ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΥΠΟΥΡΓΕΙΟ ΕΣΩΤΕΡΙΚΩΝ

ΕΘΝΙΚΟ ΚΕΝΤΡΟ ΔΗΜΟΣΙΑΣ ΔΙΟΙΚΗΣΗΣ & ΑΥΤΟΔΙΟΙΚΗΣΗΣ

ΕΣΠΑ 2014-2020 Επιχειρησιακό Πρόγραμμα ΜΕΤΑΡΡΥΘΜΙΣΗ ΔΗΜΟΣΙΟΥ ΤΟΜΕΑ

ΕΘΝΙΚΗ ΣΧΟΛΗ ΔΗΜΟΣΙΑΣ ΔΙΟΙΚΗΣΗΣ & ΑΥΤΟΔΙΟΙΚΗΣΗΣ

Η **Διαχείριση Δεδομένων**
και η
Στατιστική Επεξεργασία τους

Συντονιστής Δημήτριος Τσιμάρας
Αξιολογητής Ηλίας Μαραγκός
Συγγραφέας Κωνσταντίνα Ατσάρου

Πρόλογος

Στην σημερινή εποχή η ανάπτυξη των ηλεκτρονικών υπολογιστών και η συλλογή μεγάλου όγκου δεδομένων, έχει οδηγήσει στην ραγδαία ανάπτυξη της εφαρμογής της Στατιστικής. Οι αποφάσεις λαμβάνονται και τεκμηριώνονται με την παρουσίαση και ανάλυση δεδομένων. Με την χρήση διαφόρων στατιστικών προγραμμάτων που έχουν αναπτυχθεί, δίνεται η δυνατότητα σε οργανισμούς, και σε εταιρείες να αναλύουν τα δεδομένα τους και να καθορίζουν την πολιτική τους.

Το εγχειρίδιο αυτό έχει ως στόχο να γίνουν κατανοητές βασικές έννοιες της Στατιστικής καθώς επίσης να αναλυθούν στατιστικές μέθοδοι για την ανάλυση δεδομένων, χωρίς ανάλυση Μαθηματικών τύπων. Το εγχειρίδιο αποτελείται από 5 κεφάλαια τα οποία αναλύουν τα παρακάτω:

1ο. Εισαγωγή σε βασικές έννοιες της Στατιστικής
2ο. Βασικά στατιστικά μέτρα για την περιγραφή των δεδομένων
3ο. Γραφήματα για την απεικόνιση των δεδομένων,
4ο. Βασικές έννοιες της Επαγωγικής Στατιστικής και του Ελέγχου Υποθέσεων
5ο. Συσχέτιση μεταβλητών και Παλινδρόμηση.

Το πρόγραμμα που χρησιμοποιείται για την ανάλυση των δεδομένων είναι το EXCEL 2010.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1^{Ο}

ΕΙΣΑΓΩΓΗ ΣΤΗ ΣΤΑΤΙΣΤΙΚΗ

Η Στατιστική είναι η επιστήμη που ασχολείται με τον σχεδιασμό μιας έρευνας για την συλλογή, την επεξεργασία, την ανάλυση και την παρουσίαση στοιχείων με στόχο την εξαγωγή συμπερασμάτων. Ο πυρήνας της σύγχρονης Στατιστικής είναι η στατιστική συμπερασματολογία που από τις πληροφορίες που συλλέγουμε από ένα κατάλληλα επιλεγμένο δείγμα, οδηγούμαστε σε συμπεράσματα για τον υπό μελέτη πληθυσμό. Η Στατιστική χωρίζεται σε δύο μεγάλους κλάδους την *Περιγραφική Στατιστική* και την *Επαγωγική Στατιστική*.

- Η Περιγραφική Στατιστική (Descriptive Statistics) ασχολείται με την ταξινόμηση και την παρουσίαση των δεδομένων. Η παρουσίαση αυτή γίνεται μέσω διαγραμμάτων και διαφόρων μέτρων.
- Η Επαγωγική Στατιστική (Inferential Statistics) ασχολείται με την εξαγωγή συμπερασμάτων από ένα δείγμα για το σύνολο του πληθυσμού.

Η εφαρμογή των μεθόδων της Στατιστικής στο παρελθόν δεν ήταν εύκολη λόγω του μεγάλου όγκου των δεδομένων που απαιτούνται για την ανάλυση, με την εξέλιξη της τεχνολογίας και την ευρεία χρήση των ηλεκτρονικών υπολογιστών αυτό έγινε εφικτό, έχοντας σαν αποτέλεσμα, η Στατιστική να έχει μεγάλη ανάπτυξη και ευρεία χρήση σε διάφορες επιστήμες.

Η Στατιστική είναι μια επιστήμη που αξιοποιείται από πολλές άλλες επιστήμες, όπως η Μετεωρολογία, η Κοινωνιολογία, το Μάρκετινγκ, η Ιατρική, η Φυσική, η Χημεία, η Διαφήμιση κ.α. Σε μερικές περιπτώσεις μάλιστα είναι τόσο σημαντική η συμβολή της, που έχουν δημιουργηθεί ιδιαίτεροι κλάδοι επιστημών, όπως για παράδειγμα η εκτεταμένη εφαρμογή στατιστικών μεθοδολογιών στην Ιατρική και τη Βιολογία οδήγησε στην εμφάνιση της Βιοστατιστικής

Φυσικά, από τον κανόνα αυτό δεν εξαιρείται η Δημόσια Διοίκηση. Η ανάγκη δημιουργίας δομημένων και τεκμηριωμένων αποφάσεων, μέσω της χρήσης δειγμάτων, ή της εισαγωγής παραγόντων διαχείρισης κινδύνου ή πρόβλεψης μελλοντικών συνθηκών οδηγεί τον λήπτη της απόφασης στην υιοθέτηση στατιστικών εργαλείων και μεθοδολογιών.

1.1. Βασικές Στατιστικές έννοιες[1]

Στο κεφάλαιο αυτό θα εξηγηθούν κάποιες βασικές έννοιες της Στατιστικής που είναι απαραίτητες για το εν λόγω μάθημα. Αυτές είναι οι παρακάτω:

- Η παρατήρηση.
- Ο πληθυσμός,
- Το δείγμα,
- Η απογραφή,
- Η δειγματοληψία,
- Η στατιστική παράμετρος,
- Η στατιστική μεταβλητή ή απλά μεταβλητή.
- Τα μεταδεδομένα.
- Οι ελλείπουσες τιμές.

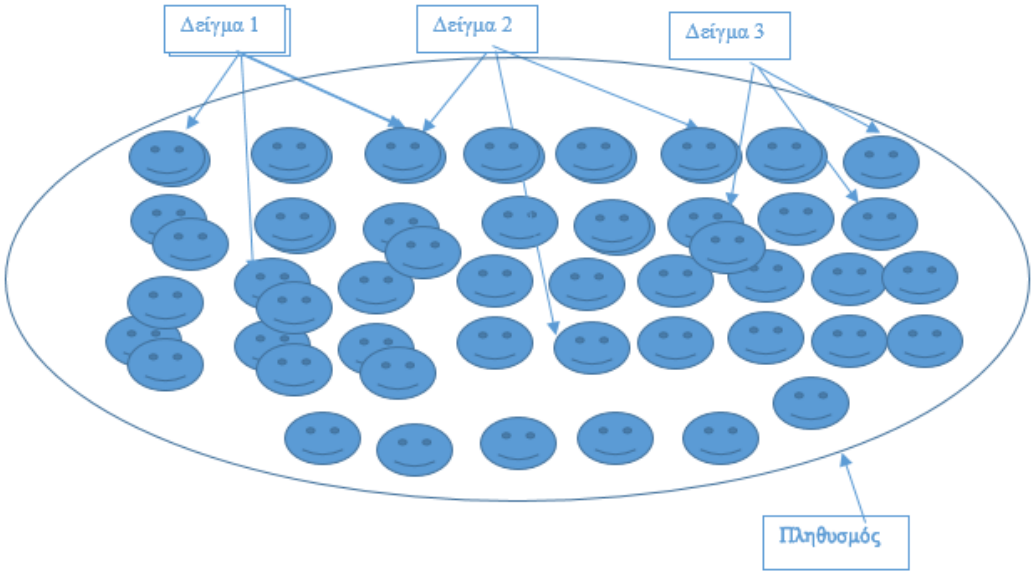
Σε όλα τα παραπάνω θα δίνονται παραδείγματα από την καθημερινότητα.
Για να γίνουν κατανοητές οι παραπάνω έννοιες θα δώσουμε ένα παράδειγμα.

Ένα υπουργείο χρησιμοποίησε μία νέα μέθοδο για την εκπαίδευση των υπαλλήλων του. Το υπουργείο εκπαίδευσε όλους τους υπαλλήλους του με την μέθοδο αυτή. Στην συνέχεια μία υπηρεσία του υπουργείου ανέθεσε σε μία ομάδα υπαλλήλων της, να μετρήσει την ικανοποίηση των υπαλλήλων από την εφαρμογή της νέας αυτής μεθόδου. Αποφασίστηκε να αποτιμηθεί η επίδραση που είχε η εκπαίδευση στην απόδοση των υπαλλήλων μέσω ενός ερωτηματολογίου που θα απαντούσε ένα δείγμα 200 εργαζομένων στο εν λόγω υπουργείο και από την μελέτη των δεδομένων να εκτιμηθεί η άποψη που έχει το σύνολο των υπαλλήλων του υπουργείου για την εκπαιδευτική δράση.
Στο παράδειγμα αυτό:

- Το σύνολο όλων των υπαλλήλων του υπουργείου αποτελεί τον **Πληθυσμό** της έρευνας.
- Το υποσύνολο των 200 υπάλληλων του Υπουργείου που θα συμμετάσχει στην έρευνα αποτελεί το **Δείγμα.**
- Το πλήθος των εργαζομένων που μετέχουν στο δείγμα ονομάζεται **Μέγεθος του δείγματος,** το οποίο στο συγκεκριμένο παράδειγμα είναι 200.
- Αν η έρευνα γινόταν σε όλους τους υπαλλήλους του Υπουργείου τότε θα επρόκειτο για **Απογραφή**, Αντίθετα, λόγω του ότι η έρευνα θα γίνει σε ένα μέρος του

[1] https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Beginners:Statistics_4_beginners/el

πληθυσμού (δείγμα) πρόκειται για έρευνα μέσω **δειγματοληψίας** και πρέπει να καθοριστεί ο τρόπος με τον οποίο θα επιλεγούν οι συμμετέχοντες και συμμετέχουσες στην έρευνα. Ο τρόπος επιλογής αποτελεί τη **μέθοδο δειγματοληψίας**.



1.1Πληθυσμός Δείγμα

Μέσω του ερωτηματολογίου οι υπάλληλοι καλούνται να απαντήσουν ερωτήσεις σχετικές με διάφορες πτυχές της εκπαίδευσης, όπως:

- Η διάρκεια της εκπαίδευσης κρίθηκε ως: υπερβολικά μεγάλη σε διάρκεια; Κανονική σε διάρκεια; μικρή σε διάρκεια;
- Τα θεματικά αντικείμενα σε ποιο βαθμό καλύφτηκαν; (Πολύ Κακό, Κακό, Μέτριο, Καλό, Πολύ καλό).

και να δώσουν προσωπικές πληροφορίες όπως:

- *Τα έτη προϋπηρεσίας στο Δημόσιο.*
- *Το Φύλο.*
- *Το εκπαιδευτικό επίπεδο υπαλλήλου (Υποχρεωτική-Δευτεροβάθμια-Τριτοβάθμια).*
- *Ο αριθμός παιδιών.*

Είναι φανερό από τα παραπάνω ότι προκύπτει η ανάγκη της μέτρησης διάφορων χαρακτηριστικών τα οποία ονομάζονται **στατιστικές μεταβλητές ή απλά μεταβλητές**.

Στο προηγούμενο παράδειγμα τα αποτελέσματα των μετρήσεων που θα προκύψουν από την έρευνα για τις διάφορες πτυχές της εκπαίδευσης αλλά και για τα χαρακτηριστικά των συμμετεχόντων/συμμετεχουσών στο δείγμα (όπως το *φύλο*, τα *έτη προϋπηρεσίας*, το

εκπαιδευτικό επίπεδο, τον *αριθμό παιδιών* κ.ά.) αποτελούν το σύνολο των **Δεδομένων.** Κάθε ένα από τα άτομα που συμμετείχαν στην έρευνα αποτελεί μια δειγματοληπτική μονάδα. Παράλληλα, το υπηρεσιακό αρχείο μέσα από το οποίο επιλέχθηκαν τα στοιχεία του δείγματος ονομάζεται δειγματοληπτικό πλαίσιο.

Χαρακτηριστικά όπως το *φύλο,* και το *εκπαιδευτικό επίπεδο* ονομάζονται **Ποιοτικά ή κατηγορικά χαρακτηριστικά**, ενώ χαρακτηριστικά όπως τα *έτη προϋπηρεσίας* και ο *αριθμός παιδιών* ονομάζονται **Ποσοτικά χαρακτηριστικά.**

Λόγω του γεγονότος ότι το 10% των υπαλλήλων δεν έδωσε απάντηση για το αν ο χρόνος εκπαίδευσης ήταν επαρκής ή όχι, στο σύνολο των δεδομένων εμφανίζονται ελλείπουσες τιμές

Γενικά, οι τιμές των ***ποιοτικών χαρακτηριστικών*** ορίζουν τις κατηγορίες στις οποίες εντάσσονται τα διάφορα μέλη του δείγματος ή του πληθυσμού.

Όταν στις κατηγορίες αυτές δεν υπάρχει διάταξη τότε οι εν λόγω ποιοτικές μεταβλητές ονομάζονται **ονομαστικές μεταβλητές**. Παραδείγματα αυτής της κατηγορίας είναι η οικογενειακή κατάσταση (με τιμές *άγαμος/η*, *έγγαμος/η*, *διαζευγμένος/η* κλπ), το θρήσκευμα (με τιμές χριστιανός/η ορθόδοξος/η, καθολικός/η, μουσουλμάνος/α, άθεος/η κλπ). Στη μέτρηση των ονομαστικών μεταβλητών χρησιμοποιούνται συχνά αριθμοί-ετικέτες κατηγοριών. Οι αριθμοί αυτοί δεν εμπεριέχουν καμία πληροφορία αναφορικά με την ιεραρχία μεταξύ των κατηγοριών αλλά, απλά, μας δίνουν τη δυνατότητα να ελέγξουμε αν ένα άτομο ανήκει ή δεν ανήκει σε μια κατηγορία. Ως εκ τούτου, θα μπορούσαν να αλλαχθούν οι αριθμοί αυτοί χωρίς να υπάρχει αλλαγή στην πληροφορία που απορρέει από τα δεδομένα. Παράδειγμα της περίπτωσης αυτής είναι ο ταχυδρομικός κώδικας. Θα μπορούσε να ανατεθεί στις περιοχές που έχουν ταχυδρομικό κώδικα το 18450 ως νέος κώδικας το 12345 και αντίστοιχα να αποδοθεί στις περιοχές που είχαν τον κώδικα 12345 ο κώδικας 18450. Άλλο παράδειγμα είναι η μεταβλητή Φύλο η οποία έπαιρνε τιμές παραδοσιακά 1 για τους άνδρες και 2 για τις γυναίκες ενώ θα μπορούσε να γίνει οποιαδήποτε άλλη απόδοση τιμών.

- Σε περίπτωση που οι μετρήσεις οδηγούν σε κατηγορίες με διάταξη τότε οι εν λόγω κατηγορικές/ποιοτικές μεταβλητές ονομάζονται **μεταβλητές διάταξης**. Παραδείγματα αυτής της κατηγορίας χαρακτηριστικών είναι ο βαθμός ικανοποίησης του εκπαιδευόμενου ή της εκπαιδευόμενης από την εκπαίδευση που έλαβε, η κατηγοριοποίηση του βάρους ενός ατόμου σε αδύνατο, κανονικό και υπέρβαρο ή η ικανοποίηση των πολιτών από τα οικονομικά μέτρα που παίρνει μια κυβέρνηση για να βοηθήσει τις μονογονεϊκές οικογένειες. Οι τιμές που χρησιμοποιούνται ως ετικέτες στα δεδομένα διάταξης ορίζονται αυθαίρετα αρκεί να διατηρείται η διάταξη μεταξύ των κατηγοριών. Για παράδειγμα, στις βαθμίδες

εκπαίδευσης, συνήθως αντιστοιχίζουμε στην πρωτοβάθμια εκπαίδευση το 1, στη δευτεροβάθμια το 2 και στην Τριτοβάθμια το 3, Εξ ίσου καλά θα μπορούσε κανείς να αποδώσει στις βαθμίδες εκπαίδευσης τις αριθμητικές ετικέτες 0 για την πρωτοβάθμια, 3 για την δευτεροβάθμια και 8 για την τριτοβάθμια. Το σημαντικό είναι οι ετικέτες που χρησιμοποιούνται να μην αλλάζουν την ιεραρχία μεταξύ των κατηγοριών. Ταυτόχρονα δεν έχει νόημα η απόσταση μεταξύ των ετικετών-αριθμών άρα δεν επιτρέπεται η αφαίρεση ως πράξη.

Από τα παραπάνω γίνεται αντιληπτό ότι η απόσταση μεταξύ των τιμών στην ονομαστική κλίμακα όπως και στην κλίμακα διάταξης δεν έχει νόημα. Για τις μεταβλητές που ανήκουν στην ονομαστική κλίμακα μέτρησης έχει νόημα ο υπολογισμός μεγεθών όπως:

- Οι συχνότητες και οι σχετικές συχνότητες εμφάνισης των τιμών τους
- Η επικρατούσα τιμή

ενώ στη διατεταγμένη κλίμακα, όπου τα δεδομένα μπορούν να διαταχθούν, εκτός από τους υπολογισμούς των *συχνοτήτων* και των *σχετικών συχνοτήτων* εμφάνισης των τιμών τους, έχουν νόημα και υπολογισμοί, οι οποίοι βασίζονται στη διάταξη, όπως είναι ο υπολογισμός της *διαμέσου*.

Η δεύτερη κατηγορία μεταβλητών, με βάση την κλίμακα μέτρησης, είναι οι ποσοτικές. Οι τιμές των **Ποσοτικών** μεταβλητών σε αντίθεση με τις ποιοτικές μεταβλητές έχουν την έννοια του αριθμού και επιτρέπονται συγκεκριμένες πράξεις. Για παράδειγμα , το βάρος και η ηλικία ενός ατόμου, ο αριθμός παιδιών σε μια οικογένεια. Οι ποσοτικές μεταβλητές διακρίνονται σε συνεχείς και διακριτές, Μία διακριτή μεταβλητή παίρνει διακεκριμένες τιμές, για παράδειγμα Αριθμός παιδιών στην οικογένεια., σε αντίθεση με μία συνεχή μεταβλητή η οποία μπορεί να πάρει οποιαδήποτε τιμή μέσα σε ένα διάστημα, για παράδειγμα το ύψος .
Οι ποσοτικές μεταβλητές, ανάλογα με τον τρόπο μέτρησης τους, διακρίνονται στις μεταβλητές:

- ***Κλίμακας διαστήματος.*** Στις μεταβλητές αυτής της κατηγορίας η μέτρηση βασίζεται σε μια κλίμακα με αυθαίρετη επιλογή του μηδενός. Παράδειγμα η θερμοκρασία. Το μηδέν στην θερμοκρασία δεν δηλώνει ανυπαρξία θερμοκρασίας.. Στην κλίμακα Κελσίου επιλέγεται το μηδέν αυθαίρετα με βάση την πήξη του νερού. Στις μεταβλητές κλίμακας διαστήματος επιτρέπεται η πρόσθεση και η αφαίρεση αλλά όχι ο πολλαπλασιασμός και η διαίρεση. Κατά συνέπεια, όταν λέμε ότι σήμερα έχουμε 2 βαθμούς Κελσίου ενώ την προηγούμενη ημέρα είχαμε έναν βαθμό μπορούμε να

ισχυριστούμε ότι η θερμοκρασία αυξήθηκε κατά ένα βαθμό αλλά δεν μπορούμε να πούμε ότι σήμερα έχει διπλάσια ζέστη από χθες.

- ***Κλίμακας λόγου***. Οι ποσοτικές μεταβλητές που ανήκουν στην κατηγορία της κλίμακας λόγου αφορούν μετρήσεις με σαφώς και μοναδικά προσδιορισμένη τιμή μηδενός. Σε αυτές, έχει νόημα, όχι μόνο το άθροισμα και η διαφορά αλλά και ο λόγος των τιμών τους. Για παράδειγμα αν Γιάννης έχει μισθό 1000€ ενώ η Αντωνία 2000€, τότε μπορούμε να ισχυριστούμε ότι η Αντωνία παίρνει:
    - 2000-1000=1000€ περισσότερα από τον Γιάννη,
    - 2000/1000=2πλάσιο μισθό από τον Γιάννη. Στο συγκεκριμένο παράδειγμα το μηδέν δηλώνει ανυπαρξία, συνεπώς ορίζεται μονοσήμαντα ως το απόλυτο μηδέν.

Ο διαχωρισμός μεταξύ διακριτών και συνεχών δεδομένων δυσχεραίνεται στην πράξη από τους περιορισμούς που επιβάλλονται από τα όργανα μέτρησης. Για παράδειγμα η μέτρηση του ύψους καταγράφεται με τη χρήση διακριτών τιμών, όπως 1,10m, 1,20m παρά το γεγονός ότι η μεταβλητή *ύψος* είναι συνεχής μεταβλητή, αφού μπορεί να πάρει οποιαδήποτε τιμή σε ένα διάστημα.

Συχνά είναι αναγκαίο να μετασχηματίσουμε ποσοτικά δεδομένα σε ποιοτικά και να τα αντιμετωπίσουμε στο εξής ως ποιοτικά. Για παράδειγμα ο δείκτης μάζας σώματος (ΔΜΣ), ο οποίος υπολογίζεται με τη διαίρεση του βάρους του ατόμου με το τετράγωνο του ύψους (σε μέτρα), είναι μία ποσοτική συνεχής μεταβλητή, Συνήθως, ακολουθούμε την παρακάτω σύμβαση για ομαδοποίηση των τιμών του ΔΜΣ:

- Κάτω από 18.5 Λιποβαρής
- 18.5 – 24.9 Κανονικός
- 25.0 – 29.9 Υπέρβαρος
- Πάνω από 30.0 Παχύσαρκος

Με τον τρόπο αυτό δημιουργείται μια νέα ποιοτική μεταβλητή με τιμές Λιποβαρής/ Κανονικός/Υπέρβαρος/Παχύσαρκος και με ταυτόχρονη απώλεια πληροφορίας, η οποία μπορεί να είναι επιθυμητή.

Ας επιστρέψουμε στο παράδειγμα του υπουργείου. Τα αρμόδια στελέχη για τη διεξαγωγή της έρευνας επιθυμούν να δημιουργήσουν συνοπτικά μέτρα από το δείγμα προκειμένου να κάνουν εκτιμήσεις για τον πληθυσμό. Ένα από αυτά είναι η μέση τιμή των ετών προϋπηρεσίας όλων των υπαλλήλων του υπουργείου. Η *μέση τιμή* των ετών προϋπηρεσίας όλων των υπαλλήλων του υπουργείου ονομάζεται **Στατιστική παράμετρος**. Ο μέσος όρος

των στοιχείων του δείγματος ονομάζεται **Στατιστική συνάρτηση.** Η έννοια της *στατιστικής συνάρτησης* και της *παραμέτρου* θα αναλυθούν στο 4ο κεφάλαιο.

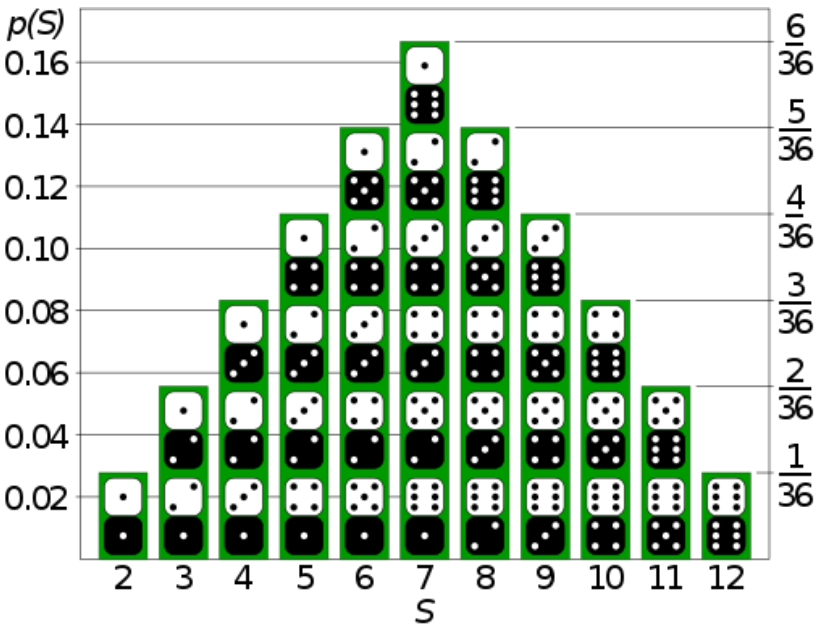
Τέλος μία πολύ βασική στατιστική έννοια είναι η έννοια της **Κατανομής**.

Η **Κατανομή** μιας μεταβλητής μας δείχνει όλες τις πιθανές τιμές (ή διαστήματα) των τιμών της μεταβλητής και πόσο συχνά αυτές εμφανίζονται.

Ας δούμε ένα παράδειγμα. Ρίχνουμε ένα μαύρο και ένα άσπρο ζάρι και σημειώνουμε το άθροισμα των ενδείξεων των δύο ζαριών. Όπως φαίνεται στο παρακάτω σχήμα αν μας ενδιαφέρει το ενδεχόμενο το άθροισμα των ενδείξεων των δύο ζαριών να είναι 3 υπάρχουν οι εξής δύο περιπτώσεις:

- Ένδειξη του άσπρου ζαριού είναι 1 και η ένδειξη του μαύρου 2 ή
- Ένδειξη άσπρου ζαριού 2 και ένδειξη του μαύρου 1.

Όλες οι δυνατές περιπτώσεις ενδείξεων των δύο ζαριών είναι 36, άρα υπάρχουν 2 στις 36 περιπτώσεις να έρθει άθροισμα ενδείξεων 3.



1.2Γράφημα κατανομής ρίψης ζαριών

Πηγή Wikipedia[2]

Μεταδεδομένα (metadata) Είναι δεδομένα τα οποία αφορούν αντικείμενα ή άλλα δεδομένα. Συνοδεύουν αδιάρρηκτα τα αντικείμενα ή τα δεδομένα στα οποία αναφέρονται, δίνοντας ζωτικές πληροφορίες για αυτά. Με την βοήθεια των μεταδεδομένων γίνεται εύκολη η

[2] https://el.wikipedia.org/wiki/%CE%9A%CE%B1%CF%84%CE%B1%CE%BD%CE%BF%CE%BC%CE%AE_%CF%80%CE%B9%CE%B8%CE%B1%CE%BD%CF%8C%CF%84%CE%B7%CF%84%CE%B1%CF%82

ανάκτηση, ο εντοπισμός ενός πληροφοριακού πόρου. Για παράδειγμα ο τίτλος, οι λέξεις κλειδιά, ο συγγραφέας, ο εκδότης, οι πίνακες περιεχομένων, πεδία με τα δικαιώματα χρήσης, ανήκουν στα μεταδεδομένα ενός βιβλίου.

Ως μεταδεδομένα σε μία στατιστική έρευνα μπορούν να θεωρούνται:

- Η υπηρεσία που πραγματοποιεί την έρευνα,
- Ο υπεύθυνος της έρευνας,
- Η περιγραφή του τρόπου μέτρησης των μεταβλητών,
- Ο πληθυσμός της έρευνας,
- Διάφορα θέματα εμπιστευτικότητας κ.λ.π.

Παράδειγμα μεταδεδομένων Στατιστικής έρευνας δίνεται στην ιστοσελίδα της ΕΛΣΤΑΤ[3]για τις *Στατιστικές Μεταναστευτικής Κίνησης*, όπου αναφέρεται το όνομα της υπηρεσίας που διενεργεί την έρευνα (στην περίπτωση αυτή η ΕΛΣΤΑΤ), το όνομα του υπευθύνου και άλλα στοιχεία χρήσιμα για την επικοινωνία με την υπηρεσία και τον υπεύθυνο της έρευνας, όπως είναι τηλέφωνα, διευθύνσεις email. Ακόμη αναφέρεται η ημερομηνία ανάρτησης και ενημέρωσης των μεταδεδομένων. Στην συνέχεια γίνεται μία σύντομη στατιστική παρουσίαση, που περιγράφονται οι πηγές, ο τύπος, η συχνότητα και η μέθοδος συλλογής των δεδομένων. Ακολούθως, υπάρχει περιγραφή των βασικών μεταβλητών που θα χρησιμοποιηθούν, όπως για παράδειγμα οι μεταβλητές που μορφοποιούν τις έννοιες της *εσωτερικής* και *εξωτερικής μετανάστευσης κ.λ.π. Στο ίδιο αρχείο* αναφέρονται πληροφορίες για τον πληθυσμό (που είναι το σύνολο των μεταναστών) και τις *στατιστικές μονάδες (* που, στην έρευνα αυτή, είναι τα άτομα) καθώς επίσης και τη χρονική κάλυψη της έρευνας (δηλαδή σε ποιες περιόδους αναφέρεται η έρευνα, η περίοδος βάσης για χρονολογικά δεδομένα, οι μονάδες μέτρησης που θα χρησιμοποιηθούν κλπ). Πέραν τούτων, αναφέρονται οι νομικές πράξεις και άλλες συμφωνίες, η πολιτική εμπιστευτικότητας, η πολιτική ανακοινώσεων (δηλαδή η συχνότητα ανακοίνωσης στοιχείων, ο τρόπος πρόσβασης των χρηστών σε αυτά, καθώς επίσης η συχνότητα και οι μορφές διάχυσης). Τέλος, στα μεταδεδομένα της έρευνας, αναφέρονται η ακρίβεια, η αξιοπιστία και τα σφάλματα της έρευνας. η συχνότητα και οι μέθοδοι συλλογής των δεδομένων,

Ελλείπουσες τιμές. Όπως ήταν φανερό και στο παράδειγμα που έχει αξιοποιηθεί για τη διασάφηση των εννοιών της Στατιστικής, σε όλες σχεδόν τις έρευνες αντιμετωπίζουμε το πρόβλημα της εμφάνισης ελλειπουσών τιμών, δηλαδή τιμών που για κάποιο λόγο λείπουν από τα δεδομένα.. Για παράδειγμα μπορεί κάποιος να συμφωνήσει να συμμετάσχει σε μία

3 http://www.statistics.gr/documents/20181/995824/%CE%9C%CE%B5%CF%84%CE%B1%CE%B4%CE%B5%CE%B4%CE%BF%CE%BC%CE%AD%CE%BD%CE%B1+%CE%A3%CF%84%CE%B1%CF%84%CE%B9%CF%83%CF%84%CE%B9%CE%BA%CF%89%CE%BD+%CE%9C%CE%B5%CF%84%CE%B1%CE%BD%CE%B1%CF%83%CF%84%CE%B5%CF%85%CF%84%CE%B9%CE%BA%CE%AE%CF%82+%CE%9A%CE%AF%CE%BD%CE%B7%CF%83%CE%B7%CF%82+%28+2008+%29.pdf/e8b00b6a-36f2-4cfc-b9f6-5336f0d92837?t=1445088749621

έρευνα, όμως να μην συμπληρώσει όλα τα πεδία του ερωτηματολογίου. Το πρόβλημα των ελλειπουσών τιμών είναι ένα πολύ σοβαρό πρόβλημα για τα αποτελέσματα της έρευνας και για τον λόγο αυτό έχουν προταθεί διάφορες μέθοδοι για την αντιμετώπισή του. Το εν λόγω θέμα δεν αποτελεί αντικείμενο μελέτης του παρόντος εγχειριδίου.

Δείγμα είναι ένα υποσύνολο του πληθυσμού.

