.5 , Σy =46 , Σx y =61.

11.2 Τα παρακάτω δεδομένα δίνουν τους ρυθμούς κατανάλωσης οξυγόνου ενός είδους πουλιών, όταν μετριούνται σε διαφορετικές περιβαλλοντικές θερμοκρασίες:

Θερμοκρασία (C^0) (x)	-15	-10	-5	0	5	10	12	19
Κατανάλωση οξυγόνου (ml/g/hr) (y)	4.4	3.8	3.6	3.3	2.6	2.3	2.2	1.8

(α) Βρείτε την ευθεία παλινδρόμησης της κατανάλωσης οξυγόνου των πουλιών στις διάφορες περιβαλλοντικές συνθήκες.
(β) Υπολογίστε τον συντελεστή συσχέτισης. Τι συμπεραίνετε;
(γ) Ελέγξτε την υπόθεση H_0: β=0, σε επίπεδο σημαντικότητας 5%. Τι συμπεραίνετε από το αποτέλεσμα του τεστ;
(δ) Ποιος είναι ο μέσος αναμενόμενος ρυθμός κατανάλωσης οξυγόνου των πουλιών σε θερμοκρασία περιβάλλοντος 15^0C. Δώστε 95% διάστημα εμπιστοσύνης για αυτό το μέσο ρυθμό κατανάλωσης οξυγόνου.

Κεφάλαιο 12

Ανάλυση διασποράς ή διακύμανσης

12.1 Εισαγωγή

Η **ανάλυση της διασποράς** (διακύμανσης) (ANalysis Of VAriance – ANOVA) είναι η μεθοδολογία εκείνη που ασχολείται με την εξέταση και τον προσδιορισμό των πηγών των αποκλίσεων που παρατηρούνται σε πειραματικά δεδομένα. Είναι δηλαδή μία στατιστική μέθοδος με την οποία η μεταβλητότητα που υπάρχει σ' ένα σύνολο δεδομένων διασπάται στις επιμέρους συνιστώσες της με στόχο την κατανόηση της σημαντικότητας των διαφορετικών πηγών προέλευσής της. Η **ανάλυση διασποράς** μελετά τη σχέση που έχει μια εξαρτημένη μεταβλητή Y, τις τιμές της οποίας μπορούμε να παρατηρήσουμε, με τις τιμές ενός παράγοντα A ή πολλών παραγόντων που είναι και οι ανεξάρτητες μεταβλητές. Είναι κατάλληλη για δεδομένα που βασίζονται πάνω στην έρευνα ή σε πειράματα και μπορεί να χρησιμοποιηθεί είτε οι ανεξάρτητες μεταβλητές είναι ποσοτικές είτε ποιοτικές. Στην απλούστερη μορφή της η ANOVA μας δίνει τη δυνατότητα να δοκιμάσουμε την υπόθεση ότι οι μέσες τιμές διαφόρων πληθυσμών είναι ίσες. Η ανάλυση διασποράς στηρίζεται στη σύγκριση της μεταβλητότητας μεταξύ των δειγματικών μέσων και της μεταβλητότητας των τιμών μέσα σε κάθε δείγμα. Αυτή η διαδικασία ανάλυσης διαιρεί τη συνολική μεταβλητότητα σε δύο βασικές συνιστώσες :

1.Μεταβλητότητα μεταξύ των δειγμάτων.

2. Μεταβλητότητα μέσα σε κάθε δείγμα.

12.2 Το κριτήριο της ανάλυσης διασποράς

Η σύγκριση της μεταβλητότητας μεταξύ των δειγματικών μέσων και της μεταβλητότητας των τιμών μέσα σε κάθε δείγμα γίνεται με τον υπολογισμό του λόγου F ο οποίος είναι:

$$F = \frac{S_0^2}{S_p^2} \qquad (12.1)$$

Η κοινή διακύμανση κ δειγμάτων S_p^2 που είναι εκτιμητής του σ^2 , δηλαδή η μεταβλητότητα μεταξύ δειγμάτων δίνεται από τον τύπο:

$$S_p^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2+\cdots+(n_\kappa-1)S_\kappa^2}{n_1+n_2+\cdots+n_\kappa-\kappa}=\frac{SSE}{N-\kappa} \qquad (12.2)$$

Όπου SSE είναι το άθροισμα των τετραγώνων των σφαλμάτων, το οποίο μετράει την συνολική μεταβλητότητα μέσα στα δείγματα, και $N=n_1+n_2+\cdots+n_k$

Το S_0^2 είναι το μέτρο της μεταβλητότητας μεταξύ των δειγμάτων και δίνεται από την σχέση:

$$S_0^2=\frac{1}{\kappa-1}\left(\sum_{\mathrm{i}=1}^{\kappa}\overline{x}_{\mathrm{i}}^2-\frac{\left(\sum\overline{x}_i\right)^2}{\kappa}\right)=\frac{\mathrm{SSA}}{\kappa\text{-}1} \qquad (12.3)$$

Όπου $\overline{x}_{\mathrm{i}}$ είναι η μέση τιμή του δείγματος i και SSA είναι το άθροισμα των αποκλίσεων των μέσων τιμών των δειγμάτων του παράγοντα από τον γενικό μέσο όρο . Ό στατιστικός λόγος F παίρνει την μορφή:

$$F=\frac{\dfrac{SSA}{\kappa-1}}{\dfrac{\mathrm{SSE}}{\mathrm{N}\text{-}\kappa}}=\frac{\mathrm{MSA}}{\mathrm{MSE}} \qquad (12.4)$$

Το MSA λέγεται μέσο τετράγωνο των δειγμάτων και δίνεται από την σχέση

$$MSA=\frac{SSA}{\kappa-1} \qquad (12.5)$$

και MSE λέγεται μέσο τετράγωνο των σφαλμάτων δίνεται από την σχέση

$$MSE=\frac{SSE}{\mathrm{N}-\kappa} \qquad (12.6)$$

.

Η ολική μεταβλητότητα (ολικό άθροισμα τετραγώνων) συμβολίζεται με SST, έχει β.ε = N-1 ,και ισούται με το άθροισμα του αθροίσματος τετραγώνων των δειγμάτων SSA με β.ε =κ-1 ,και του αθροίσματος τετραγώνων των σφαλμάτων SSE με β.ε = N-κ.

