



ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΥΠΟΥΡΓΕΙΟ ΕΣΩΤΕΡΙΚΩΝ
ΕΘΝΙΚΗ ΣΧΟΛΗ ΔΗΜΟΣΙΑΣ ΔΙΟΙΚΗΣΗΣ & ΑΥΤΟΔΙΟΙΚΗΣΗΣ

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ
ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ



Με τη χρηματοδότηση
της Ευρωπαϊκής Ένωσης
NextGenerationEU

Η ΔΙΑΧΕΙΡΙΣΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ Η ΣΤΑΤΙΣΤΙΚΗ ΕΠΕΞΕΡΓΑΣΙΑ ΤΟΥΣ ΓΙΑ ΛΗΨΗ ΑΠΟΦΑΣΕΩΝ ΣΤΗ ΔΗΜΟΣΙΑ ΔΙΟΙΚΗΣΗ

Συγγραφέας: Στέφανος Γιακουμάτος
Αξιολογήτης: Αναστασία Παπαστυλιανού
Συντονιστής: Σοφία Δημουλά

Αναλυτική περιγραφή

- 1ο. Εισαγωγικές έννοιες στη Στατιστική.
- 2ο. Μέθοδοι συλλογής των δεδομένων.
- 3ο. Οργάνωση και σύνοψη των δεδομένων.
- 4ο. Περιγραφικά στατιστικά μέτρα.
- 5ο. Γραφήματα.
- 6ο. Κατανομές και κατανομές δειγματοληψίας.
- 7ο. Διαστήματα εμπιστοσύνης και έλεγχος υποθέσεων .
- 8ο. Συσχέτιση μεταβλητών και Παλινδρόμηση

Όταν θα έχετε ολοκληρώσει την μελέτη της ενότητας θα είστε σε θέση

- να επεξεργάζεστε τα δεδομένα σας στο Excel.
- να οργανώνετε και συνοψίζετε τα δεδομένα σας με τη χρήση πινάκων, περιγραφικών στατιστικών και γραφημάτων.
- να κατασκευάζετε και ερμηνεύετε διαστήματα εμπιστοσύνης για το πληθυσμιακό μέσο όρο.
- να κάνετε έλεγχο υποθέσεων για ένα δείγμα.
- να αναγνωρίζετε την σημασία της στατιστικής ανάλυσης στην λήψη αποφάσεων
- να εφαρμόζετε τις αποκτηθείσες γνώσεις στην Δημόσια Διοίκηση

Εισαγωγικές έννοιες στη Στατιστική

Κεφάλαιο 1

Εισαγωγή

Η Στατιστική είναι η επιστήμη που ασχολείται με τη συλλογή, παρουσίαση και ανάλυση δεδομένων με στόχο την λήψη αποφάσεων

Στο πρώτο κεφάλαιο θα εξετάσουμε τις βασικές έννοιες που διέπουν τη στατιστική επιστήμη, τις μεθόδους συλλογής και ανάλυσης δεδομένων, καθώς και τις διαφορετικές κλίμακες μέτρησης που χρησιμοποιούνται.

Θα αναλύσουμε επίσης τη διάκριση μεταξύ Περιγραφικής και Επαγωγικής Στατιστικής, τους τύπους μεταβλητών και τη σημασία των μεταδεδομένων στη δημόσια διοίκηση.

Στατιστική και στατιστικές συναρτήσεις

Η Στατιστική είναι η επιστήμη που ασχολείται με τη συλλογή, παρουσίαση και ανάλυση δεδομένων με στόχο την λήψη αποφάσεων.

Σήμερα, η χρήση των στατιστικών συναρτήσεων που περιγράφουν τα αριθμητικά χαρακτηριστικά μεγάλων ομάδων δεδομένων είναι απαραίτητη για κάθε σύγχρονο διοικητικό στέλεχος.



Συλλογή δεδομένων

Τι είδους δεδομένα πρέπει να συλλέγονται και πόσα δεδομένα χρειάζονται για επαρκή τεκμηρίωση;



Οργάνωση δεδομένων

Πώς πρέπει να οργανώσουμε τα δεδομένα που συλλέξαμε;



Ανάλυση και συμπεράσματα

Πώς μπορούμε να αναλύσουμε τα δεδομένα και να βγάλουμε συμπεράσματα;



Αξιολόγηση

Πώς μπορούμε να αξιολογήσουμε την ισχύ των συμπερασμάτων;

Ορισμοί που πρέπει να γνωρίζετε

Στατιστική

Το σύνολο των μεθόδων και των θεωριών που μπορούν να χρησιμοποιηθούν για την συλλογή, ανάλυση των δεδομένων και ερμηνεία των αποτελεσμάτων.

Δεδομένα

Συλλογές πληροφοριών, κυρίως γεγονότων και αριθμών. Μπορεί να είναι αριθμητικά (π.χ. ηλικίες ατόμων) ή κατηγορικά, δηλαδή μη αριθμητικά (π.χ. φύλο ατόμων).

Μεταβλητή

Χαρακτηριστικό ή ιδιότητα που μπορεί να ποικίλλει μεταξύ των εμφανίσεων αντικειμένων, ατόμων ή καταστάσεων (π.χ. αριθμός αιτήσεων σε πρωτόκολλο).

- Η **παρατήρηση** ή **παρατηρούμενη τιμή** είναι μία τιμή μιας μεταβλητής που έχει πράγματι μετρηθεί. Για παράδειγμα, εάν 1/5/2025 υποβλήθηκαν 58 αιτήσεις, αυτό αποτελεί παρατήρηση της μεταβλητής «αριθμός αιτήσεων».
- Μια μεταβλητή συμβολίζεται συνήθως με κεφαλαίο γράμμα (π.χ. X , Y), ενώ μία παρατήρηση με το αντίστοιχο πεζό (π.χ. x , y). Αριθμητικοί υποδείκτες προστίθενται για διάκριση παρατηρήσεων (π.χ. x_1 , x_2).

Περιγραφική Στατιστική και Επαγωγική Στατιστική

Περιγραφική Στατιστική

Περιλαμβάνει τις στατιστικές τεχνικές που χρησιμοποιούνται για να οργανώσουν, συνοψίσουν και περιγράψουν τα δεδομένα.

- Μέσος όρος (μέση τιμή)
- Τυπική απόκλιση
- Διακύμανση
- Πίνακες κατανομής συχνοτήτων

Δεν μπορεί να γενικεύσει τα αποτελέσματα στο επίπεδο του πληθυσμού.

Επαγωγική Στατιστική

Περιλαμβάνει τις στατιστικές μεθόδους και τεχνικές που χρησιμοποιούν στατιστικά από τυχαία δείγματα για να γενικεύσουν ή να βγάλουν συμπεράσματα σχετικά με έναν πληθυσμό.

Είναι πιο περίπλοκη από την περιγραφική στατιστική και απαιτεί κατανόηση της τυχαιότητας που περιβάλλει τα δεδομένα και του τρόπου συλλογής τους.

Απαιτεί δείγματα που επιλέγονται τυχαία από τον πληθυσμό, ώστε να είναι αμερόληπτοι εκπρόσωποι του.

Οι Μεταβλητές

Οι μεταβλητές στη Στατιστική αφορούν χαρακτηριστικά ή ιδιότητες ατόμων, αντικειμένων ή καταστάσεων. Είναι σημαντικό να γνωρίζετε το είδος της μεταβλητής που θέλετε να αναλύσετε, διότι αυτό καθορίζει και το είδος των τεχνικών στατιστικής ανάλυσης που μπορούν να εφαρμοστούν.



Κατηγορικές (ποιοτικές)

Οι τιμές τους αντιπροσωπεύουν κατηγορίες

- Ερωτήσεις τύπου ναι/όχι
- Περιφέρεια οργανισμού
- Πολιτική προτίμηση



Αριθμητικές (ποσοτικές)

Παίρνουν τιμές που αντιπροσωπεύουν μετρήσιμη ποσότητα

- Μηνιαίο εισόδημα
- Ηλικία
- Βάρος

Αριθμητικές (ποσοτικές)



Διακριτές

Οι τιμές καθορίζονται από διαδικασία καταμέτρησης

- Αριθμός προϊόντων
- Αριθμός παιδιών



Συνεχείς

Προκύπτουν από διαδικασία μέτρησης

- Χρόνος αναμονής
- Απόσταση

Κλίμακες μέτρησης

Οι μεταβλητές μπορούν να κατηγοριοποιηθούν με βάση την κλίμακα μέτρησης που ανήκουν. Οι κλίμακες μέτρησης παρέχουν έναν τρόπο συστηματικής σκέψης σχετικά με τα χαρακτηριστικά και ιδιότητες της μέτρησης των τιμών των μεταβλητών.

Κάθε κλίμακα μέτρησης έχει ορισμένες μαθηματικές ιδιότητες που καθορίζουν τις κατάλληλες εφαρμογές των στατιστικών διαδικασιών. Σημειώστε ότι μια κλίμακα μέτρησης υψηλότερου επιπέδου περιέχει όλες τις μαθηματικές ιδιότητες που έχει μια κλίμακα μέτρησης χαμηλότερου επιπέδου συν κάτι περισσότερο.



Ονομαστική

Κατηγοριοποίηση χωρίς ιεράρχηση



Τακτική

Κατάταξη με σειρά ιεράρχησης



Διαστήματος

Ίσες αποστάσεις μεταξύ τιμών



Αναλογίας

Ίσες αποστάσεις και απόλυτο μηδέν

Μεταδεδομένα στη δημόσιο διοίκηση

Τα μεταδεδομένα (metadata) είναι δομημένες πληροφορίες που περιγράφουν, εξηγούν ή διευκολύνουν την κατανόηση, χρήση και διαχείριση άλλων δεδομένων ή πηγών πληροφορίας. Στη δημόσια διοίκηση, τα μεταδεδομένα λειτουργούν ως κρίσιμος παράγοντας για τη διαλειτουργικότητα των συστημάτων και την ασφαλή ανταλλαγή δεδομένων μεταξύ υπηρεσιών.



Στοιχεία μεταδεδομένων

Τα μεταδεδομένα συνοδεύουν τα δεδομένα μίας έρευνας, παρέχοντας σημαντικές πληροφορίες για την κατανόηση, χρήση και αξιοποίησή τους. Αποτελούν ουσιαστικά "δεδομένα για τα δεδομένα" και είναι απαραίτητα για την ορθή ερμηνεία των στατιστικών πληροφοριών.

Περιγραφή
Τίτλος και περιγραφή του συνόλου
δεδομένων

Εμπιστευτικότητα
Πολιτικές προστασίας προσωπικών
δεδομένων



Δομή

Τύποι μεταβλητών και περιοχές τιμών

Χρονολογία

Προέλευση και περίοδος συλλογής

Υπευθυνότητα

Υπεύθυνος φορέας ή άτομο και
δικαιώματα πρόσβασης

Παράδειγμα μεταδεδομένων στατιστικής έρευνας

Η ΕΛΣΤΑΤ παρέχει ένα χαρακτηριστικό παράδειγμα μεταδεδομένων για την «[Έρευνα εισοδήματος και συνθηκών διαβίωσης- EUSILC](#)». Τα μεταδεδομένα αυτά περιλαμβάνουν πλήρεις πληροφορίες για την έρευνα και τη μεθοδολογία της.

Στοιχείο μεταδεδομένων	Περιεχόμενο
Υπηρεσία	ΕΛΣΤΑΤ
Στοιχεία επικοινωνίας	Όνομα υπευθύνου, τηλέφωνα, email
Χρονολογία	Ημερομηνία ανάρτησης και ενημέρωσης
Στατιστική παρουσίαση	Πηγές, τύπος, συχνότητα και μέθοδος συλλογής
Μεταβλητές	Περιγραφή βασικών μεταβλητών (π.χ. "φτώχεια", "εισόδημα")
Πληθυσμός	Όλα τα ιδιωτικά νοικοκυριά της Ελλάδος
Στατιστικές μονάδες	Νοικοκυριά και άτομα που διαμένουν σε αυτά

Επιπλέον, περιλαμβάνονται πληροφορίες για τη χρονική κάλυψη, τις νομικές πράξεις, την πολιτική εμπιστευτικότητας, την πολιτική ανακοινώσεων, καθώς και την ακρίβεια, την αξιοπιστία και τα σφάλματα της έρευνας.

Ασκήσεις Κεφαλαίου 1

1) Προσδιορίστε εάν η δεδομένη τιμή είναι στατιστική ή παράμετρος. Το 36% όλων των ιδιοκτητών σκύλων μαζεύουν τις ακαθαρσίες από το σκυλί τους.

α) Στατιστική β) Παράμετρος

2) Προσδιορίστε εάν η μεταβλητή «Ο χρόνος που χρειάζεται ένας υπολογιστής για να ολοκληρώσει μια εργασία» είναι διακριτή ή συνεχής μεταβλητή.

α) συνεχής β) διακριτή

3) Σύμφωνα με μια έρευνα 1020 ενηλίκων που πραγματοποιήθηκε Ηνωμένες Πολιτείες, το 44% των συμμετεχόντων δήλωσε ότι πλένουν τα χέρια τους αφού χρησιμοποιήσουν κάποιο δημόσιο μέσο συγκοινωνίας

A. Προσδιορίστε το δείγμα και τον πληθυσμό.

B. Είναι η τιμή 44% στατιστικό ή παράμετρος;

Γ. Ποια κλίμακα μέτρησης θα πρέπει να χρησιμοποιηθεί για την μεταβλητή «Πλένετε τα χέρια σας εφόσον χρησιμοποιήσατε κάποιο δημόσιο μέσο συγκοινωνίας»; (ονομαστική, τακτική, διαστήματος, αναλογίας)

Μέθοδοι Συλλογής Δεδομένων στη Στατιστική

Κεφάλαιο 2

Εισαγωγή

Η Στατιστική μελετά τα χαρακτηριστικά ενός πληθυσμού συλλέγοντας δεδομένα από αυτόν. Όταν συλλέγουμε δεδομένα από όλα τα μέλη ενός πληθυσμού, αυτό ονομάζεται απογραφή. Ωστόσο, επειδή ο αριθμός των μελών σε έναν πληθυσμό μπορεί να είναι τεράστιος ή άπειρος, τις περισσότερες φορές η έρευνα διεξάγεται με ένα προσεκτικά επιλεγμένο δείγμα.

Ο τρόπος επιλογής του δείγματος από έναν πληθυσμό ονομάζεται δειγματοληψία. Υπάρχουν δύο βασικοί τύποι: η μη πιθανοτική δειγματοληψία και η δειγματοληψία πιθανότητας. Στην πιθανοτική δειγματοληψία, όλα τα μέλη του πληθυσμού έχουν γνωστές πιθανότητες επιλογής, ενώ στη μη πιθανοτική δειγματοληψία, τα μέλη επιλέγονται χωρίς μηχανισμό τυχαιότητας.

Πλεονεκτήματα της Δειγματοληψίας

Χαμηλότερο Κόστος

Η χρήση δείγματος αντί για απογραφή μειώνει σημαντικά το κόστος συλλογής και ανάλυσης των δεδομένων, καθιστώντας την έρευνα οικονομικά εφικτή.

Εξοικονόμηση Χρόνου

Η συλλογή δεδομένων από ένα δείγμα απαιτεί λιγότερο χρόνο σε σύγκριση με την απογραφή ολόκληρου του πληθυσμού, επιτρέποντας ταχύτερα αποτελέσματα.

Πρακτικότητα

Δεν είναι πάντα δυνατό ή εφικτό να αποκτήσουμε πρόσβαση σε κάθε μέλος του πληθυσμού και να έχουμε τη συγκατάθεση όλων για συμμετοχή στην έρευνα.

Μέθοδοι Μη Πιθανοτικής Δειγματοληψίας



Δειγματοληψία Ευκολίας

Οι ερευνητές χρησιμοποιούν στο δείγμα οποιονδήποτε είναι πρόθυμος να συμμετάσχει ή επιλέγουν μονάδες με βάση την εύκολη προσβασιμότητα. Παράδειγμα: συλλογή δεδομένων μέσω φίλων στα μέσα κοινωνικής δικτύωσης.



Ποσοστιαία Δειγματοληψία

Ο ερευνητής ορίζει κατηγορίες του πληθυσμού και συλλέγει δείγμα ώστε κάθε κατηγορία να αντιπροσωπεύεται με το αντίστοιχο ποσοστό της στον πληθυσμό. Χρησιμοποιείται ευρέως στην έρευνα αγοράς.



Δειγματοληψία Κρίσης

Οι μονάδες του δείγματος επιλέγονται με βάση την προσωπική κρίση ή την εμπειρία του ερευνητή. Χρησιμοποιείται όταν ο ερευνητής έχει εξειδικευμένες γνώσεις και στην ποιοτική έρευνα.

Μέθοδοι Πιθανοτικής Δειγματοληψίας



Απλή Τυχαία Δειγματοληψία

Κάθε μέλος του πληθυσμού έχει την ίδια πιθανότητα να επιλεγεί.
Απαιτεί πλήρες δειγματοληπτικό πλαίσιο και τυχαία επιλογή.



Συστηματική Δειγματοληψία

Επιλέγεται ένα τυχαίο σημείο εκκίνησης και στη συνέχεια κάθε "k" μέλος του πληθυσμού συμπεριλαμβάνεται στο δείγμα.



Στρωματοποιημένη Δειγματοληψία

Ο πληθυσμός χωρίζεται σε ομοιογενείς υποομάδες (στρώματα) και λαμβάνεται τυχαίο δείγμα από κάθε στρώμα ανάλογα με το μέγεθός του.



Δειγματοληψία Συστάδων

Επιλέγονται τυχαία ομάδες (συστάδες) και όλα τα μέλη των επιλεγμένων συστάδων περιλαμβάνονται στο δείγμα.

Απλή Τυχαία και Συστηματική Δειγματοληψία

Απλή Τυχαία Δειγματοληψία

Ένα απλό τυχαίο δείγμα μεγέθους n δημιουργείται έτσι ώστε όλα τα δείγματα με το ίδιο μέγεθος να έχουν τις ίδιες πιθανότητες επιλογής. Για παράδειγμα, σε ένα πανεπιστήμιο με 10.000 φοιτητές, μπορούμε να δώσουμε σε κάθε φοιτητή έναν μοναδικό αριθμό και να επιλέξουμε τυχαία 200 αριθμούς.

Συστηματική Δειγματοληψία

Στη συστηματική δειγματοληψία, επιλέγεται ένα τυχαίο σημείο εκκίνησης και στη συνέχεια κάθε " k " μέλος του πληθυσμού επιλέγεται. Για παράδειγμα, για να επιλέξουμε 200 φοιτητές από 10.000, επιλέγουμε τυχαία έναν αριθμό μεταξύ 1 και 50 ως σημείο εκκίνησης και μετά κάθε 50ό φοιτητή.

Στρωματοποιημένη και Δειγματοληψία Συστάδων



Στρωματοποιημένη

Ο πληθυσμός χωρίζεται σε ομοιογενείς υποομάδες και λαμβάνεται τυχαίο δείγμα από κάθε στρώμα. Αν το 60% του πληθυσμού είναι γυναίκες, τότε το 60% του δείγματος θα είναι γυναίκες.



Συστάδων

Επιλέγονται τυχαία ομάδες και όλα τα μέλη των επιλεγμένων ομάδων περιλαμβάνονται στο δείγμα. Παράδειγμα: επιλογή 100 γυμνασίων από 1521 και συμπερίληψη όλων των μαθητών τους.



Με **Excel**

Το Excel προσφέρει εργαλεία για τυχαία δειγματοληψία μέσω του πρόσθετου "Ανάλυση Δεδομένων", επιτρέποντας την επιλογή τυχαίου ή περιοδικού δείγματος από μια περιοχή δεδομένων.

Σφάλματα στην Ποσοτική Έρευνα



Σφάλμα Δειγματοληψίας

Διαφορές μεταξύ στατιστικών μέτρων του δείγματος και παραμέτρων του πληθυσμού



Σφάλματα Μέτρησης

Λάθη στην καταγραφή των δεδομένων



Σφάλματα Μη Απόκρισης

Όταν σημαντικό ποσοστό δεν απαντά ή παραλείπει ερωτήσεις



Σφάλματα Επεξεργασίας

Λάθη κατά την εισαγωγή, κωδικοποίηση ή ανάλυση δεδομένων

Το ποσοστό απόκρισης είναι κρίσιμο και ορίζεται ως ο λόγος του πλήθους των ολοκληρωμένων απαντήσεων προς το συνολικό δείγμα. Οι ελλείπουσες τιμές αποτελούν συχνό φαινόμενο και η σωστή διαχείρισή τους είναι απαραίτητη για την αποφυγή μεροληπτικών αποτελεσμάτων.

Σύγκριση Μεθόδων Δειγματοληψίας

Μέθοδος	Πλεονεκτήματα	Μειονεκτήματα	Εφαρμογές
Απλή Τυχαία	Εύκολη κατανόηση, αμερόληπτη	Απαιτεί πλήρες δειγματοληπτικό πλαίσιο	Μικροί, ομοιογενείς πληθυσμοί
Συστηματική	Απλή εφαρμογή, καλή κάλυψη	Κίνδυνος περιοδικότητας	Λίστες, αρχεία δεδομένων
Στρωματοποιημένη	Αντιπροσωπευτικότητα, ακρίβεια	Πολυπλοκότητα, κόστος	Ετερογενείς πληθυσμοί
Συστάδων	Οικονομία, πρακτικότητα	Μεγαλύτερο σφάλμα δειγματοληψίας	Γεωγραφικά διασκορπισμένοι πληθυσμοί

Η επιλογή της κατάλληλης μεθόδου δειγματοληψίας εξαρτάται από τους στόχους της έρευνας, τους διαθέσιμους πόρους, τα χαρακτηριστικά του πληθυσμού και την ανάγκη για γενίκευση των αποτελεσμάτων. Οι πιθανοτικές μέθοδοι προσφέρουν το πλεονέκτημα της γενίκευσης των αποτελεσμάτων στον πληθυσμό, ενώ οι μη πιθανοτικές είναι συχνά πιο οικονομικές και πρακτικές.