Μέγεθος δείγματος είναι το πλήθος των στοιχείων του δείγματος

Παράμετρος είναι ένα χαρακτηριστικό της κατανομής του πληθυσμού, για παράδειγμα η μέση τιμή.

Στατιστική συνάρτηση είναι μία συνάρτηση των στοιχείων του δείγματος

Μεταβλητή είναι ένα χαρακτηριστικό- μία ιδιότητα- ενός ατόμου, αντικειμένου, κατάστασης που μπορεί να μεταβάλλεται. Οι μεταβλητές διακρίνονται σε *ποιοτικές* όταν οι τιμές τους είναι κατηγορίες και σε *ποσοτικές* όταν οι τιμές τους είναι αριθμοί.

Οι ποιοτικές μεταβλητές διακρίνονται σε *ονομαστικές*, π.χ. πρόθεση ψήφου, φύλο ,χρώμα μαλλιών και σε *διάταξης* π.χ. το μορφωτικό επίπεδο. Οι ποσοτικές μεταβλητές χωρίζονται σε *συνεχείς* όταν μπορούν να πάρουν οποιαδήποτε τιμή σε ένα διάστημα π.χ. βάρος και σε *διακριτές* όταν παίρνουν διακεκριμένες τιμές π.χ. ο αριθμός παιδιών σε μια οικογένεια.

Δεδομένα είναι οι μετρήσεις που προέρχονται από κάποιο πείραμα, κάποια έρευνα. Τα δεδομένα χωρίζονται σε ***ποιοτικά*** και ***ποσοτικά***.

Η **Κατανομή** μιας μεταβλητής μας δείχνει όλες τις πιθανές τιμές (ή διαστήματα) των τιμών της μεταβλητής και πόσο συχνά αυτές εμφανίζονται.

Μεταδεδομένα (metadata) Είναι δεδομένα τα οποία αφορούν αντικείμενα ή άλλα δεδομένα. Συνοδεύουν αδιάρρηκτα τα αντικείμενα ή τα δεδομένα στα οποία αναφέρονται, δίνοντας ζωτικές πληροφορίες για αυτά

1.3. Συλλογή δεδομένων

Ένας τρόπος συλλογής δεδομένων είναι η **Απογραφή**. Η μελέτη γίνεται επί ολόκληρου του πληθυσμού. Μια τέτοια μελέτη γίνεται από την Ελληνική Στατιστική Υπηρεσία (ΕΛΣΤΑΤ) κάθε 10 χρόνια για την απογραφή του πληθυσμού. Η τελευταία απογραφή στη χώρα μας έγινε το 2011.

Υπάρχουν όμως περιπτώσεις όπου η απογραφή είναι πολύ δύσκολο ή και αδύνατο να πραγματοποιηθεί. Για παράδειγμα τα πολιτικά κόμματα επιθυμούν να γνωρίζουν την άποψη των πολιτών για διάφορα θέματα της επικαιρότητας και, όπως είναι φυσικό, δεν είναι δυνατό να εξετασθεί η άποψη όλων των πολιτών μέσω απογραφής. Ακόμη υπάρχουν περιπτώσεις όπου για να γίνει μια μελέτη είναι αναγκαίο να καταστραφεί η προς μελέτη μονάδα, έτσι η απογραφή είναι αδύνατη.

Θα μπορούσαμε να αναφέρουμε ως παράδειγμα την μελέτη της αντοχής ράβδων όπου για να γίνει η σχετική μελέτη πρέπει οι μονάδες που θα χρησιμοποιηθούν να καταστραφούν. Έτσι, είναι φανερό ότι ο ερευνητής υποχρεώνεται να συγκεντρώσει πληροφορίες από ένα δείγμα, δηλαδή ένα μέρος του πληθυσμού κατάλληλα επιλεγμένο και στη συνέχεια με διάφορες στατιστικές τεχνικές να γενικεύσει τα συμπεράσματά του για το σύνολο/πληθυσμό.

Η διαδικασία επιλογής του δείγματος ονομάζεται **Δειγματοληψία**. Η Δειγματοληψία σε σχέση με την Απογραφή έχει μικρότερο κόστος και απαιτεί λιγότερο χρόνο για την εξαγωγή των συμπερασμάτων. Ένα επιπλέον πρόβλημα που εμφανίζει η Απογραφή είναι ότι η υλοποίηση της απαιτεί την εκπαίδευση μεγάλου αριθμού συνεντευκτών, ενώ στην Δειγματοληψία, το πλήθος των συνεντευκτών είναι σαφώς πολύ μικρότερο και πιο εύκολη η εκπαίδευση τους.

Αν η επιλογή του δείγματος είναι κατάλληλη ώστε το δείγμα μας να είναι αντιπροσωπευτικό, δηλαδή το δείγμα να είναι μια 'μικρογραφία του πληθυσμού' τότε τα συμπεράσματα θα ισχύουν με αρκετά μεγάλη ακρίβεια για τον πληθυσμό. Ένα αντιπροσωπευτικό δείγμα συχνά, μας δίνει πιο ακριβή αποτελέσματα ακόμη και από ένα μεγάλο δείγμα το οποίο όμως δεν έχει επιλεγεί κατάλληλα.

Η συλλογή των δεδομένων από το δείγμα μπορεί να γίνει με διάφορους τρόπους, όπως:

- Τηλεφωνική ή προσωπική συνέντευξη,
- Αλληλογραφία,
- Μέσω διαδικτύου,
- Παρατήρηση και καταγραφή των δεδομένων που προκύπτουν από ένα πείραμα στο εργαστήριο κλπ.

(Για περισσότερες πληροφορίες επί του θέματος μπορείτε να ανατρέξετε στο διαδικτυακό πόρο: https://repository.kallipos.gr/bitstream/11419/5075/1/00_master_document_with-cover.pdf)

1.4. Τεχνικές δειγματοληψίας[4]

Για την πραγματοποίηση της δειγματοληψίας πρέπει να οριστούν οι **δειγματοληπτικές μονάδες,** δηλαδή οι μονάδες που επιλέγονται για να επιτευχθεί η πρόσβαση στα άτομα του υπό μελέτη πληθυσμού. Για παράδειγμα για την πραγματοποίηση μιας έρευνας για την ικανοποίηση των κατοίκων ενός Δήμου για κάποιο θέμα, ως δειγματοληπτική μονάδα μπορεί να οριστεί το οικοδομικό τετράγωνο. Η επιλογή των δειγματοληπτικών μονάδων απαιτεί, κατά βάση, ένα οργανωμένο αρχείο, που περιέχει τα μέλη του πληθυσμού. Αυτό το σύνολο των δειγματοληπτικών μονάδων είναι το δειγματοληπτικό πλαίσιο ή αλλιώς μητρώο.

Οι τεχνικές δειγματοληψίας διακρίνονται σε δύο κατηγορίες. Αυτές είναι οι:

- **Δειγματοληψίες Πιθανότητας** και
- **Δειγματοληψίες μη Πιθανότητας**.

Στις **δειγματοληψίες πιθανότητας** κάθε μονάδα του πληθυσμού έχει μία μη μηδενική πιθανότητα επιλογής στο δείγμα. Η πιθανότητα αυτή έχει καθορισθεί πριν την επιλογή του δείγματος. Για παράδειγμα ένας φορέας του Δημοσίου πρόκειται να αντικαταστήσει μια διαδικασία εξυπηρέτησης του Πολίτη με μια άλλη και θέλει να εξετάσει τα αποτελέσματα της αντικατάστασης αυτής επί των υπαλλήλων. Έτσι, επιλέγονται από το σύνολο των εκατό υπαλλήλων του φορέα που εξυπηρετούν τη διαδικασία εντελώς τυχαία δέκα. Στην επιλογή αυτή χρησιμοποιείται το αρχείο της διεύθυνσης προσωπικού που αποτελεί εδώ το δειγματοληπτικό πλαίσιο.

1.4.1. Δειγματοληψία πιθανότητας

Κάποιες μορφές δειγματοληψίας πιθανότητας είναι οι παρακάτω:

Απλή τυχαία Δειγματοληψία: Είναι η μέθοδος κατά την οποία όλες οι μονάδες του πληθυσμού έχουν την ίδια πιθανότητα να περιληφθούν στο δείγμα. Για παράδειγμα για την μελέτη της χρήσης του διαδικτύου σε ένα σχολείο, αποφασίστηκε να επιλεγεί ένα δείγμα 50 μαθητών. Στην συνέχεια έγινε εισαγωγή των ονομάτων όλων των μαθητών του σχολείου σε

[4] https://repository.kallipos.gr/handle/11419/1296
(σελ.1-6)

κληρωτίδα ώστε όλοι οι μαθητές να έχουν την ίδια πιθανότητα επιλογής στο δείγμα. Πλεονεκτήματα της απλής τυχαίας Δειγματοληψίας είναι η αμεροληψία και η αξιοπιστία που εξασφαλίζει, Βέβαια υπάρχει περίπτωση υπό ή υπέρ εκπροσώπησης κάποιων ομάδων. Στο προηγούμενο παράδειγμα υπάρχει περίπτωση να επιλεγούν μόνο μαθητές της Γ΄Λυκείου.

***Στρωματοποιημένη Δειγματοληψία*:** Είναι η μέθοδος κατά την οποία ο πληθυσμός χωρίζεται σε ομάδες (στρώματα) και στη συνέχεια επιλέγεται ένα απλό τυχαίο δείγμα από κάθε στρώμα. Ο χωρισμός του πληθυσμού σε στρώματα γίνεται έτσι ώστε να υπάρχει η μέγιστη δυνατή ομοιογένεια ως προς το χαρακτηριστικό για το οποίο γίνεται η μελέτη. Αυτό έχει ως αποτέλεσμα ένα σχετικά μικρό δείγμα από κάθε στρώμα να είναι αντιπροσωπευτικό του κάθε στρώματος.

Η στρωματοποιημένη δειγματοληψία πολλές φορές είναι αναγκαία για λόγους *οικονομίας* αφού όσο μεγαλύτερη είναι η ομοιογένεια στο κάθε στρώμα, τόσο πιο μικρό δείγμα είναι κατάλληλο για την εξαγωγή συμπερασμάτων, καθώς επίσης για λόγους *εκπροσώπησης* όλων των ομάδων στο δείγμα οπότε να μην υπάρχει περίπτωση υπό ή υπέρ εκπροσώπησης στο δείγμα κάποιας ομάδας, Επιπλέον με την στρωματοποιημένη δειγματοληψία υπάρχει δυνατότητα αναφοράς σε κάθε στρώμα χωριστά. Για παράδειγμα, ας θεωρήσουμε κάθε νομό της χώρας ως στρώμα. Στη συνέχεια, επιλέγουμε με απλή τυχαία δειγματοληψία ένα δείγμα από κάθε νομό με αποτέλεσμα στο τελικό δείγμα μας να έχουμε στατιστικές μονάδες από όλους τους νομούς σε κατάλληλα ποσοστά. Επιπρόσθετα, θα είναι δυνατή και η αναφορά στα αποτελέσματα ανά νομό. Στο παράδειγμα που αναφέρθηκε προηγουμένως με την χρήση του διαδικτύου από τους μαθητές ενός σχολείου, αν ως στρώμα επιλέγαμε την τάξη και στην συνέχεια με απλή τυχαία δειγματοληψία επιλέγαμε ένα δείγμα από κάθε τάξη δεν θα υπήρχε το πρόβλημα της υπό ή υπερεκπροσώπησης κάποιας τάξης.

Συστηματική Δειγματοληψία: Η Συστηματική Δειγματοληψία χρησιμοποιείται συνήθως σε περιπτώσεις που η επιλογή του δείγματος γίνεται από μια λίστα. Για παράδειγμα αν θέλουμε να επιλέξουμε ένα δείγμα μεγέθους 100 καρτελών από 1000 καρτέλες, επιλέγουμε έναν τυχαίο αριθμό από 1 έως 10 (10=1000/100) π.χ. το 4 και στην συνέχεια επιλέγουμε την $4^{η}$ κάρτα, την $14^{η}$ κάρτα κ.ο.κ. Το πλεονέκτημα της είναι η ευκολία στην επιλογή του δείγματος αφού χρειάζεται η τυχαία επιλογή ενός μόνο αριθμού. Τα αποτελέσματά της είναι συχνά πολύ ακριβή. Πρόβλημα υπάρχει με αυτήν την μέθοδο δειγματοληψίας, όταν ο πληθυσμός περικλείει μια περιοδική μορφή διακύμανσης. Τότε υπάρχει περίπτωση να πάρουμε ένα δείγμα πολύ μεροληπτικό.

Παράδειγμα: Έστω ότι θέλουμε να εκτιμήσουμε τον αριθμό των συναλλαγών στο ταμείο μιας τράπεζας μέσα σε μία ημέρα, χρησιμοποιώντας την μέθοδο της συστηματικής δειγματοληψίας. Αν επιλεγεί η πρώτη εργάσιμη ημέρα κάθε μήνα τότε το δείγμα θα είναι μεροληπτικό αφού είναι γνωστό ότι η κίνηση στα ταμεία των τραπεζών είναι αυξημένη την πρώτη εργάσιμη ημέρα κάθε μήνα.

1.4.2. Δειγματοληψία μη πιθανότητας

Στην δειγματοληψία μη πιθανότητας το δείγμα επιλέγεται από τον πληθυσμό, χωρίς να είναι εκ των προτέρων γνωστή η πιθανότητα ενός στοιχείου να περιληφθεί στο δείγμα. Οι μέθοδοι δειγματοληψίας μη πιθανότητας είναι πιθανό να οδηγήσουν σε αντιπροσωπευτικά δείγματα αλλά στερούνται της τυχαιότητας με αποτέλεσμα να μην είναι δυνατή η ανάπτυξη κάποιας δειγματοληπτικής θεωρίας ούτε και η εκτίμηση της ακρίβειας της έρευνας. Παρακάτω παρουσιάζονται κάποιες μορφές δειγματοληψίας μη πιθανότητας.

Δειγματοληψία ευκολίας: Το δείγμα επιλέγεται από ένα τμήμα του πληθυσμού στο οποίο υπάρχει εύκολη πρόσβαση. Ένα τέτοιο δείγμα δεν μπορεί να είναι αντιπροσωπευτικό του πληθυσμού. Ας υποθέσουμε ότι ένας συνεντεύκτης θέλει να αποτυπώσει την αντίληψη που έχουν οι καταναλωτές στην πρόσφατη αλλαγή των τιμών ΦΠΑ που αποφάσισε η ηγεσία του Υπουργείου Οικονομικών. Για το λόγο αυτό αποφασίζει να επιλέξει συνεντευξιαζόμενους από την οδό Ερμού της πλατείας Συντάγματος. Είναι προφανές ότι επιλέγει έναν πολυσύχναστο δρόμο για την ευκολία του αλλά το δείγμα δεν είναι απαραίτητα αντιπροσωπευτικό του πληθυσμού.
Δειγματοληψία κρίσης: Σε αυτή τη μορφή Δειγματοληψίας ο ερευνητής επιλέγει το δείγμα υποκειμενικά βασιζόμενος στην προσωπική του κρίση, εμπειρία και γνώση του υπό μελέτη πληθυσμού. Αν η κρίση του ερευνητή είναι καλή τα αποτελέσματα αυτής της μορφής της δειγματοληψίας μπορεί να είναι πολύ καλά. Για παράδειγμα έχει διαπιστωθεί για σειρά ετών ότι τα αποτελέσματα των εκλογών από κάποια περιοχή δεν διαφέρουν πολύ από τα τελικά αποτελέσματα. Έτσι ο ερευνητής επιλέγει την περιοχή αυτήν για την εκτίμηση των εκλογικών αποτελεσμάτων.

Δειγματοληψία χιονοστιβάδας: Σε αυτή την μορφή δειγματοληψίας, ο ερευνητής προσδιορίζει ένα ή περισσότερα άτομα από τον πληθυσμό τους οποίους περιλαμβάνει στο δείγμα. Στην συνέχεια μέσω αυτών προσπαθεί να προσδιορίσει άλλα μέλη του πληθυσμού κ.ο.κ. Με αυτόν τον τρόπο οι συμμετέχοντες στο δείγμα γίνονται και βοηθοί του ερευνητή.

Συχνά εφαρμόζεται σε δύσκολα προσβάσιμους πληθυσμούς, λόγω του στιγματισμού και της περιθωριοποίησης (Π.χ. άτομα που κάνουν χρήση ναρκωτικών ουσιών, μετέχουν σε trafficking κλπ)

Δειγματοληψία με προκαθορισμένα ποσοστά: Σύμφωνα με αυτήν την μέθοδο δειγματοληψίας, ο ερευνητής επιλέγει στοιχεία από κάθε κατηγορία του πληθυσμού, και μάλιστα σε αναλογία ίδια με αυτήν των κατηγοριών στον πληθυσμό, όμως η επιλογή των στοιχείων σε κάθε στρώμα δεν γίνεται τυχαία. Για παράδειγμα για την μελέτη της χρήσης του διαδικτύου σε μία πόλη, αποφασίστηκε να επιλεγεί ένα δείγμα 100 ατόμων. Γνωρίζουμε ότι η το 40% των κατοίκων της πόλης είναι ανήλικοι και το 60% ενήλικοι και θέλουμε στο δείγμα μας το 40% να είναι ανήλικοι και το 60% ενήλικοι. Για την συγκέντρωση των στοιχείων του δείγματος επιλέγουμε σε μία στάση λεωφορείου τους πρώτους 40 ανήλικους και τους πρώτους 60 ενήλικους. Αυτή η μέθοδος δειγματοληψίας μοιάζει με την στρωματοποιημένη δειγματοληψία, στερείται όμως του στοιχείου της τυχαιότητας.

Άσκηση 1ου κεφαλαίου

Επισκεφθείτε τον υπερσύνδεσμο ΕΛΣΤΑΤ

Στην εν λόγω τοποθεσία υπάρχει ένα αρχείο που αφορά τα μεταδεδομένα για την έρευνα με τίτλο:

Έρευνα Οικογενειακών Προϋπολογισμών, 2019

1. Προσδιορίστε τον/την υπεύθυνο/η για τη δημιουργία του σχετικού ερωτηματολογίου.
2. Ποιος είναι ο βασικός σκοπός και στόχοι της έρευνας;
3. Η έρευνα είναι σε επίπεδο πληθυσμού (απογραφή) ή αφορά δειγματοληψία; Αν πρόκειται για δειγματοληψία προσδιορίστε με ποιο τρόπο οι υπεύθυνοι της έρευνας ισχυρίζονται ότι εξασφαλίζουν την αντιπροσωπευτικότητα του δείγματος;
4. Προσδιορίστε τις στατιστικές ή δειγματοληπτικές μονάδες της έρευνας καθώς και την περιοχή αναφοράς. Είναι απαραίτητο να προσδιορίζεται η περιοχή αναφοράς και γιατί;
5. Ποια είναι η πολιτική εμπιστευτικότητας που έχει υιοθετηθεί στην έρευνα αυτή; Είναι απαραίτητο να συνοδεύεται μια έρευνα από πολιτική εμπιστευτικότητας; Δικαιολογείστε.
6. Αναφέρετε κάποιες από τις μεταβλητές που αποτυπώνονται από την έρευνα και προσδιορίστε την κλίμακα μέτρησής τους.

.ΚΕΦΑΛΑΙΟ 2ο

ΑΡΙΘΜΗΤΙΚΗ ΠΕΡΙΓΡΑΦΗ ΔΕΔΟΜΕΝΩΝ

Στο κεφάλαιο αυτό θα μελετήσουμε στατιστικά μεγέθη που χρησιμοποιούμε για την αριθμητική περιγραφή δεδομένων. Τα μεγέθη αυτά θα παρουσιαστούν μέσα από παραδείγματα της καθημερινής μας ζωής ώστε να γίνουν κατανοητά. Τέλος για κάθε έννοια θα παρουσιαστεί ο υπολογισμός των διαφόρων στατιστικών μεγεθών μέσω του excel και θα παρουσιαστούν οι βασικές εντολές χρήσης του excel.

2.1 Πίνακες Συχνοτήτων – Σχετικών Συχνοτήτων

Συχνά σε μία έρευνα μας ενδιαφέρει να καταγράψουμε πόσο συχνά εμφανίζεται η τιμή μιας μεταβλητής στον πληθυσμό ή το δείγμα. Ενδιαφερόμαστε δηλαδή για την **απόλυτη συχνότητα** ή απλά **συχνότητα** της τιμής της μεταβλητής.

Αν καταγράψουμε τις *συχνότητες* όλων των τιμών μιας μεταβλητής σε έναν πίνακα, τότε έχουμε δημιουργήσει έναν πίνακα συχνοτήτων. Οι πίνακες συχνοτήτων είναι συγκεντρωτικές παρουσιάσεις των δεδομένων μας.

Ο πίνακας συχνοτήτων είναι ένας πίνακας συνήθως δύο στηλών. Στην πρώτη στήλη του πίνακα τοποθετούνται οι διαφορετικές τιμές της μεταβλητής και στην δεύτερη στήλη οι αντίστοιχες συχνότητές τους.

Από την περιγραφή του πίνακα συχνοτήτων γίνεται κατανοητό ότι δεν μπορεί να δημιουργηθεί πίνακας συχνοτήτων για μία συνεχή μεταβλητή.

Πολλές φορές εκτός από το πόσο συχνά εμφανίζεται η τιμή μιας μεταβλητής στον πληθυσμό ή το δείγμα, μας ενδιαφέρει , και το ποσοστό εμφάνισης της τιμής αυτής, δηλαδή μας ενδιαφέρει η **σχετική συχνότητα** των τιμών της μεταβλητής.

Η δομή του πίνακα σχετικών συχνοτήτων είναι παρόμοια αυτής του πίνακα συχνοτήτων με τη μόνη διαφορά ότι αντί των απόλυτων συχνοτήτων εμφανίζονται οι σχετικές συχνότητες κάθε τιμής της μεταβλητής. Με άλλα λόγια, ο πίνακας σχετικών συχνοτήτων είναι και αυτός πίνακας συνήθως δύο στηλών εκ των οποίων η πρώτη στήλη του πίνακα φιλοξενεί τις διαφορετικές τιμές της μεταβλητής ενώ η δεύτερη στήλη τις αντίστοιχες σχετικές συχνότητες των τιμών της μεταβλητής.

Συχνά οι συχνότητες και οι σχετικές συχνότητες των τιμών μιας μεταβλητής αποτυπώνονται σε έναν ενιαίο πίνακα, ο οποίος συνήθως στην πρώτη στήλη περιέχει τις τιμές της μεταβλητής και στις υπόλοιπες στήλες την συχνότητα και την σχετική συχνότητα των τιμών

της μεταβλητής. Η σχετική συχνότητα μιας τιμής, όπως θα δούμε στην συνέχεια, προκύπτει από μία διαίρεση, κάτι το οποίο τις περισσότερες φορές έχει ως αποτέλεσμα την εμφάνιση πολλών δεκαδικών ψηφίων .Για την καλύτερη εμφάνιση του πίνακα σχετικών συχνοτήτων καλό είναι να αποφεύγεται η καταγραφή μεγάλου αριθμού δεκαδικών ψηφίων και να εμφανίζεται η τιμή της σχετικής συχνότητας κατάλληλα στρογγυλοποιημένη.

2.1.1 Συχνότητες

Όπως αναφέραμε προηγουμένως, *Συχνότητα* μιας τιμής της μεταβλητής είναι ένας αριθμός που μας δείχνει πόσες φορές έχει εμφανισθεί η τιμή αυτή στα δεδομένα μας.
Ας δούμε ένα παράδειγμα πίνακα συχνοτήτων και πως αντιμετωπίζεται μέσω του EXCEL.

Παράδειγμα 1ο

Μία διεύθυνση σε ένα δημόσιο οργανισμό απαρτίζεται από 20 εργαζόμενους/ες. Ο υπεύθυνος προσωπικού θέλει να καταγράψει τις ημέρες άδειας πού έλαβαν οι υπάλληλοι μέσα στο πρώτο τρίμηνο του έτους. Τα δεδομένα που προέκυψαν για τους 20 υπαλλήλους ήταν:

Ημέρες άδειας: 0,0,0,0,0,0,0,0,0,0,2,2,2,2,2,2,2,5,5,5

Παρατηρούμε ότι

- 10 εργαζόμενοι δεν πήραν άδεια μέσα στο πρώτο τρίμηνο του έτους,
- 7 εργαζόμενοι πήραν δύο (2) ημέρες άδειας ο καθένας.
- 3 εργαζόμενοι πήραν πέντε (5) ημέρες άδεια ο καθένας.

Με βάση τα παραπάνω, η τιμή **μηδέν ημέρες** (0) της μεταβλητής «Ημέρες άδειας» έχει συχνότητα εμφάνισης 10, η τιμή **δύο ημέρες** (2) έχει συχνότητα εμφάνισης 7 και η **τιμή πέντε ημέρες** (5) έχει συχνότητα εμφάνισης 3. Τα παραπάνω θα μπορούσαν να παρασταθούν σε έναν πίνακα συχνοτήτων, όπως φαίνεται παρακάτω.

Ημέρες άδειας	Συχνότητες
0	10
2	7
5	3
Σύνολο	**20**

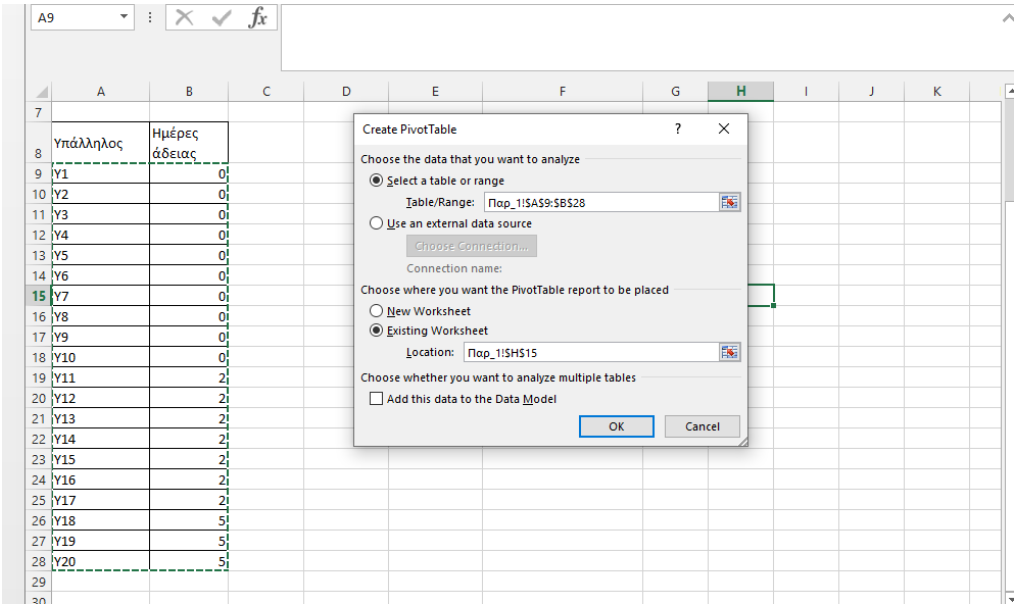
2.1Πίνακας συχνοτήτων Ημέρες άδειας

Υπολογισμός με το EXCEL

Στην πρώτη στήλη του παρακάτω πίνακα παρακάτω πίνακα εμφανίζονται οι τιμές που λαμβάνει η μεταβλητή *Ημέρες άδειας* στο σύνολο των είκοσι υπαλλήλων ενώ στη δεύτερη στήλη καταγράφονται οι απόλυτες συχνότητες εμφάνισης των τιμών αυτών. Για να δημιουργήσουμε από τα πρωτογενή δεδομένα έναν πίνακα συχνοτήτων ακολουθούμε τα παρακάτω βήματα.

Καταχωρούμε τα πρωτογενή δεδομένα που αφορούν τις ημέρες άδειας των υπαλλήλων.

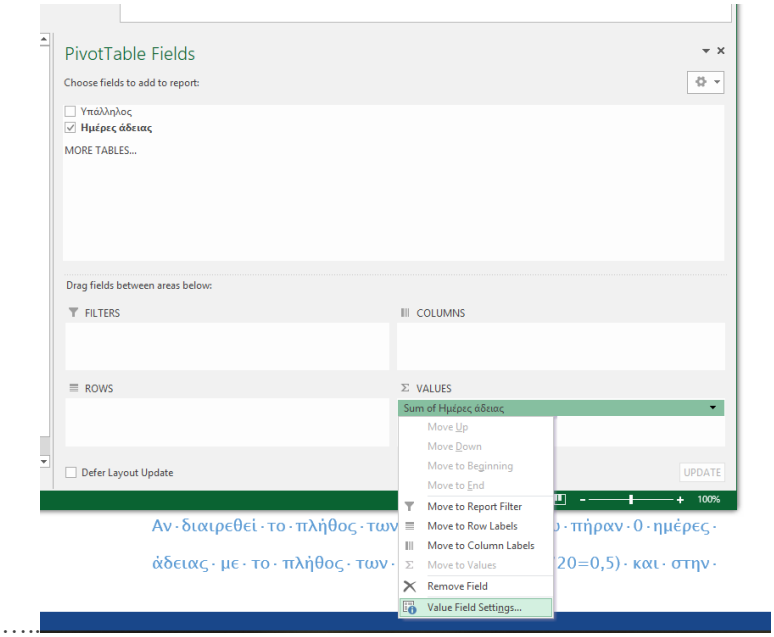
Από την καρτέλα Εισαγωγή επιλέγουμε Συγκεντρωτικοί πίνακες οπότε εμφανίζεται το παράθυρο όπου καταχωρούμε τον πίνακα και την θέση όπου θέλουμε να εμφανισθεί ο συγκεντρωτικός πίνακας όπως φαίνεται παρακάτω.



Υπάλληλος	Ημέρες άδειας
Υ1	0
Υ2	0
Υ3	0
Υ4	0
Υ5	0
Υ6	0
Υ7	0
Υ8	0
Υ9	0
Υ10	0
Υ11	2
Υ12	2
Υ13	2
Υ14	2
Υ15	2
Υ16	2
Υ17	2
Υ18	5
Υ19	5
Υ20	5

2.2Εικόνα Συγκεντρωτικοί πίνακες

Πατάμε ***ΟΚ*** οπότε εμφανίζεται η παρακάτω οθόνη του βοηθού συμπλήρωσης πεδίων.



2.3Εικόνα συγκεντρωτικοί πίνακες

Στην συνέχεια σύρουμε το ημέρες άδειας στο πεδίο των γραμμών και σύρουμε το ημέρες άδειας στο πεδίο τιμές και επιλέγουμε τη συνάρτηση count στις Ρυθμίσεις πεδίου τιμών, την συνάρτηση καταμέτρησης (count).

Είναι φανερό ότι, εκτός από τη συνάρτηση καταμέτρησης (count), μπορούμε στους συγκεντρωτικούς πίνακες να αξιοποιήσουμε και άλλες συναρτήσεις υπολογισμού όπως είναι αυτές του:

- Αθροίσματος (sum)
- Μέσου όρου (average) κλπ

Το αποτέλεσμα των παραπάνω χειρισμών είναι η εμφάνιση ενός πίνακα που εχει την παρακάτω μορφή:

Ημέρες Άδειας	Συχνότητα
0	10
2	7
5	3
Σύνολο	**20**

2.4Πίνακας Συχνότητες Ημέρες άδειας

2.1.2 Σχετικές συχνότητες

Όπως αναφέραμε στην αρχή της ενότητας, πολλές φορές μας ενδιαφέρει να αποτυπωθεί, εκτός της απόλυτης συχνότητας εμφάνισης των τιμών μιας μεταβλητής, και η σχετική εμφάνιση αυτών. Η σχετική συχνότητα μας ενδιαφέρει και αναφορικά με τον πληθυσμό αλλά και σε σχέση με το υπό εξέταση δείγμα. Αυτή εκφράζει το ποσοστό εμφάνισης της κάθε τιμής της μεταβλητής επί του συνόλου των τιμών.

Για τον υπολογισμό της σχετικής συχνότητας της τιμής μιας μεταβλητής, διαιρούμε την συχνότητα της τιμής αυτής με το άθροισμα των συχνοτήτων όλων των τιμών της μεταβλητής.