SST = SSA + SSE (12.7)

12.3 Ανάλυση διασποράς κατά ένα παράγοντα(one way ANOVA)

Η ανάλυση διασποράς κατά ένα παράγοντα πραγματοποιείται όταν έχουμε μια ποιοτική μεταβλητή με κ τυχαία δείγματα από κ ανεξάρτητους πληθυσμούς. Τα δείγματα δεν είναι ανάγκη να έχουν τον ίδιο αριθμό

παρατηρήσεων. Κύριος στόχος είναι να διερευνήσουμε κατά πόσο επιδρά ο παράγοντας στην εξαρτημένη μεταβλητή y.
Ο πίνακας ανάλυσης διασποράς κατά ένα παράγοντα είναι:

Πηγή Μεταβλητότητας (διασποράς)	Άθροισμα τετραγώνων SS	Β.ε	Μέσο άθροισμα τετραγώνων MS	Συνάρτηση ελέγχου F
Μεταξύ δειγμάτων	SSA	K-1	$MSA = \frac{SSA}{\kappa - 1}$	F= $\frac{MSA}{MSE}$
Μέσα στα δείγματα	SSE	N – κ	$MSE = \frac{SSE}{N - \kappa}$	
Σύνολο	SST	N -1		

Παράδειγμα (12.1)
Η απόδοση σε γάλα (kg/24h) μιας προβατίνας που έχει γεννήσει, υπολογίζεται ζυγίζοντας το νεογνό πριν και μετά το θηλασμό. Πάρθηκαν δείγματα από τρεις διαφορετικούς πληθυσμούς (φυλές) προβάτων και τα αποτελέσματα ήταν ως εξής:

Φυλή	1	2	3	4	5	6	7
A_1	3,2	3,4	4,1	2,8	2,9		
A_2	2,4	2,7	1,8	3,2	3,4	2,6	
A_3	3,9	4,2	3,6	2,8	3,4	3,7	3,5

Υπάρχει διαφορά μεταξύ των τριών φυλών ως προς την απόδοση γάλακτος σε επίπεδο σημαντικότητας 5%; Να διατυπώστε με ακρίβεια τις υποθέσεις που ελέγχετε.
(Δίδονται: SSA =2,6 και SST =6,5)
Λύση:
Από τον πίνακα των πειραματικών δεδομένων βλέπουμε ότι έχουμε πάνω από δύο πληθυσμούς, και θέλουμε να εξετάσουμε αν υπάρχει διαφορά μεταξύ τους ως προς την μέση τιμή απόδοσης γάλακτος. Η πιο κατάλληλη μέθοδος αξιολόγησης σε τέτοια περίπτωση είναι η ανάλυση διασποράς με ένα παράγοντα.
Διατυπώνουμε τις υποθέσεις:

$H_0 : \mu_1 = \mu_2 = \mu_3$
δηλαδή ότι δεν υπάρχουν διαφορές μεταξύ των μέσων αποδόσεων των τριών φυλών .
H_1 :τουλάχιστον μία από τις μέσες αποδόσεις μ_i είναι διαφορετική.
Έχουμε $n_1 = 5$, $n_2 = 6$, $n_3 = 7$, άρα $N = 5 + 6 + 7 = 18$ και $\kappa = 3$.
Το κριτήριο ελέγχου είναι :

$$\text{F}=\frac{MSA}{MSE} \text{ όπου } \quad \text{MSA} = \frac{SSA}{\kappa - 1}, \text{ και } \ \text{MSE} = \frac{\text{SSE}}{\text{N - κ}}$$

Ξέρουμε όμως ότι ισχύει: SST = SSA + SSE (12.7)
Από τα δεδομένα έχουμε SSA = 2,6 και SST =6,5 άρα
6,5=2,6 +SSE $\Rightarrow$SSE = 3,9

$$\text{MSA} = \frac{SSA}{\kappa - 1} = \frac{2,6}{3-1} = 1,3$$

$$\text{MSE} = \frac{\text{SSE}}{\text{N - κ}} = \frac{3,9}{18-3} = 0,26$$

$$\text{F}=\frac{MSA}{MSE} = \frac{1,3}{0,26} = 5$$

Συγκρίνουμε την πειραματική τιμή του F με την θεωρητική τιμή F για β.ε. (κ -1, N-κ) σε επίπεδο 5%. Η τιμή $F_{\kappa-1,\text{N}-\kappa} = F_{2,15(0,05)} = 3,68$

$$\text{F} = 5 > F_{2,15(0,05)} = 3,68$$

Ή H_0 απορρίπτεται σε επίπεδο σημαντικότητας 5%, που σημαίνει ότι υπάρχει διαφορά μεταξύ των τριών φυλών ως προς την απόδοση γάλακτος σε επίπεδο σημαντικότητας (α =0,05).

12.4 Ανάλυση διασποράς με δύο παράγοντες(two way ANOVA)

Η ανάλυση διασποράς κατά δύο παράγοντες πραγματοποιείται όταν θέλουμε να εξετάσουμε αν δύο παράγοντες επηρεάζουν την εξαρτημένη μεταβλητή. Οι δύο αυτοί παράγοντες που επηρεάζουν την εξαρτημένη μεταβλητή είναι δυνατό να αλληλεπιδρούν μεταξύ τους, ή να μη αλληλεπιδρούν.

12.4.1 Ανάλυση διασποράς με δύο παράγοντες χωρίς αλληλεπίδραση

με μια παρατήρηση σε κάθε περίπτωση(r=1).
Ο πίνακας ανάλυσης διασποράς είναι:

Πηγή Μεταβλητότητας (διασποράς)	Άθροισμα τετραγώνων SS	Β.ε	Μέσο άθροισμα τετραγώνων MS	Συνάρτηση ελέγχου F
Παράγοντας Α	SSA	Κ-1	$MSA = \frac{SSA}{\kappa - 1}$	$F_A = \frac{MSA}{MSE}$
Παράγοντας Β	SSB	λ-1	$MSB = \frac{SSB}{\lambda - 1}$	$F_B = \frac{MSB}{MSE}$
Σφάλμα	SSE	(κ-ι)(λ-1)	$MSE = \frac{SSE}{(\kappa - 1)(\lambda - 1)}$	
Σύνολο	SST	Ν-1		

$$SST = SSA + SSB + SSE \qquad (12.8)$$

Παράδειγμα (12.2)
Οι ακόλουθες πληροφορίες αντιπροσωπεύουν κομμάτια παραχθέντα καθημερινά από 5 υπαλλήλους που χρησιμοποιούνε 4 διαφορετικούς τύπους μηχανών. Ελέγξτε σε επίπεδο σημαντικότητας α = 0.05 εάν:
i) Η παραγωγικότητα είναι ίδια για τους 4 τύπους μηχανών (Α);
ii) Εάν οι 5 υπάλληλοι διαφέρουν σε σχέση με την παραγωγικότητα (Β);