Ασκήσεις κεφαλαίου

1) Προσδιορίστε εάν η δεδομένη τιμή είναι στατιστική ή παράμετρος. «Το 36% όλων των ιδιοκτητών σκύλων μαζεύουν τις ακαθαρσίες από το σκυλί τους».

α) Στατιστική β) Παράμετρος

2) Προσδιορίστε εάν η μεταβλητή «Ο χρόνος που χρειάζεται ένας υπολογιστής για να ολοκληρώσει μια εργασία» είναι διακριτή ή συνεχής μεταβλητή.

α) συνεχής β) διακριτή

3) Σύμφωνα με μια έρευνα 1020 ενηλίκων που πραγματοποιήθηκε Ηνωμένες Πολιτείες, το 44% των συμμετεχόντων δήλωσε ότι πλένουν τα χέρια τους αφού χρησιμοποιήσουν κάποιο δημόσιο μέσο συγκοινωνίας

A. Προσδιορίστε το δείγμα και τον πληθυσμό.

B. Είναι η τιμή 44% στατιστικό ή παράμετρος;

Γ. Ποια κλίμακα μέτρησης θα πρέπει να χρησιμοποιηθεί για την μεταβλητή «Πλένετε τα χέρια σας εφόσον χρησιμοποιήσατε κάποιο δημόσιο μέσο συγκοινωνίας»; (ονομαστική, τακτική, διαστήματος, αναλογία

Οργάνωση και Σύνοψη Δεδομένων

Κεφάλαιο 3

Εισαγωγή

Η οργάνωση και σύνοψη δεδομένων αποτελεί θεμελιώδες βήμα στη στατιστική ανάλυση. Μετά τη συλλογή δεδομένων, είναι απαραίτητο να τα οργανώσουμε με συνοπτικό τρόπο ώστε να εξάγουμε χρήσιμα συμπεράσματα για τα χαρακτηριστικά τους. Ένας από τους βασικότερους τρόπους οργάνωσης είναι ο πίνακας κατανομής συχνότητας, ο οποίος παραθέτει όλες τις κατηγορίες ή τιμές μιας μεταβλητής μαζί με τις αντίστοιχες συχνότητες, σχετικές συχνότητες και ποσοστά.

Οργάνωση Κατηγορικών Μεταβλητών

Καταγραφή Κατηγοριών

Καταγράφουμε τις κατηγορίες της μεταβλητής ως γραμμές σε πίνακα και καταμετρούμε τις απαντήσεις κάθε κατηγορίας στη στήλη συχνότητων.

Υπολογισμός Σχετικών Συχνοτήτων

Διαιρούμε τη συχνότητα κάθε κατηγορίας με το σύνολο των απαντήσεων για να βρούμε τις σχετικές συχνότητες.

Υπολογισμός Ποσοστών

Πολλαπλασιάζουμε τις σχετικές συχνότητες με το 100 για να υπολογίσουμε τα ποσοστά κάθε κατηγορίας.

Τομέας	Συχνότητα f	Σχετική συχνότητα	Ποσοστό
Μισθωτός ΔΤ	3	0.15	15
Μισθωτός ΙΤ	6	0.30	30
Ελ. Επαγγελματίας	5	0.25	25
Άνεργος	4	0.20	20
Συνταξιούχος	2	0.10	10
Σύνολο	20	1.00	100

Οργάνωση Διακριτών Αριθμητικών Μεταβλητών



Καταγραφή Τιμών

Καταγράφουμε τις τιμές της διακριτής μεταβλητής ως γραμμές σε πίνακα, ξεκινώντας από τη μικρότερη προς τη μεγαλύτερη.



Καταμέτρηση Συχνοτήτων

Καταμετρούμε πόσες φορές εμφανίζεται κάθε τιμή στο δείγμα και καταγράφουμε το αποτέλεσμα στη στήλη συχνοτήτων.



Υπολογισμός Ποσοστών

Υπολογίζουμε τις σχετικές συχνότητες και τα ποσοστά όπως και στις κατηγορικές μεταβλητές.

Κατανομή συχνότητας διακοπών δικτύου

Αριθμός δικτύου	διακοπών	Συχνότητα f	Σχετική συχνότητα	Ποσοστό
0		9	0.300	30.0
1		10	0.333	33.3
2		5	0.167	16.7
3		3	0.100	10.0
4		2	0.067	6.8
5		0	0.000	0.0
6		1	0.033	3.2
Σύνολο		30	1.000	100

Για παράδειγμα, στην περίπτωση των διακοπών δικτύου internet, η διακριτή αριθμητική μεταβλητή (αριθμός διακοπών) οργανώνεται σε πίνακα με τις τιμές από 0 έως 6, όπου η συχνότητα 0 διακοπών είναι 9 (30%), 1 διακοπής είναι 10 (33,3%), κ.ο.κ.

Οργάνωση Συνεχών Αριθμητικών Μεταβλητών



Επιλογή Αριθμού Τάξεων

Επιλέγουμε τον αριθμό των τάξεων (διαστημάτων), συνήθως μεταξύ 5 και 20, χρησιμοποιώντας τον εμπειρικό τύπο του Sturges.



Υπολογισμός Πλάτους Τάξης

Υπολογίζουμε το πλάτος κάθε τάξης διαιρώντας το εύρος των δεδομένων (μέγιστη - ελάχιστη τιμή) με τον αριθμό των τάξεων.



Δημιουργία Τάξεων

Επιλέγουμε το κάτω όριο της πρώτης τάξης και δημιουργούμε τις υπόλοιπες τάξεις προσθέτοντας το πλάτος τάξης.



Καταγραφή Συχνοτήτων

Καταγράφουμε πόσες τιμές ανήκουν σε κάθε τάξη και υπολογίζουμε τις σχετικές συχνότητες και τα ποσοστά.

Παράδειγμα Κατανομής Συχνότητας Υψών

134

Μέγεθος Δείγματος

Συνολικός αριθμός μετρήσεων ύψους
πατέρων

8

Αριθμός Τάξεων

Επιλεγμένος αριθμός διαστημάτων για την
κατανομή

0.053

Πλάτος Τάξης

Σε μέτρα, μετά από στρογγυλοποίηση προς
τα πάνω

Στο παράδειγμα των υψών των πατέρων, η συνεχής αριθμητική μεταβλητή οργανώνεται σε 8 τάξεις με πλάτος 0,053 μέτρα. Η πλειοψηφία των τιμών (35,82%) βρίσκεται στο διάστημα 1,729-1,782 μέτρα, ενώ το 67,91% του δείγματος έχει ύψος μέχρι 1,782 μέτρα.

Κατανομή συχνότητας για τα ύψη

Τάξεις ύψους		Συχνότητές	Σχετικές συχνότητες	Ποσοστά	Αθροιστικά ποσοστά
1.57	1.623	2	0.0149	1.49	1.49
1.623	1.676	13	0.0970	9.70	11.19
1.676	1.729	28	0.2090	20.90	32.09
1.729	1.782	48	0.3582	35.82	67.91
1.782	1.835	31	0.2313	23.13	91.04
1.835	1.888	7	0.0522	5.22	96.27
1.888	1.941	4	0.0299	2.99	99.25
1.941	1.994	1	0.0075	0.75	100.00
Σύνολο		134	1.0000	100.00	

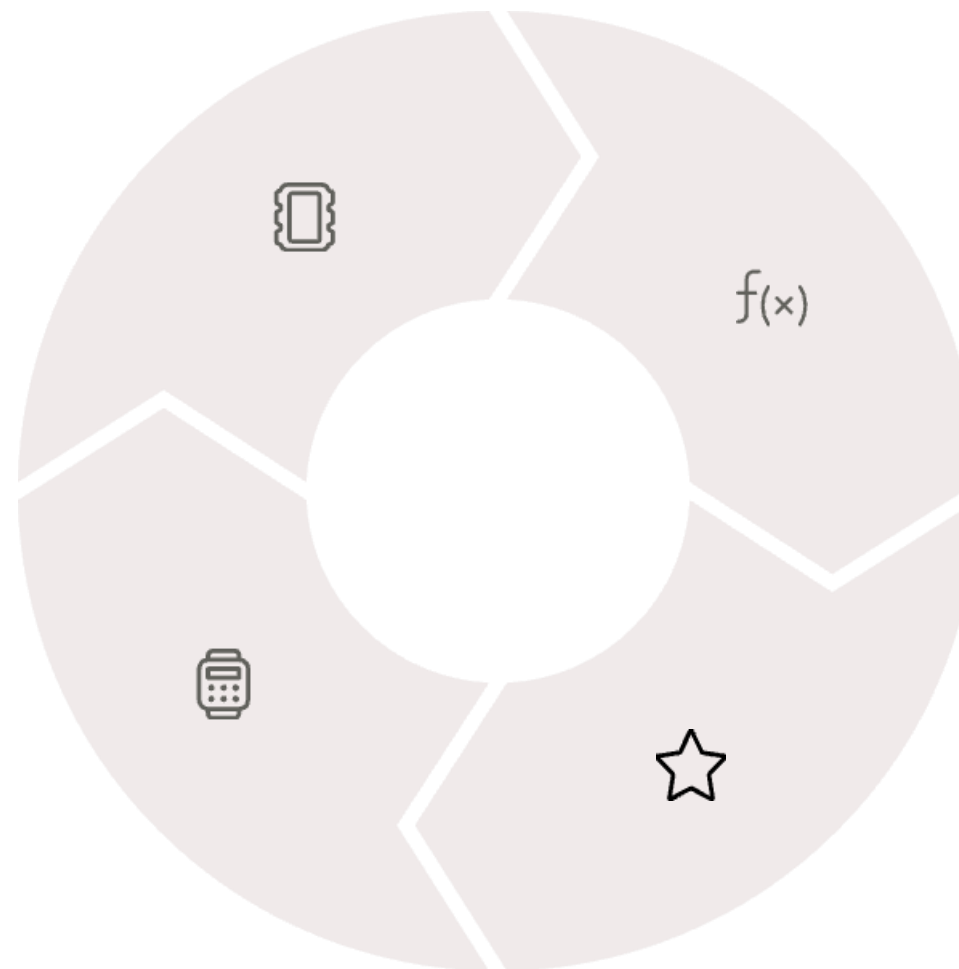
Κατασκευή Πίνακα Κατανομής στο **Excel**

Δημιουργία Τάξεων

Κατασκευάζουμε τις τάξεις σύμφωνα με τα βήματα που παρουσιάστηκαν για τις συνεχείς μεταβλητές.

Υπολογισμός Ποσοστών

Υπολογίζουμε τις σχετικές συχνότητες και τα ποσοστά με απλές πράξεις στο Excel.



Χρήση Συνάρτησης **FREQUENCY**

Χρησιμοποιούμε τη συνάρτηση **FREQUENCY** για να υπολογίσουμε τις συχνότητες των τάξεων.

Χρήση **Data Analysis**

Εναλλακτικά, χρησιμοποιούμε το πρόσθετο **Data Analysis** και επιλέγουμε **Ιστόγραμμα** για αυτόματη δημιουργία πίνακα συχνοτήτων.

Συνάρτηση FREQUENCY

Στο κελί της πρώτης συχνότητας πληκτρολογώ την εντολή **=FREQUENCY(B2:B135, K11:K18)**, όπου

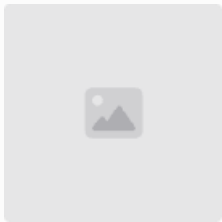
- το πρώτο πεδίο, B2:B135, περιλαμβάνει τα δεδομένα (τα δεδομένα στο αρχείο είναι από το κελί B2 μέχρι το B135), και
- το δεύτερο πεδίο περιλαμβάνει τα άνω όρια των διαστημάτων τάξεων που έχω κατασκευάσει

Με χρήση του πρόσθετου Ανάλυση Δεδομένων

- Στην καρτέλα **Δεδομένα** (Data), κάντε κλικ στην **Ανάλυση Δεδομένων** (Data Analysis).
- Επιλέξτε **Ιστόγραμμα** (Histogram) και πατήστε OK.
- Στο παράθυρο «Ιστόγραμμα»:
 - **Περιοχή εισόδου** (Input Range): Επιλέξτε την περιοχή των δεδομένων σας (π.χ. B2:B135).
 - **Περιοχή κλάσης δεδομένων** (Bin Range): Επιλέξτε την περιοχή με τα άνω όρια των τάξεων που δημιουργήσατε (π.χ. K11:K18).
 - Επιλέξτε **Περιοχή εξόδου** (Output Range) και κάντε κλικ σε ένα κενό κελί όπου θέλετε να εμφανιστεί ο πίνακας συχνοτήτων (π.χ. G1).
 - (Προαιρετικά) Μπορείτε να επιλέξετε **Έξοδος γραφήματος** (Chart Output) αν θέλετε να δείτε και το ιστόγραμμα (η έννοια θα παρουσιαστεί σε κατοπινό κεφάλαιο).
- Πατήστε **OK**.

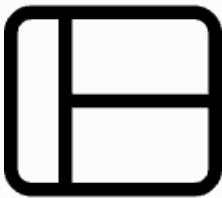
Ανάλυση Πινάκων Συνάφειας

Οι πίνακες συνάφειας χρησιμοποιούνται για την από κοινού παρουσίαση δύο κατηγορικών μεταβλητών. Στο παράδειγμα της σχέσης φύλου και καπνίσματος, ο πίνακας συνάφειας δείχνει ότι από τους 6 άνδρες, οι 2 είναι καπνιστές και οι 4 μη καπνιστές, ενώ από τις 7 γυναίκες, οι 3 είναι καπνίστριες και οι 4 μη καπνίστριες.



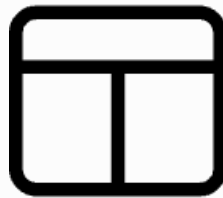
Ποσοστά επί του συνόλου

Δείχνουν το ποσοστό κάθε κελιού ως προς το συνολικό αριθμό παρατηρήσεων



Ποσοστά γραμμής

Υπολογίζονται διαιρώντας τη συχνότητα κάθε κελιού με το άθροισμα της γραμμής



Ποσοστά στήλης

Υπολογίζονται ως ποσοστό κάθε τιμής σε μία στήλη ως προς το άθροισμα της στήλης

Παράδειγμα

Δεδομένα

A/A	Φύλο	Καπνιστής	A/A	Φύλο	Καπνιστής
1	Γυναίκα	Ναι	8	Άνδρας	Ναι
2	Γυναίκα	Ναι	9	Άνδρας	Ναι
3	Άνδρας	Όχι	10	Άνδρας	Όχι
4	Άνδρας	Όχι	11	Γυναίκα	Όχι
5	Άνδρας	Όχι	12	Γυναίκα	Όχι
6	Γυναίκα	Ναι	13	Γυναίκα	Όχι
7	Γυναίκα	Όχι			

Πίνακας συνάφειας

Count of Καπνιστής	Column Labels		
Row Labels	Ναι	Όχι	Grand Total
Άνδρας	2	4	6
Γυναίκα	3	4	7
Grand Total	5	8	13

Δημιουργία Πινάκων Συνάφειας στο Excel



Επιλογή δεδομένων

Επιλέγουμε τα δεδομένα των δύο κατηγορικών μεταβλητών στο φύλλο του Excel



Δημιουργία συγκεντρωτικού πίνακα

Πηγαίνουμε στην καρτέλα "Εισαγωγή" και επιλέγουμε "Συγκεντρωτικός Πίνακας"



Διαμόρφωση πεδίων

Τοποθετούμε τις μεταβλητές στα πεδία "Γραμμές", "Στήλες" και "Τιμές"

4

Υπολογισμός ποσοστών

Από το μενού "Εμφάνιση τιμών ως", επιλέγουμε το είδος ποσοστού που θέλουμε να υπολογίσουμε

Η δημιουργία πινάκων συνάφειας στο Excel είναι μια απλή διαδικασία που μας επιτρέπει να αναλύσουμε τη σχέση μεταξύ δύο κατηγορικών μεταβλητών. Η επιλογή του κατάλληλου τύπου ποσοστού (επί του συνόλου, γραμμής ή στήλης) εξαρτάται από το είδος της ερώτησης που προσπαθούμε να απαντήσουμε με την ανάλυσή μας.

Ασκήσεις κεφαλαίου 3

Χρησιμοποιώντας το αρχείο *Data Family Heights v1.xlsx*, κατασκευάστε τον πίνακα κατανομής συχνότητας για

- το ύψος του γιού (στήλη D) και
- το ύψος της κόρης (στήλη E).

Περιγραφικά Στατιστικά Μέτρα

Κεφάλαιο 4

Εισαγωγή

Τα περιγραφικά στατιστικά μέτρα είναι βασικά εργαλεία της Στατιστικής που χρησιμοποιούνται για να συνοψίσουν και να περιγράψουν ένα σύνολο δεδομένων με απλό και κατανοητό τρόπο. Μας βοηθούν να κατανοήσουμε τη συνολική εικόνα των δεδομένων, δηλαδή, που είναι το «κέντρο» τους, πόσο διαφέρουν μεταξύ τους, και πώς είναι κατανεμημένες οι τιμές τους.

Τα περιγραφικά στατιστικά μέτρα χωρίζονται σε τρεις βασικές κατηγορίες: μέτρα θέσης ή κεντρικής τάσης, μέτρα μεταβλητότητας, και μέτρα μορφής της κατανομής. Αυτά τα μέτρα δίνουν μια πλήρη εικόνα του τρόπου με τον οποίο τα δεδομένα «συμπεριφέρονται» και είναι απαραίτητα για την ανάλυση, σύγκριση και ερμηνεία οποιουδήποτε στατιστικού συνόλου δεδομένων.

Μέτρα Θέσης ή Κεντρικής Τάσης



Αριθμητικός Μέσος Όρος

Άθροισμα όλων των παρατηρήσεων διαιρεμένο με το πλήθος τους



Διάμεσος

Η μεσαία τιμή σε ένα ταξινομημένο σύνολο δεδομένων



Επικρατούσα Τιμή

Η τιμή που εμφανίζεται πιο συχνά στα δεδομένα

Τα μέτρα θέσης χρησιμοποιούνται για να περιγράψουν την «τυπική» ή την «κεντρική» τιμή ενός συνόλου δεδομένων. Ο μέσος όρος είναι το πιο δημοφιλές μέτρο, αλλά επηρεάζεται από ακραίες τιμές. Η διάμεσος είναι ανθεκτική στις ακραίες τιμές και χρησιμοποιείται συχνά για ασύμμετρες κατανομές. Η επικρατούσα τιμή είναι το μόνο μέτρο που μπορεί να χρησιμοποιηθεί για κατηγορικά δεδομένα.

Αριθμητικός Μέσος Όρος

Ο αριθμητικός μέσος όρος είναι το άθροισμα όλων των παρατηρήσεων διαιρεμένο με το πλήθος τους. Συμβολίζεται με $\bar{X}$ και ο τύπος του είναι $\bar{X} = (1/n) \cdot \sum x_i$. Είναι εύκολος στον υπολογισμό και χρησιμοποιεί όλες τις παρατηρήσεις, αλλά επηρεάζεται από ακραίες τιμές.

12.8

Μέσος Όρος Βαθμών

Παράδειγμα μαθητή με βαθμούς 10, 12, 13, 14, 15

1700€

Μέσος Μισθός

Παράδειγμα με μισθούς 800€, 850€, 900€, 950€, 5000€

Ειδικές περιπτώσεις αποτελούν ο σταθμισμένος μέσος όρος που λαμβάνει υπόψη τη βαρύτητα κάθε τιμής, και ο περικομμένος μέσος όρος που αφαιρεί ένα ποσοστό από τις μικρότερες και μεγαλύτερες τιμές για να μειώσει την επίδραση των ακραίων τιμών.

Διάμεσος και Επικρατούσα Τιμή

Διάμεσος

Η διάμεσος είναι η μεσαία τιμή σε ένα ταξινομημένο σύνολο δεδομένων. Περίπου το 50% των τιμών είναι μικρότερες και το 50% μεγαλύτερες από αυτήν.

Η θέση της διαμέσου βρίσκεται ως $(n+1)/2$. Για περιττό αριθμό παρατηρήσεων είναι η μεσαία τιμή, ενώ για άρτιο είναι ο μέσος των δύο κεντρικών τιμών.

Πλεονέκτημα: Δεν επηρεάζεται από ακραίες τιμές.

Επικρατούσα Τιμή

Η επικρατούσα τιμή είναι η τιμή που εμφανίζεται πιο συχνά σε ένα σύνολο δεδομένων. Αντιπροσωπεύει την πιο συχνή τιμή στα δεδομένα.

Δεν έχει μεγάλη πρακτική χρησιμότητα για αριθμητικά δεδομένα, αλλά είναι ιδιαίτερα χρήσιμη για κατηγορικά δεδομένα.

Μειονέκτημα: Μπορεί να μην υπάρχει, ή να υπάρχουν πολλαπλές επικρατούσες τιμές.

Ποσοστιαία Σημεία και Τεταρτημόρια

Τα ποσοστιαία σημεία χωρίζουν τα δεδομένα σε 100 ίσα μέρη. Το p -οστό ποσοστιαίο σημείο είναι η τιμή κάτω από την οποία βρίσκεται περίπου το $p\%$ των παρατηρήσεων και πάνω στην οποία βρίσκεται το $(1-p)\%$ των παρατηρήσεων.

Πρώτο Τεταρτημόριο (Q_1)

25ο ποσοστιαίο σημείο. Το 25% των παρατηρήσεων είναι μικρότερες από αυτό.