Για την κατανόηση του υπολογισμού της σχετικής συχνότητας της τιμής μιας μεταβλητής, θα χρησιμοποιήσουμε το παράδειγμα της παραγράφου 2.1.1.

Αν διαιρεθεί το πλήθος των εργαζομένων που πήραν 0 ημέρες άδειας με το πλήθος των εργαζομένων (10/20=0,5) προκύπτει η *σχετική συχνότητα* της τιμής 0, της μεταβλητής *Ημέρες άδειας*. Όμοια υπολογίζονται οι σχετικές συχνότητες των υπολοίπων τιμών της μεταβλητής και η καταχώρηση των τιμών αυτών ΄στην τρίτη στήλη του παρακάτω πίνακα δημιουργεί έναν επαυξημένο πίνακα στον οποίο, εκτός από τις απόλυτες συχνότητες καταγράφονται και οι σχετικές συχνότητες σε μορφή δεκαδικών αριθμών.

Αν στην συνέχεια πολλαπλασιαστούν τα στοιχεία της τρίτης στήλης του πίνακα επί 100 και τα αποτελέσματα τεθούν στην τέταρτη στήλη του πίνακα που ακολουθεί τότε λαμβάνονται οι σχετικές συχνότητες σε γραφή επί τοις εκατό (%). , π.χ. αν η σχετική συχνότητα της τιμής 0 ημέρες άδειας που ισούται με 0,5 πολλαπλασιαστεί με το 100 (0,5*100=50%) θα προκύψει το ποσοστό εμφάνισης της τιμής «0 ημέρες άδειας» ως 50%, δηλαδή θα εμφανίζεται η σχετική συχνότητα της τιμής σε μορφή εκατοστιαίου ποσοστού. Η % σχετική συχνότητα των τιμών της μεταβλητής, ερμηνεύεται ως εξής:

Αν οι εργαζόμενοι ήταν 100 τότε θα αναμενόταν οι 50 από αυτούς να πάρουν 0 ημέρες άδεια, οι 35 από αυτούς 2 ημέρες άδειας και 15 από αυτούς 5 ημέρες άδειας.

Ο παρακάτω πίνακας είναι ο τελικός πίνακας συχνοτήτων και σχετικών συχνοτήτων για την μεταβλητή *Ημέρες άδειας*.

Ημέρες άδειας	Συχνότητα	Σχετική συχνότητα	% Σχετική συχνότητα
0	10	0,50	50
2	7	0,35	35
5	3	0,15	15
Σύνολο	**20**	1	100

2.5*Πίνακας Σχετικές συχνότητες*

Υπολογισμός με το EXCEL

1ος τρόπος

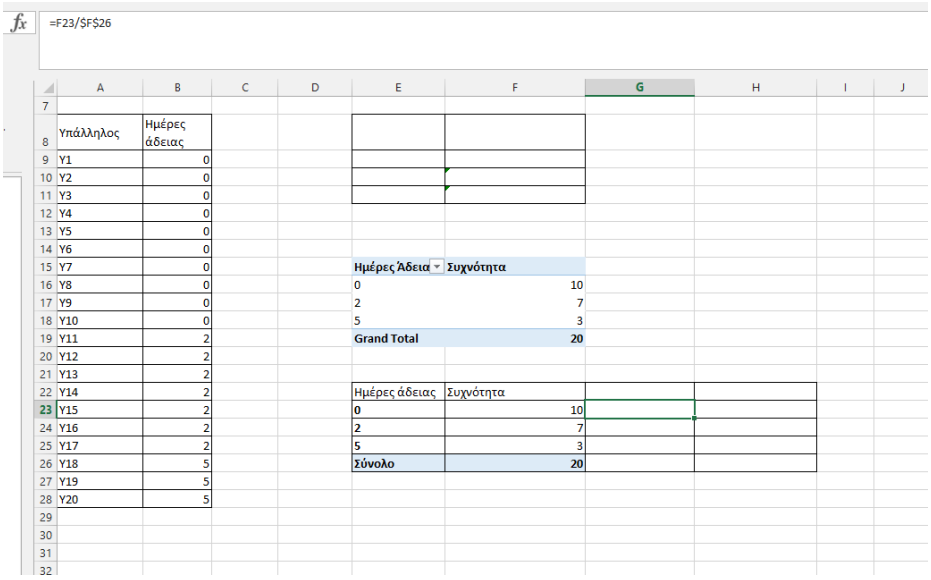
Εισάγουμε τα πρωτογενή δεδομένα

Επιλέγουμε την καρτέλα Εισαγωγή και μετά Συγκεντρωτικοί πίνακες οπότε εμφανίζεται το παράθυρο όπου καταχωρούμε τον πίνακα και την θέση όπου θέλουμε να εμφανισθεί ο συγκεντρωτικός πίνακας όπως και στην προηγούμενη παράγραφο και στο πεδίο *Τιμές* στις *Ρυθμίσεις πεδίου* στο πεδίο *Εμφάνιση τιμών* επιλέγουμε την εμφάνιση ποσοστού.

2ος τρόπος

Υπολογίζουμε τις συχνότητες, όπως στο παράδειγμα στην παράγραφο 2.1.1. στην συνέχεια αντιγράφουμε τις τιμές του πίνακα συχνοτήτων

Στο κελί που θέλουμε να γίνει ο υπολογισμός της σχετικής συχνότητας (για το παράδειγμά μας το G23) πληκτρολογούμε ***=F23/$F$26*** δηλαδή το κελί της συχνότητας της τιμής 0 ημέρες άδειας με το κελί του συνόλου. Ανάλογα, συμπληρώνονται και τα υπόλοιπα κελιά υπολογισμού των σχετικών συχνοτήτων εμφάνισης των τιμών της μεταβλητής.



2.6*Εικόνα Σχετικές συχνότητες*

Αν θέλουμε οι τιμές να εμφανίζονται στα κελιά με τη μορφή εκατοστιαίου ποσοστού και όχι δεκαδικού αριθμού τότε επιλέγουμε από τη μορφοποίηση κελιού την μορφή ποσοστού καθώς και το πλήθος των δεκαδικών ψηφίων που θα επιθυμούμε να εμφανίζονται, π.χ. δύο δεκαδικά ψηφία.

Ο πίνακας που θα προκύψει στην περίπτωση αυτή θα είναι ο ακόλουθος.

Ημέρες άδειας	Συχνότητα	Σχετική Συχνότητα	% Σχετική συχνότητα
0	10	0,5	50,00
2	7	0,35	35,00
5	3	0,15	15,00
Σύνολο	**20**	**1**	**100,00**

2.7Πίνακας Σχετικές συχνότητες

2.1.3 Αθροιστική συχνότητα –Σχετική αθροιστική συχνότητα

Πολλές φορές χρειάζεται να υπολογιστεί το πλήθος των παρατηρήσεων, των οποίων οι τιμές είναι μικρότερες ή ίσες, ενός συγκεκριμένου επίπεδου τιμής μιας μεταβλητής. Το πλήθος των παρατηρήσεων των οποίων οι τιμές είναι μικρότερες ή ίσες με μία τιμή της μεταβλητής, ονομάζεται *Αθροιστική συχνότητα*. Το ποσοστό των παρατηρήσεων που είναι μικρότερες ή ίσες με μία συγκεκριμένη τιμή της μεταβλητής ονομάζεται *Σχετική αθροιστική συχνότητα*.

Ο παραπάνω υπολογισμός δεν μπορεί να πραγματοποιηθεί όταν η μεταβλητή είναι ονομαστική, γιατί στις ονομαστικές μεταβλητές δεν υπάρχει διάταξη. Στις δύο τελευταίες στήλες του παρακάτω πίνακα παρουσιάζονται η αθροιστική συχνότητα και η σχετική αθροιστική συχνότητα για την μεταβλητή *Ημέρες άδειας* του πρώτου παραδείγματος λαμβάνοντας ως τιμές επιπέδων τις δύο και πέντε ημέρες.

Ημέρες άδειας	Συχνότητα	Σχετική συχνότητα	Αθροιστική συχνότητα	Σχετική αθροιστική συχνότητα
0	10	50,00%	10	50,00%
2	7	35,00%	17	85,00%
5	3	15,00%	20	100,00%
Σύνολο	20	100,00%		

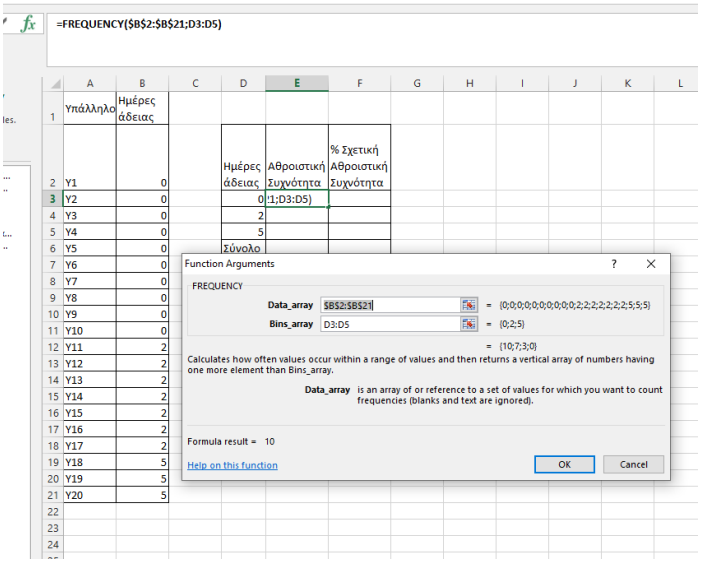
2.8Πίνακας Αθροιστική Συχνότητα-Σχετική συχνότητα

Από τον παραπάνω πίνακα γίνεται αντιληπτό ότι το 85% (=50%+35%), των εργαζομένων έχουν πάρει το πολύ 2 ημέρες άδεια και το 100% έχουν πάρει το πολύ 5 ημέρες άδεια. Ένα ερώτημα που θα μπορούσε να απαντηθεί με την βοήθεια της σχετικής συχνότητας είναι το εξής*: «Αν επιλεγεί τυχαία ένας εργαζόμενος, πόσο πιθανό θα ήταν να είχε πάρει μέσα στο τρίμηνο 2 ημέρες άδεια;»* Η απάντηση θα ήταν 35%, αφού το 35% των εργαζομένων έχουν πάρει άδεια.

Υπολογισμός με το EXCEL

Για την δημιουργία του πίνακα αθροιστικών συχνοτήτων εκτελούμε τα παρακάτω βήματα.

- Δημιουργούμε μία στήλη με τις διαφορετικές τιμές της μεταβλητής «Ημέρες Άδειας»(D3:D5), όπου καταχωρούμε τις τιμές 0, 2 και 5.
- Στο κελί της αθροιστικής συχνότητας επιλέγουμε την συνάρτηση ***Frequency (=FREQUENCY($B$2:$B$21;D3:D5).*** Τα κελιά B2 έως B21 είναι τα κελιά που περιέχονται τα δεδομένα και τα κελιά D3:D5 είναι τα κελιά που περιέχουν τις διαφορετικές τιμές της μεταβλητής.



2.9*Εικόνα Υπολογισμός σχετικής συχνότητες*

Έτσι προκύπτει ο παρακάτω πίνακας

Ημέρες άδειας	Αθροιστική Συχνότητα
0	10
2	17
5	20

2.10*Πίνακας Αθροιστική συχνότητα*

Ο υπολογισμός της % αθροιστικής συχνότητας γίνεται με παρόμοιο τρόπο, με τον υπολογισμό της σχετικής συχνότητας στο παράδειγμα της παραγράφου 2.1.2. Δηλαδή διαιρούμε την αθροιστική συχνότητα μιας τιμής με την αθροιστική συχνότητα της μεγαλύτερης παρατήρησης.

Για το παραπάνω παράδειγμα η σχετική αθροιστική συχνότητα της τιμής 0 ημέρες άδειας ισούται με 10/20=0,5 και στην μορφή ποσοστού 50% (0,5*100).

Πίνακες συνάφειας (Διασταύρωσης)

Συνήθως μελετούμε τα στοιχεία του πληθυσμού ή του δείγματος ως προς περισσότερα του ενός χαρακτηριστικά. Για ποιοτικά κυρίως χαρακτηριστικά μας ενδιαφέρει να απεικονίσουμε μέσω ενός πίνακα, την σχέση μεταξύ αυτών των χαρακτηριστικών.

Για παράδειγμα, θεωρείστε ότι θέλουμε να μελετήσουμε την σχέση καπνίσματος και φύλου. Για τον λόγο αυτό ερωτήθηκαν αν καπνίζουν ή όχι 13 άτομα ενώ καταγράφηκε και το φύλο τους. Κάθε γραμμή του παρακάτω πίνακα αντιστοιχεί σε ένα άτομο. Στην πρώτη στήλη είναι ο αύξων αριθμός, στην δεύτερη στήλη σημειώνεται το φύλο και στην τρίτη στήλη αν είναι ή όχι καπνιστής.

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

2.11*Πίνακες συνάφειας*

Ο παραπάνω πίνακας αποτελείται από τα ακατέργαστα δεδομένα που προέκυψαν από την έρευνά μας. Ο συγκεντρωτικός πίνακας που θέλουμε να δημιουργήσουμε πρέπει να έχει δύο μεταβλητές. Η μία μεταβλητή είναι η μεταβλητή «Φύλο» και η άλλη είναι η μεταβλητή «Καπνιστής». Και οι δύο μεταβλητές του παραδείγματος είναι ονομαστικές. Ελπίζουμε ότι ο

συγκεντρωτικός πίνακας θα απεικονίζει, με κάποιο τρόπο, την σχέση/εξάρτηση που έχουν οι δυο μεταβλητές δηλαδή το φύλο με το κάπνισμα.

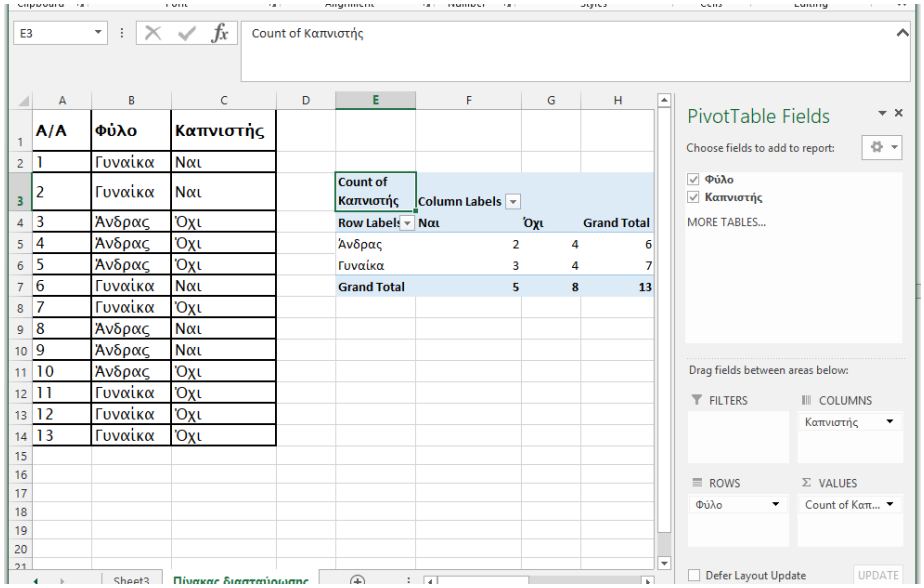
Ένας τέτοιος πίνακας ονομάζεται **Πίνακας Συνάφειας/Διασταύρωσης.**

Ένας πίνακας *διασταύρωσης* αποτελείται από γραμμές και στήλες οι οποίες ορίζονται από τις κατηγορίες των μεταβλητών. Στα κελιά του πίνακα μπορεί να εμφανίζονται οι συχνότητες ή οι σχετικές συχνότητες κάθε περίπτωσης.

Πίνακας Διασταύρωσης με το EXCEL: Για την δημιουργία Πίνακα Διασταύρωσης στο excel ακολουθούμε τα παρακάτω βήματα:

- Επιλογή πίνακα →***Καρτέλα Εισαγωγή*** →***Συγκεντρωτικοί πίνακες***→
- Στις γραμμές σύρουμε την μεταβλητή «Φύλο» και στις στήλες την μεταβλητή «Καπνιστής».
- Στο πλαίσιο Τιμές σύρουμε την μεταβλητή «Καπνιστής» →***κλικ*** στο *βελάκι* →***Επιλογή*** του ***count.***

Έτσι προκύπτει ο παρακάτω πίνακας διασταύρωσης



2.12 Εικόνα Υπολογισμός Πίνακα συνάφειας

Αντίστοιχα, αν θέλουμε να δημιουργήσουμε και πίνακα σχετικών συχνοτήτων επαναλαμβάνουμε την παραπάνω διαδικασία και αφού επιλέξουμε το *count,* στην συνέχεια επιλέγουμε από το ίδιο παράθυρο το *εμφάνιση τιμών* και διαλέγουμε μία από τις επιλογές: ως ποσοστό γραμμής ή στήλης ή επί του συνόλου.

2.2 Μέτρα θέσης

Τα μέτρα θέσης μιας κατανομής, είναι αριθμητικά μεγέθη που δίνουν τη θέση του "κέντρου" των παρατηρήσεων στον οριζόντιο άξονα. Με άλλα λόγια, εκφράζουν την "κατά μέσο όρο" απόστασή τους από την αρχή των αξόνων.

2.2.1 Αριθμητικός μέσος

Ο **Αριθμητικός Μέσος** ή απλά **Μέσος** ενός συνόλου παρατηρήσεων ισούται με το πηλίκο του αθροίσματος των τιμών της θεωρούμενης μεταβλητής δια του πλήθους των παρατηρήσεων. Όταν υπολογίζουμε το μέσο του δείγματος αναφερόμαστε στον δειγματικό μέσο ή απλά μέσο όρο, ενώ όταν υπολογίζουμε το μέσο του πληθυσμού αναφερόμαστε στη μέση τιμή του πληθυσμού η οποία είναι στατιστική παράμετρος. Είναι φανερό ότι ο υπολογισμός του αριθμητικού μέσου έχει νόημα μόνο στις ποσοτικές μεταβλητές.

Παράδειγμα 2ο

Οι βαθμολογίες 10 σπουδαστών στο μάθημα της Στατιστικής, είναι οι παρακάτω.
80, 45,.70, 100, 90, 20, 85, 73, 50, 64. Για να υπολογιστεί ο αριθμητικός μέσος των επιδόσεων των σπουδαστών προσθέτουμε όλες τις επιδόσεις και διαιρούμε με το πλήθος των σπουδαστών. Σε συμβολική μορφή:

$$\boldsymbol{Αριθμητικός\ Μέσος} = \frac{80 + 45 + 70 + 100 + 90 + 20 + 85 + 73 + 50 + 64}{10} = \mathbf{67,7}$$

Παράδειγμα 3ο

Οι μηνιαίοι μισθοί 5 εργαζομένων είναι: 350, 500, 500, 500, 4000, άρα ο μέσος μισθός είναι:

$$\boldsymbol{Μέσος\ Μισθός} = \frac{350 + 500 + 500 + 500 + 4000}{5} = \frac{5850}{5} = \mathbf{1170}$$

Τα άθροισμα των μισθών των εργαζομένων είναι 5850 €. Αν οι εργαζόμενοι μοίραζαν εξίσου τα χρήματα αυτά μεταξύ τους, τότε καθένας θα είχε μισθό 1170 €.
Από το παραπάνω παράδειγμα φαίνεται καθαρά ότι η τιμή του αριθμητικού μέσου επηρεάζεται πολύ από τις ακραίες τιμές (πολύ μικρές ή πολύ μεγάλες).

Υπολογισμός μέσου όρου με την βοήθεια πίνακα συχνοτήτων

Στο δεύτερο παράδειγμα όλες οι τιμές είχαν την ίδια βαρύτητα. Πολλές φορές όμως δεν συμβαίνει κάτι τέτοιο, όπως στο 3ο παράδειγμα αλλά και στο 4ο παράδειγμα, που ακολουθεί. Στην περίπτωση αυτή δεν είναι σωστό για τον υπολογισμό του μέσου να προσθέσουμε τις τιμές, δηλαδή στο δεύτερο παράδειγμα ο υπολογισμός του μέσου να γίνει ((0+2+5)/3).

Ο υπολογισμός του αριθμητικού μέσου, στην περίπτωση αυτή, έχει ως ακολούθως.

Υπολογίζουμε τις συχνότητες ή τις σχετικές συχνότητες. Έτσι, για τα δεδομένα (Αριθμός ημερών άδειας) (Παράδειγμα 4ο) για τον υπολογισμό του μέσου αθροίζουμε τα γινόμενα των τιμών των παρατηρήσεων με την αντίστοιχη συχνότητα και το άθροισμα το διαιρούμε με το άθροισμα των συχνοτήτων.

Παράδειγμα 4ο

Ημέρες άδειας	Συχνότητα	Σχετική συχνότητα
0	10	50,00%
2	7	35,00%
5	3	15,00%
Σύνολο	**20**	**100,00%**

Με χρήση των συχνοτήτων ο υπολογισμός του μέσου γίνεται ως εξής:

$$\boldsymbol{Μέσος} = \frac{0 * 10 + 2 * 7 + 5 * 3}{20} = \frac{29}{20} = \mathbf{1,45}\ \boldsymbol{ημέρες\ άδειας}$$

Με χρήση των σχετικών συχνοτήτων έχουμε:

$$\boldsymbol{Μέσος} = 0 * 50\% + 2 * 35\% + 5 * 15\% = \mathbf{1,45}\ \boldsymbol{ημέρες\ άδειας}$$

Όπως φαίνεται από τα παραπάνω, όταν οι τιμές της μεταβλητής ομαδοποιούνται τότε είναι δυνατό να υπολογίζεται ο αριθμητικός μέσος ως **Σταθμικός μέσος**,.

Παράδειγμα 5ο Η πρώτη στήλη του παρακάτω πίνακα περιλαμβάνει τους παράγοντες για την αξιολόγηση των υπαλλήλων. Στην δεύτερη στήλη του πίνακα, παρουσιάζεται η βαρύτητα τριών παραγόντων (Α, Β, Γ) στην αξιολόγηση των υπαλλήλων.

Παράγοντες	Βαρύτητα
Α	50
Β	70
Γ	30

Για τον υπολογισμό της μέσης βαθμολογίας ενός υπαλλήλου ο οποίος βαθμολογήθηκε ως προς τον παράγοντα Α με 65, ως προς τον παράγοντα Β με 54 και ως προς τον παράγοντα Γ με 90, θα υπολογισθεί ο *σταθμικός μέσος*, δηλαδή $\frac{65*50+54*70+90*30}{50+70+30} = 64{,}9$

Παρατηρήσεις:

1. Ο αριθμητικός μέσος δεν εκφράζει, υποχρεωτικά, κάποια τιμή που παρατηρήθηκε.
2. Το μεγαλύτερο μέρος των παρατηρήσεων δεν βρίσκεται απαραίτητα κοντά στον μέσο.
3. Ο μέσος επηρεάζεται από την ύπαρξη ακραίων τιμών, δηλαδή τιμών με μικρή συχνότητα εμφάνισης και μακριά από τις τιμές του κύριου όγκου των δεδομένων, όπως είναι η τιμή 4000 στο 2ο παράδειγμα.

Περικομμένος ή Τετριμμένος (trimmed mean) Μέσος

Όπως είδαμε στο τρίτο παράδειγμα ο μέσος επηρεάζεται από την ύπαρξη ακραίων τιμών, με αποτέλεσμα να μην είναι αντιπροσωπευτικός σε κάποιες περιπτώσεις. Ένας τρόπος για να έχουμε μία καλύτερη εικόνα αναφορικά με τον μέσο που εκφράζει τα δεδομένα μας, είναι να εξαιρέσουμε ένα ποσοστό από τις μεγαλύτερες και το ίδιο ποσοστό από τις μικρότερες τιμές των δεδομένων μας και να υπολογίσουμε τον μέσο με τη βοήθεια των δεδομένων που που απομένουν. Ο μέσος που προκύπτει από αυτή την διαδικασία ονομάζεται **Τετριμμένος** ή αλλιώς **Περικομμένος Μέσος.**

Παράδειγμα 6ο

Οι βαθμολογίες 20 σπουδαστών της Εθνικής Σχολής Δημόσιας Διοίκησης και Αυτοδιοίκησης στο μάθημα της στατιστικής είναι: 1, 2, 75, 75, 89, 87, 90, 90, 92, 93, 94, 94, 94, 95, 97, 100, 100,100,100, 100. Η μέση βαθμολογία των 20 σπουδαστών είναι 83,4 και λαμβάνεται α=*0,1 (10%)*. Στην περίπτωση αυτή ο περικομμένος μέσος, δηλαδή ο μέσος που θα υπολογιστεί αν εξαιρεθεί το 5% (5%*20=1 παρατήρηση) των μικρότερων και το 5% των μεγαλύτερων παρατηρήσεων, είναι 87,06. Από τον υπολογισμό του περικομμένου μέσου εξαιρέθηκαν οι τιμές 1 και 100.

Είναι φανερό ότι ο περικομμένος μέσος είναι ο αριθμητικός μέσος των παρατηρήσεων που προκύπτει αν εξαιρέσουμε το 50*α% των μεγαλύτερων τιμών των παρατηρήσεων και το 50*α% των μικρότερων τιμών των παρατηρήσεων.

Λόγω του ότι από τον υπολογισμό του περικομμένου μέσου έχουν εξαιρεθεί οι πολύ ακραίες τιμές (άνω και κάτω) είναι προφανές ότι ο περικομμένος μέσος επηρεάζεται λιγότερο από τις ακραίες τιμές των παρατηρήσεων από ότι ο αριθμητικός μέσος.

2.2.2 Διάμεσος

Η **Διάμεσος** ενός συνόλου παρατηρήσεων είναι η τιμή εκείνη για την οποία ισχύει ότι το πολύ το 50% των παρατηρήσεων είναι μικρότερες από αυτήν και το πολύ το 50% των παρατηρήσεων είναι μεγαλύτερο από αυτήν.

Για τον υπολογισμό της διαμέσου διατάσσουμε τις παρατηρήσεις κατά αύξουσα σειρά (από την μικρότερη στην μεγαλύτερη παρατήρηση). Αν το πλήθος των παρατηρήσεων είναι:

- Π*εριττός αριθμός* τότε η διάμεσος ισούται με την *μεσαία παρατήρηση*.
- *Άρτιος αριθμός* τότε η διάμεσος ισούται με το *ημιάθροισμα των δύο μεσαίων παρατηρήσεων*.

Για τον υπολογισμό της θέσης της μεσαίας παρατήρησης (ή των δύο μεσαίων παρατηρήσεων) στην διατεταγμένη κατά αύξουσα σειρά των παρατηρήσεων, προσθέτουμε στο πλήθος των παρατηρήσεων την μονάδα και στην συνέχεια διαιρούμε με το 2.

Παράδειγμα 7ο

- **Περιττός αριθμός παρατηρήσεων**

Ο αριθμός των μελών 9 νοικοκυριών σε μία πολυκατοικία παρουσιάζεται στον παρακάτω πίνακα κατά αύξουσα σειρά.

Νοικοκυριό	1ο	2ο	3ο	4ο	**5ο**	6ο	7ο	8ο	9ο
Αριθμός μελών	1	1	2	2	**2**	3	4	5	5

2.13Πίνακας Αριθμός μελών νοικοκυριού παράδειγμα-1

Θέση Διαμέσου=(9+1)/2=5, άρα η 5η παρατήρηση είναι η διάμεσος των δεδομένων και έχει 2 μέλη. Συνεπώς το πολύ το 50% των νοικοκυριών έχει λιγότερα από 2 μέλη και το πολύ το 50% των νοικοκυριών έχει περισσότερα από 2 μέλη.

- **Άρτιος αριθμός παρατηρήσεων**

Αν ο αριθμός των νοικοκυριών ήταν 10 όπως φαίνεται στον πίνακα:

Νοικοκυριό	1ο	2ο	3ο	4ο	5ο	6ο	7ο	8ο	9ο	10ο
Αριθμός μελών	1	1	2	2	2	3	4	5	5	6

2.14Αριθμός μελών νοικοκυριού παράδειγμα 2

τότε η *Θέση Διαμέσου=(10+1)/2=5,5*, άρα υπάρχουν δύο μεσαίες παρατηρήσεις η 5η και η 6η με τιμές 2 και 3 αντίστοιχα. Συνεπώς η διάμεσος θα προκύψει από το ημιάθροισμα της 5ης και της 6ης παρατήρησης, και θα έχει τιμή (2+3)/2=2,5. Συνεπώς το πολύ το 50% των νοικοκυριών έχει λιγότερα από 2,5 μέλη, δηλαδή από δύο και κάτω μέλη και το πολύ το 50% των νοικοκυριών έχει περισσότερα από 2,5 μέλη δηλαδή από 3 και πάνω μέλη.

Η διάμεσος δεν επηρεάζεται από ακραίες τιμές σε αντίθεση με το αριθμητικό μέσο. Για παράδειγμα αν η μεγαλύτερη τιμή στα δεδομένα μας ήταν 15 και όχι 6, ο προηγούμενος πίνακας θα διαμορφωνόταν ως εξής:

Νοικοκυριό	1ο	2ο	3ο	4ο	**5ο**	**6ο**	7ο	8ο	9ο	10ο
Αριθμός μελών	1	1	2	2	**2**	**3**	4	5	5	15

Λόγω του ότι η διάμεσος είναι πάλι το ημιάθροισμα της 5ης και της 6ης παρατήρησης, η τιμή της παραμένει 2,5.

2.2.3 Επικρατούσα τιμή

Επικρατούσα τιμή είναι η τιμή με την μεγαλύτερη συχνότητα εμφάνισης. Από τον παρακάτω πίνακα συχνοτήτων παρατηρούμε ότι το 0 είναι το πλήθος ημερών άδειας με την μεγαλύτερη συχνότητα εμφάνισης.

Ημέρες άδειας	**Συχνότητα**
0	10
2	7
5	3
Σύνολο	**20**

2.15*Πίνακας Επικρατούσα τιμή 1ο παράδειγμα*

Σε μια σειρά παρατηρήσεων μπορούμε να έχουμε περισσότερες από μια επικρατούσες τιμές. Θα μπορούσε για παράδειγμα ο παρακάτω πίνακας να είχε διαμορφωθεί ως εξής:

Ημέρες άδειας	**Συχνότητα**
0	8
2	4
5	8
Σύνολο	**20**

2.16*Πίνακας Επικρατούσα τιμή 2ο παράδειγμα*

Οπότε η τιμή 0 και η τιμή 5 θα είχαν την ίδια συχνότητα η οποία θα ήταν και η μεγαλύτερη σε σχέση με τις συχνότητες των υπολοίπων παρατηρήσεων.

Συνοψίζοντας τα παραπάνω, θα πρέπει κανείς να θυμάται για

τον **Αριθμητικό μέσο**

- Χρησιμοποιείται μόνο σε ποσοτικά δεδομένα
- Είναι ευαίσθητος στην επίδραση ακραίων τιμών
- Είναι κατάλληλος για θεωρητική ανάπτυξη
- Χρησιμοποιείται στην συμπερασματολογία για άθροισμα τιμών.
- Ο υπολογισμός του βασίζεται σε όλες τις παρατηρούμενες τιμές.

τη **Διάμεσο**

- Χρησιμοποιείται σε ποσοτικά δεδομένα και σε δεδομένα διάταξης
- Δεν είναι ευαίσθητη στην επίδραση ακραίων τιμών
- Δεν είναι εύκολη η θεωρητική ανάπτυξη
- Δεν χρησιμοποιείται στην συμπερασματολογία για άθροισμα τιμών.
- Ο υπολογισμός της δεν συμπεριλαμβάνει όλες τις παρατηρούμενες τιμές.

την **Επικρατούσα Τιμή**

- Χρησιμοποιείται σε ποσοτικά και ποιοτικά δεδομένα
- Δεν είναι ευαίσθητη στην επίδραση ακραίων τιμών
- Δεν είναι εύκολη η θεωρητική ανάπτυξη
- Δεν χρησιμοποιείται στην συμπερασματολογία για άθροισμα τιμών.
- Ο υπολογισμός της δεν συμπεριλαμβάνει όλες τις παρατηρούμενες τιμές.