	Τύποι μηχανής			
Υπάλληλοι	Α1	Α2	Α3	Α4
Β1	44	38	47	36
Β2	46	40	52	43
Β3	34	36	44	32
Β4	43	38	46	33
Β5	38	42	49	39

Δίνονται: SSA=338, SSB=160, SST=574

Λύση:
Από τον πίνακα των πειραματικών δεδομένων βλέπουμε ότι έχουμε πάνω από δύο πληθυσμούς, και δύο παράγοντες, και θέλουμε να εξετάσουμε αν η παραγωγικότητα είναι ίδια για τους 4 τύπους μηχανών (Α), και εάν οι 5

υπάλληλοι διαφέρουν σε σχέση με την παραγωγικότητα (Β).Άρα έχουμε πρόβλημα ανάλυσης διασποράς με δύο παράγοντες χωρίς αλληλεπίδραση.

Οι υποθέσεις είναι:

H_{0A}: Η παραγωγικότητα είναι ίδια για τους 4 τύπους μηχανών (Α)

H_{1A}: Η παραγωγικότητα δεν είναι ίδια για τους 4 τύπους μηχανών (Α)

H_{0B}: Οι 5 υπάλληλοι δεν διαφέρουν σε σχέση με την παραγωγικότητα (Β)

H_{1B} :Οι 5 υπάλληλοι διαφέρουν σε σχέση με την παραγωγικότητα (Β)

Ο πίνακας ανάλυσης διασποράς με δύο παράγοντες χωρίς αλληλεπίδραση είναι:

Πηγή Μεταβ (διασπ)	Άθρ τετρ SS	Β.ε	Μέσο άθροισμα τετραγώνων MS	Συνάρ Ελέγχ F
Παράγ Α	SSA	Κ-1	$MSA = \frac{SSA}{\kappa - 1}$	$F_A = \frac{MSA}{MSE}$
Παράγ Β	SSB	λ-1	$MSB = \frac{SSB}{\lambda - 1}$	$F_B = \frac{MSB}{MSE}$
Σφάλμα	SSE	(κ-ι)(λ-1)	$MSE = \frac{SSE}{(\kappa - 1)(\lambda - 1)}$	
Σύνολο	SST	N-1		

Για τον πίνακα ανάλυσης διασποράς υπολογίζουμε πρώτα το SSE.

SST = SSA+ SSB + SSE $\Rightarrow$ SSE = SST - SSA - SSB =574-338-160= 76

Αντικαθιστούμε τα πειραματικά δεδομένα στο παραπάνω πίνακα

Πηγή	SS	Β.ε	MS	F
Παράγ Α	338	4-1=3	$MSA = \frac{338}{3} = 118,67$	$F_A = \frac{118,67}{6,33} = 18,64$
Παράγ Β	160	5-1=4	$MSB = \frac{160}{4} = 40$	$F_B = \frac{40}{6,33} = = 6,32$
Σφάλμα	76	3.4=12	$MSE = \frac{76}{12} = 6,33$	
Σύνολο	574	19		

Για τον παράγοντα Α(η παραγωγικότητα) συγκρίνουμε την τιμή F_A= 18,64 με τιμή $F_{3,12(0,05)} = 3,49$.

Επειδή F_A =18,64 > $F_{3,12(0,05)}$= 3,49 απορρίπτουμε την H_{0A} και δεχόμαστε την H_{1A} .Δηλαδή η παραγωγικότητα δεν είναι ίδια για τους 4 τύπους μηχανών (Α) .
Για τον παράγοντα Β (υπάλληλοι) συγκρίνουμε την τιμή F_B= 6,32 με την τιμή $F_{4,12(0,05)} = 3,26$.

Επειδή F_B = 6,32 > $F_{4,12(0,05)}$ = 3,26 , απορρίπτουμε την H_{0B} και δεχόμαστε την H_{1AB} .Δηλαδή Οι 5 υπάλληλοι διαφέρουν σε σχέση με την παραγωγικότητα (Β)

Παράδειγμα (12.3)
Σε τυριά τύπου τελεμέ εφαρμόστηκαν τρεις διαφορετικές επεξεργασίες παρασκευής(Α) και αφέθηκαν να ωριμάσουν για πέντε μήνες. Κάθε μήνα υπολογιζόταν η συγκέντρωση του μη πρωτεϊνικού αζώτου (εκφρασμένο ως προς ολικό άζωτο ΝΠΝ/ΤΝ) ανά τελεμέ. Παρασκευάστηκαν 3 τεμάχια τελεμέ, ένα για κάθε μέθοδο παρασκευής που ακολούθως τεμαχίστηκαν σε πέντε υποτεμάχια ένα για κάθε μήνα . Τα αποτελέσματα του πειράματος φαίνονται στο παρακάτω πίνακα:

Β \ Α	Παραδοσ A_1	Με υπερδιήθ & θέρμανση A_2	Με υπερδιήθ A_3
1 μήνα B_1	11,5	21,1	17,3
2 μήνες B_2	14,4	23,3	17,5
3 μήνες B_3	16,1	24,2	29,5
4 μήνες B_4	19,5	25,6	22,9
5 μήνες B_5	20,7	27,6	24,6

Να εξετάσετε σε επίπεδο σημαντικότητας α=0,05 εάν η συγκέντρωση του μη πρωτεϊνικού αζώτου επηρεάζεται από τον χρόνο ωρίμανσης του τυριού (Β), και από την επεξεργασία παρασκευής (Α) του τελεμέ. (Δίδονται: SSA=156, SSB=120, SST=284).
Λύση:
Από τον πίνακα των πειραματικών δεδομένων βλέπουμε ότι έχουμε πάνω από δύο πληθυσμούς, και δύο παράγοντες, και θέλουμε να εξετάσουμε αν η