Τρίτο Τεταρτημόριο (Q_3)

75ο ποσοστιαίο σημείο. Το 75% των παρατηρήσεων είναι μικρότερες από αυτό.

Διάμεσος (Q_2)

50ο ποσοστιαίο σημείο. Το 50% των παρατηρήσεων είναι μικρότερες από αυτό.

Τα τεταρτημόρια είναι μια ειδική περίπτωση ποσοστιαίων σημείων που χωρίζουν τα δεδομένα σε τέσσερα ίσα μέρη. Παρέχουν μια λεπτομερή εικόνα της κατανομής και είναι ιδιαίτερα ανθεκτικά στις ακραίες τιμές.

Μέτρα Μεταβλητότητας

Εύρος

Η διαφορά μεταξύ της μέγιστης και της ελάχιστης τιμής ενός συνόλου δεδομένων. Δείχνει τη συνολική έκταση των τιμών.

$$\text{Range} = X_{\max} - X_{\min}$$

Ενδοτεταρτημοριακό Εύρος (**IQR**)

Η διαφορά μεταξύ του τρίτου και του πρώτου τεταρτημορίου. Αντιπροσωπεύει το εύρος των μεσαίων 50% των τιμών.

$$\text{IQR} = Q3 - Q1$$

Διακύμανση

Η μέση τετραγωνική απόκλιση των δεδομένων από τον μέσο όρο τους. Χρησιμοποιεί όλες τις τιμές των δεδομένων.

$$s^2 = \sum (x_i - \bar{x})^2 / (n-1)$$

Τυπική Απόκλιση

Η τετραγωνική ρίζα της διακύμανσης. Αντιπροσωπεύει τη μέση απόκλιση των τιμών από τον μέσο όρο.

$$s = \sqrt{s^2}$$

Διακύμανση και Τυπική Απόκλιση



Υπολογισμός Μέσου Όρου

Βρίσκουμε το μέσο όρο των δεδομένων



Υπολογισμός Αποκλίσεων

Αφαιρούμε το μέσο όρο από κάθε τιμή



Τετραγωνισμός

Υψώνουμε στο τετράγωνο τις αποκλίσεις



Διαίρεση

Διαιρούμε με $n-1$ για τη διακύμανση



Τετραγωνική Ρίζα

Εξάγουμε τη ρίζα για την τυπική απόκλιση

- ✓ Η διακύμανση αντιπροσωπεύει τη μέση τετραγωνική απόκλιση των δεδομένων από τον μέσο όρο τους. Είναι από τα πλέον δημοφιλή μέτρα μεταβλητότητας και χρησιμοποιεί όλες τις τιμές των δεδομένων. Το μειονέκτημά της είναι ότι η μονάδα μέτρησης είναι διαφορετική από τα αρχικά δεδομένα.
- ✓ Η τυπική απόκλιση είναι η τετραγωνική ρίζα της διακύμανσης και αντιπροσωπεύει τη μέση απόκλιση των τιμών από τον μέσο όρο. Έχει την ίδια μονάδα μέτρησης με τα δεδομένα, γεγονός που την καθιστά πιο εύκολα κατανοητή και ευρέως χρησιμοποιούμενη.

Μέτρα Σχετικής Μεταβλητότητας

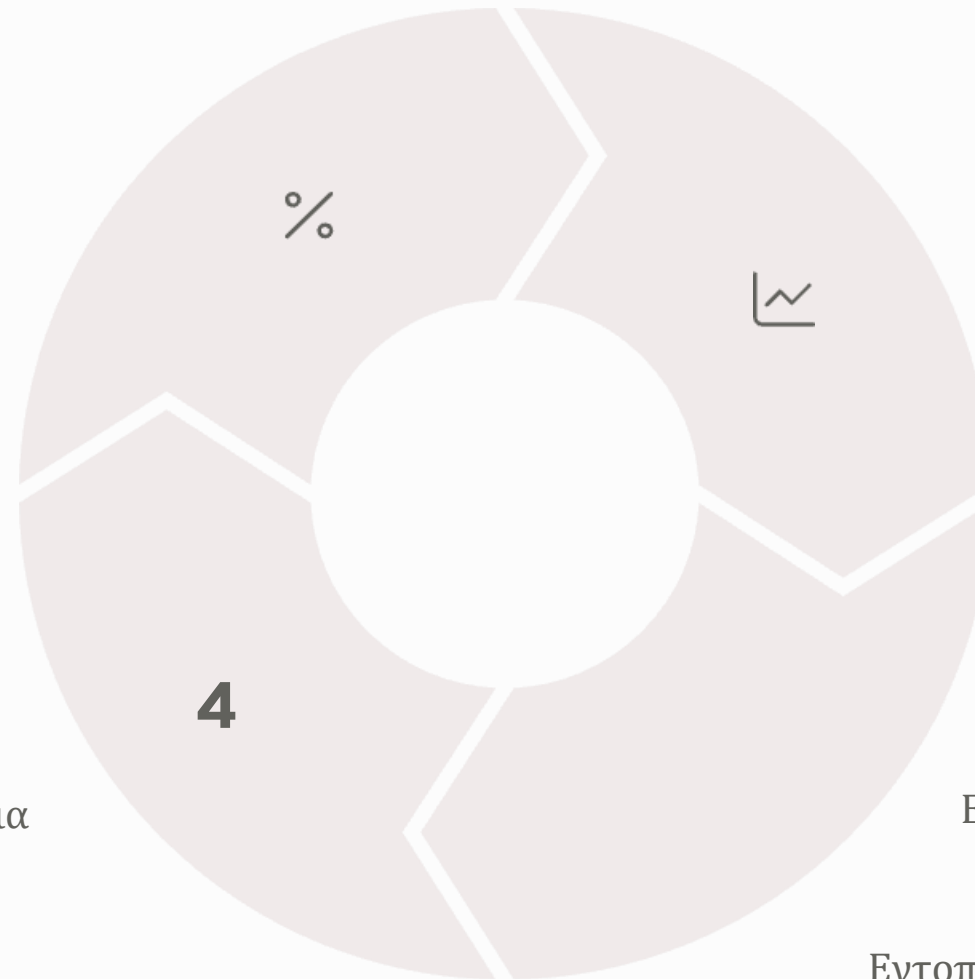
Συντελεστής Μεταβλητότητας (**CV**)

Το πηλίκο της τυπικής απόκλισης προς τον μέσο όρο, εκφρασμένο ως ποσοστό

$$CV = (s/\bar{x}) \cdot 100$$

Τυποποιημένες Τιμές (**Z-scores**)

Δείχνουν πόσες τυπικές αποκλίσεις απέχει μια τιμή από το μέσο όρο

$$Z_i = (X_i - \bar{X})/s$$


Ερμηνεία **CV**

Χαμηλός CV (<10%): Σταθερότητα και ομοιογένεια

Υψηλός CV (>30%): Αστάθεια/μεγάλη διακύμανση

Χρήση **Z-scores**

Επιτρέπουν τη σύγκριση τιμών από διαφορετικά δείγματα

Εντοπίζουν πιθανές ακραίες τιμές ($|z| > 3$)

Υπολογισμός με το EXCEL

Μέτρο	Συνάρτηση
Αριθμητικός μέσος	=AVERAGE(δεδομένα)
Περιοκομμένος μέσος Τετριμμένος μέσος	=TRIMMEAN(δεδομένα, ποσοστό περικοπής)
Διάμεσος	=MEDIAN(δεδομένα)
Επικρατούσα τιμή	=MODE.SNGL(δεδομένα)
Τεταρτημόρια	=QUARTILE.INC(δεδομένα, τεταρτημόριο)
Κ-ποσοστημόριο	=PERCENTILE.INC(δεδομένα, p)
Μέγιστο	=MAX(δεδομένα)
Ελάχιστο	=MIN(δεδομένα)

Υπολογισμός των περιγραφικών μέτρων με το πρόσθετο Ανάλυση Δεδομένων

- Στην καρτέλα Δεδομένα (Data), κάντε κλικ στην Ανάλυση Δεδομένων (Data Analysis).
- Επιλέξτε Περιγραφικά Στατιστικά (Descriptive Statistics) και πατήστε OK.
- Στο παράθυρο έχετε τις εξής επιλογές:
 - Input Range: Επιλέξτε την περιοχή των δεδομένων σας (π.χ. C2:C135).
 - Επιλέξτε Output Range και κάντε κλικ σε ένα κενό κελί όπου θέλετε να εμφανιστεί ο πίνακας των στατιστικών μέτρων (π.χ. G1), ή κάντε κλικ στο να αποθηκεύσει τα στατιστικά σε ένα νέο φύλλο (New Worksheet Ply).
 - Επιλέξτε "Summary statistics" (Περίληπτικά στατιστικά) για να υπολογίσει το Excel τα περιγραφικά στατιστικά
- Πατήστε OK

Ασκήσεις Κεφαλαίου 4

1. Υπολογίσετε τα μέτρα θέσης, τα μέτρα μεταβλητότητας και σχετικής μεταβλητότητας του ύψος του γιού (στήλη D) στο αρχείο *Data Family Heights v1.xlsx*, με τη χρήση συναρτήσεων στο Excel.
2. Υπολογίσετε τα περιγραφικά στατιστικά μέτρα για το ύψος της κόρης (στήλη E) στο αρχείο *Data Family Heights v1.xlsx*, στο Excel με τη χρήση του πρόσθετου «Ανάλυση Δεδομένων» (Data Analysis) .

Γραφική Απεικόνιση Δεδομένων

Κεφάλαιο 5



Εισαγωγή

Η γραφική παρουσίαση των δεδομένων αποτελεί έναν απλό και αποτελεσματικό τρόπο απεικόνισης που επιτρέπει τη μελέτη των χαρακτηριστικών τους. Στα προηγούμενα κεφάλαια παρουσιάστηκε η οργάνωση και παρουσίαση των δεδομένων με πίνακες και περιγραφικά στατιστικά μέτρα. Τώρα θα εξερευνήσουμε τους διάφορους τύπους γραφημάτων και τη χρήση τους για την αποτελεσματική οπτικοποίηση των δεδομένων.

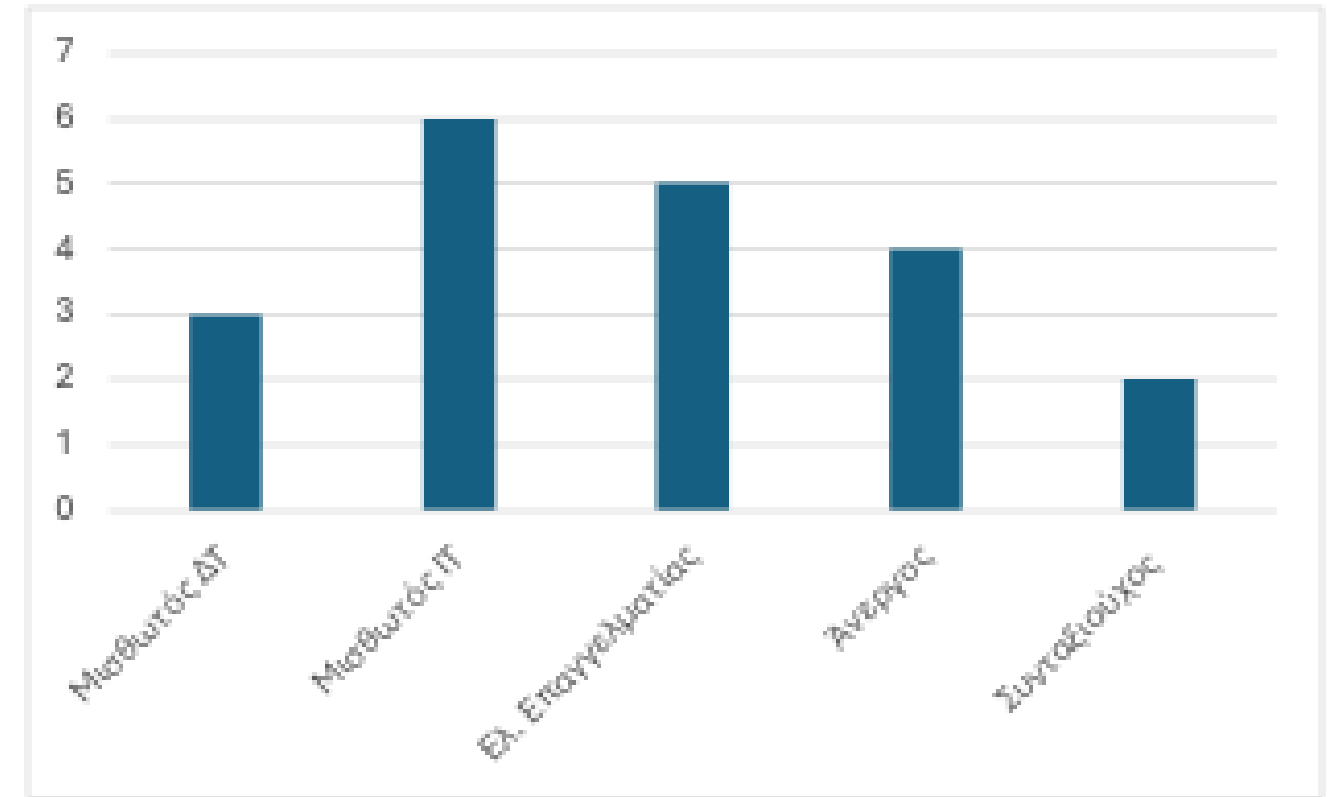
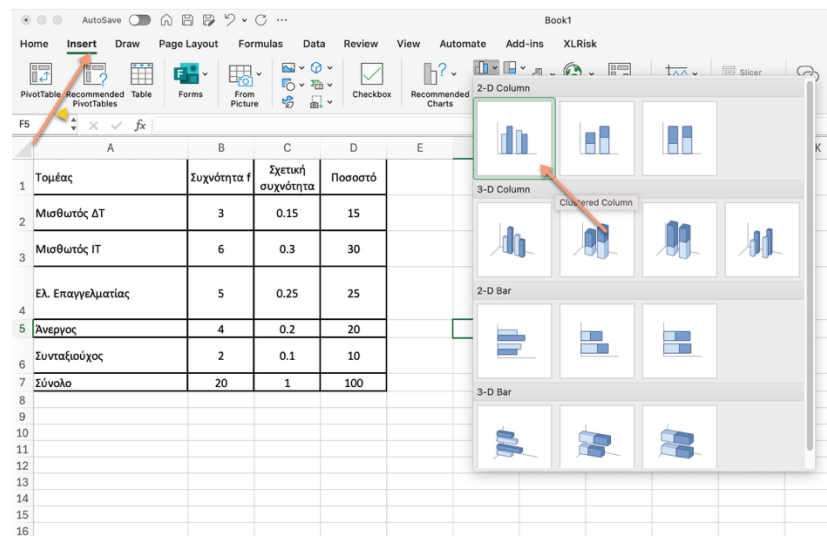
Για τη δημιουργία γραφημάτων στο Excel, ακολουθούμε τα βήματα: Επιλογή δεδομένων → Εισαγωγή → Γράφημα → Επιλογή του κατάλληλου τύπου γραφήματος. Κάθε τύπος γραφήματος εξυπηρετεί διαφορετικούς σκοπούς ανάλυσης και παρουσίασης.

Ραβδόγραμμα: Απεικόνιση Κατηγορικών Δεδομένων

Χρήση Ραβδογραμμάτων

Το ραβδόγραμμα (bar chart) χρησιμοποιείται για την παρουσίαση κατηγορικών ή διακριτών αριθμητικών δεδομένων. Στον οριζόντιο άξονα (x) τοποθετούνται οι κατηγορίες της μεταβλητής, ενώ στον κατακόρυφο άξονα (y) απεικονίζεται η συχνότητα ή η σχετική συχνότητα κάθε κατηγορίας.

Τα ραβδογράμματα είναι ιδανικά για τη σύγκριση διαφορετικών κατηγοριών και την ανάδειξη των διαφορών μεταξύ τους. Μπορούν επίσης να χρησιμοποιηθούν για την απεικόνιση διακριτών αριθμητικών δεδομένων.



Για τη δημιουργία ραβδογράμματος στο Excel, επιλέγουμε τα δεδομένα και στη συνέχεια πηγαίνουμε στο μενού "Εισαγωγή" και επιλέγουμε το πρώτο εικονίδιο στην περιοχή των γραφημάτων. Το Excel προσφέρει διάφορες παραλλαγές ραβδογραμμάτων, συμπεριλαμβανομένων των κατακόρυφων και οριζόντιων ραβδογραμμάτων.

Ιστόγραμμα: Απεικόνιση Συνεχών Δεδομένων



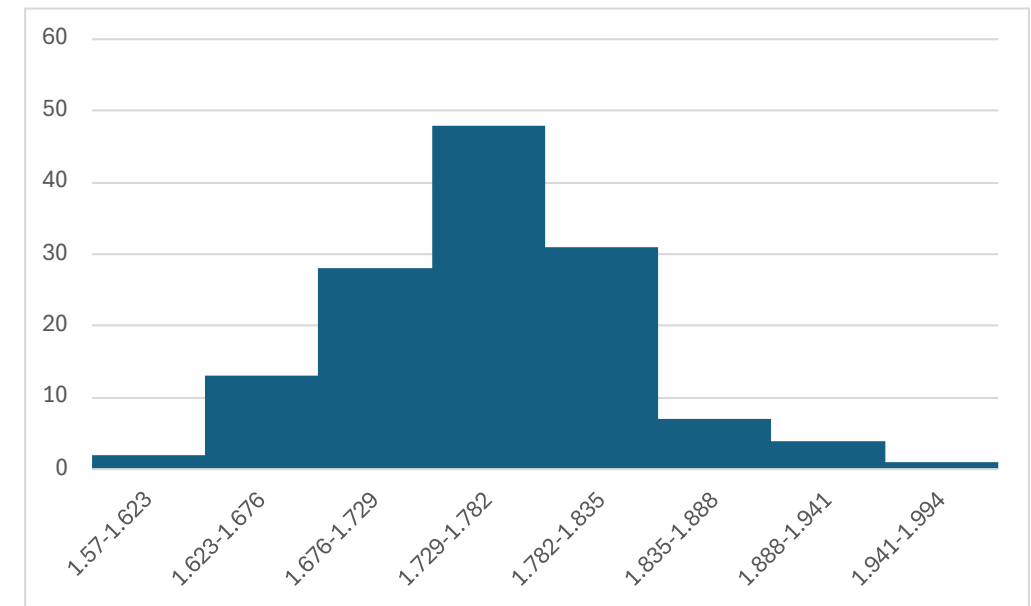
Ορισμός Ιστογράμματος

Το ιστόγραμμα (histogram) είναι μια γραφική παρουσίαση με μορφή ραβδογράμματος για συνεχείς μεταβλητές. Οι ράβδοι έχουν ίσο πλάτος για να αντιπροσωπεύουν ίσα διαστήματα τάξης και εφάπτονται η μία στην άλλη για να δείξουν ότι οι τιμές είναι συνεχείς.



Χαρακτηριστικά

Το ύψος κάθε ράβδου αντιπροσωπεύει τη συχνότητα ή το ποσοστό για κάθε διάστημα τάξης. Σε αντίθεση με το ραβδόγραμμα, δεν υπάρχουν κενά μεταξύ των ράβδων, καθώς αναπαριστούν συνεχή δεδομένα.



Ιστόγραμμα: Απεικόνιση Συνεχών Δεδομένων



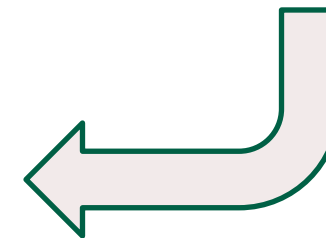
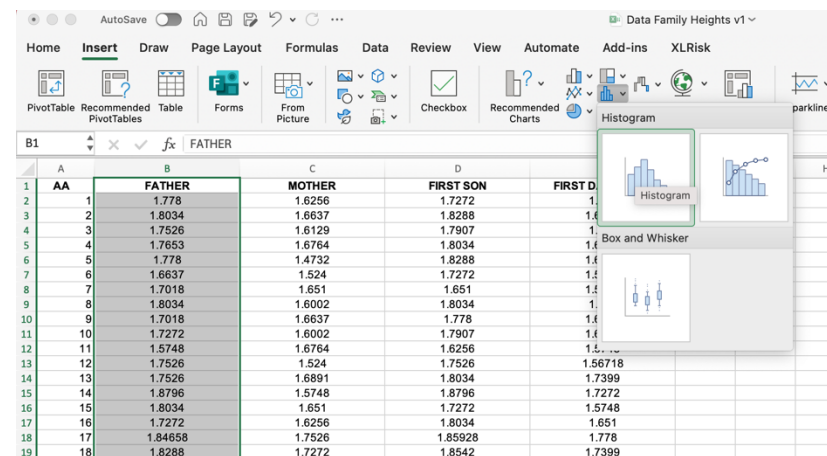
Δημιουργία στο **Excel** από πίνακα

Για τη δημιουργία ιστογράμματος στο Excel, μπορούμε είτε να κατασκευάσουμε αρχικά ένα ραβδόγραμμα και να θέσουμε το "Gap Width" ίσο με μηδέν, είτε να χρησιμοποιήσουμε την ειδική επιλογή "Ιστόγραμμα" από τα στατιστικά γραφήματα.



Δημιουργία στο **Excel** από τα δεδομένα

1. Αρχικά επιλέγουμε όλα τα αριθμητικά δεδομένα της μεταβλητής.
2. Στη συνέχεια πηγαίνουμε στο μενού «Εισαγωγή» (Insert).
3. Στην περιοχή των γραφημάτων κάνουμε κλικ στο εικονίδιο «Εισαγωγή στατιστικού γραφήματος» (Insert Statistic Chart) και επιλέγουμε «Ιστόγραμμα» (Histogram).