2.2.4 Ποσοστιαία Σημεία

Για την μελέτη μιας σειράς δεδομένων, μας ενδιαφέρει συχνά η θέση μιας τιμής σε σχέση με τις υπόλοιπες. Δηλαδή μας ενδιαφέρει να γνωρίζουμε ποιο ποσοστό παρατηρήσεων είναι μικρότερο από μια συγκεκριμένη τιμή. Η πληροφορία αυτή παρέχεται από τα **Ποσοστιαία Σημεία ή ποσοστημόρια**. Θα λέμε ότι μια τιμή χ είναι το α ποσοστιαίο σημείο της κατανομής αν το πολύ το α% των παρατηρήσεων είναι μικρότερες από το χ και το πολύ το (100-α)% των παρατηρήσεων είναι μεγαλύτερες από το χ.

Είναι φανερό ότι η **Διάμεσος** είναι το **50ο Ποσοστιαίο Σημείο** αφού το πολύ το 50% των παρατηρήσεων είναι μικρότερες από αυτήν και το πολύ το 50% των παρατηρήσεων είναι μεγαλύτερες από αυτήν.

Το **25ο Ποσοστιαίο Σημείο** ονομάζεται **πρώτο τεταρτημόριο Q_1**, αφού το πολύ το 25% των παρατηρήσεων είναι μικρότερες από αυτό και το πολύ το 75% των παρατηρήσεων είναι μεγαλύτερες από αυτό.

Αντίστοιχα ορίζεται **το 3ο τεταρτημόριο Q_3** για το οποίο ισχύει ότι το πολύ το 75% των παρατηρήσεων είναι μικρότερες από αυτό και το πολύ το 25% των παρατηρήσεων είναι μεγαλύτερες από αυτό.

Ας εξετάσουμε παρακάτω μια εφαρμογή υπολογισμού των Ποσοστιαίων Σημείων για τα παρακάτω δεδομένα που αφορούν στις βαθμολογίες 20 σπουδαστών στο μάθημα της Στατιστικής. Συγκεκριμένα έχουμε:

Σπουδαστής/ια	Βαθμολογία	Σπουδαστής/ια	Βαθμολογία
Σπ1	1	Σπ11	94
Σπ2	2	Σπ12	94
Σπ3	75	Σπ13	94
Σπ4	75	Σπ14	95
Σπ5	89	Σπ15	97
Σπ6	87	Σπ16	100
Σπ7	90	Σπ17	100
Σπ8	90	Σπ18	100
Σπ9	92	Σπ19	100
Σπ10	93	Σπ20	100

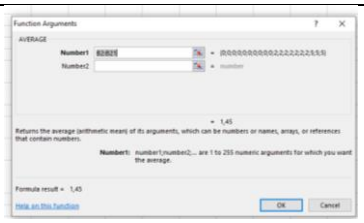
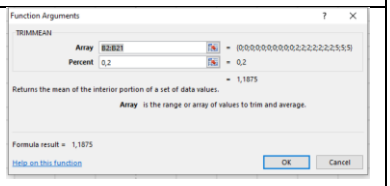
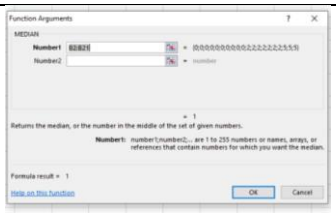
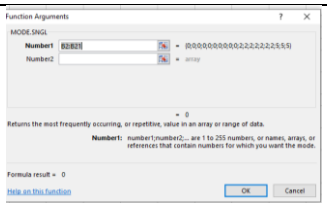
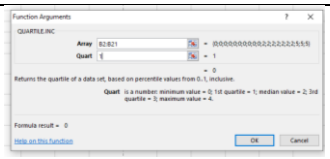
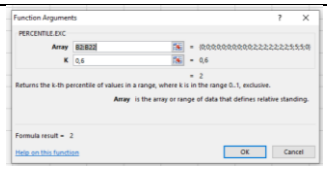
2.17Πίνακας Βαθμολογία σπουδαστών

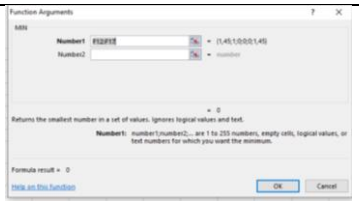
Το 10ο Ποσοστιαίο Σημείο είναι 67,7, το οποίο σημαίνει ότι το πολύ το 10% των σπουδαστών της σχολής είχε μικρότερη βαθμολογία από 67,7 και το πολύ το 90% της σχολής είχε

βαθμολογία μεγαλύτερη από 67,7. Παρατηρούμε ότι το 10ο ποσοστιαίο σημείο είναι βαθμολογία μεταξύ της 2ης και 3ης παρατήρησης. Το πρώτο τεταρτημόριο Q1= 88,5, το δεύτερο τεταρτημόριο δηλαδή η διάμεσος=93,5 και το τρίτο τεταρτημόριο Q3=97,75

Πίνακας υπολογισμού με το EXCEL

Στον παρακάτω πίνακα, οι αναφορές των κελιών είναι ενδεικτικές των αναφορών των κελιών που θα χρησιμοποιηθούν.

Μέτρο	Συνάρτηση	
Αριθμητικός μέσος	=AVERAGE(B2:B21)	
Σταθμικός μέσος	=SUMPRODUCT(A3:A5;B3:B5)/SUM(B3:B5)	
Περικομμένος μέσος Τετριμμένος μέσος	=TRIMMEAN(B2:B21;0,2)	
Διάμεσος	=MEDIAN(B2:B21)	
Επικρατούσα τιμή	=MODE.SNGL(B2:B21)	
Τεταρτημόρια	=QUARTILE.INC(B2:B21;Q) Όπου Q το δεκαδικό που εκφράζει το υπολογιζόμενο τεταρτημόριο	
Κ-ποσοστημόριο	=PERCENTILE.INC(B2:B21;Q) Όπου Q το δεκαδικό που εκφράζει το υπολογιζόμενο ποσοστημόριο	

Μέγιστο	=MAX(B2:B21)	
Ελάχιστο	=MIN(B2:B21)	

2.18Πίνακας συναρτήσεων μέτρων θέσης

2.3 Μέτρα μεταβλητότητας

Μέχρι τώρα αναφερθήκαμε σε μέτρα θέσης για μια σειρά δεδομένων. Όμως μας ενδιαφέρει και η διασπορά των δεδομένων. Δηλαδή μας ενδιαφέρει να μελετήσουμε και να μετρήσουμε την μεταβλητότητα των δεδομένων. Παρακάτω παρουσιάζονται τα πιο συνήθη μέτρα μεταβλητότητας, αυτά είναι τα εξής:

- Το **Εύρος**,
- Το **Ενδοτεταρτημοριακό Εύρος**,
- Η **Διακύμανση** και
- Η **Τυπική Απόκλιση**.

2.3.1 Εύρος

Ως **Εύρος (R)** ορίζουμε *την διαφορά της μικρότερης από την μεγαλύτερη τιμή των δεδομένων*. Για τα δεδομένα του προηγούμενου πίνακα με τις βαθμολογίες των σπουδαστών, η ελάχιστη τιμή είναι το 1 και μέγιστη το 100 άρα *Εύρος (range) R=100-1=99*. Είναι απλό στον υπολογισμό του αλλά, όπως είναι προφανές, η τιμή του Εύρους εξαρτάται μόνο από δύο παρατηρήσεις και πιο ειδικά αυτή που έδωσε τη μικρότερη τιμή και αυτή που αντιστοιχεί στη μεγαλύτερη τιμή. Ως εκ τούτου, δεν δίνει καμία πληροφορία για την μεταβλητότητα των υπολοίπων τιμών.

2.3.2 Ενδοτεταρτημοριακό Εύρος

Το **Ενδοτεταρτημοριακό Εύρος (IQR)** ορίζεται ως *η διαφορά του πρώτου τεταρτημορίου από το τρίτο,* δηλαδή $IQR=Q_3-Q_1$.

Στο παράδειγμα με τις βαθμολογίες των σπουδαστών Το πρώτο τεταρτημόριο Q1= 88,5, και το τρίτο τεταρτημόριο Q3=97,75. Άρα $IR=Q_3-Q_1=97,75-88,5=9,25$

Το Ενδοτεταρτημοριακό Εύρος είναι το εύρος των τιμών που θα προκύψουν αν εξαιρεθεί από το αρχικό σύνολο δεδομένων το 25% των μεγαλύτερων και το 25% των μικρότερων τιμών των παρατηρήσεων. Κατά συνέπεια, σε αντίθεση με το Εύρος, το Ενδοτεταρτημοριακό Εύρος δεν επηρεάζεται σοβαρά από τις πολύ ακραίες τιμές.

2.3.3 Διακύμανση Διασπορά

Ένα μειονέκτημα που έχουν τα προηγούμενα μέτρα μεταβλητότητας είναι ότι η τιμή τους διαμορφώνεται από υποσύνολο των δεδομένων. Αυτό δεν προσφέρει καλή εικόνα για την εξεταζόμενη κατανομή τιμών. Το τελευταίο γίνεται φανερό όταν μελετήσει κανείς το παρακάτω παράδειγμα.

Δίνονται οι τρεις παρακάτω σειρές δεδομένων, για τις οποίες είναι φανερό ότι το εύρος είναι το ίδιο δηλαδή E=100-0=100. Το ίδιο φανερό είναι ότι η μεταβλητότητα δεν είναι ίδια σε αυτές τις τρεις σειρές δεδομένων.

Σειρά 1η	**Σειρά 2η**	**Σειρά 3η**
0	0	0
0	10	41
0	20	41
0	30	41
0	40	42
0	50	42
100	60	43
100	75	44
100	80	44
100	85	44
100	90	44
100	100	100

Συνεπώς υπάρχει η ανάγκη για ένα μέτρο μεταβλητότητας το οποίο θα εξαρτάται από όλες τις τιμές των δεδομένων. Ένας τρόπος για τον υπολογισμό της Μεταβλητότητας θα ήταν να υπολογίσουμε τον Μέσο Όρο των αποκλίσεων των τιμών από τον μέσο όρο τους.

Έστω ότι 5 ράβδοι έχουν μήκη σε cm 1, 5, 7, 7, 10

Ο μέσος τους είναι $\boldsymbol{Μέσος\ όρος} = \frac{1+5+7+7+10}{5} = \mathbf{6,}$ οι αποκλίσεις από τον μέσο προκύπτουν από την αφαίρεση του μέσου από καθεμία τιμή, για παράδειγμα η απόκλιση για την τιμή 1 είναι 1-6=-5. Τα αποτελέσματα αυτού του υπολογισμού φαίνονται στον παρακάτω πίνακα.

Τιμές	1	5	7	7	10
Απόκλιση από τον μέσο	-5	-1	1	1	4

Παρατηρούμε ότι το άθροισμα των αποκλίσεων αξιοποιεί όλες τις παρατηρήσεις μια και βρίσκει το σύνολο των διαφορών από τον μέσο αλλά ισούται με:

(-5)+(-1)+1+1+4=0

Αυτό είναι ένα γεγονός που ισχύει για οποιαδήποτε ομάδα δεδομένων. Για το λόγο αυτό, ως μέτρο μεταβλητότητας προτάθηκε το μέτρο της **Μέσης Απόλυτης Απόκλισης**, δηλαδή να χρησιμοποιηθούν οι απόλυτες τιμές-οι τιμές χωρίς πρόσημο- των παραπάνω αποκλίσεων.

Το μέτρο αυτό παρουσιάζει διάφορα προβλήματα και έτσι προτάθηκε το μέτρο της **Διακύμανσης** ή αλλιώς **Διασποράς**, όπου υπολογίζεται ο μέσος των τετραγώνων των αποκλίσεων, δηλαδή

$$\frac{(-5)^2 + (-1)^2 + (1)^2 + (1)^2 + (4)^2}{5} = \frac{44}{5} = 8{,}8$$

Το πρόβλημα με τη Διακύμανση είναι ότι υπολογίζεται μέσω των τετραγώνων των αποκλίσεων και ως εκ τούτου οι μονάδες μέτρησης της δεν συμπίπτουν και αυτές της μέσης τιμής. Δηλαδή στο παραπάνω παράδειγμα το μήκος της ράβδου εκφράζεται σε cm ενώ η διακύμανση εκφράζεται σε cm^2.

2.3.4 Τυπική Απόκλιση

Προκειμένου να δημιουργηθεί ένα μέτρο μεταβλητότητας το οποίο να:

- Λαμβάνει υπόψη όλες τις παρατηρούμενες τιμές και
- Έχει τις ίδιες μονάδες μέτρησης με τη μέση τιμή των παρατηρούμενων τιμών

Προτάθηκε η χρήση της τυπικής απόκλισης. Αυτή ορίζεται ως η τετραγωνική ρίζα της Διακύμανσης. Ως εκ τούτου, η *Τυπική Απόκλιση* εκφράζεται σε μονάδες που εκφράζονται και οι μετρήσεις των παρατηρήσεων.

Σχόλια:

Συγκρίνοντας τα μεγέθη της Τυπικής Απόκλισης και του Εύρους διαπιστώνει κανείς ότι:

- Το Εύρος εξαρτάται μόνο από τις δύο πιο ακραίες τιμές, δεν χρησιμοποιείται για συμπερασματολογία και είναι δύσκολη η θεωρητική ανάπτυξη μέσω της αξιοποίησής του.
- Η Τυπική Απόκλιση είναι λιγότερο ευαίσθητη στην επίδραση των ακραίων τιμών από το εύρος. Χρησιμοποιούνται όλες οι τιμές για τον υπολογισμό της, χρησιμοποιείται για συμπερασματολογία και είναι κατάλληλο μέτρο για θεωρητική ανάπτυξη.

Πρόταση για παραπέρα μελέτη

Ένα σημαντικό θεώρημα της Στατιστικής το οποίο αξιοποιεί τα μεγέθη της Τυπικής Απόκλισης και του Αριθμητικού μέσου είναι το **θεώρημα του Chebyshev**.

Διατύπωση του θεωρήματος του Chebyshev.

Για ένα σύνολο παρατηρήσεων με μέσο $\bar{\chi}$ και τυπική απόκλιση s το ποσοστό των παρατηρήσεων που περιέχεται σε απόσταση κ τυπικών αποκλίσεων από το μέσο ($k \geq 1$ είναι τουλάχιστον $(1-\frac{1}{k^2})$ *100% των παρατηρήσεων.

Δηλαδή για κ=1 στο διάστημα $(\bar{x}-s\,,\bar{x}+s)$ περιέχεται τουλάχιστον το 0% των παρατηρήσεων.

Για κ=2 στο διάστημα$(\bar{x}-2s\,,\bar{x}+2s)$ περιέχεται τουλάχιστον το (1-1/4)100%=75% των παρατηρήσεων. Τα παραπάνω εφαρμόζονται χωρίς να γνωρίζουμε κάποια πληροφορία για την κατανομή των παρατηρήσεων.

Ας δούμε ένα παράδειγμα του θεωρήματος που η μέση βαθμολογία των σπουδαστών μιας σχολής είναι 90 με τυπική απόκλιση 2. Γίνεται ο υπολογισμός των άκρων των διαστημάτων:

- μέσος∓1 τυπική απόκλιση: 90∓1*2→ Στο (88 , 92) ανήκει τουλάχιστον το 0%
- μέσος∓2 τυπικές αποκλίσεις: 90∓2*2→Στο (86 , 94) ανήκει τουλάχιστον το 75%
- μέσος∓3 τυπικές αποκλίσεις: 90∓3*2→Στο (84 , 96) ανήκει τουλάχιστον το 88,9%

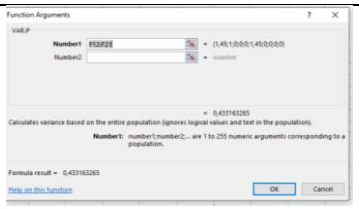
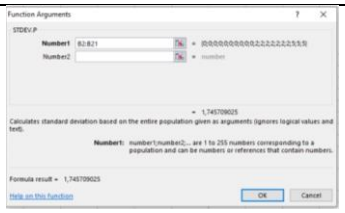
Σύμφωνα με το παραπάνω θεώρημα, αυτό σημαίνει ότι:

- Από 88 έως 92 έχει γράψει τουλάχιστον το 0% των σπουδαστών
- Από 86 έως 94 έχει γράψει τουλάχιστον το 75% των σπουδαστών

Από 84 έως 96 έχει γράψει τουλάχιστον το 88,9% των σπουδαστών.

Έτσι αν ένας σπουδαστής έχει γράψει 80, η οποία είναι μία καλή βαθμολογία, σε σχέση με τους υπόλοιπους σπουδαστές δεν έχει πάει καλά αφού πολύ μεγάλο ποσοστό των σπουδαστών έχει καλύτερη βαθμολογία από αυτόν.

Υπολογισμός των μέτρων μεταβλητότητας με το EXCEL

Μέτρο	Συνάρτηση	
Εύρος	=MAX(B2:B21)-MIN(B2:B21)	
Ενδοτεταρτημοριακό εύρος	=QUARTILE.INC(B2:B21;0,75)-QUARTILE.INC(B2:B21;0,25)	
Διακύμανση	=VAR.P(F12:F21)	
Τυπική απόκλιση	=STDEV.P(B2:B21)	

2.19*Πίνακας συναρτήσεων υπολογισμού μέτρων μεταβλητότητας*

2.4 Μέτρα σχετικής θέσης - μεταβλητότητας

2.4.1 Τυποποιημένες τιμές

Πολλές φορές υπάρχει η ανάγκη εξέτασης τιμών οι οποίες ανήκουν σε διαφορετικές κατανομές. Έστω ότι έχουμε, για παράδειγμα, ένα μαθητή ο οποίος εμφάνισε επίδοση σε τρία μαθήματα Α, Β, Γ τους βαθμούς 16, 12, 14 αντίστοιχα. Θα ήταν ορθό να ισχυριστούμε ότι ο συγκεκριμένος μαθητής είναι καλύτερος στο μάθημα Α από ότι στα Β και Γ. Η απάντηση είναι όχι γιατί οι επιδόσεις αυτές ανήκουν σε διαφορετικές κατανομές με διαφορετικά χαρακτηριστικά όπως αυτά καταγράφονται από τον μέσο όρο και την τυπική απόκλιση.

Έστω ότι για το μάθημα:

- Α η μέση επίδοση ήταν 17 με τυπική απόκλιση 1.
- Β η μέση επίδοση ήταν 10 με τυπική απόκλιση 1 και
- Γ η μέση επίδοση ήταν 13 με τυπική απόκλιση 1.

τότε παρατηρούμε ότι ο μαθητής στο μάθημα Α που έχει την μεγαλύτερη βαθμολογία είναι κάτω από τη μέση βαθμολογία κατά μια Τυπική Απόκλιση, ενώ στο μάθημα Β είναι πάνω από τη μέση βαθμολογία κατά δύο Τυπικές Αποκλίσεις και στο μάθημα Γ είναι πάνω από τη μέση βαθμολογία κατά μία Τυπική Απόκλιση.

Έτσι ορίστηκαν οι τυποποιημένες τιμές **Z-τιμές** ή αλλιώς **Z-score** ώστε να είναι δυνατή η σύγκριση τιμών από διαφορετικές κατανομές. Στο παράδειγμα με τις βαθμολογίες έχουμε:

$$Z_A = \frac{16-17}{1} = -1$$
$$Z_B = \frac{12-10}{1} = 2$$
$$Z_\Gamma = \frac{14-13}{1} = 1$$

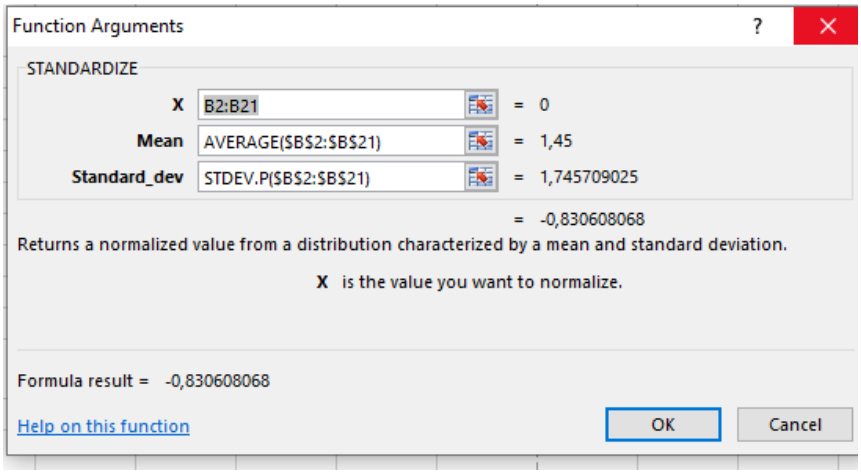
Οι τυποποιημένες τιμές δηλαδή μας δείχνουν πόσες τυπικές Αποκλίσεις απέχει μια συγκεκριμένη τιμή από το Μέσο. Μια *αρνητική τυποποιημένη τιμή* δηλώνει ότι η παρατήρηση είναι *μικρότερη του μέσου* ενώ μια *θετική τυποποιημένη τιμή* ότι η παρατήρηση είναι *μεγαλύτερη του μέσου*.

Ένα άλλο πλεονέκτημα της τυποποίησης των τιμών είναι ότι οι τυποποιημένες τιμές είναι καθαροί αριθμοί απαλλαγμένοι από μονάδες οπότε υπάρχει η δυνατότητα σύγκρισης μεταξύ τιμών οι οποίες εκφράζονται με διαφορετικές μονάδες μέτρησης.

Η τυποποίηση τιμών με το EXCEL

Έστω ότι τα δεδομένα για τα οποία θέλουμε να υπολογίσουμε τις τυποποιημένες τιμές βρίσκονται στα κελιά από *B2 έως και B21*

Στο κελί C2 εισάγουμε την συνάρτηση **STANDARDIZE** και στο παράθυρο που εμφανίζεται συμπληρώνουμε τα πεδία όπως φαίνεται παρακάτω. Οι αναφορές των πεδίων είναι ενδεικτικές.



2.20Εικόνα: Υπολογισμός τυποποιημένων τιμών

2.4.2 Συντελεστής μεταβλητότητας

Η γνώση του Μέσου και της Τυπικής Απόκλισης για μια σειρά παρατηρήσεων μας δίνει αρκετές πληροφορίες για αυτές. Όμως όταν θέλουμε να συγκρίνουμε την ομοιογένεια διαφορετικών ομάδων σε σχέση με ένα ποσοτικό χαρακτηριστικό, η γνώση του Μέσου και της Τυπικής Απόκλισης δεν είναι αρκετή αφού τα δεδομένα στις ομάδες αυτές μπορεί να εκφράζονται σε διαφορετικές μονάδες μέτρησης ή να έχουν διαφορετικούς μέσους.

Για παράδειγμα αν ο μέσος μηνιαίος μισθός των υπαλλήλων μιας εταιρίας Α είναι 1000€ με Τυπική Απόκλιση 150€ και σε μια εταιρία Β ο μέσος μισθός των υπαλλήλων είναι 2000$ με Τυπική Απόκλιση 150 $ ποια σύγκριση θα μπορούσε να κάνει κανείς για τα δύο είδη αμοιβών;

Έτσι υπήρξε ανάγκη δημιουργίας ενός άλλου μέτρου για την δυνατότητα σύγκρισης της ομοιογένειας των δύο ομάδων. Αυτό το μέτρο είναι ο **Συντελεστής Μεταβλητότητας CV-Coefficient of Variation.**

Ο Συντελεστής Μεταβλητότητας ισούται με *το πηλίκο της τυπικής απόκλισης των τιμών προς την απόλυτη τιμή του μέσου.*[5]

$$\textbf{Συντελεστής Μεταβλητότητας CV}=\frac{\textbf{τυπική απόκλιση}}{\textbf{απόλυτη τιμή του μέσου}}$$

Ο *Συντελεστής Μεταβλητότητας* εκφράζει την *Τυπική Απόκλιση* ως *ποσοστό του μέσου*. Είναι απαλλαγμένος από μονάδες άρα μπορεί να χρησιμοποιηθεί για την σύγκριση ομάδων τα δεδομένα των οποίων είναι εκφρασμένα σε διαφορετικές μονάδες.

Ακόμη υπάρχει δυνατότητα σύγκρισης ομάδων που οι μέσοι τους διαφέρουν σημαντικά. Αν ο Συντελεστής Μεταβλητότητας μιας ομάδας είναι μικρότερος από τον Συντελεστή Μεταβλητότητας μιας άλλης τότε μπορούμε να πούμε ότι η ομάδα με τον μικρότερο Συντελεστή Μεταβλητότητας είναι περισσότερο ομοιογενής.

Στο παράδειγμα των δύο εταιριών $CV_A = \frac{150}{1000} = 0{,}15$, $CV_B = \frac{150}{2000} = 0{,}075$

δηλαδή στην εταιρία Α η Τυπική Απόκλιση ισούται με το 15% του μέσου, ενώ στη δεύτερη η Τυπική Απόκλιση ισούται με το 7,5% του μέσου., άρα οι μισθοί των υπαλλήλων στην εταιρία Α έχουν μεγαλύτερη σχετική μεταβλητότητα από ότι στην εταιρία Β. Άρα οι μισθοί στην εταιρία Β έχουν μεγαλύτερη ομοιογένεια σε σχέση με την εταιρία Α.

Ισχύει ότι αν ο Συντελεστής Μεταβλητότητας είναι μικρότερος ή ίσος του 10% τα δεδομένα θεωρούνται ομοιογενή διαφορετικά ανομοιογενή.

Ας δούμε τις παρακάτω ομάδες δεδομένων

A	1	9	17	22	30
B	1	1	1	30	30

2.21*Πίνακας Συντελεστής μεταβλητότητας*

[5] Απόλυτη τιμή είναι ο αριθμός χωρίς το πρόσημό του. π.χ. η απόλυτη τιμή του -5 ισούται με 5, όπως και η απόλυτη τιμή του 5 ισούται με 5.

Παρατηρώντας τις δύο ομάδες δεδομένων διαπιστώνουμε ότι η πρώτη ομάδα έχει μεγαλύτερη ομοιογένεια από την δεύτερη. Πράγματι όπως φαίνεται και από τον παρακάτω πίνακα, ο Συντελεστής Μεταβλητότητας της πρώτης ομάδας είναι 63,72%, ενώ ο Συντελεστής Μεταβλητότητας της δεύτερης ομάδας ισούται με 112,75%, δηλαδή τα δεδομένα της δεύτερης ομάδας είναι περισσότερο ανομοιογενή από τα δεδομένα της πρώτης ομάδας.

Τυπική Απόκλιση	Μέσος	Συντελεστής Μεταβλητότητας	
10,07	15,80	0,6372	63,72%
14,21	12,60	1,1275	112,75%

2.22Πίνακας: Σύγκριση συντελεστή μεταβλητότητας μεταξύ δύο ομάδων

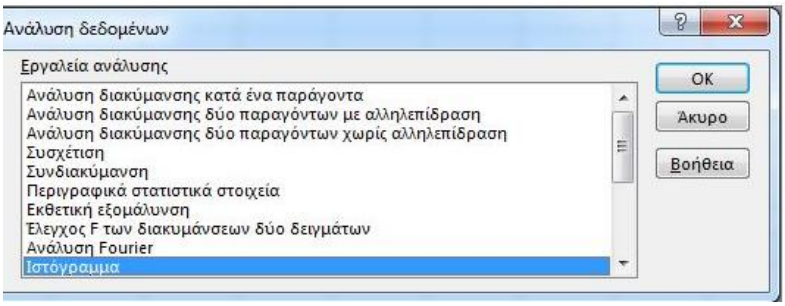
2.5 Υπολογισμός μέτρων θέσης και μεταβλητότητας με χρήση του πρόσθετου «Πακέτο Ανάλυσης Δεδομένων»

Το EXCEL είναι μια εφαρμογή λογιστικών φύλλων για καθημερινές ανάγκες. Για να χρησιμοποιήσουμε το πακέτο ανάλυσης δεδομένων σε ειδικότερες περιπτώσεις καταφεύγουμε στο πρόσθετο πακέτο ανάλυσης δεδομένων του EXCEL. Αν είναι ενεργοποιημένο τότε εμφανίζεται στην καρτέλα ***«Δεδομένα»*** , το πρόσθετο *Ανάλυση Δεδομένων*. Εάν δεν εμφανίζεται το πρόσθετο αυτό, τότε εκτελούμε τις παρακάτω ενέργειες.

Αρχείο→Επιλογές→Πρόσθετα→Πακέτο Ανάλυσης Δεδομένων

Αν θέλουμε να δούμε συγκεντρωτικά τα περιγραφικά μέτρα, εκτελούμε τα παρακάτω βήματα.

Δεδομένα→Ανάλυση Δεδομένων→Περιγραφικά στατιστικά μέτρα

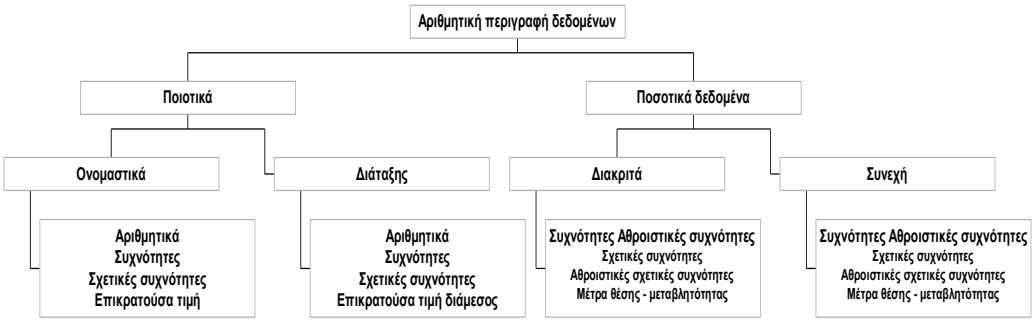


2.23Εικόνα Πακέτο ανάλυσης δεδομένων

Στην περιοχή εισόδου εισάγουμε τα δεδομένα μας και επιλέγουμε το περιληπτικά μέτρα, οπότε εμφανίζονται όλα τα περιγραφικά μέτρα.

Συνοψίζοντας

Στο παρακάτω διάγραμμα απεικονίζονται τα μέτρα που μπορούν να υπολογισθούν σε σχέση με το είδος των δεδομένων.



2.24Συγκεντρωτικό διάγραμμα Αριθμητικής περιγραφής δεδομένων

Άσκηση 2ου κεφαλαίου

Στον ακόλουθο υπερσύνδεσμο ΕΕ θα βρείτε στοιχεία που αφορούν τα ποσοστά απασχόλησης στις χώρες της ΕΕ.

1. Μελετήστε τα μεταδεδομένα του αρχείου δεδομένων.
2. Κατεβάστε το αρχείο των δεδομένων σε μορφή xls.
3. Υπολογίστε τα συνήθη στατιστικά μέτρα θέσης και μεταβλητότητας για κάθε εγγραφή του αρχείου που αφορούν την επίδοση της για την περίοδο από 2008 έως και 2018. Σε περίπτωση που εμφανίζονται πολλαπλές ελλείπουσες τιμές στα δεδομένα μιας εγγραφής εξαιρέστε την εγγραφή αλλιώς αντικαταστήστε την ελλείπουσα τιμή από τη μέση τιμή των δύο γειτονικών ετών.
4. Υπολογίστε την μέση επίδοση όλων των μελών της ΕΕ για το 2015 καθώς και την μέση επίδοση για τις χώρες-μέλη της ΕΕ που ανήκουν στη Βαλκανική. Αξιοποιείστε μόνο τις χώρες που προσφέρονται δεδομένα. Εμφανίζεται διαφορά μεταξύ της στατιστικής παραμέτρου (δηλ. σε επίπεδο ΕΕ) και της εκτίμησης μέσω του δείγματος της Βαλκανικής; Πόσο τις εκατό είναι η διαφορά αυτή; Θεωρείτε ότι θα ήταν αντιπροσωπευτικό δείγμα του πληθυσμού οι εν λόγω χώρες; Δικαιολογείστε την απάντηση σας.