συγκέντρωση του μη πρωτεϊνικού αζώτου επηρεάζεται από την επεξεργασία παρασκευής (A) του τελεμέ και από τον χρόνο ωρίμανσης του τυριού (B). Άρα έχουμε πρόβλημα ανάλυσης διασποράς με δύο παράγοντες χωρίς αλληλεπίδραση.
Οι υποθέσεις είναι:
H_{0A}: Η συγκέντρωση του μη πρωτεϊνικού αζώτου είναι ίδια για τις 3 μεθόδους παρασκευής (A) .
H_{1A} : Η συγκέντρωση του μη πρωτεϊνικού αζώτου δεν είναι ίδια για τις 3 μεθόδους παρασκευής (A) .
H_{0B}: Η συγκέντρωση του μη πρωτεϊνικού αζώτου είναι ίδια για τους 5 μήνες ωρίμανσης B.
H_{1B}: Η συγκέντρωση του μη πρωτεϊνικού αζώτου δεν είναι ίδια για τους 5 μήνες ωρίμανσης B.
Ο πίνακας ανάλυσης διασποράς με δύο παράγοντες χωρίς αλληλεπίδραση είναι:

Πηγή Μεταβλ (διασπ)	Άθρ τετρ SS	Β.ε	Μέσο άθροισμα τετραγώνων MS	Συνάρ ελέγχ F
Παράγ A	SSA	κ-1	$\mathrm{MSA} = \dfrac{SSA}{\kappa - 1}$	$F_A = \dfrac{MSA}{MSE}$
Παράγ B	SSB	λ-1	$\mathrm{MSB} = \dfrac{SSB}{\lambda - 1}$	$F_B = \dfrac{MSB}{MSE}$
Σφάλμα	SSE	(κ-ι)(λ-1)	$MSE = \dfrac{SSE}{(\kappa - 1)(\lambda - 1)}$	
Σύνολο	SST	N-1		

Για τον πίνακα ανάλυσης διασποράς υπολογίζουμε πρώτα το SSE.
SST = SSA+ SSB + SSE $\Rightarrow$ SSE = SST - SSA - SSB =284-156-120= 8.
Αντικαθιστούμε τα πειραματικά δεδομένα στο παραπάνω πίνακα :

Πηγή Μεταβ (διασπ)	Άθρ τετρ SS	Β.ε	Μέσο άθροισμα τετραγ MS	Συνάρτηση ελέγχου F
Παράγ A	156	3-1=2	$MSA=\frac{156}{2}=78$	$F_A=\frac{78}{1}=78$
Παράγ B	120	5-1=4	$MSB=\frac{120}{4}=30$	$F_B=\frac{30}{1}=30$
Σφάλμα	8	2.4=8	$MSE=\frac{8}{8}=1$	
Σύνολο	284	14		

Για τον παράγοντα A (μέθοδος παρασκευής) συγκρίνουμε την τιμή F_A= 78 με τιμή $F_{2,14(0,05)}=3{,}74$.

$F_A=78 > F_{2,14(0,05)}=3{,}74$

Απορρίπτουμε την H_{0A} και δεχόμαστε την H_{1A}. Δηλαδή η συγκέντρωση του μη πρωτεϊνικού αζώτου δεν είναι ίδια για τις 3 μεθόδους.

Για τον παράγοντα B(χρόνος ωρίμανσης) συγκρίνουμε την τιμή F_B= 30 με τιμή $F_{4,14(0,05)}=3{,}11$.

$F_B=30 > F_{4,14(0,05)}=3{,}11$ απορρίπτουμε την H_{0B} και δεχόμαστε την H_{1B}.

Δηλαδή η συγκέντρωση του μη πρωτεϊνικού αζώτου δεν είναι ίδια για τους 5 μήνες ωρίμανσης.

12.4.2 Ανάλυση διασποράς με δύο παράγοντες με αλληλεπίδραση και με περισσότερο από μια παρατήρηση σε κάθε περίπτωση ($r \rangle 1$).Ο πίνακας ανάλυσης διασποράς είναι:

Πηγή Μεταβ (διασπ)	Άθρ τετρ SS	Β.ε	Μέσο άθρ τετρ MS	Συνάρ ελέγχ F
Παράγ Α	SSA	Κ-1	$\mathrm{MSA} = \dfrac{SSB}{\kappa - 1}$	$F_A = \dfrac{MSA}{MSE}$
Παράγ Β	SSB	λ-1	$\mathrm{MSB} = \dfrac{SSB}{\lambda - 1}$	$F_B = \dfrac{MSB}{MSE}$
Αλληλεπ ΑΒ	SSAB	(κ-ι)(λ-1)	$MSAB = \\ = \dfrac{SSAB}{(\kappa - 1)(\lambda - 1)}$	$F_{AB} = \\ = \dfrac{MSAB}{MSE}$
Σφάλμα	SSE	Κλ(r-1)	$MSE = \dfrac{SSE}{\kappa\lambda(\mathrm{r} - 1)}$	
Σύνολο	SST	Κλr-1		

SST = SSA+ SSB + SSAB + SSE

Παράδειγμα (12.4)

Τα παρακάτω στοιχεία αποτελούν μετρήσεις μικροβιακών πληθυσμών που αναπτύχθηκαν σε δοκιμαστικούς σωλήνες για 3 ημέρες σε διαφορετικά επίπεδα επώασης (θερμοκρασίας) και με βάση 2 θρεπτικών συστατικών (υποστρωμάτων)

Μικροβιακός πληθυσμός (αριθμός ανά ml θρεπτικού υγρού)

	Θερμοκρασία (Α)				
Υπόστρωμα (Β)	5^o	15^o	25^o	35^o	45^o
Β1	4.7 5.0 5.0	4.8 5.7 6.1	15.7 16.6 13.7	21.6 19.0 20.5	5.4 3.9 4.4
Β2	6.2 5.5 6.6	6.2 6.4 6.9	17.1 17.4 16.8	18.7 19.0 18.7	3.8 3.5 2.6

Να εξετάσετε σε επίπεδο σημαντικότητας α=0.05 εάν η ανάπτυξη των μικροβίων επηρεάζεται από: ι)τη θερμοκρασία(Α), ιι)το υπόστρωμα(Β). ιιι) Επίσης να εξετάσετε εάν υπάρχει αλληλεπίδραση των 2 παραγόντων.
(Δίνονται SSA=1220, SSB=0.5, SSAB=16, SST=1246.5
Λύση:

Από τον πίνακα των πειραματικών δεδομένων βλέπουμε ότι έχουμε πάνω από δύο πληθυσμούς, και δύο παράγοντες, και θέλουμε να εξετάσουμε αν η ανάπτυξη των μικροβίων επηρεάζεται από τους δύο παράγοντες ξεχωριστά ,αλλά και αν υπάρχει αλληλεπίδραση μεταξύ των δύο παραγόντων. Άρα έχουμε πρόβλημα ανάλυσης διασποράς με δύο παράγοντες και αλληλεπίδραση. **Υποθέτουμε** :
H_{0A} : ο παράγοντας Α (θερμοκρασία) δεν επηρεάζει την ανάπτυξη των μικροβίων.
H_{1A} : ο παράγοντας Α (θερμοκρασία) επηρεάζει την ανάπτυξη των μικροβίων.
H_{0B} : ο παράγοντας Β (υπόστρωμα) δεν επηρεάζει την ανάπτυξη των μικροβίων.
H_{1B} : ο παράγοντας Β (υπόστρωμα) επηρεάζει την ανάπτυξη των μικροβίων.
H_{0AB}: δεν υπάρχει αλληλεπίδραση μεταξύ θερμοκρασίας Α και υποστρώματος Β.
H_{1AB}: υπάρχει αλληλεπίδραση μεταξύ θερμοκρασία Α και υπόστρωμα Β
Ο πίνακας ανάλυσης διασποράς με δύο παράγοντες με αλληλεπίδραση είναι

$\frac{MSAB}{MSE}$ Πηγή Μεταβ (διασπ)	Άθρ τετρ SS	Β.ε	Μέσο άθρ τετρ MS	Συνάρ ελέγχ F
Παράγ Α	SSA	Κ-1	$MSA = \frac{SSB}{\kappa - 1}$	$F_A = \frac{MSA}{MSE}$
Παράγ Β	SSB	λ-1	$MSB = \frac{SSB}{\lambda - 1}$	$F_B = \frac{MSB}{MSE}$
Αλληλεπ ΑΒ	SSAB	(κ-ι)(λ-1)	$MSAB = = \frac{SSAB}{(\kappa-1)(\lambda-1)}$	$F_{AB} = = \frac{MSAB}{MSE}$

Σφάλμα	SSE	Κλ(r-1)	$MSE = \frac{SSE}{\kappa\lambda(r-1)}$	
Σύνολο	SST	Κλr-1		

Για τον πίνακα ανάλυσης διασποράς υπολογίζουμε πρώτα το SSE.
SST = SSA+ SSB + SSAB + SSE $\Rightarrow$ SSE = SST - SSA - SSB - SSAB $\Rightarrow$
SSE =1246.5 – 1220 - 0.5 – 16 = 10
Αντικαθιστούμε τα πειραματικά δεδομένα στο παραπάνω πίνακα:

Πηγή διασπ	Άθρο τετραγ. SS	Β.ε	Μέσο άθροισμα τετραγώνων MS	Συνάρτηση ελέγχου F
Παράγ A	1220	5-1= 4	$MSA = \frac{1220}{4} = 305$	$F_A = \frac{305}{0,5} = 610$
Παράγ B	0.5	2-1=1	$MSB = \frac{0,5}{1} = 0,5$	$F_B = \frac{0,5}{0,5} = 1$
AB	16	4.1= 4	$MSAB = \frac{16}{4} = 4$	$F_{AB} = \frac{4}{0,5} = 8$
Σφάλ	10	20	$MSE = \frac{10}{20} = 0,5$	
Σύνολ	1246.5	29		

Για τον παράγοντα Α(θερμοκρασία) συγκρίνουμε την τιμή F_A =610 με την θεωρητική τιμή $F_{4,20,0.05}$ = 2,87

Επειδή $F_A = 610 \rangle F_{4,20,0,05} = 2,87$ απορρίπτουμε την H_{0A} και δεχόμαστε την H_{1A} δηλαδή δεχόμαστε ότι η θερμοκρασία) επηρεάζει την ανάπτυξη των μικροβίων.
Για τον παράγοντα Β (υπόστρωμα) συγκρίνουμε την τιμή F_B = 1 με την θεωρητική τιμή $F_{1,20,0.05}$= 4,35.
Επειδή $F_B = 1 \langle F_{1,20,0.05} = 4,35$ δεχόμαστε την H_{0B} δηλαδή ο παράγοντας Β (υπόστρωμα) δεν επηρεάζει την ανάπτυξη των μικροβίων.
Για την αλληλεπίδραση μεταξύ των δύο παραγόντων συγκρίνουμε την τιμή F_{AB} = 8 με την θεωρητική τιμή $F_{4,20,0,05}$= 2,87.

Επειδή $F_{AB} = 8 \rangle\ F_{4,20,0,05} = 2,87$ απορρίπτουμε την H_{0AB} και δεχόμαστε τη H_{1AB} , δηλαδή υπάρχει αλληλεπίδραση μεταξύ θερμοκρασίας Α και υποστρώματος Β.

Παράδειγμα (12.5)
Τρεις ποικιλίες σίτου (Α) δοκιμάστηκαν με 3 είδη λιπασμάτων: φωσφορική αμμωνία, θειική αμμωνία και καθόλου λίπανση (Β). οι αποδώσεις σε κάθε συνδυασμό ποικιλίας και λιπάσματος φαίνονται στον παρακάτω πίνακα:

Λίπανση (Β)	A_1	A_2	A_3
Φωσφορική αμμωνία	112 152 128 118	112 48 81 108	134 128 112 116
Θειική αμμωνία	168 80 116 144	61 98 98 58	125 110 106 110
Μάρτυρας (καμία λίπανση)	106 128 84 68	97 66 86 92	62 87 60 99

Ελέγξτε σε επίπεδο σημαντικότητας 5%, εάν υπάρχει διαφορά στις αποδόσεις του σίτου που να οφείλεται: ι) Στις διαφορετικές ποικιλίες , ιι)στην διαφορετική λίπανση, ιιι)στην αλληλεπίδραση ποικιλίας και λιπάσματος. (Δίνονται: SSA=6743.4, SSB=4481.7,SSAB=2789.4, SST=27584.5).