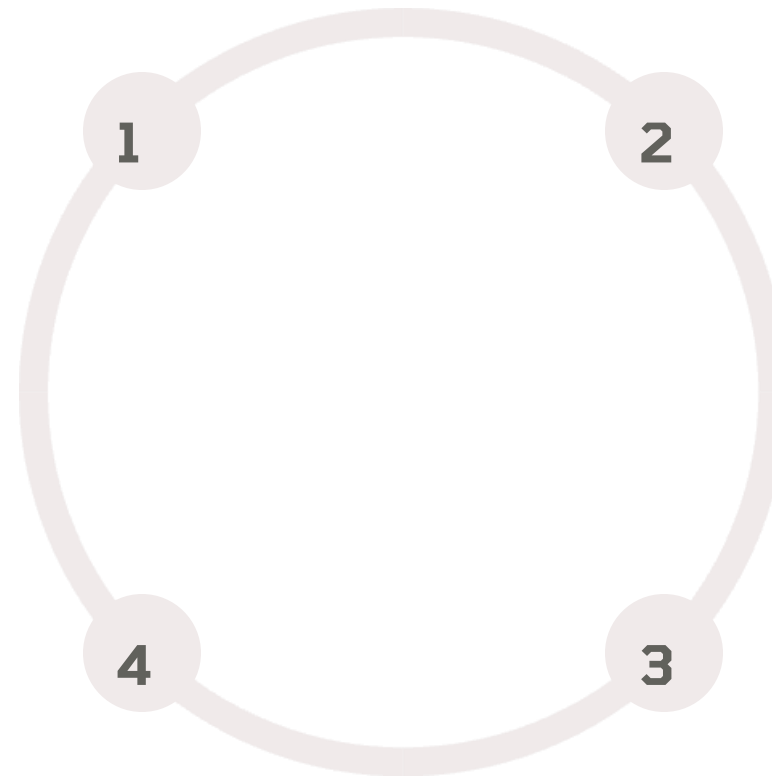
Κυκλικό Διάγραμμα

Χρήση

Το κυκλικό διάγραμμα (ή γράφημα πίτας) απεικονίζει κατηγορικά δεδομένα σε έναν κυκλικό δίσκο, όπου το μέγεθος κάθε φέτας αντιστοιχεί στη συχνότητα ή τη σχετική συχνότητα κάθε κατηγορίας.

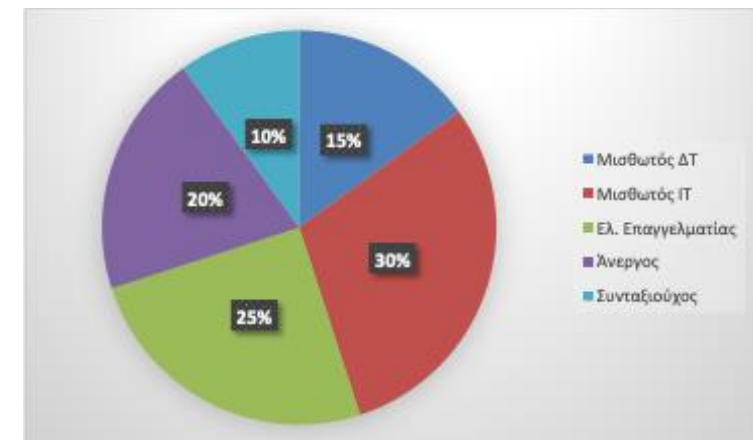
Δημιουργία στο **Excel**

Στο Excel, επιλέγουμε τα δεδομένα και πηγαίνουμε στο μενού "Εισαγωγή" και επιλέγουμε το εικονίδιο με το κυκλικό διάγραμμα στην περιοχή των γραφημάτων.



Πλεονεκτήματα

Είναι ιδανικό για την απεικόνιση της αναλογίας κάθε κατηγορίας ως προς το σύνολο, επιτρέποντας την άμεση οπτική σύγκριση των μεγεθών.



Θηκόγραμμα – Box Plot

Ορισμός και Δομή

Το θηκόγραμμα (box plot) είναι ένα σύγχρονο εργαλείο γραφικής απεικόνισης που παρουσιάζει τη διασπορά και την κεντρική τάση ενός συνόλου δεδομένων. Βασίζεται σε πέντε βασικά περιγραφικά στατιστικά μέτρα: την ελάχιστη τιμή, το πρώτο τεταρτημόριο ($Q1$), τη διάμεσο, το τρίτο τεταρτημόριο ($Q3$) και τη μέγιστη τιμή.

Στοιχεία Θηκογράμματος

Περιλαμβάνει ένα κουτί (box) που εκτείνεται από το $Q1$ έως το $Q3$, μια γραμμή στη θέση της διαμέσου, "κεραίες" (whiskers) που εκτείνονται ως τις ακραίες τιμές (εντός του εύρους $Q1-1.5 \times IQR$ έως $Q3+1.5 \times IQR$), και ακραίες τιμές (outliers) που σημειώνονται ξεχωριστά συνήθως με αστεράκια.

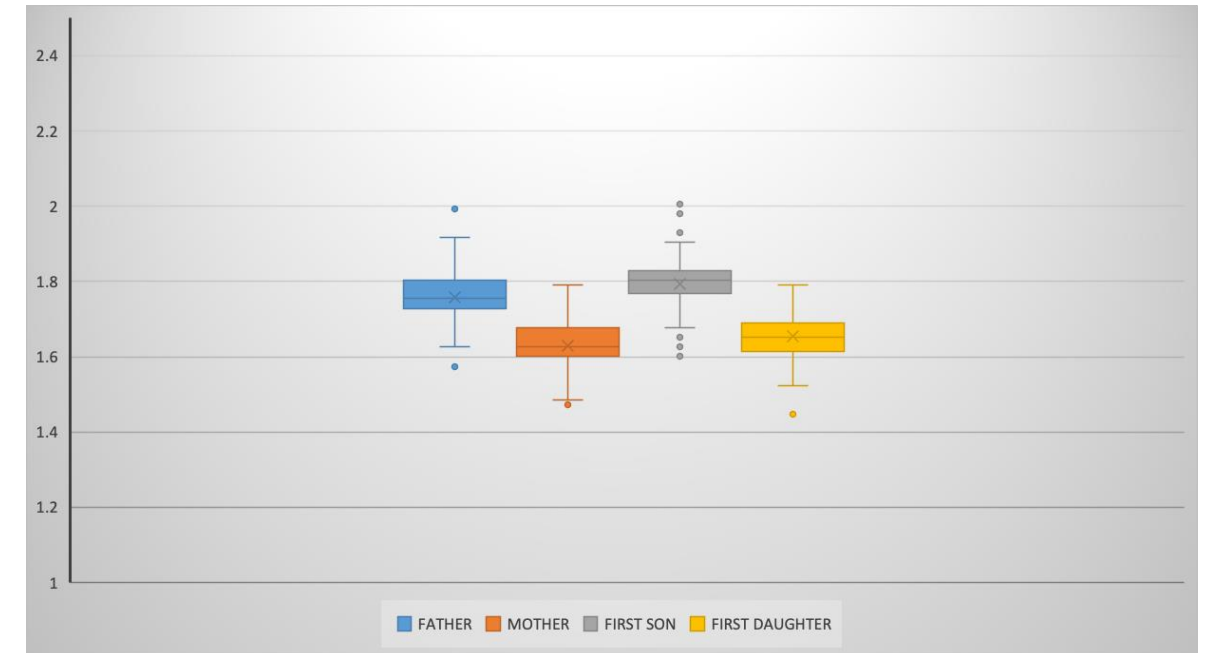
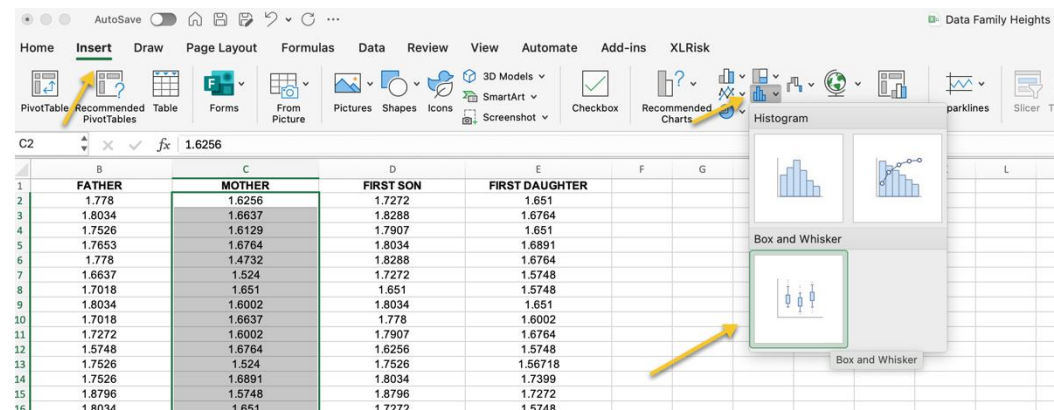
Χρήση και Πλεονεκτήματα

Χρησιμοποιείται για τη σύγκριση διαφορετικών συνόλων δεδομένων και τη μελέτη της συμμετρίας ή ασυμμετρίας στην κατανομή των δεδομένων. Είναι ιδιαίτερα χρήσιμο για την αναγνώριση ακραίων τιμών και την οπτικοποίηση της διασποράς των δεδομένων.

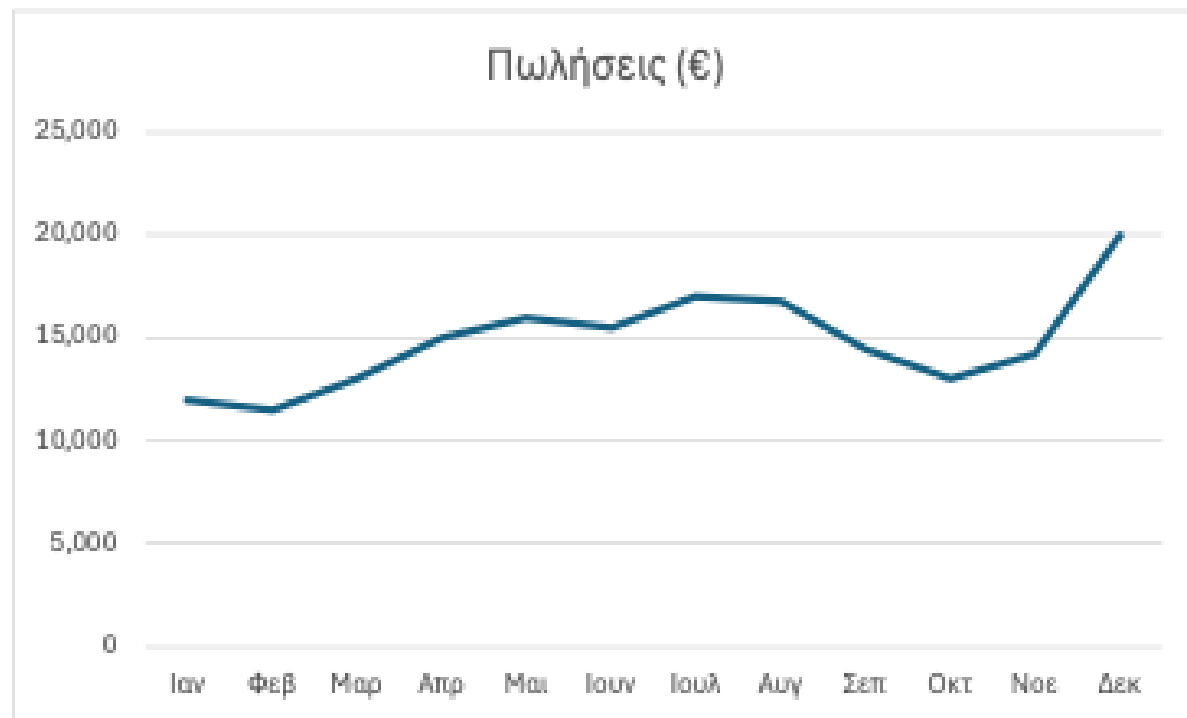


Δημιουργία στο **Excel**

1. Επιλέγουμε τα δεδομένα
2. Στην καρτέλα «Εισαγωγή» (Insert), κάνουμε κλικ στην περιοχή των γραφημάτων και επιλέγουμε τα στατιστικά γραφήματα (Statistical).
3. Επιλέγουμε «Box and Wisker».



Γράφημα Γραμμής: Απεικόνιση Χρονοσειρών



1

Ορισμός

Το γράφημα γραμμής (line plot) χρησιμοποιείται κυρίως για την απεικόνιση χρονοσειρών, δηλαδή μεταβλητών που μετρώνται σε τακτά χρονικά διαστήματα.

2

Χρήση

Χρησιμοποιείται για τη μελέτη της διαχρονικής εξέλιξης μιας μεταβλητής και για τον οπτικό προσδιορισμό της τάσης και της εποχικότητας στη χρονοσειρά.

3

Δομή

Στο γράφημα γραμμής, ο οριζόντιος άξονας αφορά το χρόνο και ο κατακόρυφος άξονας αφορά τις τιμές της μεταβλητής. Τα σημεία συνδέονται με γραμμές για να αναδείξουν την εξέλιξη.

4

Δημιουργία στο **Excel**

Για τη δημιουργία γραφήματος γραμμής στο Excel, επιλέγουμε τα δεδομένα και πηγαίνουμε στην καρτέλα "Εισαγωγή" και επιλέγουμε "Γράφημα Γραμμής".

Διάγραμμα Διασποράς: Απεικόνιση Σχέσεων

Ορισμός

Απεικονίζει τη σχέση μεταξύ δύο αριθμητικών μεταβλητών

Σκοπός

Διερευνά οπτικά τη σχέση μεταξύ μεταβλητών

Ερμηνεία

Αναγνωρίζει θετικές, αρνητικές ή μηδενικές συσχετίσεις

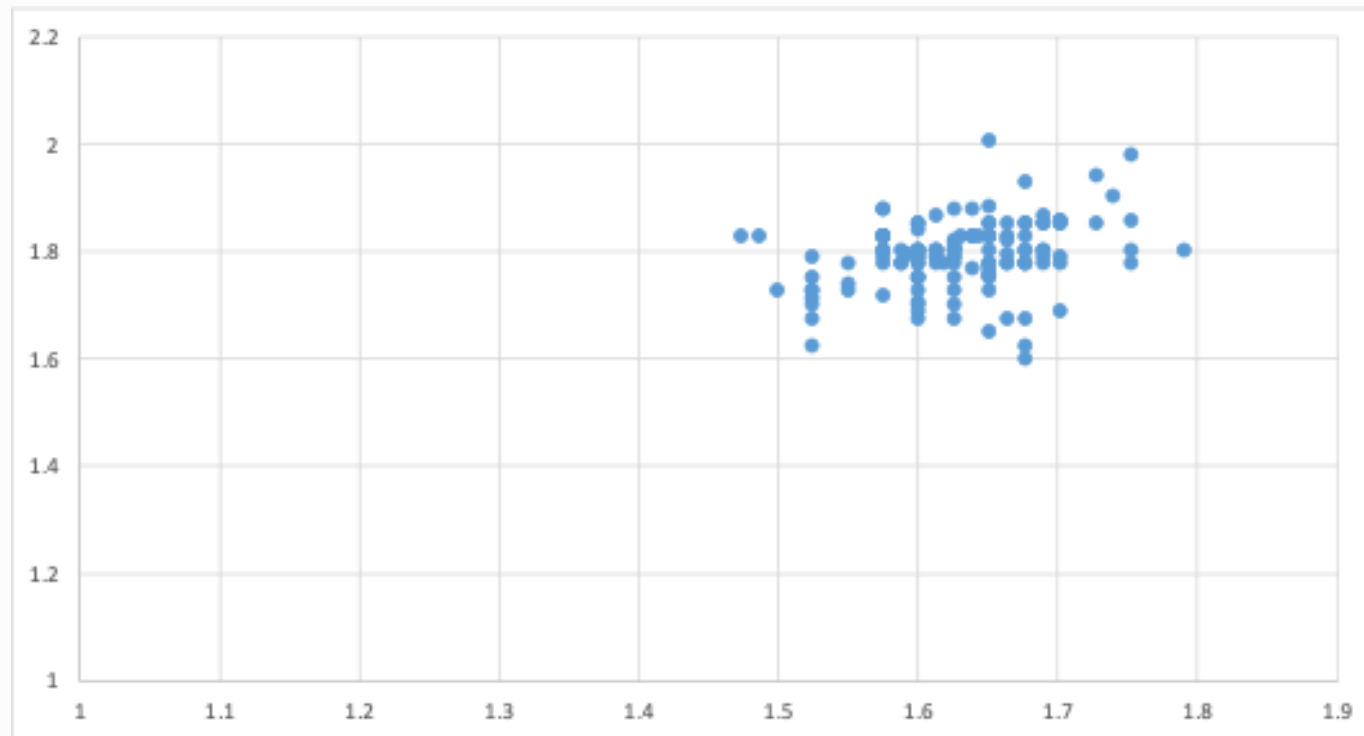
Δημιουργία

Επιλογή δεδομένων → Εισαγωγή → Διάγραμμα → Διασποράς

Το διάγραμμα διασποράς είναι ιδιαίτερα χρήσιμο για την αναγνώριση μοτίβων στα δεδομένα. Όταν τα σημεία έχουν κατεύθυνση από κάτω αριστερά προς πάνω δεξιά, αυτό αποτελεί ένδειξη θετικής συσχέτισης. Αντίθετα, όταν τα σημεία κινούνται από πάνω αριστερά προς κάτω δεξιά, αυτό υποδηλώνει αρνητική συσχέτιση. Είναι σημαντικό να τονιστεί ότι η ύπαρξη συσχέτισης δεν αποδεικνύει αιτιότητα.

Διάγραμμα Διασποράς: Απεικόνιση Σχέσεων

Διάγραμμα διασποράς υψών μητέρας και γιού



Το διάγραμμα που προκύπτει μας δείχνει μία θετική σχέση, άρα οι ψηλές μητέρες τείνουν να έχουν ψηλούς γιούς.

Πρακτική Εφαρμογή και Ασκήσεις

Επιλογή Κατάλληλου Γραφήματος

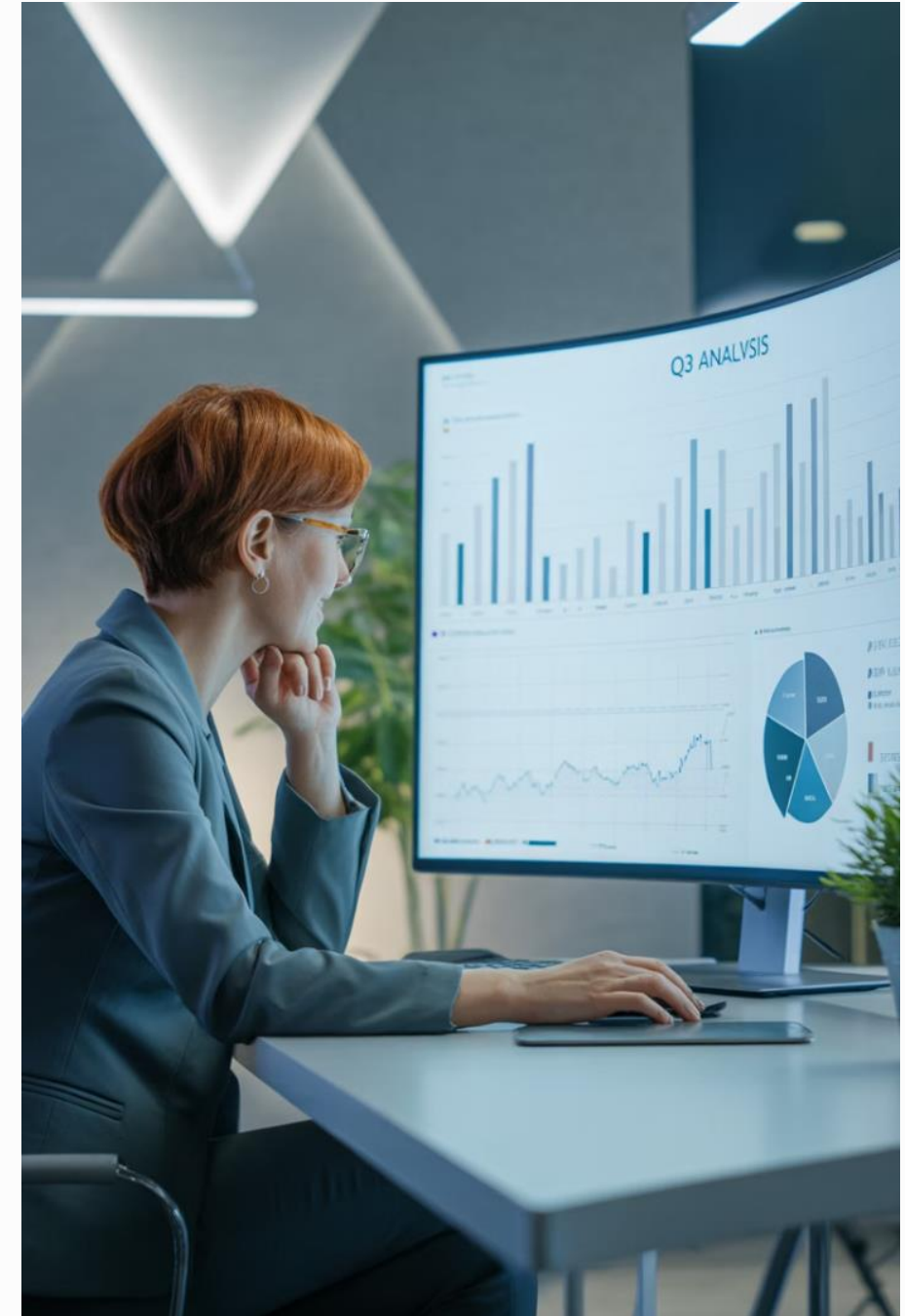
Για κάθε τύπο δεδομένων και ερώτημα ανάλυσης, υπάρχει ένας βέλτιστος τύπος γραφήματος. Για κατηγορικά δεδομένα, προτιμήστε ραβδογράμματα ή κυκλικά διαγράμματα. Για συνεχή δεδομένα, χρησιμοποιήστε ιστογράμματα ή θηκογράμματα. Για χρονοσειρές, επιλέξτε γραφήματα γραμμής.

