

# Μάθημα 2<sup>ο</sup>

## Σύγκριση Μέσων Όρων

### Μέρος Α



# Στόχοι Ενότητας



- ✓ Μεταβλητές
- ✓ Συντελεστές Συσχέτισης
  - ✓ Pearson (Παραμετρικό τεστ)
  - ✓ Spearman – Kendal (Μη Παραμετρικό Τεστ)
- ✓ Έλεγχος για την μέση τιμή ενός πληθυσμού
  - ✓ One sample T test (Παραμετρικό τεστ)
  - ✓ Wilcoxon (Μη Παραμετρικό Τεστ)





## Ανεξάρτητες Μεταβλητές

- Είναι **σταθερές** που ο ερευνητής χειρίζεται ανεξάρτητα από τι τις επηρεάζει, ή προσδιορίζει. Αυτές καθορίζονται εκ των προτέρων.
- Για παράδειγμα εξετάζουμε μια στατιστική μονάδα (π.χ. τον φοιτητή) ανεξάρτητα των μεταβλητών ύψους, βάρους, οικογενειακής κατάστασης, κλπ.
- Η **ανεξάρτητη** μεταβλητή μπορεί να είναι ποσοτική ή ποιοτική
- Στα μαθηματικά αποδίδεται ως συνάρτηση της ανεξάρτητης ( $x$ ) από την εξαρτημένη ( $y$ ) μεταβλητή, όπου  $y = f(x)$ .



## Εξαρτημένη Μεταβλητή

- Είναι η μεταβλητή που **ερευνάται** στη μελέτη. Σε ορισμένες περιπτώσεις είναι το αποτέλεσμα μιας πειραματικής διαδικασίας.
- Η μεταβλητότητα στην εξαρτημένη μεταβλητή λογικά εξαρτάται από τις συνθήκες που χειρίζεται ο ερευνητής στη μελέτη. Στις πιο πολλές μελέτες, οι εξαρτημένες μεταβλητές είναι εκείνες που ο ερευνητής στοχεύει να κατανοήσει, να εξηγήσει ή να προβλέψει. Αποτελούν εκείνο που ο ερευνητής μετρά στα άτομα μετά την έκθεση τους στην ανεξάρτητη μεταβλητή.



## Ποιοτικές

θεωρούνται οι μεταβλητές που **δεν** μπορούν να εκφραστούν αριθμητικά αλλά διακρίνονται σε συγκεκριμένες κατηγορίες ή ομάδες που ονομάζονται *διαβαθμίσεις ή κλάσεις ή ιδιότητες*.

- στις ποιοτικές μεταβλητές δεν υπάρχουν κλίμακες μέτρησης.
- Οι τιμές των ποιοτικών μεταβλητών δεν δίδονται με αριθμούς αλλά με διακριτικό είδος, για παράδειγμα το "φύλο" παίρνει τιμές άρρεν θήλυ, η ομάδα αίματος παίρνει τιμές A, B, AB και O, κ.λπ..



## Ποιοτικές

Οι ποιοτικές μεταβλητές **διακρίνονται** σε

- **Διατάξιμες** μεταβλητές οι οποίες λαμβάνουν ως τιμές - έννοιες που ιεραρχούνται (π.χ. διαγωγή μαθητή σε καλή, κοσμία, κοσμιωτάτη, ή η βαθμίδα αξιωματικού: κατώτερος, ανώτερος, ανώτατος, ή δικαστικού: πρωτοδίκης, εφέτης, αρεοπαγίτης, ή η κατάσταση υγείας: καλή, πολύ καλή άριστη κ.λπ. και σε
- **Μη- διατάξιμες** ή κατηγορικές μεταβλητές, οι οποίες λαμβάνουν τιμές - έννοιες που δεν ιεραρχούνται όπως το επάγγελμα, το φύλο, η υπηκοότητα, το θρήσκευμα κ.λπ.



## Ποσοτικές

θεωρούνται οι μεταβλητές στις οποίες μπορούμε να αντιστοιχίσουμε έναν αριθμό μιας συγκεκριμένης κλίμακας. Τα στατιστικά δεδομένα αποτελούν στην περίπτωση αυτή μια σειρά αριθμητικών μετρήσεων. Π.χ. Η μεταβλητή **Ηλικία** με τιμές 18, 35, 67...



## Ποσοτικές

Οι ποσοτικές μεταβλητές διακρίνονται

- Οι **διακριτές** παίρνουν μόνο "μεμονωμένες" αριθμητικές τιμές, είναι δηλαδή στοιχεία ενός συνόλου τα οποία μπορούν να αντιστοιχηθούν ένα προς ένα με στοιχεία του συνόλου των θετικών ακέραιων αριθμών. Τέτοια δεδομένα είναι π.χ. ο αριθμός των παιδιών σε μία οικογένεια, ο αριθμός των δωματίων μιας κατοικίας κλπ.
- Οι **συνεχείς** μπορούν να πάρουν αριθμητικές τιμές που καλύπτουν ολόκληρο διάστημα τιμών των πραγματικών αριθμών Π.χ. η ηλικία, η διάρκεια μιας τηλεφωνικής συνδιάλεξης, η θερμοκρασία κλπ.





# Σχέση Μεταβλητών

Έστω ότι έχουμε δύο ερωτήσεις **Q1, Q2**

- Θέλουμε να δούμε αν αυτές οι ερωτήσεις έχουν κάποια σχέση μεταξύ τους
- Σχέση δεν σημαίνει ομοιότητα, αφού οι ερωτήσεις είναι διαφορετικές !
- Σχέση σημαίνει πως οι απαντήσεις στην **Q1**, επηρεάζουν ή καθορίζουν σε κάποιο βαθμό και τις απαντήσεις στην **Q2**



# Σχέση Μεταβλητών

Ανάλογα με το είδος των μεταβλητών έχουμε τις παρακάτω περιπτώσεις

- Ποσοτική με Ποσοτική
- Ποιοτική με Ποσοτική
- Ποιοτική με Ποιοτική



## Σχέση μεταξύ Μεταβλητών

Όταν γίνεται διερεύνηση μιας σχέσης μεταξύ δύο μεταβλητών αρχικά θέλουμε να δώσουμε απαντήσεις στις παρακάτω

### Ερωτήσεις

- Υπάρχει σχέση;
- Τι σχέση είναι αυτή; (π.χ. όταν αυξάνει το ένα, αυξάνει και το άλλο ή αντίστροφα;)
- Ποιος είναι ο βαθμός αυτής της σχέσης;
- Ποια είναι η μορφή αυτής της σχέσης;



## Συσχέτιση

- ✓ Πολλές φορές θέλουμε να **συγκρίνουμε** δυο μεταβλητές, (π.χ. *Υπάρχει σχέση μεταξύ των ποτών (μέσος όρος) που πίνει κάποιος φοιτητής ή φοιτήτρια την εβδομάδα με το πόσο συχνά χάνει το μάθημα της Ανάλυσης Δεδομένων εξαιτίας του πονοκεφάλου που έχει την επόμενη μέρα;*)
- ✓ Η ύπαρξη **συσχέτισης** μεταξύ δυο μεταβλητών **δεν** ορίζει και την ύπαρξη αιτιακής σχέσης (π.χ *Αν η τιμή μιας μεταβλητής αυξάνεται όσο και η τιμή της άλλης, αυτό δεν σημαίνει ότι η μια μεταβλητή εξηγεί την άλλη ή μπορεί η μια να προβλέψει την άλλη...* )



## Συσχέτιση

Με τον όρο **συσχέτιση** (**correlation**) εννοούμε το βαθμό στον οποίο **συμμεταβάλλονται** δύο ποσοτικές μεταβλητές υπό την προϋπόθεση ότι η σχέση τους είναι γραμμική (Δηλαδή αν κάθε μεταβολή των τιμών της μιας μεταβλητής συνοδεύεται από συστηματική μεταβολή στις τιμές της άλλης).

- ✓ **Αιτιώδης** σχέση : Χρονικά η αιτία προηγείται του αποτελέσματος. (πχ. Αίτιο κάπνισμα, αποτέλεσμα καρκίνος. Με τη συσχέτιση δεν προϋποθέτουμε καμιά χρονική σειρά).
- ✓ **Αιτιώδης** σχέση : όταν απουσιάζει το αίτιο δεν εμφανίζεται το αποτέλεσμα. Άνθρωποι που καπνίζουν, δεν εμφανίζουν καρκίνο, και το αντίθετο.



# Συσχέτιση

Η **Απλή Συσχέτιση** εκφράζεται με έναν αριθμό, που ονομάζεται **Συντελεστής Συσχέτισης** (Correlation Coefficient). Οι πιο γνωστοί συντελεστές γραμμικής συσχέτισης είναι οι συντελεστές του **Pearson**, του **Spearman** και του **Kendall**.

Η μηδενική και η εναλλακτική υπόθεση εδώ είναι οι εξής:

- **$H_0$** :  $\rho=0$  ή δεν υπάρχει γραμμική συσχέτιση μεταξύ των δύο μεταβλητών
- **$H_1$** :  $\rho \neq 0$  ή υπάρχει γραμμική συσχέτιση μεταξύ των δύο μεταβλητών



# Συσχέτιση

## Παρατηρήσεις

- Ο συντελεστής συσχέτισης του **Pearson** “**χρειάζεται**” την υπόθεση της **κανονικότητας** των δεδομένων, σε αντίθεση με τους άλλους δύο που δεν “**χρειάζονται**” την υπόθεση της κανονικότητας των δεδομένων.
- Βέβαια, για μεγάλα δείγματα, μεγέθους 30 παρατηρήσεων και πάνω και όσο το μέγεθος του δείγματος μεγαλώνει η θεωρία μας λέει ότι οι τιμές των συντελεστών “**πλησιάζουν**” η μία την άλλη.
- ο συντελεστής του **Kendall** μπορεί να χρησιμοποιηθεί και στην περίπτωση που έχουμε κατηγορικές μεταβλητές οι οποίες όμως είναι **υποχρεωτικά** σε κλίμακα διάταξης. Είναι δηλαδή διατακτικές κατηγορικές μεταβλητές.



# Συσχέτιση

## Παρατηρήσεις

- Η κύρια **διαφορά** των συντελεστών είναι ότι ο συντελεστής του **Pearson** υπολογίζεται με βάση τα δεδομένα, ενώ οι άλλοι δύο υπολογίζονται με βάση τις τάξεις μεγέθους των δεδομένων. Ειδικότερα, ο συντελεστής του **Spearman** είναι ο συντελεστής του **Pearson** στην ουσία υπολογισμένος για τις τάξεις μεγέθους των δεδομένων.
- Το γεγονός λοιπόν ότι οι συντελεστές του **Spearman** και του **Kendall** υπολογίζονται με βάση τις τάξεις μεγέθους των δεδομένων είναι που επιτρέπει την ελευθερία ως προς τη μη ικανοποίηση της κανονικότητας των μεταβλητών.





# Συσχέτιση

## Παρατηρήσεις

- Με το συντελεστή γραμμικής συσχέτισης ελέγχουμε αν σε ένα ζεύγος μεταβλητών υπάρχει **γραμμική συσχέτιση** μόνο. Δηλαδή μπορεί να υπάρχει συσχέτιση μεταξύ των δύο μεταβλητών, αλλά όχι γραμμικής φύσεως. Σε αυτήν την περίπτωση αυτή η σχέση που συνδέει τις δύο μεταβλητές δεν μπορεί να ανιχνευτεί με το συντελεστή γραμμικής συσχέτισης.
- Οπότε **προσοχή** στην ερμηνεία που δίνουμε στο συντελεστή συσχέτισης. Να υπενθυμίσουμε επίσης ότι η λογική με την οποία απορρίπτουμε ή όχι μία υπόθεση είναι πάντα η ίδια. Αν το παρατηρηθέν επίπεδο στατιστικής σημαντικότητας είναι μικρότερο του 0,05 η υπόθεση απορρίπτεται. Στην αντίθετη περίπτωση δεν απορρίπτεται

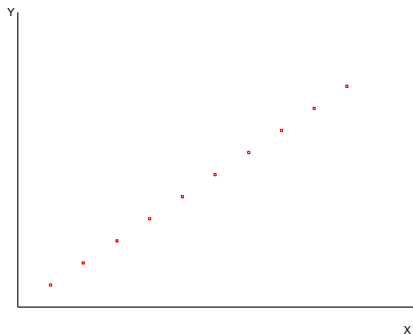


# Συσχέτιση

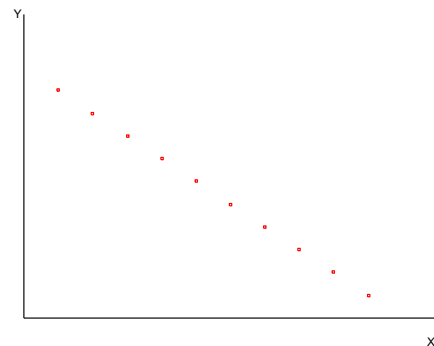
## Παρατηρήσεις

- Ο συντελεστής Συσχέτισης είναι ένας **αριθμός** ανάμεσα στο  $-1$  και το  $1$ .
- Οι τιμές  $r = -1$  και  $r = 1$  προκύπτουν όταν υπάρχει πλήρης γραμμική σχέση μεταξύ των δύο μεταβλητών  $X$  και  $Y$ . Όταν, δηλαδή, τα σημεία του αντίστοιχου διαγράμματος διασποράς που ορίζεται από τα ζεύγη των τιμών  $(x_i, y_i)$ , βρίσκονται κατά μήκος μιας ευθείας γραμμής

Πλήρης θετική συσχέτιση  $r = 1$



Πλήρης αρνητική συσχέτιση  $r = -1$





# Συσχέτιση

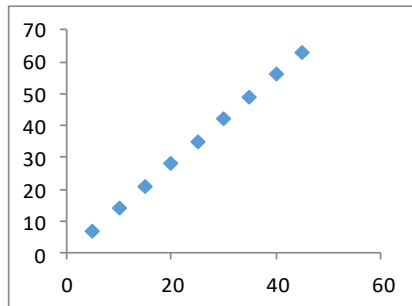
## Παρατηρήσεις

- Όταν ο **συντελεστής** είναι κοντά στο **0**, **δεν** υπάρχει σχέση ή η σχέση είναι **πολύ μικρή**. Δηλαδή όσο η σχέση μεταξύ των  $X$  και  $Y$  αποκλίνει από την πλήρη γραμμικότητα, η τιμή του  $r$  τείνει να απομακρύνεται από τις τιμές **-1** και **1** και να πλησιάζει το **0**.
- Όταν οι τιμές της  $Y$  τείνουν να **αυξάνουν** όσο **αυξάνουν** και οι αντίστοιχες τιμές της  $X$ , η τιμή του  $r$  είναι **θετική** και οι μεταβλητές χαρακτηρίζονται **θετικά συσχετιζόμενες**.
- Στην αντίστροφη περίπτωση, όπου οι τιμές της  $Y$  **ελαττώνονται** όσο οι τιμές της  $X$  **αυξάνουν**, ο συντελεστής συσχέτισης  $r$  παίρνει **αρνητικές** τιμές και οι δύο μεταβλητές χαρακτηρίζονται **αρνητικά συσχετιζόμενες**.

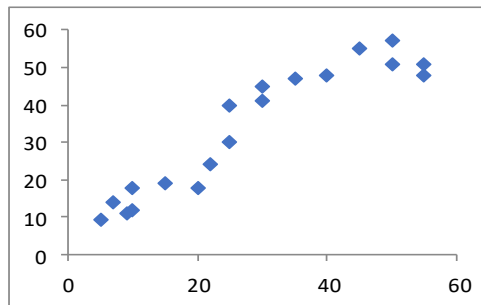


# Συσχέτιση

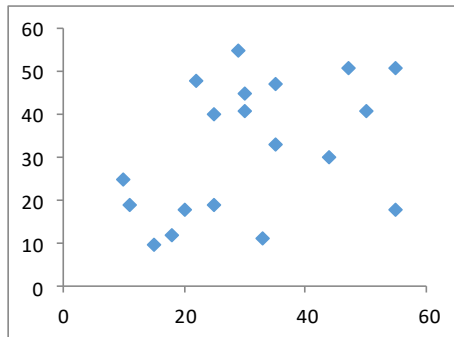
## Παρατηρήσεις



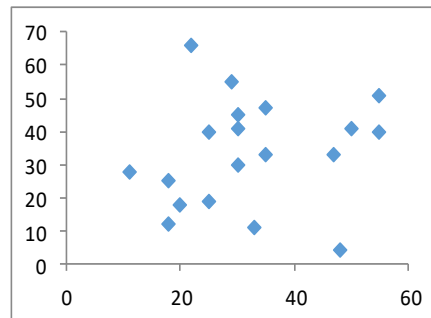
$r = 1$



$r = 0,94$



$r = 0,41$



$r = 0,15$

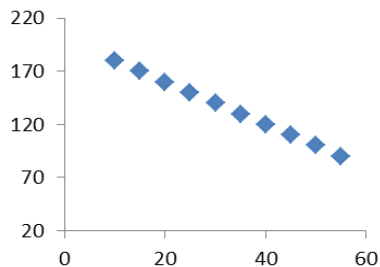
**Θετική Συσχέτιση**  
**( $r > 0$ )**

**όταν αυξάνει το ένα  
χαρακτηριστικό  
αυξάνει και το άλλο**

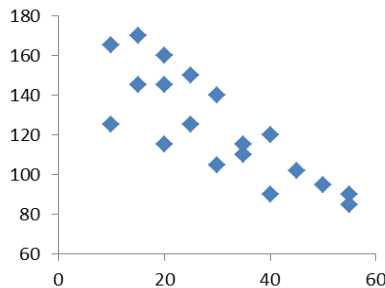


# Συσχέτιση

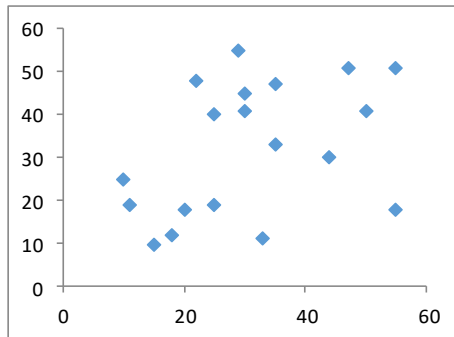
## Παρατηρήσεις



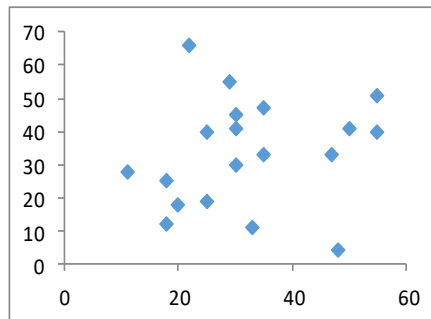
$$r = -1$$



$$r = -0,83$$



$$r = -0,55$$



$$r = -0,14$$

**Αρνητική Συσχέτιση**  
**( $r < 0$ )**

**όταν αυξάνει το ένα  
χαρακτηριστικό  
μειώνεται το άλλο**



# Συσχέτιση

## Παρατηρήσεις

Συμπερασματικά :

- 1) Όταν  $0 < r \leq 0,25$ , τότε είναι πολύ **μικρή** συσχέτιση
- 2) Όταν  $0,25 \leq r \leq 0,50$ , τότε έχουμε **ελαφρά** συσχέτιση
- 3) Όταν  $0,50 \leq r \leq 0,75$ , η συσχέτιση είναι **σχετικά ισχυρή** και
- 4) Όταν  $0,75 \leq r \leq 1$ , τότε η συσχέτιση είναι **πολύ ισχυρή**.

*Αντίστοιχα μπορούμε να πούμε και για τις αρνητικές τιμές του  $r$*



# Συσχέτιση

## Προϋποθέσεις

- Γραμμικότητα.
- Κανονικότητα
- Συνέχεια
- Ανεξαρτησία των παρατηρήσεων (ατόμων)

## Περιορισμοί

- Όταν οι μεταβλητές έχουν περιορισμένο εύρος (μειώνεται η τιμή του συντελεστή)
- Όταν λείπουν πολλές παρατηρήσεις
- Όταν έχετε μικρά δείγματα, συνεπώς και μικρή δύναμη ανίχνευσης στατιστικά σημαντικών σχέσεων, αλλά και λίγη αυτοπεποίθηση για εύρεση της πραγματικής σχέσης



## Παράδειγμα I

Έχουμε δύο Ερωτήσεις  $Q_1$ : “Πόσες ώρες κοιμήθηκες χθες” και  $Q_2$ : “Πόσο όμορφα αισθάνεσαι τώρα” σε κλίμακα από 0 «καθόλου καλά», 8 «Υπερβολικά καλά» (αρχείο lecture2\_1.sav)

Θέλουμε να διερευνήσουμε το είδος σχέσης μεταξύ των δύο μεταβλητών

| Ωρες Ύπνου | Διάθεση |
|------------|---------|
| 7          | 4       |
| 5          | 2       |
| 8          | 7       |
| 6          | 2       |
| 6          | 3       |
| 10         | 6       |
| 7          | 5       |
| 5          | 3       |
| 8          | 8       |
| 9          | 8       |
| 7          | 5       |
| 9          | 7       |
| 5          | 3       |
| 6          | 4       |
| 7          | 4       |





## Παράδειγμα I

Αρχικά εξετάζουμε αν οι μεταβλητές ακολουθούν την κανονική κατανομή (κάνουμε και έλεγχο ακραίων τιμών).

- Από το μενού επιλέγουμε **Analyze** → **Descriptive Statistics** → **Explore**
- Στην συνέχεια
  - από την επιλογή **Statistics** τσεκάρουμε τις επιλογές **Descriptives**, **Outliers**, **Percentiles** και ορίζουμε το **Διάστημα Εμπιστοσύνης**.
  - από την επιλογή **Plots** τσεκάρουμε τις επιλογές **Histogram**, και **Normality plots with tests**

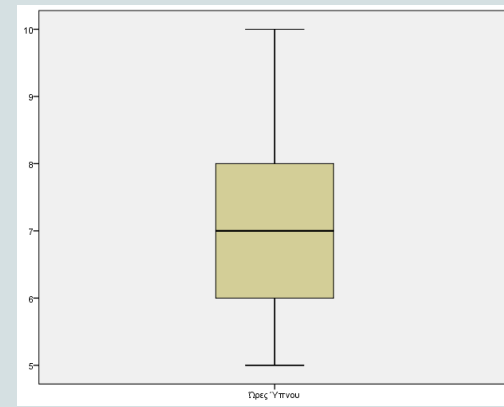
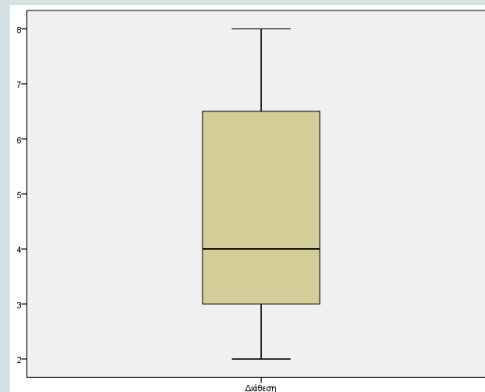


## Παράδειγμα I

|                                 |            | Percentiles |      |      |      |      |      |    |
|---------------------------------|------------|-------------|------|------|------|------|------|----|
|                                 |            | Percentiles |      |      |      |      |      |    |
|                                 |            | 5           | 10   | 25   | 50   | 75   | 90   | 95 |
| Weighted Average (Definition 1) | Ώρες Ύπνου | 5,00        | 5,00 | 6,00 | 7,00 | 8,00 | 9,40 | .  |
|                                 | Διάθεση    | 2,00        | 2,00 | 3,00 | 4,00 | 7,00 | 8,00 | .  |
| Tukey's Hinges                  | Ώρες Ύπνου |             |      | 6,00 | 7,00 | 8,00 |      |    |
|                                 | Διάθεση    |             |      | 3,00 | 4,00 | 6,50 |      |    |

### Tests of Normality

|            | Kolmogorov-Smirnov <sup>a</sup> |    |       | Shapiro-Wilk |    |      |
|------------|---------------------------------|----|-------|--------------|----|------|
|            | Statistic                       | df | Sig.  | Statistic    | df | Sig. |
| Ώρες Ύπνου | ,167                            | 15 | ,200* | ,932         | 15 | ,292 |
| Διάθεση    | ,173                            | 15 | ,200* | ,919         | 15 | ,188 |



Από τα αποτελέσματα παρατηρούμε ότι **δεν** υπάρχουν ακραίες τιμές και ότι δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση. Επομένως οι μεταβλητές μας κατανέμονται κανονικά και μπορούμε να χρησιμοποιήσουμε **παραμετρικό** τεστ.

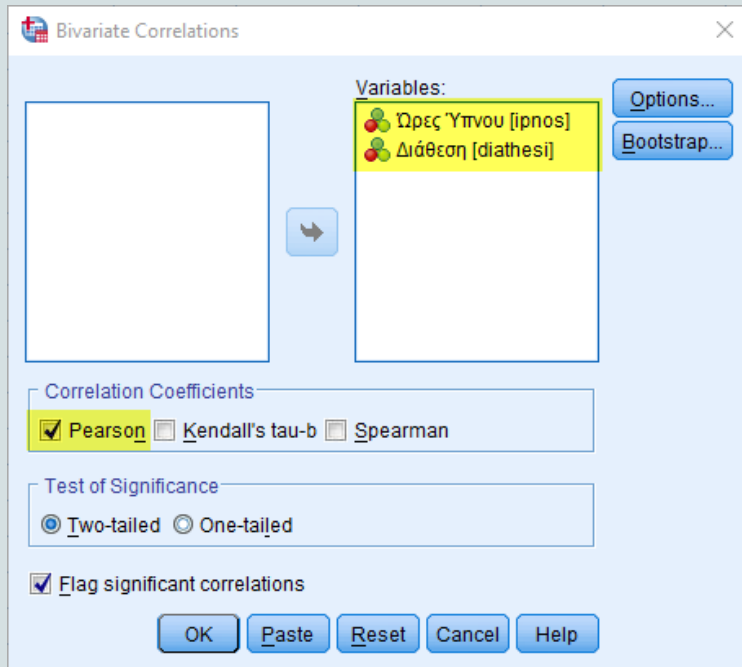
Από την εκφώνηση παρατηρούμε την συνέχεια και την ανεξαρτησία των παρατηρήσεων



## Παράδειγμα I

Από το μενού επιλέγουμε **Analyze** → **Correlate** → **Bivariate**

Στην θυρίδα διαλόγου που παρουσιάζεται μεταφέρουμε τις δύο μεταβλητές υπό εξέταση στην περιοχή **Variables** και εφόσον στο προηγούμενο βήμα διαπιστώσαμε ότι οι μεταβλητές κατανέμονται **κανονικά** επιλέγουμε σαν συντελεστή συσχέτισης τον **συντελεστή Pearson** και εκτελούμε την ανάλυση





## Παράδειγμα I

|            |                     | Ώρες Ύπνου | Διάθεση |
|------------|---------------------|------------|---------|
| Ώρες Ύπνου | Pearson Correlation | 1          | ,849**  |
|            | Sig. (2-tailed)     |            | ,000    |
|            | N                   | 15         | 15      |
| Διάθεση    | Pearson Correlation | ,849**     | 1       |
|            | Sig. (2-tailed)     | ,000       |         |
|            | N                   | 15         | 15      |

\*\* . Correlation is significant at the 0.01 level (2-tailed).

Συντελεστής Συσχέτισης

P value

Ο βαθμός συσχέτισης είναι **0,849** ο οποίος δείχνει μια **πολύ ισχυρή θετική** σχέση μεταξύ των δύο μεταβλητών.

Το **p-value** είναι μικρότερο από το **0,05** άρα μπορούμε να πούμε ότι σχέση αυτή είναι στατιστικά σημαντική.



## Παράδειγμα II

Σε ένα δείγμα 474 εργαζομένων θέλουμε να εξετάσουμε αν υπάρχει σχέση μεταξύ του αρχικού (Salbegin) και του τωρινού μισθού (Salary) (αρχείο employee.sav)

Αρχικά εξετάζουμε αν οι μεταβλητές ακολουθούν την κανονική κατανομή.

- Από το μενού επιλέγουμε **Analyze** → **Descriptive Statistics** → **Explore**
  - από την επιλογή **Statistics** τσεκάρουμε τις επιλογές **Descriptives**, **Outliers**, **Percentiles** και ορίζουμε το **Διάστημα Εμπιστοσύνης**.
  - από την επιλογή **Plots** τσεκάρουμε τις επιλογές **Histogram**, και **Normality plots with tests**

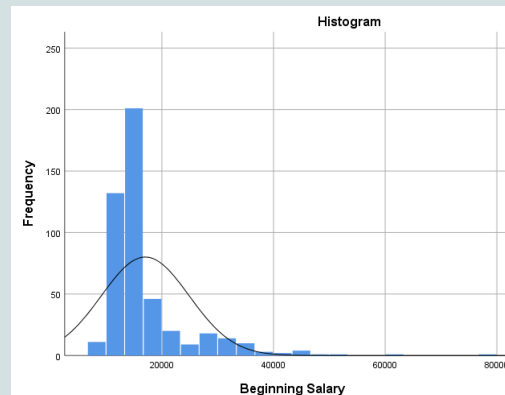
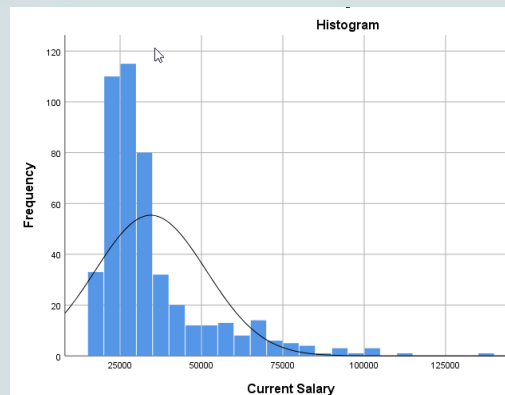


## Παράδειγμα II

### Tests of Normality

|                  | Kolmogorov-Smirnov <sup>a</sup> |     |      | Shapiro-Wilk |     |      |
|------------------|---------------------------------|-----|------|--------------|-----|------|
|                  | Statistic                       | df  | Sig. | Statistic    | df  | Sig. |
| Current Salary   | ,208                            | 474 | ,000 | ,771         | 474 | ,000 |
| Beginning Salary | ,252                            | 474 | ,000 | ,715         | 474 | ,000 |

a. Lilliefors Significance Correction



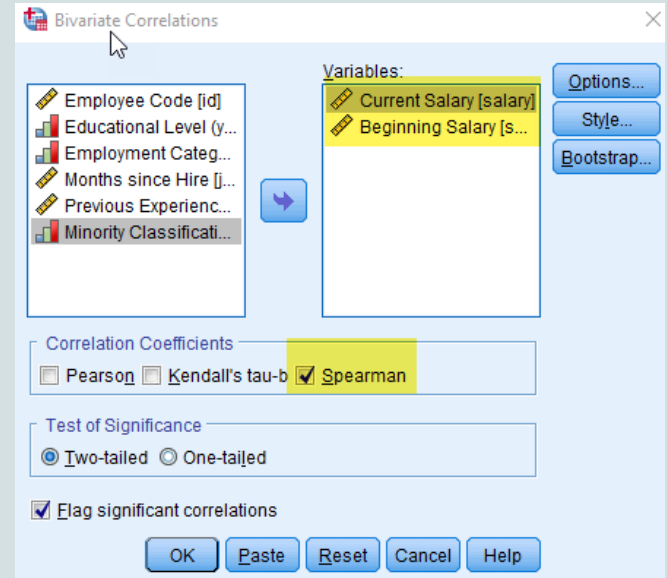
Από τα αποτελέσματα παρατηρούμε ότι καμία από τις δύο μεταβλητές υπό εξέταση **δε ακολουθεί** την **κανονική κατανομή**, εφόσον η τιμή **p-value** είναι μικρότερη του επιπέδου στατιστικής σημαντικότητας ( $\alpha=0,05$ ) που θέσαμε. Κατά συνέπεια πρέπει να εκτελέσουμε ένα μη **παραμετρικό** τεστ.



## Παράδειγμα II

Από το μενού επιλέγουμε **Analyze** → **Correlate** → **Bivariate**

Στην θυρίδα διαλόγου που παρουσιάζεται μεταφέρουμε τις δύο μεταβλητές υπό εξέταση στην περιοχή **Variables** και εφόσον στο προηγούμενο βήμα διαπιστώσαμε ότι οι μεταβλητές **δεν** κατανέμονται κανονικά επιλέγουμε σαν συντελεστή συσχέτισης τον **συντελεστή Spearman** και εκτελούμε την ανάλυση





## Παράδειγμα II

|                  |                  | Correlations            |                |                  |
|------------------|------------------|-------------------------|----------------|------------------|
|                  |                  |                         | Current Salary | Beginning Salary |
| Spearman's rho   | Current Salary   | Correlation Coefficient | 1,000          | ,826**           |
|                  |                  | Sig. (2-tailed)         | .              | ,000             |
|                  | N                |                         | 474            | 474              |
| Beginning Salary | Beginning Salary | Correlation Coefficient | ,826**         | 1,000            |
|                  |                  | Sig. (2-tailed)         | ,000           | .                |
|                  | N                |                         | 474            | 474              |

\*\* . Correlation is significant at the 0.01 level (2-tailed).

Συντελεστής Συσχέτισης

P value

Ο βαθμός συσχέτισης είναι **0,826** ο οποίος δείχνει μια **πολύ ισχυρή θετική** σχέση μεταξύ αρχικού και τωρινού μισθού.

Το **p-value** είναι μικρότερο από το **0,05** άρα μπορούμε να πούμε ότι σχέση αυτή είναι στατιστικά σημαντική.





## Άσκηση I

Έχουμε στον διπλανό πίνακα το ύψος και το βάρος 15 φοιτητών (αρχείο `lecture2_2.sav`).

Διερευνήστε το είδος σχέσης μεταξύ των δύο μεταβλητών

| Βάρος | Ύψος |
|-------|------|
| 58    | 170  |
| 60    | 172  |
| 67    | 173  |
| 72    | 175  |
| 65    | 172  |
| 81    | 175  |
| 73    | 171  |
| 74    | 174  |
| 73    | 174  |
| 68    | 171  |
| 70    | 173  |
| 75    | 175  |
| 78    | 176  |
| 80    | 177  |
| 79    | 176  |



# Έλεγχος Μέσης Τιμής

Χρησιμοποιείται σε περιπτώσεις προβλημάτων στα οποία θέλουμε να ελέγξουμε αν ένα δείγμα προέρχεται από κάποιο πληθυσμό με γνωστό μέσο όρο ή να ελέγξουμε αν ο μέσος όρος ενός δείγματος είναι ίσος με τον μέσο όρο του γενικού πληθυσμού που θεωρούμε ότι είναι γνωστός.

## Βασικές Προϋποθέσεις είναι:

- Θα πρέπει στις τιμές να μην υπάρχουν σημαντικά **ακραίες τιμές** (outliers).
- Τα στοιχεία του δείγματος θα πρέπει να είναι **ανεξάρτητα** (μη συσχετιζόμενα), το οποίο σημαίνει ότι δεν υπάρχει σχέση μεταξύ των παρατηρήσεων
- Αν το δείγμα ακολουθεί την **κανονική κατανομή** εκτελούμε **παραμετρικό** τεστ (One sample t test) διαφορετικά εκτελούμε το αντίστοιχο **μη παραμετρικό** τεστ (Wilcoxon)



# Έλεγχος Μέσης Τιμής

## Μεθοδολογία

- Θέτουμε το Ερευνητικό Ερώτημα (π.χ. ο μέσος όρος ενός δείγματος 60 ανδρών διαφέρει από τα 174 εκατοστά)
- Γράφουμε την μηδενική και την εναλλακτική υπόθεση ( $H_0: \mu = \mu_0 = 174$  και  $H_1: \mu \neq \mu_0 = 174$ )
- Ορίζουμε το επίπεδο στατιστικής σημαντικότητας (π.χ.  $\alpha=0,05$ )
- Εξετάζουμε τις προϋποθέσεις εκτέλεσης του τεστ όσο αφορά την ανεξαρτησία και τις ακραίες τιμές
- Εξετάζουμε αν πρέπει να εκτελέσουμε παραμετρικό τεστ (κανονική κατανομή) ή όχι



# Έλεγχος Μέσης Τιμής

## Μεθοδολογία

- Βρίσκουμε την τιμή **p-value** του τεστ που επιλέξαμε
- Συγκρίνουμε την τιμή **p-value** με το επίπεδο στατιστικής σημαντικότητας  $\alpha$  που θέσαμε (Αν  $p > \alpha$  αποτυγχάνουμε να απορρίψουμε τη μηδενική υπόθεση και συμπεραίνουμε ότι δεν υπάρχει στατιστικά σημαντική διαφορά μεταξύ του μέσου όρου και της συγκρινόμενης τιμής)
- Διατυπώνουμε την **απάντηση** στο **Ερευνητικό Ερώτημα**



## Παράδειγμα I

Έχουμε ένα δείγμα 60 ανδρών (αρχείο `lecture2_3.sav`) και θέλουμε να ελέγξουμε αν ο μέσος όρος του δείγματος είναι ίσος με 174 εκατοστά σε επίπεδο στατιστικής σημαντικότητας  $\alpha=0,05$ .

### Βήμα 1<sup>ο</sup>

Αρχικά εφόσον έχουμε την ερευνητική υπόθεση συντάσσουμε την μηδενική και την εναλλακτική υπόθεση

**$H_0$ : Μηδενική υπόθεση  $\mu=\mu_0$  έναντι της  $H_a$ : Εναλλακτικής υπόθεσης  $\mu\neq\mu_0$**

### Βήμα 2<sup>ο</sup>

Εφόσον έχει οριστεί το επίπεδο στατιστικής σημαντικότητας εξετάζουμε το δείγμα όσο αφορά την ανεξαρτησία (τυχειότητα), τις ακραίες τιμές και την κανονικότητα για να αποφασίσουμε πιο τεστ θα εκτελέσουμε



## Παράδειγμα I

### Βήμα 2<sup>ο</sup>

Η ανεξαρτησία προκύπτει από τον τρόπο συλλογής του δείγματος επομένως ελέγχουμε το δείγμα όσο αφορά την ύπαρξη ακραίων τιμών και την κανονικότητα

- Από το μενού επιλέγουμε **Analyze** → **Descriptive Statistics** → **Explore**
  - ✓ από την επιλογή **Statistics** τσεκάρουμε τις επιλογές **Descriptives**, **Outliers**, **Percentiles** και ορίζουμε το **Διάστημα Εμπιστοσύνης**.
  - ✓ από την επιλογή **Plots** τσεκάρουμε τις επιλογές **Histogram**, και **Normality plots with tests**

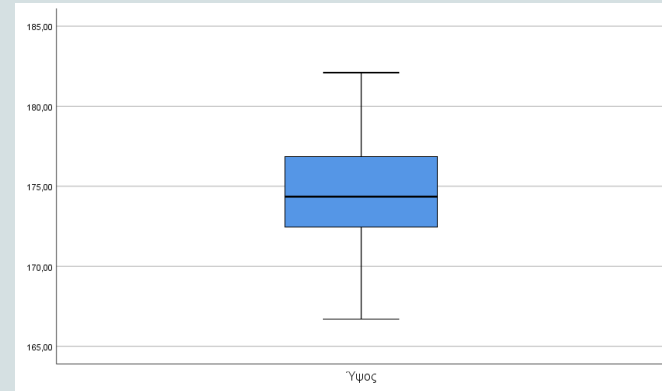
# Έλεγχος Μέσης Τιμής



## Παράδειγμα I

|                                 |      | Percentiles |          |          |          |          |          |          |
|---------------------------------|------|-------------|----------|----------|----------|----------|----------|----------|
|                                 |      | Percentiles |          |          |          |          |          |          |
|                                 |      | 5           | 10       | 25       | 50       | 75       | 90       | 95       |
| Weighted Average (Definition 1) | Υψος | 169,6000    | 170,8500 | 172,3750 | 174,3500 | 176,9250 | 179,0000 | 179,4850 |
| Tukey's Hinges                  | Υψος |             |          | 172,4500 | 174,3500 | 176,8500 |          |          |

| Tests of Normality              |           |    |                   |           |    |      |
|---------------------------------|-----------|----|-------------------|-----------|----|------|
| Kolmogorov-Smirnov <sup>a</sup> |           |    | Shapiro-Wilk      |           |    |      |
|                                 | Statistic | df | Sig.              | Statistic | df | Sig. |
| Υψος                            | ,093      | 60 | ,200 <sup>*</sup> | ,988      | 60 | ,826 |



Από τα αποτελέσματα παρατηρούμε ότι **δεν** υπάρχουν ακραίες τιμές και ότι δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση για την κανονικότητα. Επομένως η μεταβλητή κατανέμεται κανονικά επομένως μπορούμε να χρησιμοποιήσουμε **παραμετρικό** τεστ.

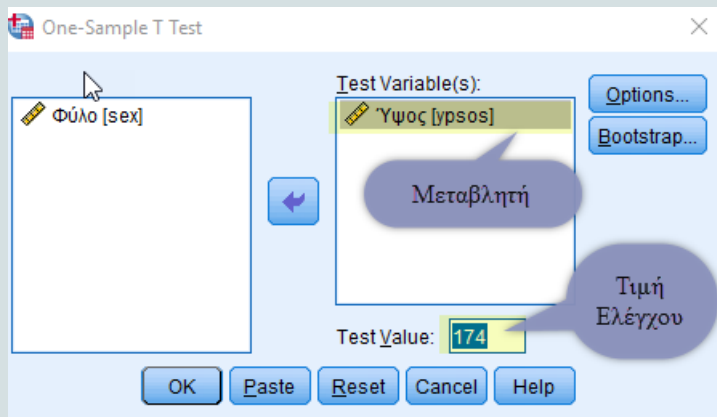


## Παράδειγμα I

### Βήμα 3<sup>ο</sup>

Εφόσον το δείγμα μας ακολουθεί την κανονική κατανομή θα εκτελέσουμε **παραμετρικό** τεστ για να βρούμε την τιμή **p-value**

**Analyze** → **Compare Means** → **One sample t test**



Στο παράθυρο διαλόγου βάζουμε την μεταβλητή υπό εξέταση στο κουτί «test Variables» και γράφουμε την τιμή ελέγχου στο πεδίο “Test Value”



# Έλεγχος Μέσης Τιμής



## Παράδειγμα I

| One-Sample Test |       |    |                 |                  |   |        |
|-----------------|-------|----|-----------------|------------------|---|--------|
|                 |       |    |                 | Test Value = 174 |   |        |
|                 |       |    |                 |                  | 95% Confidence Interval of the Difference |        |
|                 | t     | df | Sig. (2-tailed) | Mean Difference  | Lower                                     | Upper  |
| Ύψος            | 1,716 | 59 | ,091            | ,66833           | -,1112                                    | 1,4478 |

### Βήμα 4<sup>ο</sup>

Η τιμή **p-value** είναι **0,091**. Εφόσον η τιμή **p-value** είναι μεγαλύτερη του επιπέδου στατιστικής σημαντικότητας που θέσαμε ( $\alpha=0,05$ ) δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση ότι ο μέσος όρος του δείγματος δεν διαφέρει από τα **174** εκατοστά

# Έλεγχος Μέσης Τιμής



## Παράδειγμα I

| One-Sample Test |       |    |                 |                  |   |        |
|-----------------|-------|----|-----------------|------------------|---|--------|
|                 |       |    |                 | Test Value = 174 |   |        |
|                 |       |    |                 |                  | 95% Confidence Interval of the Difference |        |
|                 | t     | df | Sig. (2-tailed) | Mean Difference  | Lower                                     | Upper  |
| Ύψος            | 1,716 | 59 | ,091            | ,66833           | -,1112                                    | 1,4478 |

### Βήμα 5<sup>ο</sup>

Θέλοντας να γενικεύσουμε τα αποτελέσματα στον πληθυσμό → αν πάρουμε ένα τυχαίο δείγμα ανδρών είμαστε 95% σίγουροι ότι ο μέσος όρος του ύψους τους θα είναι 174 εκατοστά σε ένα διάστημα εμπιστοσύνης από  $174 - 0,1112$  έως  $174 + 1,4478$  εκατοστά

# Έλεγχος Μέσης Τιμής



## Παράδειγμα II

Το αρχείο adratings.sav περιέχει τα δεδομένα με τις βαθμολογίες τριών διαφημιστικών αυτοκινήτων 18 ατόμων καθώς επίσης το φύλλο και η ηλιακή κατηγορία κάθε ερωτώμενου. Το ερευνητικό ερώτημα είναι αν ο μέσος βαθμολογίας για κάθε διαφήμιση είναι 79,5.

### Βήμα 1<sup>ο</sup>

Αρχικά εφόσον έχουμε την ερευνητική υπόθεση συντάσσουμε την μηδενική και την εναλλακτική υπόθεση

$H_0$ : Μηδενική υπόθεση  $\mu = \mu_0 = 79,5$  έναντι της  $H_a$ : Εναλλακτικής υπόθεσης  $\mu \neq \mu_0 = 79,5$

### Βήμα 2<sup>ο</sup>

Εφόσον έχει οριστεί το επίπεδο στατιστικής σημαντικότητας εξετάζουμε το δείγμα όσο αφορά την ανεξαρτησία (τυχειότητα), τις ακραίες τιμές και την κανονικότητα για να αποφασίσουμε πιο τεστ θα εκτελέσουμε



## Παράδειγμα II

Η ανεξαρτησία προκύπτει από τον τρόπο συλλογής του δείγματος επομένως ελέγχουμε το δείγμα όσο αφορά την ύπαρξη ακραίων τιμών και την κανονικότητα

➤ Από το μενού επιλέγουμε **Analyze** → **Descriptive Statistics** → **Explore**

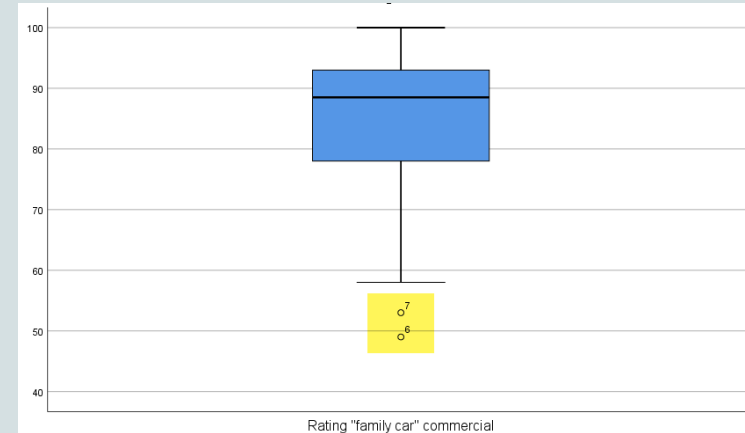
- ✓ από την επιλογή **Statistics** τσεκάρουμε τις επιλογές **Descriptives**, **Outliers**, **Percentiles** και ορίζουμε το **Διάστημα Εμπιστοσύνης**.
- ✓ από την επιλογή **Plots** τσεκάρουμε τις επιλογές **Histogram**, και **Normality plots with tests**

# Έλεγχος Μέσης Τιμής



## Παράδειγμα II

|  | Tests of Normality              |    |       |              |    |      |
|--|---------------------------------|----|-------|--------------|----|------|
|  | Kolmogorov-Smirnov <sup>a</sup> |    |       | Shapiro-Wilk |    |      |
|  | Statistic                       | df | Sig.  | Statistic    | df | Sig. |
| Rating "family car" commercial                   | ,237                            | 18 | ,009  | ,826         | 18 | ,004 |
| Rating "youngster car" commercial                | ,159                            | 18 | ,200* | ,956         | 18 | ,522 |
| Rating "environmentally friendly car" commercial | ,135                            | 18 | ,200* | ,974         | 18 | ,876 |



Από τα αποτελέσματα παρατηρούμε ότι **υπάρχουν** ακραίες τιμές για την μεταβλητή **ad1** (Family Car) και ότι **δεν** ακολουθεί την **κανονική κατανομή**. Οι άλλες δύο μεταβλητές ακολουθούν την κανονική κατανομή. Επομένως μπορούμε να εκτελέσουμε ένα **μη παραμετρικό** τεστ για την πρώτη μεταβλητή και ένα **παραμετρικό** για τις άλλες δύο.

# Έλεγχος Μέσης Τιμής



Το αντίστοιχο μη παραμετρικό τεστ καλείται έλεγχος των προσημασμένων τάξεων μεγέθους του **Wilcoxon** για τη **διάμεσο** ενός πληθυσμού.

Επειδή βασίζεται στις τάξεις μεγέθους των παρατηρήσεων και όχι στις παρατηρήσεις αυτές κάθε αυτές δε χρειάζεται **καμία προϋπόθεση** ως προς την κατανομή των παρατηρήσεων.

Ο έλεγχος όμως εδώ βασίζεται στη **διάμεσο** και όχι στο μέσο του δείγματος. Αυτό βέβαια **δε** σημαίνει ότι αυτού του είδους τα τεστ είναι χαμηλότερης αξιοπιστίας σε σύγκριση με τα παραμετρικά. Απεναντίας, σε πολλές περιπτώσεις είναι **πιο ισχυρά**.

Ένα **πλεονέκτημα** όμως των μη παραμετρικών ελέγχων που θα εξεταστούν είναι ότι εφαρμόζονται και στις περιπτώσεις που οι μεταβλητές είναι **ποιοτικές διατεταγμένης κλίμακας** (καλό, καλύτερο, πολύ καλό). Σε αυτήν την περίπτωση οι μέθοδοι ελέγχων υποθέσεων που θα δούμε είναι από **τις πιο ισχυρές** μεθόδους που υπάρχουν



## Παράδειγμα II

### Βήμα 3<sup>ο</sup>

Για την πρώτη μεταβλητή η οποία δεν ακολουθεί την κανονική κατανομή ακολουθούμε την παρακάτω διαδικασία για να βρούμε την τιμή **p-value**

**Analyze** → **Non Parametric Test** → **One sample**

One-Sample Nonparametric Tests

Objective Fields Settings

Identifies differences in single fields using one or more nonparametric tests. Nonparametric tests do not assume your data follow the normal distribution.

What is your objective?

Each objective corresponds to a distinct default configuration on the Settings Tab that you can further customize, if desired.

- Automatically compare observed data to hypothesized
- Test sequence for randomness
- Customize analysis

Από την καρτέλα **Objective** του παραθύρου διαλόγου που εμφανίζεται επιλέγουμε **Customize Analysis**”



## Παράδειγμα II

### Βήμα 3<sup>ο</sup>

The screenshot shows the 'One-Sample Nonparametric Tests' dialog box in SPSS, with the 'Fields' tab selected. The 'Use custom field assignments' radio button is chosen. In the 'Fields:' list, the variable 'Rating "family car" commercial' is highlighted in yellow. A blue arrow button is visible between the 'Fields:' and 'Test Fields:' lists. The 'Test Fields:' list is currently empty.

Στην καρτέλα **Fields** του παραθύρου διαλόγου βάζουμε τις μεταβλητές υπό εξέταση στο κουτί **Test Fields**





## Παράδειγμα II

### Βήμα 3<sup>ο</sup>

One-Sample Nonparametric Tests

Objective Fields Settings

Select an item:

- Choose Tests
- Test Options
- User-Missing Values

Automatically choose the tests based on the data

Customize tests

- Compare observed binary probability to hypothesized (Binomial test)  
Options...
- Compare observed probabilities to hypothesized (Chi-Square test)  
Options...
- Test observed distribution against hypothesized (Kolmogorov-Smirnov test)  
Options...
- Compare median to hypothesized (Wilcoxon signed-rank test)  
Hypothesized median: 79,5

Στην καρτέλα **Settings** του παραθύρου διαλόγου Στο πεδίο **Hypothesized median** βάζουμε την τιμή του μέσου όρου (79,5) που θέλουμε να ελέγξουμε και πατάμε το πλήκτρο **Run** για να εκτελέσουμε την ανάλυση

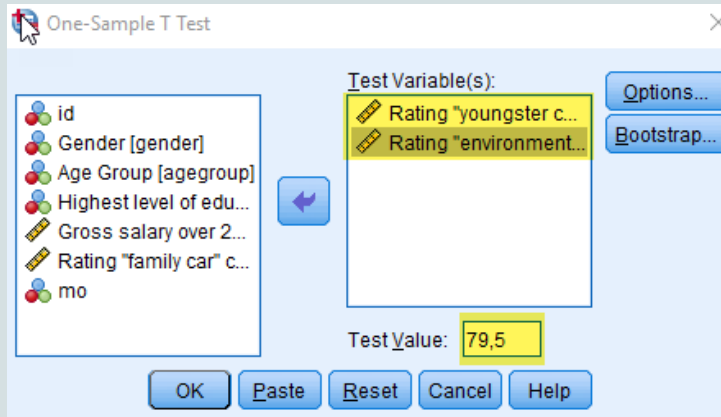


## Παράδειγμα II

### Βήμα 3<sup>ο</sup>

Εφόσον οι άλλες δύο μεταβλητές ακολουθούν την κανονική εκτελούμε **παραμετρικό** τεστ για να βρούμε την τιμή **p-value**

**Analyze** → **Compare Means** → **One sample t test**



Στο παράθυρο διαλόγου βάζουμε τις μεταβλητές υπό εξέταση στο κουτί «**test Variables**» και γράφουμε την τιμή ελέγχου στο πεδίο “**Test Value**”



## Παράδειγμα II

### Βήμα 4<sup>ο</sup>

Όσο αφορά την πρώτη μεταβλητή παρατηρούμε ότι η τιμή **p-value** είναι **0,191**. Εφόσον η τιμή **p-value** είναι μεγαλύτερη του επιπέδου στατιστικής σημαντικότητας που θέσαμε ( $\alpha=0,05$ ) δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση ότι ο μέσος όρος του δείγματος δεν διαφέρει από το **79,5**

Σ

### Hypothesis Test Summary

|   | Null Hypothesis   | Test                                 | Sig. | Decision                    |
|---|---|--------------------------------------|------|-----------------------------|
| 1 | The median of Rating "family car" commercial equals 80. | One-Sample Wilcoxon Signed Rank Test | ,191 | Retain the null hypothesis. |

Asymptotic significances are displayed. The significance level is ,05.