Λύση:
Από τον πίνακα των πειραματικών δεδομένων βλέπουμε ότι έχουμε πάνω από δύο πληθυσμούς, και δύο παράγοντες, και θέλουμε να εξετάσουμε αν υπάρχει διαφορά στις αποδόσεις που οφείλεται στους δύο παράγοντες ξεχωριστά ,αλλά και στην αλληλεπίδραση μεταξύ τους. Άρα έχουμε πρόβλημα ανάλυσης διασποράς με δύο παράγοντες και αλληλεπίδραση.
Οι υποθέσεις :
H_{0A} : ο παράγοντας Α (ποικιλία) δεν επηρεάζει την απόδοση.
H_{1A} : ο παράγοντας Α (ποικιλία) επηρεάζει την απόδοση
H_{0B} : ο παράγοντας Β (λίπασμα) δεν επηρεάζει την απόδοση.
H_{1B} : ο παράγοντας Β (λίπασμα) επηρεάζει την απόδοση
H_{0AB}: δεν υπάρχει αλληλεπίδραση μεταξύ ποικιλίας Α και λιπάσματος Β
H_{1AB}: υπάρχει αλληλεπίδραση μεταξύ ποικιλίας Α και λιπάσματος Β
Για τον πίνακα ανάλυσης διασποράς υπολογίζουμε πρώτα το SS_E.
$SST= SSA+ SSB + SSAB + SSE \quad \Rightarrow SSE = SST - SSA - SSB - SSAB \Rightarrow$
$SSE=27584,5 -6743,4 -4481,7 -2789,4 = 13570$

Ο πίνακας ανάλυσης διασποράς με δύο παράγοντες με αλληλεπίδραση είναι:

Πηγή Μεταβ (διασπ)	Άθρ τετρ SS	Β.ε	Μέσο άθρ τετρ MS	Συνάρ ελέγχ F
Παράγ Α	SSA	Κ-1	$MSA = \frac{SSB}{\kappa - 1}$	$F_A = \frac{MSA}{MSE}$
Παράγ Β	SSB	λ-1	$MSB = \frac{SSB}{\lambda - 1}$	$F_B = \frac{MSB}{MSE}$
Αλληλεπ ΑΒ	SSAB	(κ-ι)(λ-1)	$MSAB = = \frac{SSAB}{(\kappa-1)(\lambda-1)}$	$F_{AB} = = \frac{MSAB}{MSE}$
Σφάλμα	SSE	Κλ(r-1)	$MSE = \frac{SSE}{\kappa\lambda(r-1)}$	
Σύνολο	SST	Κλr-1		

Αντικαθιστούμε τα πειραματικά δεδομένα στο παραπάνω πίνακα:

	.SS	Β.ε	MS	F
Α	6743,4	3-1=2	$MSA = \frac{6743,4}{2} = 3371,7$	$F_A = \frac{3371,7}{502,59} = = 6,71$
Β	4481,7	3-1=2	$MSB = \frac{4481,7}{2} = 2240,85$	$F_B = \frac{2240,85}{502,59} = 4,46$
ΑΒ	2789,4	2.2=4	$MSAB = \frac{2789,4}{4} = 697,35$	$F_{AB} = \frac{697,35}{502,59} = 1,39$
Σφάλμα	13570	3.3.3 =27	$MSE = \frac{13570}{27} = = 502,59$	
Σύνολο	27584,5	3.3.41=35		

Για τον παράγοντα Α(ποικιλία) συγκρίνουμε την τιμή F_A =6,71 με την θεωρητική τιμή $F_{2,27,0,05}$= 3,35.
Επειδή F_A =6,71 $\rangle$ $F_{2,27,0,05}$= 3,35 απορρίπτουμε την H_{0A} και δεχόμαστε την H_{1A} δηλαδή δεχόμαστε ότι η ποικιλία επιδρά στην απόδοση .
Για τον παράγοντα Β (λίπασμα) συγκρίνουμε την τιμή F_B = 4,46 με την θεωρητική τιμή $F_{2,27,0,05}$= 3,35.
Επειδή F_B = 4,46 $\rangle$ $F_{2,27,0,05}$= 3,35 απορρίπτουμε την H_{0B} και δεχόμαστε την $H_{1\mathbf{B}}$ δηλαδή δεχόμαστε ότι το λίπασμα επιδρά στην απόδοση .
Για την αλληλεπίδραση μεταξύ των δύο παραγόντων συγκρίνουμε την τιμή F_{AB} = 1,39 με την θεωρητική τιμή $F_{4,27,0,05}$= 2,73.
Επειδή F_{AB} = 1,39 $\langle$ $F_{4,27,0,05}$= 2,73 δεχόμαστε την H_{0AB} δηλαδή δεν υπάρχει αλληλεπίδραση μεταξύ των δύο παραγόντων.

ΑΣΚΗΣΕΙΣ

12.1. Εκτελέστηκαν 24 πειράματα από δύο για κάθε συνδυασμό 4 αντιδραστηρίων B_i και 3 καταλυτών A_i. Σε κάθε περίπτωση μετρήθηκε ο ρυθμός παραγωγής της παραγόμενης ουσίας. Κωδικοποιημένες τιμές παραγωγής δίδονται στον παρακάτω πίνακα:

B_i / $\mathbf{A_i}$	$\mathbf{A_1}$		$\mathbf{A_2}$		$\mathbf{A_3}$	
$\mathbf{B_1}$	4	6	11	7	5	9
$\mathbf{B_2}$	6	4	13	15	9	7
$\mathbf{B_3}$	13	15	15	9	13	13
$\mathbf{B_4}$	12	12	12	14	7	9

Να ελεγχθεί σε επίπεδο σημαντικότητας α= 0,05, εάν υπάρχει επίδραση του είδους του καταλύτη, του είδους του αντιδραστηρίου, καθώς και εάν υπάρχει αλληλεπίδραση αυτών στον ρυθμό παραγωγής της παραγόμενης ουσίας. (Δίδονται: SSA=48, SSB=120, SSAB=84, SST=300.)

12.2 Σε μια μελέτη για αντοχή καλωδίων παρασκευάσθηκαν καλώδια με τρις διαφορετικές μεθόδους (Α), Ι, ΙΙ, ΙΙΙ, σε τέσσερα διαφορετικά μεγέθη (Β): μικρό ($M_ι$), μεσαίο προς μικρό ($MM_ι$), μεσαίο προς μεγάλο (ΜΜε), και μεγάλο (Με).Οι αντοχές που μετρήθηκαν είχαν ως εξής:

A / B	Μι	ΜΜι		ΜΜε		Με
Ι	341	344	344	340	342	343
	342					343
ΙΙ	339	341	341	340	339	340
	340					341
ΙΙΙ	340	339	337	337	339	336
	338					339

Σε επίπεδο σημαντικότητας 5% μπορεί κανείς με αυτά τα δεδομένα να ισχυριστεί ότι :(α) οι μέσες αντοχές καλωδίων κατασκευασμένων με τις τρις μεθόδους διαφέρουν μεταξύ τους; (β) το μήκος του καλωδίου επηρεάζει την αντοχή; (γ) υπάρχει αλληλεπίδραση μεταξύ μεθόδου και μήκους στην αντοχή των καλωδίων;

(Δίδονται: SSA= 72,4, SSB= 7,1 , SSAB= 10 , SST= 104)

12.3 Τα παρακάτω δεδομένα αποτελούν μετρήσεις της πρωτεΐνης του πλάσματος χελωνών ενός συγκεκριμένου είδους. Μετρήθηκε η πρωτεΐνη σε αρσενικές και θηλυκές χελώνες, αφού είχαν διατραφεί κανονικά, μετά από νηστεία 10 ημερών και μετά από νηστεία 20 ημερών.