Προσαρμογή Γραφημάτων

Το Excel προσφέρει πολλές δυνατότητες προσαρμογής των γραφημάτων. Μπορείτε να αλλάξετε χρώματα, γραμματοσειρές, τίτλους αξόνων, και να προσθέσετε υπομνήματα για να βελτιώσετε την αναγνωσιμότητα και την αποτελεσματικότητα των γραφημάτων σας.

Ασκήσεις Εξάσκησης

Χρησιμοποιώντας τα δεδομένα από το αρχείο Data Family Heights v1.xlsx: 1) Δημιουργήστε γράφημα διασποράς για τις μεταβλητές "Υψος μητέρας" και "Υψος κόρης" και σχολιάστε τη σχέση των δύο μεταβλητών. 2) Κατασκευάστε το ιστόγραμμα για τη μεταβλητή "Υψος κόρης".



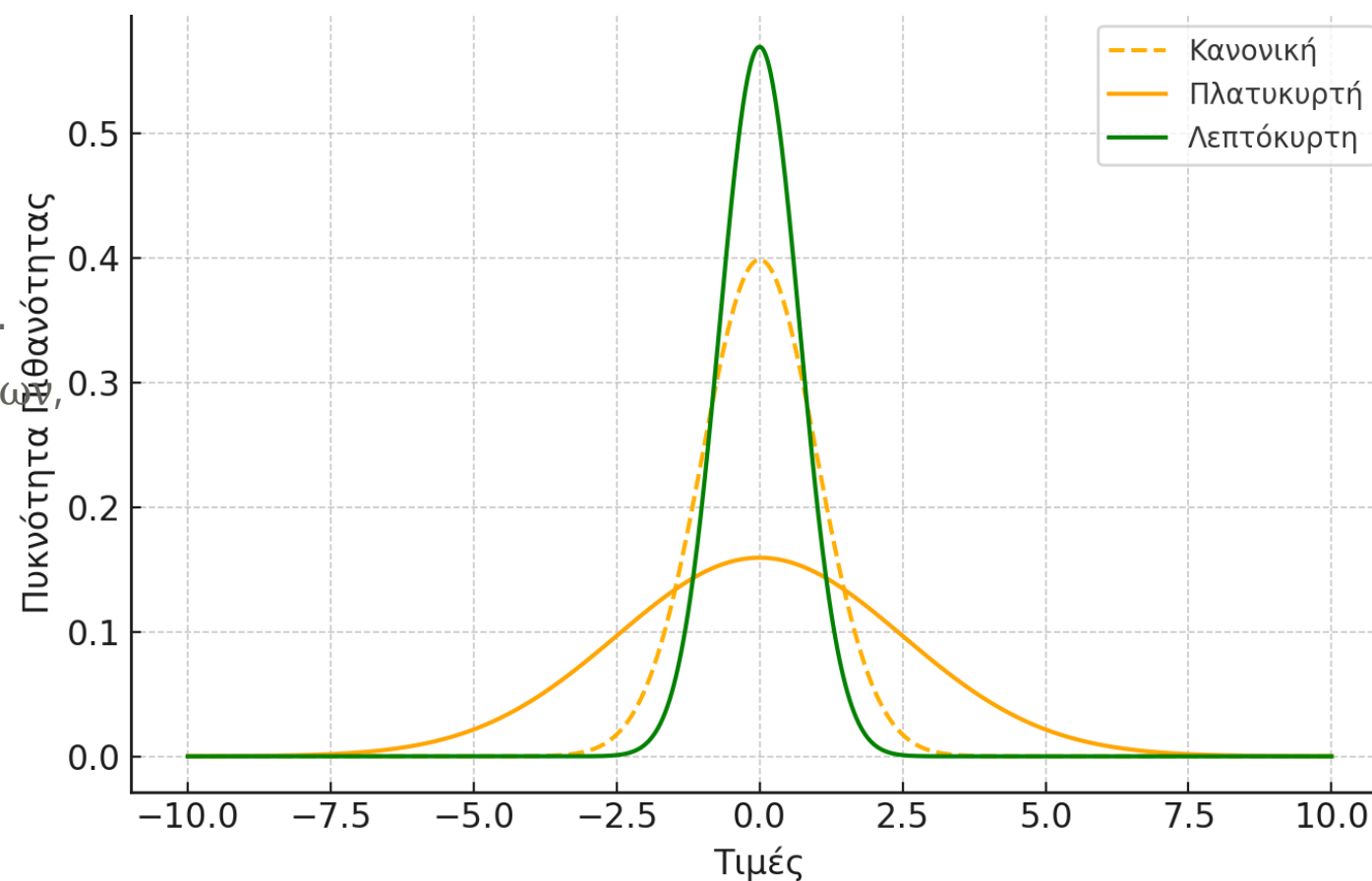
Κατανομές και Κατανομές Δειγματοληψίας

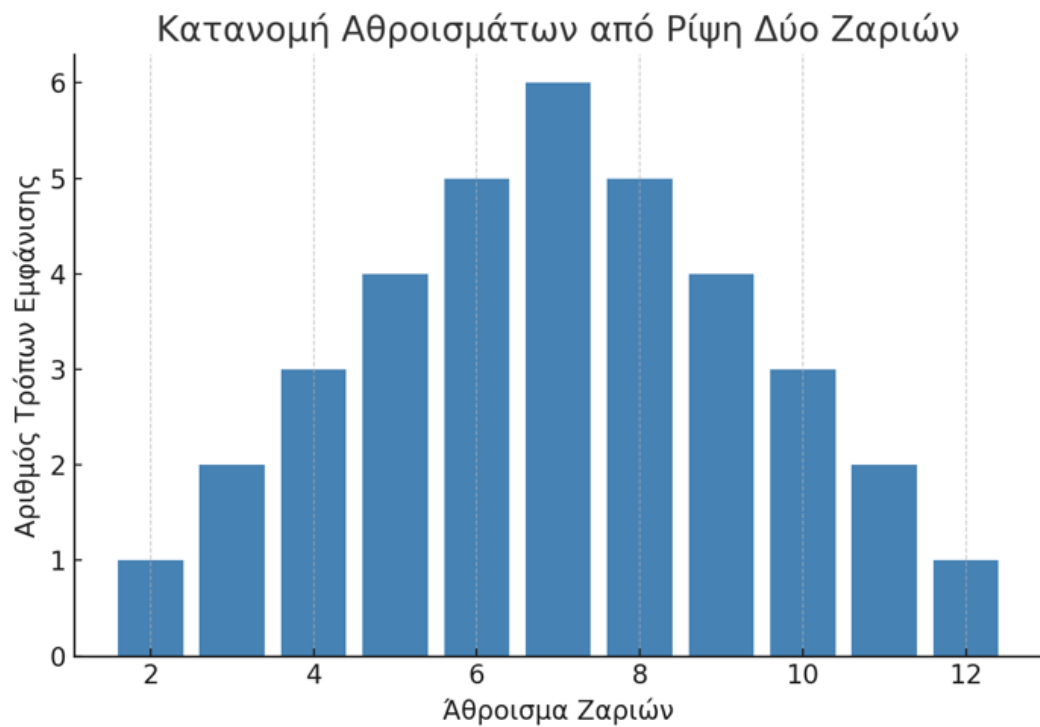
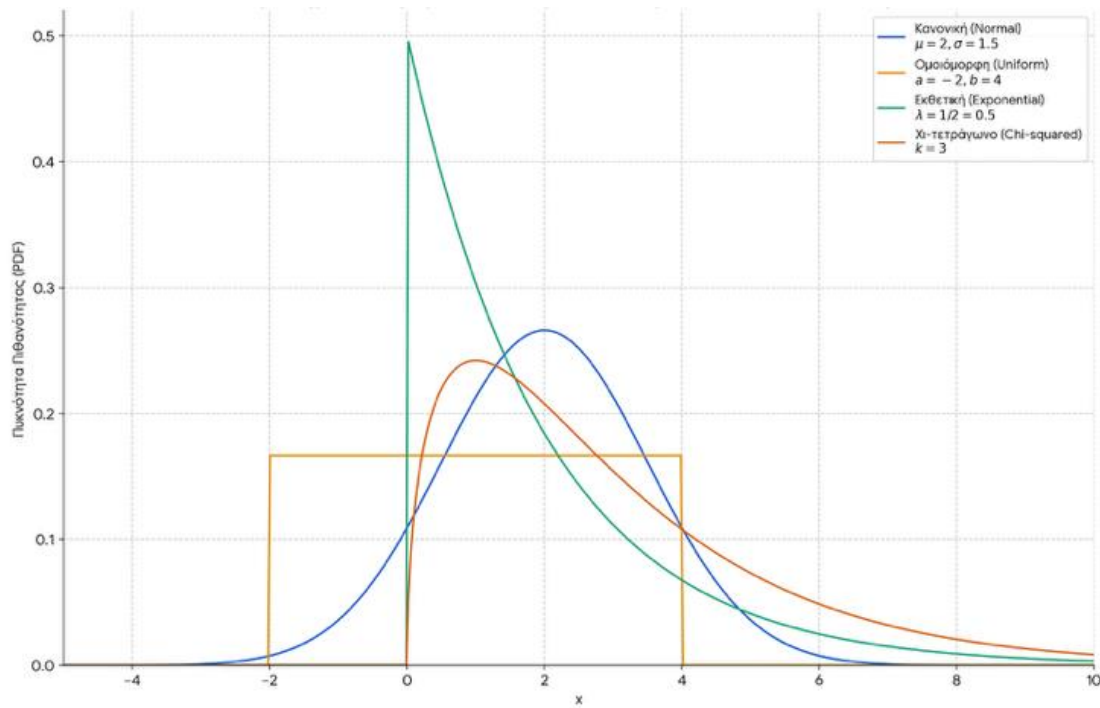
Κεφάλαιο 6

Κατανομές

Στη στατιστική, η κατανομή αναφέρεται στον τρόπο με τον οποίο κατανέμονται οι τιμές μίας μεταβλητής στον πληθυσμό ή στο δείγμα. Αποτελεί θεμελιώδη έννοια για την περιγραφή και ανάλυση δεδομένων, περιγράφοντας τη συχνότητα εμφάνισης των διάφορων τιμών.

Υπάρχουν δύο βασικοί τύποι κατανομών: οι εμπειρικές, που βασίζονται σε πραγματικά δεδομένα, και οι θεωρητικές, που βασίζονται σε μαθηματικά μοντέλα όπως η κανονική κατανομή, η διωνυμική και η t-student.





Η Έννοια της Κατανομής στη Στατιστική

Εμπειρικές Κατανομές

Βασίζονται σε πραγματικά δεδομένα και μπορούν να παρουσιαστούν με πίνακες κατανομής συχνοτήτων, ραβδογράμματα ή ιστογράμματα.

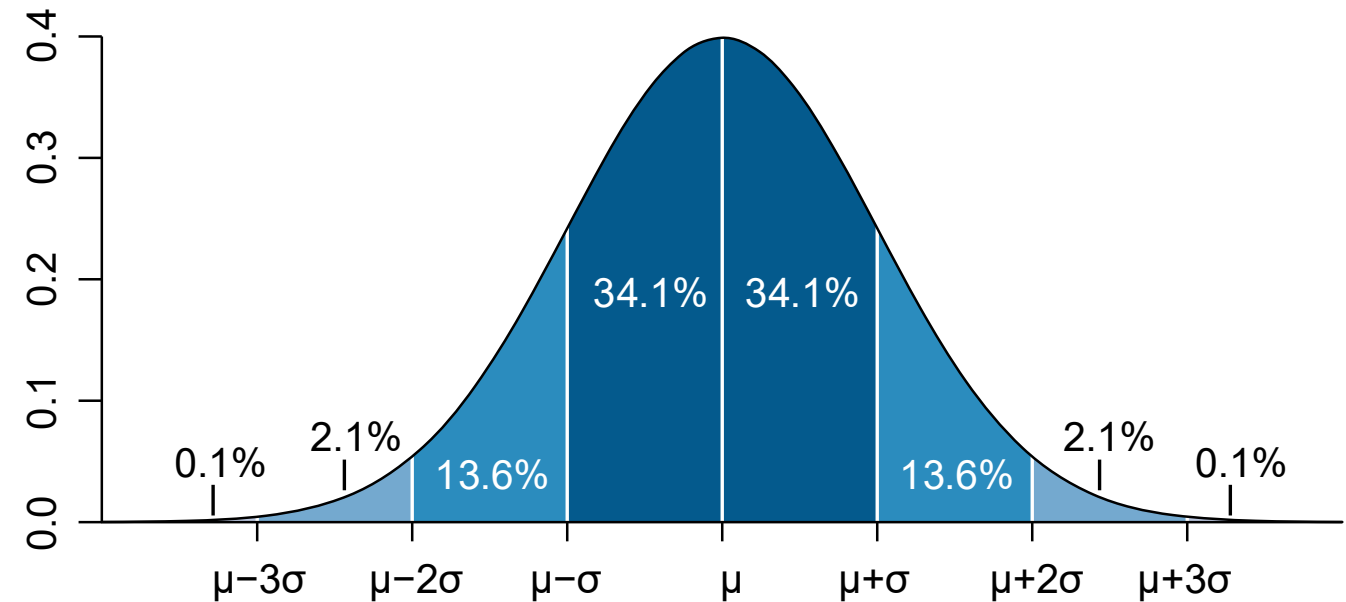
Θεωρητικές Κατανομές

Βασίζονται σε μαθηματικά μοντέλα και μπορούν να παρουσιαστούν με πίνακες, γραφήματα ή συνεχείς καμπύλες.

Παράδειγμα Κατανομής

Το άθροισμα από τη ρίψη δύο ζαριών αποτελεί χαρακτηριστικό παράδειγμα εμπειρικής κατανομής με συγκεκριμένες πιθανότητες για κάθε δυνατό άθροισμα.

Κανονική Κατανομή



Χαρακτηριστικά

Η κανονική κατανομή είναι συμμετρική γύρω από τον μέσο όρο και έχει το χαρακτηριστικό σχήμα καμπάνας. Περιγράφεται πλήρως από δύο παραμέτρους: τον μέσο όρο (μ) και την τυπική απόκλιση (σ).



Μαθηματική Εξίσωση

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



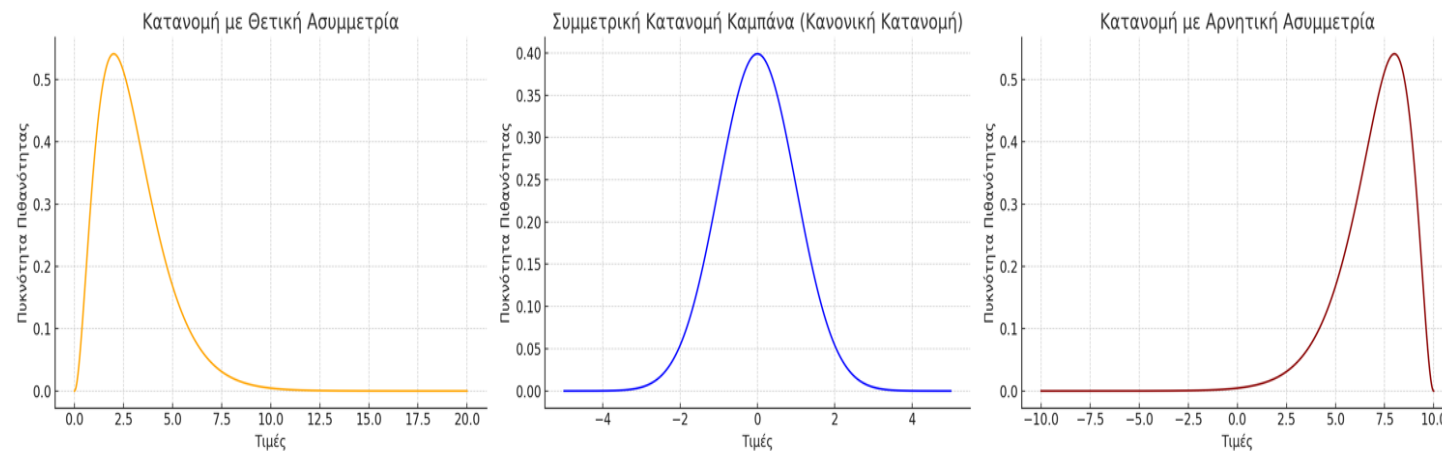
Κατανομή Τιμών

- Στο διάστημα ($\mu-\sigma$, $\mu+\sigma$) περιλαμβάνεται περίπου το 68% των παρατηρήσεων,
- στο ($\mu-2\sigma$, $\mu+2\sigma$) το 95%, και
- στο ($\mu-3\sigma$, $\mu+3\sigma$) το 99,7% των παρατηρήσεων.

Μέτρα Μορφής της Κατανομής

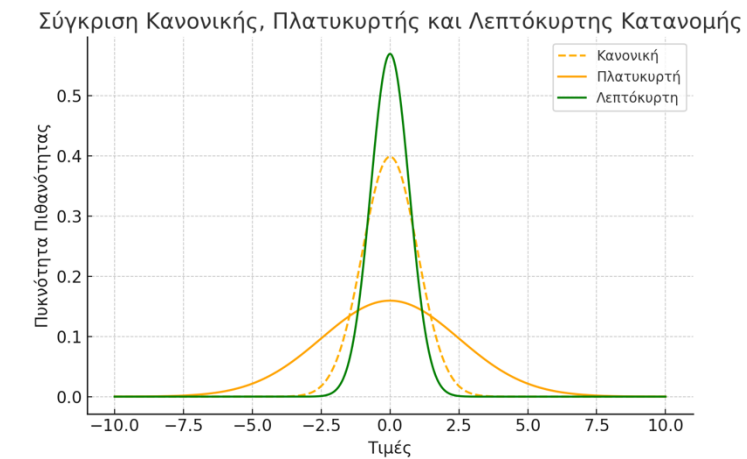
Συντελεστής Ασυμμετρίας

Ο συντελεστής ασυμμετρίας (skewness) μετρά τον βαθμό ασυμμετρίας μιας κατανομής γύρω από τον μέσο όρο. Όταν είναι μηδέν, η κατανομή είναι συμμετρική. Θετικός συντελεστής υποδηλώνει θετική ασυμμετρία με μακριά ουρά προς τα δεξιά, ενώ αρνητικός συντελεστής υποδηλώνει αρνητική ασυμμετρία με ουρά προς τα αριστερά.



Συντελεστής Κύρτωσης

Ο συντελεστής κύρτωσης (kurtosis) περιγράφει το ύψος και την αιχμή της "καμπούρας" της κατανομής. Κατανομές με σχήμα καμπάνας έχουν συντελεστή κύρτωσης μηδέν. Λεπτόκυρτες κατανομές έχουν αιχμηρότερη κορυφή και συντελεστή μεγαλύτερο του μηδενός, ενώ πλατύκυρτες είναι πιο επίπεδες με συντελεστή μικρότερο του μηδενός.



Υπολογισμός Συντελεστών στο Excel

Συντελεστής Ασυμμετρίας

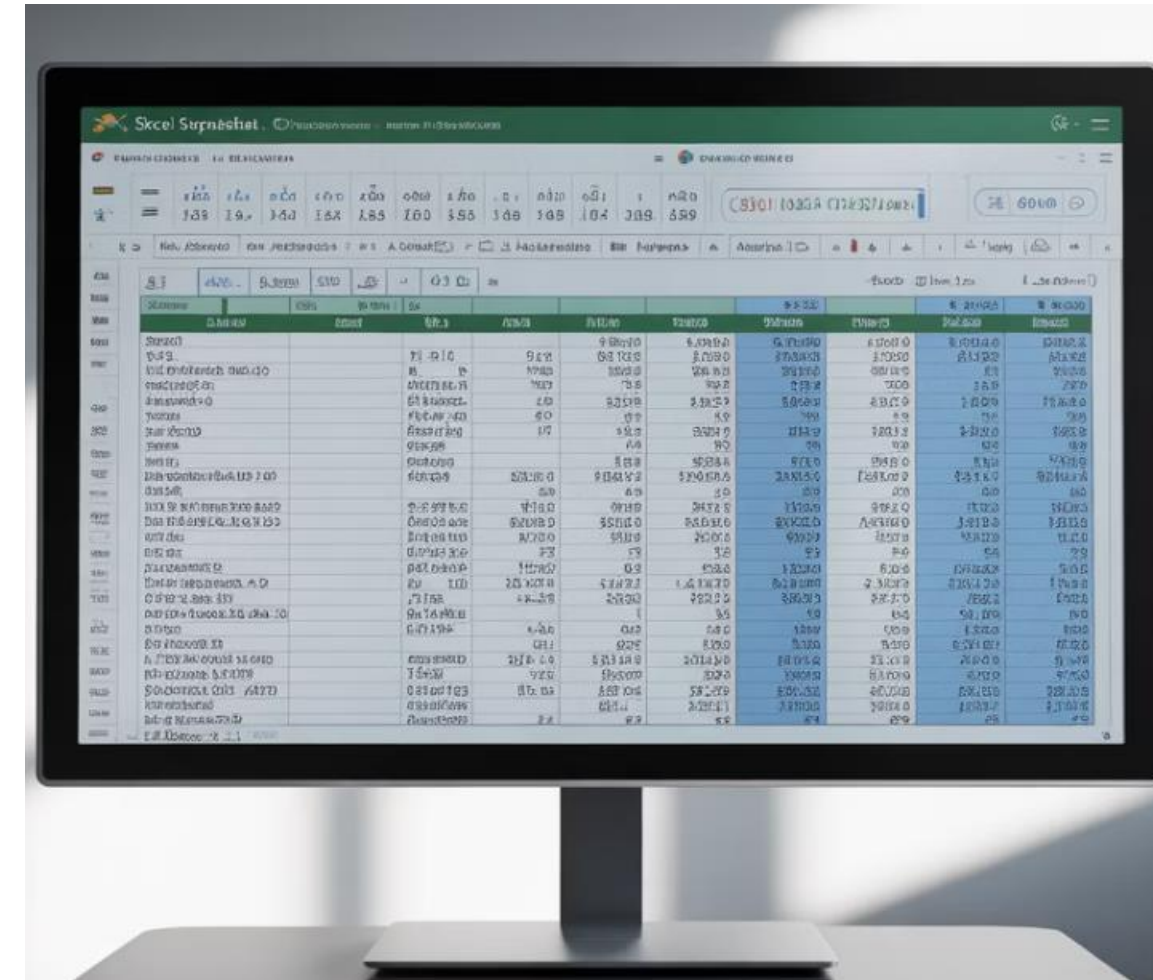
Χρησιμοποιούμε τη συνάρτηση **SKEW(δεδομένα)** του Excel. Για παράδειγμα, **=SKEW(c2:c135)** για τον υπολογισμό του συντελεστή ασυμμετρίας του ύψους των μητέρων. Το αποτέλεσμα **-0.08** δείχνει ότι η κατανομή είναι σχεδόν συμμετρική.