ΚΕΦΑΛΑΙΟ 3ο

ΓΡΑΦΗΜΑΤΑ ΚΑΤΑΝΟΜΕΣ

Στο προηγούμενο κεφάλαιο εξετάσθηκε ο τρόπος δημιουργίας περιλήψεων για σύνολα δεδομένων μέσω στατιστικών συναρτήσεων. Ένας άλλος τρόπος αφορά τη γραφική αναπαράσταση των δεδομένων. Η γραφική παρουσίαση των δεδομένων είναι ένας απλός τρόπος απεικόνισης των δεδομένων που, εκτός των άλλων, επιτρέπει να γίνονται εύκολα αντιληπτές συγκρίσεις και συσχετίσεις μεταξύ των δεδομένων.

Γενικά, η εισαγωγή ενός διαγράμματος στο EXCEL επιτυγχάνεται μέσω των λειτουργιών της καρτέλας Εισαγωγή.

Για να δημιουργήσουμε ένα γράφημα με το **EXCEL**, ακολουθούμε τα παρακάτω βήματα ***Επιλογή δεδομένων → Εισαγωγή → Γράφημα → Επιλογή του κατάλληλου γραφήματος***. Μετά την δημιουργία του γραφήματος έχουμε την δυνατότητα να αλλάξουμε χρώματα μορφή γραφήματος υπόμνημα, δεδομένα κ.λ.π. Για να πραγματοποιήσουμε αλλαγές στο γράφημά μας πρέπει αρχικά να το επιλέξουμε και στην συνέχεια από την λίστα επιλογών που εμφανίζεται με *δεξί κλικ* (ή από το μενού των *'Εργαλείων γραφήματος')* να εφαρμόσουμε τις επιθυμητές αλλαγές στο γράφημα. Είναι σημαντικό να γνωρίζει κανείς ότι τα διαγράμματα που είναι κατάλληλα για την αναπαράσταση ποιοτικών μεταβλητών δεν είναι πάντα κατάλληλα να χρησιμοποιηθούν σε ποσοτικές μεταβλητές και το αντίστροφο. Ως εκ τούτου, πρέπει να κατανοεί κανείς ποιο είναι το προσφορότερο είδος διαγράμματος σε συνδυασμό με το είδος των δεδομένων που θέλουμε να αναπαραστήσουμε.

3.1 Γραφήματα Ποιοτικών Μεταβλητών

Όπως έχει αναφερθεί στο πρώτο κεφάλαιο, οι τιμές των ποιοτικών δεδομένων αφορούν κατηγορίες και, ιδιαίτερα για τα ονομαστικά δεδομένα, ο μόνος υπολογισμός που επιτρέπεται, είναι αυτός των συχνοτήτων και των σχετικών συχνοτήτων. Γραφήματα μέσω των οποίων απεικονίζονται οι συχνότητες και οι σχετικές συχνότητες είναι το ραβδόγραμμα και το κυκλικό διάγραμμα

3.1.1 Ραβδόγραμμα

Το ραβδόγραμμα είναι ένα γράφημα, το οποίο αποτελείται από μία στήλη για κάθε κατηγορία της μεταβλητής ή των μεταβλητών που περιγράφει. Οι στήλες είναι κάθετες στον οριζόντιο ή στον κατακόρυφο άξονα και το ύψος τους ισούται με την συχνότητα ή την σχετική συχνότητα κάθε κατηγορίας (τιμής της μεταβλητής).

Για την δημιουργία Ραβδογράμματος στο EXCEL, ακολουθούμε τα παρακάτω βήματα:

Επιλογή δεδομένων → Εισαγωγή → Ραβδόγραμμα → Επιλογή του τύπου του ραβδογράμματος.

Παράδειγμα

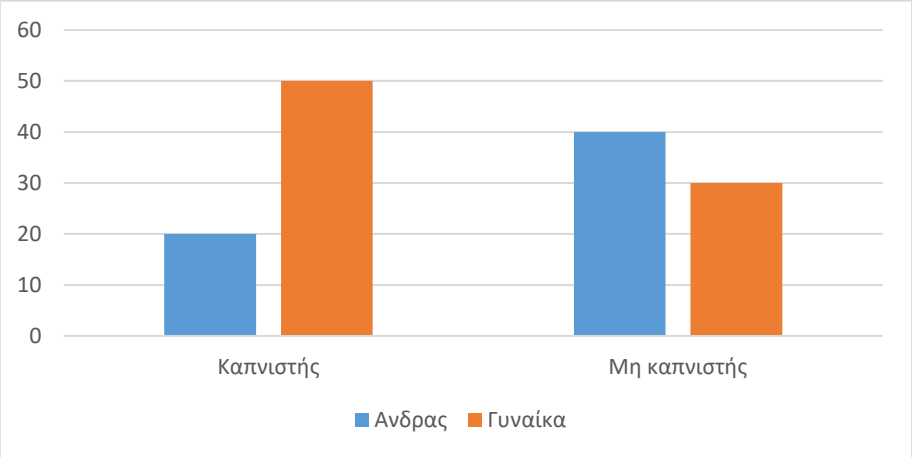
Στον παρακάτω πίνακα, παρουσιάζεται η καπνιστική συνήθεια σε σχέση με το φύλο.

Φύλο	Καπνιστής	Μη καπνιστής
Άνδρας	20	40
Γυναίκα	50	30

3.1Πίνακας Φύλο Καπνιστής

Επιθυμητό είναι να δημιουργηθεί ραβδόγραμμα για τα δεδομένα. Αν ακολουθηθούν τα βήματα που έχουν καταγραφεί προηγουμένως τότε το EXCEL δημιουργεί το ραβδόγραμμα που ακολουθεί και στο οποίο τα ύψη των στηλών του ραβδογράμματος, είναι ίσα με την συχνότητα της αντίστοιχης κατηγορίας. (Για παράδειγμα το ύψος της πρώτης στήλης ισούται με 20, δηλαδή με το πλήθος των ανδρών που είναι καπνιστές.)

Με το ραβδόγραμμα αυτό μπορεί να γίνει εύκολα και η σύγκριση της καπνιστικής συνήθειας μεταξύ ανδρών και γυναικών.



3.2Γράφημα Ραβδόγραμμα

3.1.2 **Κυκλικό διάγραμμα**

Το **κυκλικό διάγραμμα** (ή διάγραμμα πίτας) είναι ένας κυκλικός δίσκος ο οποίος είναι χωρισμένος σε τόσους κυκλικούς τομείς όσες και οι κατηγορίες τιμών της μεταβλητής. Το γωνιακό άνοιγμα κάθε κυκλικού τομέα που αντιπροσωπεύει κάποια τιμή είναι ίσο με το ποσοστό της τιμής αυτής πολλαπλασιασμένο επί τις 360 μοίρες. Για παράδειγμα, αν μια τιμή ποιοτικής μεταβλητής εμφανίζει σχετική συχνότητα 25% τότε ο κυκλικός τομέας με τον οποίο αναπαρίσταται έχει γωνιακό άνοιγμα 90 μοιρών. Το κυκλικό διάγραμμα χρησιμοποιείται για την παρουσίαση της κατανομής των τιμών τόσο σε ποιοτικές μεταβλητές (ονομαστικές και διάταξης) όσο και σε διακριτές μεταβλητές που λαμβάνουν λίγες τιμές (π.χ. το πλήθος των παιδιών μιας οικογένειας).

Για την δημιουργία του κυκλικού γραφήματος, ακολουθούμε τα παρακάτω.

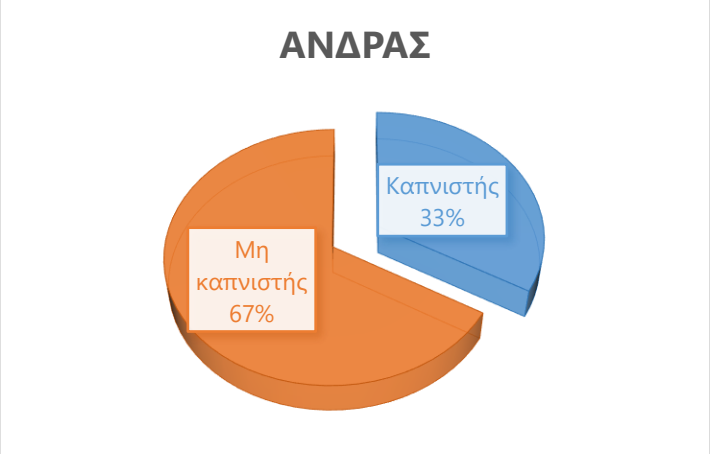
Βήματα:

Επιλογή δεδομένων → Εισαγωγή → Pie chart → *Επιλογή του τύπου του κυκλικού διαγράμματος.*

Παράδειγμα

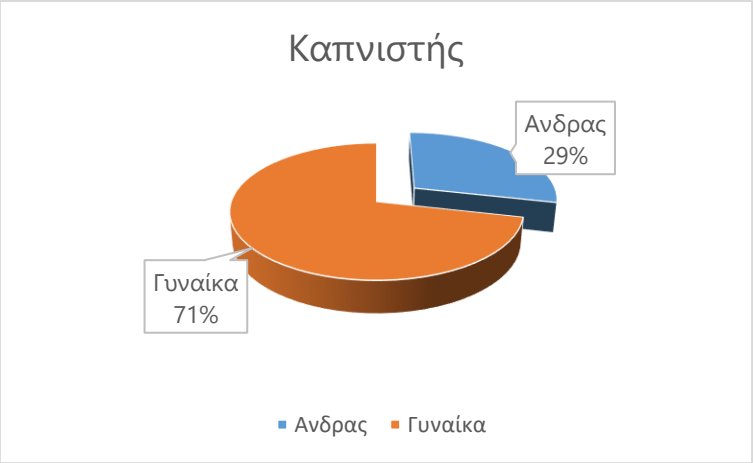
Ας χρησιμοποιήσουμε τα δεδομένα του παραδείγματος που αφορά το κάπνισμα σε σχέση με το φύλο. Η απεικόνιση των δεδομένων του σχετικού πίνακα μπορεί να γίνει και μέσω κυκλικού γραφήματος.

Στο πρώτο γράφημα απεικονίζεται η κατανομή της καπνιστικής συνήθειας των ανδρών. Σε σύνολο 60 ανδρών οι 20 καπνίζουν και οι 40 δεν καπνίζουν. Οπότε το (20/60)*100=33% των ανδρών είναι καπνιστές και το 67% των ανδρών δεν είναι καπνιστές.



3.3Γράφημα Κυκλικό διάγραμμα παράδειγμα 1

Στο δεύτερο κυκλικό διάγραμμα απεικονίζεται η κατανομή του φύλου των καπνιστών. Το 29% των καπνιστών είναι άνδρες ενώ το 71% των καπνιστών είναι γυναίκες.



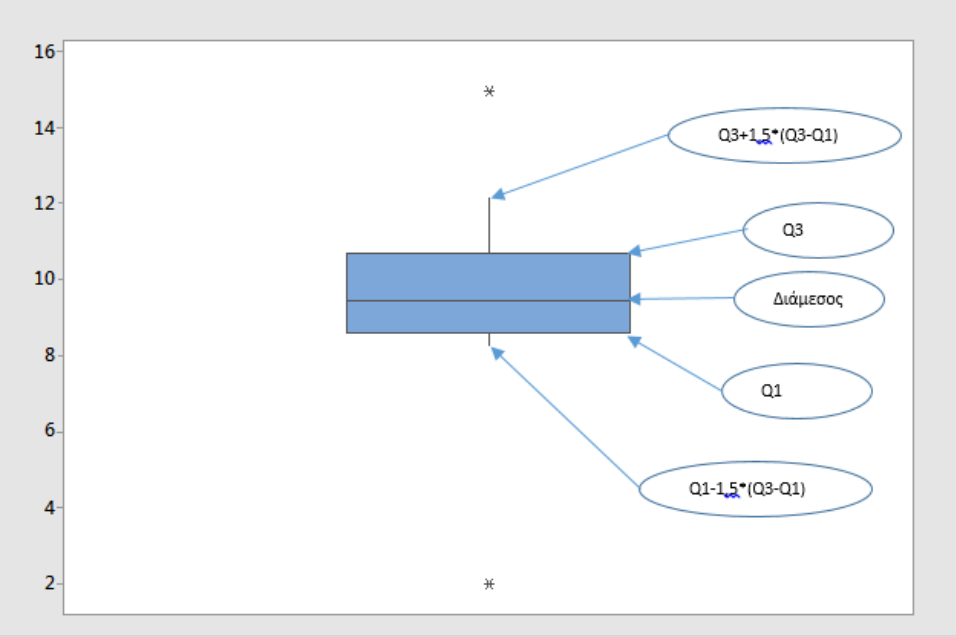
3.4Γράφημα Κυκλικό διάγραμμα παράδειγμα 2

3.2 Γραφήματα Ποσοτικών Μεταβλητών

Η γκάμα από γραφήματα που μπορούν να χρησιμοποιηθούν για την αναπαράσταση ποσοτικών δεδομένων είναι πολυπληθέστερη από εκείνη που αφορούν τα ποιοτικά δεδομένα. Τα γραφήματα των ποσοτικών μεταβλητών που θα παρουσιασθούν σε αυτήν την ενότητα, είναι το θηκόγραμμα (Boxplot) και το ιστόγραμμα.

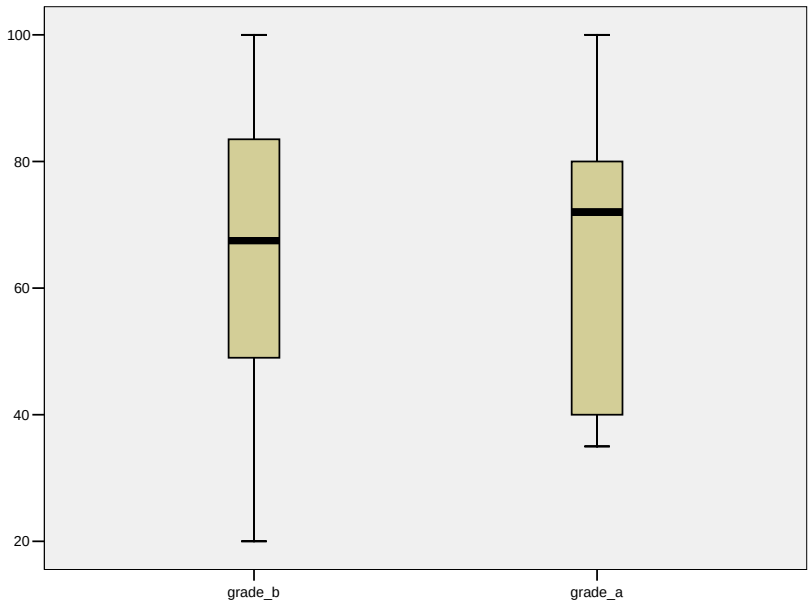
3.2.1 Θηκόγραμμα

Το **Θηκόγραμμα** αποτελείται από 5 βασικά σημεία, το πρώτο τεταρτημόριο (Q1), το τρίτο τεταρτημόριο (Q3), τη διάμεσο, την κάτω οριακή τιμή (Q1-1,5*(Q3-Q1)) και την άνω οριακή τιμή (Q3+1,5*(Q3-Q1)). Η άνω και η κάτω οριακή γραμμή ονομάζονται και άνω και κάτω φράκτης αντίστοιχα. Τιμές μικρότερες από την κάτω οριακή τιμή ή μεγαλύτερες από την πάνω οριακή τιμή θεωρούνται ακραίες και σημειώνονται συνήθως με αστερίσκο (*).



3.5Γράφημα Θηκόγραμμα μιας μεταβλητής

Το Θηκόγραμμα, μέσω των πέντε αυτών σημείων, μας παρέχει τη δυνατότητα της σύγκρισης μεταξύ κατανομών διαφορετικών πληθυσμών. Ως παράδειγμα, θεωρείστε το παρακάτω γράφημα που έχει δημιουργηθεί από τις βαθμολογίες δύο ομάδων σε μια δοκιμασία για την οποία εκπαιδεύτηκαν με δύο διαφορετικές μεθόδους.



3.6*Γράφημα Θηκόγραμμα σύγκριση μεταβλητών*

3.2.2 Ιστόγραμμα

Οι συνεχείς μεταβλητές όπως έχουμε αναφέρει και σε προηγούμενα κεφάλαια μπορούν να πάρουν οποιαδήποτε τιμή μέσα σε ένα διάστημα. Με το **Ιστόγραμμα** απεικονίζονται γραφικά ποσοτικά δεδομένα αφού ομαδοποιηθούν σε υποδιαστήματα που ονομάζονται κλάσεις. Στον παρακάτω πίνακα παρουσιάζεται το σύνολο των δεδομένων που αφορά τη μάζα 160 παιδιών μετρημένη σε κιλά. Τα παιδιά έχουν κατανεμηθεί σε έξι κλάσεις.

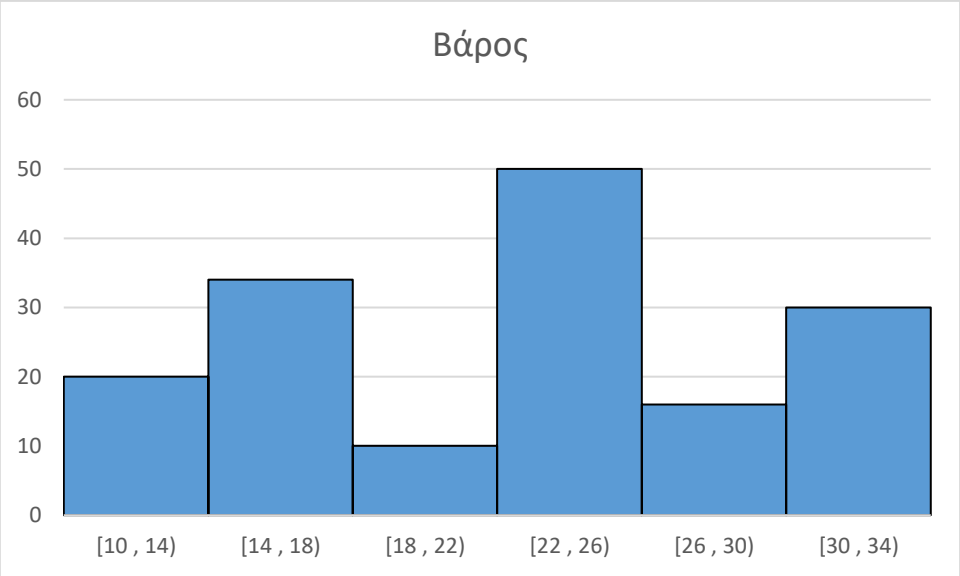
Βάρος (σε κιλά)	**Κέντρο κλάσης**	**Συχνότητα**
[10 , 14)	12	20
[14 , 18)	16	34
[18 , 22)	20	10
[22 , 26)	24	50
[26 , 30)	28	16
[30 , 34)	32	30

3.7Πίνακας Βάρος

Οι ομαδοποιήσεις [10 , 14) , [14 , 18), [18 , 22), [22 , 26),[26 , 30) και [30 , 34) ονομάζονται κλάσεις. Το κέντρο της κλάσης ισούται με το μέσον του διαστήματος που αντιπροσωπεύει η κάθε κλάση. Το εύρος της κλάσης προκύπτει από την διαφορά των άκρων της, στο

συγκεκριμένο παράδειγμα το εύρος όλων των κλάσεων ισούται με 4, όσο και η διαφορά των κέντρων δύο διαδοχικών κλάσεων.

Η συχνότητα κάθε κλάσης συμπίπτει με το πλήθος των παιδιών που έχουν βάρος μεγαλύτερο ή ίσο από το κάτω όριό της και μικρότερο από το άνω όριό της. Για παράδειγμα στην πρώτη κλάση, 20 παιδιά έχουν βάρος μεγαλύτερο ή ίσο των 10 κιλών και μικρότερο των 14 κιλών. Η γραφική παρουσίαση των στοιχείων του παραπάνω πίνακα γίνεται μέσω ιστογράμματος. Το ύψος της στήλης του ιστογράμματος, ισούται με την συχνότητα της αντίστοιχης κλάσης



3.8*Γράφημα Ιστόγραμμα*

Παράδειγμα

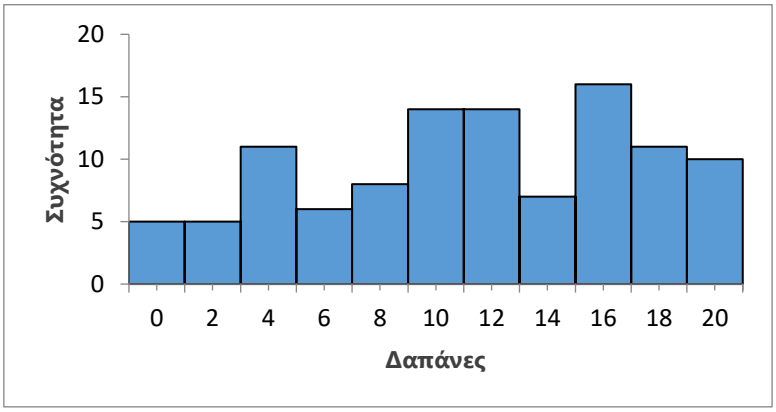
Στο φύλλο *Ιστόγραμμα* του αρχείου *Δεδομένα*, η στήλη Α περιέχει τις δαπάνες σε εκατοντάδες ευρώ 107 οικογενειών για διακοπές. Η μικρότερη δαπάνη είναι 0 και η μεγαλύτερη 2000 ευρώ. Για την δημιουργία ιστογράμματος ακολουθούμε τα παρακάτω βήματα:

Δεδομένα→Ανάλυση δεδομένων→ Ιστόγραμμα

Στην περιοχή εισόδου επιλέγουμε τα δεδομένα μας, στην συνέχεια επιλέγουμε το *Έξοδος γραφήματος*. Το EXCEL ομαδοποιεί τα δεδομένα σε κλάσεις, και εμφανίζει έναν πίνακα με τα κέντρα των κλάσεων. Έτσι προκύπτουν ο παρακάτω πίνακας, ο οποίος περιέχει στην πρώτη στήλη, τα κέντρα των κλάσεων και στην δεύτερη στήλη την συχνότητα της αντίστοιχης κλάσης. Ακόμη το EXCEL εμφανίζει και το αντίστοιχο ιστόγραμμα.

Κέντρο κλάσης	*Συχνότητα*
0	5
2	5
4	11
6	6
8	8
10	14
12	14
14	7
16	16
18	11
20	10

3.9Πίνακας Κέντρα κλάσεων ιστογράμματος



3.10Γράφημα Ιστόγραμμα Δαπάνες

3.2.3 Διάγραμμα αράχνης

Όταν τα δεδομένα παρουσιάζονται σε γραμμές και στήλες όπως στον παρακάτω πίνακα, ένα γράφημα για την παρουσίαση και την σύγκριση των τιμών του πίνακα είναι το αραχνοειδές.

Παράδειγμα: Ο παρακάτω πίνακας εμφανίζει το μέσο όρο ημερών άδειας για τρεις υπηρεσίες (ΥΑ,ΥΒ,ΥΓ) και τρία έτη (2016, 2017, 2018). Ζητείται η γραφική παράσταση των δεδομένων μέσω διαγράμματος αράχνης.

Μέσος όρος ημερών άδειας υπαλλήλου από την υπηρεσία			
Έτη	ΥΑ	ΥΒ	ΥΓ
2016	20	30	24
2017	25	30	32
2018	14	10	21

3.11Πίνακας Μέσος όρος ημερών άδειας

Για την δημιουργήσουμε το διάγραμμα αράχνης για τον παραπάνω πίνακα, ακολουθούμε τα παρακάτω βήματα:

Επιλογή δεδομένων → Εισαγωγή → Γράφημα Αράχνης→ *Επιλογή του τύπου του γραφήματος.*

3.12*Γράφημα αράχνης Μέσος όρος ημερών άδειας*

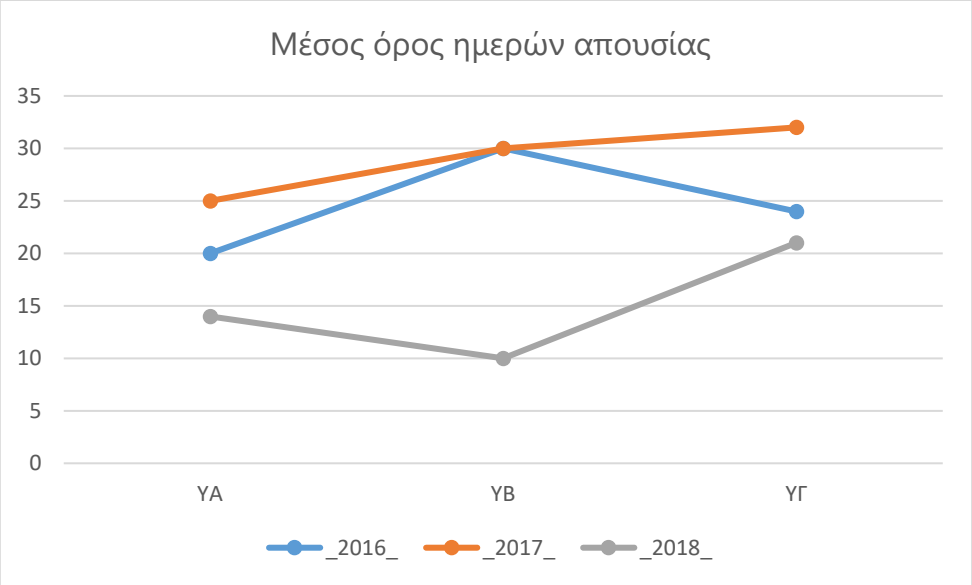
Οι κορυφές των τριγώνων, στο διάγραμμα αράχνης αποτυπώνουν τον μέσο αριθμό ημερών άδειας για κάθε υπηρεσία, σε καθένα από τα έτη 2016 έως και 2018. Παρατηρούμε ότι σημειώθηκε μείωση του μέσου αριθμού των ημερών άδειας και στις τρείς υπηρεσίες το 2018 (εσωτερικό γκρι τρίγωνο). Παρατηρούμε ακόμη, ότι στην υπηρεσία ΥΒ τα δύο πρώτα έτη διατηρήθηκε σταθερός ο μέσος αριθμός ημερών άδειας, ενώ στην συνέχεια μειώθηκε περισσότερο από τις άλλες υπηρεσίες. Για τις υπηρεσίες ΥΑ και ΥΓ, ο μέσος όρος των ημερών άδειας, αυξήθηκε το 2017 σε σχέση με το 2016 και στην συνέχεια το 2018 μειώθηκε.

3.2.4 Γράφημα γραμμής

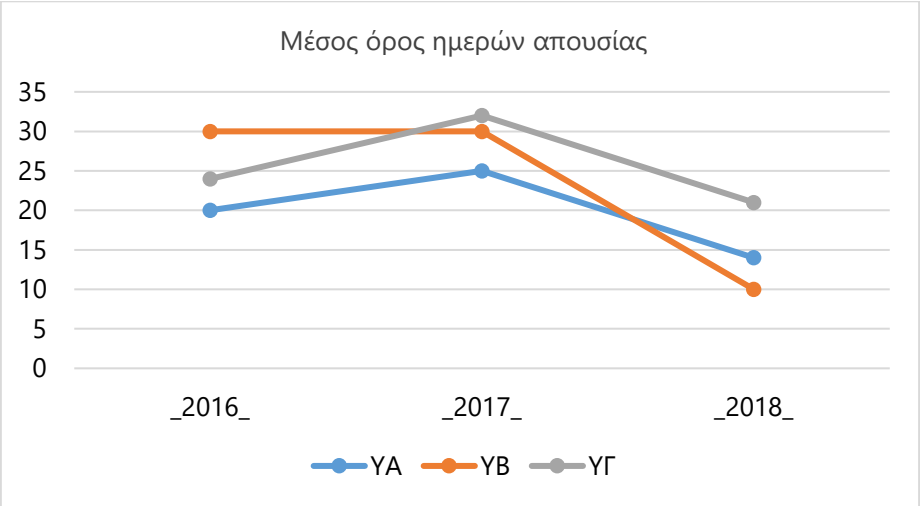
Με το γράφημα γραμμής μπορούμε να αποτυπώσουμε την εξέλιξη ενός χαρακτηριστικού μέσα στο χρόνο, καθώς επίσης να συγκρίνουμε την εξέλιξη του σε διαφορετικές ομάδες. Παράλληλα, με διάγραμμα γραμμής θα μπορούσαν να αναπαρασταθούν και τα δεδομένα του παραδείγματος των τριών υπηρεσιών που χρησιμοποιήθηκαν για το διάγραμμα αράχνης.

Για την δημιουργία του εν λόγω Γραφήματος Γραμμής ακολουθούμε τα παρακάτω βήματα ***Επιλογή δεδομένων →Εισαγωγή→Γράφημα Γραμμής→*** *Επιλογή του τύπου του γραφήματος.*

Στα δύο παρακάτω γραφήματα απεικονίζεται ο μέσος όρος των ημερών απουσίας από την υπηρεσία, ανά υπηρεσία και ανά έτος.



3.13Γράφημα Γραμμής Μέσος όρος ημερών απουσίας ανά υπηρεσία



3.14Γράφημα γραμμής Μέσος όρος ημερών απουσίας ανά έτος

3.2.5 Γράφημα διασποράς

Για να αποκτήσουμε μια πρώτη ιδέα για το αν και με ποιο τρόπο, δυο ποσοτικές μεταβλητές συσχετίζονται, μπορούμε να κατασκευάσουμε ένα διάγραμμα διασποράς. Στον οριζόντιο άξονα του γραφήματος διασποράς σημειώνουμε τις τιμές της μίας μεταβλητής και στον κατακόρυφο τις τιμές της άλλης μεταβλητής.

Παράδειγμα

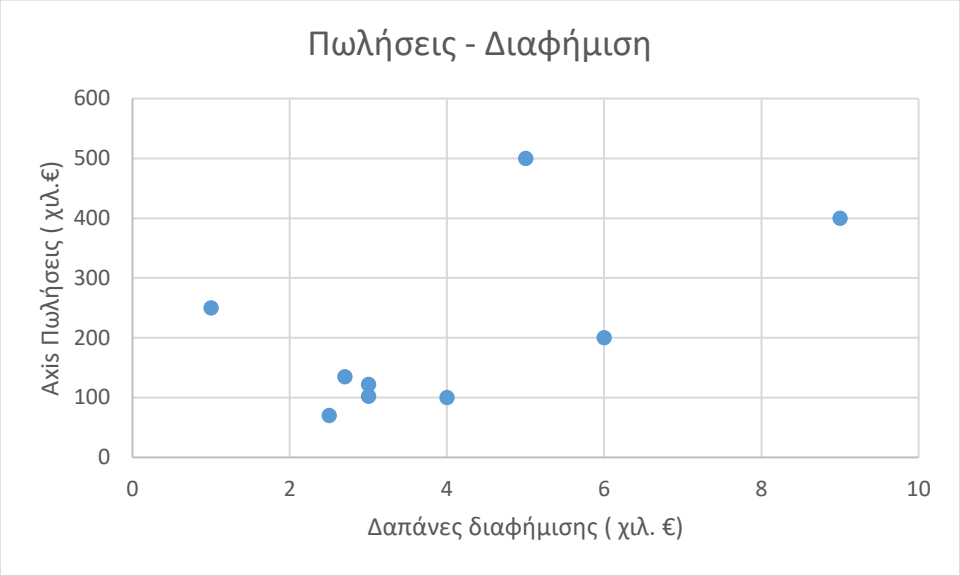
Στον παρακάτω πίνακα καταγράφονται οι δαπάνες για διαφήμιση μιας εταιρείας και οι αντίστοιχες πωλήσεις της εταιρείας σε χιλιάδες ευρώ, κατά τα έτη 2010 έως και 2018.