Φύλο (B)	Διατροφή (A_1)	Νηστεία 10 ημερών (A_2)	Νηστεία 20 ημερών (A_3)
Αρσενικά	42.8	42.4	38.9
	43.1	42.2	40.3
	40.4	40.8	37.5
	46.6	45.9	34.2
Θηλυκά	42.2	42.4	39.7
	38.7	38.1	35.8
	35.3	34.3	32.3
	40.5	40.1	37.3

Ελέγξτε σε επίπεδο σημαντικότητας 5% εάν υπάρχει διαφοροποίηση της πρωτεΐνης του πλάσματος που να οφείλεται: στη διατροφή ή τη νηστεία 10 ή 20 ημερών (A), το φύλο (B), καθώς και αν υπάρχει αλληλεπίδραση των παραγόντων αυτών.
(Δίνονται : SSA = 85.5, SSB = 61.5, SSAB = 9, SST = 300).

ΠΑΡΑΡΤΗΜΑ

ΣΤΑΤΙΣΤΙΚΟΙ ΠΙΝΑΚΕΣ

ΠΙΝΑΚΑΣ 1 : τυποποιημένης κανονικής κατανομής

ΠΙΝΑΚΑΣ 2 : κατανομή t

ΠΙΝΑΚΑΣ 3 : κατανομή X^2

ΠΙΝΑΚΑΣ 4 : κατανομή F

ΠΙΝΑΚΑΣ 1

Τιμές των πιθανοτήτων $\boldsymbol{\Phi}(z) = \boldsymbol{P}(\boldsymbol{Z} \le z) = \boldsymbol{P}(\boldsymbol{Z} < z)$ της τυποποιημένης κανονικής κατανομής $\boldsymbol{N}(\mathbf{0},\mathbf{1})$ για $z \ge \mathbf{0}$. Για $z < \mathbf{0}$ ισχύει $\boldsymbol{\Phi}(z) = \mathbf{1} - \boldsymbol{\Phi}(-z)$.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84850	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92786	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99897	0.99900

α	0.0005	0.001	0.005	0.01	0.025	0.05	0.10
Z_α	3.29	3.09	2.576	2.326	1.960	1.645	1.282

ΠΙΝΑΚΑΣ 2

Τιμών $t_{\nu;a}$ της t_ν-κατανομής ώστε $P(T_\nu > t_{\nu;a}) = P(T_\nu \geq t_{\nu;a}) = a$.

ν	α = 0.10	α = 0.05	α = 0.025	α = 0.01	α = 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
∞	1.282	1.645	1.960	2.326	2.576

ΠΙΝΑΚΑΣ 3

Των τιμών $\chi^2_{v;1-a}$ της χ^2 κατανομής για τις οποίες

$$P(X < \chi^2_{v\,;\,1\text{-}a}) = P(X \le \chi^2_{v\,;\,1\text{-}a}) = a\,.$$

v	*α* = 0.005	*α* = 0.01	*α* = 0.025	*α* = 0.05	*α* = 0.10
1	0.0000393	0.0001571	0.0009821	0.0039321	0.0157908
2	0.0100251	0.0201007	0.0506356	0.102587	0.210720
3	0.0717212	0.114832	0.215795	0.351846	0.584375
4	0.206990	0.297110	0.484419	0.710721	1.063623
5	0.411740	0.554300	0.831211	1.145476	1.61031
6	0.675727	0.872085	1.237347	1.63539	2.20413
7	0.989265	1.239043	1.68987	2.16735	2.83311
8	1.344419	1.646482	2.17973	2.73264	3.48954
9	1.734926	2.087912	2.70039	3.32511	4.16816
10	2.15585	2.55821	3.24697	3.94030	4.86518
11	2.60321	3.05347	3.81575	4.57481	5.57779
12	3.07382	3.57056	4.40379	5.22603	6.30380
13	3.56503	4.10691	5.00874	5.89186	7.04150
14	4.07468	4.66043	5.62872	6.57063	7.78953
15	4.60094	5.22935	6.26214	7.26094	8.54675
16	5.14224	5.81221	6.90766	7.96164	9.31223
17	5.69724	6.40776	7.56418	8.67176	10.0852
18	6.26481	7.01491	8.23075	9.39046	10.8649
19	6.84398	7.63273	8.90655	10.1170	11.6509
20	7.43386	8.26040	9.59083	10.8508	12.4426
21	8.03366	8.89720	10.28293	11.5913	13.2396
22	8.64272	9.54249	10.9823	12.3380	14.0415
23	9.26042	10.19567	11.6885	13.0905	14.8479
24	9.88623	10.8564	12.4011	13.8484	15.6587
25	10.5197	11.5240	13.1197	14.6114	16.4734
26	11.1603	12.1981	13.8439	15.3791	17.2919
27	11.8076	12.8786	14.5733	16.1513	18.1138
28	12.4613	13.5648	15.3079	16.9279	18.9392
29	13.1211	14.2565	16.0471	17.7083	19.7677
30	13.7867	14.9535	16.7908	18.4926	20.5992
40	20.7065	22.1643	24.4331	26.5093	29.0505
50	27.9907	29.7067	32.3574	34.7642	37.6886
60	35.5346	37.4848	40.4817	43.1879	46.4589
70	43.2752	45.4418	48.7576	51.7393	55.3290
80	51.1720	53.5400	57.1532	60.3915	64.2778
90	59.1963	61.7541	65.6466	69.1260	73.2912
100	67.3276	70.0648	74.2219	77.9295	82.3581