Συντελεστής Κύρτωσης

Χρησιμοποιούμε τη συνάρτηση **KURT(δεδομένα)** του Excel. Για παράδειγμα, **=KURT(c2:c135)** για τον υπολογισμό του συντελεστή κύρτωσης του ύψους των μητέρων. Το αποτέλεσμα **0.206** δείχνει ότι η κατανομή έχει ελαφρώς πιο αιχμηρή καμπούρα από την κανονική.

Ερμηνεία Αποτελεσμάτων

Οι τιμές των συντελεστών μας βοηθούν να κατανοήσουμε καλύτερα τη μορφή της κατανομής και συμπληρώνουν την περιγραφή της κεντρικής τάσης. Πολλές στατιστικές τεχνικές απαιτούν συμμετρική κατανομή με σχήμα καμπάνας.



Κατανομή Δειγματοληψίας

Ορισμός

Η κατανομή δειγματοληψίας είναι η κατανομή όλων των πιθανών τιμών μιας στατιστικής συνάρτησης (π.χ. μέσου όρου), οι οποίες προκύπτουν από πολλαπλά τυχαία δείγματα ίδιου μεγέθους από έναν πληθυσμό.

Κεντρικό Οριακό Θεώρημα

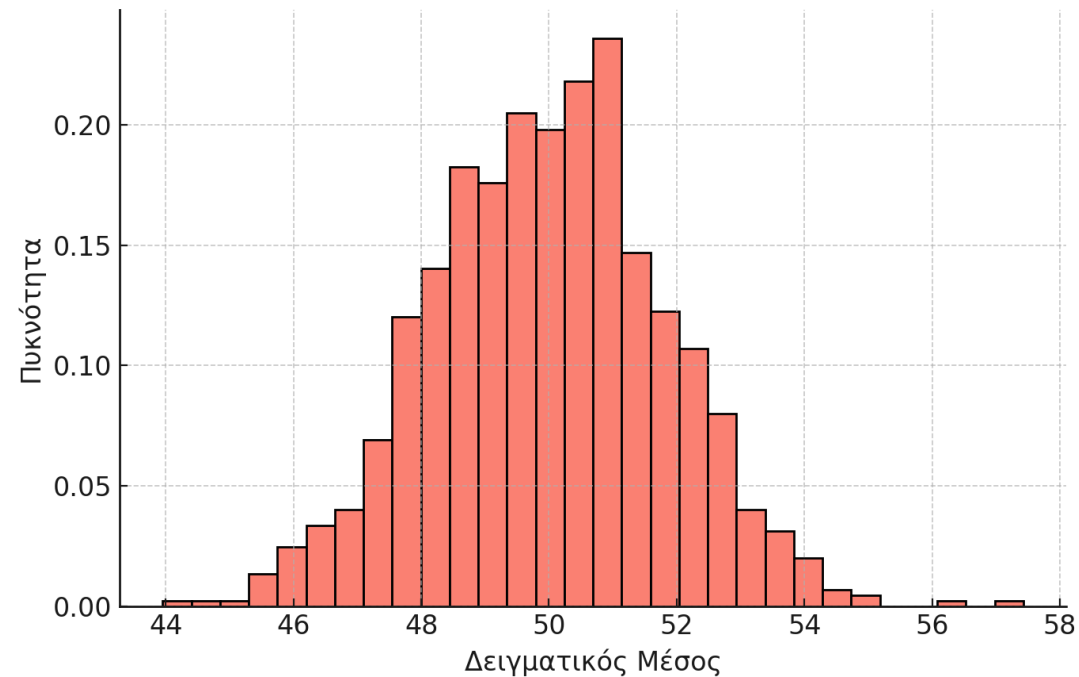
Ανεξάρτητα από την κατανομή του πληθυσμού, η κατανομή του μέσου όρου του δείγματος τείνει, όταν το μέγεθος του δείγματος μεγαλώνει ($n \geq 30$), στην κανονική κατανομή με κέντρο τον πραγματικό μέσο του πληθυσμού.



Χαρακτηριστικά

Ο μέσος όρος της κατανομής δειγματοληψίας του δειγματικού μέσου όρου ισούται με τον πληθυσμιακό μέσο όρο (μ), ενώ η τυπική απόκλιση (τυπικό σφάλμα) είναι ίση με $\sigma/\sqrt{n}$.

Εφαρμογή του Κεντρικού Οριακού Θεωρήματος



Πληθυσμός



Έχουμε έναν άπειρο θεωρητικό πληθυσμό, με μέσο όρο 50 και τυπική απόκλιση 10.

Δειγματοληψία



Επιλέγουμε 1000 τυχαία δείγματα, το καθένα μεγέθους 30, και υπολογίζουμε τον μέσο όρο κάθε δείγματος.

Κατανομή Δειγματοληψίας



Οι δειγματικοί μέσοι συγκεντρώνονται γύρω από τον πληθυσμιακό μέσο (50) και σχηματίζουν μια κατανομή που πλησιάζει την κανονική.

Τυπικό Σφάλμα



Η τυπική απόκλιση της κατανομής δειγματοληψίας (τυπικό σφάλμα) είναι περίπου 1.81, που επιβεβαιώνει τον τύπο $\sigma/\sqrt{n} = 10/\sqrt{30} \approx 1.83$.

Σημασία και Εφαρμογές

Το κεντρικό οριακό θεώρημα έχει δεσπόζουσα θέση στη Στατιστική διότι επιτρέπει την εφαρμογή επαγωγικών στατιστικών μεθόδων, οι οποίες συχνά βασίζονται στην υπόθεση της κανονικότητας, ακόμη και όταν η κατανομή του αρχικού πληθυσμού δεν είναι γνωστή ή δεν είναι κανονική.



Επαγωγική Στατιστική

Επιτρέπει την εφαρμογή επαγωγικών μεθόδων

$$f(x)$$

Εκτίμηση Παραμέτρων

Βάση για διαστήματα εμπιστοσύνης



Έλεγχος Υποθέσεων

Θεμέλιο για στατιστικούς ελέγχους

Η κατανόηση των κατανομών δειγματοληψίας είναι κρίσιμη για την εκτίμηση παραμέτρων του πληθυσμού και τη διεξαγωγή στατιστικών ελέγχων με ακρίβεια και αξιοπιστία.

Ασκήσεις κεφαλαίου

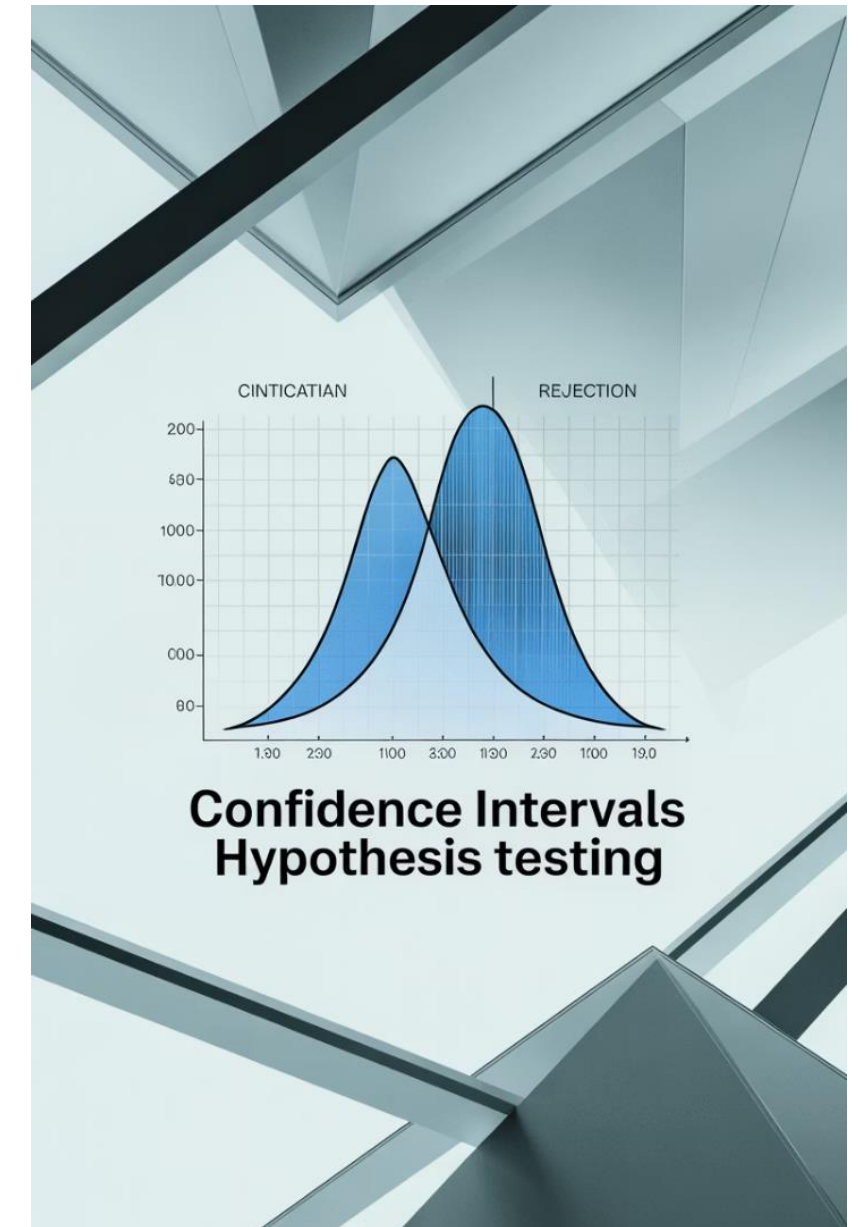
Χρησιμοποιώντας τα δεδομένα από το αρχείο Family Heights v1.xlsx:

1. Να υπολογίσετε τον συντελεστή ασυμμετρίας για τις μεταβλητές, «Ύψος μητέρας» και «Ύψος κόρης» και να σχολιασθεί η ασυμμετρία της κατανομής των δύο μεταβλητών.
2. Να υπολογίσετε τον συντελεστή κύρτωσης για τις μεταβλητές, «Ύψος πατέρα» και «Ύψος γιου» και να σχολιασθεί η κυρτότητα της κατανομής των δύο μεταβλητών.

Διαστήματα Εμπιστοσύνης και Έλεγχος Υποθέσεων

Η επαγωγική στατιστική στοχεύει στην εξαγωγή συμπερασμάτων για έναν ευρύτερο πληθυσμό, βασιζόμενη στην ανάλυση ενός δείγματος. Σε αντίθεση με την περιγραφική στατιστική που απλώς συνοψίζει δεδομένα, η επαγωγική στατιστική μας επιτρέπει να κάνουμε προβλέψεις και να αξιολογήσουμε την εγκυρότητα ισχυρισμών σχετικά με πληθυσμιακές παραμέτρους.

Στο κεφάλαιο αυτό θα εξετάσουμε δύο βασικά εργαλεία της επαγωγικής στατιστικής: τα διαστήματα εμπιστοσύνης και τον έλεγχο υποθέσεων, καθώς και τις μεθοδολογίες που χρησιμοποιούνται για την εφαρμογή τους στην πράξη.



Σημειακή Εκτίμηση

Ορισμός

Η σημειακή εκτίμηση είναι η τιμή που χρησιμοποιούμε για να εκτιμήσουμε μία άγνωστη πληθυσμιακή ποσότητα, παρέχοντας την καλύτερη εικασία για την πραγματική τιμή μιας παραμέτρου.

Περιορισμοί

Οι σημειακές εκτιμήσεις είναι απίθανο να ταυτίζονται απόλυτα με την πραγματική τιμή της παραμέτρου και δεν παρέχουν πληροφορίες για την ακρίβεια ή την αβεβαιότητα της εκτίμησης.

Χρησιμότητα

Παρά τους περιορισμούς, οι σημειακές εκτιμήσεις αποτελούν τη βάση για πιο σύνθετες στατιστικές μεθόδους και συχνά συμπληρώνονται με διαστήματα εμπιστοσύνης.

Ο πρωταρχικός σκοπός της σημειακής εκτίμησης είναι να προσδιορίσει μια προσεγγιστική τιμή για μια άγνωστη παράμετρο του πληθυσμού, αξιοποιώντας τις πληροφορίες που περιέχονται σε ένα τυχαίο δείγμα. Για παράδειγμα, ο δειγματικός μέσος όρος $\bar{X}$ χρησιμοποιείται συχνά ως σημειακός εκτιμητής του πληθυσμιακού μέσου όρου (μ).

Πληθυσμιακές Παράμετροι και Σημειακοί Εκτιμητές

Πληθυσμιακή Παράμετρος	Σύμβολο Παραμέτρου	Σημειακός Εκτιμητής	Σύμβολο Εκτιμητή
Μέση Τιμή Πληθυσμού	μ	Δειγματικός μέσος όρος	$\bar{X}$
Αναλογία Πληθυσμού (Ποσοστό)	π	Δειγματική Αναλογία	p ή $\hat{p}$
Διακύμανση Πληθυσμού	σ^2	Δειγματική Διακύμανση	s^2
Τυπική Απόκλιση Πληθυσμού	σ	Δειγματική Τυπική Απόκλιση	s

Εκτίμηση με Διαστήματα Εμπιστοσύνης

Οι σημειακές εκτιμήσεις σπάνια ταυτίζονται με την πραγματική τιμή της παραμέτρου και δεν παρέχουν πληροφορίες για την ακρίβεια ή την αβεβαιότητα. Για την αντιμετώπιση αυτού του προβλήματος, χρησιμοποιούμε την εκτίμηση με διάστημα εμπιστοσύνης.

Ένα διάστημα εμπιστοσύνης είναι ένα εύρος τιμών που χρησιμοποιείται για την εκτίμηση μιας άγνωστης πληθυσμιακής παραμέτρου. Κατασκευάζεται έτσι ώστε, εάν η διαδικασία επαναλαμβανόταν πολλές φορές, ένα συγκεκριμένο ποσοστό αυτών των διαστημάτων (ο βαθμός εμπιστοσύνης) θα περιείχε την πραγματική τιμή της άγνωστης παραμέτρου.

Βασικά Συστατικά Διαστήματος Εμπιστοσύνης

Η γενική μορφή ενός διαστήματος εμπιστοσύνης είναι: **Σημειακή Εκτίμηση $\pm$ Περιθώριο Σφάλματος**. Με αυτό τον τρόπο προσδιορίζουμε το άνω και κάτω όριο του διαστήματος εμπιστοσύνης.

Σημειακή Εκτίμηση

Η τιμή του σημειακού εκτιμητή που υπολογίζεται από το δείγμα, π.χ. ο δειγματικός μέσος όρος.

Περιθώριο Σφάλματος

Η ποσότητα που προστίθεται και αφαιρείται από τη σημειακή εκτίμηση για να δημιουργηθούν το κάτω και το άνω όριο του διαστήματος.

Βαθμός Εμπιστοσύνης

Η πιθανότητα (εκφρασμένη ως ποσοστό) ότι η μέθοδος κατασκευής του διαστήματος θα οδηγήσει σε ένα διάστημα που περιέχει την πραγματική παράμετρο. Συνήθεις τιμές είναι το 95% ή το 99%.

Εκτίμηση με Διαστήματα Εμπιστοσύνης



Σημειακή Εκτίμηση

Υπολογισμός του σημειακού εκτιμητή από το δείγμα (π.χ. δειγματικός μέσος όρος)



Περιθώριο Σφάλματος

Προσδιορισμός του περιθωρίου σφάλματος με βάση το τυπικό σφάλμα και την κρίσιμη τιμή



Κατασκευή Διαστήματος

Δημιουργία του διαστήματος προσθέτοντας και αφαιρώντας το περιθώριο σφάλματος από τη σημειακή εκτίμηση



Ερμηνεία

Ερμηνεία του διαστήματος με βάση το βαθμό εμπιστοσύνης

Ένα διάστημα εμπιστοσύνης είναι ένα διάστημα τιμών που χρησιμοποιείται για να εκτιμηθεί μια άγνωστη πληθυσμιακή παράμετρος. Κατασκευάζεται με τέτοιο τρόπο ώστε, εάν η διαδικασία δειγματοληψίας και κατασκευής του διαστήματος επαναλαμβανόταν πολλές φορές, ένα συγκεκριμένο ποσοστό αυτών των διαστημάτων (ο βαθμός εμπιστοσύνης) θα περιείχε την πραγματική τιμή της άγνωστης παραμέτρου.

Ερμηνεία Διαστημάτων Εμπιστοσύνης

Ο βαθμός εμπιστοσύνης 95% αναφέρεται στην αξιοπιστία της μεθόδου που χρησιμοποιήθηκε για την κατασκευή του διαστήματος, δηλαδή, αν επαναλαμβάναμε τη διαδικασία πολλές φορές, το 95% των διαστημάτων που θα προέκυπταν θα περιείχαν την πραγματική τιμή της πληθυσμιακής παραμέτρου μ .

Ορθή Ερμηνεία

Αν υπολογίσουμε ένα 95% Δ.Ε. για τον μέσο όρο ενός πληθυσμού και βρούμε ότι είναι $[10, 14]$, η ορθή ερμηνεία είναι:

"Έχουμε 95% εμπιστοσύνη (ή είμαστε 95% σίγουροι) ότι ο πραγματικός μέσος όρος (μ) του πληθυσμού βρίσκεται μεταξύ 10 και 14."

Λανθασμένη Ερμηνεία

Δεν είναι σωστό να ερμηνεύονται τα διαστήματα εμπιστοσύνης με όρους πιθανοτήτων, δηλαδή ότι υπάρχει 95% πιθανότητα η συγκεκριμένη, άγνωστη τιμή μ να βρίσκεται μέσα στο συγκεκριμένο διάστημα.

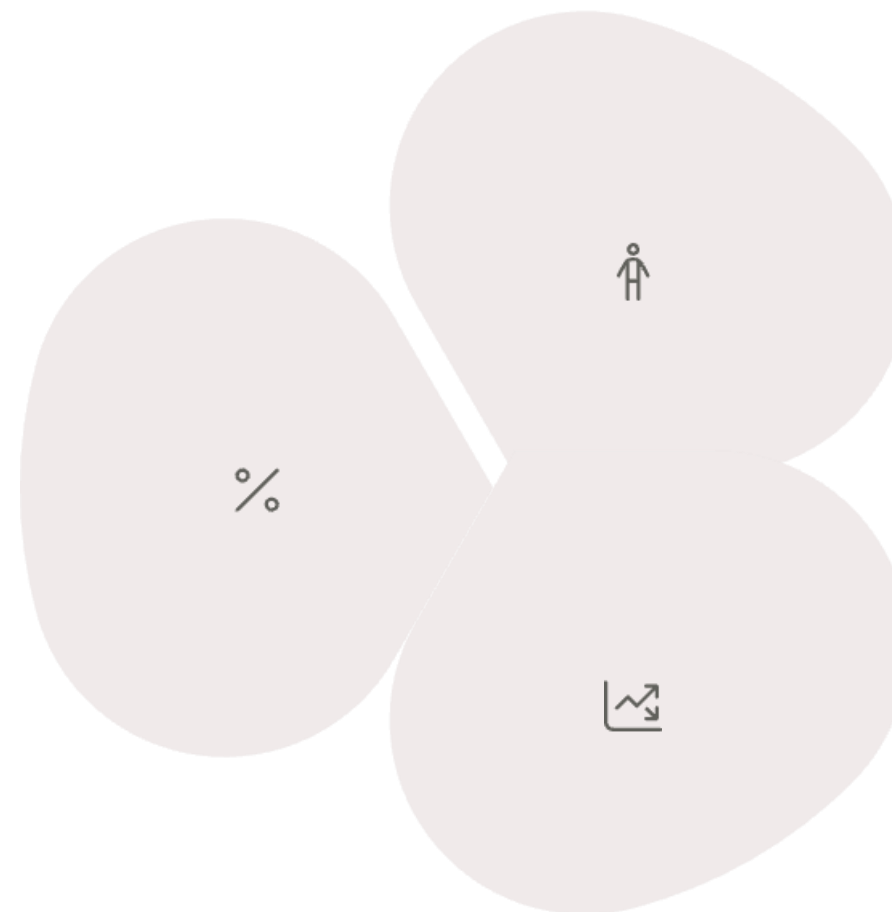
Αφού το διάστημα εμπιστοσύνης έχει υπολογιστεί, η παράμετρος μ είτε είναι μέσα σε αυτό το διάστημα είτε δεν είναι (η πιθανότητα είναι 0 ή 1).

Παράγοντες που Επηρεάζουν το Εύρος του Διαστήματος Εμπιστοσύνης

Ένα διάστημα εμπιστοσύνης για να είναι χρήσιμο στην λήψη αποφάσεων, θα πρέπει το εύρος του να είναι μικρό. Το εύρος (ή πλάτος) ενός διαστήματος εμπιστοσύνης είναι η διαφορά μεταξύ του άνω και του κάτω ορίου του. Ένα στενότερο διάστημα υποδηλώνει μεγαλύτερη ακρίβεια στην εκτίμηση της πληθυσμιακής παραμέτρου.

Βαθμός Εμπιστοσύνης

Αύξηση του βαθμού εμπιστοσύνης (π.χ., από 95% σε 99%) οδηγεί σε διεύρυνση του διαστήματος.



Μέγεθος Δείγματος

Αύξηση του μεγέθους του δείγματος οδηγεί σε στένωση του διαστήματος, καθώς μεγαλύτερα δείγματα παρέχουν περισσότερες πληροφορίες και μειώνουν το τυπικό σφάλμα.

Μεταβλητότητα Πληθυσμού

Μεγαλύτερη μεταβλητότητα στον πληθυσμό οδηγεί σε διεύρυνση του διαστήματος. Όταν τα δεδομένα είναι πιο διεσπαρμένα, η εκτίμηση είναι λιγότερο ακριβής.

Κατασκευή Διαστήματος Εμπιστοσύνης με το **Excel**

Κατασκευάστε το 95% διάστημα εμπιστοσύνης για το μέσο ύψος των πατέρων (στήλη B) στο αρχείο Data Family Heights v1.xlsx,



Επιλογή Δεδομένων

Στην καρτέλα «Δεδομένα», κάντε κλικ στην «Ανάλυση Δεδομένων» και επιλέξτε «Περιγραφικά Στατιστικά Στοιχεία».



Ρύθμιση Παραμέτρων

Επιλέξτε την περιοχή δεδομένων, ορίστε την περιοχή εξόδου και επιλέξτε «Περίληπτικά στατιστικά» και «Βαθμός εμπιστοσύνης για τον μέσο» με τιμή 95.



Ανάλυση Αποτελεσμάτων

Το Excel δημιουργεί πίνακα με τα περιγραφικά στατιστικά, συμπεριλαμβανομένου του μέσου όρου και του περιθωρίου σφάλματος.

Με αυτά τα στοιχεία μπορούμε εύκολα να υπολογίσουμε το διάστημα εμπιστοσύνης προσθέτοντας και αφαιρώντας το περιθώριο σφάλματος από τον μέσο όρο. Η διαδικασία αυτή μας επιτρέπει να έχουμε μια αξιόπιστη εκτίμηση για την πραγματική τιμή της παραμέτρου του πληθυσμού.

Παράδειγμα **95%** Διαστήματος Εμπιστοσύνης

Κατασκευάστε το 95% διάστημα εμπιστοσύνης για το μέσο ύψος των πατέρων (στήλη B) στο αρχείο Data Family Heights v1.xlsx,

Υπολογισμός Μέσου Όρου

Από το δείγμα υπολογίζουμε το μέσο ύψος των πατέρων που είναι 1.759 μέτρα.

Εύρεση Περιθωρίου Σφάλματος

Το Excel υπολογίζει το περιθώριο σφάλματος για 95% διάστημα εμπιστοσύνης, που είναι 0.01165.

Κατασκευή Διαστήματος

Κάτω όριο: $1.759 - 0.01165 = 1.747$ μέτρα

Άνω όριο: $1.759 + 0.01165 = 1.771$ μέτρα

Άρα, έχουμε 95% εμπιστοσύνη ότι το πραγματικό μέσο ύψος (μ) του πληθυσμού των πατέρων βρίσκεται μεταξύ 1.747 και 1.771 μέτρα.

Έλεγχος Υποθέσεων στη Στατιστική

Ο έλεγχος υποθέσεων αποτελεί μια από τις σημαντικότερες μεθόδους της επαγωγικής στατιστικής. Μας επιτρέπει να λαμβάνουμε αποφάσεις ή να εξάγουμε συμπεράσματα σχετικά με έναν πληθυσμό, βασιζόμενοι σε δεδομένα από ένα δείγμα.

Αντί να εκτιμούμε απλώς μια άγνωστη παράμετρο, ο έλεγχος υποθέσεων μας βοηθά να αξιολογήσουμε την εγκυρότητα μιας συγκεκριμένης δήλωσης ή ισχυρισμού σχετικά με αυτή την άγνωστη παράμετρο. Η διαδικασία περιλαμβάνει τη διατύπωση δύο ανταγωνιστικών υποθέσεων και στη συνέχεια τη χρήση δειγματικών δεδομένων για να καθορίσουμε ποια υπόθεση υποστηρίζεται περισσότερο από τα στοιχεία.

Μηδενική και Εναλλακτική Υπόθεση

Μηδενική Υπόθεση (**H₀**)

Είναι μια δήλωση για μια πληθυσμιακή παράμετρο που συνήθως αντιπροσωπεύει την τρέχουσα πεποίθηση, την κατάσταση "καμία αλλαγή", "καμία επίδραση" ή "καμία διαφορά". Είναι η υπόθεση που ως τώρα θεωρείται αληθής και προσπαθούμε να βρούμε στοιχεία εναντίον της.

Η μηδενική υπόθεση περιέχει πάντα μια μορφή της ισότητας ($=$, $\leq$, ή $\geq$).

Εναλλακτική Υπόθεση (**H₁** ή **H_a**)

Είναι μια δήλωση που έρχεται σε αντίθεση με τη μηδενική υπόθεση. Αντιπροσωπεύει αυτό που ο ερευνητής πιστεύει ότι μπορεί να είναι αληθές ή αυτό που θέλει να αποδείξει.

Η εναλλακτική υπόθεση δεν περιέχει ποτέ κάποια μορφή ισότητας ($\neq$, $<$, ή $>$).

Παραδείγματα Ελέγχων Υποθέσεων

Διάρκεια Ζωής Μπαταρίας

$H_0: \mu=500$ (η μέση διάρκεια ζωής της μπαταρίας είναι 500 ώρες)

$H_1: \mu>500$ (η μέση διάρκεια ζωής της μπαταρίας είναι μεγαλύτερη από 500 ώρες)

Αποτελεσματικότητα Φαρμάκων

$H_0: \pi_1=\pi_2$ (δεν υπάρχει διαφορά στην αποτελεσματικότητα των δύο φαρμάκων)

$H_1: \pi_1\neq\pi_2$ (υπάρχει διαφορά στην αποτελεσματικότητα των δύο φαρμάκων)

Ποσοστό Ελαττωματικών Προϊόντων

$H_0: \pi\geq 0.05$ (το ποσοστό ελαττωματικών προϊόντων είναι τουλάχιστον 5%)

$H_1: \pi<0.05$ (το ποσοστό ελαττωματικών προϊόντων είναι μικρότερο από 5%)

Τύποι Σφαλμάτων στον Έλεγχο Υποθέσεων

Σφάλμα Τύπου I

Συμβαίνει όταν απορρίπτουμε τη μηδενική υπόθεση (H_0) ενώ αυτή είναι αληθής. Η πιθανότητα διάπραξης σφάλματος Τύπου I συμβολίζεται με α και ονομάζεται επίπεδο σημαντικότητας του ελέγχου.



Επίπεδο Σημαντικότητας

Το επίπεδο σημαντικότητας του ελέγχου καθορίζεται από τον ερευνητή εκ των προτέρων του ελέγχου, και συνήθως επιλέγεται $\alpha=0.05$, δηλαδή 5%, (ή 0.01).

Σφάλμα Τύπου II

Συμβαίνει όταν δεν απορρίπτουμε τη μηδενική υπόθεση (H_0) ενώ αυτή είναι ψευδής. Αυτό σημαίνει ότι η εναλλακτική υπόθεση H_1 είναι αληθής. Η πιθανότητα διάπραξης σφάλματος Τύπου II συμβολίζεται με β .

Τύποι Σφαλμάτων στον Έλεγχο Υποθέσεων

	H_0 είναι Αληθής	H_0 είναι Ψευδής (H_1 είναι Αληθής)
Δεν απορρίπτεται η H_0	Σωστή Απόφαση (Πιθανότητα $1-\alpha$)	Σφάλμα Τύπου II (Πιθανότητα β)
Απορρίπτεται η H_0	Σφάλμα Τύπου I (Πιθανότητα α)	Σωστή Απόφαση (Ισχύς = $1-\beta$)

Στον έλεγχο υποθέσεων, υπάρχει πάντα η πιθανότητα να κάνουμε λάθος, καθώς η απόφασή μας βασίζεται σε ένα δείγμα και όχι σε ολόκληρο τον πληθυσμό. Το Σφάλμα Τύπου I συμβαίνει όταν απορρίπτουμε τη μηδενική υπόθεση ενώ αυτή είναι αληθής, με πιθανότητα α (επίπεδο σημαντικότητας). Το Σφάλμα Τύπου II συμβαίνει όταν δεν απορρίπτουμε τη μηδενική υπόθεση ενώ αυτή είναι ψευδής, με πιθανότητα β .

Διαδικασία Λήψης Απόφασης



Υπολογισμός Στατιστικής Συνάρτησης

Υπολογίζουμε την κατάλληλη στατιστική συνάρτηση ελέγχου από τα δεδομένα του δείγματος. Αυτή μετράει πόσο απέχει η εκτίμηση από το δείγμα από την τιμή της παραμέτρου στην μηδενική υπόθεση.



Καθορισμός Περιοχής Απόρριψης

Με βάση το επίπεδο σημαντικότητας α καθορίζεται η περιοχή απόρριψης. Η περιοχή απόρριψης είναι το σύνολο των τιμών της στατιστικής συνάρτησης ελέγχου για τις οποίες θα απορρίψουμε τη μηδενική υπόθεση.



Σύγκριση και Απόφαση

Συγκρίνουμε την τιμή της στατιστικής συνάρτησης ελέγχου με τα όρια της περιοχής απόρριψης. Εναλλακτικά, συγκρίνουμε την τιμή- p με το επίπεδο σημαντικότητας α .



Εξαγωγή Συμπεράσματος

Αν η τιμή- $p \leq \alpha$, τότε απορρίπτουμε την H_0 . Αν η τιμή- $p > \alpha$, τότε δεν απορρίπτουμε την H_0 .

Τιμή- p (p -value)



Ορισμός

Η πιθανότητα, υποθέτοντας ότι η μηδενική υπόθεση H_0 είναι αληθής, να παρατηρήσουμε μια τιμή της στατιστικής συνάρτησης ελέγχου τόσο ακραία ή περισσότερο ακραία από αυτή που πράγματι υπολογίστηκε από το δείγμα.



Σύγκριση με α

Εάν τιμή- $p \leq \alpha$, τότε απορρίπτουμε την H_0 . Εάν τιμή- $p > \alpha$, τότε δεν απορρίπτουμε την H_0 .



Ερμηνεία

Μικρή τιμή- p υποδηλώνει ισχυρές ενδείξεις κατά της μηδενικής υπόθεσης. Μεγάλη τιμή- p υποδηλώνει ότι τα δεδομένα είναι συμβατά με τη μηδενική υπόθεση.

Τύποι Ελέγχων Υποθέσεων

Οι έλεγχοι υποθέσεων κατηγοριοποιούνται ως μονόπλευροι έλεγχοι και δίπλευροι έλεγχοι. Αναλυτικά,

- Δίπλευρος Έλεγχος (two-tailed test) είναι ο έλεγχος στον οποίο η εναλλακτική υπόθεση είναι της μορφής $H_1: \text{παράμετρος} \neq \text{τιμή}$. Η περιοχή απόρριψης χωρίζεται περιοχές που βρίσκονται στις δύο «ουρές» της κατανομής, με $\alpha/2$ σε κάθε ουρά.
- Μονόπλευρος Έλεγχος (one-tailed test) είναι ο έλεγχος στον οποίο η εναλλακτική υπόθεση είναι της μορφής $H_1: \text{παράμετρος} > \text{τιμή}$ (δεξιόπλευρος) ή $H_1: \text{παράμετρος} < \text{τιμή}$ (αριστερόπλευρος). Στους μονόπλευρους έλεγχοι η περιοχή απόρριψης α βρίσκεται σε μία ουρά της κατανομής.

Παράδειγμα Ελέγχου Υποθέσεων με **Excel**

- Θέλουμε να ελέγξουμε σε επίπεδο σημαντικότητας $\alpha=5\%$, κατά πόσο το πραγματικό μέσο ύψος των πατέρων στον πληθυσμό διαφέρει από το 1.80 m. Για το σκοπό αυτό θα χρησιμοποιήσουμε τα δεδομένα (δείγμα μεγέθους 134) που περιέχονται στο αρχείο Family Heights v1.xlsx στη στήλη B

1ος τρόπος

- Θα κατασκευάσουμε το 95% διάστημα εμπιστοσύνης για το μέσο ύψος.
 - Αν η τιμή 1.80 (τιμή στην H_0) δεν περιλαμβάνεται σε αυτό το διάστημα τότε θα απορρίψουμε την μηδενική υπόθεση, και θα θεωρήσουμε ότι το μέσο ύψος των πατέρων στον πληθυσμό είναι διαφορετικό από το 1.80.
 - Αν η τιμή 1.80 περιλαμβάνεται σε αυτό το διάστημα, τότε δεν θα απορρίψουμε την μηδενική υπόθεση, και θα θεωρήσουμε ότι το ύψος των πατέρων στον πληθυσμό είναι το 1.80. Για να κατασκευάσουμε το διάστημα εμπιστοσύνης ακολουθούμε τα βήματα που παρουσιάστηκαν στην ενότητα για το διάστημα εμπιστοσύνης.

2ος τρόπος

- Αντιγράφουμε τα δεδομένα του δείγματος σε ένα νέο φύλλο εργασίας στην στήλη Α.
- Στην διπλανή στήλη σε τόσα κελιά όσα και τα δεδομένα του δείγματος πληκτρολογούμε την τιμή της μηδενικής υπόθεσης, δηλαδή το 1.80

Παράδειγμα Ελέγχου Υποθέσεων με Excel



Διατύπωση Υποθέσεων

Έλεγχος σε επίπεδο σημαντικότητας $\alpha=5\%$ κατά πόσο το πραγματικό μέσο ύψος των πατέρων στον πληθυσμό διαφέρει από το 1.80 m.

$H_0: \mu=1.80$

$H_1: \mu \neq 1.80$



Ανάλυση με Excel

Χρησιμοποιούμε την "Ανάλυση Δεδομένων" (Data Analysis) και επιλέγουμε "t-test του μέσου δυο δειγμάτων συσχετισμένων ζευγών" (t-test paired two sample for the mean).



Αποτελέσματα

Η p-τιμή ($P(T \leq t)$ δίπλευρη) είναι 1.45E-10, που είναι μικρότερη από το 5%.



Συμπέρασμα

Απορρίπτουμε τη μηδενική υπόθεση. Υπάρχουν επαρκείς ενδείξεις ότι το μέσο ύψος των πατέρων στον πληθυσμό διαφέρει από το 1.80 m.

Ο Έλεγχος Ανεξαρτησίας χ^2

Ο έλεγχος ανεξαρτησίας χ^2 αποσκοπεί στον προσδιορισμό του κατά πόσον υπάρχει στατιστικά σημαντική συσχέτιση μεταξύ δύο κατηγορικών μεταβλητών. Εξετάζει εάν η κατανομή των συχνοτήτων της μιας μεταβλητής είναι ανεξάρτητη από την κατανομή των συχνοτήτων της άλλης μεταβλητής.

Ο έλεγχος αυτός, γνωστός και ως Έλεγχος Συσχέτισης χ^2 , χρησιμοποιείται ευρέως σε διάφορους τομείς έρευνας, όπως οι κοινωνικές επιστήμες, η ιατρική, το μάρκετινγκ και η βιολογία, για την ανάλυση δεδομένων που προέρχονται από έρευνες, πειράματα ή παρατηρήσεις.

Βασική Αρχή του Ελέγχου χ^2

Παρατηρούμενες Συχνότητες

Ο έλεγχος ανεξαρτησίας χ^2 συγκρίνει τις παρατηρούμενες συχνότητες σε κάθε κελί ενός πίνακα συνάφειας με τις αναμενόμενες συχνότητες που θα προέκυπταν αν οι δύο μεταβλητές ήταν απόλυτα ανεξάρτητες.

Αναμενόμενες Συχνότητες

Όταν οι διαφορές μεταξύ των παρατηρούμενων και των αναμενόμενων συχνοτήτων είναι αρκετά μεγάλες, τότε απορρίπτεται η υπόθεση της ανεξαρτησίας και συνάγεται το συμπέρασμα ότι υπάρχει κάποια μορφή σχέσης μεταξύ των μεταβλητών.

Χαρακτηριστικά του Ελέγχου χ^2



Μη Παραμετρικός Έλεγχος

Ο έλεγχος χ^2 είναι ένας μη παραμετρικός στατιστικός έλεγχος, που δεν απαιτεί συγκεκριμένες παραδοχές σχετικά με την κατανομή των δεδομένων στον πληθυσμό.



Κατάλληλος για Κατηγορικές Μεταβλητές

Αυτή η ιδιότητα καθιστά τον έλεγχο χ^2 ιδιαίτερα ευέλικτο και κατάλληλο για ένα ευρύ φάσμα περιπτώσεων, όταν οι μεταβλητές είναι κατηγορικές.



Περιορισμοί Χρήσης

Δεν είναι κατάλληλος όταν οι κατηγορικές μεταβλητές αντιπροσωπεύουν μετρήσεις "πριν-μετά" στα ίδια άτομα, καθώς παραβιάζεται η παραδοχή της ανεξαρτησίας των παρατηρήσεων.

Υποθέσεις του Ελέγχου χ^2

Μηδενική Υπόθεση (**H₀**)

Δεν υπάρχει σχέση (συσχέτιση ή εξάρτηση) μεταξύ των δύο κατηγορικών μεταβλητών στον πληθυσμό από τον οποίο προήλθε το δείγμα.

Ισοδύναμα, H₀: Η [Μεταβλητή 1] είναι ανεξάρτητη από την [Μεταβλητή 2].

Εναλλακτική Υπόθεση (**H₁**)

Υπάρχει σχέση (συσχέτιση ή εξάρτηση) μεταξύ των δύο κατηγορικών μεταβλητών στον πληθυσμό.

Ισοδύναμα, H₁: Η [Μεταβλητή 1] δεν είναι ανεξάρτητη από την [Μεταβλητή 2].

Υπολογισμός Αναμενόμενων Συχνοτήτων

Για κάθε κελί του πίνακα συνάφειας, η αναμενόμενη συχνότητα E_{ij} (για το κελί στη γραμμή i και στήλη j) υπολογίζεται με τον παραπάνω τύπο. Οι διαφορές μεταξύ παρατηρούμενων και αναμενόμενων συχνοτήτων οφείλονται στο γεγονός ότι χρησιμοποιούμε δείγμα.



Τύπος Υπολογισμού

$$E_{ij} = (\text{Άθροισμα γραμμής } i) \times (\text{Άθροισμα στήλης } j) / \text{Συνολικό μέγεθος δείγματος } (N)$$



Εφαρμογή σε Πίνακα Συνάφειας

Εφαρμόζεται σε κάθε κελί του πίνακα



Έλεγχος Ανεξαρτησίας

Αν οι μεταβλητές είναι ανεξάρτητες, οι αναμενόμενες συχνότητες θα είναι κοντά στις παρατηρούμενες

Στατιστική Συνάρτηση Ελέγχου

Τύπος Στατιστικής Συνάρτησης

$$\chi^2_{\text{stat}} = \sum (O_{ij} - E_{ij})^2 / E_{ij}$$

Ερμηνεία

Μετρά τη διαφορά μεταξύ παρατηρούμενων και αναμενόμενων συχνοτήτων

Υπολογισμός στο **Excel**

Χρήση της συνάρτησης

CHISQ.TEST(παρατηρούμενες τιμές, αναμενόμενες τιμές)

Η στατιστική συνάρτηση ελέγχου συγκρίνει τις παρατηρούμενες συχνότητες (O_{ij}) με τις αναμενόμενες συχνότητες (E_{ij}) για κάθε κελί του πίνακα συνάφειας. Όσο μεγαλύτερες είναι οι διαφορές, τόσο μεγαλύτερη είναι η τιμή της στατιστικής συνάρτησης, υποδεικνύοντας πιθανή συσχέτιση μεταξύ των μεταβλητών.

Παράδειγμα Εφαρμογής

	Προτιμά Α	Προτιμά Β	Προτιμά Γ	Σύνολο
Νέοι	20	15	10	45
Μεσαίας ηλικίας	25	30	20	75
Ηλικιωμένοι	15	25	40	80
Σύνολο	60	70	70	200

Σε μία έρευνα αγοράς ρωτήθηκαν 200 άτομα σχετικά με ποιο από τα 3 προϊόντα (Α, Β ή Γ) προτιμούν. Τα άτομα επίσης κατηγοριοποιήθηκαν με βάση την ηλικία τους ως νέοι, μεσήλικες και ηλικιωμένοι. Ο παραπάνω πίνακας συνάφειας παρουσιάζει τις προτιμήσεις σε σχέση με την ηλικία τους.