Έτος	Δαπάνες διαφήμισης (σε χιλ. €)	Πωλήσεις (σε χιλ.€)
2010	1	250
2011	3	122
2012	2,7	135
2013	6	200
2014	9	400
2015	5	500
2016	4	100
2017	3	102
2018	2,5	70

3.15Πίνακας Δαπάνες διαφήμισης Πωλήσεις

Για την δημιουργία ενός Γραφήματος Διασποράς ακολουθούμε τα παρακάτω βήματα ***Επιλογή δεδομένων →Εισαγωγή→Γράφημα Διασποράς→*** *Επιλογή του τύπου του γραφήματος.*

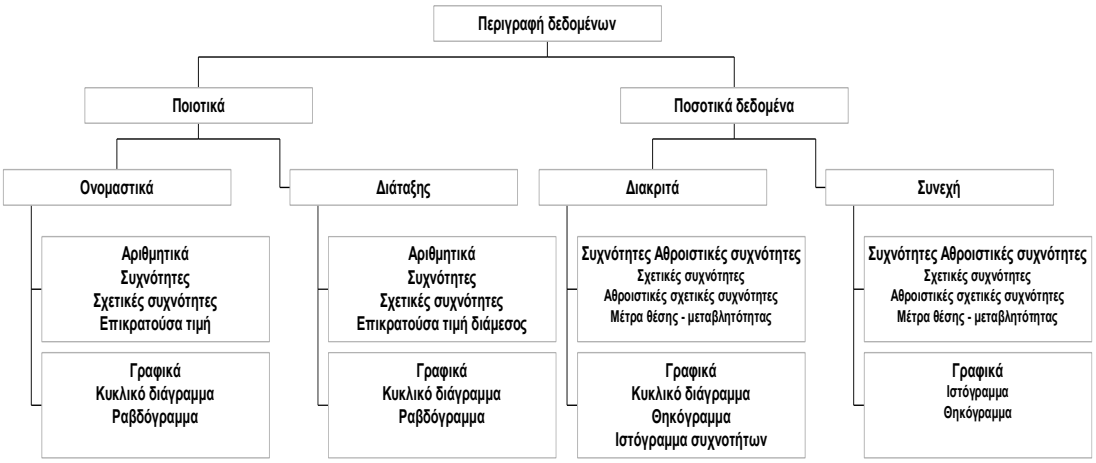
Στο παρακάτω γράφημα διασποράς παρατηρείται μία ανοδική *τάση* των πωλήσεων σε σχέση με τις διαφημιστικές δαπάνες.



3.16 *Γράφημα διασποράς Διαφήμιση Πωλήσεις*

Συνοψίζοντας

Η περιγραφή των δεδομένων μπορεί να γίνει γραφικά ή αριθμητικά. Το πιο σημαντικό βήμα για την επιλογή του κατάλληλου μέτρου και του κατάλληλου γραφήματος, για την περιγραφή και την απεικόνιση των δεδομένων, είναι η γνώση του είδους των δεδομένων. Στον παρακάτω πίνακα απεικονίζονται το είδος των δεδομένων και τα αντίστοιχα μέτρα καθώς επίσης και τα κατάλληλα γραφήματα, για την παρουσίαση των δεδομένων.



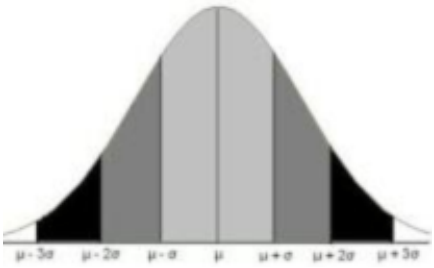
3.17 *Διάγραμμα Μέτρα κα γραφήματα περιγραφής δεδομένων*

3.3 Κατανομή

Στο πρώτο κεφάλαιο αναφερθήκαμε στην έννοια της κατανομής και είδαμε ένα ενδεικτικό παράδειγμα με το άθροισμα των ενδείξεων των δύο ζαριών. Όπως αναφέρθηκε στο 1ο κεφάλαιο η **Κατανομή** μιας μεταβλητής μας δείχνει όλες τις πιθανές τιμές (ή διαστήματα) των τιμών της μεταβλητής και πόσο συχνά αυτές εμφανίζονται. Η Κατανομή μιας ποιοτικής μεταβλητής, μας δείχνει το ποσοστό εμφάνισης των τιμών της.

Για τις *συνεχείς μεταβλητές* δεν είναι δυνατή η εύρεση της πιθανότητας κάθε τιμής. Έτσι, η κατανομή μιας συνεχούς μεταβλητής αναφέρεται στην πιθανότητα εμφάνισης διαστημάτων τιμών της μεταβλητής.

Μία πολύ σημαντική Κατανομή για την επαγωγική στατιστική και τη στατιστική συμπερασματολογία είναι η **Κανονική Κατανομή**. Είναι γνωστή και ως κατανομή του Gauss. Αξίζει να σημειώσουμε ότι η Κανονική Κατανομή υποκρύπτεται σε πολλά φυσικά φαινόμενα



3.18Γράφημα Κανονική κατανομή

και ότι αξιοποιείται για την προσέγγιση πολλών άλλων κατανομών.

Στο παραπάνω σχήμα παρουσιάζεται η καμπύλη της Κανονικής Κατανομής με *Μέσο* μ και *Τυπική Απόκλιση* σ.

Για δεδομένα που ακολουθούν την Κανονική Κατανομή με μέσο μ και τυπική απόκλιση σ ισχύουν τα παρακάτω:

- Στο διάστημα $(\mu - \sigma\,, \mu + \sigma\,)$ περιλαμβάνεται το 68% των παρατηρήσεων
- Στο διάστημα $(\mu - 2\sigma\,, \mu + 2\sigma\,)$ περιλαμβάνεται το 95% των παρατηρήσεων
- Στο διάστημα $(\mu - 3\sigma\,, \mu + 3\sigma\,)$ περιλαμβάνεται το 99,7% των παρατηρήσεων
- Η Κανονική Κατανομή είναι συμμετρική ως προς το μέσο.
- Το εμβαδόν που περικλείεται από την καμπύλη της κανονικής κατανομής, τον οριζόντιο άξονα και τις ευθείες για παράδειγμα, με χ=μ-2σ και χ=μ+2σ εκφράζει το ποσοστό των παρατηρήσεων που ανήκουν στο διάστημα $(\mu - 2\sigma\,, \mu + 2\sigma\,)$
- Ο Μέσος και η Διάμεσος της Κανονικής Κατανομής έχουν ίσες τιμές.

Η κανονική κατανομή με μέσο 0 και τυπική απόκλιση 1 ονομάζεται *Τυποποιημένη κανονική κατανομή*. Αν μία μεταβλητή ακολουθεί κανονική κατανομή με μέσο μ και τυπική απόκλιση σ, τότε οι τυποποιημένες τιμές της, ακολουθούν την τυποποιημένη κανονική κατανομή.

3.4 Μέτρα σχηματικής μορφής

Πολλές φορές μας ενδιαφέρει εκτός από μέτρα θέσης και διασποράς να μελετήσουμε μέσω στατιστικών συναρτήσεων και τη μορφή, -το σχήμα – μιας κατανομής. Μας ενδιαφέρει δηλαδή να μελετήσουμε την ασυμμετρία και την κυρτότητα της κατανομής. Για το λόγο αυτό χρησιμοποιούμε τους **συντελεστές ασυμμετρίας**. και **κύρτωσης**

3.4.1 Συντελεστής ασυμμετρίας

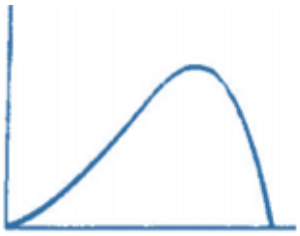
Μια Κατανομή θα λέγεται *Συμμετρική* αν οι τιμές κατανέμονται συμμετρικά γύρω από τον μέσο. Σε αυτήν την περίπτωση ο μέσος και η διάμεσος ταυτίζονται. Στην περίπτωση που έχουμε ασυμμετρία ο μέσος επηρεάζεται από τις ακραίες τιμές σε αντίθεση με την διάμεσο με αποτέλεσμα να μην είναι ίσες οι τιμές τους.

Για την μέτρηση της ασυμμετρίας μιας Κατανομής έχει οριστεί ο **συντελεστής ασυμμετρίας**. Ο συντελεστής ασυμμετρίας είναι:

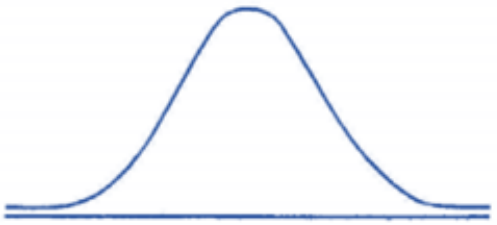
- Ίσος με 0 αν οι τιμές της μεταβλητής κατανέμονται συμμετρικά γύρω από το μέσο
- Θετικός αν έχουμε δεξιά ασυμμετρία, δηλαδή αν οι ακραίες τιμές βρίσκονται προς τα δεξιά.
- Αρνητικός αν έχουμε αριστερή ασυμμετρία δηλαδή αν οι ακραίες τιμές βρίσκονται προς τα αριστερά του κύριου όγκου των δεδομένων

Στο παρακάτω γράφημα απεικονίζονται δεδομένα με θετική ασυμμετρία, στο δεύτερο γράφημα δεδομένα με αρνητική ασυμμετρία, και στο τρίτο δεδομένα κατανεμημένα συμμετρικά γύρω από τον μέσο.

3.19Γράφημα Δεξιά ασυμμετρία

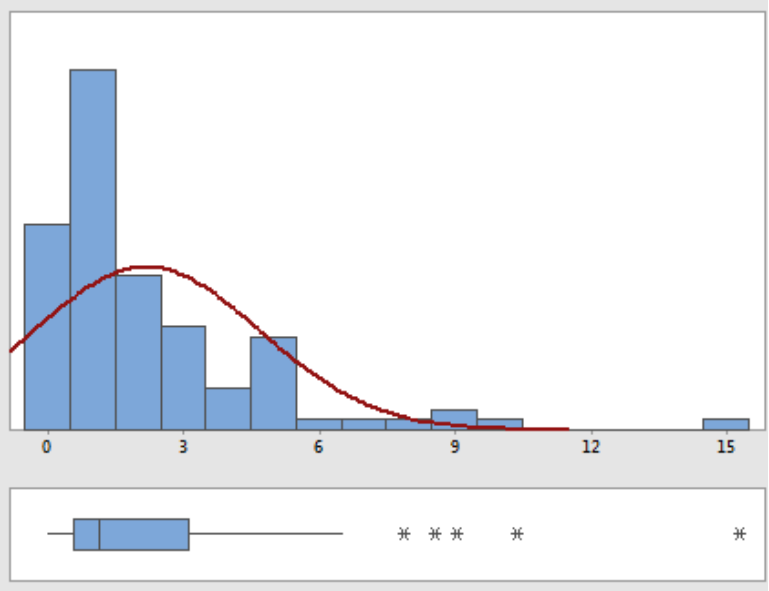


3.20*Γράφημα Αριστερή ασυμμετρία*



3.21Γράφημα Συμμετρική κατανομή

Στο παρακάτω γράφημα, εμφανίζεται το ιστόγραμμα και το αντίστοιχο θηκόγραμμα για δεδομένα με δεξιά ασυμμετρία.



3.22Γράφημα Ιστόγραμμα- Θηκόγραμμα

Στην περίπτωση της συμμετρίας η γραμμή της διαμέσου είναι στο κέντρο του ορθογωνίου του θηκογράμματος.

Μία σημαντική παρατήρηση για την σχέση μέσου διαμέσου και ασυμμετρίας μιας κατανομής, είναι ότι αν παρουσιάζεται:

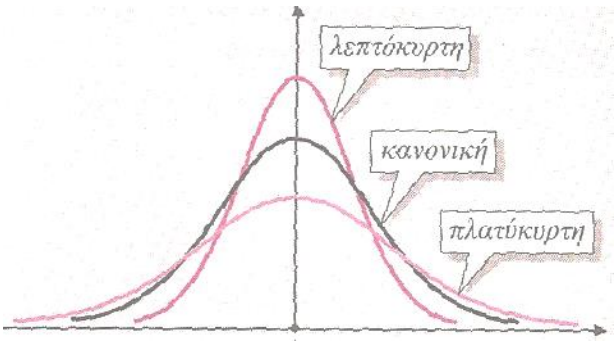
- Αριστερή ασυμμετρία η διάμεσος είναι μεγαλύτερη από τον μέσο,
- Δεξιά ασυμμετρία η διάμεσος είναι μικρότερη του μέσου

- Συμμετρία, η διάμεσος ισούται με τον μέσο. Αυτό οφείλεται στο γεγονός που συζητήθηκε στο δεύτερο κεφάλαιο σχετικά με τον επηρεασμό του μέσου από τις ακραίες τιμές.

3.4.2 Συντελεστής κύρτωσης

Για την μελέτη του σχήματος της κατανομής των τιμών μιας μεταβλητής μας ενδιαφέρει ακόμα και ο **Βαθμός Κύρτωσης** δηλαδή κατά πόσο απότομη είναι καμπύλη. Κατά πόσο δηλαδή συγκεντρώνονται οι τιμές γύρω από το μέσο.

Αν ο συντελεστής κύρτωσης είναι ίσος με μηδέν τότε η καμπύλη είναι μεσόκυρτη, αν είναι μεγαλύτερος του μηδενός τότε ονομάζεται λεπτόκυρτη ενώ αν είναι μικρότερος του μηδενός τότε ονομάζεται πλατύκυρτη.



3.23*Γράφημα Κύρτωση*

Υπολογισμός με το EXCEL

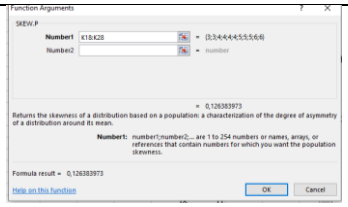
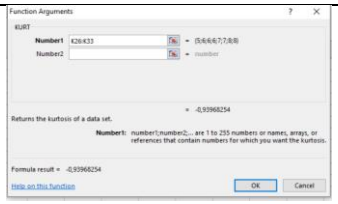
1ος τρόπος

Μενού Δεδομένα →Ανάλυση Δεδομένων → Περιγραφικά στατιστικά στοιχεία → Περιληπτικά μέτρα

Στα περιληπτικά μέτρα εμφανίζονται ο συντελεστής ασυμμετρίας και ο συντελεστής κύρτωσης

2ος τρόπος

Με την χρήση συναρτήσεων του EXCEL

Μέτρο	Συνάρτηση	
Συντελεστής ασυμμετρίας	=SKEW.P(K18:K28)	
Συντελεστής κύρτωσης	=KURT(K26:K33)	

3.24Πίνακας συναρτήσεων υπολογισμού μέτρων σχηματικής μορφής

Άσκηση του 3ου κεφαλαίου

Χρησιμοποιώντας τα δεδομένα της άσκησης του 2ου κεφαλαίου δημιουργείστε:

1. Διάγραμμα γραμμής για την Ελλάδα και τη Γερμανία στο διάστημα ετών από 2008 έως και 2016 αναφορικά με το βαθμό απασχόλησης στις δύο χώρες.
2. Με τις τιμές όλων των κρατών-μελών της ΕΕ για τα έτη 2008 και 2016 δημιουργείστε θηκογράμματα και συγκρίνετε τις δύο κατανομές. Εξετάστε αν εμφανίζονται ακραίες τιμές και αν οι εν λόγω κατανομές θα μπορούσαν να χαρακτηριστούν ως συμμετρικές.
3. Υπολογίστε τα μέτρα ασυμμετρίας και κύρτωσης για τις ετήσιες κατανομές των τιμών της απασχόλησης στην ΕΕ , αναφορικά με τα έτη 2010, 2012 και 2014. Πως χαρακτηρίζετε τις κατανομές αυτές; Σημειώνεται κάποια σημαντική διαφοροποίηση; Συζητήστε την άποψη σας.

ΚΕΦΑΛΑΙΟ 4ο

ΕΠΑΓΩΓΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

Ο κλάδος της Επαγωγικής Στατιστικής αξιοποιεί τα δεδομένα που διατίθενται από ένα δείγμα πληθυσμού για να κάνει εκτιμήσεις και να βγάλει συμπεράσματα αναφορικά με τον πληθυσμό. Είναι ιδιαίτερα σημαντικός κλάδος και βρίσκει εφαρμογή σε πολλές επιστήμες όπως η Ιατρική, η Οικονομία, η Διοίκηση κλπ. Ίσως, το πιο γνωστό παράδειγμα αξιοποίησης των τεχνικών αυτών από τον κρατικό μηχανισμό είναι του σκάνδαλο της VW.
Στο κεφάλαιο αυτό θα ασχοληθούμε μερικά βασικά θέματα της Επαγωγικής Στατιστικής που έχουν να κάνουν με την εκτίμηση παραμέτρων και τον έλεγχο υποθέσεων.

4.1 Εκτίμηση παραμέτρων

Η εκτίμηση παραμέτρου είναι η διαδικασία μέσω της οποίας προσπαθούμε να εκτιμήσουμε την τιμή μιας παραμέτρου ενός πληθυσμού μέσω ενός δείγματός του. Οι πιο συνηθισμένες περιπτώσεις εκτίμησης παραμέτρου αφορούν τον μέσο όρο, την τυπική απόκλιση μιας ποσοτικής μεταβλητής ή το ποσοστό εμφάνισης κάποιας τιμής μεταβλητής στον πληθυσμό. Η αξιοποίηση του μεθοδολογικού πλαισίου των εκτιμήσεων κρίνονται αναγκαία για την καλύτερη λειτουργία Κρατών, Οργανισμών, Εταιρειών, ακόμη και στην καθημερινότητα ενός ατόμου. Για παράδειγμα έχουμε εκτίμηση του ποσοστού ανεργίας, της μείωσης θερμοκρασίας ασθενών μετά την λήψη κάποιου φαρμάκου, του χρόνου αναμονής σε ένα ΚΕΠ ή σε μία Τράπεζα. Τα δύο είδη στατιστικών εκτιμήσεων είναι η εκτίμηση *σημείου* και η εκτίμηση *διαστήματος*.

4.1.1 Σημειακή εκτίμηση

Στην εκτίμηση σημείου υπολογίζεται μία μόνο τιμή από το δείγμα ως εκτιμώμενη τιμή μιας παραμέτρου του πληθυσμού.

Αν για παράδειγμα θα θέλαμε να εκτιμήσουμε την μέση βαθμολογία στο μάθημα της Στατιστικής στο Μαθηματικό τμήμα Αθηνών, θα μπορούσαμε να χρησιμοποιήσουμε ως εκτίμηση της μέσης τιμής του πληθυσμού, τον μέσο όρο των βαθμολογιών των φοιτητών ενός δείγματος. Έστω ότι ένα τυχαίο δείγμα φοιτητών είχε τις ακόλουθες βαθμολογίες

8,4,7,10,5,5,7,9,10,7, τότε η εκτίμηση για την μέση τιμή του πληθυσμού από αυτό το συγκεκριμένο δείγμα θα ήταν (8+4+7+10+5+5+7+9+10+7)/10=7,2.
Αν επιλέγαμε ένα άλλο δείγμα από τον ίδιο πληθυσμό των φοιτητών και υπολογίζαμε τον μέσο όρο του νέου δείγματος, όπως καταλαβαίνουμε το αποτέλεσμα δεν είναι απαραίτητο να ήταν 7,2. Έτσι αν υπολογίζαμε τους μέσους όλων των δυνατών δειγμάτων που μπορούν να προκύψουν από τον πληθυσμό, θα παίρναμε πολλές διαφορετικές τιμές.
Στο προηγούμενο παράδειγμα χρησιμοποιήσαμε ως τύπο για την εκτίμηση της μέσης τιμής τον μέσο όρο του δείγματος, που είναι ο ίδιος τύπος με αυτόν της μέσης τιμής του πληθυσμού, κάτι το οποίο δεν είναι απαραίτητο να συμβαίνει πάντα στην εκτίμηση των παραμέτρων. Θέματα σχετικά με τις ιδιότητες των εκτιμητριών των παραμέτρων και την επιλογή της καλλίτερης εκτιμήτριας είναι εκτός των στόχων του μαθήματος.

4.1.2 Διάστημα εμπιστοσύνης

Στην προηγούμενη παράγραφο αναφερθήκαμε στην σημειακή εκτίμηση μιας παραμέτρου. Είναι φανερό ότι η σημειακή αυτή εκτίμηση ενδεχομένως θα διαφοροποιηθεί αν χρησιμοποιηθεί κάποιο διαφορετικό δείγμα, το οποίο θα ληφθεί από τον ίδιο πληθυσμό.
Έτσι θα ήταν προτιμότερο να κατασκευάζαμε ένα διάστημα, το οποίο θα περιείχε την πραγματική τιμή της παραμέτρου. Τα διαστήματα αυτά ονομάζονται **διαστήματα εμπιστοσύνης**. Σημαντική παράμετρος στον προσδιορισμό των διαστημάτων εμπιστοσύνης είναι και ο προσδιορισμός του **βαθμού εμπιστοσύνης** στο διάστημα αυτό, δηλαδή την πιθανότητα το διάστημα αυτό να περιέχει την πραγματική τιμή της παραμέτρου που μας ενδιαφέρει.

Όπως αναφέρθηκε στην προηγούμενη παράγραφο, αν υπολογίζαμε τους μέσους όλων των δυνατών δειγμάτων που μπορούν να προκύψουν από έναν πληθυσμό, θα παίρναμε πολλές διαφορετικές τιμές. Γίνεται κατανοητό ότι ο μέσος όρος του δείγματος είναι μία μεταβλητή με τιμές τις διαφορετικές τιμές των μέσων των δειγμάτων. Συνεπώς μπορούμε να μιλήσουμε για την κατανομή των τιμών της μεταβλητής αυτής. Η μεταβλητή αυτή ονομάζεται **εκτιμήτρια.**

Αν είναι γνωστή η κατανομή της εκτιμήτριας, τότε είναι εφικτό να προσδιορίσουμε ένα διάστημα το οποίο με συγκεκριμένη πιθανότητα, θα περιλαμβάνει την πραγματική τιμή της παραμέτρου. Για τις ανάγκες του μαθήματος, θα παρουσιαστεί η εύρεση των διαστημάτων εμπιστοσύνης για την μέση τιμή μ ενός πληθυσμού.

Για έναν πληθυσμό με *μέσο μ* και *τυπική απόκλιση σ* ο *δειγματικός μέσος* είναι μια *εκτιμήτρια του μ*, για την οποία ισχύει ότι *η μέση τιμή της* (δηλαδή ο μέσος των μέσων όλων των δειγμάτων) είναι *ίση με την μέση τιμή του πληθυσμού μ*. Αν μάλιστα η μεταβλητή στον πληθυσμό ακολουθεί την κανονική κατανομή (ή το δείγμα είναι μεγάλο) τότε ο δειγματικός μέσος ακολουθεί την κανονική κατανομή με μέσο μ και τυπική απόκλιση σ/√n , όπου n το μέγεθος του δείγματος και σ η τυπική απόκλιση του πληθυσμού.

Έτσι μπορούμε να προσδιορίσουμε δύο τιμές L,U τέτοιες ώστε με πιθανότητα (1-α)*100% το διάστημα (L ,U) να περιέχει την μέση τιμή του πληθυσμού. Άρα η πιθανότητα να μην περιέχει το διάστημα (L ,U) την μέση τιμή του πληθυσμού είναι α.

Το *α* είναι *η πιθανότητα* για το συγκεκριμένο δείγμα να μην προκύψει διάστημα που να περιέχει την μέση τιμή μ του πληθυσμού. Με αυτόν τον τρόπο προσδιορίσαμε ένα διάστημα, το οποίο με πιθανότητα (1-α)100% θα περιέχει την μέση τιμή του πληθυσμού. Ακόμη ισχύει ότι το ποσοστό των δειγμάτων των οποίων ο μέσος θα περιέχεται στο παραπάνω διάστημα είναι (1-α)100% Το 1-α ονομάζεται **επίπεδο εμπιστοσύνης**.

Υπολογισμός διαστήματος εμπιστοσύνης με το EXCEL

Για τον υπολογισμό διαστήματος εμπιστοσύνης τα βήματα που ακολουθούμε είναι τα εξής:

Δεδομένα→Ανάλυση δεδομένων→Περιγραφικά στατιστικά,

- Στην περιοχή εισόδου εισάγουμε τα κελιά με τα δεδομένα .
- Επιλογή των Περιληπτικών μέτρων
- Επιλογή του επιπέδου εμπιστοσύνης

Παράδειγμα

Παρακάτω θα εξετάσουμε ένα παράδειγμα για την εκτίμηση του μέσου χρόνου αναμονής στο ταμείο μιας τράπεζας. Λαμβάνουμε δείγμα 90 πελατών και σημειώνουμε τον χρόνο αναμονής στο ταμείο. Θα κατασκευάσουμε ένα 95% διάστημα εμπιστοσύνης για τον μέσο χρόνο αναμονής στο ταμείο της τράπεζας.

Ακολουθώντας τα παραπάνω βήματα σημειώνουμε τον μέσο χρόνο αναμονής στο δείγμα, που είναι 31,66 λεπτά και την ένδειξη του πεδίου για το διάστημα εμπιστοσύνης που είναι 3,64. Στην συνέχεια υπολογίζουμε τα όρια του διαστήματος εμπιστοσύνης όπως φαίνεται παρακάτω.

Κάτω όριο	=31,65-3,64
Άνω όριο	=31,65+3,64

Οπότε προκύπτουν τα όρια

Κάτω όριο	28,01
Άνω όριο	35,29

Οπότε ο μέσος χρόνος αναμονής στο ταμείο της τράπεζας, περιέχεται στο διάστημα (28,01 , 35,29), με εμπιστοσύνη 95%. Αυτό σημαίνει ότι με 95% πιθανότητα αναμένουμε το διάστημα (28,01 , 35,29), να περιέχει την μέση τιμή του χρόνου αναμονής στο ταμείο της τράπεζας. Ακόμη θα μπορούσαμε να πούμε ότι αν πάρουμε 100 δείγματα, αναμένουμε στα 95 δείγματα, ο μέσος όρος του χρόνου αναμονής να περιέχεται στο διάστημα (28,01 , 35,29).

4.2 Έλεγχος Υποθέσεων

Στις προηγούμενες παραγράφους του κεφαλαίου ασχοληθήκαμε με την σημειακή εκτίμηση παραμέτρων και την δημιουργία διαστημάτων εμπιστοσύνης. Σε αυτήν την ενότητα θα ασχοληθούμε με τον έλεγχο υποθέσεων. Δηλαδή θα προσπαθήσουμε να βρούμε τρόπους για να αποφασίσουμε κατά πόσο μια τιμή που υποθέτουμε ότι έχει μία παράμετρος του πληθυσμού, είναι αποδεκτή με βάση τις παρατηρήσεις ενός δείγματος.

Για παράδειγμα η υπόθεση που κάνει κάποιος ότι η μέση παραγωγή είναι 1000 μονάδες προϊόντος ανά εβδομάδα ή ότι η μέση κατανάλωση βενζίνης για δύο μάρκες αυτοκινήτων είναι η ίδια.

Στον έλεγχο υποθέσεων δημιουργούνται δύο υποθέσεις. Στο παράδειγμα που αναφέρθηκε παραπάνω που αφορά στην μέση εβδομαδιαία παραγωγή προϊόντος, η μία υπόθεση είναι ότι *η μέση εβδομαδιαία παραγωγή προϊόντος είναι 1000 μονάδες* και η άλλη ότι η *μέση εβδομαδιαία παραγωγή προϊόντος δεν είναι 1000 μονάδες*. Στο παράδειγμα με την κατανάλωση βενζίνης η μία υπόθεση είναι ότι *η μέση κατανάλωση βενζίνης για δύο μάρκες αυτοκινήτων* είναι η ίδια και η άλλη ότι *η μέση κατανάλωση βενζίνης για τις δύο μάρκες αυτοκινήτων δεν είναι η ίδια*. Η μία από τις δύο υποθέσεις ονομάζεται μηδενική και συμβολίζεται με H_0 και η υπόθεση ότι η H_0 είναι λανθασμένη ονομάζεται εναλλακτική υπόθεση και συμβολίζεται με H_1.

Ο έλεγχος αυτής της υπόθεσης θα γίνει με βάση ένα δείγμα οπότε τα αποτελέσματα που θα προκύψουν θα έχουν μια πιθανότητα λάθους. Τα λάθη που μπορεί να κάνει ο ερευνητής στην απόφαση του- αν θα απορρίψει ή όχι την H_0 -είναι να απορρίψει την H_0 ενώ η H_0 είναι σωστή ή να μην απορρίψει την H_0 ενώ είναι λανθασμένη. Στον παρακάτω πίνακα έχουμε παραστατικά τις δυνατές αποφάσεις που μπορεί να πάρει ο ερευνητής και τα λάθη.

		Πραγματική κατάσταση	
		H_0 σωστή	**H_0 λάθος**
Απόφαση	**Απορρίπτω την H_0**	Λάθος απόφαση Λάθος τύπου I	Σωστή απόφαση
	Δεν απορρίπτω την H_0	Σωστή απόφαση	Λάθος απόφαση Λάθος τύπου II

4.1Πίνακας Λήψης αποφάσεων

Η πιθανότητα να κάνουμε το λάθος τύπου Ι δηλαδή η πιθανότητα

$$P(\text{να απορρίψουμε την } H_0 \text{ ενώ στην πραγματικότητα η } H_0 \text{ είναι σωστή }) = P(\text{να κάνω λάθος τύπου I}) = \alpha$$

ονομάζεται ***επίπεδο σημαντικότητας*** *του ελέγχου*. Το α είναι η πιθανότητα λάθους που επιτρέπει ο ερευνητής στον εαυτό του να απορρίψει την μηδενική υπόθεση ενώ αυτή είναι σωστή, δηλαδή να κάνει λάθος τύπου Ι.

Όταν θέλουμε να ελέγξουμε μια υπόθεση το ζήτημα είναι ποια υπόθεση θα θεωρήσουμε ως H_0 και ποια ως H_1. Ως H_0 θα θεωρήσουμε την υπόθεση για την οποία θα θέλαμε να έχουμε ισχυρές ενδείξεις για να την απορρίψουμε. Για παράδειγμα όταν δικάζεται ένας πολίτης για κάποιο αδίκημα ο δικαστής θεωρεί ότι ο κατηγορούμενος είναι αθώος και με αυτό δεδομένο, εξετάζει αν θα απορρίψει ή όχι την αθωότητα του κατηγορουμένου. Δηλαδή ο δικαστής ορίζει ως H_0 την υπόθεση ότι ο κατηγορούμενος είναι αθώος και ως H_1 ότι ο κατηγορούμενος δεν είναι αθώος.
Από τα παραπάνω γίνεται κατανοητό ότι είναι πολύ σημαντικός ο καθορισμός της πιθανότητας του λάθους τύπου Ι. Συνήθως οι έλεγχοι γίνονται σε επίπεδο σημαντικότητας 5%. Βέβαια δεν υπάρχει ένας γενικός κανόνας για τον καθορισμό του *επιπέδου σημαντικότητας*.
Για την πραγματοποίηση ενός ελέγχου υπόθεσης, αφού ορισθεί η μηδενική υπόθεση και το επίπεδο σημαντικότητας του ελέγχου, στην συνέχεια προσδιορίζεται η στατιστική συνάρτηση η οποία θα χρησιμοποιηθεί για τον έλεγχο. Κατόπιν γίνεται η συλλογή των στοιχείων του δείγματος και υπολογίζεται η τιμή της στατιστικής συνάρτησης., βάσει της οποίας λαμβάνεται η απόφαση να απορριφθεί ή να μην απορριφθεί η μηδενική υπόθεση.

Όπως αναφέρθηκε προηγουμένως δεν υπάρχει ένας γενικός κανόνας για τον καθορισμό του επιπέδου σημαντικότητας, αυτό οδήγησε στον ορισμό ενός άλλου μεγέθους, το οποίο ονομάζεται *παρατηρούμενο επίπεδο σημαντικότητας(p-value)*.

Το (p-value) ορίζεται ως η πιθανότητα να προκύψει μια τιμή τόσο ή περισσότερο ακραία από αυτήν που προέκυψε από το δείγμα για την στατιστική συνάρτηση που χρησιμοποιείται για τον έλεγχο, θεωρώντας ότι η μηδενική υπόθεση είναι σωστή. Συνεπώς αν το p-value είναι μικρότερο από το επίπεδο σημαντικότητας του ελέγχου σημαίνει ότι απορρίπτω την μηδενική υπόθεση, διαφορετικά δεν απορρίπτω την H_0.