ΠΙΝΑΚΑΣ 3

Των τιμών $\chi^2_{v;a}$ της χ^2 κατανομής για τις οποίες

$$P(X > \chi^2_{v;a}) = P(X \geq \chi^2_{v;a}) = a$$

v	*A* = 0.10	*α* = 0.05	*α* = 0.025	*α* = 0.01	*α* = 0.005
1	2.70554	3.84146	5.02389	6.63490	7.87944
2	4.60517	5.99147	7.37776	9.21034	10.5966
3	6.25139	7.81473	9.34840	11.3449	12.8381
4	7.77944	9.48773	11.1433	13.2767	14.8602
5	9.23635	11.0705	12.8325	15.0863	16.7496
6	10.6446	12.5916	14.4494	16.8119	18.5476
7	12.0170	14.0671	16.0128	18.4753	20.2777
8	13.3616	15.5073	17.5346	20.0902	21.9550
9	14.6837	16.9190	19.0228	21.6660	23.5893
10	15.9871	18.3070	20.4831	23.2093	25.1882
11	17.2750	19.6751	21,9200	24.7250	26.7569
12	18.5494	21.0261	23.3367	26.2170	28.2995
13	19.8119	22.3621	24.7356	27.6883	29.8194
14	21.0642	23.6848	26.1190	29.1413	31.3193
15	22.3072	24.9958	27.4884	30.5779	32.8013
16	23.5418	26.2962	28.8454	31.9999	34.2672
17	24.7690	27.5871	30.1910	33.4087	35.7185
18	25.9894	28.8693	31.5264	34.8053	37.1564
19	27.2036	30.1435	32.8523	36.1908	38.5822
20	28.4120	31.4104	34.1696	37.5662	39.9968
21	29.6151	32.6705	35.4789	38.9321	41.4010
22	30.8133	33.9244	36.7807	40.2894	42.7956
23	32.0069	35.1725	38.0757	41.6384	44.1813
24	33.1963	36.4151	39.3641	42.9798	45.5585
25	34.3816	37.6525	40.6465	44.3141	46.9278
26	35.5631	38.8852	41.9232	45.6417	48.2899
27	36.7412	40.1133	43.1944	46.9630	49.6449
28	37.9159	41.3372	44.4607	48.2782	50.9933
29	39.0875	42.5569	45.7222	49.5879	52.3356
30	40.2560	43.7729	46.9792	50.8922	53.6720
40	51.8050	55.7585	59.3417	63.6907	66.7659
50	63.1671	67.5048	71.4202	76.1539	79.4900
60	74.3970	79.0819	83.2976	88.3794	91.9517
70	85.5271	90.5312	95.0231	100.425	104.215
80	96.5782	101.879	106.629	112.329	116.321
90	107.565	113.145	118.136	124.116	128.299
100	118.498	124.342	129.561	135.807	140.169

ΠΙΝΑΚΑΣ 4

Τιμές $F_{v_1,v_2;a}$ της F κατανομής για τις οποίες $P(X > F_{v_1,v_2;a}) = P(X \geq F_{v_1,v_2;a}) = a \quad (a = 0.01)$. Για τα α - κάτω ποσοστιαία σημεία $F_{v_1,v_2;1-a}$ ισχύει η σχέση $F_{v_1,v_2;1-a} = 1\big/F_{v_2,v_1;a}$.

v_2 \ v_1	1	2	3	4	5	6	7	8	9
1	4052	4999.5	5403	5625	5764	5859	5928	5982	6022
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41

ΠΙΝΑΚΑΣ 4 (συνέχεια)

Τιμές $F_{v_1,v_2;a}$ της F κατανομής για τις οποίες

$$P(X > F_{v_1,v_2;a}) = P(X \geq F_{v_1,v_2;a}) = a \qquad (a = 0.01).$$

Για τα α - κάτω ποσοστιαία σημεία $F_{v_1,v_2;1-a}$ ισχύει η σχέση

$$F_{v_1,v_2;1-a} = 1 \big/ F_{v_2,v_1;a} .$$

v_2 \ v_1	10	12	15	20	24	30	40	60	120	∞
1	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

ΠΙΝΑΚΑΣ 4 (συνέχεια)

Τιμές $F_{v_1,v_2;a}$ της F κατανομής για τις οποίες

$P(X > F_{v_1,v_2;a}) = P(X \geq F_{v_1,v_2;a}) = a \quad (a = 0.05)$.Για τα α - κάτω ποσοστιαία σημεία $F_{v_1,v_2;1-a}$ ισχύει η σχέση $F_{v_1,v_2;1-a} = 1/F_{v_2,v_1;a}$.

v_2 \ v_1	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

ΠΙΝΑΚΑΣ 4 (συνέχεια)

Τιμές $F_{\nu_1,\nu_2;a}$ της F κατανομής για τις οποίες

$P(X > F_{\nu_1,\nu_2;a}) = P(X \geq F_{\nu_1,\nu_2;a}) = a \quad (a = 0.05)$. Για τα α - κάτω ποσοστιαία σημεία $F_{\nu_1,\nu_2;1-a}$ ισχύει η σχέση $F_{\nu_1,\nu_2;1-a} = 1/F_{\nu_2,\nu_1;a}$.

ν_2 \ ν_1	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
1	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Βιβλιογραφία

1. **Αθανασόπουλος, Δ. (1980).** Στατιστική, εκδόσεις Σταμούλης.

2. **Αθανασόπουλος, Δ. και Μπένος, Β(1990).** Εφαρμογές Στατιστικής Τόμος Α′, εκδόσεις Σταμούλης.

3. **Αρμενάκης, Α. (1991).** Σημειώσεις Στατιστικής.

4. **Δαμιανός, Χ. και Κούτρας, Μ(1988).** Εισαγωγή στην Στατιστική Μέρος Ι, εκδόσεις Συμμετρίας.

5. **Ζαχαροπούλου, Χ (1993).**Στατιστική , μέθοδοι – εφαρμογές Τόμος Α′, εκδοτικό κέντρο βόρειας Ελλάδας.

6. **Καραγεώργος**, **Δ.** Στατιστική (Περιγραφική & Επαγωγική) εκδόσεις Σαββάλας

7. **Κονόμος , Γ. (2009).** Πιθανότητες & Στατιστική Γκιούρδας Εκδοτική

8. **Κούνια, Σ. Κουλυβά – Μαχαίρα, Φ. Μπαγιάτη, και Κ. Μπόρα – Σέντα, Ε.** (2003) Εισαγωγή στην Στατιστική. Εκδόσεις Χριστοδουλίδη.

9.**Τζώνου, Α . και Κατσουγιάννη, Κ.(1997).** Ασκήσεις Βιοστατιστικής . Εκδόσεις Συμμετρία , Αθήνα

10.**Τριχόπουλος , Δ. , Τζώνου, Α . και Κατσουγιάννη, Κ. (2000).** Βιοστατιστική, εκδόσεις Παρισιανού, Αθήνα.

11.**Χαλικιάς , Ι. (2003) .** Στατιστική – Μέθοδοι Ανάλυσης.