Ερμηνεία Αποτελεσμάτων

Υπολογισμός Αναμενόμενων
Συχνοτήτων

Για το πρώτο κελί: $E11 = (45 \times 60) / 200 = 13.5$

Συμπέρασμα

Απορρίπτουμε τη μηδενική υπόθεση της
ανεξαρτησίας



Χρήση **Excel**

=CHISQ.TEST(παρατηρούμενες
συχνότητες, αναμενόμενες συχνότητες)

Αποτέλεσμα

$p\text{-τιμή} = 0.0025 < \alpha = 0.05$

Το Excel υπολογίζει την p -τιμή ίση με 0.0025 που είναι μικρότερη από το επίπεδο σημαντικότητας $\alpha = 0.05$. Κατά συνέπεια απορρίπτουμε την μηδενική υπόθεση, δηλαδή την υπόθεση ότι οι δύο μεταβλητές είναι ανεξάρτητες. Άρα έχουμε επαρκείς ενδείξεις ότι οι δύο μεταβλητές σχετίζονται.

Ασκήσεις κεφαλαίου

- Χρησιμοποιώντας τα δεδομένα από το αρχείο Family Heights v1.xlsx, κατασκευάστε το 95% διάστημα εμπιστοσύνης για το μέσο ύψος στο πληθυσμό
 - των μητέρων (στήλη C)
 - των γιών (στήλη D)
 - των κοριτσιών (στήλη E)
- Ελέγξτε σε επίπεδο σημαντικότητας $\alpha=0.05$, κατά πόσο το μέσο ύψος των γιων διαφέρει από το 1.80
- Ο παρακάτω πίνακας συνάφειας παρουσιάζει το φύλο και το εκπαιδευτικό επίπεδο, 394 ατόμων που απάντησαν σε μία σχετική έρευνα. Εφαρμόστε την κατάλληλη στατιστική τεχνική για να απαντήσετε στο ερευνητικό ερώτημα, αν το επίπεδο εκπαίδευσης είναι ανεξάρτητο από το φύλο;

Φύλο	Απόφοιτος Λυκείου	Πτυχίο	Μεταπτυχιακό	Διδακτορικό	Σύνολο Γραμμής
Γυναίκα	60	54	46	41	201
Άνδρας	40	44	53	57	194
Σύνολο Στήλης	100	98	99	98	395 (N)

Συσχέτιση και Απλή Γραμμική Παλινδρόμηση

Κεφάλαιο 8

Εισαγωγή

Η κατανόηση των σχέσεων μεταξύ μεταβλητών αποτελεί θεμελιώδη επιδίωξη σε ένα ευρύ φάσμα επιστημονικών κλάδων, από τις θετικές επιστήμες έως τις κοινωνικές επιστήμες και τη διοίκηση επιχειρήσεων.

Η Στατιστική παρέχει τις κατάλληλες μεθόδους για την αξιολόγηση της ύπαρξης σχέσης και της ισχύος της μεταξύ δύο αριθμητικών συνεχών μεταβλητών. Αυτή η σχέση, στα πλαίσια της Στατιστικής, ονομάζεται συσχέτιση.

Συσχέτιση **vs** Αιτιότητα

Συσχέτιση

Υποδεικνύει ότι δύο ή περισσότερες μεταβλητές τείνουν να μεταβάλλονται από κοινού με προβλέψιμο τρόπο, είτε προς την ίδια κατεύθυνση (θετική) είτε προς την αντίθετη (αρνητική).

Αιτιότητα

Υποδηλώνει ότι η μεταβολή μιας μεταβλητής προκαλεί τη μεταβολή μιας άλλης. Απαιτεί πολύ πιο αυστηρές μεθόδους από την απλή παρατήρηση συσχέτισης.

Κοινή Παρανόηση

Η ανθρώπινη φύση τείνει να αναζητά μοτίβα και να συνάγει αιτίες όταν δύο γεγονότα συμβαίνουν συστηματικά μαζί, αλλά χωρίς περαιτέρω ανάλυση, μια υπόθεση αιτιότητας είναι επισφαλής.

Η Έννοια της Συσχέτισης

Η συσχέτιση, με τη στατιστική έννοια, αναφέρεται σε οποιαδήποτε στατιστική σχέση μεταξύ δύο τυχαίων μεταβλητών. Στη γραμμική συσχέτιση, εξετάζουμε τον βαθμό στον οποίο δύο συνεχείς μεταβλητές μεταβάλλονται μαζί με γραμμικό τρόπο.

Η συσχέτιση εκτιμάται μεταξύ δύο μεταβλητών (X και Y) που είναι συνεχείς, δηλαδή μεταβλητών που μετρούνται σε κλίμακα διαστήματος ή αναλογίας.

Ο σκοπός της ανάλυσης συσχέτισης είναι να προσδιοριστεί αν οι μεταβλητές συν-μεταβάλλονται και σε ποιο βαθμό, παρέχοντας μια ποσοτική εκτίμηση της δύναμης και της κατεύθυνσης αυτής της γραμμικής σχέσης.

Ο Συντελεστής Συσχέτισης **Pearson**



Μαθηματικός Τύπος

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$



Εύρος Τιμών

Από -1 έως +1



Ερμηνεία

Κατεύθυνση και ισχύς της γραμμικής σχέσης

Ο πιο διαδεδομένος δείκτης για τη μέτρηση της γραμμικής συσχέτισης μεταξύ δύο συνεχών μεταβλητών είναι ο συντελεστής συσχέτισης του Pearson, ο οποίος συμβολίζεται με r . Μια τιμή $r=+1$ υποδηλώνει μια τέλεια θετική γραμμική σχέση, ενώ $r=-1$ υποδηλώνει μια τέλεια αρνητική γραμμική σχέση. Τέλος, $r=0$ υποδηλώνει την απουσία γραμμικής συσχέτισης.

Κατεύθυνση Συσχέτισης



Θετική Συσχέτιση ($r > 0$)

Οι μεταβλητές κινούνται προς την ίδια κατεύθυνση. Όταν η μία αυξάνεται, και η άλλη τείνει να αυξάνεται.



Μηδενική Συσχέτιση ($r = 0$)

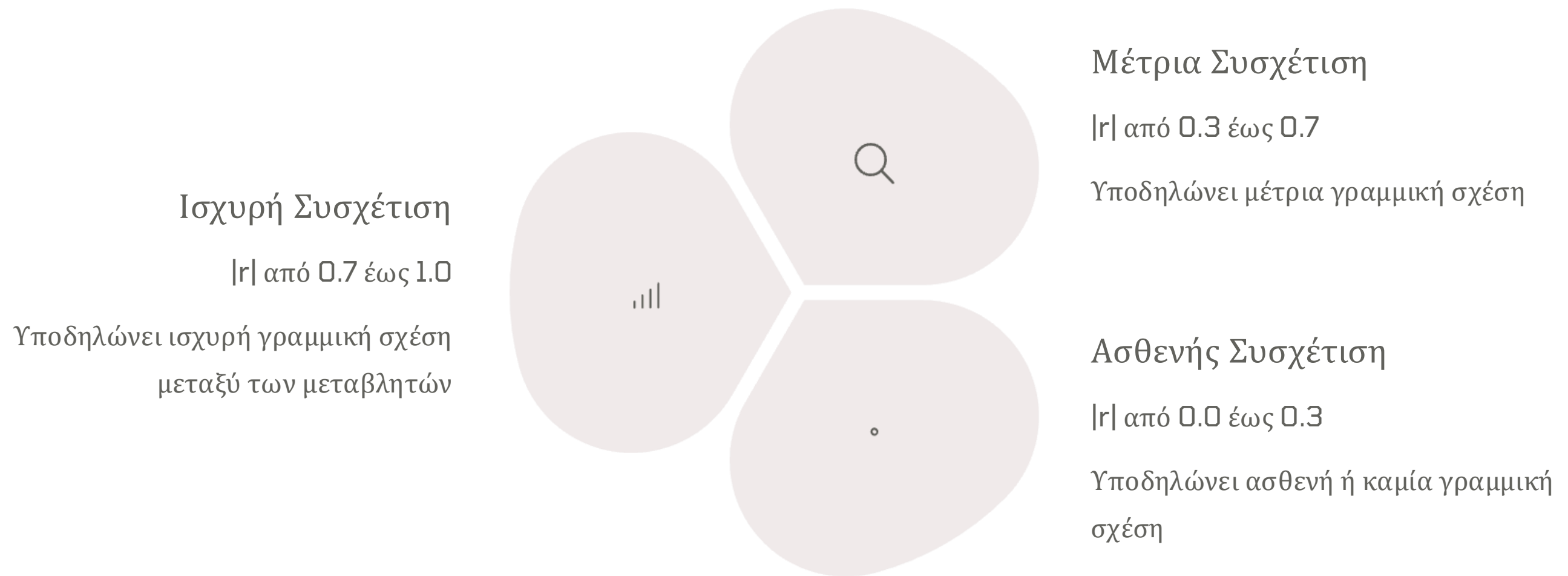
Δεν υπάρχει γραμμική σχέση μεταξύ των μεταβλητών. Αυτό δεν αποκλείει την ύπαρξη μη γραμμικής σχέσης.



Αρνητική Συσχέτιση ($r < 0$)

Οι μεταβλητές κινούνται προς αντίθετες κατευθύνσεις. Όταν η μία αυξάνεται, η άλλη τείνει να μειώνεται.

Ισχύς της Συσχέτισης



Είναι κρίσιμο να θυμόμαστε ότι η παραπάνω κατηγοριοποίηση των τιμών του συντελεστή συσχέτισης είναι σχετική και η πρακτική σημασία μιας συγκεκριμένης τιμής r εξαρτάται πάντα από το πλαίσιο της έρευνας.

Παράδειγμα Συσχέτισης: Ώρες Μελέτης και Βαθμοί

Φοιτητής	Ώρες Μελέτης (X)	Βαθμός Εξέτασης (Y)
1	2	55
2	3	60
3	1	45
4	5	75
5	4	70
6	6	85
7	2	50
8	3	65
9	7	90
10	4	68

Υπολογισμός Συσχέτισης στο Excel

Εισαγωγή Δεδομένων

Τα δεδομένα εισάγονται σε δύο στήλες στο Excel, π.χ., οι ώρες μελέτης στη στήλη Α (A2:A11) και οι βαθμοί στη στήλη Β (B2:B11).

Επιλογή Συνάρτησης

Για τον υπολογισμό του συντελεστή συσχέτισης Pearson r στο Excel, μπορούν να χρησιμοποιηθούν δύο ισοδύναμες συναρτήσεις: CORREL() ή PEARSON().

Υπολογισμός

Σε ένα κενό κελί, π.χ. D2, εισάγεται ο τύπος: PEARSON(A2:A11, B2:B11) ή CORREL(A2:A11, B2:B11).

Για τα παραπάνω δεδομένα, ο συντελεστής συσχέτισης r είναι περίπου 0.96, υποδεικνύοντας μια πολύ ισχυρή θετική γραμμική συσχέτιση μεταξύ των ωρών μελέτης και των βαθμών εξέτασης. Αυτό σημαίνει ότι, γενικά, όσο περισσότερες ώρες μελετούν οι φοιτητές, τόσο υψηλότερους βαθμούς τείνουν να λαμβάνουν.

Εισαγωγή στην Απλή Γραμμική Παλινδρόμηση

Η απλή γραμμική παλινδρόμηση είναι μια στατιστική μέθοδος που χρησιμοποιείται για τη μοντελοποίηση της γραμμικής σχέσης μεταξύ μιας εξαρτημένης (Y) και μιας ανεξάρτητης (X) μεταβλητής. Αυτό επιτυγχάνεται με την προσαρμογή μιας ευθείας γραμμής στα παρατηρούμενα δεδομένα.

Ο σκοπός της απλής γραμμικής παλινδρόμησης είναι διττός: αφενός η πρόβλεψη της τιμής της εξαρτημένης μεταβλητής (Y) για μια τιμή της ανεξάρτητης μεταβλητής (X), και αφετέρου η κατανόηση της σχέσης μεταξύ των δύο μεταβλητών.

Το Μοντέλο της Απλής Γραμμικής Παλινδρόμησης

$f(x)$

Μαθηματική Εξίσωση

$Y = \beta_0 + \beta_1 X + \varepsilon$, όπου Y είναι η εξαρτημένη μεταβλητή, X η ανεξάρτητη, β_0 ο σταθερός όρος, β_1 ο συντελεστής κλίσης και ε ο όρος σφάλματος.



Μέθοδος Υπολογισμού

Για τον προσδιορισμό των συντελεστών β_0 και β_1 χρησιμοποιείται συνήθως η μέθοδος των ελαχίστων τετραγώνων, που ελαχιστοποιεί το άθροισμα των τετραγώνων των σφαλμάτων.



Ερμηνεία Συντελεστών

Ο β_0 αντιπροσωπεύει την εκτιμώμενη μέση τιμή της Y όταν $X=0$, ενώ ο β_1 δείχνει την εκτιμώμενη μεταβολή στη μέση τιμή της Y για κάθε μοναδιαία αύξηση της X .

Βασικές Παραδοχές της Γραμμικής Παλινδρόμησης

Γραμμικότητα

Η σχέση μεταξύ της ανεξάρτητης μεταβλητής (X) και της εξαρτημένης μεταβλητής (Y) πρέπει να είναι γραμμική.

Ανεξαρτησία Σφαλμάτων

Τα σφάλματα (ϵ_i) πρέπει να είναι ανεξάρτητα μεταξύ τους. Το σφάλμα για μια παρατήρηση δεν πρέπει να σχετίζεται με το σφάλμα για οποιαδήποτε άλλη παρατήρηση.

Ομοσκεδαστικότητα

Η διακύμανση των σφαλμάτων (ϵ_i) πρέπει να είναι σταθερή για όλες τις τιμές της ανεξάρτητης μεταβλητής X .

Κανονικότητα Σφαλμάτων

Τα σφάλματα (ϵ_i) πρέπει να ακολουθούν μια κανονική κατανομή, με μέση τιμή μηδέν.

Εφαρμογή Ανάλυσης Παλινδρόμησης με το **Excel**



Προετοιμασία Δεδομένων

Οργανώστε τα δεδομένα σας σε στήλες, με την εξαρτημένη μεταβλητή (Y) και την ανεξάρτητη μεταβλητή (X) σε ξεχωριστές στήλες.



Άνοιγμα Εργαλείου Ανάλυσης

Από την καρτέλα "Δεδομένα" (Data), κάντε κλικ στο "Ανάλυση Δεδομένων" (Data Analysis).



Ρύθμιση Παραμέτρων

Επιλέξτε "Παλινδρόμηση" (Regression), καθορίστε τις περιοχές εισόδου Y και X, και επιλέξτε τις επιθυμητές επιλογές εξόδου.



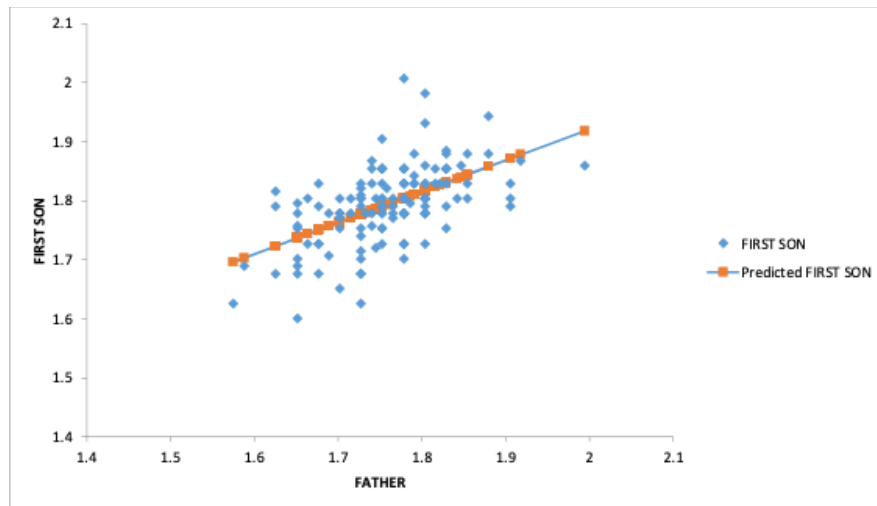
Ανάλυση Αποτελεσμάτων

Εξετάστε τους πίνακες με τα στατιστικά της παλινδρόμησης, τους συντελεστές και το γράφημα της ευθείας παλινδρόμησης.

Παράδειγμα: Πρόβλεψη Ύψους Γιου από Ύψος Πατέρα



Στο παράδειγμα χρησιμοποιούνται δεδομένα από 134 ζεύγη πατέρων-γιων, με τα ύψη των πατέρων στη στήλη B και τα ύψη των γιων στη στήλη D του αρχείου Excel.



Το γράφημα της ευθείας παλινδρόμησης δείχνει τη σχέση μεταξύ των υψών πατέρων και γιων, με τα σημεία να αντιπροσωπεύουν τα πραγματικά δεδομένα και την ευθεία γραμμή την εκτιμημένη σχέση.

Ερμηνεία των Στατιστικών Παλινδρόμησης

0.304

Συντελεστής Προσδιορισμού (R^2)

Το 30.4% της μεταβλητότητας του ύψους του γιου εξηγείται από το ύψος του πατέρα.

0.552

Συντελεστής Συσχέτισης (R)

Υπάρχει μέτρια θετική συσχέτιση μεταξύ των υψών πατέρων και γιων.

Regression Statistics		
Πολλαπλό R	Multiple R	0.55156
R Τετράγωνο	R Square	0.30422
Προσαρμοσμένο R Τετράγωνο	Adjusted R Square	0.29895
Τυπικό σφάλμα	Standard Error	0.05469
Μέγεθος δείγματος	Observations	134

Ερμηνεία των Συντελεστών Παλινδρόμησης

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0.8642	0.1224	7.0577	0.0000	0.6220	1.1064
FATHER	0.5284	0.0696	7.5970	0.0000	0.3908	0.6660

Η εξίσωση παλινδρόμησης είναι: $\text{[Ύψος γιού]}^{\wedge} = 0.8642 + 0.5284 \cdot \text{[Ύψος πατέρα]}$. Ο συντελεστής κλίσης 0.5284 σημαίνει ότι για κάθε αύξηση του ύψους του πατέρα κατά 1 εκατοστό, το ύψος του γιου αυξάνεται κατά 0.5284 εκατοστά.

Η p-τιμή του συντελεστή κλίσης είναι 0.0000 (< 0.05), που σημαίνει ότι η κλίση της ευθείας παλινδρόμησης είναι στατιστικά σημαντική σε επίπεδο σημαντικότητας 5%. Άρα, το μοντέλο παλινδρόμησης είναι στατιστικά σημαντικό.

Χρήση του Μοντέλου για Πρόβλεψη



Εφαρμογή της Εξίσωσης

Χρησιμοποιούμε την εξίσωση (Ύψος γιού) = $0.8642 + 0.5284 \cdot (\text{Ύψος πατέρα})$



Εισαγωγή Δεδομένων

Για πατέρα με ύψος 1.80 μέτρα



Υπολογισμός Πρόβλεψης

$0.8642 + 0.5284 \cdot (1.8) = 1.815$ μέτρα

Το μοντέλο προβλέπει ότι το μέσο ύψος των γιων που έχουν πατέρα με ύψος 1.80 μέτρα θα είναι περίπου 1.815 μέτρα.

Έλεγχος **Wald**



Μηδενική Υπόθεση (**H0**)

Η μηδενική υπόθεση του ελέγχου Wald είναι ότι ο συντελεστής κλίσης ισούται με μηδέν ($\beta_1 = 0$), που σημαίνει ότι δεν υπάρχει γραμμική σχέση μεταξύ των μεταβλητών.



Εναλλακτική Υπόθεση (**H1**)

Η εναλλακτική υπόθεση είναι ότι ο συντελεστής κλίσης διαφέρει από το μηδέν ($\beta_1 \neq 0$), υποδηλώνοντας την ύπαρξη γραμμικής σχέσης μεταξύ των μεταβλητών.



Στατιστικός Έλεγχος

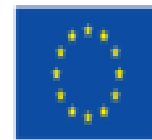
Ο έλεγχος Wald εξετάζει αν ο συντελεστής β_1 διαφέρει στατιστικά σημαντικά από το μηδέν, επιτρέποντάς μας να αξιολογήσουμε την αξιοπιστία του μοντέλου παλινδρόμησης.

- Στον πίνακα των συντελεστών του παραδείγματός μας, υπάρχει στήλη με τις p-τιμές (p-value) τόσο για αυτόν το συντελεστή κλίσης (β_1) όσο και για την σταθερά (β_0). Στο συγκεκριμένο παράδειγμα, η p-τιμή είναι $0.00 < 0.05$, άρα η κλίση της ευθείας παλινδρόμησης είναι στατιστικά σημαντική σε επίπεδο σημαντικότητας 5%.
- Στα πλαίσια της απλής γραμμικής παλινδρόμησης αυτό σημαίνει ότι το μοντέλο παλινδρόμησης είναι στατιστικά σημαντικό.

Ασκήσεις κεφαλαίου

Χρησιμοποιώντας τα δεδομένα από το αρχείο *Family Heights v1.xlsx*:

1. Να υπολογισθεί ο συντελεστής συσχέτισης για τις μεταβλητές, «Ύψος μητέρας» και «Ύψος κόρης» και να σχολιασθεί η συσχέτιση των δύο μεταβλητών.
2. Να κατασκευασθεί μοντέλο παλινδρόμησης που να εξηγεί το ύψος της κόρης σε σχέση με το ύψος της μητέρας.
3. Να γραφεί η εξίσωση της παλινδρόμησης και να ερμηνευθούν οι συντελεστές.
4. Είναι ο συντελεστής κλίσης της παλινδρόμησης στατιστικά σημαντικός;
5. Να δοθεί η ερμηνεία του συντελεστή προσδιορισμού



Με τη χρηματοδότηση
της Ευρωπαϊκής Ένωσης
NextGenerationEU