## Παράδειγμα II

### Βήμα 4<sup>ο</sup>

Όσο αφορά τις δύο άλλες μεταβλητές οι τιμές **p-value** είναι **0,000** και **0,006** αντίστοιχα και εφόσον οι τιμές **p-value** είναι μικρότερες του επιπέδου στατιστικής σημαντικότητας που θέσαμε ( $\alpha=0,05$ ) απορρίπτουμε την μηδενική υπόθεση ότι ο μέσος όρος των δύο μεταβλητών δεν διαφέρει από το **79,5** που θέσαμε ως τιμή ελέγχου

↳

| One-Sample Test                                  |        |    |                 |                 |   |        |
|--|--------|----|-----------------|-----------------|---|--------|
| Test Value = 79.5                                |        |    |                 |                 |   |        |
|  | t      | df | Sig. (2-tailed) | Mean Difference | 95% Confidence Interval of the Difference |        |
|  |        |    |                 |                 | Lower                                     | Upper  |
| Rating "youngster car" commercial                | -5,120 | 17 | ,000            | -24,500         | -34,60                                    | -14,40 |
| Rating "environmentally friendly car" commercial | -3,135 | 17 | ,006            | -14,000         | -23,42                                    | -4,58  |



## Παράδειγμα II

### Βήμα 5<sup>ο</sup>

Το γενικά συμπεράσματα του ελέγχου είναι :

- ✓ Η πρώτη μεταβλητή παρουσιάζει δύο ακραίες τιμές και δεν ακολουθεί την κανονική κατανομή επομένως πρέπει να εκτελέσουμε ένα μη παραμετρικό τεστ.
- ✓ Οι άλλες δύο μεταβλητές δεν περιέχουν ακραίες τιμές και ακολουθούν την κανονική κατανομή επομένως μπορούμε να τρέξουμε ένα παραμετρικό τεστ για τον έλεγχο της μηδενικής υπόθεσης (ότι ο μέσος όρος των εκτιμήσεων για τα διαφημιστικά σποτ είναι **79,5**)
- ✓ Σύμφωνα με τα αποτελέσματα **αποδεχόμαστε** την μηδενική υπόθεση για το πρώτο διαφημιστικό σποτ και την **απορρίπτουμε** για τα άλλα δύο

# Έλεγχος Μέσης Τιμής



## Άσκηση 1<sup>η</sup>

Το αρχείο `testscores.sav` περιλαμβάνει την απόδοση 25 φοιτητών στα μαθήματα Μαθηματικά, Προγραμματισμός, Άλγεβρα, Βάσεις Δεδομένων και Στατιστική.

- ✓ Να διερευνήσετε την σχέση μεταξύ των μεταβλητών
- ✓ Να ελέγξετε κατά πόσο ο μέσος όρος των μεταβλητών είναι ίσος με 55

| math | programming | algebra | database | statistic |
|------|-------------|---------|----------|-----------|
| 46   | 68          | 53      | 46       | 47        |
| 72   | 64          | 60      | 56       | 62        |
| 41   | 52          | 51      | 50       | 39        |
| 86   | 88          | 79      | 76       | 91        |
| 56   | 66          | 62      | 66       | 50        |
| 22   | 52          | 48      | 48       | 38        |
| 49   | 56          | 61      | 64       | 51        |
| 40   | 61          | 64      | 62       | 42        |
| 32   | 42          | 53      | 38       | 32        |
| 19   | 50          | 57      | 40       | 34        |
| 42   | 59          | 64      | 47       | 62        |
| 50   | 72          | 61      | 50       | 59        |
| 74   | 85          | 80      | 76       | 73        |
| 46   | 48          | 68      | 72       | 72        |
| 34   | 56          | 54      | 65       | 59        |
| 60   | 60          | 64      | 62       | 61        |
| 52   | 52          | 62      | 48       | 60        |
| 12   | 45          | 42      | 32       | 26        |
| 66   | 63          | 52      | 50       | 42        |
| 69   | 82          | 80      | 76       | 72        |
| 38   | 60          | 60      | 52       | 73        |
| 29   | 56          | 59      | 50       | 40        |
| 46   | 66          | 66      | 64       | 62        |
| 64   | 67          | 69      | 72       | 68        |
| 24   | 45          | 48      | 39       | 30        |



## Άσκηση 2<sup>η</sup>

Στον παρακάτω πίνακα φαίνονται οι μισθοί (σε χρηματικές μονάδες) 40 τυχαία επιλεγμένων υπαλλήλων που εργάζονται στον ιδιωτικό τομέα.

|        |        |        |        |        |        |        |        |        |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 784,53 | 844,21 | 801,23 | 793,45 | 798,53 | 803,46 | 823,88 | 837,13 | 809,43 |
| 856,34 | 789,21 | 859,32 | 807,67 | 798,12 | 800,56 | 840,43 | 787,13 | 815,43 |
| 806,24 | 784,12 | 844,44 | 825,53 | 779,23 | 824,12 | 813,56 | 817,43 | 825,51 |
| 816,34 | 803,72 | 804,12 | 835,37 | 849,43 | 834,29 | 826,21 | 827,10 | 828,78 |

Να ελέγξετε, με επίπεδο σημαντικότητας 5%, αν ο μέσος μισθός  $\mu$  του πληθυσμού από τον οποίο προέρχεται το παραπάνω δείγμα είναι ίσος με  $\mu_0 = 817$  χρηματικές μονάδες.



## Άσκηση 3<sup>η</sup>

Έρευνες έχουν δείξει ότι ο μέσος όρος ύπνου είναι 8 ώρες στο γενικό πληθυσμό (Knutson, et al., 2010).

Μας ενδιαφέρει αν οι φοιτητές έχουν διαφορετικές συνήθειες από το γενικό πληθυσμό σε σχέση με τις ώρες ύπνου καθώς έχει παρατηρηθεί ότι έρχονται συχνά αργοπορημένοι στα μαθήματά τους λόγω του ότι έχουν ξενυχτήσει.

Ρωτήθηκαν 10 φοιτητές για τις ώρες ύπνου τους και απάντησαν:

6, 5, 4, 3, 10, 7, 6, 5, 8, 7

Να βρεθεί αν όντως οι φοιτητές κοιμούνται διαφορετικές ώρες από το γενικό πληθυσμό.





## Άσκηση 4<sup>η</sup>

Ένας παραγωγός σταφυλιών ισχυρίζεται ότι τα κιβώτια σταφυλιών που συσκευάζει έχουν βάρος **19** κιλά κατά μέσο όρο. Για να ελεγχθεί ο ισχυρισμός του, λαμβάνουμε ένα τυχαίο δείγμα από 10 κιβώτια. Οι μετρήσεις δίδονται στον πίνακα που ακολουθεί.

**Βάρος κιβωτίων του δείγματος (σε κιλά)**

|      |      |      |      |      |      |      |      |      |      |
|------|------|------|------|------|------|------|------|------|------|
| 12,0 | 20,0 | 18,5 | 17,6 | 15,0 | 14,8 | 16,7 | 21,6 | 15,2 | 16,0 |
|------|------|------|------|------|------|------|------|------|------|

Βάσει των πληροφοριών του δείγματος, μπορεί ο ισχυρισμός του να θεωρηθεί εύλογος σε  $\alpha=5\%$ ;



Ερωτήσεις ?????????????????????????????????