Τέλος να αναφέρουμε ότι υπάρχουν δύο είδη στατιστικών υποθέσεων η *Απλή στατιστική υπόθεση* και η *Σύνθετη στατιστική υπόθεση*.

- Απλή στατιστική υπόθεση: Είναι η υπόθεση όπου η H_0 αναφέρεται σε μια μόνο τιμή για παράδειγμα H_0: μ=20 έναντι της H_1: μ≠20, όπου μ η μέση τιμή του πληθυσμού. Ο έλεγχος σε αυτήν την περίπτωση ονομάζεται δίπλευρος
- Σύνθετη στατιστική υπόθεση: Είναι η υπόθεση όπου η H_0 αναφέρεται σε περισσότερες από μια τιμές, για παράδειγμα H_0: μ≤20 έναντι της H_1: μ≥20, όπου μ η μέση τιμή. Ο έλεγχος σε αυτήν την περίπτωση ονομάζεται μονόπλευρος.

4.2.1 Έλεγχος υπόθεσης για την μέση τιμή ενός πληθυσμού

Στην παράγραφο αυτήν θα ελέγξουμε αν η μέση τιμή του πληθυσμού ισούται με μια συγκεκριμένη τιμή. Θέλουμε δηλαδή να πραγματοποιήσουμε τον έλεγχο

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0.$$

Για τον έλεγχο υπόθεσης. για την μέση τιμή θα χρησιμοποιήσουμε δυο διαφορετικές μεθόδους όπως περιγράφονται παρακάτω

Μέθοδος 1

Δημιουργούμε ένα (1-α)*100% διάστημα εμπιστοσύνης για την μέση τιμή (που α το επίπεδο σημαντικότητας του ελέγχου) όπως στην παράγραφο 4.1.2 και αν η τιμή μ_0 περιέχεται στο διάστημα, δεν απορρίπτουμε την μηδενική υπόθεση διαφορετικά την απορρίπτουμε.

Μέθοδος 2

Δημιουργούμε μία στήλη με τόσα κελιά όσα και τα δεδομένα του δείγματος και πληκτρολογούμε την τιμή μ_0 σε όλα τα κελιά της στήλης στην συνέχεια ***Δεδομένα→Ανάλυση δεδομένων→t-test κατά ζεύγη→*** στο παράθυρο που εμφανίζεται στο ένα πεδίο εισόδου, *εισάγουμε τα δεδομένα μας* και στο άλλο *την στήλη που δημιουργήσαμε με το* μ_0.

Στον *πίνακα των αποτελεσμάτων το p-value* εμφανίζεται στο πεδίο

P(T<=t) two-tail

Αν το p-value είναι μικρότερο από το επίπεδο σημαντικότητας του ελέγχου απορρίπτουμε την μηδενική υπόθεση, ενώ αν το p-value είναι μεγαλύτερο από το επίπεδο σημαντικότητας του ελέγχου δεν απορρίπτουμε την μηδενική υπόθεση.

Για το παράδειγμα του μέσου χρόνου αναμονής στο ταμείο μιας τράπεζας, θέλουμε να ελέγξουμε την υπόθεση ότι ο μέσος χρόνος αναμονής στο ταμείο είναι 34 λεπτά έναντι της υπόθεσης ότι ο μέσος χρόνος αναμονής δεν είναι 34 λεπτά, σε επίπεδο σημαντικότητας 5%. Για την πραγματοποίηση του ελέγχου λαμβάνουμε δείγμα 90 πελατών και σημειώνουμε τον χρόνο αναμονής στο ταμείο. Έχουμε τις παρακάτω υποθέσεις

$$H_0: \mu = 34$$

$$H_1: \mu \neq 34$$

Μέθοδος 1

Δημιουργούμε το 95% διάστημα εμπιστοσύνης για την μέση τιμή του χρόνου αναμονής, το οποίο είναι (28,01 , 35,29), όπως υπολογίστηκε στην παράγραφο 4.1.2. παρατηρούμε ότι το 34 περιέχεται στο διάστημα, οπότε δεν απορρίπτουμε την μηδενική υπόθεση σε επίπεδο σημαντικότητας 5%.

Μέθοδος 2

Δημιουργούμε μία στήλη με 90 κελιά όσα και τα στοιχεία του δείγματος και πληκτρολογούμε την τιμή 34 σε όλα τα κελιά της στήλης αυτής, στην συνέχεια

Δεδομένα→Ανάλυση δεδομένων→t-test κατά ζεύγη→ στο παράθυρο που εμφανίζεται στο ένα πεδίο εισόδου, εισάγουμε τα δεδομένα μας και στο άλλο την στήλη που δημιουργήσαμε. Στον πίνακα των αποτελεσμάτων εμφανίζεται το p-value το οποίο ισούται με 0,20>0,05, άρα δεν απορρίπτουμε την υπόθεση ότι ο μέσος χρόνος αναμονής στο ταμείο της τράπεζας είναι 34 λεπτά σε επίπεδο σημαντικότητας 5%.

Αν ο έλεγχος επαναληφθεί με διαφορετικό επίπεδο σημαντικότητας, η υπόθεση μπορεί να απορριφθεί.

4.2.2 Έλεγχος ανεξαρτησίας X^2

Ο έλεγχος ανεξαρτησίας μας βοηθά να διαπιστώσουμε αν δύο κριτήρια ταξινόμησης ως προς τα οποία εξετάζουμε τον πληθυσμό μας είναι ανεξάρτητα το ένα από το άλλο. Για παράδειγμα θέλουμε να εξετάσουμε αν είναι το επίπεδο ικανοποίησης από τις υπηρεσίες μιας υπηρεσίας του Δημοσίου είναι ανεξάρτητο του φύλου, ή αν η καπνιστική συνήθεια είναι ανεξάρτητη του φύλου ή του επιπέδου εκπαίδευσης. Με υποθέσεις

H_0: Ανεξαρτησία

H_1: εξάρτηση.

Έλεγχος με το EXCEL

Για να υλοποιηθεί ο έλεγχος ανεξαρτησίας, στο EXCEL, πρέπει να δημιουργήσουμε *δύο πίνακες*. Ο ένας είναι ο *πίνακας συχνοτήτων* τον οποίο αναλύσαμε στην παράγραφο 2.2 και ο άλλος είναι ο *πίνακας αναμενόμενων συχνοτήτων*, δηλαδή των συχνοτήτων που θα είχαν τα κελιά αν ίσχυε η υπόθεση της ανεξαρτησίας.

Για κάθε κελί του αρχικού πίνακα συχνοτήτων *η αναμενόμενη συχνότητα* ισούται με το *γινόμενο του αθροίσματος της στήλης του κελιού με το άθροισμα της γραμμής του κελιού δια του αθροίσματος όλων των κελιών*. Μ εαυτόν τον τρόπο δημιουργείται ένας νέος πίνακας διασταύρωσης που ως τιμές έχει τις αναμενόμενες συχνότητες των κελιών.
Στην συνέχεια καλούμε την συνάρτηση ***CHISQ.TEST*** η οποία υπολογίζει το p-value του ελέγχου.

Για να είναι έγκυρα τα αποτελέσματα ενός ελέγχου ανεξαρτησίας, θα πρέπει να πληρούνται οι παρακάτω προϋποθέσεις

1. Καμία αναμενόμενη συχνότητα να μην είναι μικρότερη του 1
2. Το πολύ το 20% των κελιών του πίνακα να έχει αναμενόμενη συχνότητα μικρότερη του 5.

Αν κάποια από τις προϋποθέσεις δεν πληρείται, τότε κάνουμε συγχώνευση κελιών, και υπολογίζουμε εκ νέου τις αναμενόμενες συχνότητες. Στην περίπτωση των πινάκων με δύο γραμμές και δύο στήλες, όπου δεν είναι εφικτή η συγχώνευση κελιών εφαρμόζουμε το *Fisher's exact test*, το οποίο όμως δεν δίνεται από το EXCEL.

Παράδειγμα 1

Για την μελέτη της καπνιστικής συνήθειας, επελέγη δείγμα 49 ατόμων, στα οποία δόθηκε ερωτηματολόγιο, στο οποίο, μεταξύ άλλων, οι ερωτώμενοι σημείωναν το φύλο και αν είναι καπνιστές ή όχι.

Στον παρακάτω πίνακα συχνοτήτων απεικονίζονται τα αποτελέσματα για την σχέση μεταξύ φύλου και καπνιστικής συνήθειας.

	Καπνιστής		
Φύλο	**Ναι**	**Όχι**	**Σύνολο**
Άνδρας	20	10	30
Γυναίκα	15	4	19
Σύνολο	35	14	49

Έλεγχος υπόθεσης

H0: Ανεξαρτησία φύλου και καπνίσματος

H1: Εξάρτηση φύλου και καπνίσματος

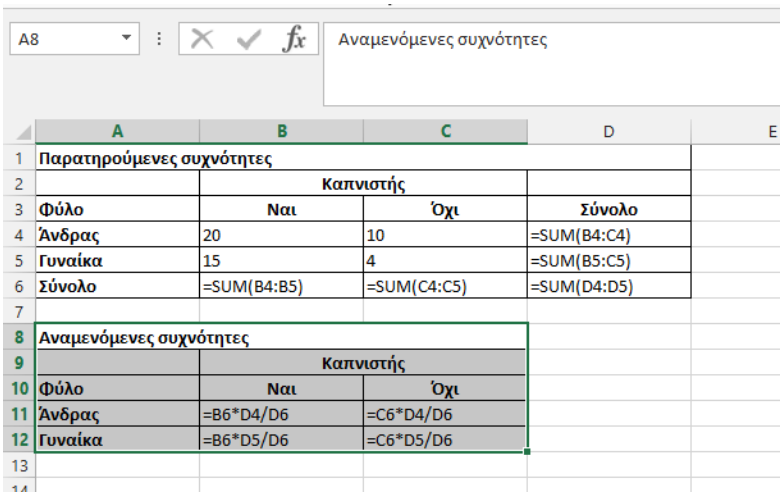
Για την υλοποίηση του ελέγχου υπολογίζουμε τις αναμενόμενες συχνότητες, ελέγχουμε αν ισχύουν οι προϋποθέσεις του ελέγχου και εκτελούμε τον έλεγχο.

Για τον υπολογισμό του πίνακα αναμενόμενων συχνοτήτων ακολουθούμε τα παρακάτω βήματα

1. Δημιουργούμε τον παρακάτω πίνακα

	Αναμενόμενες συχνότητες	
	Καπνιστής	
Φύλο	**Ναι**	**Όχι**
Άνδρας		
Γυναίκα		

2. Δημιουργούμε τις αναμενόμενες συχνότητες όπως φαίνεται παρακάτω φύλλο

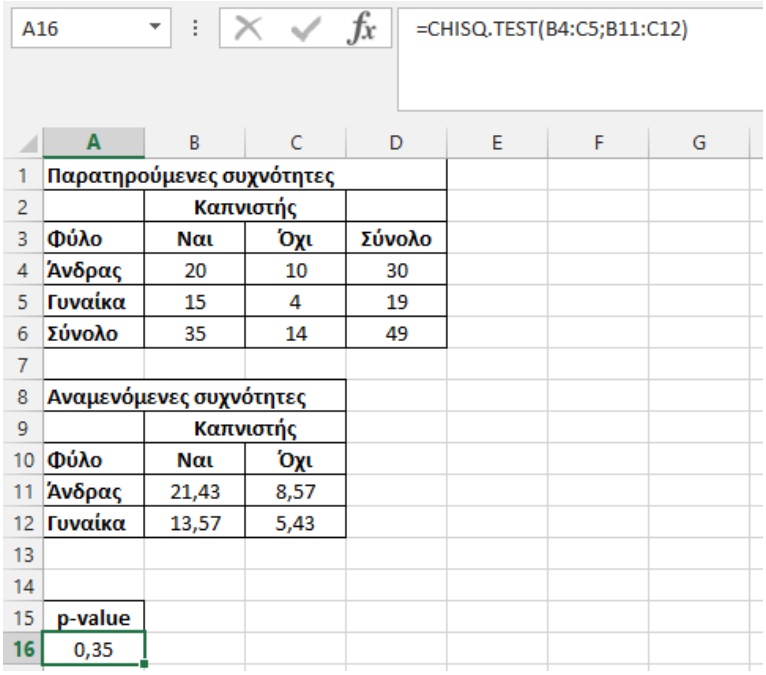


	A	B	C	D
1	Παρατηρούμενες συχνότητες			
2		Καπνιστής		
3	Φύλο	Ναι	Όχι	Σύνολο
4	Άνδρας	20	10	=SUM(B4:C4)
5	Γυναίκα	15	4	=SUM(B5:C5)
6	Σύνολο	=SUM(B4:B5)	=SUM(C4:C5)	=SUM(D4:D5)
7				
8	Αναμενόμενες συχνότητες			
9		Καπνιστής		
10	Φύλο	Ναι	Όχι	
11	Άνδρας	=B6*D4/D6	=C6*D4/D6	
12	Γυναίκα	=B6*D5/D6	=C6*D5/D6	
13				

4.2Εικόνα Συναρτήσεις υπολογισμού Αναμενόμενων συχνοτήτων

Έτσι προκύπτει ο πίνακας των αναμενόμενων συχνοτήτων, όπως φαίνεται παρακάτω. Παρατηρούμε ότι όλες οι αναμενόμενες συχνότητες είναι μεγαλύτερες του 5 άρα ισχύουν οι προϋποθέσεις του ελέγχου.

Στο κελί ***A16*** καλούμε την συνάρτηση ***CHISQ.TEST*** και στο παράθυρο που εμφανίζεται, στο 'Πραγματικό εύρος πληκτρολογούμε ***B4:C5*** ενώ στο αναμενόμενο εύρος ***B11:C12***. Από την τιμή 0,35>0,05, για το P-value του ελέγχου καταλαβαίνουμε ότι δεν απορρίπτεται η υπόθεση της ανεξαρτησίας μεταξύ φύλου και καπνίσματος σε επίπεδο σημαντικότητας 5%.



	A	B	C	D
1	Παρατηρούμενες συχνότητες			
2		Καπνιστής		
3	Φύλο	Ναι	Όχι	Σύνολο
4	Άνδρας	20	10	30
5	Γυναίκα	15	4	19
6	Σύνολο	35	14	49
7				
8	Αναμενόμενες συχνότητες			
9		Καπνιστής		
10	Φύλο	Ναι	Όχι	
11	Άνδρας	21,43	8,57	
12	Γυναίκα	13,57	5,43	
13				
14				
15	p-value			
16	0,35			

4.3Εικόνα Έλεγχος X^2

Παράδειγμα 2

Για την μελέτη της καπνιστικής συνήθειας, επελέγη δείγμα 87 ατόμων, στα οποία δόθηκε ερωτηματολόγιο στο οποίο μεταξύ άλλων οι ερωτώμενοι σημείωναν το φύλο και αν καπνίζουν πολύ, αρκετά, λίγο ή καθόλου. Τα αποτελέσματα παρουσιάζονται στον παρακάτω πίνακα

Παρατηρούμενες συχνότητες					
	Καθόλου	**Λίγο**	**Αρκετά**	**Πολύ**	**Σύνολο**
Γυναίκα	20	1	10	40	71
Άνδρας	10	1	0	5	16
Σύνολο	30	2	10	45	87

4.4Πίνακας παρατηρούμενων συχνοτήτων Κάπνισμα Φύλο

Από τον πίνακα των αναμενομένων συχνοτήτων προκύπτει ότι δεν ισχύουν οι προϋποθέσεις του ελέγχου.

Αναμενόμενες συχνότητες				
	Καθόλου	**Λίγο**	**Αρκετά**	**Πολύ**
Γυναίκα	24,48276	1,632184	8,16092	36,72414
Άνδρας	5,517241	0,367816	1,83908	8,275862

4.5Πίνακας αναμενόμενες συχνότητες Κάπνισμα Φύλο

Θα συγχωνεύσουμε τα κελιά αρκετά και λίγο σε ένα και θα το ονομάσουμε μέτρια. Έτσι προκύπτει ο πίνακας συγχώνευσης κελιών

Παρατηρούμενες συχνότητες				
	Καθόλου	**Μέτρια**	**Πολύ**	**Σύνολο**
Γυναίκα	20	11	40	71
Άνδρας	10	1	5	16
Σύνολο	30	12	45	87

4.6Πίνακας συγχώνευσης Κάπνισμα Φύλο

Δημιουργούμε τον νέο πίνακα αναμενόμενων συχνοτήτων, στον οποίο παρατηρούμε ότι ισχύουν οι προϋποθέσεις του ελέγχου και επαναλαμβάνουμε τον έλεγχο.

Αναμενόμενες συχνότητες			
	Καθόλου	**Μέτρια**	**Πολύ**
Γυναίκα	24,48	9,79	36,72
Άνδρας	5,51	2,21	8,28

4.7Πίνακας συγχώνευσης Κάπνισμα Φύλο

Άσκηση 4ου κεφαλαίου

Ένας φορέας του Δημοσίου θέλει να ερευνήσει αν τα έτη προϋπηρεσίας πριν τον διορισμό στο Δημόσιο επηρεάζουν την απόδοση των υπαλλήλων Για τον λόγο αυτό διεξήγαγε δειγματοληπτική έρευνα,, στην οποία εκτός άλλων χαρακτηριστικών, καταγράφηκε το φύλο (μεταβλητή ***Φύλο***), το εκπαιδευτικό επίπεδο (μεταβλητή ***Εκπαιδευτικό επίπεδο***), τα έτη προϋπηρεσίας πριν τον διορισμό στο Δημόσιο (μεταβλητή ***Έτη προϋπηρεσίας***), και η απόδοση του υπαλλήλου (μεταβλητή ***Απόδοση),***.

Τα δεδομένα της άσκησης, βρίσκονται στο αρχείο με όνομα ***Δεδομένα*** στο φύλλο με το όνομα ***Άσκηση_4-5***

1. Να υπολογισθεί ένα διάστημα εμπιστοσύνης για την μέση απόδοση των υπαλλήλων.

2. Να ελεγχθεί αν η μέση απόδοση των υπαλλήλων διαφέρει από το 80.

3. Να ελεγχθεί η ανεξαρτησία φύλου και εκπαιδευτικού επιπέδου

ΚΕΦΑΛΑΙΟ 5^{Ο}

ΣΥΣΧΕΤΙΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Στην παράγραφο αυτή θα παρουσιασθούν οι έννοιες της **Συσχέτισης** μεταξύ δύο μεταβλητών και της **Γραμμικής παλινδρόμησης**.

5.1 Συσχέτιση

Η συσχέτιση είναι μια στατιστική τεχνική που μας βοηθά να εξετάσουμε σε ποιο βαθμό σχετίζονται οι τιμές δύο μεταβλητών. Συσχέτιση μπορεί να αναζητηθεί σε ζεύγη ποιοτικών ή ποσοτικών μεταβλητών. Παράλληλα, στις ποσοτικές μεταβλητές η συσχέτιση μπορεί να αφορά γραμμική ή μη γραμμική σχέση μεταξύ των τιμών των μεταβλητών.

Θα πρέπει να τονιστεί εδώ ότι η συσχέτιση μεταξύ δύο μεταβλητών δεν σημαίνει απαραίτητα και σχέση αιτίας-αιτιατού απαραίτητα.

Στο παρόν κεφάλαιο θα μας απασχολήσει μόνο η γραμμική συσχέτιση μεταξύ ποσοτικών μεταβλητών. Ας πάρουμε για παράδειγμα τη συσχέτιση μεταξύ των αναστημάτων δύο ενηλίκων αδελφών ίδιου φύλου. Έστω ότι έχουμε δύο ποσοτικές μεταβλητές (λόγου), η μια αφορά το ύψος του ενός ατόμου και η άλλη το ύψος του αδελφού του μια ότι τα ζεύγη τιμών είναι n. Αν οι δύο μεταβλητές συσχετίζονται γραμμικά τότε το διάγραμμα διασποράς αυτών θα αποτελείται από σημεία τα οποία θα τείνουν να συσσωρεύονται γύρω από μία ευθεία γραμμή.

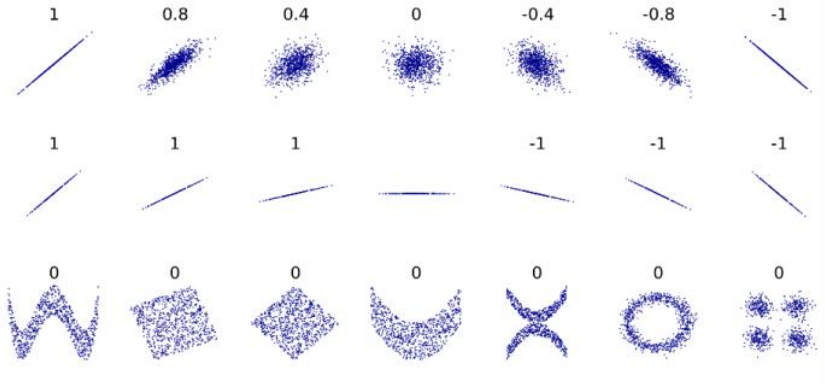
Η ένταση της γραμμικής σχέσης μεταξύ δύο μεταβλητών, δηλαδή πόσο κοντά συσσωρεύονται τα σημεία στο γράφημα διασποράς των δύο μεταβλητών γύρω από μία ευθεία γραμμή, μετριέται με τον ***συντελεστή συσχέτισης Pearson***, ο οποίος συμβολίζεται με **r** για το δείγμα και **ρ** για τον πληθυσμό. Οι τιμές του συντελεστή συσχέτισης Pearson βρίσκονται στο διάστημα από *-1 έως και +1*.

Αρνητικές τιμές του συντελεστή συσχέτισης δηλώνουν ότι η αύξηση των τιμών της μίας μεταβλητής συνοδεύεται με μια (μάλλον) γραμμικού τύπου μείωση των τιμών της άλλης. Στην περίπτωση αυτή λέμε ότι οι μεταβλητές συσχετίζονται αρνητικά, δηλαδή ότι έχουμε *αρνητική συσχέτιση*.

Θετικές τιμές του συντελεστή συσχέτισης δηλώνουν ότι η αύξηση των τιμών της μίας μεταβλητής συνοδεύεται με μια (μάλλον) γραμμικού τύπου αύξηση των τιμών της άλλης. Στην περίπτωση αυτή λέμε ότι οι μεταβλητές συσχετίζονται θετικά, δηλαδή ότι έχουμε *θετική συσχέτιση*.

Η *τιμή -1* δηλώνει πλήρη *αρνητική συσχέτιση*, δηλαδή ότι στο διάγραμμα διασποράς των δύο μεταβλητών, τα σημεία συγκεντρώνονται πάνω σε μία ευθεία γραμμή με αρνητική κλίση. Η *τιμή 1* δηλώνει πλήρη *θετική συσχέτιση*, δηλαδή ότι στο διάγραμμα διασποράς των δύο μεταβλητών, τα σημεία συγκεντρώνονται πάνω σε μία ευθεία γραμμή με θετική κλίση. Όσο πιο κοντά στο 1 ή το -1 είναι η τιμή του συντελεστή συσχέτισης, τόσο πιο έντονη είναι η γραμμική σχέση μεταξύ των δύο μεταβλητών. Η τιμή *0 ή τιμές κοντά στο 0* δηλώνουν ότι μεταξύ των μεταβλητών δεν υπάρχει *γραμμική σχέση*.

Όπως αναφέρθηκε και στην αρχή της παραγράφου, αν διαπιστωθεί ότι οι τιμές δύο μεταβλητών δεν συνδέονται με γραμμική σχέση αυτό δεν σημαίνει ότι αυτές δεν μπορούν να συνδέονται με κάποια άλλη μορφή σχέσης. Υπάρχουν πολλοί τρόποι σχέσης δύο μεταβλητών. Στην παρακάτω εικόνα παρουσιάζονται διάφορες μορφές σχέσης δύο μεταβλητών και ο αντίστοιχος συντελεστής Pearson. Όπως φαίνεται στην τρίτη γραμμή της εικόνας, μπορεί ο συντελεστής συσχέτισης Pearson να είναι ίσος με το μηδέν για ένα σύνολο τιμών δύο μεταβλητών αλλά μπορεί να ενυπάρχει μεταξύ των μεταβλητών κάποια σχέση, όχι όμως γραμμική.



5.1*Εικόνα Γραμμικές μη γραμμικές συσχετίσεις*

Πηγή Wikipedia[6]

Υπολογισμός συντελεστή συσχέτισης με το EXCEL

Στην παρακάτω εικόνα αποτυπώνονται οι δαπάνες για διαφήμιση και οι αντίστοιχες πωλήσεις μιας εταιρείας. Για τον υπολογισμό του Συντελεστή Συσχέτισης με το EXCEL, θα χρησιμοποιήσουμε δυο διαφορετικές μεθόδους όπως περιγράφονται παρακάτω.

[6] By DenisBoigelot, original uploader was Imagecreator - Έργο αυτού που το ανεβάζει, original uploader was Imagecreator, CC0, https://commons.wikimedia.org/w/index.php?curid=15165296

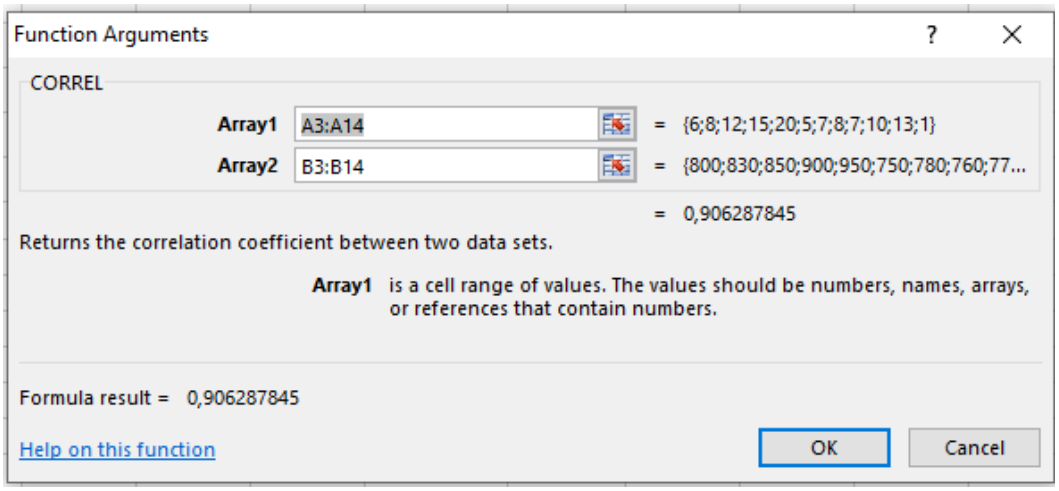
Δαπάνες διαφήμισης (σε χιλ.€)	Πωλήσεις (σε χιλ €)
6	800
8	830
12	850
15	900
20	950
5	750
7	780
8	760
7	770
10	810
13	820
1	580

5.2Πίνακας Διαφήμιση Πωλήσεις

Μέθοδος 1η

Μέσω του EXCEL, μπορεί να υπολογισθεί η γραμμική συσχέτιση με δύο συναρτήσεις. Αυτές δίνουν παρόμοια αποτελέσματα και είναι οι *Pearson* και *Correl*. Στο παράδειγμα που ακολουθεί θα χρησιμοποιήσουμε την συνάρτηση Correl

Στο κελί D3 καλούμε την συνάρτηση ***correl***, οπότε εμφανίζεται το παράθυρο



5.3Εικόνα Η συνάρτηση Correl

Στις περιοχές εισόδου εισάγουμε τα δεδομένα από Α3 έως Α14, για την πρώτη μεταβλητή και από Β3 έως Β14 για την δεύτερη.

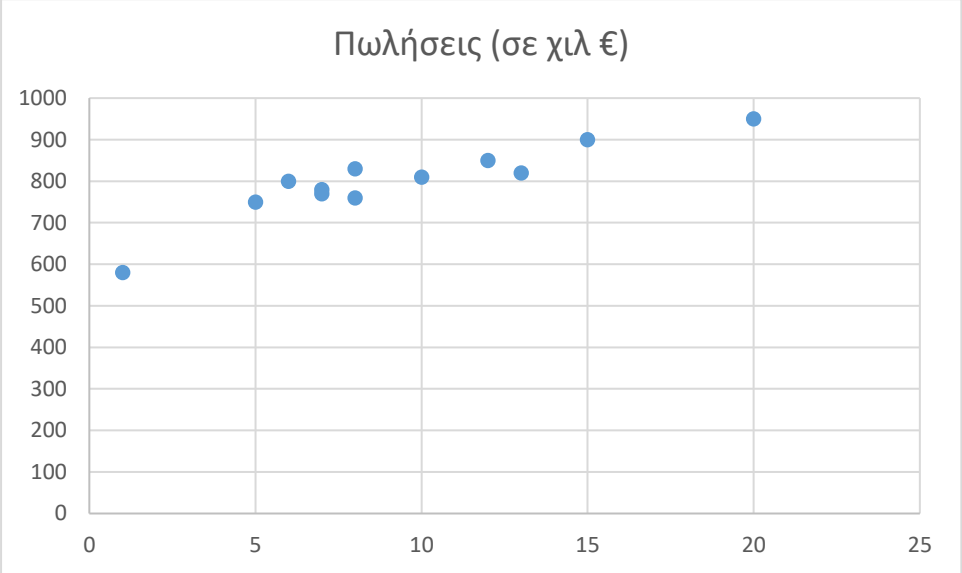
Μέθοδος 2

Δεδομένα →Ανάλυση δεδομένων→Συσχέτιση→Στις περιοχές εισόδου εισαγωγή των δύο μεταβλητών. Οπότε προκύπτει ο παρακάτω πίνακας αποτελεσμάτων

	Δαπάνες διαφήμισης (σε χιλ. €)	**Πωλήσεις (σε χιλ.€)**
Δαπάνες διαφήμισης (σε χιλ. €)	1	
Πωλήσεις (σε χιλ.€)	0,91	1

5.4Πίνακας Υπολογισμός συντελεστή συσχέτισης

Ο συντελεστής συσχέτισης είναι πολύ κοντά στο 1 , άρα υπάρχει ισχυρή θετική συσχέτιση. Κάτι το οποίο φαίνεται και από το γράφημα διασποράς.



5.5Γράφημα Διασποράς Πωλήσεις Διαφήμιση

5.2 Παλινδρόμηση

Στην προηγούμενη παράγραφο μελετήσαμε την συσχέτιση μεταξύ δύο μεταβλητών. Στην περίπτωση που διαπιστώσουμε ότι υπάρχει σχέση μεταξύ των μεταβλητών θα ήταν χρήσιμο να μπορέσουμε να χρησιμοποιήσουμε αυτήν την σχέση για πρόβλεψη.

Με την **Ανάλυση παλινδρόμησης** εξετάζουμε την σχέση μεταξύ δύο ή περισσοτέρων μεταβλητών με στόχο την πρόβλεψη των τιμών μιας από τις μεταβλητές (που την ονομάζουμε **εξαρτημένη μεταβλητή**) μέσω των τιμών των υπολοίπων μεταβλητών (που ονομάζονται **ανεξάρτητες**). Η παραγόμενη εξίσωση αποτελεί το **μοντέλο παλινδρόμησης.**

Για παράδειγμα το Υπουργείο Πολιτισμού θέλει να προβλέψει τα αποτελέσματα μιας καμπάνιας για την προσέλκυση τουριστών. Αν η προσέλκυση τουριστών εξαρτάται μόνο από τα έξοδα διαφήμισης της καμπάνιας αυτής, τότε αναφερόμαστε στην πιο απλή μορφή παλινδρόμησης που είναι η απλή γραμμική παλινδρόμηση. Στην περίπτωση αυτή, η πρόβλεψη της εξαρτημένης μεταβλητής γίνεται μόνο από μία ανεξάρτητη μεταβλητή. Βέβαια η προσέλκυση τουριστών δεν εξαρτάται ενδεχομένως μόνο από τις δαπάνες για την καμπάνια. Στην περίπτωση αυτή, δημιουργούμε μία εξίσωση με περισσότερες από μία ανεξάρτητες μεταβλητές. Η περίπτωση αυτή αφορά την *Πολλαπλή παλινδρόμηση*.

