Οι αριθμοί και η ζωή μας

ΤΕΥΧΟΣ V

ΕΛΛΗΝΙΚΗ ΣΤΑΤΙΣΤΙΚΗ ΑΡΧΗ

Συγγραφική ομάδα: Γεωργία Λύτρα (Ειδική Σύμβουλος - Αυτοτελές Γραφείο Προέδρου ΕΛΣΤΑΤ)
Μαρία Λαφτσίδου (Τμήμα Σύνθεσης Εθνικών Λογαριασμών ΕΛΣΤΑΤ)

Συμβουλευτική ομάδα: Απόστολος Κασάπης (Διευθυντής Αυτοτελούς Γραφείου Προέδρου ΕΛΣΤΑΤ)
Ιωάννης Μοσχάκης (Προϊστάμενος Γενικής Διεύθυνσης Διοίκησης και Οργάνωσης ΕΛΣΤΑΤ)
Αθανάσιος Σταυρόπουλος (Προϊστάμενος Διεύθυνσης Στατιστικής Πληροφόρησης και Εκδόσεων ΕΛΣΤΑΤ)
Ιωάννης Νικολαΐδης (Προϊστάμενος Τμήματος Μεθοδολογίας και Ονοματολογιών ΕΛΣΤΑΤ)
Γεώργιος Ντούρος (Προϊστάμενος Τμήματος Ειδικών Στατιστικών Νοικοκυριών ΕΛΣΤΑΤ)

Επιμέλεια Κειμένου: Κωνσταντίνος Καμπανάκης (Προϊστάμενος Τμήματος Επιμέλειας Εκδόσεων και Μεταφράσεων ΕΛΣΤΑΤ)
Παναγιώτα Βαληνδρά (Τμήμα Επιμέλειας Εκδόσεων και Μεταφράσεων ΕΛΣΤΑΤ)

Φωτοστοιχειοθεσία: Διονύσης Καπότσης (Τμήμα Φωτοστοιχειοθεσίας και Τυπογραφικής Διαμόρφωσης Εκδόσεων ΕΛΣΤΑΤ)

Εκτυπωτικές και Βιβλιοδετικές Εργασίες: Χρυσάνθη Δροσοπούλου (Προϊσταμένη Τμήματος Εκτυπώσεων ΕΛΣΤΑΤ), Ιουλία Αποστολάκη, Θεόδωρος Αποστολόπουλος, Γεώργιος Ζαφείρης, Ευαγγελία Κάτσουρα, Φλώρα Καψήλη, Ελένη Μαρίνου, Ιωάννης Μυλωνάκης, Μενέλαος Παπαγεωργίου, Ευτυχία Παπαζή (Τμήμα Εκτυπώσεων ΕΛΣΤΑΤ)

Έτος Έκδοσης: 2024

Το παρόν έντυπο εκδόθηκε από την Ελληνική Στατιστική Αρχή (ΕΛΣΤΑΤ) και είναι διαθέσιμο και ηλεκτρονικά στην ιστοσελίδα της Αρχής: www.statistics.gr

ISBN: 978-618-5884-07-9 (Έντυπο)

ISBN: 978-618-5884-08-6 (Ηλεκτρονικό)

Αγαπητέ/ή αναγνώστη/στρια,

Βασικός μας στόχος στην Ελληνική Στατιστική Αρχή (ΕΛΣΤΑΤ) είναι η διάδοση και η προώθηση, σε στοχευμένες ομάδες (π.χ. μαθητές, φοιτητές), της αξίας των επίσημων στατιστικών, αλλά και η δημιουργία στατιστικής συνείδησης σε ολόκληρη την κοινωνία, μέσω της στρατηγικής και των δράσεών μας για την ανάπτυξη της «Στατιστικής Παιδείας» στην Ελλάδα.

Το τετράδιο ασκήσεων «Οι αριθμοί και η ζωή μας, Τεύχος V», που κρατάτε στα χέρια σας, αποσκοπεί στην εξοικείωση των μαθητών/τριών με τις βασικές έννοιες της θεωρίας πιθανοτήτων και της στατιστικής, ενώ δίνει έμφαση στην ικανότητά τους να κατανοούν την έννοια της στατιστικής πληροφορίας και να τη χρησιμοποιούν με τρόπο ωφέλιμο για τη ζωή τους.

Εύχομαι να διασκεδάσετε μαθαίνοντας,

Αθανάσιος Κων. Θανόπουλος
Πρόεδρος της Ελληνικής Στατιστικής Αρχής

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

«ΑΝΑΠΤΥΞΗ ΤΗΣ ΣΤΑΤΙΣΤΙΚΗΣ ΠΑΙΔΕΙΑΣ»

Το τετράδιο ασκήσεων «Οι αριθμοί και η ζωή μας, τεύχος V» εντάσσεται στο πλαίσιο των δράσεων της στρατηγικής της Ελληνικής Στατιστικής Αρχής (ΕΛΣΤΑΤ) για την «Ανάπτυξη της Στατιστικής Παιδείας» στην Ελλάδα. Στόχος των δράσεων είναι να εισάγουν, με εκπαιδευτικό και ψυχαγωγικό χαρακτήρα, μαθητές/ήτριες και φοιτητές/ήτριες στον κόσμο της Στατιστικής, βοηθώντας τους να κατανοήσουν την εφαρμογή της στις καθημερινές τους συνήθειες. Στοχεύουν, επίσης, να ενημερώσουν το ευρύ κοινό για τον ρόλο, τη χρήση και τον τρόπο παραγωγής, βάσει ευρωπαϊκών και διεθνών στατιστικών προτύπων, των επίσημων στατιστικών.

Στο πρόγραμμα περιλαμβάνονται δράσεις, όπως διαγωνισμοί, εκπαιδευτικές επισκέψεις, η ένταξη του EMOS (European Master in Official Statistics) σε Τμήματα Ελληνικών Πανεπιστημίων, διαδικτυακά εκπαιδευτικά παιχνίδια και πολλά άλλα.

Όραμά μας είναι η εξέλιξη και ανάπτυξη του προγράμματος, ώστε να οικοδομήσουμε στατιστική κατανόηση σε ολόκληρη την κοινωνία, διασφαλίζοντας ότι όλοι έχουμε τη δυνατότητα να κατανοούμε και να αξιοποιούμε πλήρως τα δεδομένα.

Όλες οι πληροφορίες και οι δράσεις της στρατηγικής της ΕΛΣΤΑΤ για τη Στατιστική Παιδεία αναρτώνται στην ειδική ιστοσελίδα:

Στατιστική Παιδεία

Hellenic Statistical Authority

@StatisticsGR

ELSTAT

statistics.gr

1. ΕΡΕΥΝΑ ΟΙΚΟΓΕΝΕΙΑΚΩΝ ΠΡΟΫΠΟΛΟΓΙΣΜΩΝ ΚΑΙ ΔΕΙΓΜΑΤΟΛΗΨΙΑ

Η «Έρευνα Οικογενειακών Προϋπολογισμών» (ΕΟΠ) αποτελεί μια στατιστική έρευνα που πραγματοποιείται από την ΕΛΣΤΑΤ. Μέσω αυτής, συλλέγονται πληροφορίες από αντιπροσωπευτικό δείγμα νοικοκυριών που επικεντρώνονται στη σύνθεση των νοικοκυριών, την απασχόληση των μελών τους, τις συνθήκες στέγασης και, κυρίως, στις δαπάνες διαβίωσής τους, καθώς και στα εισοδήματά τους.

Οι πληροφορίες που συλλέγονται για τις δαπάνες των νοικοκυριών είναι λεπτομερείς. Αυτό σημαίνει ότι δεν περιορίζονται μόνο στις βασικές κατηγορίες δαπανών, όπως τρόφιμα, ένδυση και υπόδηση, υγεία κ.λπ. Αντίθετα, παρέχονται λεπτομερείς πληροφορίες για κάθε επιμέρους κατηγορία, όπως λευκό ψωμί, φρέσκο γάλα, νωπό μοσχαρίσιο κρέας (στην κατηγορία τροφίμων), ανδρική και γυναικεία υπόδηση (στην κατηγορία ένδυσης και υπόδησης), φαρμακευτικά προϊόντα, υπηρεσίες εργαστηρίων ιατρικών αναλύσεων (στην κατηγορία υγείας) και άλλα παρόμοια.

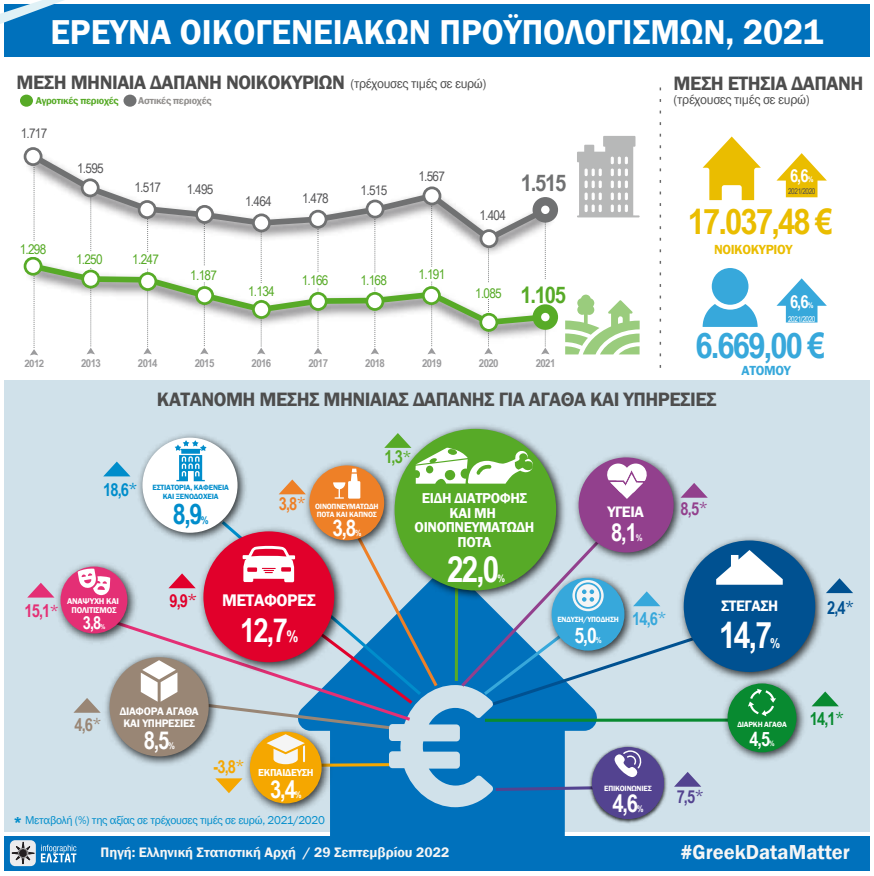
Τα αποτελέσματα της ΕΟΠ χρησιμοποιούνται τόσο στην κατανόηση της συμπεριφοράς των νοικοκυριών της Χώρας όσο και στην εκτίμηση του Ακαθάριστου Εγχώριου Προϊόντος (ΑΕΠ) και τον προσδιορισμό των σταθμίσεων στο «καλάθι του νοικοκυριού» για την εκτίμηση του Δείκτη Τιμών Καταναλωτή και, κατά συνέπεια, του πληθωρισμού.

1.1 Infographic

Το έτος 2021 είναι η χρονική περίοδος στην οποία αναφέρονται τα στοιχεία. Στις επίσημες στατιστικές η περίοδος αναφοράς των στοιχείων συνήθως είναι: μήνας (π.χ. Ιανουάριος 2023) ή τρίμηνο (π.χ. Β΄ Τρίμηνο 2023) ή έτος (π.χ. 2023).

Τα αποτελέσματα της ΕΟΠ 2021 δημοσιεύτηκαν με ανακοίνωση της ΕΛΣΤΑΤ, στις 29 Σεπτεμβρίου 2022. Επιπλέον, την ίδια ημέρα η ΕΛΣΤΑΤ παρουσίασε, μέσω της ιστοσελίδας της www.statistics.gr και των λογαριασμών της στα μέσα κοινωνικής δικτύωσης, το παρακάτω infographic με πολύ βασικά αποτελέσματα της έρευνας:

Infographic λέμε την οπτική αναπαράσταση πληροφοριών με την ελάχιστη χρήση κειμένου.



1.1.α

Χρησιμοποιείστε τις πληροφορίες που αποτυπώνονται στο infographic και κυκλώστε τη σωστή απάντηση στις ερωτήσεις που ακολουθούν:

i. Κατά το 2021 η μέση ετήσια δαπάνη των νοικοκυριών ήταν περίπου:

Α. 1.515 ευρώ στις αστικές περιοχές και 1.105 στις αγροτικές περιοχές

Β. 17.037 ευρώ

Γ. 6.669 ευρώ

ii. Κατά τη δεκαετία 2012 - 2021 η μέση μηνιαία δαπάνη των νοικοκυριών στις αστικές περιοχές, σε σχέση με τις αγροτικές περιοχές, ήταν διαχρονικά:

Α. περίπου ίδια

Β. μεγαλύτερη

Γ. μικρότερη

iii. Κατά το 2021 οι περισσότερες δαπάνες των νοικοκυριών αφορούσαν σε:

Α. Μεταφορές

Β. Είδη διατροφής και μη οινοπνευματώδη ποτά

Γ. Στέγαση, ύδρευση, καύσιμα και φωτισμό κύριας και δευτερεύουσας ή εξοχικής κατοικίας

Πώς σημειώνεται συνοπτικά αυτή η κατηγορία στο infographic;

Το 2020 υπήρξαν ιδιαίτερες συνθήκες (ποιες;), με μεγάλη επίδραση στη λειτουργία κάποιων υπηρεσιών (ποιων;).

iv. Κατά το 2021 το ποσοστό (%) της μέσης μηνιαίας δαπάνης των νοικοκυριών της Χώρας για «Υγεία» ήταν:

Α. 8,1%

Β. 8,5%

Γ. 6,6%

v. Κατά το 2021, συγκριτικά με το 2020, οι δαπάνες των νοικοκυριών που κατέγραψαν τη μεγαλύτερη ποσοστιαία (%) αύξηση αφορούσαν σε:

Α. Είδη διατροφής και μη οινοπνευματώδη ποτά

Β. Εστιατόρια, καφενεία και ξενοδοχεία

Γ. Αναψυχή και πολιτισμό

vi. Κατά το 2021, συγκριτικά με το 2020, μείωση δαπανών καταγράφεται στην κατηγορία:

Α. Εκπαίδευση

Β. Είδη ένδυσης και υπόδησης

Γ. Διαρκή αγαθά οικιακής χρήσης - οικιακά είδη άμεσης κατανάλωσης και οικιακές υπηρεσίες

Πώς σημειώνεται συνοπτικά αυτή η κατηγορία στο infographic;

1.2 Μεταδεδομένα

1.2.α

Εντοπίστε στην ιστοσελίδα της ΕΛΣΤΑΤ www.statistics.gr την τοποθεσία όπου είναι αναρτημένοι οι πίνακες αποτελεσμάτων και μεθοδολογικά κείμενα της ΕΟΠ 2021. Στη συνέχεια, συμβουλευτείτε την «Ενότητα 3: Στατιστική παρουσίαση»[1] των μεταδεδομένων της έρευνας, για να συμπληρώσετε τα κενά στις προτάσεις που ακολουθούν:

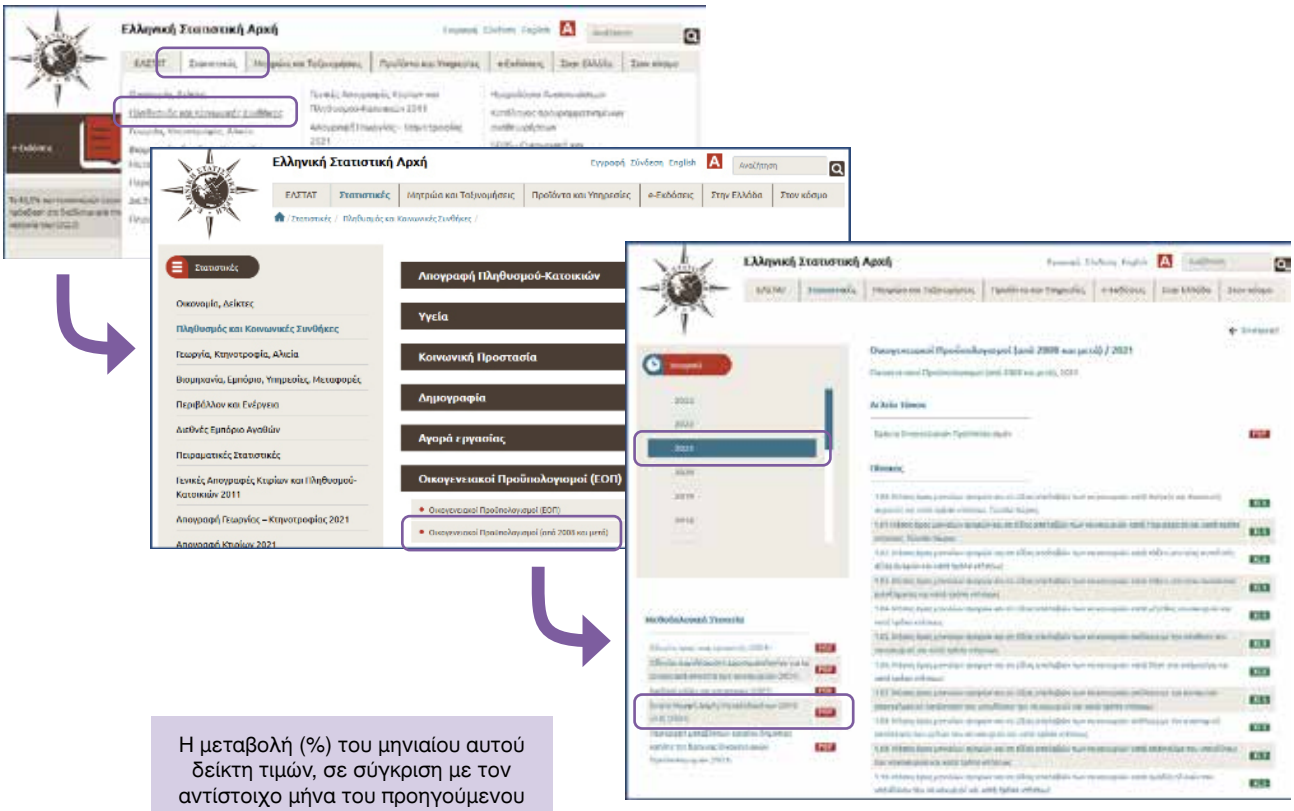
Μαθαίνω

Μεταδεδομένα μιας έρευνας είναι οι πληροφορίες που χρειαζόμαστε για να χρησιμοποιήσουμε και να επεξηγήσουμε τις στατιστικές.

Τα μεταδεδομένα περιγράφουν στατιστικά δεδομένα, δίνοντας τους ορισμούς των πληθυσμών, των μεταβλητών, καθώς και πληροφορίες για τη μεθοδολογία και την ποιότητα των μεθόδων και των αποτελεσμάτων των ερευνών.

Η Eurostat και το Ευρωπαϊκό Στατιστικό Σύστημα (ΕΣΣ) ανέπτυξαν την Ενιαία Ολοκληρωμένη Δομή Μεταδεδομένων (Single Integrated Metadata Structure - SIMS) για την τυποποίηση της ανταλλαγής και της διάδοσης εναρμονισμένων μεταδεδομένων αναφοράς.

Πώς βρίσκω τα αρχεία με τα μεταδεδομένα της ΕΟΠ 2021 στην ιστοσελίδα www.statistics.gr;



Η μεταβολή (%) του μηνιαίου αυτού δείκτη τιμών, σε σύγκριση με τον αντίστοιχο μήνα του προηγούμενου έτους, ονομάζεται «πληθωρισμός».

1. Βασικός σκοπός της ΕΟΠ είναι ο προσδιορισμός του καταναλωτικού προτύπου των νοικοκυριών για την αναθεώρηση (ενημέρωση των σταθμίσεων) του δείκτη τιμών .
2. Ως ορίζεται ένα άτομο που ζει μόνο του σε μία κατοικία ή μία ομάδα ατόμων συγγενικών ή μη, τα οποία διαμένουν στην ίδια κατοικία.
3. Στην έρευνα συμμετέχουν περίπου νοικοκυριά.
4. Η πρώτη ΕΟΠ στην Ελλάδα διενεργήθηκε κατά τα έτη .
5. Από την έρευνα εξαιρούνται οι κάθε είδους , όπως οικοτροφεία, νοσοκομεία κ.λπ.

(1) Στο Παράρτημα Ι, σελ. 72, παρατίθενται όλα τα αποσπάσματα των μεταδεδομένων της ΕΟΠ 2021, στα οποία γίνεται αναφορά στην Ενότητα αυτή.

1.3 Δειγματοληπτικές έρευνες και τρόποι δειγματοληψίας

Η ΕΟΠ είναι μια έρευνα που εξετάζει τις δαπάνες, τα εισοδήματα και τα χαρακτηριστικά των νοικοκυριών της Χώρας.

Ο πληθυσμός μιας έρευνας αποτελείται από όλες τις μονάδες (άτομα, νοικοκυριά, επιχειρήσεις κ.λπ.), τις οποίες ενδιαφερόμαστε να εξετάσουμε ως προς ένα ή περισσότερα χαρακτηριστικά.

Ο πληθυσμός της, δηλαδή, είναι το σύνολο όλων των νοικοκυριών της Χώρας και των ατόμων που διαμένουν σε αυτά.

Στην ΕΟΠ, όμως, δεν ερευνάται όλος ο πληθυσμός, αλλά μόνο ένα υποσύνολό του, το οποίο ονομάζουμε δείγμα.

Η έρευνα που διεξάγεται σε ένα δείγμα του πληθυσμού ονομάζεται δειγματοληπτική. Η έρευνα που διεξάγεται σε όλο τον πληθυσμό ονομάζεται απογραφική.

Η ΕΟΠ βασίζεται σε δισταδιακή στρωματοποιημένη δειγματοληψία νοικοκυριών από πλαίσιο δειγματοληψίας, που έχει δημιουργηθεί με βάση τα στοιχεία για τον μόνιμο πληθυσμό, από την πιο πρόσφατη Απογραφή Πληθυσμού - Κατοικιών.

Η διαδικασία επιλογής του δείγματος ονομάζεται δειγματοληψία.

Το μητρώο - κατάλογος των μονάδων του πληθυσμού της έρευνας, το οποίο χρησιμοποιούμε για να κάνουμε την επιλογή του δείγματος, ονομάζεται πλαίσιο δειγματοληψίας.

Από τα στοιχεία που συλλέγονται για το δείγμα των νοικοκυριών, εξάγονται αποτελέσματα / συμπεράσματα για το σύνολο των νοικοκυριών της Χώρας.

Μαθαίνω

Πληθυσμός είναι το σύνολο των μονάδων (ατόμων, νοικοκυριών, επιχειρήσεων κ.λπ.) που ενδιαφερόμαστε να εξετάσουμε ως προς ένα ή περισσότερα χαρακτηριστικά (π.χ. ηλικία, εισόδημα, κύκλος εργασιών).

Δείγμα είναι το υποσύνολο του πληθυσμού που μελετάται, προκειμένου να εξαχθούν συμπεράσματα / αποτελέσματα για το σύνολο του πληθυσμού.

Δειγματοληψία ονομάζουμε τη διαδικασία επιλογής των μονάδων του δείγματος (ατόμων, νοικοκυριών, επιχειρήσεων κ.λπ.) από τον πληθυσμό.

Δειγματοληπτικό πλαίσιο ή **πλαίσιο δειγματοληψίας** ονομάζεται το μητρώο - κατάλογος των μονάδων του πληθυσμού, με τη βοήθεια του οποίου γίνεται η επιλογή του δείγματος.

Δειγματοληπτική μονάδα ονομάζεται η μονάδα (στοιχείο ή συλλογή στοιχείων) που μπορεί να επιλεγεί σε κάποιο στάδιο της δειγματοληψίας.

Απογραφική έρευνα ονομάζεται η έρευνα που διεξάγεται στο σύνολο του πληθυσμού.

Δειγματοληπτική έρευνα ονομάζεται η έρευνα που διεξάγεται σε δείγμα του πληθυσμού.

1.3.α

Διαβάστε τις πληροφορίες που συνοδεύουν τις ακόλουθες πέντε έρευνες και αποφασίστε αν πρόκειται για δειγματοληπτική (Δ) ή απογραφική έρευνα (Α).

i. __

Η **Απογραφή Γεωργίας - Κτηνοτροφίας** διενεργείται κάθε δέκα χρόνια σε όλα τα κράτη μέλη της Ευρωπαϊκής Ένωσης, με σκοπό τη συλλογή στοιχείων αναφορικά με τη διάρθρωση των γεωργικών και κτηνοτροφικών εκμεταλλεύσεων και τα ειδικά χαρακτηριστικά τους, καθώς και την απασχόληση του αγροτικού πληθυσμού σε αυτές. Για τον σκοπό αυτόν, συμπληρώνεται ειδικό ερωτηματολόγιο από κάθε γεωργική ή κτηνοτροφική εκμετάλλευση.

ii. __

Η **Έρευνα Χρήσης Τεχνολογιών Πληροφόρησης, Επικοινωνίας και Ηλεκτρονικού Εμπορίου στις Επιχειρήσεις** καλύπτει όλες τις επιχειρήσεις της Χώρας, με απασχόληση δέκα άτομα και άνω, που ανήκουν στους διψήφιους κλάδους οικονομικής δραστηριότητας (NACE Rev.2)* 10 - 63, 68 - 82 και 95.1.

Στην έρευνα εφαρμόζεται μονοσταδιακή στρωματοποιημένη δειγματοληψία, με μονάδα έρευνας την επιχείρηση που απασχολεί δέκα άτομα και άνω. Η στρωματοποίηση αυτή γίνεται βάσει:

- Περιφέρειας (NUTS 2)**,
- ομάδων των κλάδων οικονομικής δραστηριότητας,
- τάξης μεγέθους απασχόλησης της επιχείρησης.

* Η Στατιστική Ταξινόμηση των Οικονομικών Δραστηριοτήτων στην Ευρωπαϊκή Κοινότητα είναι γνωστή με τη συντομογραφία NACE. Ο όρος NACE προέρχεται από τη γαλλική έκφραση «Nomenclature statistique des activités économiques dans la Communauté européenne».

** Η κοινή ονοματολογία των εδαφικών στατιστικών μονάδων (NUTS) (από το γαλλικό: Nomenclature d' Unités Territoriales Statistiques) είναι το γεωκωδικό πρότυπο της Ευρωπαϊκής Ένωσης για την κωδικοποίηση της διοικητικής διαίρεσης κάθε περιοχής, για στατιστικούς λόγους. Οι περιοχές δε μεταφράζονται, αλλά αποτυπώνονται όπως είναι στη γλώσσα της κάθε χώρας.

iii. __

Η **Απογραφή Πληθυσμού - Κατοικιών** αποσκοπεί στην καταγραφή του συνόλου του μόνιμου και του νόμιμου πληθυσμού της Χώρας, των κατοικιών κάθε είδους και των χαρακτηριστικών τους, καθώς και των μεταναστευτικών ροών από και προς την Ελλάδα.

2021
Απογραφή
Πληθυσμού - Κατοικιών
Ελληνική Στατιστική Αρχή

iv. __

Η **Έρευνα Οδικών Τροχαίων Ατυχημάτων** διενεργείται σε μηνιαία βάση και παρακολουθεί, κατά Περιφέρεια και Νομό, τον αριθμό όλων των οδικών τροχαίων ατυχημάτων κατά βαρύτητα (θανατηφόρα και/ή με τραυματισμούς) και τον αριθμό των παθόντων ατόμων κατά κατηγορία αυτών (οδηγοί, μεταφερόμενοι, πεζοί).
Η συλλογή των στοιχείων γίνεται από τις κατά τόπους Αστυνομικές και Λιμενικές Αρχές.

v. __

Η **Έρευνα Εισοδήματος και Συνθηκών Διαβίωσης των Νοικοκυριών** αποτελεί τη βασική πηγή αναφοράς συγκριτικών στατιστικών για την κατανομή του εισοδήματος και τον κοινωνικό αποκλεισμό, σε ευρωπαϊκό επίπεδο. Η συγκρισιμότητα των στοιχείων θεωρείται εξασφαλισμένη, αφού η έρευνα διενεργείται σε όλα τα κράτη μέλη, χρησιμοποιώντας κοινές μεταβλητές και ορισμούς. Κατά το έτος 2021 η έρευνα διενεργήθηκε σε τελικό δείγμα 12.617 νοικοκυριών και σε 27.710 μέλη των νοικοκυριών αυτών, εκ των οποίων 24.333 ηλικίας 16 ετών και άνω.

1.3.6

Αντιστοιχίστε την περιγραφή κάθε τρόπου δειγματοληψίας της στήλης Α με το όνομα του τρόπου δειγματοληψίας στη στήλη Β.

Στήλη Α Περιγραφή τρόπου δειγματοληψίας	**Στήλη Β** Όνομα τρόπου δειγματοληψίας
1. Οι δειγματοληπτικές μονάδες επιλέγονται από τον πληθυσμό, με κριτήριο την ευκολία πρόσβασης σε αυτές.	**i.** Απλή τυχαία δειγματοληψία
2. Κάθε μονάδα ενός αρχικού δείγματος προσκαλεί σε συμμετοχή στην έρευνα (και συμπερίληψη στο δείγμα) άλλες μονάδες, που με τη σειρά τους προσκαλούν κάποιες άλλες μονάδες κ.ο.κ.	**ii.** Συστηματική δειγματοληψία
3. Κάθε μονάδα του δείγματος επιλέγεται από τον πληθυσμό με τυχαίο τρόπο, δηλαδή έχει ίση πιθανότητα να επιλεγεί.	**iii.** Δειγματοληψία ευκολίας
4. Ξεκινώντας από μία τυχαία επιλεγμένη μονάδα, οι υπόλοιπες μονάδες του δείγματος επιλέγονται από τον πληθυσμό ανά ίσα, διαδοχικά διαστήματα, δηλαδή με συστηματικό τρόπο.	**iv.** Δειγματοληψία χιονοστιβάδας
5. Τα στοιχεία του πληθυσμού διαιρούνται σε ομοιογενείς υποπληθυσμούς ή στρώματα (strata), στη βάση κάποιου σημαντικού χαρακτηριστικού. Η επιλογή του δείγματος γίνεται χωριστά σε κάθε στρώμα, με απλή τυχαία ή συστηματική δειγματοληψία.	**v.** Στρωματοποιημένη δειγματοληψία
6. Οι μονάδες του πληθυσμού διαιρούνται σε συστάδες (clusters) και στη συνέχεια, με απλή τυχαία δειγματοληψία, επιλέγεται να διερευνηθεί ένα δείγμα από τις συστάδες αυτές. Από τις επιλεγμένες συστάδες ερευνάται το σύνολο των μονάδων που περιέχουν.	**vi.** Δειγματοληψία κατά συστάδες

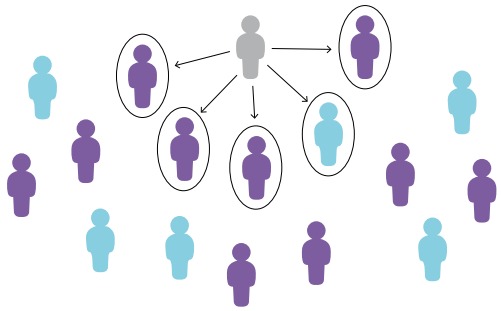
Ω, υπάρχουν τόσοι διαφορετικοί τρόποι δειγματοληψίας!

Και δεν είναι μόνο αυτοί! Υπάρχουν και άλλοι που δεν έχουμε αναφέρει. Επιπλέον, οι τρόποι αυτοί μπορούν να συνδυαστούν μεταξύ τους σε δύο ή περισσότερα στάδια δειγματοληψίας και να προκύψουν νέοι. Κατά τον σχεδιασμό μιας έρευνας, διερευνούμε ποιος είναι ο κατάλληλος τρόπος δειγματοληψίας για την περίπτωσή μας.

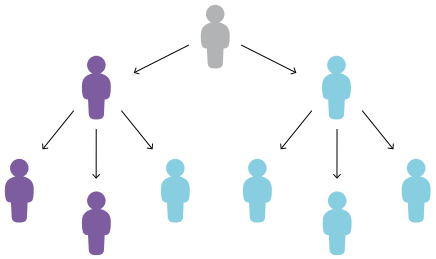
1.3.γ

Συμπληρώστε το όνομα του τρόπου δειγματοληψίας που ταιριάζει σε καθεμία από τις ακόλουθες εικόνες:

Α.

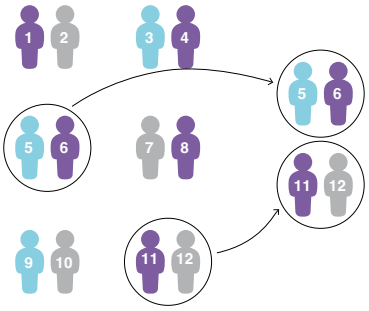


Β.

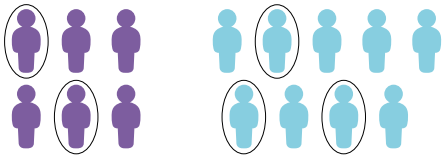


Γ.

Δ.



Ε.



Υπόδειξη: Για την περιγραφή και το όνομα των διαφόρων τρόπων δειγματοληψίας, συμβουλευτείτε τη δραστηριότητα 1.3.β.

Και μπορούμε να εξάγουμε συμπεράσματα για όλο τον πληθυσμό, ρωτώντας μόνο ένα δείγμα;

Ναι, με μία βασική προϋπόθεση: πρέπει το δείγμα μας να είναι δείγμα πιθανότητας, δηλαδή κάθε μονάδα του πληθυσμού να έχει μία καθορισμένη (γνωστή) μη μηδενική πιθανότητα επιλογής στο δείγμα.

Νοικοκυριό, άτομο, επιχείρηση κ.λπ.

Και γιατί είναι τόσο σημαντικό να γνωρίζουμε την πιθανότητα με την οποία κάθε μονάδα του πληθυσμού μπορεί να επιλεγεί στο δείγμα;

Επειδή με βάση την πιθανότητα αυτή υπολογίζουμε πόσες μονάδες του πληθυσμού αντιπροσωπεύει η κάθε μονάδα του δείγματος, ώστε να μπορέσουμε να κάνουμε την αναγωγή των αποτελεσμάτων μας από το δείγμα στον πληθυσμό. Θα το συζητήσουμε με περισσότερη λεπτομέρεια παρακάτω, όταν θα μιλήσουμε για τους αναγωγικούς συντελεστές.

Μαθαίνω

Δείγμα πιθανότητας ονομάζουμε το δείγμα που έχει επιλεγεί με μεθόδους πιθανότητας, δηλαδή, όταν η κάθε μονάδα του πληθυσμού έχει μία καθορισμένη (μη μηδενική) πιθανότητα να συμπεριληφθεί στο δείγμα. Όταν έχουμε δείγμα πιθανότητας, μπορούμε να εφαρμόσουμε στατιστικές μεθόδους και τεχνικές για να εξάγουμε συμπεράσματα / αποτελέσματα από το δείγμα, για το σύνολο του πληθυσμού.

1.3.δ

Στην παρακάτω λίστα τρόπων δειγματοληψίας, εντοπίστε και κυκλώστε τους τρόπους δειγματοληψίας που βασίζονται σε μεθόδους πιθανότητας:

Απλή τυχαία δειγματοληψία	Δειγματοληψία κατά συστάδες
Δειγματοληψία χιονοστιβάδας	Στρωματοποιημένη δειγματοληψία
Δειγματοληψία ευκολίας	Συστηματική δειγματοληψία

Υπόδειξη: Για την περιγραφή και το όνομα των διαφόρων τρόπων δειγματοληψίας, συμβουλευτείτε τη δραστηριότητα 1.3.β.

1.4 Δειγματοληπτικός σχεδιασμός της ΕΟΠ

1.4.α

Συμβουλευτείτε την «Ενότητα 18.1: Τύπος πρωτογενών δεδομένων»[2] των μεταδεδομένων της ΕΟΠ 2021, για να συμπληρώσετε τα κενά στις προτάσεις και για να απαντήσετε στις ερωτήσεις σχετικά με τον δειγματοληπτικό σχεδιασμό της έρευνας.

i. Πού βασίζεται το πλαίσιο δειγματοληψίας της ΕΟΠ;

Θυμάμαι

Το πλαίσιο δειγματοληψίας, δηλαδή ο κατάλογος των μονάδων του πληθυσμού που χρησιμοποιούμε για να επιλέξουμε το δείγμα μας, θέλουμε να:

- καλύπτει όλο τον πληθυσμό,
- περιλαμβάνει κάθε μονάδα του πληθυσμού ακριβώς μία φορά.

Προκειμένου να έχει τις παραπάνω ιδιότητες, θα πρέπει ο κατάλογος αυτός να είναι ενημερωμένος.

ii. Ποιος τρόπος δειγματοληψίας εφαρμόζεται στην ΕΟΠ;

iii. Με βάση ποια χαρακτηριστικά (κριτήρια στρωμάτωσης) διαιρείται ο πληθυσμός (νοικοκυριά) σε στρώματα (υποπληθυσμούς);

1.
2.

Τα χαρακτηριστικά αυτά αφορούν στον τόπο μόνιμης κατοικίας των νοικοκυριών.

iv. Τι είναι **«βαθμός αστικότητας»**;

Ο βαθμός αστικότητας χαρακτηρίζει μια Κοινότητα με βάση τον πληθυσμό της. Διακρίνονται 3 βαθμοί αστικότητας:

1. Αστικές περιοχές: πληθυσμός ≥
2. Ημιαστικές περιοχές: ≤ πληθυσμός ≤
3. Αγροτικές περιοχές: πληθυσμός ≤

Συμμετέχω στην Απογραφή Πληθυσμού - Κατοικιών

Εκπροσωπούμαι στις επίσημες στατιστικές που αφορούν σε νοικοκυριά και άτομα που διαμένουν σε αυτά

Αυτό το γνώριζες;

Με την Απογραφή Κτιρίων επικαιροποιείται ο κατάλογος των οικοδομικών τετραγώνων και ο χάρτης της Χώρας, δηλαδή ενσωματώνονται τυχόν αλλαγές (δημιουργίες, συγχωνεύσεις ή διασπάσεις τετραγώνων) σε σχέση με την προηγούμενη Απογραφή. Τα οικοδομικά τετράγωνα, μεμονομένα ή ομαδοποιημένα, δημιουργούν τις μονάδες επιφανείας. Με την Απογραφή Πληθυσμού - Κατοικιών προκύπτει το πλήθος των νοικοκυριών - κατοικιών κάθε οικοδομικού τετραγώνου.

Για τις μονάδες επιφανείας που περιλαμβάνονται στα δείγματα των ερευνών, ερευνητές επισκέπτονται την κάθε μονάδα επιφανείας και καταγράφουν τα υπάρχοντα νοικοκυριά. Με αυτόν τον τρόπο, δημιουργείται ο επικαιροποιημένος κατάλογος νοικοκυριών που διαμένουν στα οικοδομικά τετράγωνα, τα οποία περιλαμβάνονται στα δείγματα.

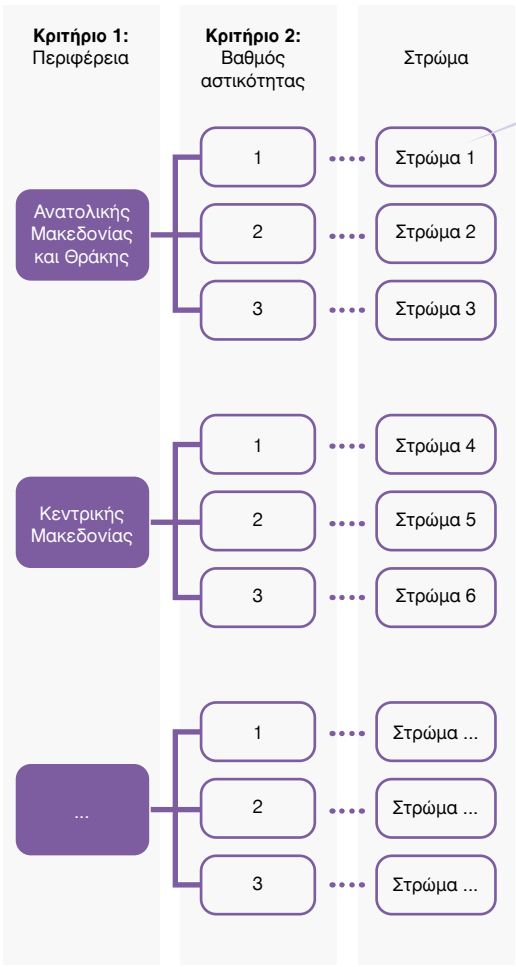
(2) Στο Παράρτημα Ι, σελ. 72, παρατίθενται όλα τα αποσπάσματα των μεταδεδομένων της ΕΟΠ 2021, στα οποία γίνεται αναφορά στην Ενότητα αυτή.

Ας δούμε αναλυτικά τον δειγματοληπτικό σχεδιασμό της ΕΟΠ.

Αρχικά γίνεται ο διαμερισμός του πληθυσμού της έρευνας (νοικοκυριά) σε στρώματα που έχουν κοινά χαρακτηριστικά (Περιφέρεια και βαθμός αστικότητας).

Στην ΕΟΠ 2021 δημιουργήθηκαν (v) στρώματα.

Στρωμάτωση



Περιλαμβάνει τα νοικοκυριά που διαμένουν σε αστικές περιοχές (με πληθυσμό ≥ 10.000) της Περιφέρειας Ανατολικής Μακεδονίας και Θράκης.

Μονάδα επιφανείας είναι ένας όρος που χρησιμοποιείται από την ΕΛΣΤΑΤ για να περιγράψει μια εδαφική περιοχή που περιλαμβάνει ένα ή περισσότερα (συνήθως γειτονικά) οικοδομικά τετράγωνα ενός Οικισμού. Οι μονάδες επιφανείας έχουν συνήθως παρόμοιο πλήθος κατοικιών - νοικοκυριών και διακριτά μη επικαλυπτόμενα όρια.

Παρατίθενται παρακάτω υποδείγματα μονάδων επιφανείας του Κεντρικού Τομέα Αθηνών. Το υπόδειγμα στα αριστερά περιλαμβάνει 4 γειτονικά οικοδομικά (γραμμοσκιασμένα) τετράγωνα, ενώ το υπόδειγμα στα δεξιά περιλαμβάνει 1 οικοδομικό τετράγωνο (γραμμοσκιασμένο).

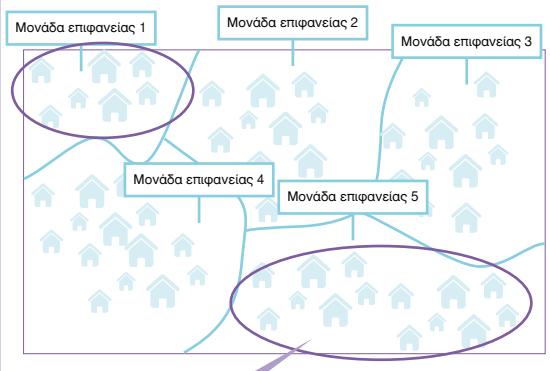
Μετά τη στρωμάτωση ακολουθεί η δειγματοληψία σε κάθε (vi), η οποία πραγματοποιείται σε δύο στάδια.

Αυτό το γνώριζες;

Στην ΕΛΣΤΑΤ υπάρχει **Τμήμα Χαρτογραφίας και Γεωχωρικών Δεδομένων** για την υποστήριξη των στατιστικών εργασιών, μέσω της παροχής του αναγκαίου χαρτογραφικού υλικού για την διεξαγωγή της εκάστοτε Γενικής Απογραφής Πληθυσμού - Κατοικιών και των ειδικών ερευνών και απογραφών.

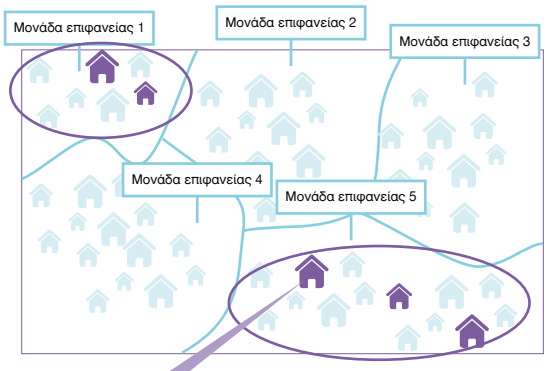
Στάδιο 1	Στάδιο 2
Στο 1ο στάδιο δειγματοληψίας επιλέγονται από κάθε στρώμα οι πρωτογενείς μονάδες επιφανείας που θα διερευνηθούν. Η πιθανότητα (vii) μίας μονάδας επιφανείας είναι ανάλογη του μεγέθους της (πλήθος νοικοκυριών κατά την τελευταία Απογραφή).	Στο 2ο στάδιο δειγματοληψίας, σε κάθε επιλεγείσα μονάδα επιφανείας του 1ου σταδίου, εφαρμόζεται (viii) δειγματοληψία και επιλέγεται δείγμα (ix).

Στάδιο 1: Επιλογή μονάδων επιφανείας



Επιλεγμένη μονάδα επιφανείας

Στάδιο 2: Επιλογή νοικοκυριών



Επιλεγμένο νοικοκυριό

Η πιθανότητα επιλογής p_i του νοικοκυριού i είναι ίση με το γινόμενο της πιθανότητας επιλογής της μονάδας επιφανείας, στην οποία περιλαμβάνεται το νοικοκυριό (στάδιο 1), επί την πιθανότητα επιλογής του νοικοκυριού εντός της μονάδας επιφανείας (στάδιο 2).

Ας δούμε τώρα ένα απλοποιημένο παράδειγμα συστηματικής δειγματοληψίας. Σε μία μονάδα επιφανείας υπάρχουν 28 (N_1) νοικοκυριά και θα επιλέξουμε δείγμα μεγέθους $n_1 = 4$ με συστηματική δειγματοληψία.

Βήμα 1: Αριθμούμε τα νοικοκυριά της μονάδας επιφανείας από το 1 ως και το 28.

Βήμα 2: Προσδιορίζουμε το **διάστημα δειγματοληψίας** k, διαιρώντας το πλήθος όλων των νοικοκυριών με το μέγεθος του δείγματος, δηλαδή:

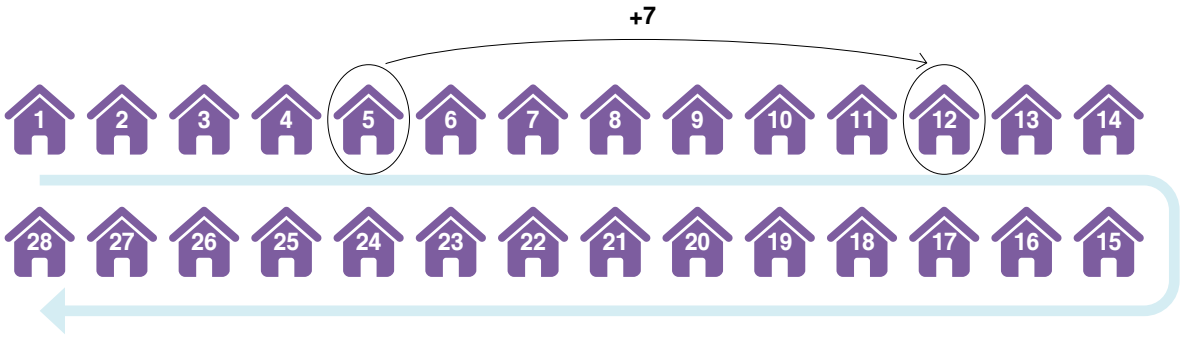
διάστημα δειγματοληψίας $k = \frac{N_1}{n_1} = \frac{28}{4} = 7$

Βήμα 3: Επιλέγουμε έναν τυχαίο ακέραιο αριθμό (με κλήρωση) από το 1 έως και το k, δηλαδή στην περίπτωσή μας από το 1 έως και το 7. Το νοικοκυριό που έχει αυτόν τον αύξοντα αριθμό είναι το πρώτο νοικοκυριό του δείγματός μας και το σημείο εκκίνησης για την επιλογή των υπόλοιπων.

Βήμα 4: Στη συνέχεια, ξεκινώντας από αυτόν τον αριθμό, προσθέτουμε κάθε φορά το διάστημα δειγματοληψίας για να βρούμε τον επόμενο αύξοντα αριθμό του νοικοκυριού που θα συμπεριλάβουμε στο δείγμα μας.

1.4.6

Αν στο βήμα 3 κληρώθηκε το νοικοκυριό 5, κυκλώστε τα νοικοκυριά που θα συμπεριληφθούν στο δείγμα.

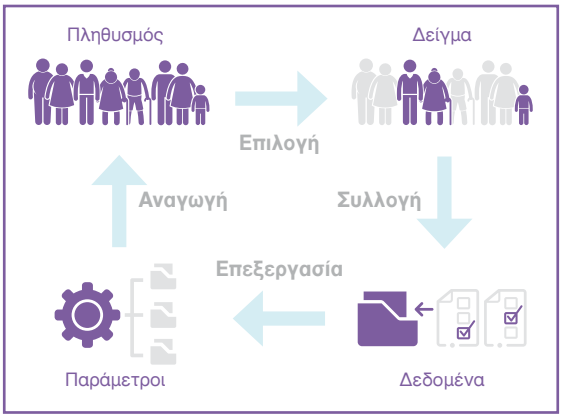


1.5 Αναγωγή αποτελεσμάτων στον πληθυσμό

Ωραία! Συγκεντρώνουμε τις πληροφορίες από το δείγμα πιθανότητας και έχουμε τα στοιχεία της έρευνας. Εμείς, όμως, θέλουμε να βγάλουμε συμπεράσματα για όλο τον πληθυσμό. Πώς θα το κάνουμε αυτό; Δηλαδή πώς θα ανάγουμε στον πληθυσμό τα δεδομένα του δείγματος;

Θα προσπαθήσω να περιγράψω απλά και συνοπτικά τη διαδικασία, αλλά θα χρειαστεί να κάνετε και την εφαρμογή που ακολουθεί (1.8.β) για να την καταλάβετε καλύτερα.

- Για κάθε μονάδα i του δείγματος πιθανότητας γνωρίζουμε την πιθανότητα επιλογής p_i.
- Ο αντίστροφος της πιθανότητας επιλογής $w_i = \frac{1}{p_i}$ ονομάζεται **αναγωγικός συντελεστής**.
- Ο αναγωγικός συντελεστής δηλώνει πόσες μονάδες στον πληθυσμό αντιπροσωπεύονται από την εκάστοτε μονάδα του δείγματος.



- Επομένως, τα χαρακτηριστικά / στοιχεία που συλλέχθηκαν για κάθε μονάδα i του δείγματος εκτιμούμε ότι υπάρχουν w_i φορές στον πληθυσμό.

Για παράδειγμα, ας υποθέσουμε ότι κάνουμε μία δειγματοληπτική έρευνα για να εξετάσουμε ποιο είναι το «αγαπημένο μάθημα» των μαθητών/τριών σε κάποιο σχολείο και ότι ένα από τα παιδιά στο δείγμα μας απάντησε «Ιστορία».

Αν το παιδί αυτό είχε πιθανότητα 0,1 να επιλεγεί στο δείγμα, τότε ο αναγωγικός συντελεστής του είναι 1/0,1=10, δηλαδή το εν λόγω παιδί αντιπροσωπεύει 10 παιδιά στον πληθυσμό (σχολείο).

Επομένως, η μία (1) απάντηση «Ιστορία» στο δείγμα ανάγεται σε 10 απαντήσεις «Ιστορία» στον πληθυσμό.

Με τον ίδιο τρόπο ανάγουμε στον πληθυσμό και τις απαντήσεις των υπόλοιπων παιδιών του δείγματος, ώστε να καταλήξουμε στις εκτιμήσεις μας για το αγαπημένο μάθημα των μαθητών/τριών στο σχολείο.

Μαθαίνω

Ο αναγωγικός συντελεστής υπολογίζεται ως ο αντίστροφος της πιθανότητας επιλογής της εκάστοτε μονάδας του δείγματος και δηλώνει πόσες μονάδες στον πληθυσμό αντιπροσωπεύει η μονάδα του δείγματος.
Το άθροισμα των αναγωγικών συντελεστών των n μονάδων του δείγματος αποτελεί την εκτίμηση της δειγματοληπτικής έρευνας για το μέγεθος N του πληθυσμού.
Αναμένουμε λοιπόν ότι $\sum_{i=1}^{n} w_i = N$.

Συντόμευση που παριστάνει το άθροισμα πολλών στοιχείων.

1.6 Δειγματοληπτικά σφάλματα

Στα μεταδεδομένα της ΕΟΠ διαβάζω ότι το δείγμα αντιστοιχεί περίπου στο 1,5‰ του συνόλου των νοικοκυριών της Χώρας. Τα αποτελέσματα που προκύπτουν από το δείγμα αυτό είναι ακριβή;

Όταν εκτιμούμε τις παραμέτρους ενός πληθυσμού με βάση τις πληροφορίες ενός μόνο δείγματος, τότε, αναπόφευκτα, κάνουμε κάποιο σφάλμα, το οποίο ονομάζουμε δειγματοληπτικό σφάλμα.

Πράγματι! Αν πάρουμε άλλο δείγμα με ίδιο μέγεθος και χωρίς καμία αλλαγή στον σχεδιασμό της έρευνας, μπορεί να προκύψουν διαφορετικές εκτιμήσεις των παραμέτρων του πληθυσμού!

Μαθαίνω

Η **πραγματική τιμή μιας παραμέτρου** του πληθυσμού (π.χ. μέση ηλικία, μέσος αριθμός ατόμων ανά νοικοκυριό, μέσο εισόδημα) μπορεί να προσδιοριστεί, συλλέγοντας στοιχεία από ολόκληρο τον πληθυσμό (απογραφή).

Η **εκτίμηση της τιμής μιας παραμέτρου** του πληθυσμού γίνεται, όταν, με βάση τα στοιχεία ενός δείγματος, εκτιμούμε την τιμή της παραμέτρου αυτής.

Δειγματοληπτικό σφάλμα ή σφάλμα δειγματοληψίας ονομάζεται η διαφορά μεταξύ της εκτίμησης που βασίζεται σε ένα δείγμα και της πραγματικής τιμής της παραμέτρου.

Τα δειγματοληπτικά σφάλματα είναι στη φύση των δειγματοληπτικών ερευνών.

Τα καλά νέα είναι ότι στη Στατιστική, και εφόσον πρόκειται για δείγματα πιθανότητας, διαθέτουμε διάφορα μέτρα τα οποία ποσοτικοποιούν τα δειγματοληπτικά σφάλματα και έτσι μπορούμε να έχουμε μια καλή ιδέα για την ακρίβεια των αποτελεσμάτων μας!

Απλοποιημένο παράδειγμα παρουσίασης και ερμηνείας δειγματοληπτικού σφάλματος

Εκτιμήθηκε ότι η μέση ετήσια δαπάνη (αγορές) των νοικοκυριών για είδη διατροφής και μη οινοπνευματώδη ποτά είναι 3.742 ± 73 ευρώ, δηλαδή από 3.669 έως 3.815 ευρώ, σε επίπεδο εμπιστοσύνης 95%.

Η εκτίμηση της τιμής της παραμέτρου στον πληθυσμό, με βάση τις παρατηρήσεις στο δείγμα.

Το μέτρο δειγματοληπτικού σφάλματος.

Το διάστημα τιμών εντός του οποίου εκτιμούμε ότι βρίσκεται η πραγματική τιμή της παραμέτρου.

Ονομάζεται διάστημα εμπιστοσύνης, καθώς αναφέρεται σε ένα συγκεκριμένο επίπεδο εμπιστοσύνης.

Το επίπεδο εμπιστοσύνης (%) αντιπροσωπεύει την πιθανότητα το διάστημα που υπολογίσαμε να περιέχει την πραγματική τιμή της παραμέτρου.

Διάστημα εμπιστοσύνης 95% σημαίνει ότι, αν επαναλαμβάναμε την δειγματοληψία πολλές φορές, τότε αναμένεται ότι 95 στις 100 φορές η πραγματική τιμή της παραμέτρου θα ήταν μέσα στο διάστημα που θα εκτιμούσαμε κάθε φορά.

Το βασικό χαρακτηριστικό των δειγματοληπτικών σφαλμάτων είναι ότι τείνουν να μειώνονται (όχι αναλογικά) με την αύξηση του μεγέθους του δείγματος.

Συνήθως στις δειγματοληπτικές έρευνες προσδιορίζουμε από την αρχή το επιθυμητό επίπεδο ακρίβειας των αποτελεσμάτων μας και στη συνέχεια επιλέγουμε δείγμα κατάλληλου μεγέθους για να το πετύχουμε.

Υπάρχει κάποιος τύπος για να υπολογίζω κάθε φορά το μέγεθος δείγματος που θα μου διασφαλίζει την επιθυμητή ακρίβεια;

Ο τύπος υπολογισμού του μεγέθους του δείγματος εξαρτάται από τη μέθοδο δειγματοληψίας που επιλέγεται, και άρα δεν υπάρχει ένας μοναδικός τύπος. Το μέγεθος του δείγματος, προκειμένου να πετύχουμε το επιθυμητό επίπεδο ακρίβειας, εξαρτάται από πολλούς παράγοντες, όπως:

ο τρόπος δειγματοληψίας

και οι μέθοδοι που θέλουμε να εφαρμόσουμε κατά την ανάλυση των στοιχείων.

η μεταβλητότητα των χαρακτηριστικών που μελετάμε στον πληθυσμό

Στη στρωματοποιημένη δειγματοληψία υπάρχει διαφορετική μεταβλητότητα σε κάθε στρώμα.

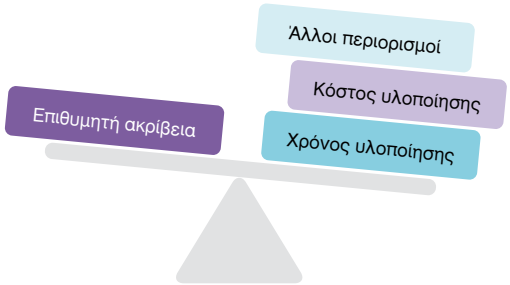
Μεγαλύτερη μεταβλητότητα απαιτεί μεγαλύτερο δείγμα για το ίδιο επίπεδο ακρίβειας.

το μέγεθος του πληθυσμού

Για μεγαλύτερο πληθυσμό απαιτείται γενικά μεγαλύτερο δείγμα, προκειμένου να πετύχουμε το επιθυμητό επίπεδο ακρίβειας.

Ωστόσο, από ένα σημείο και μετά, η αύξηση του μεγέθους του πληθυσμού δεν επηρεάζει σημαντικά το μέγεθος του δείγματος.

Στην πράξη, το μέγεθος του δείγματος αποτελεί έναν συμβιβασμό ανάμεσα στην επιθυμητή ακρίβεια των αποτελεσμάτων και σε περιορισμούς, όπως το κόστος και ο χρόνος υλοποίησης της έρευνας.



Στις έρευνες της ΕΛΣΤΑΤ, οι οποίες διεξάγονται για την παραγωγή των Επίσημων Στατιστικών της Χώρας, επιλέγονται δείγματα κατάλληλου μεγέθους, ώστε να έχουμε την επιθυμητή ακρίβεια αποτελεσμάτων με τη λιγότερη δυνατή επιβάρυνση σε κόστος, χρόνο υλοποίησης και συνολικό φόρτο ερευνωμένων.

Αυτό το γνώριζες;

Φόρτος ερευνωμένου είναι το βάρος που επωμίζεται ο ερευνώμενος εξαιτίας της συμμετοχής του στην έρευνα. Ο φόρτος ερευνωμένου μετριέται συνήθως με τον χρόνο που δαπάνησε για να απαντήσει στο ερωτηματολόγιο, συμπεριλαμβανομένου του χρόνου που χρειάστηκε για τη συγκέντρωση ή ανάκτηση των στοιχείων/πληροφοριών που του ζητήθηκαν.

1.6.α

Ετοίμασα ένα σταυρόλεξο για τους συμμαθητές και τις συμμαθήτριές μου. Δοκιμάστε να το λύσετε.

Υπόδειξη: όλες οι λέξεις είναι στην ονομαστική πτώση.

Οριζόντια

2. Το επιθυμητό επίπεδό της έχει καθοριστικό ρόλο στον προσδιορισμό του μεγέθους του δείγματος.
4. Περιγράφει τη διαφοροποίηση των τιμών ενός χαρακτηριστικού.
5. Το μέγεθός του παίζει ρόλο και στο μέγεθος του δείγματος.

Κάθετα

1. Ο τρόπος, με τον οποίο γίνεται, επηρεάζει το μέγεθος του δείγματος που εξασφαλίζει ένα συγκεκριμένο επίπεδο ακρίβειας.
3. Στην απόφασή μας για το μέγεθος του δείγματος μιας έρευνας ενδέχεται να ληφθούν υπόψη και περιορισμοί, όπως αυτός.

1.7 Μη δειγματοληπτικά σφάλματα

Σε μια έρευνα όμως, είτε είναι δειγματοληπτική είτε απογραφική, ενδέχεται να έχουμε, και πολύ συχνά έχουμε, και **μη δειγματοληπτικά σφάλματα**, τα οποία προσπαθούμε κάθε φορά να περιορίσουμε.

1.7.α

Στη στήλη Α παρακάτω θα βρείτε τις περιγραφές διαφόρων μη δειγματοληπτικών σφαλμάτων. Αντιστοιχίστε τις περιγραφές με το κατάλληλο όνομα από τη στήλη Β.

Στήλη Α Περιγραφή μη δειγματοληπτικών σφαλμάτων	**Στήλη Β** Όνομα σφαλμάτων
1. Σφάλματα που σχετίζονται με το δειγματοληπτικό πλαίσιο και τον βαθμό που αυτό καλύπτει ή όχι τον πληθυσμό.	**i.** Σφάλματα μη απόκρισης
2. Σφάλματα που σχετίζονται με αδυναμίες του ερωτηματολογίου ή του ερευνητή ή του ερευνωμένου. Λόγου χάρη, μια ασαφής ερώτηση δημιουργεί τέτοια σφάλματα.	**ii.** Σφάλματα κάλυψης
3. Σφάλματα που δημιουργούνται κατά τη διάρκεια της κωδικοποίησης, εισαγωγής και επεξεργασίας των στοιχείων της έρευνας.	**iii.** Σφάλματα επεξεργασίας
4. Σφάλματα που οφείλονται στην άρνηση/αδυναμία κάποιων μονάδων του δείγματος να απαντήσουν σε ορισμένα ερωτήματα ή να συμμετέχουν στην έρευνα.	**iv.** Σφάλματα μέτρησης

1.7.β

Ανατρέξτε στην «Ενότητα 13.3: Μη δειγματοληπτικά σφάλματα» των μεταδεδομένων της ΕΟΠ (Παράρτημα Ι), για να δείτε ποιες ενέργειες γίνονται για την αντιμετώπιση των μη δειγματοληπτικών σφαλμάτων στην ΕΟΠ.

1.7.γ

Για καθεμία από τις περιπτώσεις που περιγράφονται στη στήλη Α, σημειώστε στη στήλη Β τι είδους μη δειγματοληπτικό σφάλμα αναμένουμε να προκληθεί:

	Στήλη Α Περιγραφή καταστάσεων	**Στήλη Β** Πιθανά σφάλματα
i.	Το δειγματοληπτικό πλαίσιο δεν είναι επικαιροποιημένο.	
ii.	Στις ερωτήσεις χρησιμοποιούνται όροι με διαφορετική σημασία από αυτή που είναι γνωστή στους ερωτώμενους.	
iii.	Οι ερευνητές δεν έχουν εκπαιδευτεί για τη διεξαγωγή της έρευνας.	
iv.	Κατά την επεξεργασία των δεδομένων, παραλείπεται -λόγω έλλειψης πόρων- ο έλεγχος για τον εντοπισμό λαθών.	
v.	Οι ερωτώμενοι έχουν αμφιβολίες για την τήρηση του απορρήτου των δεδομένων που παρέχουν και αρνούνται να συμμετέχουν.	
vi.	Οι ερωτώμενοι νιώθουν άβολα με κάποια ερωτήματα του ερωτηματολογίου.	
vii.	Η κωδικοποίηση των απαντήσεων γίνεται από προσωπικό που δεν έχει εκπαιδευτεί.	

Για να σχηματίσω μια καλή εικόνα για τη συνολική ακρίβεια των αποτελεσμάτων μιας έρευνας, πρέπει να έχω πληροφορίες τόσο για τα δειγματοληπτικά όσο και για τα μη δειγματοληπτικά σφάλματα.

Μαθαίνω

Μη δειγματοληπτικά σφάλματα είναι τα σφάλματα που δεν οφείλονται στη δειγματοληψία, αλλά σε άλλους παράγοντες, κατά τη διάρκεια του σχεδιασμού, της συλλογής ή της επεξεργασίας των στοιχείων της έρευνας. Τέτοια σφάλματα είναι τα σφάλματα μη απόκρισης, επεξεργασίας, κάλυψης και μέτρησης.

Μη δειγματοληπτικά σφάλματα υπάρχουν τόσο στις απογραφικές όσο και στις δειγματοληπτικές έρευνες.

Αυτό το γνώριζες;

Ο **Κώδικας Ορθής Πρακτικής** για τις Ευρωπαϊκές Στατιστικές αποτελείται από 16 αρχές, που καλύπτουν το θεσμικό πλαίσιο, τις διαδικασίες παραγωγής στατιστικών και το στατιστικό προϊόν, και καθορίζει τον τρόπο με τον οποίο θα πρέπει να παράγονται και να διαδίδονται οι ευρωπαϊκές στατιστικές, για να εξασφαλίζεται η εμπιστοσύνη των χρηστών σε αυτές.

ΑΡΧΗ 5
Στατιστικό απόρρητο και προστασία των δεδομένων
Διασφαλίζονται απολύτως η ιδιωτική ζωή των παρόχων δεδομένων, η εμπιστευτικότητα των πληροφοριών που παρέχουν, η χρήση τους μόνο για στατιστικούς σκοπούς και η ασφάλεια των δεδομένων.

1.8 Δραστηριότητες

1.8.α

Για τις προτάσεις που ακολουθούν, σημειώστε **Σ** αν είναι σωστές ή **Λ** αν είναι λανθασμένες:

Πρόταση	Σ ή Λ
i. Εφαρμόζοντας δειγματοληψία χιονοστιβάδας, παίρνουμε αντιπροσωπευτικά δείγματα.	
ii. Αν θέλω να υπολογίσω τον χρόνο ζωής των λαμπτήρων που κατασκευάζει ένα εργοστάσιο, καλύτερα να κάνω απογραφική έρευνα.	
iii. Οι δειγματοληπτικές έρευνες έχουν μόνο δειγματοληπτικά σφάλματα.	
iv. Οι απογραφικές έρευνες έχουν μόνο μη δειγματοληπτικά σφάλματα.	
v. Οι απογραφικές έρευνες έχουν μεγαλύτερο κόστος και διάρκεια υλοποίησης από τις δειγματοληπτικές.	
vi. Το δειγματοληπτικό πλαίσιο της ΕΟΠ είναι η (πλέον πρόσφατη) Απογραφή Πληθυσμού - Κατοικιών.	
vii. Αναφορικά με την ΕΟΠ, στο 2ο στάδιο της δειγματοληψίας δημιουργείται επικαιροποιημένος καταλόγος των νοικοκυριών που διαμένουν στις πρωτογενείς μονάδες του δείγματος (μονάδες επιφανείας).	
viii. Για να εκτιμήσω το επίπεδο εμπιστοσύνης των πολιτών στις επίσημες στατιστικές της ΕΛΣΤΑΤ, έκανα μία έρευνα με ερωτηματολόγιο, στην οποία συμμετείχαν 100 περαστικοί στην Πλατεία Συντάγματος. Το δείγμα μου είναι δείγμα πιθανότητας, αφού τους περαστικούς τους επέλεξα στην τύχη.	
ix. Τα σφάλματα μέτρησης συγκαταλέγονται στα μη δειγματοληπτικά σφάλματα.	
x. Ο αναγωγικός συντελεστής μιας μονάδας του δείγματος είναι ίσος με τον αντίστροφο της πιθανότητας επιλογής της μονάδας αυτής στο δείγμα.	

Αυτό το γνώριζες;

Η εμπιστοσύνη των πολιτών αποτελεί διαχρονικά σημαντική προϋπόθεση για τη διασφάλιση έγκυρης και ποιοτικής στατιστικής πληροφορίας. Η συμμετοχή των πολιτών στις Έρευνες της ΕΛΣΤΑΤ και ειδικά στις ευρείας κλίμακας στατιστικές εργασίες, όπως οι Απογραφές Γεωργίας - Κτηνοτροφίας, Κτιρίων και Πληθυσμού - Κατοικιών, συντελεί στην εκπλήρωση της αποστολής της ΕΛΣΤΑΤ για την παραγωγή και διάδοση χρήσιμων, επίκαιρων και αξιόπιστων στατιστικών στοιχείων.

1.8.6

Για τους $N = 75$ μαθητές/τριες της Γ΄ Λυκείου ενός σχολείου, θέλουμε να μάθουμε πόσο χρόνο απασχολήθηκαν στα μέσα κοινωνικής δικτύωσης το Σαββατοκύριακο. Για τον σκοπό αυτόν, επιλέξαμε τυχαίο δείγμα μεγέθους $n = 15$ μαθητών/τριών για να απαντήσουν στο ερώτημα «Πόσο χρόνο (σε λεπτά) ξόδεψες στα μέσα κοινωνικής δικτύωσης το προηγούμενο Σαββατοκύριακο;».

Επιλογή του δείγματος

Για την επιλογή του δείγματος, δημιουργήθηκε κατάλογος όλων των μαθητών/τριών της Γ΄ Λυκείου και σε κάθε μαθητή/τρια αποδόθηκε διαδοχικά ένας αριθμός από το 1 έως το 75.

Στη συνέχεια, με τη βοήθεια μίας γεννήτριας τυχαίων αριθμών (www.random.org), επιλέχθηκε μία 15άδα ακέραιων αριθμών ανάμεσα στο 1 και στο 75. Κάθε αριθμός είχε την ίδια πιθανότητα να κληρωθεί. Οι μαθητές/τριες που αντιστοιχούσαν στους αριθμούς αυτούς επιλέχθηκαν για το δείγμα μας.

Αναγωγή στον πληθυσμό

Συμπληρώστε τα κενά στις παρακάτω προτάσεις:

i. Ο τρόπος δειγματοληψίας που περιγράφεται παραπάνω είναι η απλή _____ δειγματοληψία.

ii. Επιλέξαμε συνολικά $n = 15$ μαθητές/τριες από τους $N = 75$, άρα για κάθε μαθητή/τρια i η πιθανότητα να επιλεγεί στο δείγμα ήταν $p_i = \frac{n}{N} = \frac{\_\_}{\_\_} = \_\_$

iii. Ο αναγωγικός συντελεστής w_i για κάθε μαθητή/τρια i είναι η αντίστροφη πιθανότητα επιλογής, άρα είναι $w_i = \frac{N}{n} = \frac{\_\_}{\_\_} = \_\_$

iv. Αυτό σημαίνει ότι κάθε μαθητής/τρια στο δείγμα αντιπροσωπεύει _____ μαθητές/τριες στον πληθυσμό.

v. Στο δείγμα περιλαμβάνονται $n = 15$ μαθητές/τριες και ο/η καθένας/μία έχει τον ίδιο αναγωγικό συντελεστή $w_i = 5$. Το άθροισμα των αναγωγικών συντελεστών είναι:

$$\sum_{i=1}^{15} w_i = \sum_{i=1}^{15} 5 = 15 \cdot \_\_ = \_\_ = N$$

Δηλαδή το άθροισμα των αναγωγικών συντελεστών είναι _____ με το πλήθος των μαθητών/τριών στον _____ .

vi. Οι απαντήσεις των 15 μαθητών/τριών του δείγματος, στο ερώτημα της έρευνας, δίνονται στον παρακάτω πίνακα. Συμπληρώστε τα υπόλοιπα στοιχεία:

Καταγράφουμε τον χρόνο για 1 μονάδα του δείγματος.

Καταγράφουμε για 1 παιδί στο δείγμα → 40 λεπτά

Μαθητής/τρια i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Σύνολο (Άθροισμα)
Χρόνος σε λεπτά x_i	20	0	120	30	80	240	0	60	120	180	160	40	80	0	200	1330
Αναγωγικός συντελεστής w_i	5	5	5	5				5	5	5	5	5	5	5	5	
$x_i \cdot w_i$																

Εκτιμούμε τον (συνολικό) χρόνο για w_i μονάδες του πληθυσμού.

Εκτιμούμε για 5 παιδιά στον πληθυσμό → συνολικά $5 \cdot 40 = 200$ λεπτά

vii. Ο χρόνος που δαπάνησαν συνολικά όλοι/ες οι μαθητές/τριες της Γ΄ Λυκείου του σχολείου στα μέσα κοινωνικής δικτύωσης το προηγούμενο Σαββατοκύριακο εκτιμούμε ότι είναι _____ λεπτά.

viii. Εκτιμούμε ότι ο χρόνος που δαπάνησαν οι μαθητές/τριες της Γ΄ Λυκείου του σχολείου στα μέσα κοινωνικής δικτύωσης είναι, κατά μέσο όρο, ο σταθμισμένος μέσος των παρατηρήσεων στο δείγμα, δηλαδή:

$$\frac{\sum w_i \cdot x_i}{\sum w_i} = \frac{6650}{\quad} \approx \quad$$ λεπτά, όπου οι συντελεστές στάθμισης είναι οι _____ συντελεστές.

ix. Ο απλός μέσος των παρατηρήσεων στο δείγμα είναι $\frac{\sum x_i}{n} = \frac{\quad}{15} \approx$ _____ λεπτά.

Στο παράδειγμά μας ο απλός και ο σταθμισμένος μέσος (με συντελεστές στάθμισης τους αναγωγικούς συντελεστές) των παρατηρήσεων του δείγματος ταυτίζονται. Αυτό συμβαίνει, επειδή όλες οι μονάδες του δείγματος έχουν τον ίδιο αναγωγικό συντελεστή, που είναι απόρροια του γεγονότος ότι έχουν την ίδια πιθανότητα επιλογής, λόγω της απλής τυχαίας δειγματοληψίας.

x. Για να υπολογίσουμε την πραγματική τιμή μιας παραμέτρου θα πρέπει να συλλέξουμε στοιχεία από ολόκληρο τον _____.

xi. Η διαφορά ανάμεσα στην εκτίμησή μας για την τιμή της παραμέτρου και την πραγματική τιμή της παραμέτρου αυτής ονομάζεται _____ σφάλμα.

Μαθαίνω

Απλός μέσος ενός συνόλου n παρατηρήσεων x_i είναι το άθροισμα $\sum x_i$ των παρατηρήσεων διά του πλήθους n των παρατηρήσεων: $\bar{x} = \frac{\sum x_i}{n}$

Σταθμισμένος μέσος ενός συνόλου n παρατηρήσεων x_i, όπου η καθεμία έχει βαρύτητα (συντελεστή στάθμισης) w_i, είναι το άθροισμα $\sum w_i \cdot x_i$ των γινομένων των παρατηρήσεων με τους αντίστοιχους συντελεστές στάθμισης προς το άθροισμα $\sum w_i$ των συντελεστών στάθμισης: $\bar{x} = \frac{\sum w_i \cdot x_i}{\sum w_i}$

Αναφέραμε παραπάνω ότι, στην απλή τυχαία δειγματοληψία δείγματος μεγέθους n από έναν πληθυσμό μεγέθους N, η πιθανότητα μία μονάδα i του πληθυσμού να συμπεριληφθεί στο δείγμα είναι $p_i = \frac{n}{N}$. Μπορείτε να το αποδείξετε;

Θα χρειαστεί να θυμηθείτε τι είναι συνδυασμός και ποια είναι η βασική αρχή της απαρίθμησης ή αρχή του γινομένου.

1.8.γ

Συμπληρώστε τα κενά στους συλλογισμούς που ακολουθούν:

i. Όλα τα δυνατά διαφορετικά δείγματα μεγέθους n από έναν πληθυσμό μεγέθους N (διαφορετικών στοιχείων) είναι σε πλήθος όσα οι _____ N ανά n, δηλαδή:

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

Δεν μας ενδιαφέρει η σειρά των μονάδων του δείγματος.

ii. Όλα τα δυνατά διαφορετικά δείγματα μεγέθους n από έναν πληθυσμό μεγέθους N που περιέχουν μία συγκεκριμένη μονάδα i του πληθυσμού είναι σε πλήθος όσα οι συνδυασμοί N-1 ανά n-1, δηλαδή:

$$\binom{N-1}{n-1} = \frac{(N-1)!}{(n-1)!(N-1-(n-1))!} = \frac{(N-1)!}{(n-1)!\,(N-1-n+1)!} = \frac{\qquad}{\qquad}$$

Επεξήγηση

Για να μετρήσουμε πόσα είναι τα δυνατά δείγματα μεγέθους n από έναν πληθυσμό μεγέθους N διαφορετικών στοιχείων που περιλαμβάνουν μία συγκεκριμένη μονάδα i, σκεφτόμαστε ως εξής:

Ενέργεια 1: Υπάρχει 1 τρόπος να επιλεγεί η μονάδα i.

Ενέργεια 2: Υπάρχουν $\binom{N-1}{n-1}$ τρόποι για να επιλεγούν οι (υπόλοιπες) n-1 μονάδες του δείγματος από τις υπόλοιπες N-1 μονάδες του πληθυσμού.

Άρα, με βάση την αρχή του γινομένου, υπάρχουν $1 \cdot \binom{N-1}{n-1} = \binom{N-1}{n-1}$ τρόποι να επιλεγεί ένα δείγμα που να περιλαμβάνει τη μονάδα i.

iii. Στην απλή τυχαία δειγματοληψία, κάθε μονάδα i του πληθυσμού έχει την ίδια πιθανότητα να επιλεγεί στο δείγμα, η οποία ισούται με:

$$\text{πιθανότητα επιλογής μονάδας } i = \frac{\text{πλήθος δυνατών διαφορετικών δειγμάτων που περιέχουν τη μονάδα } i}{\text{πλήθος δυνατών διαφορετικών δειγμάτων}} =$$

$$= \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{\frac{(N-1)!}{(n-1)!(N-n)!}}{\frac{N!}{n!(N-n)!}} = \frac{(N-1)!\,n!}{N!\,(n-1)!} = \frac{n}{N}$$

$n! = (n-1)! \cdot n$

$N! = (N-1)! \cdot N$

Θυμάμαι

Συνδυασμός ν διαφορετικών αντικειμένων ανά **κ** λέγεται κάθε ομάδα **κ** αντικειμένων από τα **ν**, όταν δεν μας ενδιαφέρει η σειρά τοποθέτησής τους.

Το **πλήθος των συνδυασμών ν** διαφορετικών αντικειμένων **ανά κ** συμβολίζεται $\binom{\nu}{\kappa}$ και δίνεται από τον τύπο: $\binom{\nu}{\kappa} = \frac{\nu!}{\kappa!(\nu-\kappa)!}$

Σύμφωνα με την **«αρχή του γινομένου»** ή, αλλιώς, **«βασική αρχή της απαρίθμησης»**, αν υπάρχουν

ν_1 τρόποι για να κάνω την ενέργεια ε_1
ν_2 τρόποι για να κάνω την ενέργεια ε_2
.
.
ν_κ τρόποι για να κάνω την ενέργεια ε_κ

τότε υπάρχουν $\nu_1 \cdot \nu_2 \cdot \ldots \cdot \nu_\kappa$ τρόποι για να κάνω τις ενέργειες ε_1 ε_2 ... ε_κ ταυτόχρονα.

Αυτό το γνώριζες;

Η ΕΛΣΤΑΤ παρέχει ανοιχτή πρόσβαση σε ανωνυμοποιημένα μικροδεδομένα στατιστικών ερευνών της, τα οποία έχουν ανωνυμοποιηθεί σύμφωνα με κριτήρια ανωνυμοποίησης που έχει προκαθορίσει έτσι, ώστε να μην είναι δυνατή η άμεση ή έμμεση αποκάλυψη των ερευνώμενων μονάδων (Αρχεία Δημόσιας Χρήσης).

Περισσότερες πληροφορίες: https://www.statistics.gr/el/public-use-files

ΤΑ ΒΗΜΑΤΑ ΠΟΥ ΑΚΟΛΟΥΘΩ ΣΕ ΜΙΑ ΔΕΙΓΜΑΤΟΛΗΠΤΙΚΗ ΕΡΕΥΝΑ ΜΕ ΕΡΩΤΗΜΑΤΟΛΟΓΙΟ

ΒΗΜΑ 01

Προσδιορίζω τις ανάγκες πληροφόρησης

- Προσδιορίζω τι θέλω να μάθω.
- Προσδιορίζω τι πρέπει να ρωτήσω.
- Προσδιορίζω/ορίζω τις βασικές έννοιες της έρευνας.

ΒΗΜΑ 02

Σχεδιάζω

- Προσδιορίζω τις μεταβλητές και τα αποτελέσματα που θέλω να παράγω.
- Σχεδιάζω το σύστημα για την εισαγωγή και επεξεργασία των συλλεγόμενων στοιχείων.
- Σχεδιάζω τη διαδικασία συλλογής των δεδομένων.
- Προσδιορίζω το πλαίσιο και τον τρόπο δειγματοληψίας.
- Προσδιορίζω τις μεθόδους επεξεργασίας και ανάλυσης των δεδομένων.

ΒΗΜΑ 07

Δημοσιεύω

- Παρουσιάζω τα αποτελέσματα με πίνακες, ανακοινώσεις, infographic, δημοσιεύματα κ.λπ.
- Παρέχω στους χρήστες των στατιστικών πληροφορίες για τους ορισμούς των μεταβλητών, τη μεθοδολογία και την ποιότητα των αποτελεσμάτων της έρευνας.

ΒΗΜΑ 08

Αξιολογώ

- Αξιολογώ τις διαδικασίες της έρευνας.
- Προσδιορίζω τι χρειάζεται βελτίωση.
- Κάνω προτάσεις για τη βελτίωση των διαδικασιών/μεθόδων.

ΒΗΜΑ
03

Υλοποιώ

- Δημιουργώ το ερωτηματολόγιο.
- Δημιουργώ το σύστημα για την εισαγωγή και επεξεργασία των δεδομένων.

ΒΗΜΑ
04

Συλλέγω

- Επιλέγω το δείγμα της έρευνας.
- Συλλέγω τα δεδομένα, με τη βοήθεια του ερωτηματολογίου.

ΒΗΜΑ
05

Επεξεργάζομαι

- Κωδικοποιώ και εισάγω τα δεδομένα.
- Ελέγχω τα δεδομένα και κάνω τις απαραίτητες διορθώσεις.
- Υπολογίζω τους αναγωγικούς συντελεστές.
- Υπολογίζω περιγραφικά μέτρα των στοιχείων.

ΒΗΜΑ
06

Αναλύω

- Κάνω την εκτίμηση των παραμέτρων/ αποτελεσμάτων.
- Ελέγχω, επεξηγώ και επικυρώνω τα αποτελέσματα.
- Διασφαλίζω τη μη αποκάλυψη εμπιστευτικών στοιχείων.

2. ΓΡΑΜΜΙΚΗ ΣΥΣΧΕΤΙΣΗ

2.1 Είδη μεταβλητών

Έχουμε μιλήσει για την ΕΟΠ και τις πληροφορίες που καταγράφει. Στην 1η στήλη του Πίνακα 1 που ακολουθεί, έχουν σημειωθεί μερικά από τα χαρακτηριστικά (μεταβλητές) των νοικοκυριών της Χώρας και των ατόμων που διαμένουν σε αυτά, τα οποία εξετάζει η έρευνα.

2.1.α

Για καθεμία από τις μεταβλητές της 1ης στήλης, σημειώστε, στη 2η στήλη, αν πρόκειται για ποιοτική ή ποσοτική μεταβλητή. Στην περίπτωση ποσοτικής μεταβλητής, σημειώστε, στην 3η στήλη, αν είναι διακριτή ή συνεχής.

ΠΙΝΑΚΑΣ 1

	Χαρακτηριστικό (μεταβλητή)	Είδος μεταβλητής: ποιοτική ή ποσοτική	Είδος ποσοτικής μεταβλητής: διακριτή ή συνεχής
i.	Πλήθος μελών νοικοκυριού		
ii.	Περιφέρεια μόνιμης κατοικίας του νοικοκυριού		
iii.	Ποσό που δαπανήθηκε για φωτισμό και καύσιμα (ηλεκτρικό ρεύμα, φυσικό αέριο, καυσόξυλα κ.λπ.) της κατοικίας (για ζεστό νερό, μαγείρεμα, θέρμανση, ψύξη κ.ά.)		
iv.	Ποσό που δαπανήθηκε σε οδοντιάτρους		
v.	Ποσό που δαπανήθηκε για την αγορά άσπρου ψωμιού, από καλαμπόκι, σιμιγδάλι κ.λπ.		
vi.	Αριθμός αυτοκινήτων ΙΧ που διαθέτει το νοικοκυριό		
vii.	Με τι μαγειρεύει το νοικοκυριό (π.χ. ηλεκτρική κουζίνα, συσκευή υγραερίου)		

Μαθαίνω

Μεταβλητές ονομάζονται τα διάφορα χαρακτηριστικά ως προς τα οποία εξετάζεται ένας πληθυσμός. Τέτοια χαρακτηριστικά είναι το φύλο, η ηλικία, το εισόδημα, ο αριθμός εργαζομένων, ο κύκλος εργασιών κ.λπ.

Διακρίνονται δύο είδη μεταβλητών:
1) οι **ποιοτικές**, των οποίων οι τιμές είναι κατηγορίες, π.χ. φύλο, επάγγελμα,
2) οι **ποσοτικές**, των οποίων οι τιμές είναι αριθμοί, π.χ. ο αριθμός των μελών σε ένα νοικοκυριό, η ποσότητα ψωμιού (σε γραμμάρια) που αγοράστηκε σε μια περίοδο αναφοράς.

Οι ποσοτικές μεταβλητές διακρίνονται περαιτέρω σε:
α) **διακριτές**, όταν παίρνουν «μεμονωμένες» τιμές, π.χ. αριθμός εργαζομένων,
β) **συνεχείς**, οι οποίες μπορούν να πάρουν οποιαδήποτε τιμή μέσα σε ένα διάστημα τιμών, π.χ. βάρος, κύκλος εργασιών.

2.2. Διάγραμμα διασποράς Συσχέτιση

Παρατηρώντας τις μεταβλητές που έχουμε στη διάθεσή μας από την έρευνα, μπορούμε να θέσουμε ερωτήματα προς διερεύνηση. Ένα ερώτημα είναι αν υπάρχει κάποια σχέση μεταξύ δύο μεταβλητών, όπως για παράδειγμα μεταξύ του εισοδήματος και των δαπανών των νοικοκυριών, και, αν ναι, τότε να περιγράψουμε με μαθηματικό τρόπο ποια είναι αυτή.

Πρωτογενή δεδομένα είναι αυτά που συλλέγονται από το υποκείμενο της έρευνας.

Τα πρωτογενή στοιχεία της έρευνας αφορούν σε διαφορετικές συνθέσεις/«τύπους» νοικοκυριών, όπως «μονομελές νοικοκυριό», «ζευγάρι χωρίς παιδί», «ένας ενήλικας και ένα παιδί» κ.ά.

Εμείς επιλέξαμε 15 εγγραφές, από την ΕΟΠ 2019, που αφορούν σε τύπο νοικοκυριού «4μελής οικογένεια με 2 παιδιά ηλικίας έως 15 ετών». Στον Πίνακα 2 που ακολουθεί, παρουσιάζονται 2 συγκεντρωτικές μεταβλητές, για αυτές τις εγγραφές, το «ΕΙΣΟΔΗΜΑ» και οι «ΔΑΠΑΝΕΣ». Θα εξετάσουμε αν υπάρχει κάποια σχέση μεταξύ των μεταβλητών αυτών και ποια είναι αυτή.

ΠΙΝΑΚΑΣ 2

ΕΓΓΡΑΦΕΣ	ΕΙΣΟΔΗΜΑ (€)	ΕΙΣΟΔΗΜΑ ΣΤΡΟΓΓΥΛΟΠΟΙΗΜΕΝΟ (σε χιλιάδες €)	ΔΑΠΑΝΕΣ (€)	ΔΑΠΑΝΕΣ ΣΤΡΟΓΓΥΛΟΠΟΙΗΜΕΝΕΣ (σε χιλιάδες €)
1	46.047,96	46	26.944,44	27
2	35.900,04	36	22.374,48	22
3	28.600,08	29	22.863,12	23
4	25.899,96	26	18.660,60	19
5	25.552,54	26	20.018,52	20
6	27.120,00	27	20.549,88	21
7	21.420,00	21	19.267,92	19
8	21.066,12	21	14.048,40	14
9	19.864,68	20	16.525,44	17
10	19.400,04	19	17.199,24	17
11	18.083,40	18	11.656,08	12
12	18.259,92	18	14.399,64	14
13	17.400,00	17	13.082,04	13
14	14.723,16	15	10.900,92	11
15	11.580,00	12	9.733,20	10

Εδώ μας ενδιαφέρει πως μεταβάλλονται οι «ΔΑΠΑΝΕΣ» σε σχέση με το «ΕΙΣΟΔΗΜΑ».

είναι η εξαρτημένη μεταβλητή που συνήθως συμβολίζεται με Y.

είναι η ανεξάρτητη μεταβλητή που συνήθως συμβολίζεται με X.

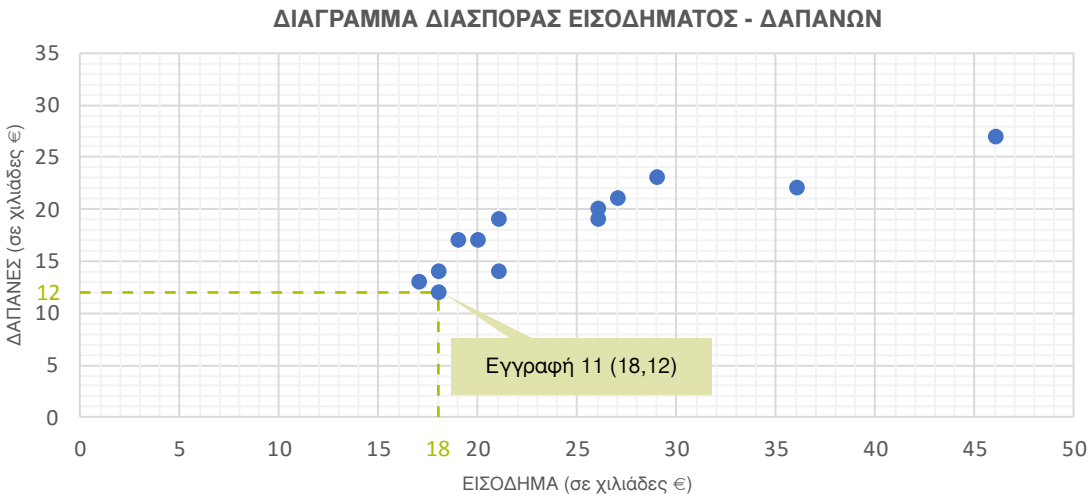
Στην περίπτωση που θέλαμε να βρούμε τι συμβαίνει με το «ΕΙΣΟΔΗΜΑ» σε σχέση με τις «ΔΑΠΑΝΕΣ», τότε η εξαρτημένη μεταβλητή θα ήταν το «ΕΙΣΟΔΗΜΑ» και η ανεξάρτητη μεταβλητή οι «ΔΑΠΑΝΕΣ».

Παρατηρώ τα στοιχεία που περιέχονται στον Πίνακα 2 και αναρωτιέμαι αν υπάρχει κάποια σχέση μεταξύ του «ΕΙΣΟΔΗΜΑΤΟΣ» και των «ΔΑΠΑΝΩΝ» των νοικοκυριών. Θα το διερευνήσω γραφικά, δημιουργώντας το διάγραμμα των σημείων/ζευγών των δύο μεταβλητών. Στον οριζόντιο άξονα αποτυπώνονται οι τιμές της ανεξάρτητης μεταβλητής «ΕΙΣΟΔΗΜΑ» και στον κάθετο οι τιμές της εξαρτημένης μεταβλητής «ΔΑΠΑΝΕΣ».

2.2.α

Στο Διάγραμμα 1, έχει παραλειφθεί η αποτύπωση των εγγραφών 14 και 15. Προσθέστε τις εγγραφές που λείπουν.

ΔΙΑΓΡΑΜΜΑ 1



Σωστή σκέψη! Παρατηρούμε λοιπόν, στο Διάγραμμα 1, τον τρόπο που «διασπείρονται» τα σημεία/ζεύγη των μεταβλητών που εξετάζουμε.

Μαθαίνω

- Το Διάγραμμα 1 ονομάζεται **διάγραμμα διασποράς** των μεταβλητών X (εδώ «ΕΙΣΟΔΗΜΑ») και Y (εδώ «ΔΑΠΑΝΕΣ»).
- Κάθε σημείο του διαγράμματος αναπαριστά ένα ζεύγος παρατηρήσεων.
- Τα διαγράμματα διασποράς μας βοηθούν να μελετήσουμε τη σχέση που ενδέχεται να έχουν δύο μεταβλητές.

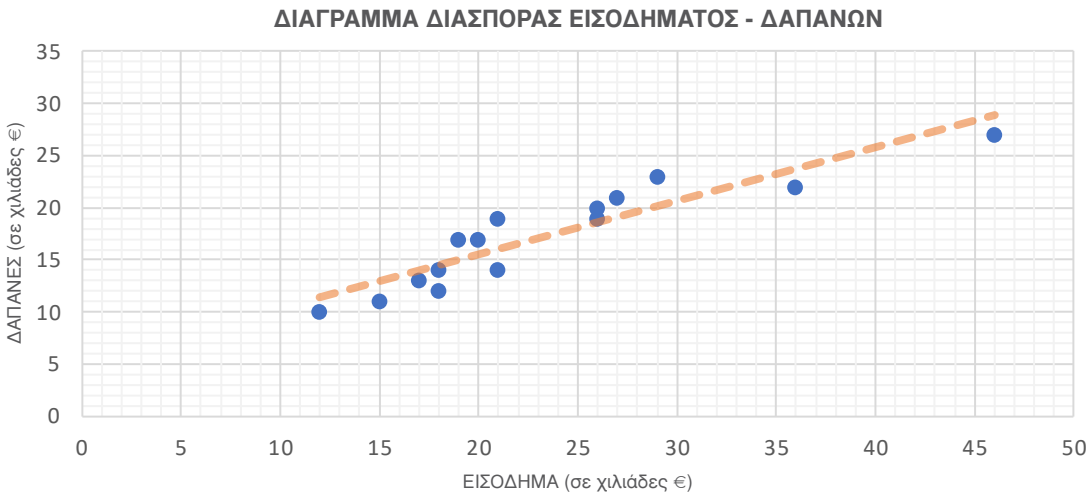
2.2.6

Συμπληρώστε τα κενά στον διάλογο που ακολουθεί:

Παρατηρώ, στο Διάγραμμα 1, ότι όσο μεγαλύτερο είναι το εισόδημα των νοικοκυριών τόσο _____ **(i)**, συνήθως, είναι και οι δαπάνες που κάνουν. Άρα υπάρχει σχέση μεταξύ τους.

Ακριβώς, και όπως παρατηρούμε στο Διάγραμμα 2, τα σημεία σχηματίζουν μία νοητή _____ **(ii)**. Οπότε, εδώ, η μεταξύ τους σχέση καλείται **γραμμική**.

ΔΙΑΓΡΑΜΜΑ 2



2.3 Συντελεστής Γραμμικής Συσχέτισης

Επαρκούν αυτές οι πληροφορίες για να περιγραφεί η σχέση των μεταβλητών;

Σωστός προβληματισμός, γιατί δε φτάνει μόνο να βρούμε το είδος της σχέσης μεταξύ των δύο ποσοτικών μεταβλητών, αλλά πρέπει να γνωρίζουμε και την ένταση της σχέσης τους. Για τον σκοπό αυτόν, υπολογίζουμε τον **Συντελεστή Γραμμικής Συσχέτισης *r*** των δύο μεταβλητών.

Μαθαίνω

Ένας συντελεστής που μετράει την ένταση της (γραμμικής) σχέσης μεταξύ δύο ποσοτικών μεταβλητών είναι ο **Συντελεστής Γραμμικής Συσχέτισης** και συμβολίζεται με ***r***.

Ο **Συντελεστής Γραμμικής Συσχέτισης *r*** είναι καθαρός αριθμός, δηλαδή δεν εκφράζεται με κάποια μονάδα μέτρησης, και παίρνει τιμές από -1 έως και 1.

Για να βρούμε την ένταση της γραμμικής σχέσης δύο ποσοτικών μεταβλητών, χρησιμοποιούμε τον Συντελεστή Γραμμικής Συσχέτισης r, ο οποίος υπολογίζεται από τον τύπο:

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

όπου

$\bar{x}$ η δειγματική μέση τιμή
$\bar{y}$ η δειγματική μέση τιμή
n το μέγεθος του δείγματος.

Θυμάμαι

Τύποι υπολογισμού:

Μεταβλητές	Δειγματική μέση τιμή
μεταβλητή X	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
μεταβλητή Y	$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$

2.4 Υπολογισμός του Συντελεστή Γραμμικής Συσχέτισης

Θα υπολογίσουμε τον Συντελεστή Γραμμικής Συσχέτισης r των μεταβλητών «ΕΙΣΟΔΗΜΑ» και «ΔΑΠΑΝΕΣ», χρησιμοποιώντας τον τύπο:

$$r = \frac{\Sigma_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\Sigma_{i=1}^{n} (y_i - \bar{y})^2}} \text{ για } n = 15$$

Οπότε θα χρειαστεί να υπολογίσουμε:

- το άθροισμα των γινομένων των αποκλίσεων των x_i και των y_i από τη δειγματική μέση τιμή $X\,(\bar{x})$ και $Y\,(\bar{y})$, αντίστοιχα.

$$\Sigma_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

- το άθροισμα των τετραγώνων των αποκλίσεων των x_i από τη δειγματική μέση τιμή $\bar{x}$.

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

- το άθροισμα των τετραγώνων των αποκλίσεων των y_i από τη δειγματική μέση τιμή $\bar{y}$.

$$\Sigma_{i=1}^{n} (y_i - \bar{y})^2$$

Αρχικά, αντιγράφουμε, από τον Πίνακα 2, τις στήλες των μεταβλητών με τις οποίες θα εργαστούμε και σχηματίζουμε τον Πίνακα 3 στην επόμενη σελίδα, ο οποίος περιέχει, συμπληρωματικά, και τους υπολογισμούς που χρειάζονται για να εφαρμόσουμε τον τύπο εύρεσης του Συντελεστή Γραμμικής Συσχέτισης r.

2.4.α

Κάντε τους απαραίτητους υπολογισμούς και συμπληρώστε τα κενά στον Πίνακα 3. Στη συνέχεια υπολογίστε τον Συντελεστή Γραμμικής Συσχέτισης r.

Αυτό το γνώριζες;

Ο Συντελεστής Γραμμικής Συσχέτισης r δύο μεταβλητών μπορεί να υπολογιστεί στο Excel με τη χρήση της συνάρτησης CORREL. Βρείτε, στο Παράρτημα ΙΙ, τα βήματα για τον υπολογισμό του r, με τη βοήθεια του Excel.

ΠΙΝΑΚΑΣ 3

ΕΓΓΡΑΦΕΣ	ΕΙΣΟΔΗΜΑ ΣΤΡΟΓΓΥΛΟ-ΠΟΙΗΜΕΝΟ (σε χιλιάδες €)	ΔΑΠΑΝΕΣ ΣΤΡΟΓΓΥΛΟ-ΠΟΙΗΜΕΝΕΣ (σε χιλιάδες €)					
	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x}) * (y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	46	27	22,6	9,7	219,2	510,8	94,1
2	36	22	12,6		59,2	158,8	22,1
3	29	23	5,6	5,7	31,9	31,4	32,5
4	26	19	2,6	1,7		6,8	2,9
5	26	20		2,7	7,0	6,8	7,3
6	27	21	3,6	3,7	13,3	13,0	13,7
7	21	19	-2,4	1,7	-4,1		2,9
8	21	14	-2,4	-3,3	7,9	5,8	10,9
9	20	17	-3,4	-0,3	1,0	11,6	0,1
10	19	17	-4,4	-0,3	1,3	19,4	0,1
11	18	12	-5,4	-5,3	28,6	29,2	28,1
12	18	14	-5,4	-3,3	17,8	29,2	10,9
13	17	13	-6,4	-4,3	27,5	41,0	18,5
14	15	11	-8,4	-6,3	52,9	70,6	39,7
15	12	10	-11,4	-7,3	83,2	130,0	
Σύνολο	**351**	**259**			**551,1**	**1070,2**	

Άθροισμα των x_i

$$\sum_{i=1}^{n} x_i$$

Άθροισμα των y_i

$$\sum_{i=1}^{n} y_i$$

Άθροισμα των γινομένων των αποκλίσεων των x_i, y_i από τη δειγματική μέση τιμή $\bar{x}$ και $\bar{y}$, αντίστοιχα

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Άθροισμα των τετραγώνων των αποκλίσεων των x_i από τη δειγματική μέση τιμή $\bar{x}$

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

Άθροισμα των τετραγώνων των αποκλίσεων των y_i από τη δειγματική μέση τιμή $\bar{y}$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2$$

Υπολογισμοί

- Δειγματική μέση τιμή των μεταβλητών X και Y
  (με στρογγυλοποίηση στο 1ο δεκαδικό ψηφίο)

$$\bar{x} = \frac{\Sigma_{i=1}^{n} x_i}{n} = \frac{\quad}{\quad} = \quad \text{(i)}$$

$$\bar{y} = \frac{\quad}{\quad} = \frac{\quad}{\quad} \approx \quad \text{(ii)}$$

- Άθροισμα των τετραγώνων των αποκλίσεων y_i από τη δειγματική μέση τιμή $\bar{y}$
  (με στρογγυλοποίηση στο 1ο δεκαδικό ψηφίο)

$\Sigma_{i=1}^{n} (y_i - \bar{y})^2 =$ (iii) οπότε $\sqrt{\Sigma_{i=1}^{n} (y_i - \bar{y})^2} = 18,4$

- Άθροισμα των τετραγώνων των αποκλίσεων x_i από τη δειγματική μέση τιμή $\bar{x}$
  (με στρογγυλοποίηση στο 1ο δεκαδικό ψηφίο)

$\Sigma_{i=1}^{n} (x_i - \bar{x})^2 = 1070,2$ οπότε $\sqrt{\Sigma_{i=1}^{n} (x_i - \bar{x})^2} = 32,7$

Στον Πίνακα 3, έχουμε υπολογίσει ότι:

$\Sigma(x_i - \bar{x}) \cdot (y_i - \bar{y}) = 551,1$

- Αντικαθιστούμε τις παραπάνω τιμές στον τύπο

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

(υπολογίζω με στρογγυλοποίηση στο 3ο δεκαδικό ψηφίο)

$$r = \frac{\quad}{\quad} \approx \frac{\quad}{\quad} \approx \quad \text{(iv)}$$

και βρίσκουμε ότι ο **Συντελεστής Γραμμικής Συσχέτισης *r*** είναι 0,916.

Το βρήκαμε, αλλά αυτός ο αριθμός τι μας δείχνει; Τι πρέπει να γνωρίζουμε για τις τιμές που μπορεί να παίρνει ο **Συντελεστής Γραμμικής Συσχέτισης *r***;

2.5
Οι τιμές του r

Ο **Συντελεστής Γραμμικής Συσχέτισης r:**

- είναι καθαρός αριθμός, δηλαδή δεν εκφράζεται με συγκεκριμένες μονάδες μέτρησης,
- παίρνει τιμές από -1 έως και 1.

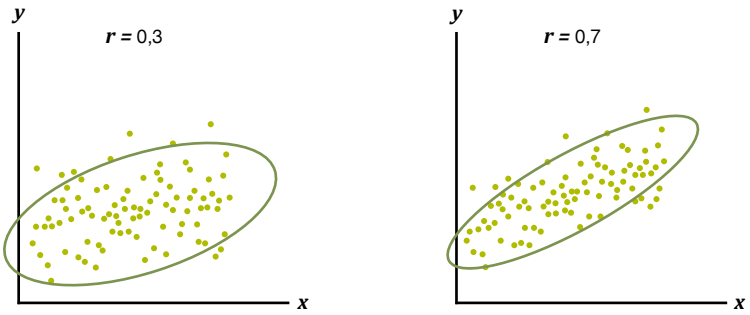
2.5.α

Παρατηρήστε τα παρακάτω διαγράμματα διασποράς, σε συνδυασμό με την αντίστοιχη τιμή του Συντελεστή Γραμμικής Συσχέτισης των μεταβλητών X, Y, και συμπληρώστε τα κενά:

■ **Όταν 0 < r ≤ +1**, τότε οι μεταβλητές X, Y είναι **θετικά** γραμμικά συσχετισμένες.

Τι σημαίνει r = 0,3 και r = 0,7;

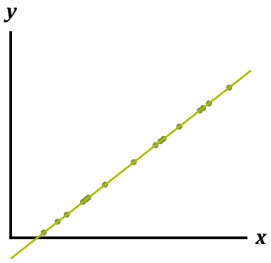
Διάγραμμα διασποράς των μεταβλητών X και Y, όπου $0 < r < +1$



- Υπάρχει **θετική γραμμική συσχέτιση**, δηλαδή οι τιμές της μεταβλητής Y τείνουν να αυξηθούν, ακολουθώντας την _____ (i) των τιμών της μεταβλητής X.
- Όσο πιο κοντά στο 1 είναι η τιμή του r _____ (ii) ισχυρότερη είναι η _____ (iii) **γραμμική συσχέτιση**.
- Όσο πιο ξεκάθαρη είναι η νοητή ευθεία που σχηματίζουν τα σημεία στο διάγραμμα διασποράς τόσο _____ (iv) είναι η _____ (v).

Τι σημαίνει r =1;

Διάγραμμα διασποράς των μεταβλητών X και Y, όπου $r = 1$

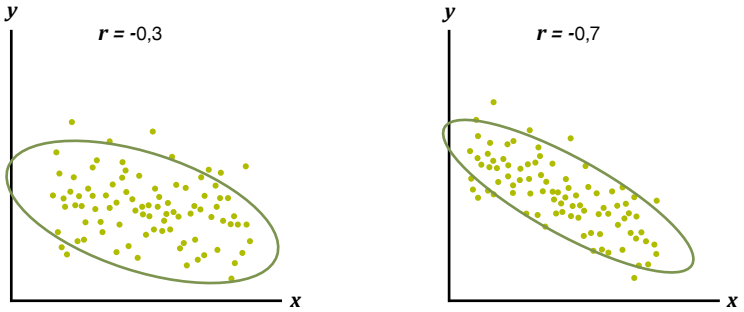


- Υπάρχει τέλεια θετική γραμμική συσχέτιση και, όταν αυξάνεται η τιμή της μίας μεταβλητής, _____ (vi) και η τιμή της άλλης.
- Όλα τα σημεία του διαγράμματος διασποράς βρίσκονται **πάνω** σε μία ευθεία, η οποία έχει θετική κλίση.

■ **Όταν -1 ≤ *r* < 0**, τότε οι μεταβλητές X, Y είναι **αρνητικά** γραμμικά συσχετισμένες.

Τι σημαίνει *r* = - 0,3 και *r* = - 0,7;

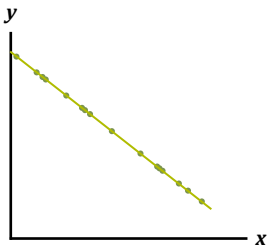
Διάγραμμα διασποράς των μεταβλητών X και Y, όπου $-1 < r < 0$



- Υπάρχει **αρνητική γραμμική συσχέτιση**, δηλαδή οι τιμές της μεταβλητής Y τείνουν να (vii) με την αύξηση των τιμών της μεταβλητής X.
- Όσο πιο κοντά στο -1 είναι η τιμή του r (viii) ισχυρότερη είναι η (ix) **γραμμική συσχέτιση**.
- Όσο πιο ξεκάθαρη είναι η νοητή ευθεία που σχηματίζουν τα σημεία στο διάγραμμα διασποράς τόσο (x) είναι η συσχέτιση.

Τι σημαίνει *r* = -1;

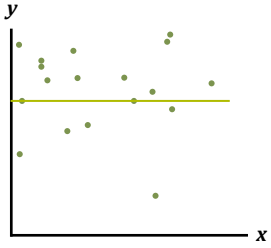
Διάγραμμα διασποράς των μεταβλητών X και Y, όπου $r = -1$



- Υπάρχει τέλεια αρνητική γραμμική συσχέτιση και, όταν αυξάνεται η τιμή της μίας μεταβλητής, (xi) η τιμή της άλλης.
- Όλα τα σημεία του διαγράμματος διασποράς βρίσκονται πάνω σε μία ευθεία, η οποία έχει (xii) κλίση.

■ **Όταν *r* = 0**

Διάγραμμα διασποράς των μεταβλητών X και Y, όπου $r = 0$



- Δεν υπάρχει καμία (xiii) σχέση που να περιγράφει τη σχέση των X και Y.
- Αυτό δεν αποκλείει ότι μπορεί να υπάρχει μη γραμμική σχέση.

Συντελεστής Γραμμικής Συσχέτισης	Ένταση της Γραμμικής Συσχέτισης
$\|r\| = 1$	Τέλεια συσχέτιση
$0,7 \leq \|r\| < 1$	Ισχυρή συσχέτιση
$0,5 \leq \|r\| < 0,7$	Μέτρια συσχέτιση
$0,3 \leq \|r\| < 0,5$	Ασθενής συσχέτιση
$0 \leq \|r\| < 0,3$	Πολύ ασθενής ή μη ύπαρξη συσχέτισης

Η ένταση της γραμμικής συσχέτισης καθορίζεται από την απόλυτη τιμή του r.
Από το πρόσημο του r καθορίζεται το είδος της, δηλαδή αν είναι θετική ή αρνητική συσχέτιση.

Στον ακόλουθο πίνακα, συνοψίζονται όσα αναφέρθηκαν για το είδος της γραμμικής συσχέτισης δύο μεταβλητών.

Μαθαίνω

ΓΡΑΜΜΙΚΗ ΣΥΣΧΕΤΙΣΗ

	ανεξάρτητη μεταβλητή X	**εξαρτημένη** μεταβλητή Y
θετική γραμμική συσχέτιση	αύξηση	αύξηση
	μείωση	μείωση
αρνητική γραμμική συσχέτιση	αύξηση	μείωση
	μείωση	αύξηση

2.5.6

Άρα, σύμφωνα με όσα ειπώθηκαν παραπάνω, **ο Συντελεστής Γραμμικής Συσχέτισης** r = 0,899 ποια πληροφορία μας δίνει; Συμπληρώστε τα κενά της ακόλουθης πρότασης:

Ο **Συντελεστής Γραμμικής Συσχέτισης** r = 0,899 δείχνει ότι η συσχέτιση μεταξύ των δύο μεταβλητών μας, του «ΕΙΣΟΔΗΜΑΤΟΣ» και των «ΔΑΠΑΝΩΝ», είναι (i) και (ii).

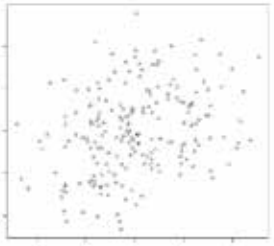
2.5.γ

Για τις προτάσεις που ακολουθούν, σημειώστε **Σ** αν είναι σωστές ή **Λ** αν είναι λανθασμένες:

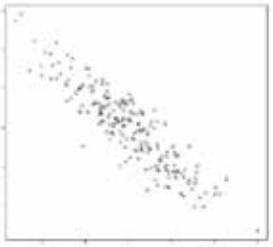
Πρόταση		Σ ή Λ
i.	Όταν r = 0,72 η γραμμική συσχέτιση είναι ισχυρή.	
ii.	Όταν r = 0,25 η γραμμική συσχέτιση είναι μέτρια.	
iii.	Όταν r = 0 η γραμμική συσχέτιση είναι ισχυρή.	
iv.	Όταν r = -1 δεν υπάρχει γραμμική συσχέτιση.	
v.	Όταν r = 1 υπάρχει πολύ ισχυρή (τέλεια) γραμμική συσχέτιση.	
vi.	Όταν υπάρχει θετική γραμμική συσχέτιση, τότε, όταν αυξάνονται οι τιμές της μίας μεταβλητής, οι τιμές της άλλης τείνουν να αυξάνονται, επίσης.	
vii.	Όταν υπάρχει αρνητική γραμμική συσχέτιση, τότε, όταν μειώνονται οι τιμές της μίας μεταβλητής, οι τιμές της άλλης τείνουν να μειώνονται, επίσης.	
viii.	Όταν υπάρχει αρνητική γραμμική συσχέτιση, τότε, όταν μειώνονται οι τιμές της μίας μεταβλητής, οι τιμές της άλλης τείνουν να αυξάνονται.	

2.5.δ

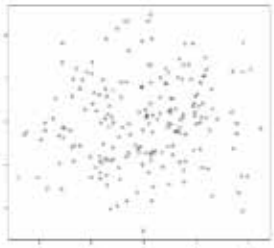
Αντιστοιχίστε τους Συντελεστές Γραμμικής Συσχέτισης r_1=-0,9, r_2=-0,5, r_3=0,1, r_4=0,3, r_5=0,5, r_6=0,8 δύο ποσοτικών μεταβλητών με τα αντίστοιχα διαγράμματα διασποράς των μεταβλητών αυτών:



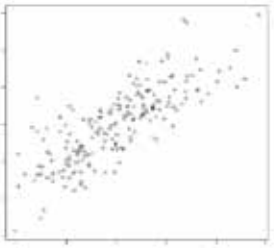
i.



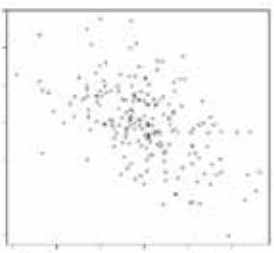
iv.



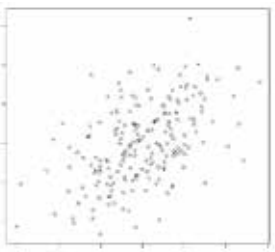
ii.



v.



iii.



vi.

2.6 Συσχέτιση και Αιτιότητα

Οι ερευνητές, στους τομείς του επιστημονικού τους πεδίου, προσπαθούν να βρουν ουσιώδεις και σημαντικές συσχετίσεις πάνω σε στατιστικά δεδομένα, με στόχο να πραγματοποιήσουν πιθανές νέες ανακαλύψεις. Υπάρχουν, όμως, δυσκολίες: **η ύπαρξη συσχέτισης μεταξύ δύο μεταβλητών δε σημαίνει απαραίτητα ότι υπάρχει και σχέση αιτιότητας.**

Δύο μεταβλητές ενδέχεται να έχουν υψηλή συσχέτιση **χωρίς**, όμως, η μεταβολή των τιμών της μίας να προκαλεί τη μεταβολή των τιμών της άλλης. Δηλαδή δεν είναι απαραίτητο να υπάρχει σχέση «αιτία - αποτέλεσμα», αλλά η σχέση τους ενδέχεται να είναι συμπτωματική ή να υπάρχει ένας τρίτος παράγοντας που να επηρεάζει τις τιμές και των δύο μεταβλητών που εξετάζουμε.

Συσχέτιση	VS	Αιτιότητα
σημαίνει ότι υπάρχει μια σχέση που συνδέει τις τιμές δύο μεταβλητών.		σημαίνει ότι η μεταβολή των τιμών μιας μεταβλητής προκαλεί τη μεταβολή στις τιμές μιας άλλης μεταβλητής.

Όλα αυτά με προβλημάτισαν και, ψάχνοντας στο διαδίκτυο, βρήκα ένα πολύ διασκεδαστικό παράδειγμα.

Τι συμβαίνει; Η κατανάλωση παγωτού προκαλεί εγκαύματα ή τα εγκαύματα ανοίγουν την όρεξη για παγωτό ή μήπως υπάρχει κάποιος τρίτος παράγοντας πίσω από αυτήν τη σχέση;

Επεξήγηση

Τους καλοκαιρινούς μήνες, λόγω της υψηλής θερμοκρασίας, πηγαίνουμε στη θάλασσα για να δροσιστούμε. Αν όμως δεν προστατευτούμε σωστά από τον ήλιο, παθαίνουμε ηλιακό έγκαυμα.

Από την άλλη, η αύξηση της θερμοκρασίας μας οδηγεί, επίσης, στην κατανάλωση παγωτών για να δροσιστούμε. Αυτό σημαίνει αύξηση των πωλήσεων παγωτού.

2.6.α

Συμπληρώστε τα κενά, στην παρακάτω πρόταση:

Από τα παραπάνω καταλαβαίνουμε ότι η αύξηση και στις δύο μεταβλητές (πωλήσεις παγωτού και πλήθος των περιστατικών ηλιακών εγκαυμάτων) οφείλεται στην αύξηση της θερμοκρασίας, η οποία είναι ο (i) παράγοντας που δημιουργεί την υψηλή (ii) μεταξύ των δύο μεταβλητών που εξετάζουμε.

θα μπορούσες να σκεφτείς ένα αντίστοιχο παράδειγμα για να το συζητήσετε στην τάξη σου;

2.7
Ακραίες τιμές

2.7.α

Για τις τιμές των μεταβλητών X και Y, που περιλαμβάνονται στον Πίνακα 4, δημιουργήστε το **διάγραμμα διασποράς**:

ΠΙΝΑΚΑΣ 4

X	Y
2	2
2	3
2	4
3	2
3	3
3	4
4	2
4	3
4	4
10	10

2.7.β

Κάντε τους απαραίτητους υπολογισμούς και συμπληρώστε τον Πίνακα 5.
Στη συνέχεια, υπολογίστε τον Συντελεστή Γραμμικής Συσχέτισης r, εφαρμόζοντας τον τύπο

$$r = \frac{\Sigma_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\Sigma_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\Sigma_{i=1}^{n} (y_i - \bar{y})^2}}$$, και συμπληρώστε τα παρακάτω κενά:

ΠΙΝΑΚΑΣ 5

ΕΓΓΡΑΦΕΣ	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})\cdot(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	2	2			2,89	2,89	
2	2	3			1,19		
3	2	4			-0,51		
4	3	2			1,19	0,49	
5	3	3			0,49		
6	3	4			-0,21		
7	4	2			-0,51	0,09	
8	4	3			-0,21		
9	4	4			0,09		
10	10	10			39,69	39,69	
ΑΘΡΟΙΣΜΑ	**37**	**37**			**44,1**	**50,1**	

Υπολογισμοί

- Δειγματική μέση τιμή των μεταβλητών X και Y

$$\bar{x} = \frac{\quad}{\quad} = \frac{\quad}{\quad} = \quad \text{(i)}$$

$$\bar{y} = \frac{\quad}{\quad} = \frac{\quad}{\quad} = \quad \text{(ii)}$$

- Το άθροισμα των αποκλίσεων x_i από τη δειγματική μέση τιμή $\bar{x}$

$$\Sigma_{i=1}^{n} (x_i - \bar{x})^2 = \quad \text{(iii)}$$

- Το άθροισμα των αποκλίσεων y_i από τη δειγματική μέση τιμή $\bar{y}$

$$\Sigma_{i=1}^{n} (y_i - \bar{y})^2 = \quad \text{(iv)}$$

- Στον Πίνακα 5, έχουμε υπολογίσει ότι: $\Sigma_{i=1}^{n} (x_i - \bar{x}) (y_i - \bar{y}) = 44{,}1$

- Αντικαθιστούμε τις παραπάνω τιμές στον τύπο
(με στρογγυλοποίηση στο 3ο δεκαδικό ψηφίο)

$$r = \frac{\Sigma_{i=1}^{n} (x_i - \bar{x}) (y_i - \bar{y})}{\sqrt{\Sigma_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\Sigma_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{\quad}{\quad} = \frac{\quad}{\quad} \approx \quad \text{(v)}$$

και βρίσκουμε ότι ο **Συντελεστής Γραμμικής Συσχέτισης *r*** είναι 0,880.

Από την τιμή του **Συντελεστή Γραμμικής Συσχέτισης *r* =** (vi) που υπολογίσαμε, φαίνεται ότι μεταξύ των δύο μεταβλητών X και Y υπάρχει (vii) θετική συσχέτιση, σύμφωνα με όσα είπαμε παραπάνω.

Πράγματι, σύμφωνα με τον αλγεβρικό υπολογισμό, φαίνεται ότι η συσχέτιση είναι ισχυρή. Παρατηρώντας, όμως, τη γραφική παράσταση των σημείων, ερμηνεύεται το ίδιο και γραφικά; Ας υπολογίσουμε τι θα συμβεί, αν εξαιρέσουμε την παρατήρηση που είναι πιο απομακρυσμένη από όλες τις υπόλοιπες.

Εξαιρώντας το ζευγάρι τιμών που βρίσκεται απομακρυσμένο από τα άλλα σημεία, δηλαδή το σημείο (10,10), προκύπτει ο Πίνακας 6 και το αντίστοιχο **διάγραμμα διασποράς** των δύο μεταβλητών X και Y.

ΠΙΝΑΚΑΣ 6

X	Y
2	2
2	3
2	4
3	2
3	3
3	4
4	2
4	3
4	4

2.7.γ

Κάντε τους απαραίτητους υπολογισμούς και συμπληρώστε τον Πίνακα 7.
Στη συνέχεια, υπολογίστε τον Συντελεστή Γραμμικής Συσχέτισης r, εφαρμόζοντας τον τύπο

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$, και συμπληρώστε τα παρακάτω κενά:

ΠΙΝΑΚΑΣ 7

ΕΓΓΡΑΦΕΣ	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})\cdot(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	2	2					
2	2	3					
3	2	4					
4	3	2					
5	3	3					
6	3	4					
7	4	2					
8	4	3					
9	4	4					
ΑΘΡΟΙΣΜΑ	**27**	**27**					

Υπολογισμοί

- Δειγματική μέση τιμή των μεταβλητών X και Y

$$\bar{x} = \frac{\quad}{\quad} = \frac{\quad}{\quad} = \quad \text{(i)}$$

$$\bar{y} = \frac{\quad}{\quad} = \frac{\quad}{\quad} = \quad \text{(ii)}$$

- Το άθροισμα των αποκλίσεων x_i από τη δειγματική μέση τιμή $\bar{x}$

$$\Sigma_{i=1}^{n} (x_i - \bar{x})^2 = \quad \text{(iii)}$$

- Το άθροισμα των αποκλίσεων y_i από τη δειγματική μέση τιμή $\bar{y}$

$$\Sigma_{i=1}^{n} (y_i - \bar{y})^2 = \quad \text{(iv)}$$

- Στον Πίνακα 7, έχουμε υπολογίσει ότι: $\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 0$

- Αντικαθιστούμε τις παραπάνω τιμές στον τύπο

$$r = \frac{\Sigma_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\Sigma_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{\quad}{\quad} = \frac{\quad}{\quad} \approx \quad \text{(v)}$$

και βρίσκουμε ότι ο **Συντελεστής Γραμμικής Συσχέτισης *r*** είναι 0.

Εξαιρώντας το απομακρυσμένο σημείο, βρήκαμε ότι ο **Συντελεστής Γραμμικής Συσχέτισης *r*** είναι 0, που σημαίνει ότι (vi) υπάρχει γραμμική (vii) μεταξύ των δύο μεταβλητών X και Y.

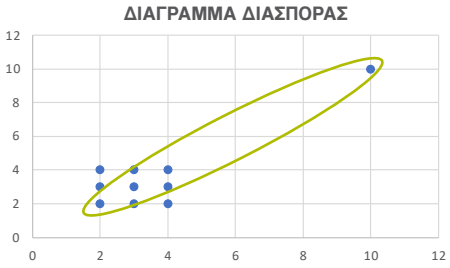
Δηλαδή, όταν αποκλείσουμε από τα δεδομένα μας την τιμή (10,10), ο **Συντελεστής Γραμμικής Συσχέτισης** γίνεται ***r* =** 0.

Όπως φαίνεται στο παράδειγμά μας, ο Συντελεστής Γραμμικής Συσχέτισης επηρεάζεται πολύ από τις τιμές που βρίσκονται πολύ μακριά, σε σχέση με το υπόλοιπο σύνολο των τιμών μας. Οι τιμές αυτές ονομάζονται ακραίες.

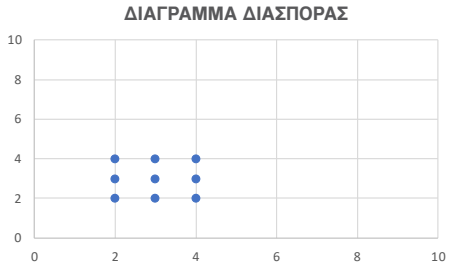
Αυτό συμβαίνει, επειδή ο υπολογισμός του Συντελεστή Συσχέτισης *r* βασίζεται στη μέση τιμή και την τυπική απόκλιση των δύο μεταβλητών, οι οποίες επηρεάζονται από την παρουσία ακραίων τιμών.

Όσο λιγότερες είναι οι παρατηρήσεις που έχουμε τόσο μεγαλύτερη είναι η επίδραση των ακραίων τιμών.

r **= 0,88**



r **= 0**



Μαθαίνω

Ακραία τιμή είναι μια τιμή ενός συνόλου δεδομένων, η οποία είναι ασυνήθιστη σε σύγκριση με τις περισσότερες τιμές του συνόλου των δεδομένων που εξετάζουμε.

Επομένως, πριν ερμηνεύσουμε τη συσχέτιση δύο μεταβλητών, θα πρέπει πρώτα να εξετάσουμε τις τυχόν ακραίες τιμές.

- Αν οι ακραίες τιμές αντιστοιχούν σε σωστές μετρήσεις (δηλαδή δεν οφείλονται σε πιθανό σφάλμα μέτρησης ή καταχώρισης των δεδομένων στη βάση μας), μπορούν να μας οδηγήσουν σε σωστά συμπεράσματα και να ανακαλύψουμε περισσότερα για τη σχέση των δύο μεταβλητών μας.
- Αν οι ακραίες τιμές είναι λάθη στο σύνολο δεδομένων, τότε ενδέχεται να δημιουργήσουν ψευδείς συσχετίσεις και είναι σωστό να τις αφαιρέσουμε. Σε αυτήν την περίπτωση, δεν παραλείπουμε να αναφέρουμε τις ακραίες τιμές, καθώς και τον λόγο για τον οποίο τις αφαιρέσαμε.

Δεν παραλείπουμε να απεικονίζουμε γραφικά τα δεδομένα μας, ώστε να μας αποκαλύπτονται τα χαρακτηριστικά τους που δεν είναι πάντα εμφανή κατά την αλγεβρική επεξεργασία τους.

στατιστική ΠΑΙΔΕΙΑ

ΕΚΠΑΙΔΕΥΤΙΚΑ ΤΕΤΡΑΔΙΑ

Επισκεφθείτε την ιστοσελίδα της ΕΛΣΤΑΤ
https://www.statistics.gr/el/edu-publications
για να βρείτε και τις άλλες εκδόσεις «Οι αριθμοί και η ζωή μας».

Το ΤΕΥΧΟΣ Ι
απευθύνεται σε παιδιά των τεσσάρων πρώτων τάξεων του Δημοτικού, με στόχο να αποκτήσουν οι μαθητές μια πρώτη επαφή με εναλλακτικές μεθόδους απεικόνισης των στατιστικών στοιχείων και, λύνοντας ασκήσεις, να προβληματιστούν με ευχάριστο τρόπο.

Το ΤΕΥΧΟΣ ΙΙ
είναι σχεδιασμένο για τους μαθητές των τελευταίων τάξεων του Δημοτικού και των πρώτων τάξεων του Γυμνασίου. Σε αυτό παρουσιάζονται επίσημα στατιστικά στοιχεία της ΕΛΣΤΑΤ με τη σύγχρονη μορφή απεικόνισης της παραγόμενης πληροφορίας, τα infographics.

Το ΤΕΥΧΟΣ ΙΙΙ
απευθύνεται σε μαθητές της δευτεροβάθμιας εκπαίδευσης. Καλύπτει, μεταξύ άλλων, τα μέτρα θέσης, τα μέτρα διασποράς και τον υπολογισμό του συντελεστή μεταβλητότητας, μέσω δραστηριοτήτων, παιχνιδιών, ασκήσεων και με τη χρήση infographics.

Το ΤΕΥΧΟΣ IV
απευθύνεται σε μαθητές Λυκείου. Περιλαμβάνει ασκήσεις και συνοπτική θεωρία στις εξής θεματικές ενότητες: πιθανότητες και συνδυαστική (διατάξεις, μεταθέσεις, συνδυασμοί).

Το ΤΕΥΧΟΣ V
απευθύνεται σε μαθητές Λυκείου. Καλύπτει τα θέματα: δειγματοληψία και δειγματοληπτικές έρευνες, γραμμική συσχέτιση και γραμμική παλινδρόμηση.

3. ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Στην Ενότητα 2, παρουσιάσαμε πώς μπορούμε να διερευνήσουμε αν υπάρχει κάποια σχέση μεταξύ δύο ποσοτικών μεταβλητών και, αν ναι, τι είδους και ποιας έντασης.

Σε αυτήν την Ενότητα, θα διερευνήσουμε **αν υπάρχει τρόπος πρόβλεψης των τιμών μίας μεταβλητής, γνωρίζοντας τις τιμές της άλλης**.

Πιο συγκεκριμένα, θα χρησιμοποιήσουμε την ανάλυση παλινδρόμησης για να δημιουργήσουμε ένα μοντέλο που θα περιγράφει τη σχέση των δύο μεταβλητών.

Συσχέτιση και Παλινδρόμηση

Στις Ενότητες 2 και 3, εξετάζουμε τη συσχέτιση και την εξίσωση της ευθείας παλινδρόμησης μεταξύ δύο μεταβλητών. Ο Συντελεστής Γραμμικής Συσχέτισης μετράει την ένταση της γραμμικής σχέσης μεταξύ δύο μεταβλητών, ενώ η ευθεία παλινδρόμησης προσδιορίζει τη σχέση εξάρτησης μεταξύ δύο μεταβλητών. Οι δύο αυτές διαδικασίες σχετίζονται και μπορούν να λειτουργούν συμπληρωματικά.

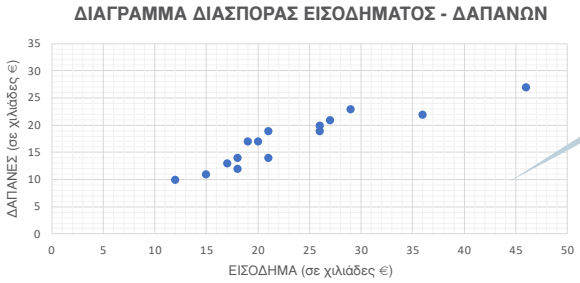
3.1 Προσεγγίζοντας την ευθεία παλινδρόμησης

Στον Πίνακα 1 που ακολουθεί, μεταφέραμε τα στοιχεία που χρησιμοποιήσαμε στην Ενότητα 2. Πρόκειται για 15 εγγραφές από την ΕΟΠ 2019, που αφορούν σε τύπο νοικοκυριού «4μελής οικογένεια με 2 παιδιά ηλικίας έως 15 ετών». Παρουσιάζονται 2 συγκεντρωτικές μεταβλητές για αυτές τις εγγραφές, το «ΕΙΣΟΔΗΜΑ» και οι «ΔΑΠΑΝΕΣ». Για αυτές τις μεταβλητές, είχαμε εξετάσει αν υπάρχει κάποια γραμμική σχέση μεταξύ τους και, αν υπάρχει, τι είδους και ποιας έντασης.

ΠΙΝΑΚΑΣ 1

	Ανεξάρτητη μεταβλητή X	Εξαρτημένη μεταβλητή Y
ΕΓΓΡΑΦΕΣ	ΕΙΣΟΔΗΜΑ ΣΤΡΟΓΓΥΛΟΠΟΙΗΜΕΝΟ (σε χιλιάδες €)	ΔΑΠΑΝΕΣ ΣΤΡΟΓΓΥΛΟΠΟΙΗΜΕΝΕΣ (σε χιλιάδες €)
1	46	27
2	36	22
3	29	23
4	26	19
5	26	20
6	27	21
7	21	19
8	21	14
9	20	17
10	19	17
11	18	12
12	18	14
13	17	13
14	15	11
15	12	10

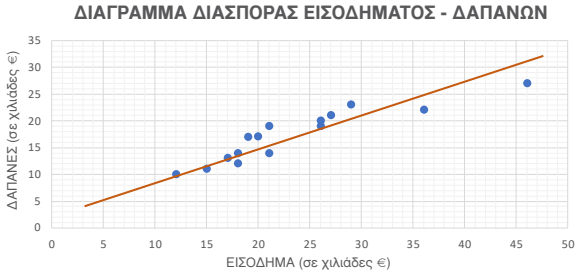
ΔΙΑΓΡΑΜΜΑ 1



Το διάγραμμα αναπαριστά τις «ΔΑΠΑΝΕΣ» σε σχέση με το «ΕΙΣΟΔΗΜΑ», για 15 εγγραφές από την ΕΟΠ 2019.

Παρατηρώντας το Διάγραμμα 1, βλέπουμε ότι τα σημεία είναι συγκεντρωμένα γύρω από μία νοητή ευθεία, την οποία έχω φέρει «με το μάτι» (κόκκινη ευθεία), στο Διάγραμμα 2.

ΔΙΑΓΡΑΜΜΑ 2



Μαθαίνω

Η **ευθεία παλινδρόμησης** προσδιορίζει τη σχέση εξάρτησης μεταξύ δύο μεταβλητών X (ανεξάρτητη) και Y (εξαρτημένη).

Το $\boldsymbol{b}$ μας δίνει το σημείο $(0, b)$ όπου η ευθεία τέμνει τον άξονα $y'y$, ενώ το $\boldsymbol{a}$ μας δίνει την κλίση της ευθείας.

Θυμόμαστε ότι η ευθεία έχει εξίσωση της μορφής $\boldsymbol{y=ax+b}$. Για να βρούμε την εξίσωση, λοιπόν, θα πρέπει να υπολογίσουμε τις παραμέτρους της εξίσωσης $\boldsymbol{a}$ και $\boldsymbol{b}$. Πώς θα μπορούσαμε να εργαστούμε;

Μπορούμε να επιλέξουμε δύο σημεία πάνω στην ευθεία, για παράδειγμα το Α(12,10) και το Β(17,13), και να αντικαταστήσουμε τις τιμές στην εξίσωση $\boldsymbol{y=ax+b}$ της ευθείας, ώστε να προκύψει ένα σύστημα δύο εξισώσεων με δύο αγνώστους.

3.1.α

Επιλύστε το σύστημα εξισώσεων και προσδιορίστε την εξίσωση της ευθείας:

$$\left.\begin{aligned} 10 &= a \cdot 12 + b \\ 13 &= a \cdot 17 + b \end{aligned}\right\}$$

Και τα αποτελέσματα είναι: $a =$ (i) και $b =$ (ii)

Οπότε, η εξίσωση της ευθείας που φαίνεται να προσαρμόζεται καλύτερα στα δεδομένα μας παίρνει την εξής μορφή: $y=$ **0,6** $\cdot x +$ **2,8**. Αυτό σημαίνει ότι:

- η ευθεία τέμνει τον άξονα $y'y$ στο σημείο (iii),
- ενώ η κλίση της ευθείας είναι (iv).

Βρήκαμε την εξίσωση της ευθείας που μοιάζει να προσεγγίζει καλύτερα τα σημεία που αποτελούν το διάγραμμα διασποράς.

Πώς μπορεί να μας φανεί χρήσιμη η εξίσωση αυτή;

Χρησιμοποιώντας την εξίσωση αυτή, μπορούμε να εκτιμήσουμε τιμές της μεταβλητής Y από τιμές της μεταβλητής X.

3.1.6

Αν μια 4μελής οικογένεια με 2 παιδιά έως 15 ετών έχει εισόδημα 25 χιλιάδες €, πόσες χιλιάδες € είναι οι δαπάνες που **αναμένουμε** να έχει;

Για να υπολογίσουμε τις δαπάνες που αναμένουμε να έχει αυτό το νοικοκυριό, θα χρησιμοποιήσουμε την εξίσωση της ευθείας που βρήκαμε και θα τη λύσουμε για τιμή του εισοδήματος ίση με 25 χιλιάδες €.

$y = 0{,}6 \cdot x + 2{,}8$, για $x=25$ έχουμε
$y =$ (i)
$y =$ (ii)

Άρα ένα νοικοκυριό που έχει εισόδημα 25 χιλιάδες €, **αναμένουμε** να έχει δαπάνες (iii) χιλιάδες €.

Θα μπορούσαμε να εκτιμήσουμε τις δαπάνες που αναμένουμε να έχει ένα νοικοκυριό με εισόδημα 200 χιλιάδες €;

Ναι! Αντικαθιστούμε στην ευθεία μας όπου $x=200$ και εκτιμούμε τις αναμενόμενες δαπάνες $y=122{,}8$ χιλιάδες €.

Προσοχή! Σε αυτήν την περίπτωση δεν μπορούμε να χρησιμοποιήσουμε τη γραμμική σχέση (εξίσωση ευθείας) που βρήκαμε, επειδή:

α) δεν γνωρίζουμε αν το νοικοκυριό είναι 4μελής οικογένεια με 2 παιδιά έως 15 ετών και
β) η τιμή x=200 είναι πολύ μακριά από το εύρος τιμών που εξετάσαμε.

Η εξίσωση της ευθείας που βρήκαμε βρίσκει εφαρμογή κοντά στο εύρος τιμών των παρατηρήσεων της μεταβλητής X.

Η ευθεία $y = \mathbf{0{,}6} \cdot x + \mathbf{2{,}8}$ που φέραμε, και η οποία μας φαίνεται ότι προσαρμόζεται καλύτερα στα δεδομένα μας, είναι αποτέλεσμα υποκειμενικής επιλογής. Αυτό σημαίνει ότι κάποιος άλλος, για το ίδιο διάγραμμα διασποράς, ίσως θα επέλεγε μια διαφορετική ευθεία.

Γι' αυτό δημιουργείται η ανάγκη για πιο αντικειμενικό και ακριβή υπολογισμό των συντελεστών της εξίσωσης!

Μια μέθοδος για την εκτίμηση των συντελεστών της εξίσωσης είναι η «**μέθοδος των ελαχίστων τετραγώνων**».

3.2 Η μέθοδος των ελαχίστων τετραγώνων

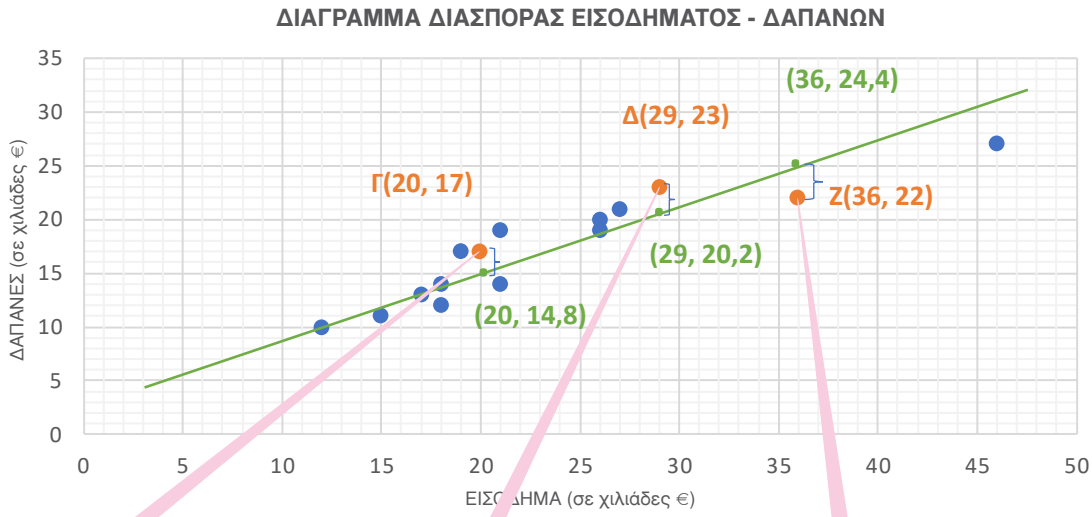
3.2.α

Συμπληρώστε τα κενά στις προτάσεις που ακολουθούν:

Ανακεφαλαίωση

- Η εξίσωση που αναζητάμε είναι της μορφής $y=ax+b$
- X _____ (i) μεταβλητή
- Y _____ (ii) μεταβλητή
- b το σημείο (0, b) όπου η ευθεία τέμνει τον άξονα _____ (iii)
- a είναι η _____ (iv) της ευθείας.

Παρατηρούμε ότι για κάθε παρατήρηση της μεταβλητής X υπάρχει διαφορά (σφάλμα) ανάμεσα στην καταγεγραμμένη τιμή της μεταβλητής Y και στην εκτίμησή της μέσω της ευθείας παλινδρόμησης.



Για το νοικοκυριό (εγγραφή 9) με εισόδημα 20 χιλιάδες €, έχουμε, από τα στοιχεία της ΕΟΠ, ότι οι δαπάνες είναι 17 χιλιάδες €, ενώ σύμφωνα με την ευθεία που φέραμε, οι δαπάνες του αναμένεται να είναι 14,8 χιλιάδες €.

Για το νοικοκυριό (εγγραφή 3) με εισόδημα 29 χιλιάδες €, έχουμε, από τα στοιχεία της ΕΟΠ, ότι οι δαπάνες είναι 23 χιλιάδες €, ενώ σύμφωνα με την ευθεία που φέραμε, οι δαπάνες του αναμένεται να είναι 20,2 χιλιάδες €.

Για το νοικοκυριό (εγγραφή 2) με εισόδημα _____ (v) χιλιάδες €, έχουμε, από τα στοιχεία της ΕΟΠ, ότι οι δαπάνες είναι _____ (vi) χιλιάδες €, ενώ σύμφωνα με την ευθεία που φέραμε, οι δαπάνες του αναμένεται να είναι _____ (vii) χιλιάδες €.

Σφάλμα είναι η διαφορά ανάμεσα στην καταγεγραμμένη τιμή της μεταβλητής Y και στην εκτίμηση της μέσω της ευθείας

3.2.6

Υπολογίστε το σφάλμα στις τρεις εγγραφές:

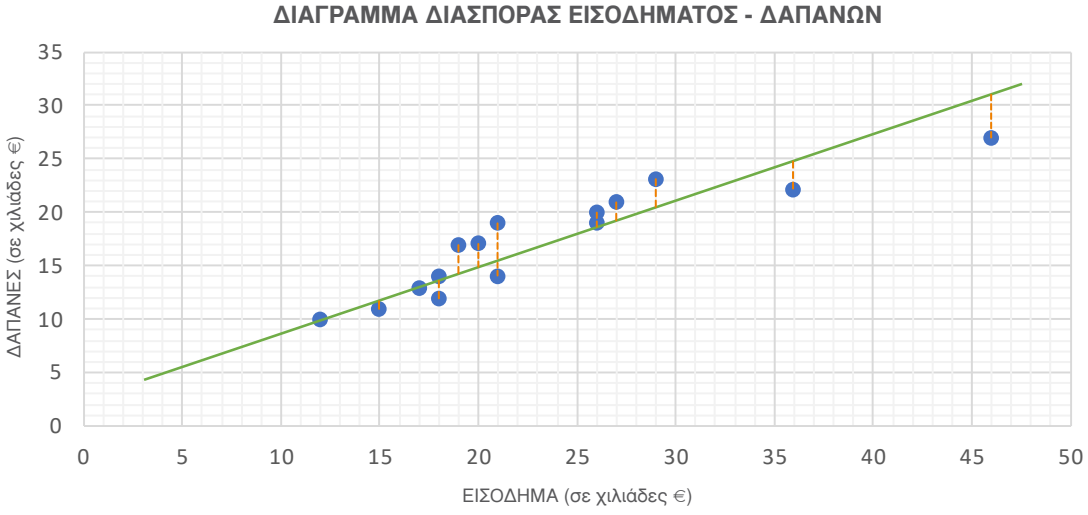
Για $x_9 = 20$, το σφάλμα υπολογίζεται: $e_9 = 17 - 14{,}8 = 2{,}2$
Για $x_3 = 29$, το σφάλμα υπολογίζεται: $e_3 =$ - = (i)
Για $x_2 = 36$, το σφάλμα υπολογίζεται: $e_2 =$ - = (ii)

Άρα παρουσιάζεται η ανάγκη για την εύρεση κάποιας μεθόδου που προσαρμόζεται καλύτερα στα δεδομένα μας (όσο το δυνατόν μικρότερα σφάλματα).

Επιδιώκουμε να βρούμε την εξίσωση της ευθείας για την οποία το σφάλμα (e) είναι όσο το δυνατόν μικρότερο. Επειδή τα επιμέρους σφάλματα (e_i) μπορεί να έχουν θετικό ή αρνητικό πρόσημο, προσπαθούμε να ελαχιστοποιήσουμε **όχι το άθροισμά τους, αλλά το άθροισμα των (μη αρνητικών) τετραγώνων των σφάλματων**.

Στην εξίσωση της ευθείας $\boldsymbol{y = ax + b}$ (1), η ύπαρξη σφάλματος απεικονίζεται με την προσθήκη ενός όρου που τον συμβολίζουμε με $\boldsymbol{e}$.
Οπότε, η νέα μορφή της εξίσωσης (1) είναι $\boldsymbol{y = ax + b + e}$ (2).
Λύνοντας τη (2) ως προς $\boldsymbol{e}$, προκύπτει η σχέση $\boldsymbol{e = y - ax - b}$ (3).

Όπως είπαμε, αναζητούμε μια εξίσωση ευθείας, η οποία να προσεγγίσει τα σημεία που αποτελούν το διάγραμμα διασποράς και όπου τα σφάλματα (δηλαδή η κατακόρυφη απόκλιση κάθε σημείου από την ευθεία) να είναι όσο το δυνατόν μικρότερα.



Ας γενικεύσουμε όσα ειπώθηκαν για n σημεία.

Αφού $e_i = y_i - ax_i - b$ είναι το σφάλμα, δηλαδή η κατακόρυφη απόκλιση κάθε σημείου από την ευθεία, το τετράγωνο κάθε σφάλματος είναι $e_i^2 = (y_i - ax_i - b)^2$ (4) και το άθροισμα των τετραγώνων όλων των σφαλμάτων εκφράζεται από τη σχέση:

$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2$ (5)

Επιθυμούμε το άθροισμα των τετραγώνων των σφαλμάτων $\sum_{i=1}^{n} e_i^2$ να γίνεται ελάχιστο έτσι, ώστε η εξίσωση της ευθείας να προσεγγίζει καλύτερα τα σημεία του διαγράμματος διασποράς.

Εκτιμήτριες ελαχίστων τετραγώνων

Για να ελαχιστοποιήσουμε το άθροισμα $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2$, σκεφτόμαστε ως εξής:

Εφόσον x_i και y_i συμβολίζουν τις τιμές που παίρνουν οι μεταβλητές μας, μπορούμε να εργαστούμε με τις παραμέτρους της εξίσωσης (το a και το b) και να τις «εκτιμήσουμε» έτσι, ώστε η ευθεία που απεικονίζει την εξίσωση να είναι η καλύτερη περιγραφή της σχέσης που υπάρχει μεταξύ των μεταβλητών X και Y.

Οι τιμές των παραμέτρων α και b, για τις οποίες ελαχιστοποιείται το άθροισμα, ονομάζονται **εκτιμήτριες ελαχίστων τετραγώνων** και τις συμβολίζουμε $\hat{a}$ και $\hat{b}$ (δηλαδή «καπελώνοντας» τις παραμέτρους a και b, παίρνουμε τις εκτιμήτριες). Για τον υπολογισμό τους, χρησιμοποιούνται οι τύποι:

$$\hat{a} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

και $\hat{b} = \bar{y} - \hat{a}\,\bar{x}$

όπου $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Το $\hat{a}$ εκφράζει το πόσο περιμένουμε να μεταβληθεί η αναμενόμενη τιμή της Y, αν η X αυξηθεί κατά μία μονάδα.

Το $\hat{b}$ εκφράζει την αναμενόμενη τιμή της Y, όταν η X παίρνει την τιμή 0. Στην περίπτωση που η τιμή $X = 0$ δεν έχει νόημα, τότε η τιμή του $\hat{b}$ δεν έχει ερμηνεία.

Μαθαίνω

Η εξίσωση της ευθείας $\hat{y} = \hat{a}x + \hat{b}$ ονομάζεται **ευθεία ελαχίστων τετραγώνων** ή **ευθεία παλινδρόμησης**, η οποία παρέχει μια **εκτίμηση** της γραμμής παλινδρόμησης. Όπου:

$\hat{y}$ είναι η εκτίμηση της τιμής που θα πάρει η μεταβλητή Y,

$\hat{a}$ είναι η εκτίμηση της κλίσης της ευθείας,

$\hat{b}$ είναι η εκτίμηση του σημείου $(0, b)$, όπου η ευθεία τέμνει τον άξονα $y'y$.

3.2.γ

Στην Ενότητα 2, είχαμε κάνει τους εξής υπολογισμούς:

$\bar{x} = 23{,}4$ $\bar{y} = 17{,}3$ $\Sigma_{i=1}^{15}(x_i - \bar{x})^2 = 1070{,}2$ $\Sigma_{i=1}^{15}(x_i - \bar{x})(y_i - \bar{y}) = 551{,}1$

Βρείτε την εξίσωση παλινδρόμησης, χρησιμοποιώντας τη μέθοδο των ελαχίστων τετραγώνων.

- Αντικαθιστώ στους τύπους και βρίσκω τις **εκτιμήτριες ελαχίστων τετραγώνων** $\hat{a}$ και $\hat{b}$ (με στρογγυλοποίηση στο 3ο δεκαδικό ψηφίο):

$\hat{a} = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}$ οπότε $\hat{a} =$ _____ $\approx$ _____ (i)

και $\hat{b} = \bar{y} - \hat{a}\,\bar{x}$ οπότε $\hat{b} =$ _____ - _____ • _____ = _____ - _____ = _____ (ii)

- Η εξίσωση παλινδρόμησης (με στρογγυλοποίηση στο 3ο δεκαδικό) θα είναι της μορφής:

$\hat{y} =$ _____ $\cdot x +$ _____ (iii)

Η αύξηση του εισοδήματος ενός νοικοκυριού με 2 ενήλικες και 2 παιδιά (έως 15 ετών), κατά 1 χιλιάδα (1.000€), σημαίνει αύξηση της αναμενόμενης δαπάνης κατά 515€.

Ένα νοικοκυριό με 2 ενήλικες και 2 παιδιά (έως 15 ετών), χωρίς εισόδημα (X=0), αναμένεται να δαπανήσει 5.249€ (από δανεισμό, αποταμιεύσεις κ.ά.).

3.2.δ

Σε αυτό το σημείο, ας θυμηθούμε την εξίσωση ευθείας $y =$ _____ $\cdot x +$ _____ (i) που βρήκαμε, χρησιμοποιώντας τα δύο σημεία από το διάγραμμα διασποράς.
Τι παρατηρείτε;

Παρατηρούμε ότι

(ii)

Ας συνοψίσουμε!

Με τη μέθοδο της γραμμικής παλινδρόμησης δημιουργείται ένα μοντέλο (ευθεία παλινδρόμησης) που μας επιτρέπει να προβλέπουμε την αναμενόμενη τιμή της εξαρτημένης μεταβλητής Y, όταν είναι γνωστή η τιμή της ανεξάρτητης μεταβλητής X. Χρησιμοποιούμε τη μέθοδο των ελαχίστων τετραγώνων για να προσδιορίσουμε την εξίσωση της ευθείας παλινδρόμησης μεταξύ δύο μεταβλητών. Η ευθεία αυτή προσαρμόζεται καλύτερα στα δεδομένα μας, καθώς προσεγγίζει τα σημεία που αποτελούν το διάγραμμα διασποράς των δύο μεταβλητών με τέτοιο τρόπο, ώστε τα σφάλματα, δηλαδή η κατακόρυφη απόκλιση κάθε σημείου από την ευθεία, να είναι όσο το δυνατό μικρότερα.

Τι θα πρέπει να προσέχουμε, όταν εφαρμόζουμε την ευθεία παλινδρόμησης για να περιγράψουμε τη σχέση που συνδέει τις δύο μεταβλητές:

1. Η εξίσωση της ευθείας που βρήκαμε να έχει την καλύτερη προσαρμογή, δηλαδή να εφαρμόζει καλύτερα στα δεδομένα μας, μας δίνει καλή εκτίμηση όταν:

   α) η συσχέτιση είναι ισχυρή και

   β) το διάγραμμα διασποράς αποτελείται από επαρκή αριθμό σημείων.

   Δηλαδή:

   - Αν **τα σημεία των δεδομένων μας βρίσκονται πολύ κοντά στη ευθεία** που βρήκαμε ότι έχει την καλύτερη προσαρμογή και **η συσχέτιση είναι πολύ ισχυρή**, τότε η πρόβλεψη **τείνει να είναι ακριβής**.

   - Αν **τα σημεία των δεδομένων μας βρίσκονται μακριά από την ευθεία** που βρήκαμε ότι έχει την καλύτερη προσαρμογή και **η συσχέτιση είναι ασθενής**, τότε η πρόβλεψη **τείνει να μην είναι ακριβής**.

2. Η ευθεία παλινδρόμησης βοηθάει να κάνουμε **εκτιμήσεις** για σημεία που βρίσκονται **στο εύρος τιμών των παρατηρήσεων της ανεξάρτητης μεταβλητής**.

3. Μια ευθεία παλινδρόμησης, η οποία βασίζεται σε **παρελθοντικές** μετρήσεις, **ενδέχεται να μην έχει** καλή εφαρμογή σε **μελλοντικά** δεδομένα.

4. Η ευθεία παλινδρόμησης που προκύπτει από το διάγραμμα διασποράς ενός πληθυσμού **δεν εφαρμόζεται σε διαφορετικό πληθυσμό**.

5. Η ευθεία παλινδρόμησης έχει νόημα, όταν υπάρχει **σημαντική συσχέτιση** και όταν η **σχέση** που συνδέει τις μεταβλητές είναι **γραμμική**.

3.2.ε

Για τις προτάσεις που ακολουθούν, σημειώστε **Σ** αν είναι σωστές ή **Λ** αν είναι λανθασμένες:

	Πρόταση	**Σ ή Λ**
i.	Η γραμμική παλινδρόμηση προσδιορίζει την αιτία της σχέσης μεταξύ των μεταβλητών.	
ii.	Η μέθοδος ελαχίστων τετραγώνων χρησιμοποιείται για να βρεθούν οι εκτιμήτριες των παραμέτρων της ευθείας παλινδρόμησης.	
iii.	Η κλίση της ευθείας παλινδρόμησης αντιπροσωπεύει τη μεταβολή των τιμών της εξαρτημένης μεταβλητής ανά μονάδα μεταβολής της ανεξάρτητης μεταβλητής.	
iv.	Ο σταθερός όρος της ευθείας παλινδρόμησης αντιπροσωπεύει την αναμενόμενη τιμή της εξαρτημένης μεταβλητής, όταν η ανεξάρτητη μεταβλητή είναι μηδέν.	
v.	Η γραμμική παλινδρόμηση είναι μια μέθοδος για την εκτίμηση της σχέσης μεταξύ δύο ανεξάρτητων μεταβλητών.	
vi.	Η εξίσωση παλινδρόμησης χρησιμοποιείται για να προβλέψει τιμές της ανεξάρτητης μεταβλητής, βάσει της εξαρτημένης μεταβλητής.	
vii.	Η γραμμική παλινδρόμηση μπορεί να χρησιμοποιηθεί μόνο για μεταβλητές που έχουν γραμμική σχέση.	

στατιστική ΠΑΙΔΕΙΑ

ΕΠΙΜΕΡΟΥΣ ΔΡΑΣΕΙΣ ΣΤΑΤΙΣΤΙΚΗΣ ΠΑΙΔΕΙΑΣ

Ψηφιακή Πύλη για τη Στατιστική Παιδεία
www.statistics.gr/el/edu

European Master in Official Statistics
Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης (ΑΠΘ)
Οικονομικό Πανεπιστήμιο Αθηνών (ΟΠΑ)

Διοργάνωση του Πανελλήνιου Διαγωνισμού στη Στατιστική

Διαλέξεις Προέδρου στις εγκαταστάσεις της ΕΛΣΤΑΤ, σε σχολεία και πανεπιστήμια

Εκπαιδευτικές επισκέψεις / παρουσιάσεις

Πρακτική άσκηση και πτυχιακές εργασίες στην ΕΛΣΤΑΤ, για φοιτητές και σπουδαστές

«Απογραφή στο Σχολείο»

Εκπαιδευτικά τετράδια

Μνημόνια Συνεργασίας με Πανεπιστήμια

Παρουσιάσεις και ενημερώσεις στη ΔΕΘ

Συμμετοχή σε δράσεις της Eurostat
EMOS Cross-border Traineeships
(EMOS-Διασυνοριακή πρακτική άσκηση)

Εκπαιδευτικό παιχνίδι StatRun

Εκπαιδευτικά quiz διαδραστικά δημοσιεύματα και video

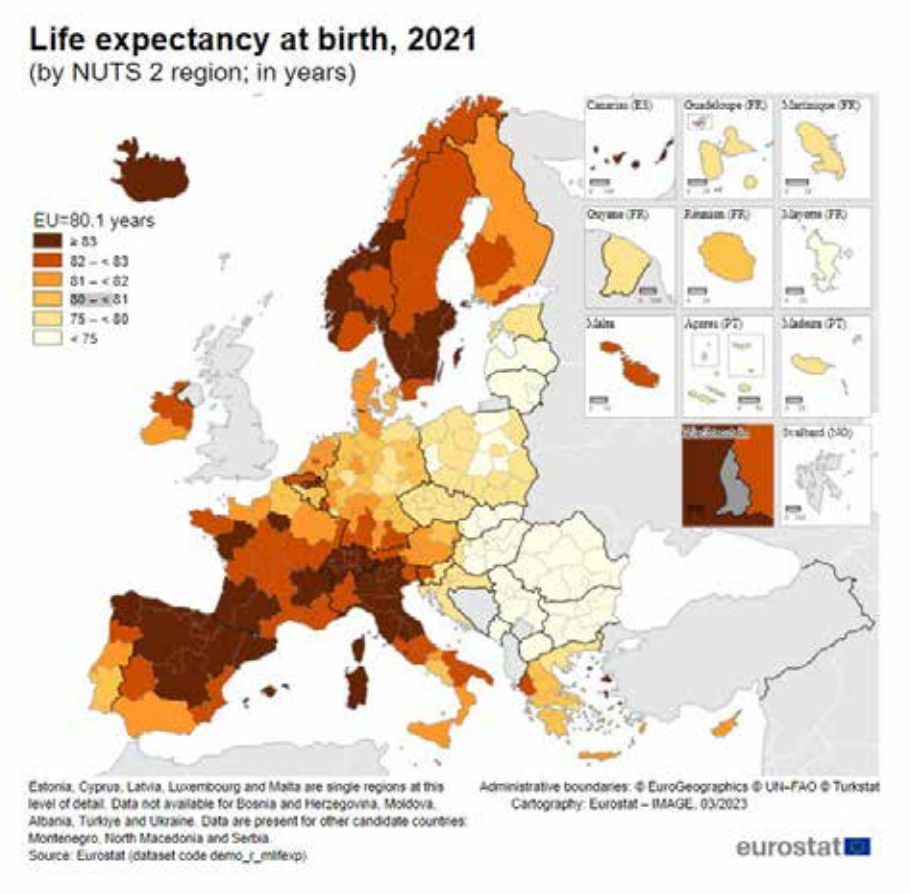
Φτιάξε την ομάδα σου και λάβε μέρος στον επόμενο Διαγωνισμό στη Στατιστική!

Μπορείς να βρεις περισσότερες πληροφορίες στην τοποθεσία:

https://www.statistics.gr/el/edu-statistics-competition

4. ΠΡΟΣΔΟΚΙΜΟ ΖΩΗΣ - ΕΠΑΝΑΛΗΠΤΙΚΗ ΕΝΟΤΗΤΑ

Το infographic που ακολουθεί δημοσιεύτηκε από τη Eurostat, τη Στατιστική Υπηρεσία της Ευρωπαϊκής Ένωσης (ΕΕ), τον Μάρτιο 2023, και αφορά στο προσδόκιμο ζωής κατά τη γέννηση (Life expectancy at birth) στις περιφέρειες των ευρωπαϊκών χωρών, για τις οποίες διατίθενται στοιχεία του έτους 2021.



Μαθαίνω

Το προσδόκιμο ζωής σε μια ορισμένη ηλικία είναι ο μέσος αριθμός επιπλέον ετών που μπορεί να αναμένεται να ζήσει ένα άτομο αυτής της ηλικίας, αν τα πρότυπα θνησιμότητας παραμείνουν αμετάβλητα σε όλη την υπόλοιπη ζωή του.

Το προσδόκιμο ζωής κατά τη γέννηση είναι ο μέσος αριθμός ετών που μπορεί να αναμένεται να ζήσει ένα νεογέννητο παιδί, αν τα πρότυπα θνησιμότητας παραμείνουν αμετάβλητα.

Το προσδόκιμο ζωής υπολογίζεται χωριστά για όλα τα επίπεδα ηλικίας, καθώς και για τους άνδρες, τις γυναίκες και το σύνολο του πληθυσμού.

Αυτό το γνώριζες;

Η Eurostat συγκεντρώνει στοιχεία και αποτελέσματα ερευνών από τις Εθνικές Στατιστικές Υπηρεσίες της ΕΕ, αλλά και άλλων χωρών, με τις οποίες η ΕΕ έχει συνεργασία. Δεδομένου ότι έχουν υιοθετηθεί κοινοί ορισμοί, πρότυπα και μεθοδολογίες, τα δημοσιεύματα της Eurostat επιτρέπουν τη σύγκριση μεταξύ των χωρών.

4.1 Infographic

4.1.α

Χρησιμοποιώντας τις πληροφορίες που αποτυπώνονται στο infographic, κυκλώστε τη σωστή απάντηση στις ερωτήσεις που ακολουθούν.

i. Στην Ελλάδα, το μεγαλύτερο προσδόκιμο ζωής κατά τη γέννηση, για το 2021, υπολογίστηκε για την Περιφέρεια:

Α. Ηπείρου Β. Βορείου Αιγαίου Γ. Νοτίου Αιγαίου

ii. Για όλες τις περιφέρειες της Βουλγαρίας, το προσδόκιμο ζωής κατά τη γέννηση, για το 2021, είναι:

Α. κάτω από 75 έτη Β. από 75 έως 80 έτη Γ. 83 ή περισσότερα έτη

iii. Για τις βόρειες περιφέρειες της Ισπανίας, το προσδόκιμο ζωής κατά τη γέννηση, για το 2021, είναι:

Α. κάτω από 75 έτη Β. από 75 έως 80 έτη Γ. 83 ή περισσότερα έτη

iv. Για το σύνολο της ΕΕ, το προσδόκιμο ζωής κατά τη γέννηση, για το 2021, είναι:

Α. 80,1έτη Β. από 75 έως 80 έτη Γ. 83 ή περισσότερα έτη

4.2 Διάγραμμα διασποράς Συσχέτιση

4.2.α

Στον Πίνακα 1, παρατίθενται τα στοιχεία για το προσδόκιμο ζωής κατά τη γέννηση, για το 2021, στις 27 χώρες της ΕΕ, χωριστά για τους άνδρες και τις γυναίκες, όπως ανακτήθηκαν από τη βάση της Eurostat, στις 10 Απριλίου 2023.

ΠΙΝΑΚΑΣ 1: ΠΡΟΣΔΟΚΙΜΟ ΖΩΗΣ ΚΑΤΑ ΤΗ ΓΕΝΝΗΣΗ, 2021

Χώρα	Κωδικός χώρας	Άνδρες (Χ)	Γυναίκες (Υ)	Χώρα	Κωδικός χώρας	Άνδρες (Χ)	Γυναίκες (Υ)
Βέλγιο	BE	79,4	84,3	Λιθουανία	LT	69,5	78,8
Βουλγαρία	BG	68,0	75,1	Λουξεμβούργο	LU	80,5	84,8
Τσεχία	CZ	74,1	80,5	Ουγγαρία	HU	70,7	77,8
Δανία	DK	79,6	83,3	Μάλτα	MT	80,8	84,3
Γερμανία	D	78,4	83,3	Ολλανδία	NL	79,7	83,0
Εσθονία	EE	72,7	81,4	Αυστρία	AT	78,8	83,7
Ιρλανδία	IE	80,5	84,3	Πολωνία	PL	71,6	79,6
Ελλάδα	EL	77,4	82,9	Πορτογαλία	PT	78,5	84,4
Ισπανία	ES	80,4	86,2	Ρουμανία	RO	69,2	76,6
Γαλλία	FR	79,3	85,5	Σλοβενία	SI	77,7	83,8
Κροατία	HR	73,6	79,8	Σλοβακία	SK	71,2	78,2
Ιταλία	IT	80,5	84,9	Φιλανδία	FI	79,3	84,6
Κύπρος	CY	79,2	83,4	Σουηδία	SE	81,3	84,9
Λετονία	LV	68,2	78,0				

Κυκλώστε τη σωστή απάντηση στις παρακάτω προτάσεις:

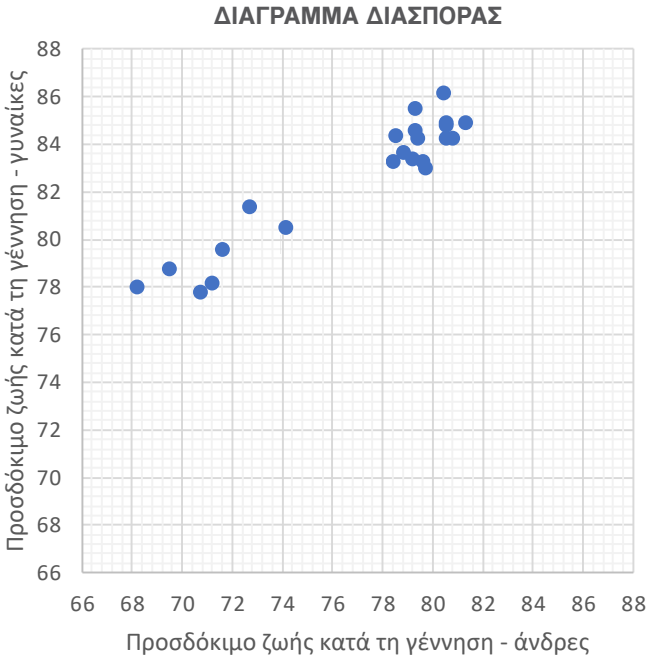
i. Το προσδόκιμο ζωής κατά τη γέννηση είναι **μεγαλύτερο** / **μικρότερο** για τους άνδρες συγκριτικά με τις γυναίκες, σε όλες τις χώρες της ΕΕ.

ii. Οι μεταβλητές «προσδόκιμο ζωής κατά τη γέννηση για τους άνδρες» και «προσδόκιμο ζωής κατά τη γέννηση για τις γυναίκες» είναι **ποσοτικές** / **ποιοτικές** μεταβλητές. Ειδικότερα, πρόκειται για **συνεχείς** / **διακριτές** μεταβλητές.

4.2.6

i. Θεωρώντας τη μεταβλητή «προσδόκιμο ζωής κατά τη γέννηση για τους άνδρες» ανεξάρτητη μεταβλητή (X) και «προσδόκιμο ζωής κατά τη γέννηση για τις γυναίκες» εξαρτημένη (Y), κατασκευάστηκε το διάγραμμα διασποράς που ακολουθεί (Διάγραμμα 1). Προσθέστε, στο διάγραμμα αυτό, τα στοιχεία για τις βαλκανικές χώρες της ΕΕ (BG, EL, HR, RO, SI), που έχουν παραλειφθεί.

ΔΙΑΓΡΑΜΜΑ 1



ii. Με βάση το διάγραμμα διασποράς, τι συμπεραίνετε για τη συσχέτιση των δύο μεταβλητών;

4.2.γ

Για τα στοιχεία του Πίνακα 1, έχει υπολογιστεί ότι:

- $\bar{x} = 76{,}300 \qquad \bar{y} = 82{,}126$
- $\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 343{,}970$
- $\Sigma_{i=1}^{n}(x_i - \bar{x})^2 = 527{,}820 \quad \Sigma_{i=1}^{n}(y_i - \bar{y})^2 = 244{,}092$

Υπολογίστε τον Συντελεστή Γραμμικής Συσχέτισης r, εφαρμόζοντας τον τύπο

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

και στη συνέχεια συμπληρώστε τα κενά, στην πρόταση που ακολουθεί:

$r =$ _____ $\approx$ _____ $\approx$ _____ (i)

Ο **Συντελεστής Γραμμικής Συσχέτισης *r*** είναι (ii), δηλαδή υπάρχει πολύ (iii) θετική συσχέτιση μεταξύ των δύο μεταβλητών.

Σε αυτήν την περίπτωση, θα μπορούσαμε να μιλήσουμε για σχέση «αιτίας - αποτελέσματος» μεταξύ των δύο μεταβλητών; Μπορείτε να σκεφθείτε κάποιους παράγοντες που επηρεάζουν το προσδόκιμο ζωής και των δύο φύλων;

4.3 Γραμμική παλινδρόμηση

4.3.α

Μεταφέρετε τα στοιχεία του Πίνακα 1 σε ένα φύλλο του Excel και ακολουθήστε τις οδηγίες που θα βρείτε στο Παράρτημα ΙΙ, για να απαντήσετε στα ερωτήματα που ακολουθούν:

i. Αναπαράγετε το διάγραμμα διασποράς (Διάγραμμα 1) των δύο μεταβλητών.

ii. Βρείτε την εξίσωση της ευθείας που προσαρμόζεται καλύτερα στα δεδομένα.

iii. Για την Αλβανία, το προσδόκιμο ζωής κατά τη γέννηση (2021) για τους άνδρες είναι 73,6 έτη. Χρησιμοποιείστε την ευθεία της παλινδρόμησης που βρήκατε στο (ii) για να εκτιμήσετε πόσο αναμένεται να είναι το προσδόκιμο ζωής για τις γυναίκες.

iv. Το προσδόκιμο ζωής κατά τα γέννηση (2021) για τις γυναίκες, στην Αλβανία, είναι 77,7 έτη. Υπολογίστε την απόλυτη τιμή του σφάλματος της εκτίμησης.

v. Για τη Νορβηγία, το προσδόκιμο ζωής κατά τη γέννηση (2021) για τους άνδρες είναι 81,7 έτη. Εκτιμήστε πόσο αναμένεται να είναι το προσδόκιμο ζωής για τις γυναίκες. Στη συνέχεια, υπολογίστε την απόλυτη τιμή του σφάλματος της εκτίμησης του προσδόκιμου ζωής κατά τη γέννηση (2021) για τις γυναίκες, λαμβάνοντας υπόψη ότι το προσδόκιμο ζωής για τις γυναίκες είναι 84,7 έτη.

ΣΗΜΕΙΩΣΕΙΣ ΘΕΩΡΙΑΣ

Μεταδεδομένα Τα μεταδεδομένα περιγράφουν στατιστικά δεδομένα, δίνοντας τους ορισμούς των πληθυσμών, των μεταβλητών, καθώς και πληροφορίες για τη μεθοδολογία και την ποιότητα των μεθόδων και των αποτελεσμάτων των ερευνών.

Πληθυσμός Είναι το σύνολο των μονάδων (ατόμων, νοικοκυριών, επιχειρήσεων κ.λπ.) που ενδιαφερόμαστε να εξετάσουμε ως προς ένα ή περισσότερα χαρακτηριστικά (π.χ. ηλικία, εισόδημα, κύκλος εργασιών).

Δείγμα Είναι το υποσύνολο του πληθυσμού που μελετάται, προκειμένου να εξαχθούν συμπεράσματα / αποτελέσματα για το σύνολο του πληθυσμού.

Δειγματοληψία Είναι η διαδικασία επιλογής των μονάδων του δείγματος (ατόμων, νοικοκυριών, επιχειρήσεων κ.λπ.) από τον πληθυσμό.

Δειγματοληπτικό πλαίσιο ή πλαίσιο δειγματοληψίας Είναι το μητρώο - κατάλογος των μονάδων του πληθυσμού, με τη βοήθεια του οποίου γίνεται η επιλογή του δείγματος.

Δειγματοληπτική μονάδα Είναι η μονάδα (στοιχείο ή συλλογή στοιχείων) που μπορεί να επιλεγεί σε κάποιο στάδιο της δειγματοληψίας.

Απογραφική έρευνα Είναι η έρευνα που διεξάγεται στο σύνολο του πληθυσμού.

Δειγματοληπτική έρευνα Είναι η έρευνα που διεξάγεται σε δείγμα του πληθυσμού.

Δείγμα πιθανότητας Είναι το δείγμα που έχει επιλεγεί με μεθόδους πιθανότητας, δηλαδή, όταν η κάθε μονάδα του πληθυσμού έχει μία καθορισμένη (μη μηδενική) πιθανότητα να συμπεριληφθεί στο δείγμα. Όταν έχουμε δείγμα πιθανότητας, μπορούμε να εφαρμόσουμε στατιστικές μεθόδους και τεχνικές για να εξάγουμε συμπεράσματα / αποτελέσματα από το δείγμα για το σύνολο του πληθυσμού.

Πραγματική τιμή παραμέτρου Η πραγματική τιμή μιας παραμέτρου του πληθυσμού (π.χ. μέση ηλικία, μέσο πλήθος ατόμων ανά νοικοκυριό, μέσο εισόδημα) μπορεί να προσδιοριστεί, συλλέγοντας στοιχεία από ολόκληρο τον πληθυσμό (απογραφή).

Εκτίμηση τιμής παραμέτρου Η εκτίμηση της τιμής μιας παραμέτρου του πληθυσμού γίνεται, όταν, με βάση τα στοιχεία ενός δείγματος, εκτιμούμε την τιμή της παραμέτρου αυτής.

Δειγματοληπτικό σφάλμα ή σφάλμα δειγματοληψίας Είναι η διαφορά μεταξύ της εκτίμησης που βασίζεται σε ένα δείγμα και της πραγματικής τιμής της παραμέτρου.

Μη δειγματοληπτικά σφάλματα Είναι τα σφάλματα που δεν οφείλονται στη δειγματοληψία, αλλά σε άλλους παράγοντες κατά τη διάρκεια του σχεδιασμού, της συλλογής ή της επεξεργασίας των στοιχείων της έρευνας. Τέτοια σφάλματα είναι τα σφάλματα μη απόκρισης, επεξεργασίας, κάλυψης και μέτρησης. Μη δειγματοληπτικά σφάλματα υπάρχουν τόσο στις απογραφικές όσο και στις δειγματοληπτικές έρευνες.

Απλός μέσος Απλός μέσος ενός συνόλου n παρατηρήσεων x_i είναι το άθροισμα $\sum x_i$ των παρατηρήσεων διά του πλήθους n των παρατηρήσεων: $\bar{x} = \frac{\sum x_i}{n}$

(Όπου $\sum$ η συντόμευση που παριστάνει το άθροισμα πολλών στοιχείων.)

Σταθμισμένος μέσος

Σταθμισμένος μέσος ενός συνόλου n παρατηρήσεων x_i, όπου η καθεμία έχει βαρύτητα (συντελεστή στάθμισης) w_i, είναι το άθροισμα $\sum w_i \cdot x_i$ των γινομένων των παρατηρήσεων με τους αντίστοιχους συντελεστές στάθμισης προς το άθροισμα $\sum w_i$ των συντελεστών στάθμισης: $\bar{x} = \frac{\sum w_i \cdot x_i}{\sum w_i}$

Μεταβλητές

Είναι τα διάφορα χαρακτηριστικά ως προς τα οποία εξετάζεται ένας πληθυσμός. Τέτοια χαρακτηριστικά είναι το φύλο, η ηλικία, το εισόδημα, ο αριθμός εργαζομένων, ο κύκλος εργασιών κ.λπ.

Διακρίνονται δύο είδη μεταβλητών:

1) οι **ποιοτικές**, των οποίων οι τιμές είναι κατηγορίες, π.χ. φύλο, επάγγελμα,
2) οι **ποσοτικές**, των οποίων οι τιμές είναι αριθμοί, π.χ. ο αριθμός των μελών σε ένα νοικοκυριό, η ποσότητα ψωμιού (σε γραμμάρια) που αγοράστηκε σε μια περίοδο αναφοράς.

   Οι **ποσοτικές μεταβλητές** διακρίνονται περαιτέρω σε:
   α) **διακριτές**, όταν παίρνουν «μεμονωμένες» τιμές, π.χ. αριθμός εργαζομένων,
   β) **συνεχείς**, οι οποίες μπορούν να πάρουν οποιαδήποτε τιμή μέσα σε ένα διάστημα τιμών, π.χ. βάρος, κύκλος εργασιών.

Συντελεστής Γραμμικής Συσχέτισης r

Είναι ένας συντελεστής που μετράει την ένταση της (γραμμικής) σχέσης μεταξύ δύο ποσοτικών μεταβλητών. Είναι καθαρός αριθμός, δηλαδή δεν εκφράζεται με κάποια μονάδα μέτρησης, και παίρνει τιμές από -1 έως και 1.

Τύποι υπολογισμού του Συντελεστή Γραμμικής Συσχέτισης *r* μεταξύ δύο ποσοτικών μεταβλητών X, Y:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \quad \text{ή εναλλακτικά} \quad r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{(n-1)s_X s_Y}$$

$\bar{x}$ η δειγματική μέση τιμή και s_X η δειγματική τυπική απόκλιση της μεταβλητής X

$\bar{y}$ η δειγματική μέση τιμή και s_Y η δειγματική τυπική απόκλιση της μεταβλητής Y

n το μέγεθος του δείγματος.

Μεταβλητές	Δειγματική μέση τιμή	Δειγματική τυπική απόκλιση
μεταβλητή X	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$	$s_X = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$
μεταβλητή Y	$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$	$s_Y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}$

Παρατηρούμε ότι στον τύπο της δειγματικής τυπικής απόκλισης διαιρούμε με το n-1, όπου n είναι το μέγεθος του δείγματος. Η χρήση του n - 1 στον τύπο της δειγματικής τυπικής απόκλισης προκύπτει από τους βαθμούς ελευθερίας της εκτίμησης της τυπικής απόκλισης. Αρχικά οι βαθμοί ελευθερίας, ήταν όσοι και το μέγεθος του δείγματος, δηλαδή n, αλλά χάθηκε ένας βαθμός ελευθερίας επειδή στον ορισμό της δειγματικής τυπικής απόκλισης περιλαμβάνεται η δειγματική μέση τιμή και, επομένως, οι n τιμές του δείγματος δεσμεύονται να δίνουν τη δειγματική μέση τιμή.

Ακραία τιμή

Είναι μια τιμή ενός συνόλου δεδομένων, η οποία είναι ασυνήθιστη, σε σύγκριση με τις περισσότερες τιμές του συνόλου των δεδομένων που εξετάζουμε.

Ευθεία παλινδρόμησης

Η ευθεία παλινδρόμησης προσδιορίζει τη σχέση εξάρτησης μεταξύ δύο μεταβλητών X (ανεξάρτητη) και Y (εξαρτημένη).

Μέθοδος των ελαχίστων τετραγώνων

Είναι μια μέθοδος που χρησιμοποιείται για την εκτίμηση των παραμέτρων a και b της εξίσωσης $y=ax+b$ της ευθείας παλινδρόμησης. Οι εκτιμήτριες ελαχίστων τετραγώνων των παραμέτρων a και b, που συμβολίζονται $\hat{a}$ και $\hat{b}$, αντίστοιχα, υπολογίζονται από τους τύπους:

$$\hat{a} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

και $\hat{b} = \bar{y} - \hat{a}\,\bar{x}$

Η εξίσωση της ευθείας $\hat{y} = \hat{a}x + \hat{b}$ ονομάζεται **ευθεία ελαχίστων τετραγώνων**.

5. ΛΥΣΕΙΣ

1

1.1.α
i. Β, **ii.** Β, **iii.** Β, **iv.** Α, **v.** Β, **vi.** Α

1.2.α
1. καταναλωτή, **2.** νοικοκυριό, **3.** 6.400, **4.** 1957/58, **5.** συλλογικές ... συμβιώσεις / κατοικίες

1.3.α
i. Α, **ii.** Δ, **iii.** Α, **iv.** Α, **v.** Δ

1.3.β
1.→ iii, 2.→ iv, 3.→ i, 4.→ ii, 5.→ v, 6.→ vi

1.3.γ
Α. Δειγματοληψία ευκολίας, **Β.** Δειγματοληψία χιονοστιβάδας, **Γ.** Συστηματική δειγματοληψία, **Δ.** Δειγματοληψία κατά συστάδες, **Ε.** Στρωματοποιημένη δειγματοληψία

1.3.δ
Απλή τυχαία δειγματοληψία, Δειγματοληψία κατά συστάδες, Στρωματοποιημένη δειγματοληψία, Συστηματική δειγματοληψία

1.4.α
i. Στην πιο πρόσφατη (ή τελευταία) Απογραφή Πληθυσμού - Κατοικιών
ii. Δισταδιακή στρωματοποιημένη δειγματοληψία
iii. 1. Γεωγραφική διαίρεση της χώρας (Περιφέρειες), **2.** Βαθμό αστικότητας
iv. 1. 10.000, **2.** 2.000 ≤ ... ≤ 9.999, **3.** 1.999
v. 79, **vi.** στρώμα, **vii.** επιλογής, **viii.** συστηματική, **ix.** νοικοκυριών / κατοικιών

1.4.β
5, 12, 19, 26

1.6.α
1. Δειγματοληψία, **2.** Ακρίβεια, **3.** Κόστος, **4.** Μεταβλητότητα, **5.** Πληθυσμός

1.7.α
1.→ ii, 2.→ iv, 3.→ iii, 4.→ i

1.7.γ
i. Σφάλμα κάλυψης, **ii.** Σφάλμα μέτρησης, **iii.** Σφάλμα μέτρησης, **iv.** Σφάλμα επεξεργασίας, **v.** Σφάλμα μη απόκρισης, **vi.** Σφάλμα μέτρησης (απόκρυψη στοιχείων) ή μη απόκρισης (άρνηση παροχής στοιχείων) **vii.** Σφάλμα επεξεργασίας

1.8.α
i. Λ, **ii.** Λ, **iii.** Λ, **iv.** Σ, **v.** Σ, **vi.** Σ, **vii.** Σ, **viii.** Λ, **ix.** Σ, **x.** Σ,

1.8.β
i. τυχαία, **ii.** $p_i = \frac{n}{N} = \frac{15}{75} = 0{,}2$, **iii.** $w_i = \frac{N}{n} = \frac{75}{15} = 5$, **iv.** 5,

v. $\sum_{i=1}^{15} w_i = \sum_{i=1}^{15} 5 = 15 \cdot 5 = 75 = N$, ... ίσο ... πληθυσμό

vi.

Μαθητής/τρια *i*	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**Σύνολο (Άθροισμα)**
Χρόνος σε λεπτά x_i	20	0	120	30	80	240	0	60	120	180	160	40	80	0	200	1330
Αναγωγικός συντελεστής w_i	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	75
$x_i \cdot w_i$	100	0	600	150	400	1200	0	300	600	900	800	200	400	0	1000	6650

vii. 6.650, **viii.** $\frac{\sum w_i \cdot x_i}{\sum w_i} = \frac{6650}{75} \approx 89$ ... αναγωγικοί, **ix.** $\frac{\sum x_i}{n} = \frac{1330}{15} \approx 89$,

x. πληθυσμό, **xi.** δειγματοληπτικό

1.8.γ

i. συνδυασμοί

ii. $\binom{N\text{-}1}{n\text{-}1} = \frac{(N\text{-}1)!}{(n\text{-}1)!(N\text{-}1\text{-}(n\text{-}1))!} = \frac{(N\text{-}1)!}{(n\text{-}1)!\,(N\text{-}1\text{-}n\text{+}1)!} = \frac{(N\text{-}1)!}{(n\text{-}1)!(N\text{-}n)!}$

iii. $= \frac{\binom{N\text{-}1}{n\text{-}1}}{\binom{N}{n}} = \frac{\frac{(N\text{-}1)!}{(n\text{-}1)!(N\text{-}n)!}}{\frac{N!}{n!(N\text{-}n)!}} = \frac{(N\text{-}1)!n!(N\text{-}n)!}{N!(n\text{-}1)!(N\text{-}n)!} = \frac{(N\text{-}1)!n!}{N!(n\text{-}1)!} = \frac{(N\text{-}1)!(n\text{-}1)!n}{(N\text{-}1)!\,N(n\text{-}1)!} = \frac{n}{N}$

2

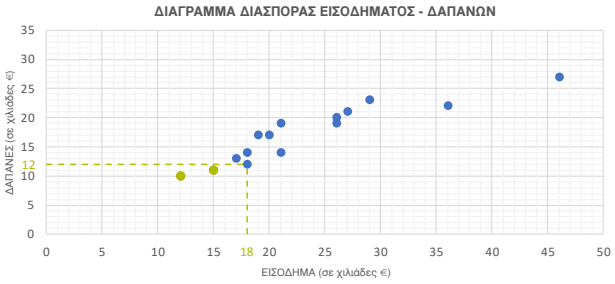
2.1.α

ΠΙΝΑΚΑΣ 1

i. ποσοτική ... διακριτή, **ii.** ποιοτική, **iii.** ποσοτική ... συνεχής, **iv.** ποσοτική ... συνεχής, **v.** ποσοτική ... συνεχής, **vi.** ποσοτική ... διακριτή, **vii.** ποιοτική

2.2.α

ΔΙΑΓΡΑΜΜΑ 1



2.2.β

i. μεγαλύτερες / περισσότερες / υψηλότερες, **ii.** ευθεία

2.4.α

ΠΙΝΑΚΑΣ 3

Εγγραφή 2: 4,7, **Εγγραφή 4:** 4,4, **Εγγραφή 5:** 2,6, **Εγγραφή 7:** 5,8, **Εγγραφή 15:** 53,3

Υπολογισμοί

i. $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{351}{15} = 23{,}4$, **ii.** $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{259}{15} \approx 17{,}3$,

iii. $\sum_{i=1}^{n} (y_i - \bar{y})^2 \approx 337{,}1$

iv. $r = \frac{551{,}1}{\sqrt{1070{,}2} \cdot \sqrt{337{,}1}} \approx \frac{551{,}1}{32{,}7 \cdot 18{,}4} \approx 0{,}916$

2.5.α

i. αύξηση, **ii.** τόσο, **iii.** θετική, **iv.** ισχυρότερη, **v.** συσχέτιση, **vi.** αυξάνεται, **vii.** μειωθούν, **viii.** τόσο, **ix.** αρνητική, **x.** ισχυρότερη, **xi.** μειώνεται, **xii.** αρνητική, **xiii.** γραμμική

2.5.β

i. θετική, **ii.** ισχυρή

2.5.γ

i. Σ, **ii.** Λ, **iii.** Λ, **iv.** Λ, **v.** Σ, **vi.** Σ, **vii.** Λ, **viii.** Σ

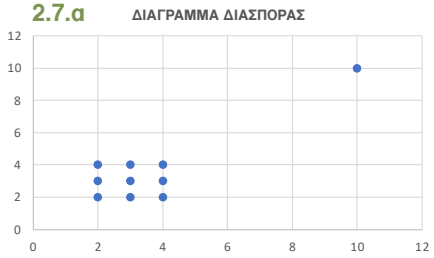
2.5.δ

i. $r_4 = 0,3$, **ii.** $r_3 = 0,1$, **iii.** $r_2 = -0,5$,
iv. $r_1 = -0,9$, **v.** $r_6 = 0,8$, **vi.** $r_5 = 0,5$

2.6.α

i. τρίτος, **ii.** συσχέτιση

2.7.α



2.7.β

ΠΙΝΑΚΑΣ 5

ΕΓΓΡΑΦΕΣ	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})\cdot(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	2	2	-1,7	-1,7	2,89	2,89	2,89
2	2	3	-1,7	-0,7	1,19	2,89	0,49
3	2	4	-1,7	0,3	-0,51	2,89	0,09
4	3	2	-0,7	-1,7	1,19	0,49	2,89
5	3	3	-0,7	-0,7	0,49	0,49	0,49
6	3	4	-0,7	0,3	-0,21	0,49	0,09
7	4	2	0,3	-1,7	-0,51	0,09	2,89
8	4	3	0,3	-0,7	-0,21	0,09	0,49
9	4	4	0,3	0,3	0,09	0,09	0,09
10	10	10	6,3	6,3	39,69	39,69	39,69
ΑΘΡΟΙΣΜΑ	**37**	**37**			**44,1**	**50,1**	**50,1**

Υπολογισμοί

i. $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{37}{10} = 3,7$, **ii.** $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{37}{10} = 3,7$

iii. $\sum_{i=1}^{n} (x_i - \bar{x})^2 = 50,1$

iv. $\sum_{i=1}^{n} (y_i - \bar{y})^2 = 50,1$

v. $r = \frac{44,1}{\sqrt{50,1} \cdot \sqrt{50,1}} = \frac{44,1}{50,1} \approx 0,880$

vi. $r = 0,88$, **vii.** ισχυρή

2.7.γ

ΠΙΝΑΚΑΣ 7

ΕΓΓΡΑΦΕΣ	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})\cdot(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	2	2	-1	-1	1	1	1
2	2	3	-1	0	0	1	0
3	2	4	-1	1	-1	1	1
4	3	2	0	-1	0	0	1
5	3	3	0	0	0	0	0
6	3	4	0	1	0	0	1
7	4	2	1	-1	-1	1	1
8	4	3	1	0	0	1	0
9	4	4	1	1	1	1	1
ΑΘΡΟΙΣΜΑ	**27**	**27**			**0**	**6**	**6**

Υπολογισμοί

i. $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{27}{9} = 3$, **ii.** $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{27}{9} = 3$

iii. $\sum_{i=1}^{n} (x_i - \bar{x})^2 = 6$, **iv.** $\sum_{i=1}^{n} (y_i - \bar{y})^2 = 6$

v. $r = \frac{0}{\sqrt{6} \cdot \sqrt{6}} = \frac{0}{6} = 0$

vi. δεν, **vii.** συσχέτιση

3

3.1.α
i. 0,6, **ii.** 2,8, **iii.** (0, 2,8), **iv.** 0,6

3.1.β
i. $y = 0,6 \cdot 25 + 2,8$, **ii.** $y = 17,8$, **iii.** 17,8 χιλιάδες €

3.2.α
i. ανεξάρτητη, **ii.** εξαρτημένη, **iii.** $y'y$, **iv.** κλίση, **v.** 36, **vi.** 22, **vii.** 24,4

3.2.β
i. $e_3 = 23 - 20,2 = 2,8$, **ii.** $e_2 = 22 - 24,4 = -2,4$

3.2.γ
i. $\hat{a} = \frac{551,5}{1070,2} \approx 0,515$, **ii.** $\hat{b} = 17,3 - 0,515 \cdot 23,4 = 17,3 - 12,051 = 5,249$

iii. $\hat{y} = 0,515 \cdot x + 5,249$

3.2.δ
i. $y = 0,6 \cdot x + 2,8$,
ii. Παρατηρούμε ότι υπάρχει διαφορά ανάμεσα στην εξίσωση της ευθείας που προσαρμόσαμε «με το μάτι» και σε εκείνη που προέκυψε από τη μέθοδο των ελαχίστων τετραγώνων.

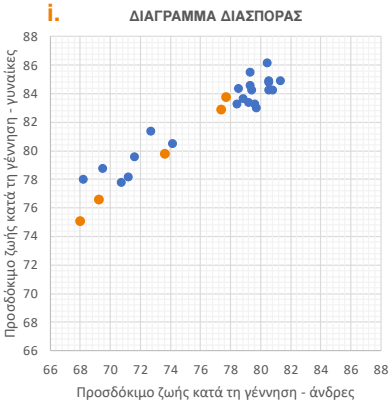
3.2.ε
i. Λ, **ii.** Σ, **iii.** Σ, **iv.** Σ, **v.** Λ, **vi.** Λ, **vii.** Σ

4

4.1.α
i. Β, **ii.** Α, **iii.** Γ, **iv.** Α

4.2.α
i. μικρότερο, **ii.** ποσοτικές ... συνεχείς

4.2.β
ΔΙΑΓΡΑΜΜΑ 1
i.



ii. Από το Διάγραμμα Διασποράς διαφαίνεται ότι υπάρχει ισχυρή θετική συσχέτιση μεταξύ των δύο μεταβλητών.

4.2.γ
i. $r = \frac{343,970}{\sqrt{527,820}\ \sqrt{244,092}} \approx \frac{343,970}{358,923} \approx 0,958$

ii. 0,958, **iii.** ισχυρή

4.3.α
ii. $y = 0,6517x + 32,403$, **iii.** $y = 0,6517x + 32,403$
Οπότε, για $x = 73,6$, εκτιμούμε ότι $y = 0,6517 \cdot 73,6 + 32,403 \approx 80,4$ έτη,

iv. απόλυτη τιμή σφάλματος = $|77,7 - 80,4| = 2,7$ έτη,

v. Εκτίμηση: $y = 0,6517 \cdot 81,7 + 32,403 \approx 85,6$ έτη,
απόλυτη τιμή σφάλματος = $|84,7 - 85,6| = 0,9$ έτη.

ΠΑΡΑΡΤΗΜΑ Ι

Αποσπάσματα Μεταδεδομένων ΕΟΠ 2021

«Ενότητα 3: Στατιστική παρουσίαση»

3. Στατιστική παρουσίαση

3.1 Σύντομη περιγραφή δεδομένων

Η «**Έρευνα Οικογενειακών Προϋπολογισμών**» (Household Budget Survey) είναι μια εθνική έρευνα με την οποία συγκεντρώνονται πληροφορίες από αντιπροσωπευτικό δείγμα νοικοκυριών, για τη σύνθεσή τους, την απασχόληση των μελών τους, τις συνθήκες στέγασης και, κυρίως, για τις δαπάνες διαβίωσής τους, καθώς και για τα εισοδήματά τους.
Οι πληροφορίες που συγκεντρώνονται για τις δαπάνες των νοικοκυριών είναι αναλυτικές. Αυτό σημαίνει ότι δε συγκεντρώνονται απλά πληροφορίες για τις βασικές κατηγορίες δαπανών, όπως για τρόφιμα, ένδυση και υπόδηση, υγεία κ.λπ., αλλά αναλυτικές πληροφορίες για τις επιμέρους κατηγορίες αυτών, π.χ. λευκό ψωμί, φρέσκο γάλα, νωπό μοσχαρίσιο κρέας (τρόφιμα), ανδρική υπόδηση, γυναικεία υπόδηση (ένδυση και υπόδηση), φαρμακευτικά προϊόντα, υπηρεσίες εργαστηρίων ιατρικών αναλύσεων (υγεία), κ.λπ.

Βασικός σκοπός της έρευνας είναι ο προσδιορισμός του καταναλωτικού προτύπου των νοικοκυριών, για την αναθεώρηση του Δείκτη Τιμών Καταναλωτή. Επιπλέον, η έρευνα αυτή είναι η καταλληλότερη πηγή πληροφοριών για:

- τη συμπλήρωση των διαθέσιμων στατιστικών στοιχείων για την εκτίμηση της συνολικής ιδιωτικής κατανάλωσης,
- τη μελέτη του ύψους και της διάρθρωσης των δαπανών των νοικοκυριών σε σχέση με το εισόδημά τους, καθώς και με άλλα οικονομικά, κοινωνικά και δημογραφικά χαρακτηριστικά τους,
- την ανάλυση των μεταβολών του επιπέδου διαβίωσης των νοικοκυριών σε σχέση με τις προηγούμενες έρευνες,
- τη μελέτη της σχέσης μεταξύ των αγορών και των σε είδος απολαβών των νοικοκυριών,
- τη μελέτη των ορίων χαμηλού εισοδήματος για διάφορες κοινωνικοοικονομικές κατηγορίες και ομάδες του πληθυσμού και
- τη μελέτη των αλλαγών στις διατροφικές συνήθειες των νοικοκυριών της Χώρας.

3.2 Χρησιμοποιούμενο σύστημα ταξινόμησης

Χρησιμοποιούνται οι εξής διεθνείς ταξινομήσεις:

- NUTS II, για τις Περιφέρειες,
- ISCED 2011, για το επίπεδο εκπαίδευσης,
- ISCO 08, για την απασχόληση και τα επαγγέλματα,
- NACE Rev.2 (από το 2008), για την οικονομική δραστηριότητα.

3.3 Κάλυψη κλάδων

Οι πληροφορίες για τις δαπάνες που συγκεντρώνονται από τα νοικοκυριά είναι πολύ αναλυτικές. Δε συγκεντρώνονται, δηλαδή, πληροφορίες για κατηγορίες δαπανών συνολικά, όπως «δαπάνες διατροφής», «είδη ένδυσης – υπόδησης», «δαπάνες για υγεία» κ.λπ., αλλά για καθεμία δαπάνη χωριστά, π.χ. ψωμί άσπρο, γάλα νωπό πλήρες, μοσχάρι νωπό, υποδήματα ανδρικά, γυναικεία ή μικροβιολογικές εξετάσεις, φάρμακα.

Η έρευνα είναι δειγματοληπτική, με σχεδιασμό rotational integrated design, που επιλέχθηκε ως ο πλέον κατάλληλος για ενιαία συγχρονική και διαχρονική έρευνα. Τελική δειγματοληπτική μονάδα είναι το νοικοκυριό, ενώ μονάδες ανάλυσης είναι τα νοικοκυριά και τα μέλη τους. Υπολογίζεται ότι θα συμπληρωθούν ερωτηματολόγια για, περίπου, 6.400 νοικοκυριά, αριθμός που αποτελεί το 1,5/1000 των νοικοκυριών, περίπου, της Χώρας.

3.4 Έννοιες και ορισμοί των βασικών μεταβλητών

1.Νοικοκυριό

Ως νοικοκυριό θεωρείται ένα άτομο που ζει μόνο του σε μία κατοικία ή μία ομάδα ατόμων συγγενικών ή μη, τα οποία διαμένουν στην ίδια κατοικία.

2.Μέλη του νοικοκυριού

Τα άτομα που αποτελούν το νοικοκυριό καλούνται μέλη του νοικοκυριού. Αυτά διαμένουν, συνήθως, στο νοικοκυριό ή μπορεί να απουσιάζουν προσωρινά από αυτό.

Άτομα που συνήθως διαμένουν στο νοικοκυριό θεωρούνται τα άτομα που κατά το χρονικό διάστημα των τελευταίων 6 μηνών πέρασαν τις περισσότερες ώρες της ημέρας και της νύκτας στο συγκεκριμένο νοικοκυριό.

Άτομα που προσωρινά απουσιάζουν από το νοικοκυριό, είτε βρίσκονται σε άλλο ιδιωτικό νοικοκυριό είτε σε συλλογική κατοικία (π.χ. νοσοκομείο, γηροκομείο), θα θεωρηθούν και θα καταγραφούν ως μέλη του νοικοκυριού, σύμφωνα με τις προϋποθέσεις που αναφέρονται παρακάτω.

3.5 Στατιστικές μονάδες

Νοικοκυριά και άτομα.

3.6 Πληθυσμός αναφοράς

Η έρευνα καλύπτει τα νοικοκυριά όλης της Χώρας, ανεξάρτητα από το μέγεθος ή τα οποιαδήποτε οικονομικά και κοινωνικά χαρακτηριστικά τους. Από την έρευνα εξαιρούνται:

1. Οι κάθε είδους συλλογικές συμβιώσεις (οικοτροφεία, γηροκομεία, νοσοκομεία, φυλακές, αναμορφωτήρια, στρατόπεδα κλπ.).

2. Τα νοικοκυριά που έχουν περισσότερους από πέντε (5) οικότροφους.

3. Τα νοικοκυριά με μέλη ξένους υπηκόους που υπηρετούν σε ξένες διπλωματικές αποστολές.

3.7 Περιοχή αναφοράς (γεωγραφική κάλυψη)

Ολόκληρη η Χώρα.

3.8 Χρονική κάλυψη

Η πρώτη ΕΟΠ στην Ελλάδα διενεργήθηκε κατά τα έτη 1957/58, είχε διάρκεια ένα χρόνο και το δείγμα ανήλθε σε 2.500, περίπου, νοικοκυριά των αστικών περιοχών της Χώρας.

Τον Απρίλιο του έτους 1963, παράλληλα με την έρευνα στις αστικές περιοχές, άρχισε ευρείας έκτασης έρευνα στις ημιαστικές και στις αγροτικές περιοχές της Χώρας, δηλαδή σε δήμους και κοινότητες με πληθυσμό κάτω των 10.000 κατοίκων, διήρκεσε ένα χρόνο, περιελήφθησαν 3.755 νοικοκυριά των περιοχών αυτών και συνεχίστηκε μέχρι το έτος 1972, σε μικρότερο όμως δείγμα νοικοκυριών.

Κατά τα έτη 1974, 1981/82, 1987/88, 1993/94, 1998/99 και 2004/05 πραγματοποιήθηκαν έρευνες Οικογενειακών Προϋπολογισμών, οι οποίες κάλυψαν όλες τις περιοχές της Χώρας, σε δείγμα, περίπου, 7.500 νοικοκυριών για την πρώτη και σε δείγμα, περίπου, 6.000 έως 6.800 νοικοκυριών για καθεμία από τις πέντε επόμενες, και είχαν διάρκεια ένα έτος.

Από το έτος 2008 αποφασίστηκε, για εθνικές ανάγκες (κατάρτιση Δείκτη Τιμών Καταναλωτή με μεγαλύτερη αξιοπιστία, παραγωγή συγκρίσιμων στατιστικών για τις ανάγκες των Εθνικών Λογαριασμών), η διενέργεια της έρευνας να είναι ετήσια και συνεχής, δηλ. να έχει διάρκεια ένα χρόνο και να πραγματοποιείται κάθε χρόνο και να γίνεται σε δείγμα, περίπου, 4.000 νοικοκυριών. Το 2014 το δείγμα αυξήθηκε, προκειμένου να παραχθούν αποτελέσματα μεγαλύτερης αξιοπιστίας, σε επίπεδο περιφέρειας.

3.9 Περίοδος βάσης

Έτος διενέργειας της έρευνας, δηλαδή 2021.

«Ενότητα 13.3: Μη δειγματοληπτικά σφάλματα»

13.3 Μη δειγματοληπτικά σφάλματα

Τα μη δειγματοληπτικά σφάλματα αναφέρονται σε: σφάλματα πλαισίου, σφάλματα λόγω μη απάντησης, σφάλματα επεξεργασίας και σφάλματα μέτρησης (θα συνταχθούν και αναρτηθούν στην Έκθεση Ποιότητας).

Ενέργειες που έγιναν για τη μείωση του ποσοστού μη ανταπόκρισης των μονάδων της έρευνας είναι:

- Η αποστολή ενημερωτικής επιστολής στα υπό έρευνα νοικοκυριά, περίπου, ένα μήνα πριν τη διενέργεια της έρευνας.
- Οι τρεις τουλάχιστον προσπάθειες για τηλεφωνική επικοινωνία με τα νοικοκυριά τα οποία δεν ανταποκρίνονταν στις κλήσεις, διαφορετικές ημέρες και ώρες της ημέρας.

α. Μη απάντηση σε επίπεδο μονάδας

β. Μη απάντηση σε επίπεδο ερώτησης

13.3.1 Σφάλμα κάλυψης

13.3.1.1 Α2. Ποσοστό υπερκάλυψης

13.3.1.2 Α3. Κοινές μονάδες (ποσοστό)

13.3.2 Σφάλμα μέτρησης

Αντιμετωπίζονται με την ύπαρξη οδηγιών για την ορθή συμπλήρωση του ερωτηματολογίου, με την πραγματοποίηση εκπαιδευτικών συγκεντρώσεων, αλλά και με την πραγματοποίηση ελέγχων (λογικών, ροής, τιμών), που υλοποιούνται από την ΕΛΣΤΑΤ.

13.3.3 Σφάλμα επεξεργασίας

Ποιοτικοί και ποσοτικοί έλεγχοι της βάσης δεδομένων πραγματοποιούνται για τη διόρθωση των λαθών καταχώρισης.

13.3.4 Σφάλμα από την εφαρμογή μοντέλου

«Ενότητα 18.1: Τύπος πρωτογενών δεδομένων»

18.1 Τύπος πρωτογενών δεδομένων

Η έρευνα είναι δειγματοληπτική, με σχεδιασμό rotational integrated design, που επιλέχθηκε ως ο πλέον κατάλληλος για ενιαία συγχρονική και διαχρονική έρευνα. Τελική δειγματοληπτική μονάδα είναι το νοικοκυριό, ενώ μονάδες ανάλυσης είναι τα νοικοκυριά και τα μέλη τους.

Η ΕΟΠ 2021 βασίζεται σε δισταδιακή στρωματοποιημένη δειγματοληψία νοικοκυριών, από πλαίσιο δειγματοληψίας, που έχει δημιουργηθεί με βάση τα στοιχεία για τον μόνιμο πληθυσμό της Απογραφής 2011, και καλύπτει πλήρως τον πληθυσμό αναφοράς, ώστε να εξασφαλίζεται η αντιπροσωπευτικότητα του δείγματος.

Ο σχεδιασμός της δειγματοληψίας περιλαμβάνει δύο κριτήρια στρωμάτωσης:

- Το πρώτο κριτήριο στρωμάτωσης του ερευνώμενου πληθυσμού είναι η γεωγραφική διαίρεση της Χώρας. Ως μείζονα στρώματα χρησιμοποιήθηκαν οι 13 Περιφέρειες (NUTS II), η πρώην Περιφέρεια Πρωτευούσης και το Πολεοδομικό Συγκρότημα Θεσσαλονίκης.
- Το δεύτερο κριτήριο στρωμάτωσης είναι ο βαθμός αστικότητας. Σε κάθε Περιφέρεια (NUTS II), τα νοικοκυριά κατανεμήθηκαν ανάλογα με τον βαθμό αστικότητας (τοπικά διαμερίσματα με πληθυσμό πάνω από 10.000 κατοίκους, από 2.000 έως 9.999 κατοίκους και μέχρι 1.999 κατοίκους). Η στρωμάτωση κατά βαθμό αστικότητας έγινε σύμφωνα με τον πληθυσμό των τοπικών διαμερισμάτων, εκτός των πολεοδομικών συγκροτημάτων των δύο μεγάλων πόλεων (πρώην Περιφέρεια Πρωτευούσης και Πολεοδομικό Συγκρότημα Θεσσαλονίκης),

Η πρώην Περιφέρεια Πρωτευούσης χωρίστηκε σε 31 στρώματα ίσου, περίπου, μεγέθους (ίσους αριθμούς νοικοκυριών), με βάση τους καταλόγους-πλαίσια με τα οικοδομικά τετράγωνα των Δήμων και σύμφωνα με κοινωνικοοικονομικά κριτήρια. Όμοια, το πρώην Πολεοδομικό Συγκρότημα Θεσσαλονίκης χωρίστηκε σε 9 ίσου μεγέθους στρώματα. Τα πρώην Πολεοδομικά Συγκροτήματα αυτών των δύο μεγάλων πόλεων αποτελούν, περίπου, το 40% του συνολικού πληθυσμού και κατέχουν ακόμα υψηλότερο ποσοστό σε κάποιες κοινωνικοοικονομικές μεταβλητές.

Ο αριθμός των στρωμάτων, που προέκυψε από την εφαρμογή των δύο κριτηρίων στρωμάτωσης του πληθυσμού, ανέρχεται σε 79.

Το μέγεθος του δείγματος νοικοκυριών ανήλθε περίπου στα 6.400 (κλάσμα δειγματοληψίας 1,5‰), το οποίο ισοκατανεμήθηκε μέσα στο έτος, ώστε να επιλεγούν 4 ισοδύναμα ανεξάρτητα δείγματα, που αντιστοιχούν στα 4 τρίμηνα του έτους.

1ο στάδιο δειγματοληψίας

Στο πρώτο στάδιο δειγματοληψίας, από κάθε στρώμα, επιλέγονται οι πρωτογενείς μονάδες (επιφάνειες) με πιθανότητα επιλογής ανάλογη του μεγέθους τους (πλήθος νοικοκυριών, σύμφωνα με την Απογραφή Πληθυσμού έτους 2011).

2ο στάδιο δειγματοληψίας

Στο δεύτερο στάδιο δειγματοληψίας, σε κάθε επιλεγείσα πρωτογενή δειγματοληπτική μονάδα (μονάδα επιφανείας), από τον ενημερωμένο κατάλογο-πλαίσιο των νοικοκυριών, επιλέγεται το δείγμα νοικοκυριών με ίσες πιθανότητες και με την εφαρμογή της συστηματικής δειγματοληψίας. Στην πραγματικότητα, στο δεύτερο στάδιο επιλέγεται δείγμα κατοικιών. Εντούτοις, στις περισσότερες περιπτώσεις, υπάρχει μία αντιστοίχιση μεταξύ νοικοκυριού και κατοικίας. Εάν η επιλεγείσα κατοικία αποτελείται από περισσότερα του ενός νοικοκυριά, τότε ερευνώνται όλα τα νοικοκυριά.

Ο συνολικός αριθμός μονάδων επιφανείας που επιλέχθηκαν ήταν 1.068.

ΠΑΡΑΡΤΗΜΑ ΙΙ

Εφαρμογή με τη βοήθεια του Excel

Στα στοιχεία του Πίνακα 1 από την ΕΟΠ 2019, ας εφαρμόσουμε τη **«μέθοδο των ελαχίστων τετραγώνων»**, για να εκτιμήσουμε τους συντελεστές της εξίσωσης ευθείας που αντιπροσωπεύει καλύτερα το διάγραμμα διασποράς των δεδομένων μας.

Σημειώνεται ότι δεν έχει πραγματοποιηθεί στρογγυλοποίηση στα δεδομένα του ακόλουθου πίνακα, διότι δεν υπάρχει η ανάγκη απλοποίησης των αριθμών προκειμένου να διευκολυνθούν οι πράξεις, εφόσον οι υπολογισμοί θα πραγματοποιηθούν στο Excel.

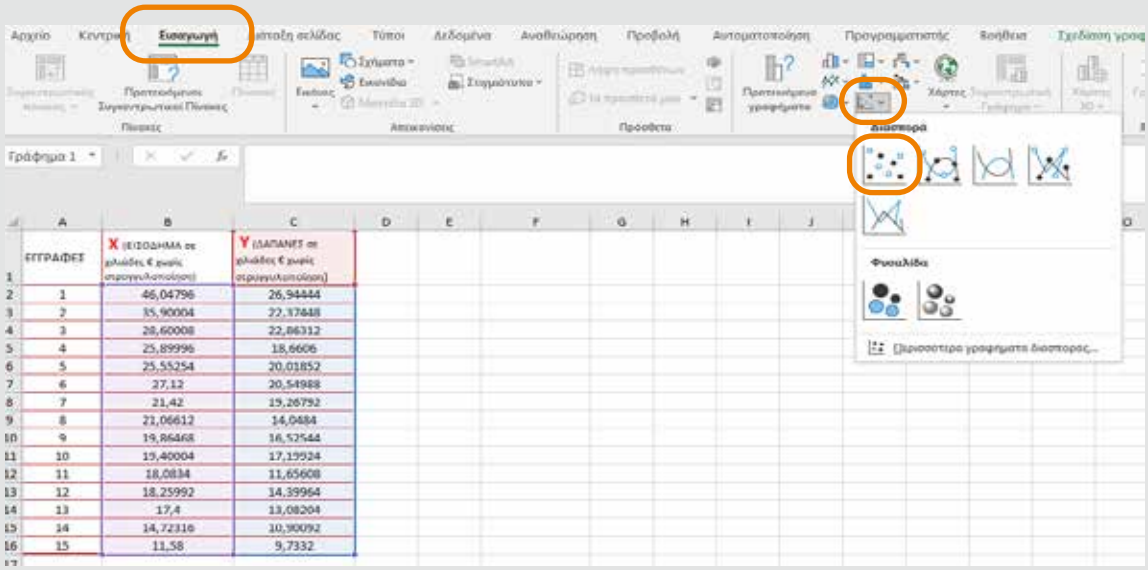
Ακολουθεί αναλυτικά και ενδεικτικά ο εναλλακτικός τρόπος με τον οποίο μπορείτε να εργαστείτε στο Excel, εφαρμόζοντας τη **«μέθοδο των ελαχίστων τετραγώνων»**

A	B	C
ΕΓΓΡΑΦΕΣ	X (ΕΙΣΟΔΗΜΑ σε χιλιάδες € χωρίς στρογγυλοποίηση)	Y (ΔΑΠΑΝΕΣ σε χιλιάδες € χωρίς στρογγυλοποίηση)
1	46,04796	26,94444
2	35,90004	22,37448
3	28,60008	22,86312
4	25,89996	18,6606
5	25,55254	20,01852
6	27,12	20,54988
7	21,42	19,26792
8	21,06612	14,0484
9	19,86468	16,52544
10	19,40004	17,19924
11	18,0834	11,65608
12	18,25992	14,39964
13	17,4	13,08204
14	14,72316	10,90092
15	11,58	9,7332

ΔΟΥΛΕΥΟΝΤΑΣ ΣΤΟ EXCEL

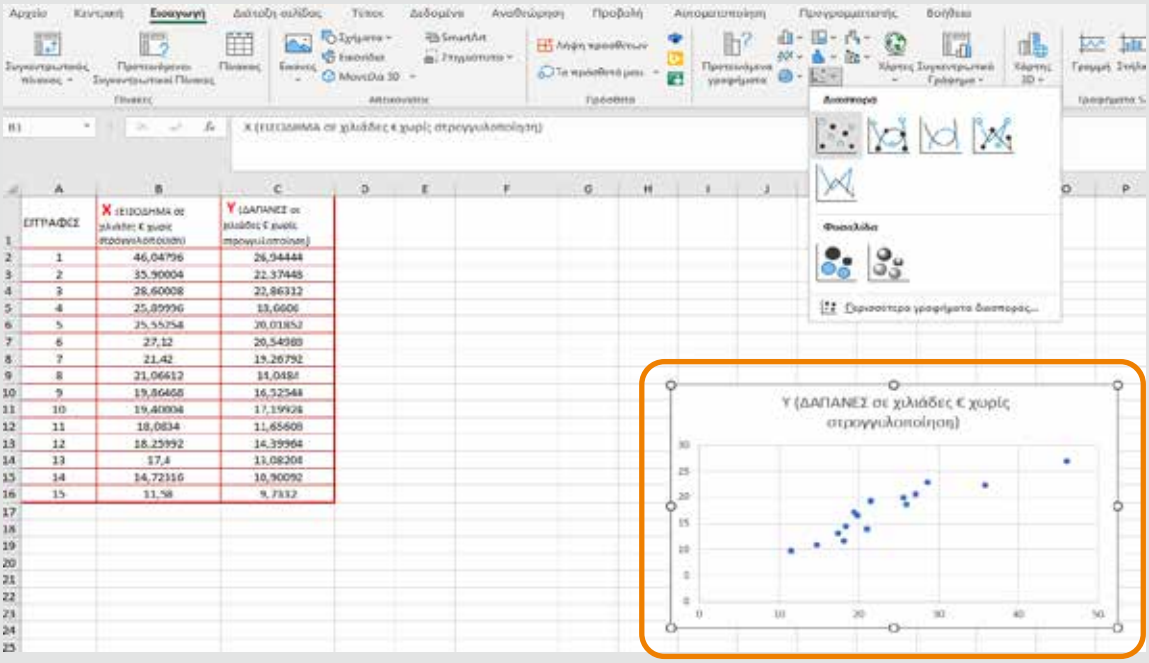
Διαδικασία δημιουργίας ενός Διαγράμματος Διασποράς (scatterplot)

1. Ανοίγετε ένα νέο φύλλο Excel.
2. Εισάγετε τα δεδομένα, τα οποία εμφανίζονται στις στήλες Β (**ΕΙΣΟΔΗΜΑ**) και C (**ΔΑΠΑΝΕΣ**) και δημιουργείτε δύο λίστες.
3. Επιλέγετε τα δεδομένα από τις στήλες (εδώ Β και C) για τις δύο μεταβλητές, με τα οποία θέλετε να δημιουργήσετε το διάγραμμα διασποράς.
4. Επιλέγετε «Εισαγωγή» και, από το πεδίο «Γραφήματα», επιλέγετε το διάγραμμα διασποράς, όπως φαίνεται στην εικόνα.

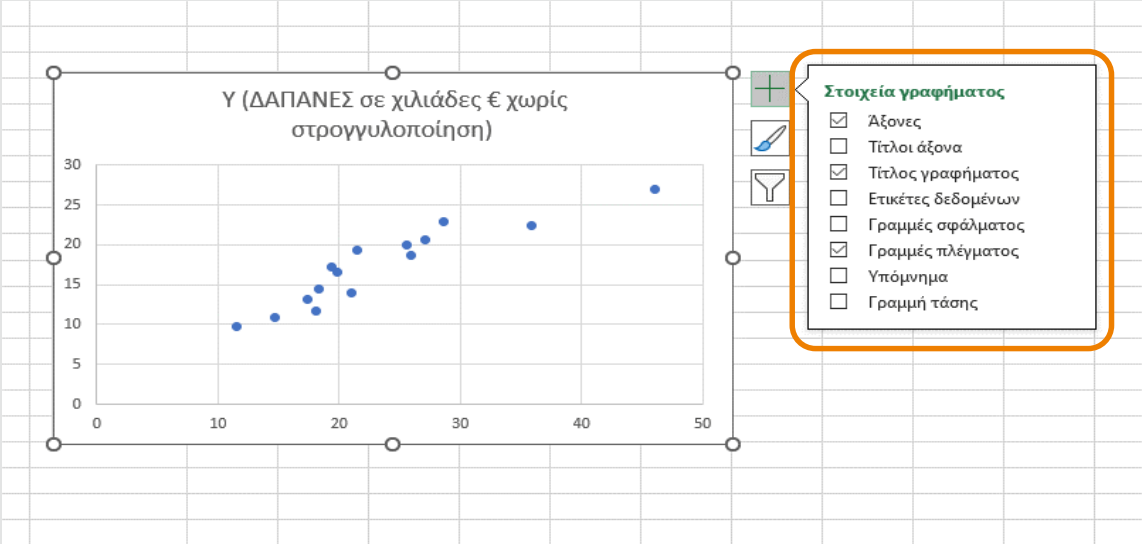


	A	B	C
1	ΕΓΓΡΑΦΕΣ	X (ΕΙΣΟΔΗΜΑ σε χιλιάδες € χωρίς στρογγυλοποίηση)	Y (ΔΑΠΑΝΕΣ σε χιλιάδες € χωρίς στρογγυλοποίηση)
2	1	46,04796	26,94444
3	2	35,90004	22,37448
4	3	28,60008	22,86312
5	4	25,89996	18,6606
6	5	25,55254	20,01852
7	6	27,12	20,54988
8	7	21,42	19,26792
9	8	21,06612	14,0484
10	9	19,86468	16,52544
11	10	19,40004	17,19924
12	11	18,0834	11,65608
13	12	18,25992	14,39964
14	13	17,4	13,08204
15	14	14,72316	10,90092
16	15	11,58	9,7332

Εμφανίζεται το διάγραμμα διασποράς.

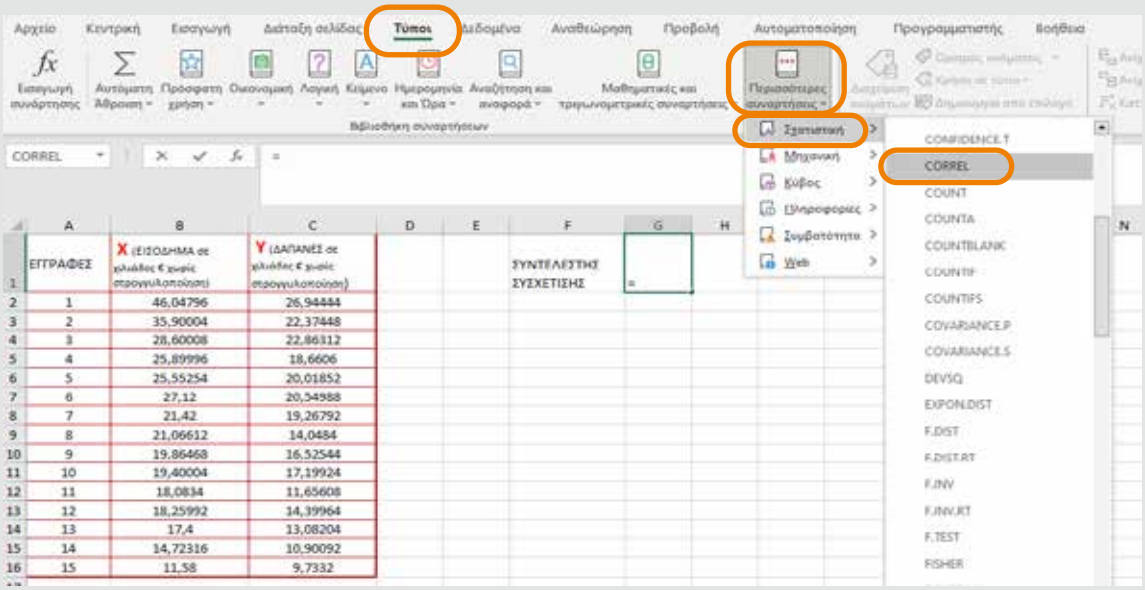


5. Αφού δημιουργηθεί το διάγραμμα διασποράς, μπορείτε να χρησιμοποιήσετε τις επιλογές από τα «Στοιχεία γραφήματος» για να προσαρμόσετε τη σχεδίαση, το εύρος αξόνων, τις ετικέτες και άλλα.

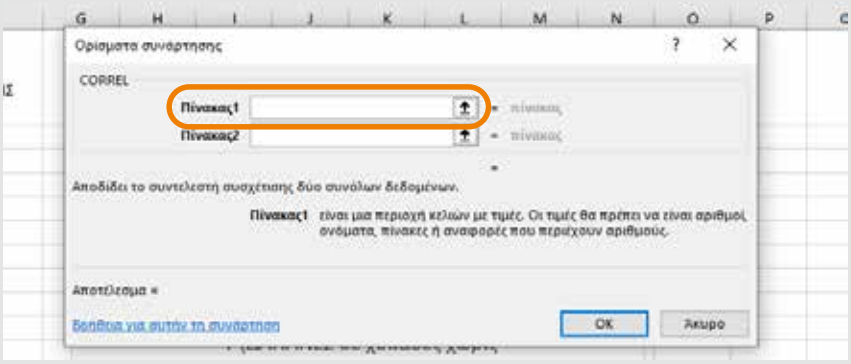


6. Υπολογίζετε τον Συντελεστή Συσχέτισης για τις δύο στήλες δεδομένων μεταξύ αυτών των δύο μεταβλητών. Επιλέγετε ένα κενό κελί που επιθυμείτε να τοποθετήσετε το αποτέλεσμα υπολογισμού, πληκτρολογείτε τον τύπο **= CORREL (B2: B16, C2: C16)** και, στη συνέχεια, πατάτε «enter».

   Εναλλακτικά, για να υπολογίσετε τον Συντελεστή Συσχέτισης, επιλέγετε από το κεντρικό μενού «Τύποι» και μετά «Περισσότερες Συναρτήσεις». Από τη λίστα που ανοίγει επιλέγετε «Στατιστική» και από την επόμενη λίστα που ανοίγει επιλέγετε τη συνάρτηση «CORREL».



Στη συνέχεια, ανοίγει το ακόλουθο παράθυρο:



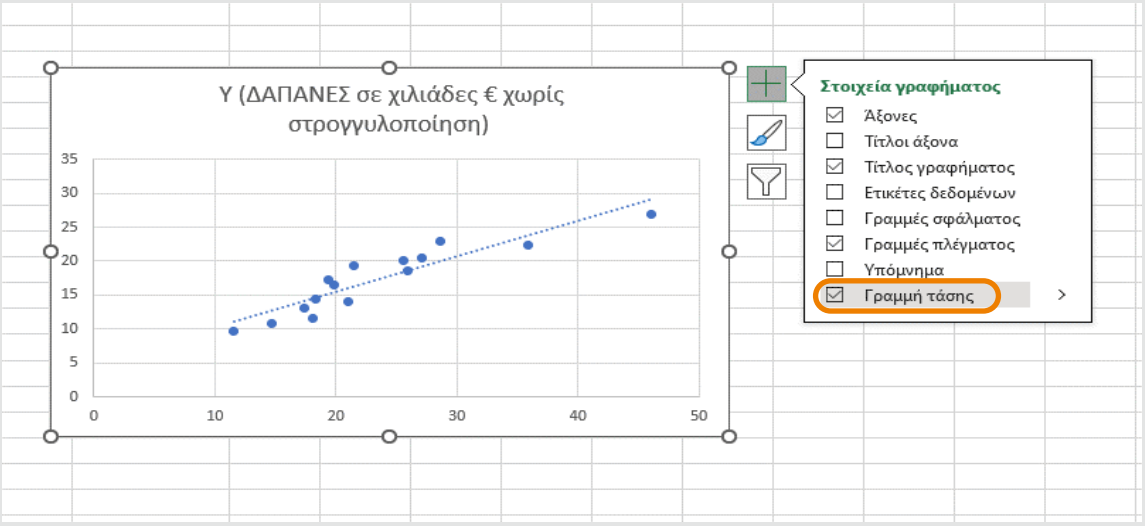
	A	B	C
1	ΕΓΓΡΑΦΕΣ	X (ΕΙΣΟΔΗΜΑ σε χιλιάδες € χωρίς στρογγυλοποίηση)	Y (ΔΑΠΑΝΕΣ σε χιλιάδες € χωρίς στρογγυλοποίηση)
2	1	46,04796	26,94444
3	2	35,90004	22,37448
4	3	28,60008	22,86312
5	4	25,89996	18,6606
6	5	25,55254	20,01852
7	6	27,12	20,54988
8	7	21,42	19,26792
9	8	21,06612	14,0484
10	9	19,86468	16,52544
11	10	19,40004	17,19924
12	11	18,0834	11,65608
13	12	18,25992	14,39964
14	13	17,4	13,08204
15	14	14,72316	10,90092
16	15	11,58	9,7332

Στο πεδίο «Πίνακας1», τοποθετείτε τον κέρσορα και επισημαίνετε τα αντίστοιχα δεδομένα (σε αυτό το παράδειγμα τα δεδομένα για τη μεταβλητή X), όπως φαίνεται στη διπλανή εικόνα.

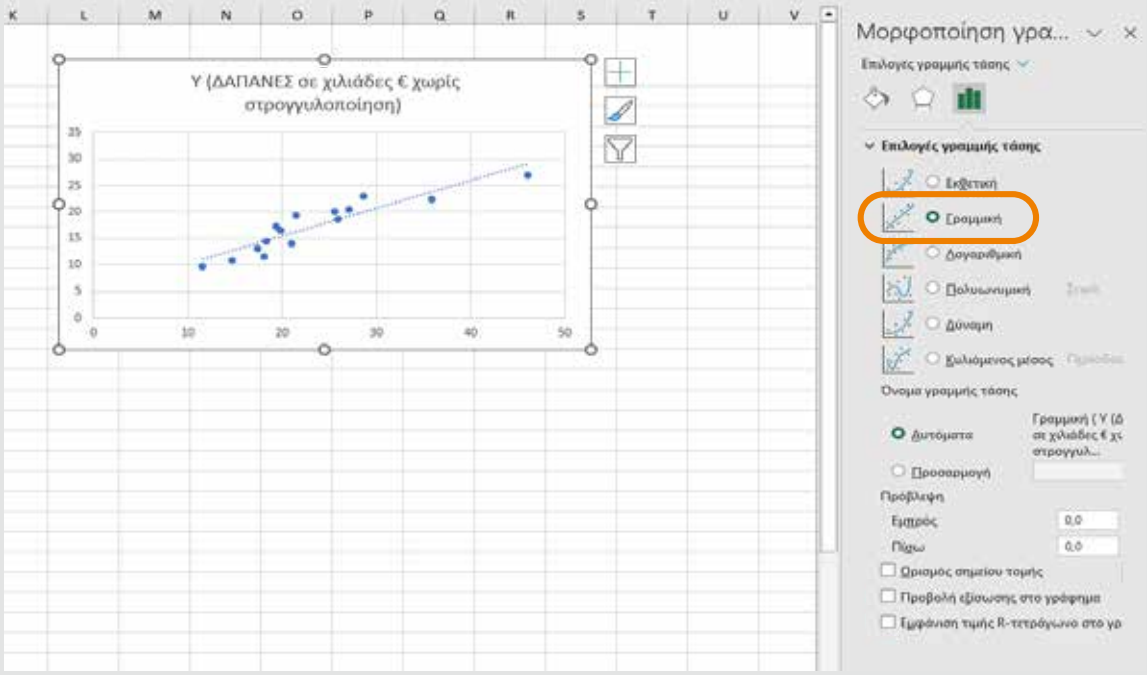
Συνεχίζετε, αντίστοιχα, για το επόμενο πεδίο «Πίνακας 2» (που θα περιλαμβάνει τα δεδομένα της μεταβλητής Y).

Βρίσκετε την τιμή του Συντελεστή Συσχέτισης $r = 0{,}920124$.

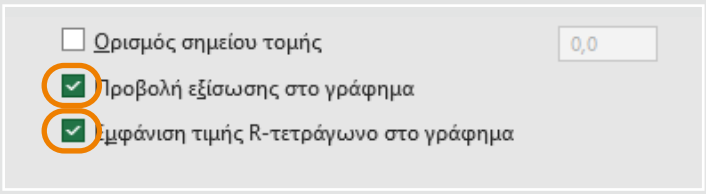
7. Πηγαίνετε στο διάγραμμα, το ενεργοποιείτε και εμφανίζεται το μενού «Στοιχεία γραφήματος» (βήμα 5). Στη συνέχεια, επιλέγετε «Γραμμή τάσης», όπως φαίνεται στο παρακάτω βήμα, και προκύπτει η ευθεία με την καλύτερη προσαρμογή (στα σημεία του γραφήματος).

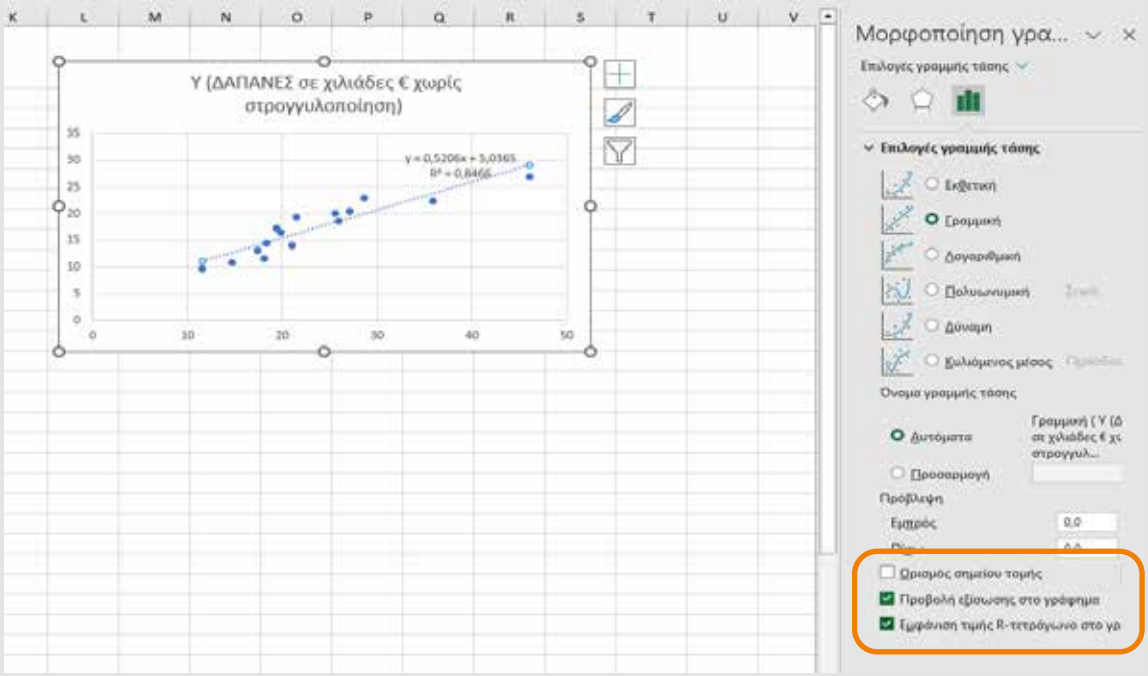


8. Με διπλό κλικ πάνω στη «Γραμμής τάσης», ανοίγει το παράθυρο «Μορφοποίηση γραμμής τάσης». Επιλέγετε «Γραμμική», στο πεδίο «Επιλογές γραμμής τάσης».



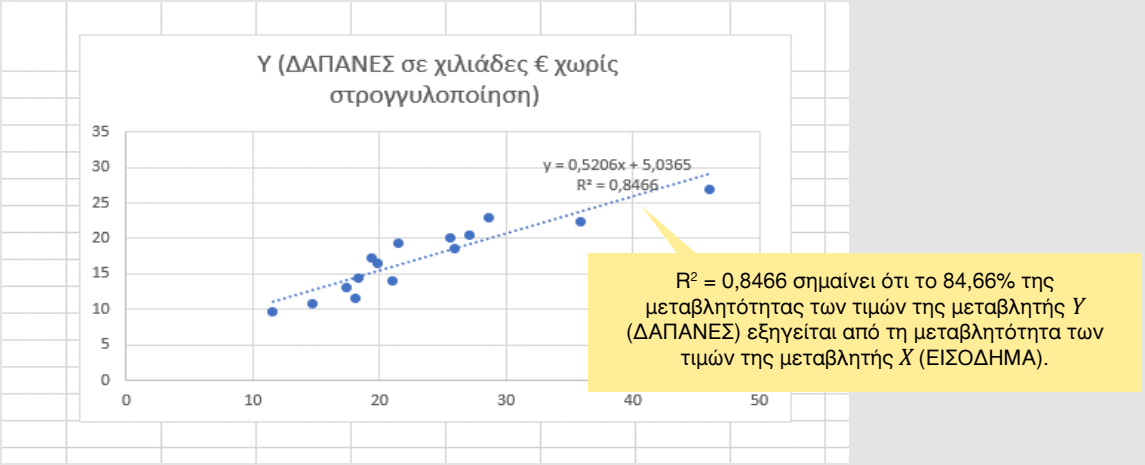
Στη συνέχεια, σημειώνετε τις επιλογές «Προβολή εξίσωσης στο γράφημα» και «Εμφάνιση τιμής R-τετράγωνο στο γράφημα».





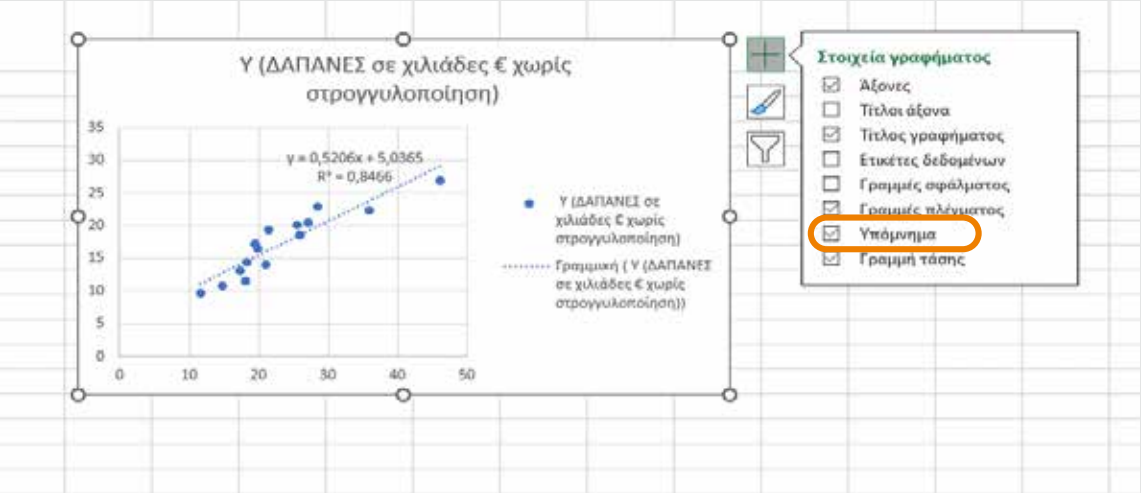
***Εναλλακτικά**, (στο πεδίο που ανοίγει δίπλα στο διάγραμμα διασποράς) μετά τη «Γραμμή τάσης», επιλέγετε «Περισσότερες επιλογές».*

Οπότε, εμφανίζεται στο διάγραμμα η εξίσωση της ευθείας με την καλύτερη προσαρμογή (στα σημεία του γραφήματος), καθώς και η τιμή του R^2.

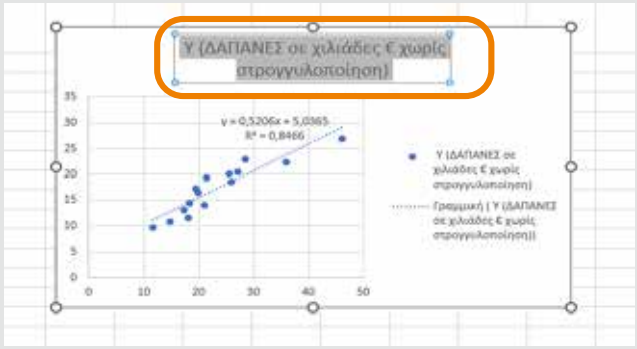


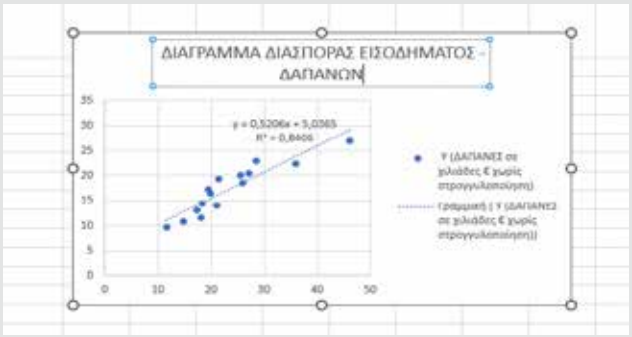
Παρατηρείτε ότι η εξίσωση της ευθείας παλινδρόμησης, στο Excel, είναι ***y* = 0,5206 • *x* + 5,0356**, ενώ, στην Ενότητα 3.2, είναι $\hat{y}$ = 0,505 • x + 5,483. Η μικρή αυτή διαφορά, που παρατηρείται, οφείλεται στις στρογγυλοποιήσεις που πραγματοποιήθηκαν, για τη διευκόλυνση των πράξεων στους υπολογισμούς «στο χαρτί».

9. Αν θέλετε στο διάγραμμά σας να υπάρχει και επεξηγηματικό υπόμνημα, στο μενού «Στοιχεία γραφήματος», επιλέγετε «Υπόμνημα».



Στη συνέχεια, μπορείτε να επεξεργαστείτε τον τίτλο του διαγράμματος, επιλέγοντας το κείμενο του τίτλου για να το ενεργοποιήσετε και, ακολούθως, να το αλλάξετε, όπως επιθυμείτε (στο παράδειγμά μας ο τίτλος έγινε «ΔΙΑΓΡΑΜΜΑ ΔΙΑΣΠΟΡΑΣ ΕΙΣΟΔΗΜΑΤΟΣ - ΔΑΠΑΝΩΝ»).

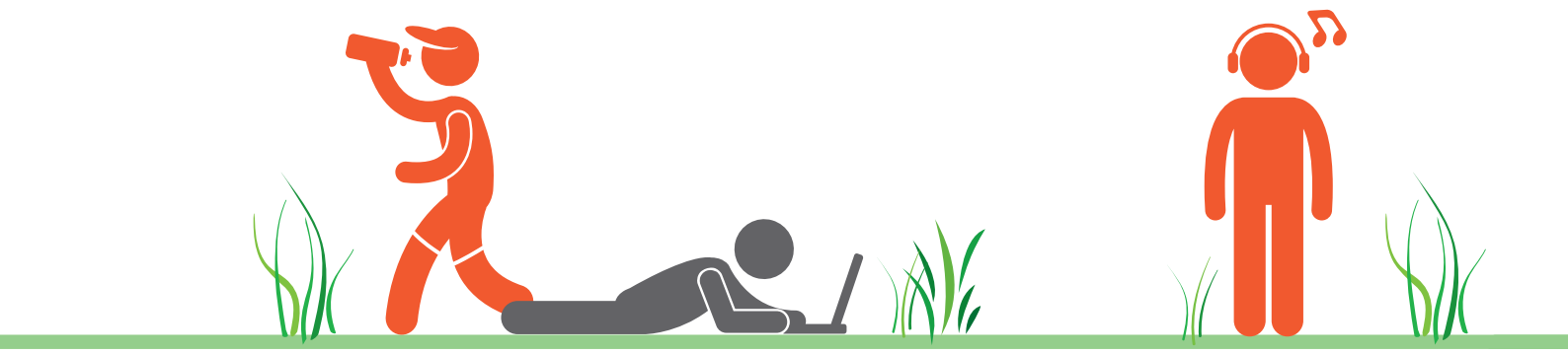




Σημειώνεται ότι το R^2*, που συναντήσατε, ονομάζεται* ***Συντελεστής Προσδιορισμού*** *και παίρνει τιμές στο διάστημα [0, 1]. Εκφράζει το ποσοστό της διασποράς της μεταβλητής Υ που εξηγείται με βάση το μοντέλο της γραμμικής παλινδρόμησης.*

Αποδεικνύεται ότι, στην περίπτωση που έχουμε μία ανεξάρτητη και μία εξαρτημένη μεταβλητή στο μοντέλο μας, το R^2 *ισούται με το τετράγωνο του Συντελεστή Συσχέτισης r.*

Όσο μεγαλύτερη είναι η τιμή του Συντελεστή Προσδιορισμού τόσο ισχυρότερη είναι η γραμμική σχέση εξάρτησης των δύο μεταβλητών του μοντέλου μας.



9 786185 884086

www.statistics.gr