Στην επόμενη παράγραφο θα παρουσιασθεί το μοντέλο μόνο της *Απλής γραμμικής παλινδρόμησης*.

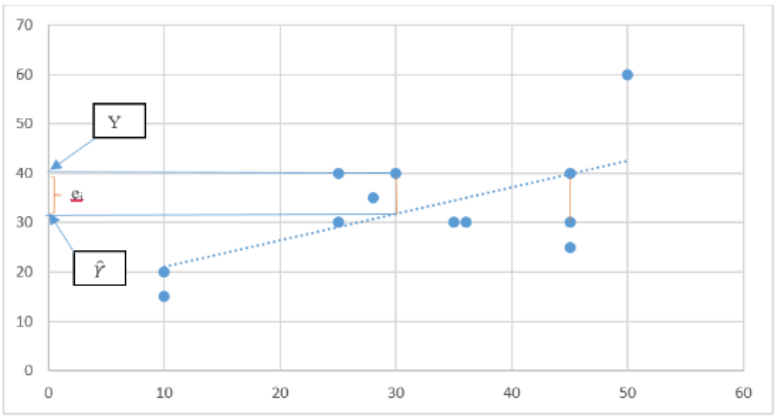
5.2.1 Απλή γραμμική παλινδρόμηση

Όπως αναφέρθηκε προηγουμένως, η απλή γραμμική παλινδρόμηση είναι η πιο απλή μορφή παλινδρόμησης. Η εξίσωση που αναζητούμε έχει μορφή $\hat{Y} = \alpha + \beta * X$, όπου $\hat{Y}$ η *εκτιμώμενη τιμή* για την εξαρτημένη μεταβλητή και X η *ανεξάρτητη*. Τα *α* και *β* είναι οι *συντελεστές*, οι οποίοι θα προσδιορισθούν από τα δεδομένα.

Στο παρακάτω γράφημα παρατηρούμε ότι για την ίδια τιμή της ανεξάρτητης μεταβλητής X, οι τιμές της οποίας απεικονίζονται στον οριζόντιο άξονα, η εξαρτημένη μεταβλητή Y, όπως ήταν αναμενόμενο μπορεί να πάρει πολλές διαφορετικές τιμές. Αν για παράδειγμα ανεξάρτητη μεταβλητή ληφθεί το ύψος του ατόμου και εξαρτημένη η μάζα του, τότε άτομα που εμφανίζουν το ίδιο ύψος μπορούν να έχουν διαφορετικές μάζες.

Η ευθεία στο παρακάτω γράφημα είναι η ευθεία που διέρχεται όσο το δυνατόν πιο κοντά από τα σημεία του διαγράμματος, η εξίσωση της οποίας θα χρησιμοποιηθεί για την πρόβλεψη των τιμών της Y.

Όπως διαπιστώνουμε από το γράφημα για κάθε τιμή της ανεξάρτητης μεταβλητής υπάρχει, μία μόνο προβλεπόμενη τιμή $\hat{Y}$, ενώ υπάρχουν διαφορετικές πραγματικές τιμές για την εξαρτημένη. Έτσι για κάθε πραγματική τιμή Yi ισχύει $Y_i = \alpha + \beta * X_i + e_i$. Το e_i ονομάζεται *κατάλοιπο* και προφανώς θέλουμε να είναι όσο το δυνατόν μικρότερο.



5.6Γράφημα διασποράς

Μετά τον υπολογισμό της ευθείας παλινδρόμησης, πρέπει να υπολογίσουμε ένα μέτρο για να διαπιστώσουμε πόσο καλά προσαρμόζεται η ευθεία στα δεδομένα μας. Αυτό το μέτρο είναι ο **συντελεστής προσδιορισμού**, ο οποίος συμβολίζεται με R^2. Οι τιμές του συντελεστή προσδιορισμού είναι από *0 μέχρι 1*. Όσο οι τιμές του συντελεστή προσδιορισμού είναι πιο

κοντά στο 1, τόσο καλύτερη είναι η προσαρμογή της γραμμής στα δεδομένα. Ο συντελεστής προσδιορισμού εκφράζει το ποσοστό της μεταβλητότητας που ερμηνεύεται από το μοντέλο.

Σχέση του συντελεστή συσχέτισης και του συντελεστή β στην εξίσωση $\hat{Y} = \alpha + \beta * X$

- Αν ο συντελεστής συσχέτισης είναι θετικός τότε και το β είναι θετικό
- Αν ο συντελεστής συσχέτισης είναι αρνητικός τότε και το β είναι αρνητικό

Όταν από ένα δείγμα θέλουμε να εκτιμήσουμε την ευθεία παλινδρόμησης στον πληθυσμό, οι συντελεστές α, β που προσδιορίζουμε από τα δεδομένα του δείγματος είναι εκτιμήσεις για τους αντίστοιχους συντελεστές της ευθείας παλινδρόμησης του πληθυσμού.

Άρα πρέπει να κάνουμε έλεγχο για την σημαντικότητα της παλινδρόμησης. Αυτός ο έλεγχος είναι

H_0: β=0

H_1: β≠0

όπου β ο συντελεστής στην εξίσωση $\hat{Y} = \alpha + \beta * X$. Για να είναι η παλινδρόμηση στατιστικά σημαντική, θα πρέπει η H_0 να απορριφθεί.

Για να μπορεί το υπόδειγμα που έχει υπολογισθεί από τα δειγματικά δεδομένα μας να χρησιμοποιηθεί για την εκτίμηση του μοντέλου στον πληθυσμό θα πρέπει ακόμη τα κατάλοιπα να πληρούν κάποιες προϋποθέσεις. Αυτές είναι

- Κανονικότητα (τα κατάλοιπα ακολουθούν κανονική κατανομή)
- Ομοσκεδαστικότητα (τα κατάλοιπα έχουν σταθερή διασπορά)

Οι παραπάνω έλεγχοι γίνονται με την βοήθεια στατιστικών ελέγχων και διαγραμμάτων. Τα διαγράμματα αυτά παρέχονται από το EXCEL από το πρόσθετο *Ανάλυση Δεδομένων*. Οι έλεγχοι καταλοίπων είναι εκτός των στόχων του μαθήματος.

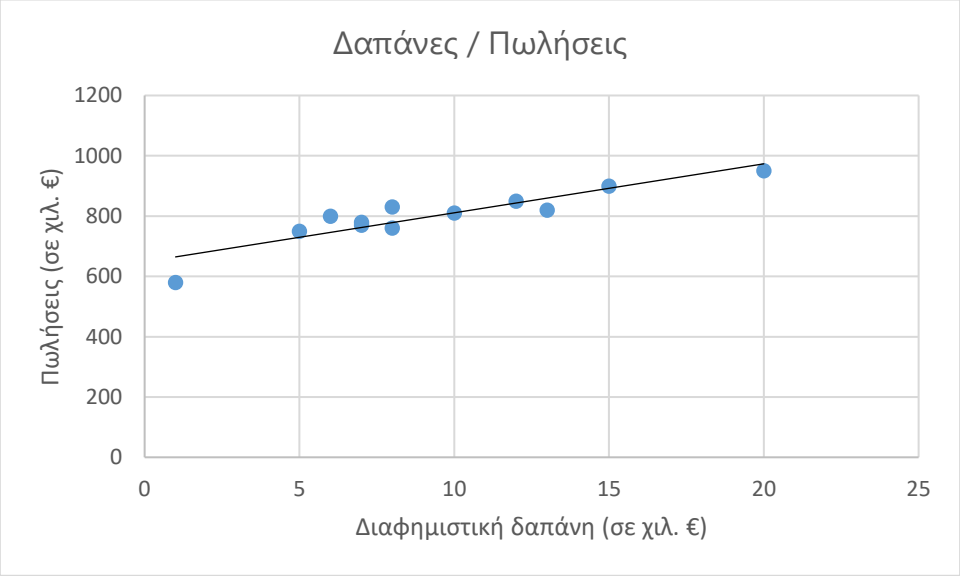
Παράδειγμα

Στον παρακάτω πίνακα αποτυπώνονται οι δαπάνες για διαφήμιση και οι αντίστοιχες πωλήσεις μιας εταιρείας σε χιλιάδες Ευρώ. Θα δημιουργήσουμε μία εξίσωση για την πρόβλεψη των πωλήσεων θεωρώντας ότι αυτές εξαρτώνται μόνο από τις διαφημιστικές δαπάνες .

A/A	Δαπάνες διαφήμισης (σε χιλ. €)	Πωλήσεις (σε χιλ.€)
1	6	800
2	8	830
3	12	850
4	15	900
5	20	950
6	5	750
7	7	780
8	8	760
9	7	770
10	10	810
11	13	820
12	1	580

5.7Πίνακας Πωλήσεις Δαπάνες διαφήμισης

Αρχικά απεικονίζουμε τις τιμές των δύο μεταβλητών σε ένα διάγραμμα διασποράς για να αποκτήσουμε μία εικόνα για τα δεδομένα μας



5.8Γράφημα Διασπορά Πωλήσεις Δαπάνες

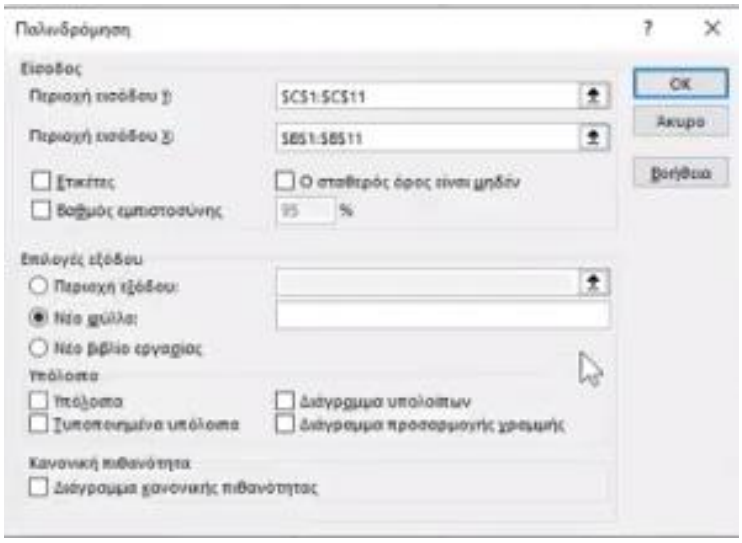
Παρατηρούμε ότι τα σημεία συγκεντρώνονται γύρω από μία ευθεία γραμμή. Στην συνέχεια υπολογίζουμε τον συντελεστή συσχέτισης για τις δύο μεταβλητές *Δαπάνες* και *Πωλήσεις*, ο οποίος ισούται με 0,91. Από την τιμή του συντελεστή συσχέτισης διαπιστώνουμε ότι υπάρχει ισχυρή θετική συσχέτιση μεταξύ δαπανών διαφήμισης και πωλήσεων. Η ανάλυση της σχέσης

των δύο μεταβλητών με στόχο την πρόβλεψη των πωλήσεων από τις διαφημιστικές δαπάνες θα πραγματοποιηθεί με την βοήθεια ενός υποδείγματος *απλής γραμμικής παλινδρόμησης*.

Υπολογισμός με το EXCEL

Δεδομένα → Ανάλυση Δεδομένων→Παλινδρόμηση→

- Στην περιοχή εισόδου Υ εισάγουμε τα κελιά με τις τιμές της εξαρτημένης μεταβλητής *Πωλήσεις* μαζί με τον τίτλο *Πωλήσεις*.
- Στην περιοχή εισόδου Χ εισάγουμε τα κελιά με τις τιμές της ανεξάρτητης μεταβλητής *Δαπάνες* μαζί με τον τίτλο *Δαπάνες*.
- Επιλογή εμφάνισης ετικετών.
- Επιλογή του βαθμού εμπιστοσύνης, για τα διαστήματα εμπιστοσύνης των συντελεστών.
- Επιλογή εξόδου, η οποία μπορεί να είναι σε κάποια θέση του ίδιου φύλλου εργασίας ή σε άλλο φύλλο εργασίας ή ακόμη σε κάποιο νέο βιβλίο εργασίας.
- Επιλογή εμφάνισης των καταλοίπων.
- Επιλογή εμφάνισης του διαγράμματος προσαρμογής γραμμής.



5.9Εικόνα Παλινδρόμηση με το πακέτο Ανάλυσης

Έτσι προκύπτει ο πίνακας με τα στατιστικά της παλινδρόμησης, ο πίνακας ανάλυσης διακύμανσης, και ο πίνακας των συντελεστών. Για την μελέτη του μοντέλου της απλής γραμμικής παλινδρόμησης, θα μελετηθούν στοιχεία από τον πίνακα με τα στατιστικά της παλινδρόμησης, και από τον πίνακα των συντελεστών.

Από τον πίνακα *Στατιστικά παλινδρόμησης* το R^2 ***(διαβάζεται R Τετράγωνο)*** είναι ο συντελεστής προσδιορισμού, ο οποίος ισούται με 0,82, από όπου καταλαβαίνουμε ότι το 82% της συνολικής μεταβλητότητας των πωλήσεων ερμηνεύεται από το μοντέλο.

Στατιστικά παλινδρόμησης	
Πολλαπλό R	0,91
R Τετράγωνο	**0,82**
Προσαρμοσμένο R Τετράγωνο	0,80
Τυπικό σφάλμα	40,14
Μέγεθος δείγματος	12

5.10Πίνακας Στατιστικά Παλινδρόμησης

Από τον πίνακα των συντελεστών που ακολουθεί προκύπτει η εξίσωση της ευθείας παλινδρόμησης η οποία είναι

Πωλήσεις=16,25**Δαπάνες διαφήμισης*+648,36 (1)

	Συντελεστές	*Τυπικό σφάλμα*	*t*	*Τιμή-P*	*Κατώτερο 95%*	*Ανώτερο 95%*
Τεταγμένη επί την αρχή	648,36	25,19	25,74	1,799E-10	592,24	704,48
Δαπάνες διαφήμισης (σε χιλ. €)	16,25	2,40	6,78	4,855E-05	10,91	21,59

5.11Πίνακας συντελεστών παλινδρόμησης

Αν θέλουμε να προβλέψουμε τις πωλήσεις στην περίπτωση που δαπανήσουμε 7000€ για διαφήμιση, θα αντικαταστήσουμε στην εξίσωση (1) στην μεταβλητή *Δαπάνες Διαφήμισης* το 7. Δηλαδή η πρόβλεψη θα είναι

*Πωλήσεις =16,25*7+648,36=762,11*

Άρα, αν δαπανήσουμε:

- 7000€ για διαφήμιση *αναμένουμε* να προκύψουν πωλήσεις 762,11*1000=762110€
- 6000€ αναμένουμε οι πωλήσεις να είναι *πωλήσεις =16,25*6+648,36=745,84 χιλ. δηλαδή745840 €*

Αν αφαιρέσουμε από τις αναμενόμενες πωλήσεις για δαπάνη 7000€ τις αναμενόμενες πωλήσεις για δαπάνη 6000€, δηλαδή (*762,11 - 745,84=16,27) προκύπτει η τιμή του συντελεστή της ανεξάρτητης μεταβλητής στο μοντέλο.*

Από τον συντελεστή των δαπανών στην εξίσωση συμπεραίνουμε ότι αν αυξηθεί η διαφημιστική δαπάνη κατά 1000 €, τότε αναμένουμε να αυξηθούν οι πωλήσεις κατά 16,25*1=16,250 χιλ €. Δηλαδή 16.250€.

Γενικά ο συντελεστής της ανεξάρτητης μεταβλητής στο μοντέλο μας, εκφράζει πόσο αναμένεται να μεταβληθεί η τιμή της εξαρτημένης μεταβλητής, σε αύξηση μιας μονάδας της τιμής της ανεξάρτητης μεταβλητής.

Ο πίνακας των προβλέψεων για τις τιμές της ανεξάρτητης τιμής του παραδείγματος δίνεται από το EXCEL μαζί με τα αντίστοιχα κατάλοιπα.

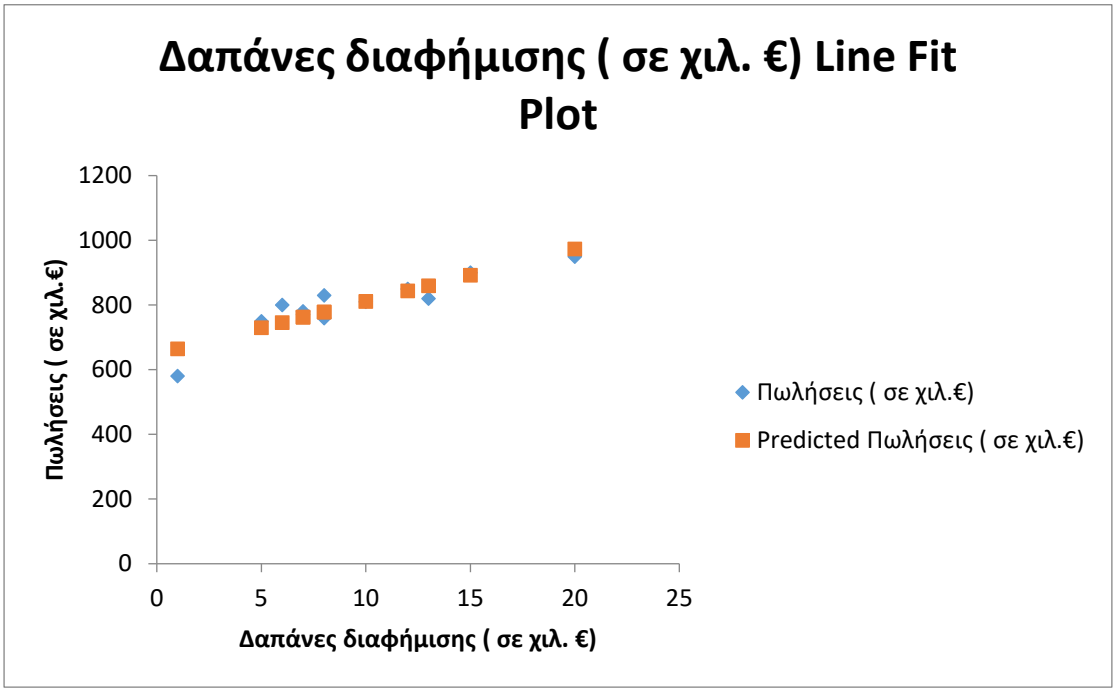
Α/Α Παρατήρησης	**Δαπάνες διαφήμισης (σε χιλ. €)**	**Πωλήσεις (σε χιλ.€)**	**Προβλεπόμενες Πωλήσεις (σε χιλ.€)**	**Κατάλοιπα**
1	6	800	745,84	54,16
2	8	830	778,34	51,66
3	12	850	843,33	6,67
4	15	900	892,07	7,93
5	20	950	973,30	-23,30
6	5	750	729,60	20,40
7	7	780	762,09	17,91
8	8	760	778,34	-18,34
9	7	770	762,09	7,91
10	10	810	810,83	-0,83
11	13	820	859,57	-39,57
12	1	580	664,61	-84,61

5.12Πίνακας: Πίνακας καταλοίπων - προβλεπομένων τιμών

Η εξίσωση (1) χρησιμοποιείται όπως προαναφέραμε για την πρόβλεψη των πωλήσεων. Οι τιμές που δίνονται στην εξίσωση στην ανεξάρτητη μεταβλητή ώστε να προβλεφθεί η εξαρτημένη δεν πρέπει να είναι πολύ μακριά από το εύρος των τιμών της ανεξάρτητης μεταβλητής, γιατί οι προβλέψεις δεν θα είναι αξιόπιστες.

Τα κατάλοιπα υπολογίζονται ως η διαφορά της πραγματικής τιμής της μεταβλητής *Πωλήσεις* από την τιμή που προκύπτει για τις πωλήσεις από την εξίσωση (1). Για παράδειγμα το κατάλοιπο για την πρώτη παρατήρηση ισούται με e_1=800-745,84=54,16.

Στο παρακάτω γράφημα απεικονίζονται οι πραγματικές πωλήσεις και οι εκτιμώμενες από την εξίσωση παλινδρόμησης



5.13Γράφημα πωλήσεων και προβλεπομένων πωλήσεων

Έλεγχος υπόθεσης για την σημαντικότητα της παλινδρόμησης

Όπως έχουμε αναφέρει αν η εξίσωση της παλινδρόμησης είναι $\hat{Y} = \alpha + \beta * X$, τότε η σημαντικότητα της παλινδρόμησης ελέγχεται με τον παρακάτω έλεγχο

H_0: β=0

H_1: β≠0

Στον πίνακα των συντελεστών του παραδείγματός μας το p-value του παραπάνω ελέγχου είναι 0,00<0,05,άρα η παλινδρόμηση είναι στατιστικά σημαντική σε επίπεδο σημαντικότητας 5%.

Για την αξιοπιστία του μοντέλου πρέπει να γίνουν έλεγχοι καταλοίπων, οι οποίοι όμως δεν θα συζητηθούν σε αυτό το μάθημα.

Άσκηση 5ου κεφαλαίου

Στην άσκηση 4 αναφερθήκαμε στην έρευνα ενός φορέα του Δημοσίου, για την απόδοση των υπαλλήλων σε σχέση με τα έτη προϋπηρεσίας πριν τον διορισμό στο Δημόσιο.

1.Να υπολογισθεί ο συντελεστής συσχέτισης για τις μεταβλητές, ***Έτη προϋπηρεσίας*** και ***Απόδοση*** και να σχολιασθεί η συσχέτιση των δύο μεταβλητών.

2.Να κατασκευασθεί μοντέλο παλινδρόμησης για την επίδραση των ετών προϋπηρεσίας πριν τον διορισμό στο Δημόσιο στην απόδοση των υπαλλήλων.

3. Είναι η παλινδρόμηση στατιστικά σημαντική;

4.Να γραφεί η εξίσωση της παλινδρόμησης και να ερμηνευθούν οι συντελεστές.

5.Να δοθεί η ερμηνεία του συντελεστή προσδιορισμού.

Βιβλιογραφία

Καστανιά, Α. & Αποστολάκης, Ι. &Πιερράκου, Χ (2003) Στατιστική επεξεργασία δεδομένων στην υγεία, Αθήνα Παπαζήσης

Κιόχος, Π. (1985). Οικονομική Στατιστική, Πειραιάς Σταμούλης

Παπαϊωάννου, Π. & Λουκάς, Σ. (2002). Εισαγωγή στην Στατιστική, Πειραιάς Σταμούλης

Douglas Downing, Jeffrey Clark ; μετάφραση Παναγιώτης Σταυρόπουλος, Γιώργος Σταυρόπουλος.(1998), Αθήνα Κλειδάριθμος

Κουτρουβέλης Ι. (1999). Πιθανότητες και Στατιστική Ι. Ελληνικό Ανοιχτό Πανεπιστήμιο.

Κουτρουβέλης Ι. (1999). Πιθανότητες και Στατιστική ΙΙ. Ελληνικό Ανοιχτό Πανεπιστήμιο.

Witte, R. and Witte, J. (2016). Statistics. 11th ed. Wiley.

Madsen, B. (2016). Statistics for non-statisticians. Springer-Verlag Berlin Heidelberg.

Cox, V. (2017). Translating statistics to make decisions. Apress

https://imegsevee.gr/wp-content/uploads/2018/02/ypologistika_baseis_dedomenwn.pdf

https://www.youtube.com/watch?v=nDxo1gT8LGY

https://www.youtube.com/watch?v=x7FNIq6FgcM

https://el.wikipedia.org/

https://www.statisticshowto.datasciencecentral.com/

https://ec.europa.eu/eurostat/statistics-explained/

http://users.auth.gr/gvasil/stat_symper.pdf

http://ebooks.edu.gr/modules/ebook/show.php/DSGL-C125/494/3206,13023/

https://stats.oecd.org/glossary/detail.asp?ID=1647

https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Beginners:Statistics_4_beginners/el

https://www.djsresearch.co.uk/glossary/item/Sampling-Unit

https://www.statisticshowto.datasciencecentral.com/sampling-frame/

https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/

https://opentextbc.ca/researchmethods/chapter/reliability-and-validity-of-measurement/

https://repository.kallipos.gr/bitstream/11419/5360/1/01_chapter_04.pdf

https://www.investopedia.com/terms/p/population.asp

https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Beginners:Statistical_concept_-_Survey,_census_and_register/el

https://www.investopedia.com/terms/s/sampling.asp

https://oceclass.aua.gr/modules/document/file.php/OCDAERD104/aoa_th_2375_07b.pdf

https://repository.kallipos.gr/bitstream/11419/5075/1/00_master_document__with-cover.pdf

https://www.youtube.com/watch?v=pTuj57uXWlk

https://www.investopedia.com/terms/s/simple-random-sample.asp

https://repository.kallipos.gr/handle/11419/1296

https://www.investopedia.com/terms/stratified_random_sampling.asp

https://www.statisticshowto.datasciencecentral.com/systematic-sampling/

https://www.questionpro.com/blog/convenience-sampling/

https://explorable.com/judgmental-sampling

https://www.statisticshowto.datasciencecentral.com/snowball-sampling/

https://www.statisticshowto.datasciencecentral.com/quota-sampling/

http://www.statistics.gr/documents/20181/985004/%CE%95%CE%BD%CE%B9%CE%B1%CE%AF%CE%B1+%CE%9C%CE%BF%CF%81%CF%86%CE%AE+%CE%94%CE%BF%CE%BC%CE%AE%CF%82+%CE%9C%CE%B5%CF%84%CE%B1%CE%B4%CE%B5%CE%B4%CE%BF%CE%BC%CE%AD%CE%BD%CF%89%CE%BD+%28SIMS+v.2.0%29+%28+2019+%29/83b74066-e774-4b14-b6f9-945dae973af2?version=1.0

https://repository.kallipos.gr/bitstream/11419/2062/3/02_chapter_03-Iliopoulou_%CE%91%CE%9D%CE%91%CE%98%CE%95%CE%A9%CE%A1%CE%97%CE%A3%CE%97.pdf

https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Beginners:Statistical_concept_-_Aggregate/el

https://users.auth.gr/~dkugiu/Teach/ElectricEngineer/descriptive.pdf

https://support.office.com/

http://ebooks.edu.gr/modules/ebook/show.php/DSGL-C125/494/3206,13025/

https://en.wikipedia.org/wiki/Contingency_table

https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Beginners:Statistical_concept_-_Mean_and_median/el

https://www.aua.gr/gpapadopoulos/files/perigrafiki11.pdf

https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Beginners:Statistical_concept_-_Mean_and_median/el

http://www2.stat-athens.aueb.gr/~jpan/statistiki-skepsi-I/chapter3.pdf

https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Beginners:Statistical_concept_-_Mean_and_median/el

http://skillsacademic.weebly.com/epsilonpiiotakapparhoalphatauomicron973sigmaalpha-tauiotamu942.html

https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Beginners:Statistical_concept_-_Quintile_and_decile/el

https://users.auth.gr/dkugiu/Teach/CivilEngineer/descriptive.pdf

https://el.wikipedia.org/wiki/%CE%A4%CF%85%CF%80%CE%B9%CE%BA%CE%AE_%CE%B1%CF%80%CF%8C%CE%BA%CE%BB%CE%B9%CF%83%CE%B7#%CE%91%CE%BD%CE%B9%CF%83%CF%8C%CF%84%CE%B7%CF%84%CE%B1_Chebyshev

https://www.statisticshowto.datasciencecentral.com/normalized/

https://exceljet.net/excel-functions/excel-standardize-function

https://el.wikipedia.org/wiki/%CE%9A%CE%AF%CE%BD%CE%B4%CF%85%CE%BD%CE%BF%CF%82_(%CE%BF%CE%B9%CE%BA%CE%BF%CE%BD%CE%BF%CE%BC%CE%B9%CE%BA%CE%AC)#%CE%A3%CF%85%CE%BD%CF%84%CE%B5%CE%BB%CE%B5%CF%83%CF%84%CE%AE

%CF%82_%CE%BC%CE%B5%CF%84%CE%B1%CE%B2%CE%BB%CE%B7%CF%84%CF%8C%CF%84%CE%B7%CF%84%CE%B1%CF%82

https://ec.europa.eu/eurostat/databrowser/view/t2020_10/default/table?lang=en

http://photodentro.edu.gr/lor/handle/8521/7872

https://el.wikipedia.org/wiki/%CE%98%CE%B7%CE%BA%CF%8C%CE%B3%CF%81%CE%B1%CE%BC%CE%BC%CE%B1

https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Government_finance_statistics/el#CE.94.CE.B7.CE.BC.CF.8C.CF.83.CE.B9.CE.B1_.CE.AD.CF.83.CE.BF.CE.B4.CE.B1_.CE.BA.CE.B1.CE.B9_.CE.B4.CE.B1.CF.80.CE.AC.CE.BD.CE.B5.CF.82

https://el.wikipedia.org/wiki/%CE%9A%CE%B1%CE%BD%CE%BF%CE%BD%CE%B9%CE%BA%CE%AE_%CE%BA%CE%B1%CF%84%CE%B1%CE%BD%CE%BF%CE%BC%CE%AE

https://el.wikipedia.org/wiki/%CE%9A%CE%B1%CE%BD%CE%BF%CE%BD%CE%B9%CE%BA%CE%AE_%CE%BA%CE%B1%CF%84%CE%B1%CE%BD%CE%BF%CE%BC%CE%AE#%CE%A6%CE%B1%CE%B9%CE%BD%CF%8C%CE%BC%CE%B5%CE%BD%CE%B1_%CF%80%CE%BF%CF%85_%CE%B1%CE%BA%CE%BF%CE%BB%CE%BF%CF%85%CE%B8%CE%BF%CF%8D%CE%BD_%CF%84%CE%B7%CE%BD_%CE%BA%CE%B1%CE%BD%CE%BF%CE%BD%CE%B9%CE%BA%CE%AE_%CE%BA%CE%B1%CF%84%CE%B1%CE%BD%CE%BF%CE%BC%CE%AE

https://stattrek.com/estimation/confidence-interval.aspx

https://www.statisticshowto.datasciencecentral.com/confidence-level/

https://repository.kallipos.gr/bitstream/11419/5363/1/01_chapter_07.pdf

https://repository.kallipos.gr/bitstream/11419/5363/1/01_chapter_07.pdf

https://el.wikipedia.org/wiki/%CE%94%CE%BF%CE%BA%CE%B9%CE%BC%CE%B1%CF%83%CE%AF%CE%B1_X2#%CE%A0%CF%81%CE%BF%CF%8B%CF%80%CE%BF%CE%B8%CE%AD%CF%83%CE%B5%CE%B9%CF%82_%CE%95%CE%BB%CE%AD%CE%B3%CF%87%CE%BF%CF%85_%CE%91%CE%BD%CE%B5%CE%BE%CE%B1%CF%81%CF%84%CE%B7%CF%83%CE%AF%CE%B1%CF%82

https://www.surveysystem.com/correlation.htm

https://users.auth.gr/dkugiu/Teach/DataAnalysis/Chp5.pdf

https://www.emathzone.com/tutorials/basic-statistics/linear-and-non-linear-correlation.html

https://repository.kallipos.gr/bitstream/11419/5082/1/08_chapter7.pdf

https://el.wikipedia.org/wiki/%CE%A3%CF%85%CF%83%CF%87%CE%AD%CF%84%CE%B9%CF%83%CE%B7_%CE%BA%CE%B1%CE%B9_%CE%B5%CE%BE%CE%AC%CF%81%CF%84%CE%B7%CF%83%CE%B7

https://en.wikipedia.org/wiki/Volkswagen_emissions_scandal

https://support.office.com/en-us/article/pearson-function-0c3e30fc-e5af-49c4-808a-3ef66e034c18

Πίνακας περιεχομένων Εικόνων Πινάκων


ΕΣΠΑ
2014-2020
Επιχειρησιακό Πρόγραμμα
ΜΕΤΑΡΡΥΘΜΙΣΗ
ΔΗΜΟΣΙΟΥ ΤΟΜΕΑ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης