Ε Λ Λ Η Ν Ι Κ Η   Δ Η Μ Ο Κ Ρ Α Τ Ι Α

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

Καθηγητής Ι. Μητρόπουλος
ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ & ΤΕΧΝΟΛΟΓΙΑΣ
Εργαστήριο Επιχειρησιακού Σχεδιασμού & Λήψης Αποφάσεων
Τηλ.: +030 2610 369213, email: imitro@upatras.gr
Διεύθυνση: Μεγάλου Αλεξάνδρου 1, 263 34 ΠΑΤΡΑ

Θέμα: ΣΗΜΕΙΩΣΕΙΣ ΓΙΑ ΤΟ ΕΡΓΑΣΤΗΡΙΟ ΤΟΥ ΜΑΘΗΜΑΤΟΣ Ε2

Γ. ΒΑΣΙΟΥ, Α. ΚΑΛΑΠΟΔΗ, Χ. ΠΑΠΑΘΑΝΑΣΟΠΟΥΛΟΥ

Περιεχόμενα

1. ΔΙΑΔΙΚΑΣΙΑ EXPLORE 3
Εφαρμογή της διαδικασίας EXPLORE 5
Άσκηση 12
2. ΕΛΕΓΧΟΣ ΚΑΝΟΝΙΚΟΤΗΤΑΣ – ΔΙΑΣΤΗΜΑ ΕΜΠΙΣΤΟΣΥΝΗΣ ΓΙΑ ΤΟΝ ΜΕΣΟ 13
Εφαρμογή της διαδικασίας 14
Άσκηση 21
3. ΕΛΕΓΧΟΣ ΥΠΟΘΕΣΕΩΝ ΓΙΑ ΤΗ ΜΕΣΗ ΤΙΜΗ ΕΝΟΣ ΔΕΙΓΜΑΤΟΣ (One Sample t-test) 22
Εφαρμογή της διαδικασίας One Sample t-test για αμφίπλευρο έλεγχο 24
Εφαρμογή της διαδικασίας One Sample t-test για μονόπλευρο έλεγχο 27
Ασκήσεις 29
4. ΕΛΕΓΧΟΣ ΥΠΟΘΕΣΕΩΝ ΓΙΑ ΤΟΥΣ ΜΕΣΟΥΣ –ΑΝΕΞΑΡΤΗΤΑ ΔΕΙΓΜΑΤΑ (Independent 30
samples t-test) 30
Εφαρμογή της διαδικασίας Independent samples t-test 33
Ασκήσεις 43
5. ΕΛΕΓΧΟΣ ΥΠΟΘΕΣΕΩΝ ΓΙΑ ΤΟΥΣ ΜΕΣΟΥΣ –ΕΞΑΡΤΗΜΕΝΑ ΔΕΙΓΜΑΤΑ (Paired samples t-test) 45
Εφαρμογή της διαδικασίας Paired samples t-test 47
Άσκηση 53
Επαναληπτικές ασκήσεις στον έλεγχο υποθέσεων 54
6. ΣΥΣΧΕΤΙΣΗ ΠΟΣΟΤΙΚΩΝ ΜΕΤΑΒΛΗΤΩΝ 56
Εφαρμογή της διαδικασίας 58
7. ΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ 63
Εφαρμογή της διαδικασίας 66
Άσκηση 74

1. ΔΙΑΔΙΚΑΣΙΑ EXPLORE

Η διαδικασία Explore στο SPSS αρχικά μας δίνει πληροφορίες κυρίως για:

- Αριθμητικά στατιστικά μέτρα
- Τις ακραίες τιμές που μπορεί να έχει μία μεταβλητή
- Διαγράμματα, τα οποία εκτός του ότι περιγράφουν τα δεδομένα μας παρέχουν ενδείξεις για το αν επιπλέον ακολουθούν την κανονική κατανομή ή όχι.

Τα βήματα που ακολουθούμε για την διαδικασία αυτή στο SPSS είναι τα παρακάτω :

1. Δημιουργούμε την μεταβλητή μας στο Variable View και εισάγουμε τα δεδομένα στο Data View.
2. Επιλέγουμε :

   Analyze → Descriptive Statistics → Explore
3. Στο Dependent List μεταφέρουμε την μεταβλητή μας.
4. Επιλέγουμε Statistics και κλικάρουμε Descriptives (στο Confidence Interval for Mean γράφουμε το διάστημα εμπιστοσύνης που μας ζητάνε π.χ. 95%), Outliers και Percentiles. Πατάμε Continue.
5. Επιλέγουμε Plots και κλικάρουμε Boxplots, Stem and Leaf και Histogram. Πατάμε Continue.
6. Στο Display επιλέγουμε Both και πατάμε OK.

Αποτελέσματα στο Output :

- **Πίνακας Descriptives** (προκύπτει από το κλικάρισμα του Descriptives). Στα στοιχεία αυτού του πίνακα υπάρχουν κυρίως τα ακόλουθα αριθμητικά μέτρα:
  - mean (αριθμητικός μέσος)
  - 95% confidence interval for mean (95% διάστημα εμπιστοσύνης για τον μέσο), δηλαδή ένα διάστημα μέσα στο οποίο είμαστε κατά 95% σίγουροι ότι βρίσκεται ο μέσος της μεταβήτής
  - 5% trimmed mean, δηλαδή ο μέσος που προκύπτει αν αποκλείσουμε από τον υπολογισμό το 5% των μικρότερων και το 5% των μεγαλύτερων τιμών του δείγματος
  - median (διάμεσος)
  - variance (διασπορά)
  - Std. Deviation (τυπική απόκλιση)

- Minimum, Maximum και Range, δηλαδή η μικρότερη τιμή, η μεγαλύτερη τιμή και το εύρος (διαφορά της μικρότερης από την μεγαλύτερη)
- Interquartile Range (ενδοτεταρτημοριακό εύρος), δηλαδή το εύρος που προκύπτει αν αποκλείσουμε από τον υπολογισμό το 25% των μικρότερων και το 25% των μεγαλύτερων τιμών.
- Ο συντελεστής Skewness (ασυμμετρίας) και ο συντελεστής Kurtosis (κύρτωσης).

● **Πίνακας Percentiles** (προκύπτει από το κλικάρισμα του Percentiles) Μας δίνει τα εκατοστημόρια.

● **Πίνακας Extreme Values** (προκύπτει από το κλικάρισμα του Outliers)
Μας δίνει τις μικρότερες και τις μεγαλύτερες τιμές. Με αυτόν τον τρόπο μπορούμε να δούμε αν υπάρχουν ύποπτες (ακραίες) τιμές στην μεταβλητή μας.

● **Διάγραμμα Histogram** (Ιστόγραμμα)

● **Διάγραμμα Stem and Leaf**
Αυτό το διάγραμμα μας δίνει πληροφορίες για τα ψηφία των τιμών τις μεταβλητής μας και αναγνωρίζει τις ακραίες τιμές
.
● **Διάγραμμα Normal Q-Q Plot**
Στον οριζόντιο άξονα έχουμε τις πραγματικές τιμές (observed values) της μεταβλητής μας και στον κατακόρυφο τις τιμές της μεταβλητής που θα έπρεπε να είχαμε (αναμενόμενες τιμές) αν αυτή ακολουθούσε ακριβώς την κανονική κατανομή (expected normal). Όσο πιο κοντά βρίσκονται τα σημεία στην ευθεία του διαγράμματος τόσο πιο πολύ πλησιάζει η μεταβλητή μας την κανονική κατανομή.

● **Διάγραμμα Detrended Normal Q-Q Plot**
Στον οριζόντιο άξονα έχουμε τις πραγματικές τιμές της μεταβλητής και στον κατακόρυφο τις αποκλίσεις τους από την κανονική κατανομή. Για να ακολουθούν τα δεδομένα την κανονική κατανομή, θα πρέπει τα σημεία να είναι κατανεμημένα τυχαία (δηλαδή να μην σχηματίζουν ευθεία ή παραβολή ή κάποιο άλλο πρότυπο) και τα περισσότερα από αυτά να είναι συγκεντρωμένα σε μία ταινία γύρω από την οριζόντια ευθεία.

● **Διάγραμμα Boxplot** (προκύπτει από την επιλογή Boxplots)
Το διάγραμμα αυτό ελέγχει την συμμετρία της κατανομής που είναι προϋπόθεση για να υπάρχει κανονικότητα. Στο ορθογώνιο παραλληλόγραμμο του διαγράμματος η πάνω πλευρά αντιστοιχεί στο Q1, η κάτω στο Q2 και η εσωτερική οριζόντια γραμμή στη διάμεσο. Αν αυτή η γραμμή είναι στο κέντρο του ορθογωνίου τότε έχουμε συμμετρική κατανομή. Αν είναι προς τα κάτω έχουμε θετική ασυμμετρία ενώ αν είναι προς τα πάνω αρνητική ασυμμετρία. Η κατακόρυφη γραμμή συμβολίζει το εύρος που προκύπτει αν αποκλείσουμε από τον υπολογισμό τις ακραίες τιμές. Αν υπάρχουν σημεία εκτός του ορθογωνίου αυτά αφορούν τις ακραίες τιμές και συμβολίζονται με κύκλους.

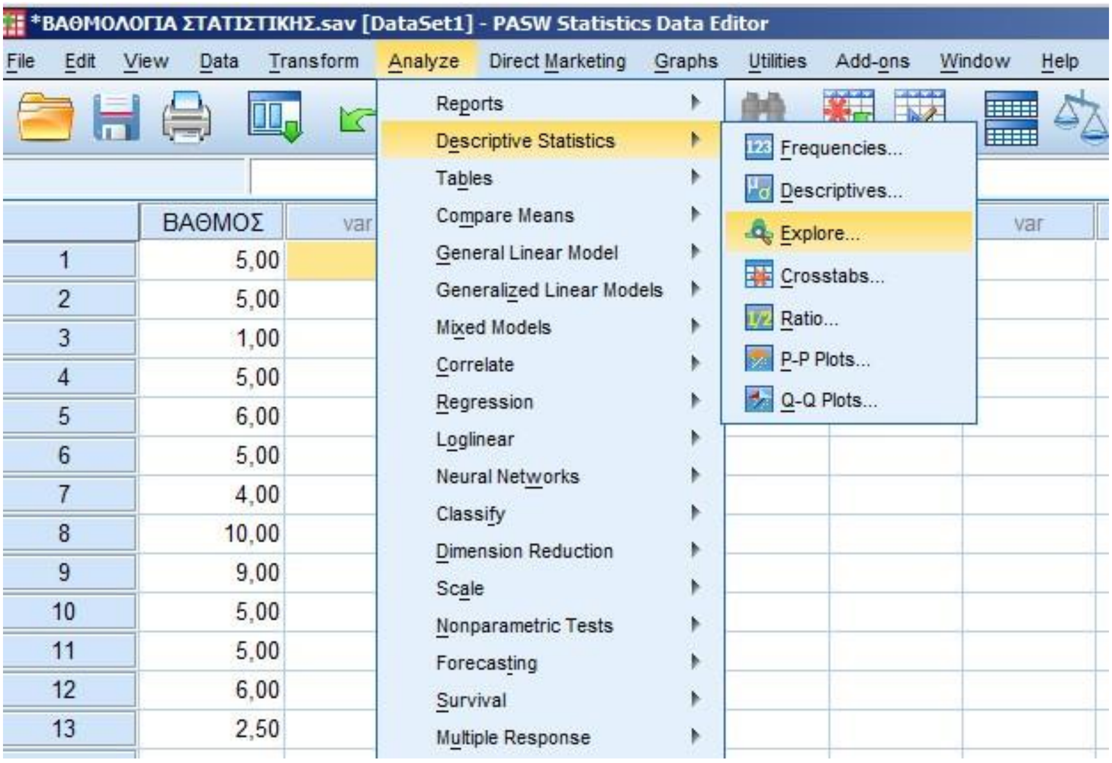
Εφαρμογή της διαδικασίας EXPLORE

Στο παράδειγμα που ακολουθεί περιγράφεται αναλυτικά η διαδικασία Explore και η ερμηνεία των αποτελεσμάτων για τα βασικά αριθμητικά στατιστικά μέτρα. Δεν γίνεται ερμηνεία των διαγραμμάτων. Τα δεδομένα του παραδείγματος μπορείτε να τα βρείτε στο αρχείο **ΒΑΘΜΟΛΟΓΙΑ ΣΤΑΤΙΣΤΙΚΗΣ.sav**

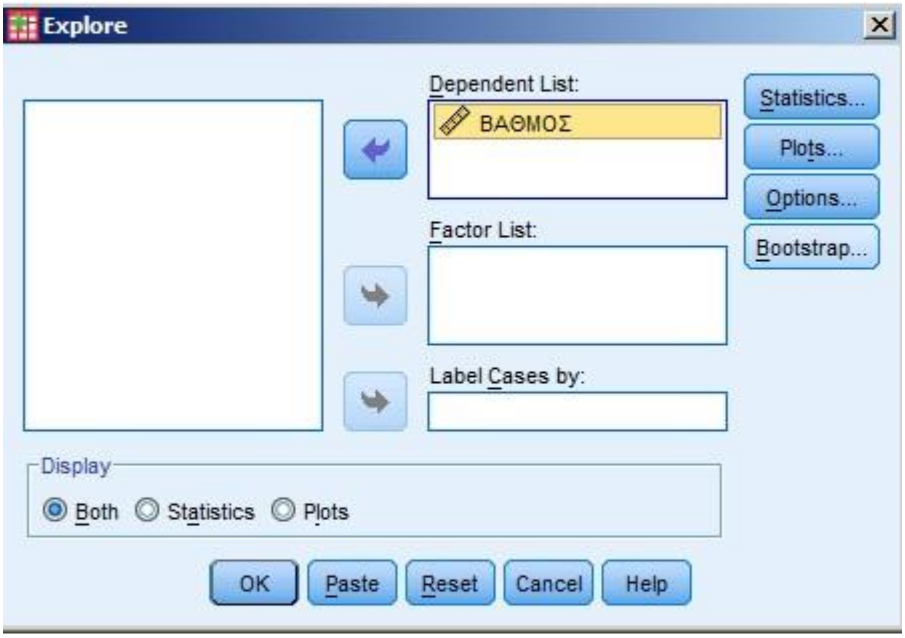
Δίνεται η βαθμολογία 416 φοιτητών στο εργαστηριακό μάθημα «Εισαγωγή στην Στατιστική των επιχειρήσεων». Να υλοποιηθεί περιγραφική ανάλυση της μεταβλητής ΒΑΘΜΟΣ με τη χρήση της διαδικασίας **EXPLORE** (υπολογισμός αριθμητικών στατιστικών μέτρων και γραφικών παραστάσεων).

Τα βήματα που ακολουθούμε είναι:

1. Από το κεντρικό παράθυρο διαλόγου επιλέγουμε:

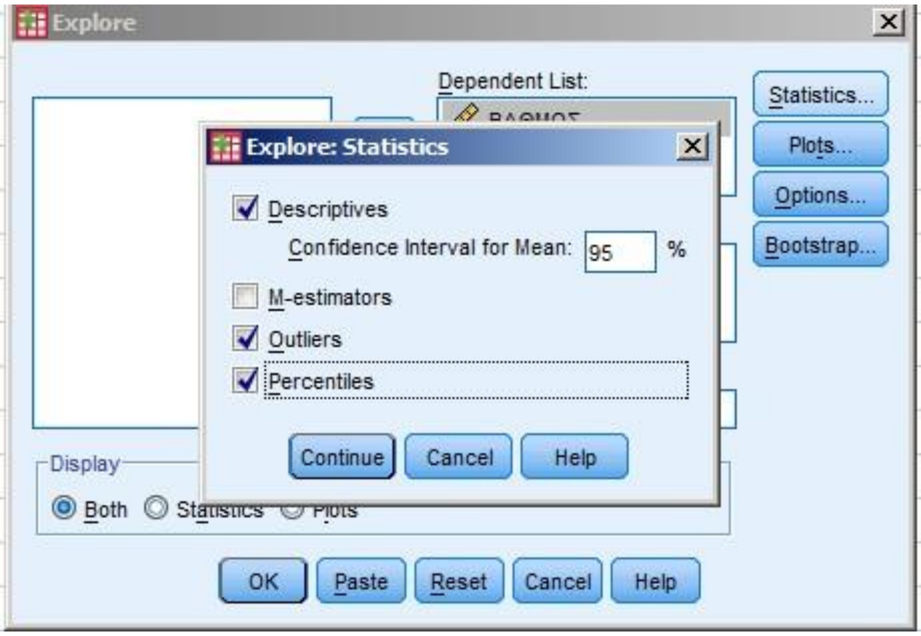


2. Στο παράθυρο διαλόγου που προκύπτει τοποθετούμε στο πλαίσιο Dependent την μεταβλητή ΒΑΘΜΟΣ που θέλουμε να αναλύσουμε.
   (Στο πλαίσιο Dependent τοποθετούμε υποχρεωτικά και μόνο ποσοτικές μεταβλητές).



3. Διατηρώντας την προεπιλογή **Display Both** (στο κάτω αριστερό άκρο του παραθύρου) έχουμε τη δυνατότητα απόκτησης τόσο στατιστικών μέτρων όσο και γραφημάτων. Το πλαίσιο Label Cases By το αφήνουμε ως έχει κενό, έτσι ώστε το S.P.S.S να χρησιμοποιήσει την προεπιλογή του αύξοντα αριθμού παρατήρησης.

4. Από την επιλογή **Statistics** επιλέγουμε τα ακόλουθα:



Descriptives: απόκτηση των κυριότερων περιγραφικών μέτρων, όπως η διάμεσος, η μέση τιμή, η τυπική απόκλιση κ.ά. καθώς και ενός π.χ. 95% διαστήματος εμπιστοσύνης για την πληθυσμιακή μέση τιμή του υπό μελέτη χαρακτηριστικού (που έχει δηλωθεί στο πλαίσιο Dependent List ΒΑΘΜΟΣ). Το διάστημα αυτό υπολογίζεται υπό την υπόθεση της κανονικότητας. Επομένως χρειάζεται προσοχή στην περίπτωση αποκλίσεων από την κανονικότητα.

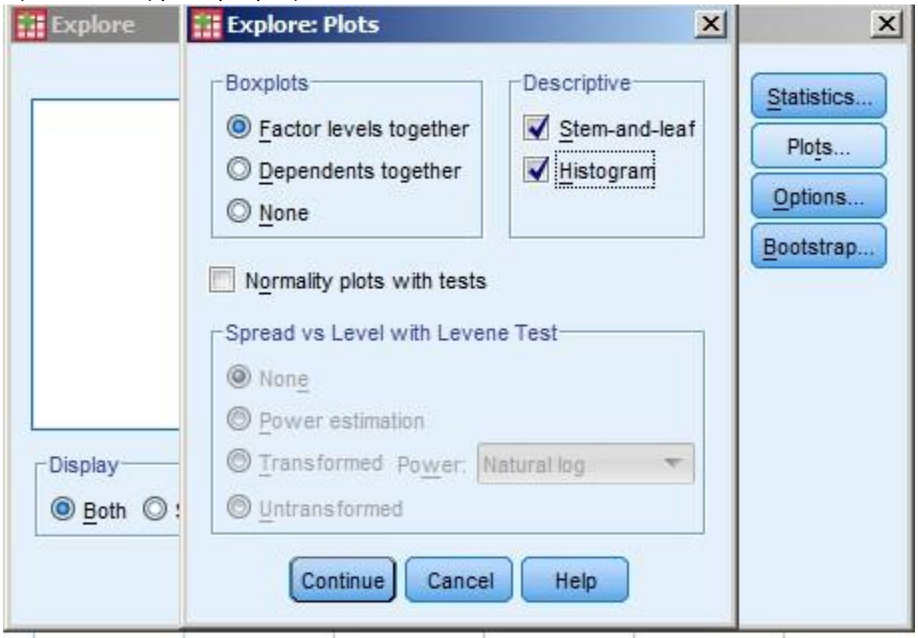
Outliers: το λογισμικό θα μας δώσει τις πέντε μικρότερες και πέντε μεγαλύτερες τιμές της μεταβλητής μας που έχει δηλωθεί στο πλαίσιο Dependent List, ως προς τις κατηγορίες της μεταβλητής που έχει δηλωθεί στο πλαίσιο Factor List.

Percentiles: υπολογίζει το 5° –95° ποσοστιαίο σημείο.

5. Από την επιλογή Plots έχουμε τη δυνατότητα για τα ακόλουθα:

   Boxplots: αποκτούμε το θηκόγραμμα.

   Descriptive: έχουμε διαθέσιμες τις επιλογές Steam-and-Leaf και Histogram, από όπου δηλαδή μπορούμε να αποκτήσουμε το φυλλογράφημα και το ιστόγραμμα για τις ποσοτικές μεταβλητές.



Ερμηνεία αποτελεσμάτων:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
ΒΑΘΜΟΣ	416	100,0%	0	,0%	416	100,0%

Ο πίνακας αυτός μας πληροφορεί ότι το δείγμα αποτελείται από 416 φοιτητές, χωρίς να υπάρχουν ελλιπείς τιμές.

Descriptives

			Statistic	Std. Error
ΒΑΘΜΟΣ	Mean		5,3743	,12642
	95% Confidence Interval for Mean	Lower Bound	5,1258	
		Upper Bound	5,6228	
	5% Trimmed Mean		5,3627	
	Median		5,0000	
	Variance		6,648	
	Std. Deviation		2,57841	
	Minimum		1,00	
	Maximum		10,00	
	Range		9,00	
	Interquartile Range		4,50	
	Skewness		-,006	,120
	Kurtosis		-,952	,239

Στον πίνακα Descriptives μας δίνονται διάφορα περιγραφικά μέτρα (και όχι μόνο) για τη μεταβλητή ΒΑΘΜΟΣ.

Χρήζουν ιδιαίτερης προσοχής και σχολιασμού τα ακόλουθα:

- Η μέση τιμή (Mean) της βαθμολογίας στο εργαστηριακό μάθημα της στατιστικής είναι 5,3743.
- Το λογισμικό μας δίνει το 95% διάστημα εμπιστοσύνης (95% Confidence Interval for Mean, Lower and Upper Bound) για το μέσο. Για τα συγκεκριμένα δεδομένα το 95% διάστημα εμπιστοσύνης για την μέση βαθμολογία είναι (5,1258, 5,6228).

- Η διάμεσος (Median) είναι 5. Αυτό σημαίνει ότι το 50% των φοιτητών έγραψαν κάτω από 5 και το υπόλοιπο 50% έγραψε πάνω από 5. Παρατηρούμε ότι ο μέσος βαθμός είναι περίπου ίσος με τη διάμεσο (median), επομένως τα δεδομένα μπορούν να θεωρηθούν ότι προέρχονται από συμμετρικό πληθυσμό.
- Η διασπορά σ^2 (Variance) της βαθμολογίας είναι 6,648, και η τυπική απόκλιση σ (Std. Deviation) είναι 2,57841.
- Η ελάχιστη βαθμολογία (Minimum) είναι 1, η μέγιστη (Maximum) είναι 10, και το εύρος μεταβολής (Range) είναι 9.

Επιπλέον, στον πίνακα Percentiles εμφανίζονται τα ποσοστιαία σημεία, ενώ στη στήλη Extreme Values οι χρόνοι των 5 πιο χαμηλών και πιο υψηλών βαθμολογιών.

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average(Definition 1)	ΒΑΘΜΟΣ	1,0000	2,0000	3,0000	5,0000	7,5000	9,0000	9,5000
Tukey's Hinges	ΒΑΘΜΟΣ			3,0000	5,0000	7,5000		

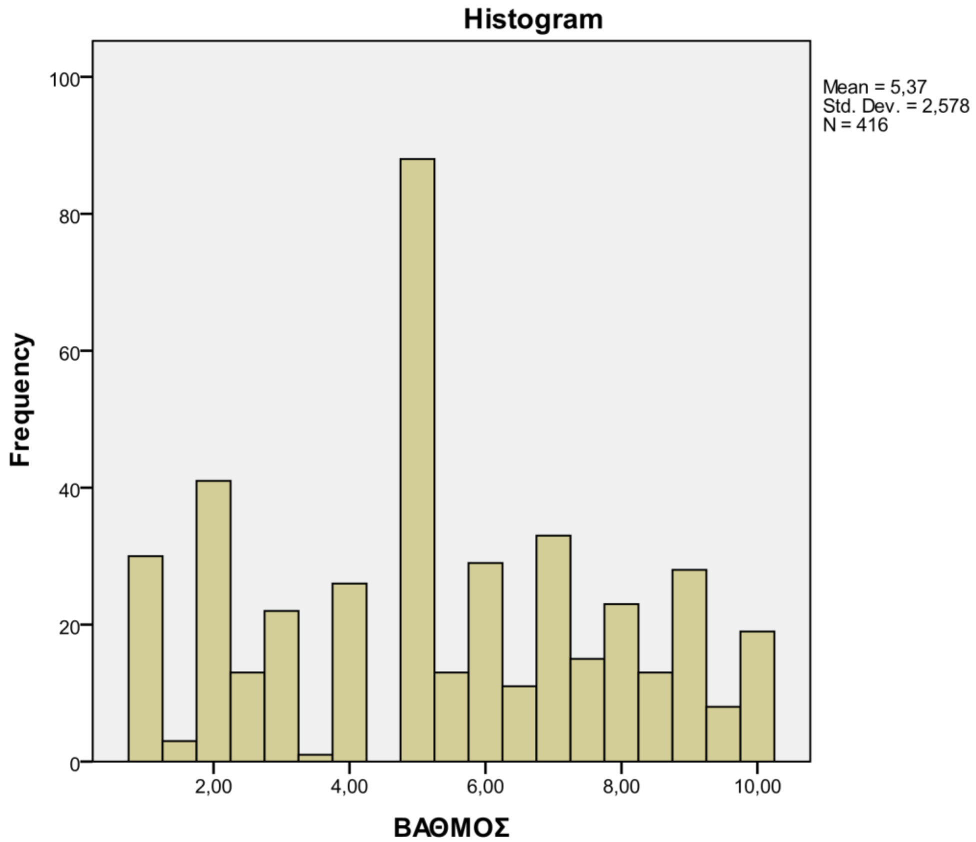
Extreme Values

			Case Number	Value
ΒΑΘΜΟΣ	Highest	1	8	10,00
		2	17	10,00
		3	33	10,00
		4	35	10,00
		5	81	10,00[a]
	Lowest	1	403	1,00
		2	383	1,00

3	380	1,00
4	377	1,00
5	371	1,00[b]

a. Only a partial list of cases with the value 10,00 are shown in the table of upper extremes.
b. Only a partial list of cases with the value 1,00 are shown in the table of lower extremes.

Επιπλέον έχουμε το ιστόγραμμα το φυλλογράφημα και το θηκόγραμμα της μεταβλητής βαθμός.



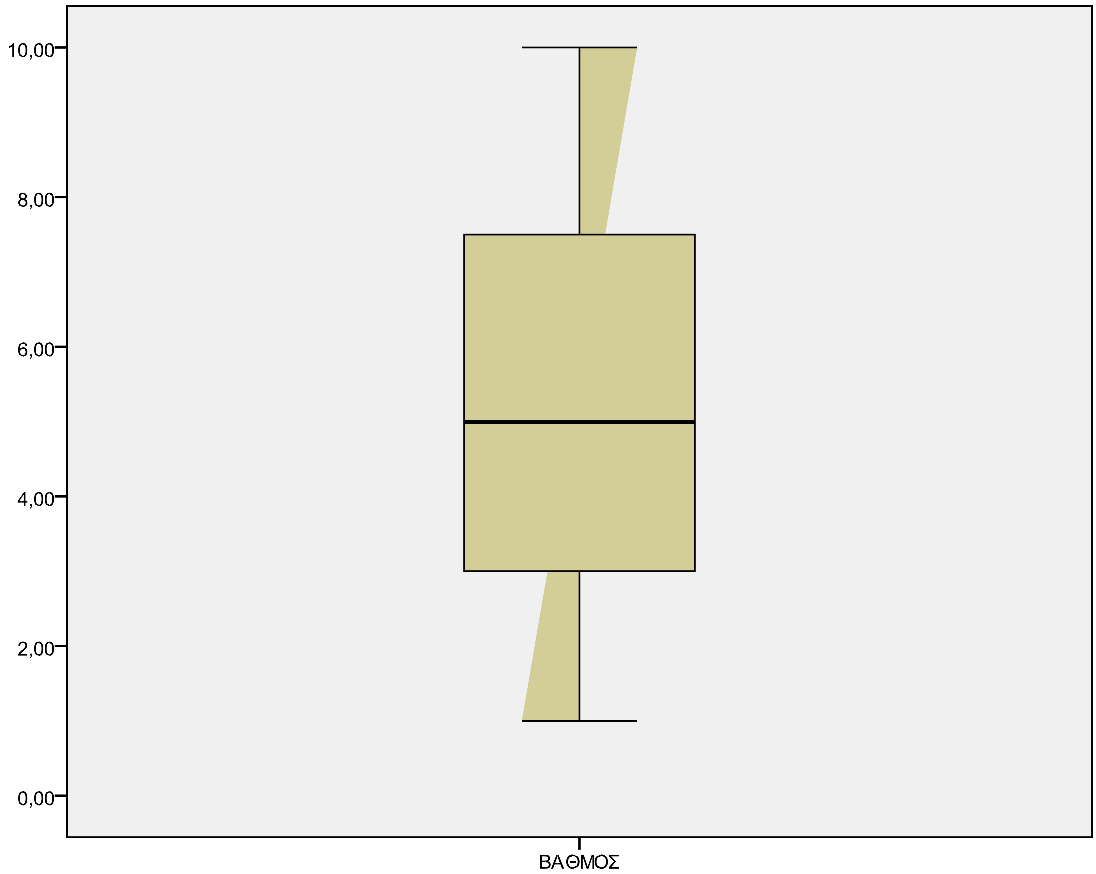
ΒΑΘΜΟΣ Stem-and-Leaf Plot

```
 Frequency    Stem &  Leaf

    30,00        1 .  000000000000000000000000000000
     3,00        1 .  555
    41,00        2 .  00000000000000000000000000000000000000000
    13,00        2 .  5555555555556
    22,00        3 .  0000000000000000000000
     1,00        3 .  5
    26,00        4 .  00000000000000000000000000
      ,00        4 .
    88,00        5 .
000000000000000000000000000000000000000000000000000000000000000000000000000
0000000000000
    13,00        5 .  5555555555555
    29,00        6 .  00000000000000000000000000000
    11,00        6 .  55555555555
    33,00        7 .  000000000000000000000000000000000
    15,00        7 .  555555555555556
    23,00        8 .  00000000000000000000000
    13,00        8 .  5555555555555
    28,00        9 .  0000000000000000000000000000
     8,00        9 .  55555555
    19,00       10 .  0000000000000000000

 Stem width:      1,00
```

Each leaf: 1 case(s)



Άσκηση

Σε ένα τυχαίο δείγμα 27 φοιτητών καταγράψαμε τον χρόνο που κάνουν για να φτάσουν στο Τ.Ε.Ι.. Με βάση τους χρόνους που δίνονται παρακάτω, να εφαρμοστεί η διαδικασία Explore και να αναφερθούν τα συμπεράσματα που προκύπτουν από αυτήν.

Α/Α	ΧΡΟΝΟΣ ΣΕ ΛΕΠΤΑ
1	45
2	28
3	33
4	18
5	65
6	24

7	31
8	60
9	15
10	18
11	36
12	38
13	42
14	29
15	44
16	37
17	45
18	27
19	20
20	19
21	9
22	68
23	32
24	35
25	38
26	40
27	41

Απάντηση :
Από τη διαδικασία Explore συμπεραίνουμε ότι υπάρχει κανονικότητα γεγονός που επιβεβαιώνεται από τα αντίστοιχα διαγράμματα. Στους πίνακες Descriptives και Extreme Values φαίνονται όλα τα στοιχεία που μπορούμε να συλλέξουμε από τη συγκεκριμένη διαδικασία.

2. ΕΛΕΓΧΟΣ ΚΑΝΟΝΙΚΟΤΗΤΑΣ – ΔΙΑΣΤΗΜΑ ΕΜΠΙΣΤΟΣΥΝΗΣ ΓΙΑ ΤΟΝ ΜΕΣΟ

Η διαδικασία Explore στο SPSS μας δίνει επιπλέον απαντήσεις στα ερωτήματα:

- Τα δεδομένα μας, δηλαδή οι παρατηρήσεις του δείγματος, προέρχονται από πληθυσμό που ακολουθεί την κανονική κατανομή;
- Μπορούμε να προσδιορίσουμε ένα διάστημα μέσα στο οποίο περιμένουμε να βρίσκεται η άγνωστη παράμετρος μ (μέσος) του πληθυσμού με μια

προκαθορισμένη πιθανότητα (συνήθως 0.95, 0.99 ή 0.90);

Τα βήματα που ακολουθούμε για την διαδικασία αυτή στο SPSS είναι τα παρακάτω :

1. Δημιουργούμε την μεταβλητή μας στο Variable View και εισάγουμε τα δεδομένα στο Data View.
2. Επιλέγουμε :

   Analyze → Descriptive Statistics → Explore
3. Στο Dependent List μεταφέρουμε την μεταβλητή μας.
4. Επιλέγουμε Statistics και κλικάρουμε Descriptives (στο Confidence Interval for Mean γράφουμε το διάστημα εμπιστοσύνης που μας ζητάνε π.χ. 95% ή 99% ή 90%), Πατάμε Continue.
5. Επιλέγουμε Plots και κλικάρουμε την επιλογή Normality plots with tests. Πατάμε Continue.

Αποτελέσματα στο Output :

● Πίνακας Tests of Normality

Ο πίνακας αυτός μας δίνει τα αποτελέσματα του ελέγχου κανονικότητας. Αν το δείγμα μας είναι μεγέθους ν>50 τότε μας ενδιαφέρει το κριτήριο Kolmogorov-Smirnov ενώ αν το δείγμα μας είναι μεγέθους ν≤50 κοιτάζουμε τα αριθμητικά αποτελέσματα του κριτηρίου Shapiro-Wilk. Σε κάθε κριτήριο, ο αριθμός στη στήλη sig (στα επόμενα για λόγους συντομίας θα αναφέρεται απλά ως sig) είναι το αριθμητικό μέτρο σύγκρισης με το επίπεδο στατιστικής σημαντικότητας (significance level) που μας ενδιαφέρει. Το επίπεδο στατιστικής σημαντικότητας συμβολίζεται συνήθως με α και σε περίπτωση που δεν προσδιορίζεται θεωρούμε ότι α= 5%=0,05.

- Αν sig>α, τότε μπορούμε να θεωρήσουμε ότι τα δεδομένα μας προέρχονται από πληθυσμό που ακολουθεί την κανονική κατανομή.
- Αν sig<α, τότε δεν είμαστε σίγουροι για το εάν είναι δυνατό να υποθέσουμε ότι τα δεδομένα μας προέρχονται από πληθυσμό που ακολουθεί την κανονική κατανομή.

● Πίνακας Descriptives

Από τον πίνακα αυτό χρειαζόμαστε το πεδίο α% confidence interval for mean (όπου α% είναι το επίπεδο σημαντικότητας που επιλέξαμε) δηλαδή το διάστημα μέσα στο οποίο είμαστε κατά α% σίγουροι ότι βρίσκεται ο μέσος

Εφαρμογή της διαδικασίας

Στο παράδειγμα που ακολουθεί περιγράφεται αναλυτικά η διαδικασία Explore για τον έλεγχο κανονικότητας και την εύρεση διαστήματος εμπιστοσύνης.

Οι υπεύθυνοι μιας αλυσίδας fast food ισχυρίζονται ότι ο μέσος χρόνος αναμονής των πελατών τους είναι 3 λεπτά. Προκειμένου το τμήμα ποιοτικού ελέγχου της επιχείρησης να πιστοποιήσει τον ισχυρισμό, παίρνει τυχαία δείγμα 50 πελατών και σημειώνει τον χρόνο αναμονής κάθε πελάτη. Οι παρατηρήσεις είναι οι ακόλουθες :

4,56	3,02	5,07	3,49	2,36	2,95	3,98	3,74	2,64	3,93	2,02
3,09	1,40	1,23	3,03	3,09	3,19	3,17	3,07	2,06	3,13	3,69
3,13	3,21	2,28	1,80	4,17	2,18	2,98	3,04	2,78	2,82	1,25
3,50	2,34	4,52	1,61	3,28	1,96	2,51	1,01	0,96	1,44	2,18
1,73	2,14	3,24	1,39	3,18	2,64					

α) Θα μπορούσατε να δεχθείτε ότι ο χρόνος αναμονής των πελατών είναι κανονικά κατανεμημένος;

β) Να βρεθεί ένα 95% διάστημα εμπιστοσύνης για την πραγματική μέση τιμή του χρόνου αναμονής των πελατών.

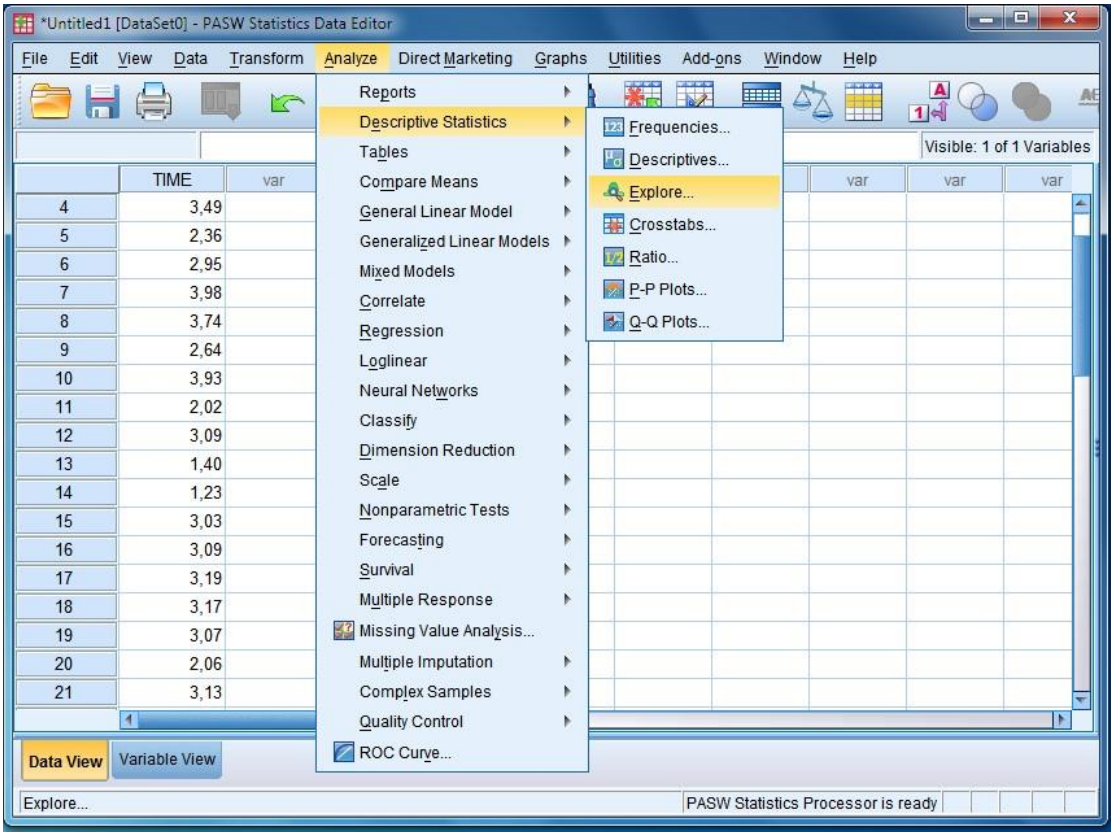
Τα βήματα που ακολουθούμε είναι:

1. Δημιουργούμε την μεταβλητή TIME και εισάγουμε τα δεδομένα.
2. Γράφουμε τον έλεγχο υπόθεσης:

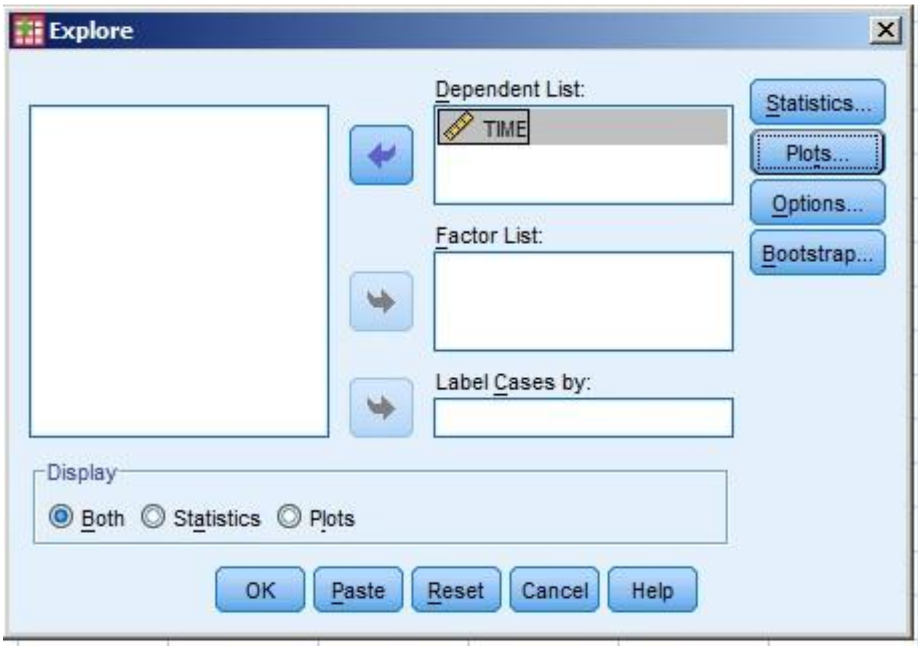
   H_O : Τα δεδομένα ακολουθούν την κανονική κατανομή
   (ή το δείγμα μας προέρχεται από κανονικά κατανεμημένο πληθυσμό)

   H_1 : Τα δεδομένα δεν ακολουθούν την κανονική κατανομή
   (ή το δείγμα μας δεν προέρχεται από κανονικά κατανεμημένο πληθυσμό)

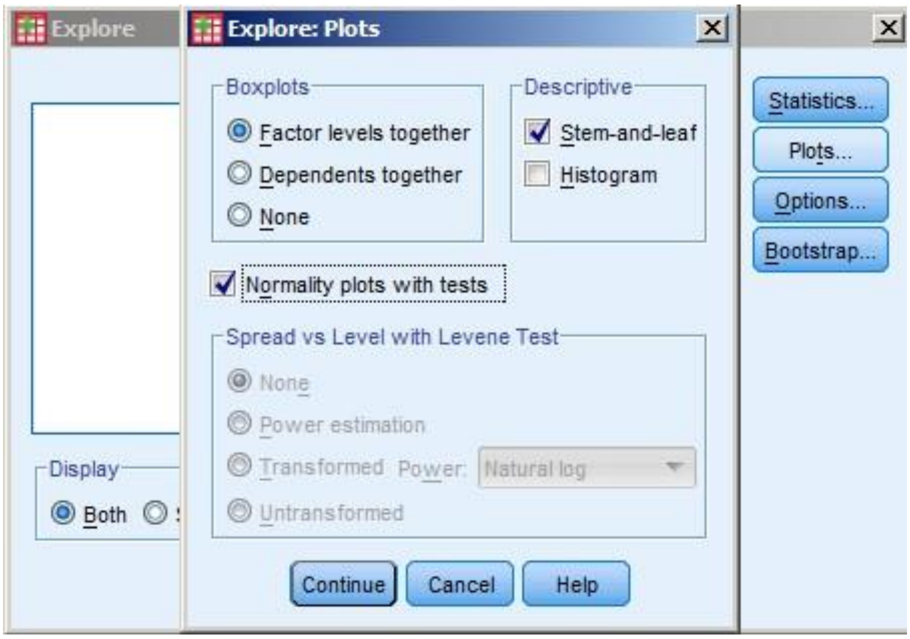
3. Από το κεντρικό παράθυρο διαλόγου επιλέγουμε:



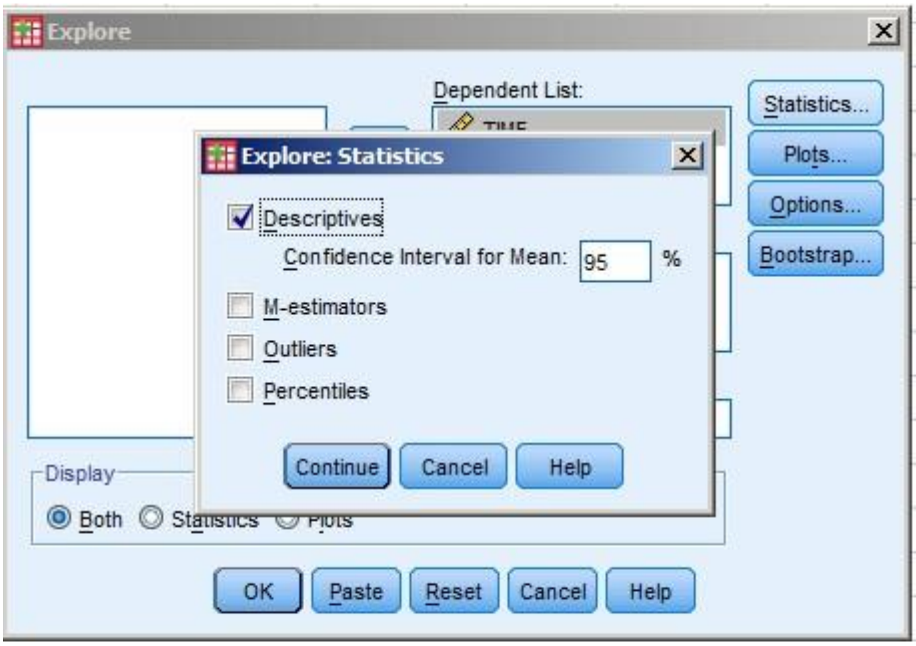
4. Στο παράθυρο διαλόγου που προκύπτει τοποθετούμε στο πλαίσιο Dependent την μεταβλητή TIME.
(Στο πλαίσιο Dependent τοποθετούμε υποχρεωτικά και μόνο ποσοτικές μεταβλητές).



5. Από την επιλογή **Plots** επιλέγουμε **Normality Plots with tests** προκειμένου να προκύψουν τα κατάλληλα test και διαγράμματα για τον έλεγχο κανονικότητας, και μετά **Continue**.



6. Από την επιλογή **Statistics** επιλέγουμε **Descriptives** και στο παράθυρο **Confidence Interval For Mean** δίνουμε την τιμή ανάλογα με το διάστημα εμπιστοσύνης που μας έχει ζητηθεί να υπολογίσουμε. Στη συνέχεια πατάμε **Continue – OK.**



Ερμηνεία αποτελεσμάτων:

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
TIME	,106	50	,200*	,977	50	,450

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

Στον πίνακα **Tests Of Normality** περιέχονται δύο κριτήρια που εξετάζουν τον έλεγχο κανονικότητας.

<u>Παρατήρηση:</u> Επειδή έχουμε δείγμα μεγέθους 50 (οριακή τιμή για επιλογή κριτηρίου) θα ελέγξουμε και τις δύο τιμές.

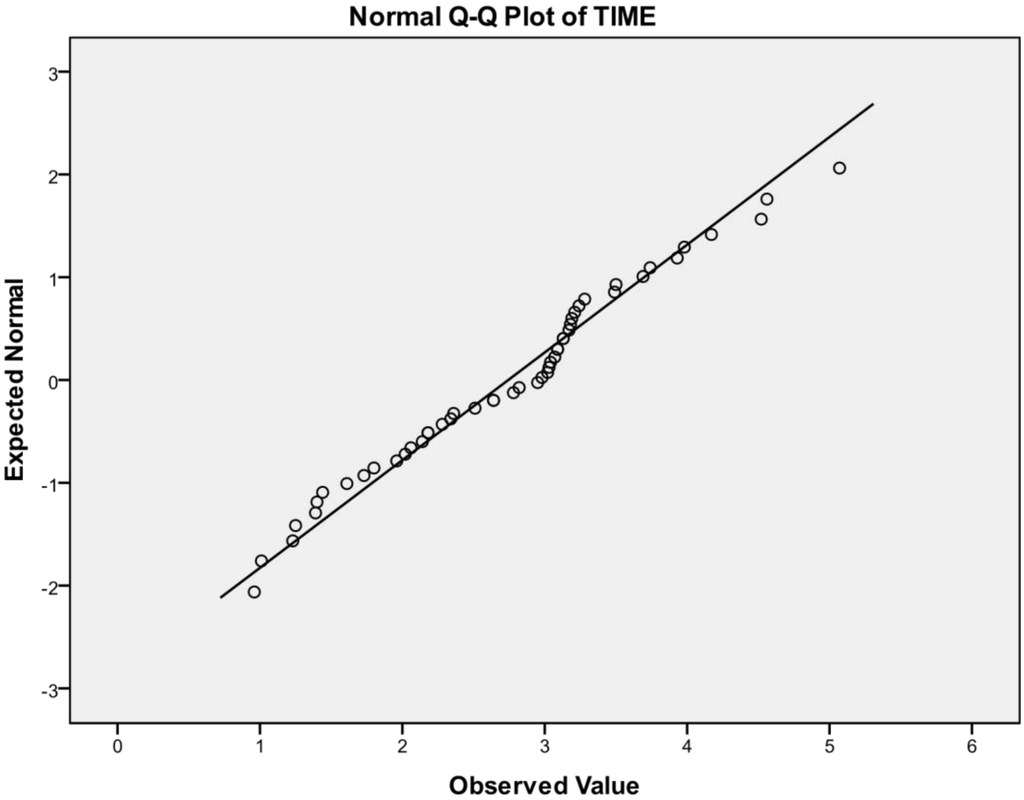
Το πρώτο κριτήριο (Kolmogorov-Smirnov) δίνει sig. = 0,200. Από την τιμή του sig. θα αποφασίσουμε αν θα απορρίψουμε ή θα δεχθούμε την μηδενική υπόθεση της κανονικότητας των δεδομένων. Στην περίπτωσή μας έχουμε 0,200 > 0,05. Άρα αποδεχόμαστε την H_0.

Το δεύτερο κριτήριο (Shapiro-Wilk) μας δίνει sig. = 0,450 > 0,05. Άρα και πάλι αποδεχόμαστε την H_0.

Δηλαδή σύμφωνα και με τα δύο κριτήρια καταλήγουμε ότι ο χρόνος αναμονής των πελατών είναι κανονικά κατανεμημένος.

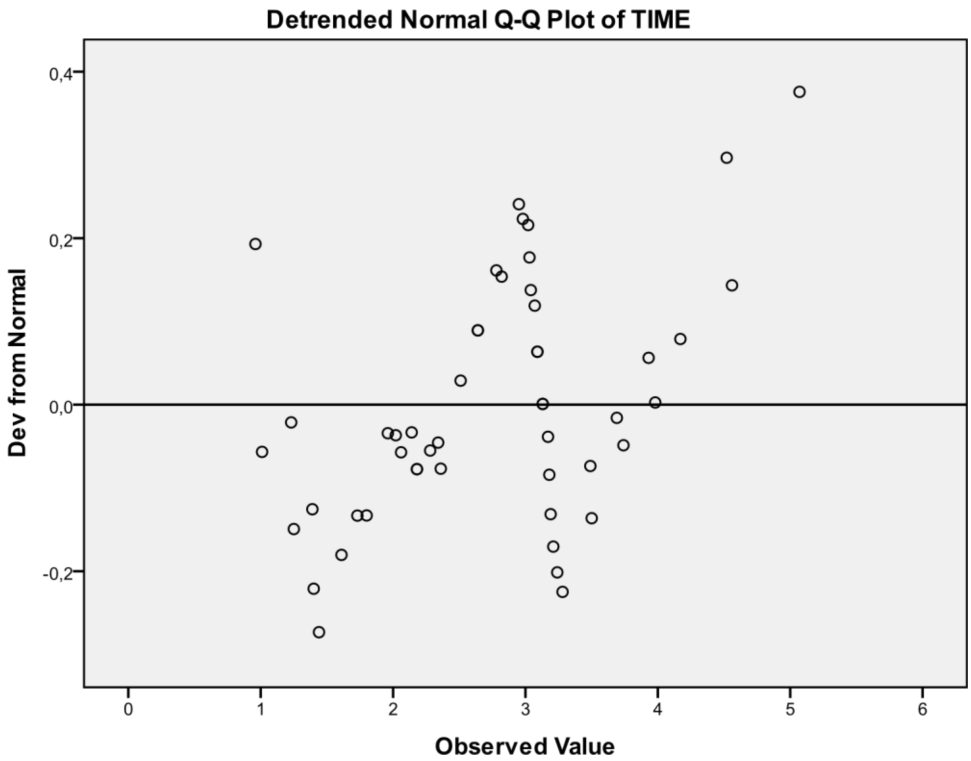
Σ' αυτό το συμπέρασμα μπορούμε επίσης να καταλήξουμε βλέποντας και ερμηνεύοντας τα διάγραμμα που σχετίζονται με τον έλεγχο κανονικότητας:

- Normal Q – Q Plot
- Detrended Normal Q-Q Plot.



Το **NORMAL Q – Q PLOT** μας δείχνει τις πραγματικές τιμές (observed values) και τις αναμενόμενες τιμές (expected values) αν τα δεδομένα ήταν δείγμα από κανονική κατανομή. Για να μπορούμε να συμπεράνουμε ότι το δείγμα ακολουθεί την κανονική κατανομή, τα σημεία του διαγράμματος θα πρέπει να είναι συγκεντρωμένα γύρω από την ευθεία γραμμή (γεγονός που ισχύει για τα δεδομένα της άσκησης).

Το επόμενο διάγραμμα **DETRENDED NORMAL Q – Q PLOT** μας δείχνει **την διαφορά** μεταξύ των αναμενόμενων και πραγματικών τιμών. Θα πρέπει οι διαφορές να τείνουν στο 0 αν τα δεδομένα μας ακολουθούν την κανονική κατανομή.



Στον πίνακα που ακολουθεί βλέπουμε τα βασικά στατιστικά μέτρα που έχουν υπολογισθεί. Το 95% διάστημα εμπιστοσύνης για την πραγματική μέση τιμή του χρόνου αναμονής των πελατών είναι (Lower Bound, Upper Bound) = (2,4724, 3,0148). Δηλ. κατά 95% είμαστε σίγουροι ότι ο μέσος χρόνος αναμονής των πελατών βρίσκεται μέσα σ’ αυτό το διάστημα.

Descriptives

			Statistic	Std. Error
TIME	Mean		2,7436	,13497
	95% Confidence Interval for Mean	Lower Bound	2,4724	
		Upper Bound	3,0148	
	5% Trimmed Mean		2,7268	
	Median		2,9650	
	Variance		,911	

Std. Deviation	,95436	
Minimum	,96	
Maximum	5,07	
Range	4,11	
Interquartile Range	1,17	
Skewness	,102	,337
Kurtosis	-,270	,662

Άσκηση

Μια εταιρεία που παράγει μπαταρίες για μικρές υπολογιστικές μηχανές θέλει να δει ποια είναι η μέση διάρκεια ζωής τους. Για το λόγο αυτό πήρε τυχαίο δείγμα από 22 μπαταρίες και κατέγραψε την διάρκεια ζωής κάθε μίας. Στον παρακάτω πίνακα φαίνονται οι χρόνοι αυτοί. Να βρεθεί το 95% διάστημα εμπιστοσύνης για τον μέσο του δείγματος, να υπολογιστούν τα αριθμητικά στατιστικά μέτρα και να ελεγχθεί αν οι χρόνοι ακολουθούν την κανονική κατανομή με βάση το συγκεκριμένο δείγμα.

Α/Α	ΔΙΑΡΚΕΙΑ ΖΩΗΣ ΣΕ ΩΡΕΣ
1	25,3
2	22,4
3	19,2
4	17
5	18,4
6	29,8
7	11,25
8	32
9	29
10	25,6
11	16,5
12	17,9
13	10,9
14	21
15	23,9
16	27,5
17	30
18	31,3
19	26,1

20	9,9
21	35
22	18

Ενδεικτική Απάντηση :
Από τη διαδικασία Explore προκύπτει ότι οι χρόνοι ακολουθούν την κανονική κατανομή. Το 95% διάστημα εμπιστοσύνης για τον μέσο είναι το [19.4395,25.8287].

3. ΕΛΕΓΧΟΣ ΥΠΟΘΕΣΕΩΝ ΓΙΑ ΤΗ ΜΕΣΗ ΤΙΜΗ ΕΝΟΣ ΔΕΙΓΜΑΤΟΣ (One Sample t-test)

Το κριτήριο One sample t-test χρησιμοποιείται όταν θέλουμε να συγκρίνουμε τον αριθμητικό μέσο μ με μία συγκεκριμένη δοσμένη τιμή μ_0.

Είδη ελέγχου

$H_0 : \mu=\mu_0$ $\quad H_1 : \mu\neq\mu_0$ (αμφίπλευρος έλεγχος)

$H_0 : \mu=\mu_0$ $\quad H_1 : \mu>\mu_0$ (μονόπλευρος έλεγχος)

$H_0 : \mu=\mu_0$ $\quad H_1 : \mu<\mu_0$ (μονόπλευρος έλεγχος)

Για να μπορούμε να χρησιμοποιήσουμε το κριτήριο, πρέπει να ισχύουν τα παρακάτω :

- το δείγμα μας θα πρέπει να έχει επιλεγεί τυχαία από τον πληθυσμό
- το δείγμα μας θα πρέπει να προέρχεται από κανονικά κατανεμημένο πληθυσμό. (Για τον λόγο αυτό πριν χρησιμοποιήσουμε το κριτήριο One sample t-test πρέπει να κάνουμε πρώτα έλεγχο κανονικότητας)
- να γνωρίζουμε επίπεδο σημαντικότητας α που μας ενδιαφέρει

Τα βήματα που ακολουθούμε για την διαδικασία αυτή στο SPSS είναι τα παρακάτω :

1. Δημιουργούμε την μεταβλητή μας στο Variable View και εισάγουμε τα δεδομένα στο Data View.

2. Επιλέγουμε :

   Analyze → Compare Means → One sample t-test

3. Στο Test variable μεταφέρουμε την μεταβλητή μας.
4. Στο παράθυρο Test Value γράφουμε την τιμή μ_0 με την οποία θέλουμε να συγκρίνουμε τον μέσο και πατάμε OK.

Αποτελέσματα στο Output :

● Πίνακας One-Sample Statistics
Από αυτόν τον πίνακα μας ενδιαφέρει μόνο το mean δηλαδή ο μέσος του δείγματος.

● Πίνακας One-Sample Test
Από αυτόν τον πίνακα μας ενδιαφέρει μόνο ο αριθμός sig.

Συμπέρασμα:

● Αμφίπλευρος έλεγχος

- αν sig>α τότε αποδεχόμαστε την υπόθεση H_0
- αν sig<α τότε απορρίπτουμε την υπόθεση H_0

● Μονόπλευρος έλεγχος

Αν ο mean ικανοποιεί την ανισότητα της H_1 τότε ισχύουν τα εξής :

- αν $\frac{sig}{2} > \alpha$ τότε αποδεχόμαστε την υπόθεση H_0
- αν $\frac{sig}{2} < \alpha$ τότε απορρίπτουμε την υπόθεση H_0

Αν ο mean δεν ικανοποιεί την ανισότητα της H_1 τότε ισχύουν τα εξής:

- αν $1 - \frac{sig}{2} > \alpha$ τότε αποδεχόμαστε την υπόθεση H_0
- αν $1 - \frac{sig}{2} < \alpha$ τότε απορρίπτουμε την υπόθεση H_0

Εφαρμογή της διαδικασίας One Sample t-test για αμφίπλευρο έλεγχο

Ο ιδιοκτήτης ενός ορυχείου ενδιαφέρεται να αξιολογήσει μια νέα μέθοδο παραγωγής συνθετικών διαμαντιών. Η μελέτη του κόστους που συνεπάγεται η διαδικασία κατασκευής έχει οδηγήσει στο συμπέρασμα ότι για να είναι επικερδής η νέα αυτή μέθοδος θα πρέπει το μέσο βάρος των συνθετικών διαμαντιών να είναι γύρω στα 0,5 καράτια. Προκειμένου να αξιολογηθεί η διαδικασία κατασκευής επιλέγεται δείγμα από 6 συνθετικά διαμάντια που έχουν κατασκευασθεί με την νέα μέθοδο παρασκευής.

Το βάρος τους βρίσκεται ότι είναι: 0,46 0,61 0,52 0,48 0,57 0,54 καράτια αντίστοιχα.

Να καθορισθεί σε ε.σ. 5% με βάση τις πληροφορίες από το δείγμα αν η νέα μέθοδος είναι επικερδής ή όχι.

Τα βήματα που ακολουθούμε είναι:

1. Δημιουργούμε την μεταβλητή **BAROS_KARATIA** και εισάγουμε τα δεδομένα. Ο έλεγχος υπόθεσης που πρέπει να γίνει είναι:

   $\mathbf{H_0 : \mu = 0,5 \quad H_1 : \mu \neq 0,5}$

Πριν την διενέργεια του παραπάνω ελέγχου θα πρέπει να διενεργηθεί έλεγχος κανονικότητας καθώς μία βασική προϋπόθεση εφαρμογής του t-test είναι ότι η κατανομή της ποσοτικής μεταβλητής πρέπει να είναι στοιχειωδώς κανονική.

<u>Παρατήρηση:</u> Η εκτροπή από την κανονικότητα_δεν δημιουργεί πρόβλημα κατά τον έλεγχο εφόσον βέβαια η κατανομή της ποσοτικής μεταβλητής δεν είναι εντελώς ασύμμετρη.

2. Διενέργεια ελέγχου κανονικότητας με τη διαδικασία **Explore**

Tests of Normality

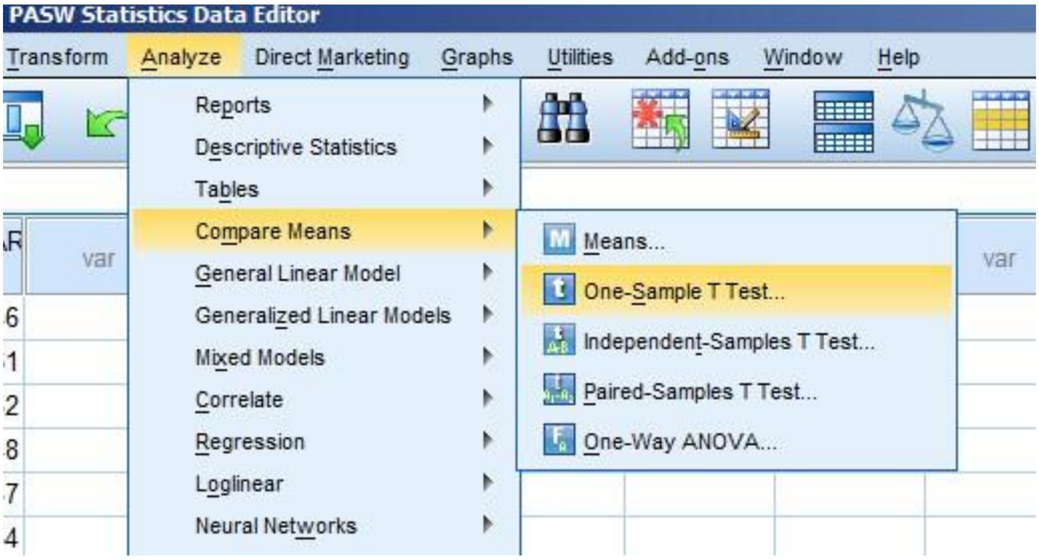
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
BAROS_KARATIA	,148	6	,200*	,977	6	,937

a. Lilliefors Significance Correction

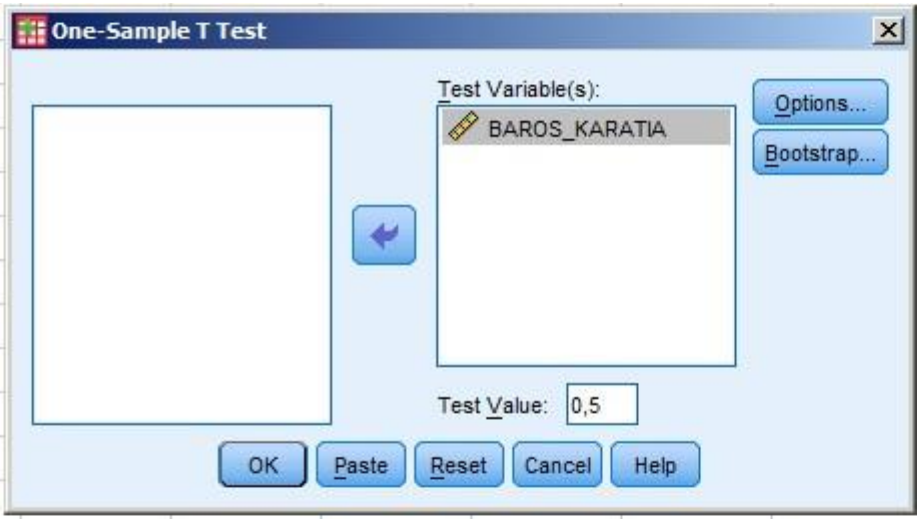
*. This is a lower bound of the true significance.

Άρα, σύμφωνα με το κριτήριο Shapiro-Wilk καταλήγουμε ότι δείγμα μας προέρχεται από κανονικά κατανεμημένο πληθυσμό.

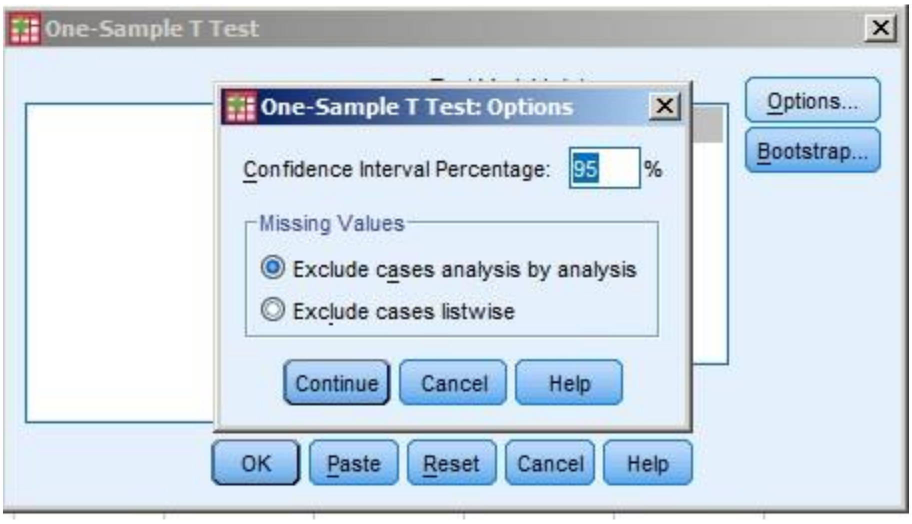
3. Από το κεντρικό παράθυρο διαλόγου επιλέγουμε:



4. Επιλέγουμε από το παράθυρο αριστερά την μεταβλητή BAROS_KARATIA και με πάτημα στο μαύρο βέλος αυτή μεταφέρεται στο παράθυρο **Test Variable(s)**. Στο παράθυρο **Test Value** βάζουμε **<u>την τιμή 0,5 με την οποία θα συγκρίνουμε τη μέση τιμή</u>**.



5. Πατάμε την επιλογή **Options** και στο παράθυρο **Confidence Interval Percentage** γράφουμε το επίπεδο εμπιστοσύνης με το οποίο θέλουμε να γίνει ο έλεγχος (έστω 95%) και στη συνέχεια CONTINUE - OK και εμφανίζεται το OUTPUT.



Ερμηνεία αποτελεσμάτων:

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
BAROS_KARATIA	6	,5300	,05586	,02280

Ο πίνακας **One-Sample Statistics** μας δίνει:

- Το πλήθος των παρατηρήσεων του δείγματος (**N** = 6)
- Τον αριθμητικό μέσο των παρατηρήσεων του δείγματος (**Mean** μ = 0,53)
- Την τυπική απόκλιση των παρατηρήσεων του δείγματος (**Std. Deviation**)
- Το τυπικό σφάλμα του αριθμητικού μέσου του δείγματος (**St. Error Mean**)

One-Sample Test

	Test Value = 0.5					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
BAROS_KARATIA	1,316	5	,245	,03000	-,0286	,0886

Ο πίνακας **One-Sample Test** μας δίνει :

- Την τιμή του t – test (**t** = 1,316)
- Το sig. του t – test (**Sig.** = 0,245)
- Την διαφορά της μέσης τιμής της μεταβλητής που ελέγχεται και της αριθμητικής τιμής που έχουμε ορίσει (**Mean Difference**=,03000)
- Το 95% διάστημα εμπιστοσύνης της διαφοράς της μέσης τιμής της μεταβλητής που ελέγχεται και της αριθμητικής τιμής που έχουμε ορίσει 95% (**Confidence Interval of the Difference** (-,0286, 0,0886)). Μπορούμε να ζητήσουμε τον υπολογισμό οποιουδήποτε άλλου διαστήματος εμπιστοσύνης, εισάγοντας μία τιμή από το 1 έως το 99 στο πεδίο Confidence Interval (συνήθως εισάγουμε 90,95 ή 99).

Από την τιμή του sig. θα αποφασίσουμε αν θα απορρίψουμε ή θα δεχθούμε την μηδενική υπόθεση. Αν ο αριθμός αυτός είναι μικρότερος από το 0,05 τότε απορρίπτουμε την μηδενική υπόθεση, ενώ αν είναι μεγαλύτερος από το 0,05 αποδεχόμαστε την μηδενική υπόθεση.

Επομένως, αφού στην περίπτωσή μας έχουμε sig. = 0,245 > 0,05 **<u>αποδεχόμαστε την H_0, δηλ. το μέσο βάρος των συνθετικών διαμαντιών του δείγματος δεν διαφέρει σημαντικά από το 0,5.</u>**

<u>Σημείωση:</u> Αν ο έλεγχος γίνεται με **Confidence Interval 90%** η σύγκριση του sig. γίνεται με το **0.1,** και όταν ο έλεγχος γίνεται με **Confidence Interval 99%** η σύγκριση του sig. γίνεται με το **0.01.**

Εφαρμογή της διαδικασίας One Sample t-test για μονόπλευρο έλεγχο

Ένα μεσιτικό γραφείο που ειδικεύεται στις πωλήσεις οικοπέδων έχει παρατηρήσει ότι κατά μέσο όρο τα οικόπεδα πωλούνται σε 90 ημέρες από την στιγμή που θα περάσουν στη δικαιοδοσία του. Τελευταία έχει δημιουργηθεί η εντύπωση ότι τα οικόπεδα «παραμένουν» περισσότερο καιρό στο γραφείο.

Για να ελεγχθεί αν συμβαίνει κάτι τέτοιο παίρνουν ένα τυχαίο δείγμα 20 πρόσφατα πουλημένων οικοπέδων. Οι μέρες μετά τις οποίες πουλήθηκαν αυτά ήταν:

98	62	99	59	83
133	99	109	93	107

91	97	99	111	134
138	125	87	94	107

Αληθεύει ο παραπάνω ισχυρισμός σε ε.σ. 5%;

Τα βήματα που ακολουθούμε είναι:

1. Δημιουργούμε την μεταβλητή **DAYS** και εισάγουμε τα δεδομένα. Ο έλεγχος υπόθεσης που πρέπει να γίνει είναι:

   $H_0 : \mu = 90 \quad H_1 : \mu > 90$

2. Διενέργεια ελέγχου κανονικότητας με τη διαδικασία **Explore**

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
DAYS	,143	20	,200*	,946	20	,305

a. Lilliefors Significance Correction *. This is a lower bound of the true significance.

Άρα, σύμφωνα με το κριτήριο Shapiro-Wilk καταλήγουμε ότι δείγμα μας προέρχεται από κανονικά κατανεμημένο πληθυσμό.

3. Από το κεντρικό παράθυρο διαλόγου επιλέγουμε

   Analyze – Compare Means – One Sample T Test.

   Επιλέγουμε από το παράθυρο αριστερά την μεταβλητή **DAYS** και με πάτημα στο μαύρο βέλος αυτή μεταφέρεται στο παράθυρο **Test Variable(s)**.

   Στο παράθυρο **Test Value** βάζουμε την **τιμή 90** με την οποία θα συγκρίνουμε τη μέση τιμή.

   Στην επιλογή **Options** και στο παράθυρο **Confidence Interval Percentage** αφήνουμε την επιλογή 95% καθώς θέλουμε να ελέγξουμε τον ισχυρισμό σε ε.σ. 5%.

   Στη συνέχεια CONTINUE - OK και εμφανίζεται το OUTPUT.

Για την διενέργεια αμφίπλευρων ελέγχων η διαδικασία υλοποίησης στο SPSS είναι ακριβώς η ίδια με αυτήν των μονόπλευρων ελέγχων. Η διαφορά έγκειται στην ερμηνεία των αποτελεσμάτων.

Ερμηνεία αποτελεσμάτων:

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
DAYS	20	101,25	20,961	4,687

One-Sample Test

	Test Value = 90					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
DAYS	2,400	19	,027	11,250	1,44	21,06

Ο αριθμητικός μέσος των παρατηρήσεων του δείγματος (**Mean** = 101,25) ικανοποιεί την ανισότητα **H_1 : μ > 90**, οπότε από την τιμή του **sig./2** θα αποφασίσουμε αν θα απορρίψουμε ή θα δεχθούμε την μηδενική υπόθεση.

Αν ο αριθμός αυτός είναι μικρότερος από το 0,05 τότε απορρίπτουμε την μηδενική υπόθεση, ενώ αν είναι μεγαλύτερος από το 0,05 αποδεχόμαστε την H_0.

Επομένως, στην περίπτωσή μας έχουμε **sig./2 =** 0,027/2 = 0,0135 < 0,05. Άρα απορρίπτουμε την μηδενική υπόθεση H_0.

~~**Αυτό σημαίνει ότι σε ε.σ. 5% αληθεύει ο ισχυρισμός ότι χρειάζονται πάνω από 90 ημέρες για να πουληθούν τα οικόπεδα.**~~

Ασκήσεις

1. Για να ελέγξουμε ένα νέο είδος πυρίτιδας μετράμε την ταχύτητα βλημάτων σε m/sec.
Από ένα δείγμα 8 βλημάτων πήραμε τις ακόλουθες μετρήσεις :

3005, 2925, 2935, 2965, 2995, 3005, 2937, 2905.

Να ελεγχθεί σε επίπεδο σημαντικότητας α=3% αν η μέση ταχύτητα είναι 4000m/sec.

Απάντηση :
Από τον έλεγχο προκύπτει ότι η μέση ταχύτητα δεν είναι 4000m/sec.

3

2. Μετρήσαμε την ετήσια βροχόπτωση (σε mm) σε μια περιοχή της Ελλάδας τα τελευταία 10 έτη και βρήκαμε τα παρακάτω αποτελέσματα :

   76.25, 85.25, 69.75, 73.5, 87.5, 67.25, 75.5, 70.75, 79.25, 64.5.

Να ελεγχθεί σε επίπεδο σημαντικότητας α=5% αν η μέση βροχόπτωση υπερβαίνει τα 75 mm^3 .

Απάντηση :

Κάνοντας τον έλεγχο, βλέπουμε ότι η μέση βροχόπτωση δεν υπερβαίνει αλλά θεωρείται ίση με 75 mm^3 .

4. ΕΛΕΓΧΟΣ ΥΠΟΘΕΣΕΩΝ ΓΙΑ ΤΟΥΣ ΜΕΣΟΥΣ –ΑΝΕΞΑΡΤΗΤΑ ΔΕΙΓΜΑΤΑ (Independent samples t-test)

Το κριτήριο Independent samples t-test χρησιμοποιείται όταν θέλουμε να συγκρίνουμε τους αριθμητικούς μέσους μ_1 και μ_2 δύο ανεξάρτητων δειγμάτων.

Ανεξάρτητα είναι δύο δείγματα όταν τα στοιχεία του ενός δεν μπορεί να είναι συγχρόνως και στοιχεία του άλλου.

Στις ασκήσεις όπου χρησιμοποιούμε το Independent samples t-test θα έχουμε δύο ανεξάρτητα δείγματα που όμως **θα εξετάζονται ως προς την ίδια μεταβλητή**.

Είδη ελέγχου

$H_0 : \mu_1 = \mu_2$ $\quad H_1 : \mu_1 \neq \mu_2$ (αμφίπλευρος έλεγχος)

$H_0 : \mu_1 = \mu_2$ $\quad H_1 : \mu_1 > \mu_2$ (μονόπλευρος έλεγχος)

$H_0 : \mu_1 = \mu_2$ $\quad H_1 : \mu_1 < \mu_2$ (μονόπλευρος έλεγχος)

Για να μπορούμε να χρησιμοποιήσουμε το κριτήριο, πρέπει να ισχύουν τα παρακάτω :

- Και τα δύο δείγματα θα πρέπει να έχουν επιλεγεί τυχαία
- Και τα δύο δείγματα θα πρέπει να προέρχονται από κανονικά κατανεμημένους πληθυσμούς. Επειδή ο έλεγχος κανονικότητας που κάνουμε έχει κάποιες μικροδιαφορές σε σχέση με την κλασική διαδικασία Explore, θα τον δούμε αναλυτικά στη συνέχεια.

  Εναλλακτικά, επιτρέπεται η χρήση τους χωρίς έλεγχο, όταν τα μεγέθη των δειγμάτων είναι αρκετά μεγάλα (> 30).
- να γνωρίζουμε επίπεδο σημαντικότητας α που μας ενδιαφέρει

Κρίσιμο είναι εδώ και το ερώτημα της ύπαρξης διαφοράς μεταξύ των διακυμάνσεων των δύο πληθυσμών, γεγονός που οδηγεί σε διαφορετικό στατιστικό test.

Τα βήματα που ακολουθούμε για την διαδικασία αυτή στο SPSS είναι τα παρακάτω :

1. Δημιουργούμε την μεταβλητή ως προς την οποία εξετάζονται τα δύο δείγματα στο Variable View και εισάγουμε τα δεδομένα και για τα δύο δείγματα στο Data View.
2. Στη συνέχεια δημιουργούμε μία νέα μεταβλητή με το όνομα group και από την οθόνη Variable View στην επιλογή Values ανοίγουμε ένα παράθυρο όπου:
   - Στο value γράφουμε 1 και στο label μία ονομασία για το πρώτο δείγμα, πατάμε add.
   - Στο value γράφουμε 2 και στο label μια ονομασία για το δεύτερο δείγμα, πατάμε add και OK.

   Πάμε στην οθόνη Data View και κάτω από την μεταβλητή group γράφουμε τον αριθμό 1 σε όσες τιμές της πρώτης μεταβλητής αφορούν το πρώτο δείγμα και τον αριθμό 2 σε όσες αφορούν το δεύτερο. Με αυτόν τον τρόπο διαχωρίζουμε τα δύο δείγματα ενώ έχουμε συμπεριλάβει τις τιμές των στοιχείων τους στην ίδια μεταβλητή.

3. Για τον έλεγχο κανονικότητας, επιλέγουμε:

   Analyze → Descriptive Statistics → Explore

   Στη συνέχεια

   - Στο Dependent List μεταφέρουμε την μεταβλητή μας και στο Factor list την μεταβλητή group.
   - Επιλέγουμε στο Display το Plots.
   - Επιλέγουμε δεξιά το Plots και κλικάρουμε μόνο το Normality plots with tests.

4. Για τον έλεγχο t-test επιλέγουμε :

   Analyze → Compare Means → Independent Samples t-test

5. Στο Test variable μεταφέρουμε την αρχική μεταβλητή μας και στο grouping variable την μεταβλητή group. Κλικάρουμε Define groups και στο group 1 γράφουμε τον αριθμό 1 ενώ στο group 2 τον αριθμό 2. Πατάμε ΟΚ.

Αποτελέσματα στο Output :

● **Πίνακας Group Statistics**
Από αυτόν τον πίνακα μας ενδιαφέρουν μόνο οι μέσοι (mean) των δύο δειγμάτων.

● **Πίνακας Independent Samples Test** Από αυτόν τον πίνακα μας ενδιαφέρουν:
- ο αριθμός sig που υπάρχει στο Levene's Test for Equality of Variances
- οι δύο αριθμοί sig που υπάρχουν στο t-test for Equality of Means

Συμπέρασμα:

Επειδή στον πίνακα Independent Samples Test στο πεδίο t-test for Equality of Means υπάρχουν δύο sig, για να ξέρουμε ποιο θα επιλέξουμε για να το συγκρίνουμε με το επίπεδο σημαντικότητας α, απαιτείται να κάνουμε πρώτα αμφίπλευρο έλεγχο ισότητας των διασπορών των δύο δειγμάτων.

Διατύπωση του ελέγχου ισότητας διασπορών:

$$H_0 : \sigma_1^2 = \sigma_2^2 \qquad H_1 : \sigma_1^2 \neq \sigma_2^2$$

Το αποτέλεσμα προκύπτει από το sig που έχει προκύψει στο πεδίο Levene's Test for Equality of Variances του πίνακα Independent Samples Test. Ειδικότερα:

- αν sig>α τότε αποδεχόμαστε την υπόθεση H_0 , δηλαδή ότι οι διασπορές των δύο δειγμάτων είναι ίσες και στη συνέχεια θα χρησιμοποιήσουμε το sig της πρώτης γραμμής από το πεδίο t-test for Equality of Means του πίνακα Independent Samples Test
- αν sig<α τότε απορρίπτουμε την υπόθεση H_0 , δηλαδή αποδεχόμαστε ότι οι διασπορές των δύο δειγμάτων είναι άνισες και στη συνέχεια θα χρησιμοποιήσουμε

το sig της δεύτερης γραμμής από το πεδίο t-test for Equality of Means του πίνακα Independent Samples Test

Το συμπέρασμα του ελέγχου για τους μέσους, προκύπτει όπως και στην περίπτωση του One Sample t-test, χρησιμοποιώντας το κατάλληλο σε κάθε περίπτωση sig, δηλαδή:

● **Αμφίπλευρος έλεγχος**

- αν sig>α τότε αποδεχόμαστε την υπόθεση H_0 ➢ αν sig<α τότε απορρίπτουμε την υπόθεση H_0

● **Μονόπλευρος έλεγχος**

Αν οι δύο mean ικανοποιούν την ανισότητα της H_1 τότε ισχύουν τα εξής :

- αν $\frac{sig}{2} > \alpha$ τότε αποδεχόμαστε την υπόθεση H_0
- αν $\frac{sig}{2} < \alpha$ τότε απορρίπτουμε την υπόθεση H_0

Αν οι δύο mean δεν ικανοποιούν την ανισότητα της H_1 τότε ισχύουν τα εξής:

- αν $1 - \frac{sig}{2} > \alpha$ τότε αποδεχόμαστε την υπόθεση H_0
- αν $1 - \frac{sig}{2} < \alpha$ τότε απορρίπτουμε την υπόθεση H_0

Εφαρμογή της διαδικασίας Independent samples t-test

θέλουμε να συγκρίνουμε τους μέσους μισθούς σε δυο διαφορετικές κατηγορίες εργαζομένων όπως οι δασκάλες και οι νοσοκόμες. Μια εταιρεία εξετάζει εάν ο μισθός των νοσοκόμων είναι υψηλότερος από τον μισθό των δασκάλων.

Για τη μελέτη αυτή συλλέχθηκε το ακόλουθο δείγμα:

ΔΑΣΚΑΛΕΣ (ΕΥΡΩ)	545	526	527	484	509	502	520	529	530	542	532	575
ΝΟΣΟΚΟΜΕΣ (ΕΥΡΩ)	541	590	521	471	550	559	525	529				

Είναι λογικό να καταλήξουμε στο συμπέρασμα ότι ο μισθός των νοσοκόμων είναι υψηλότερος από των δασκάλων; (α = 0.01).

Στην άσκηση αυτή έχουμε 2 άγνωστα δείγματα μισθών, με 12 και 8 παρατηρήσεις το κάθε δείγμα αντίστοιχα. Επειδή έχουμε το γεγονός ότι οι παρατηρήσεις μας αφορούν δείγματα τα οποία δεν εμφανίζονται σε δυο διαφορετικές χρονικές αλλά στη ίδια χρονική στιγμή και δεν μπορεί ένα στοιχείο να ανήκει και στα δύο δείγματα τα θεωρούμε ως ανεξάρτητα και δουλεύουμε με το **Independent t-test**.

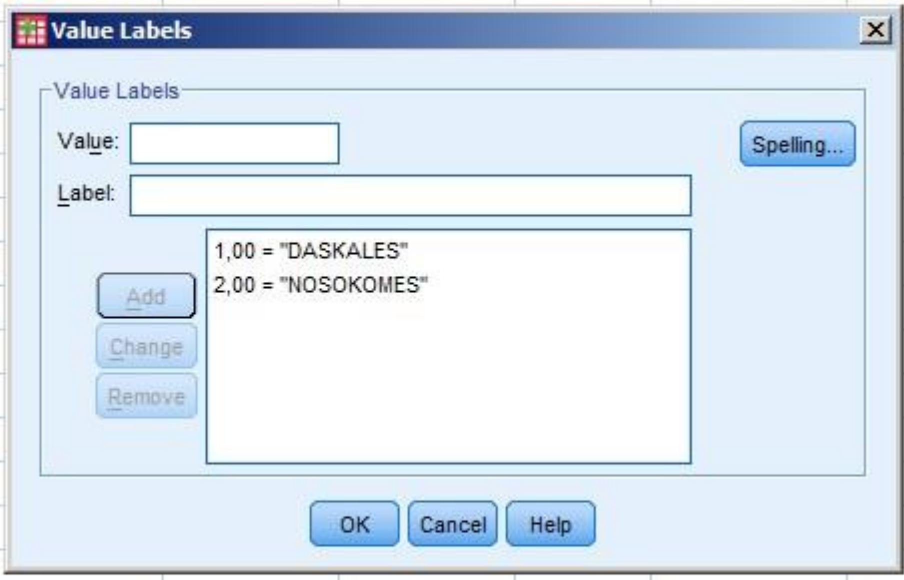
Πριν ξεκινήσουμε να εκτελούμε την διαδικασία μέσω του SPSS θα πρέπει να εξετάσουμε τον τρόπο με τον οποίο και θα εισάγουμε τα δεδομένα μας στο μενού **DATA VIEW** του SPSS. Στην περίπτωση των ανεξάρτητων δειγμάτων χρειάζεται μεγάλη προσοχή η εισαγωγή των δεδομένων, καθώς **δεν θα δημιουργούμε δυο μεταβλητές που αφορούν τον μισθό των νοσοκόμων και των δασκάλων ξεχωριστά αλλά μία μεταβλητή που θα αφορά γενικά τους μισθούς.**

Τα βήματα που ακολουθούμε είναι:

1. Δημιουργούμε την μεταβλητή **ΜΙΣΘΟΙ** και εισάγουμε όλα τα δεδομένα.

2. Στο μενού **VARIABLE VIEW** δημιουργούμε την μεταβλητή **GROUP** και στη στήλη **VALUES εισάγουμε τα εξής VALUE LABELS:**

    - Θέτουμε όπου Value την τιμή 1 και Value Labels το όνομα DASKALES και πατάμε ADD. Με βάση την διαδικασία αυτή εμφανίζεται “1=DASKALES”.
    - Ομοίως όπου Value την τιμή 2 και Value Labels το όνομα NOSOKOMES και πατάμε ADD. Με βάση την διαδικασία αυτή εμφανίζεται “2=NOSOKOMES**”**



3. Στο μενού **DATA VIEW** καταχωρούμε τις τιμές 1 και 2 για να διαχωρίσουμε τα δείγματα

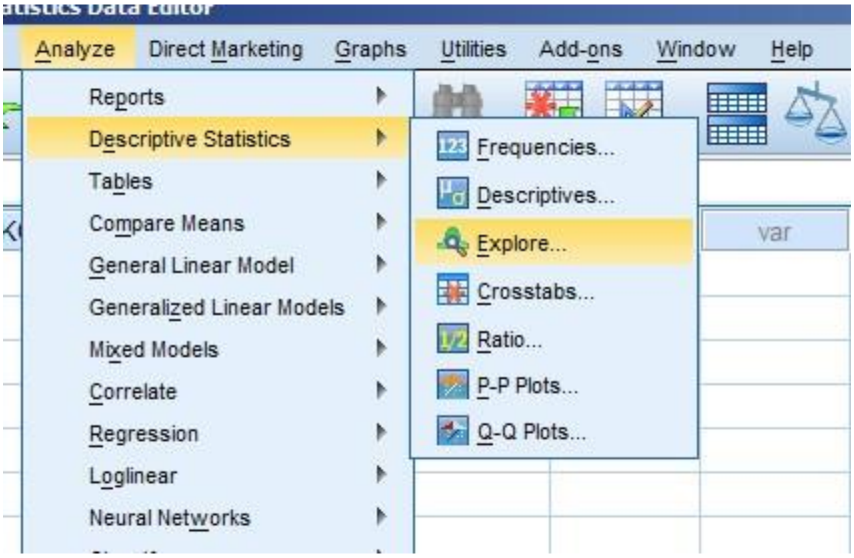
	ΜΙΣΘΟΙ	GROUP
1	545	DASKALES
2	526	DASKALES
3	527	DASKALES
4	484	DASKALES
5	509	DASKALES
6	502	DASKALES
7	520	DASKALES
8	529	DASKALES
9	530	DASKALES
10	542	DASKALES
11	532	DASKALES
12	575	DASKALES
13	541	NOSOKOMES
14	590	NOSOKOMES
15	521	NOSOKOMES
16	471	NOSOKOMES
17	550	NOSOKOMES
18	559	NOSOKOMES

4. Επειδή η άσκηση μας ρωτά για το αν ο μισθός των νοσοκόμων είναι μεγαλύτερος από αυτών των δασκάλων καταλαβαίνουμε ότι ο έλεγχος είναι μονόπλευρος και μάλιστα έχει την εξής μορφή

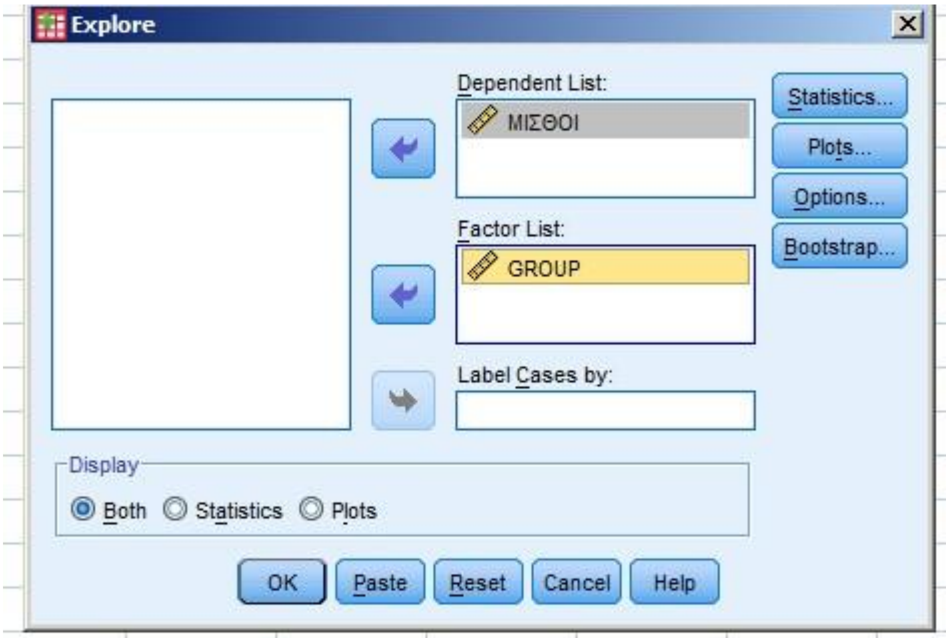
$$H_0 : \mu_\Delta = \mu_N \qquad H_1 : \mu_\Delta < \mu_N$$

5. Έλεγχος κανονικότητας

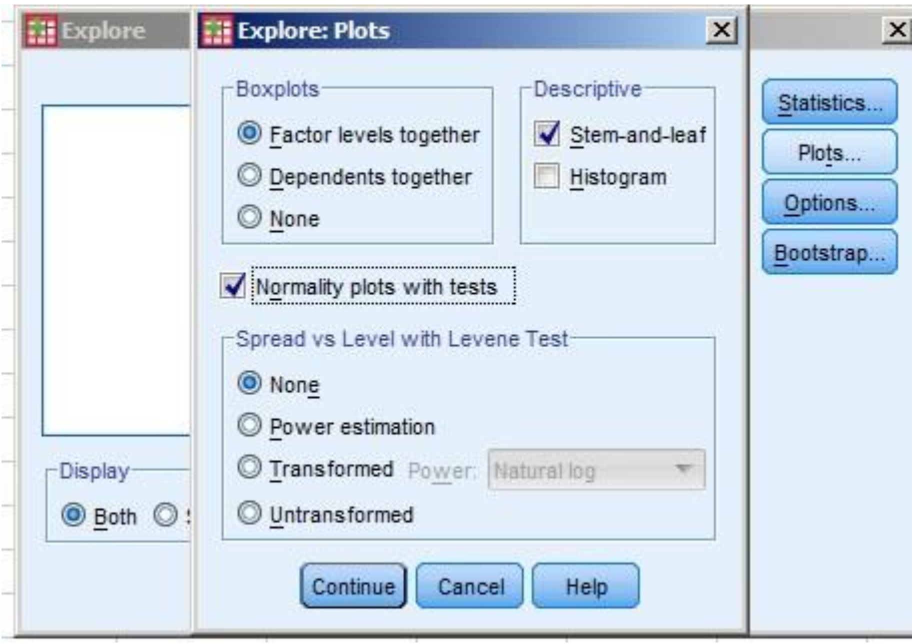
Θα πρέπει να προηγηθεί έλεγχος κανονικότητας (διότι οι παρατηρήσεις είναι λιγότερες από 30 σε κάθε δείγμα), προκειμένου να δούμε ότι τόσο ο μισθός των δασκάλων, όσο και ο μισθός των νοσοκόμων ακολουθούν κανονική κατανομή, με τη χρήση της εντολής **Explore:**



Στο Dependent List μεταφέρουμε την μεταβλητή μας **(ΜΙΣΘΟΙ)** και στο Factor list την μεταβλητή **GROUP**.



Επιλέγουμε δεξιά το **Plots** και κλικάρουμε μόνο το Normality plots with tests.



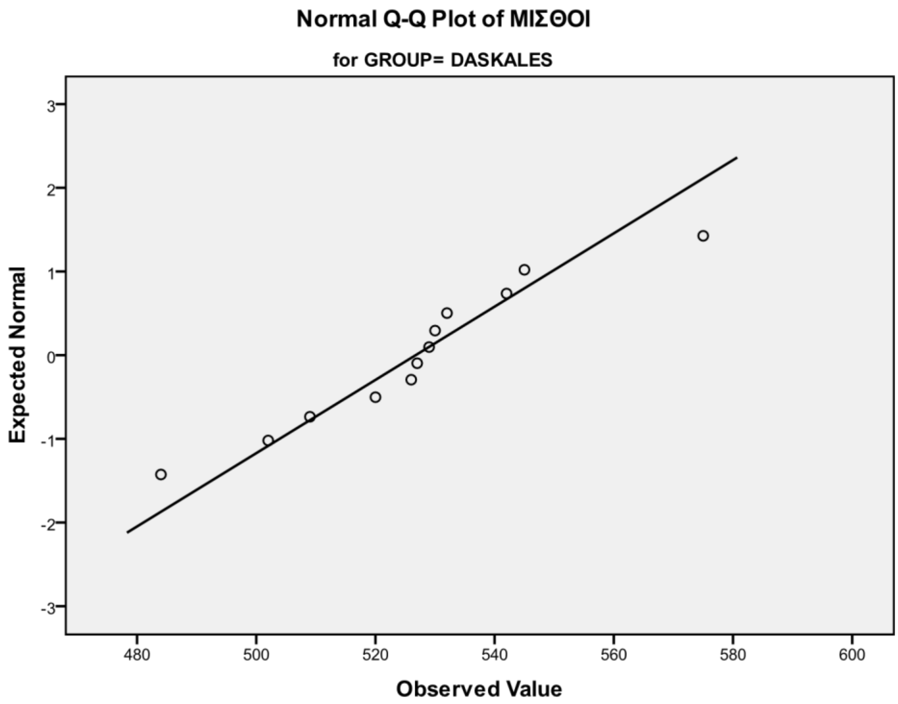
Ερμηνεία κατά τα γνωστά:

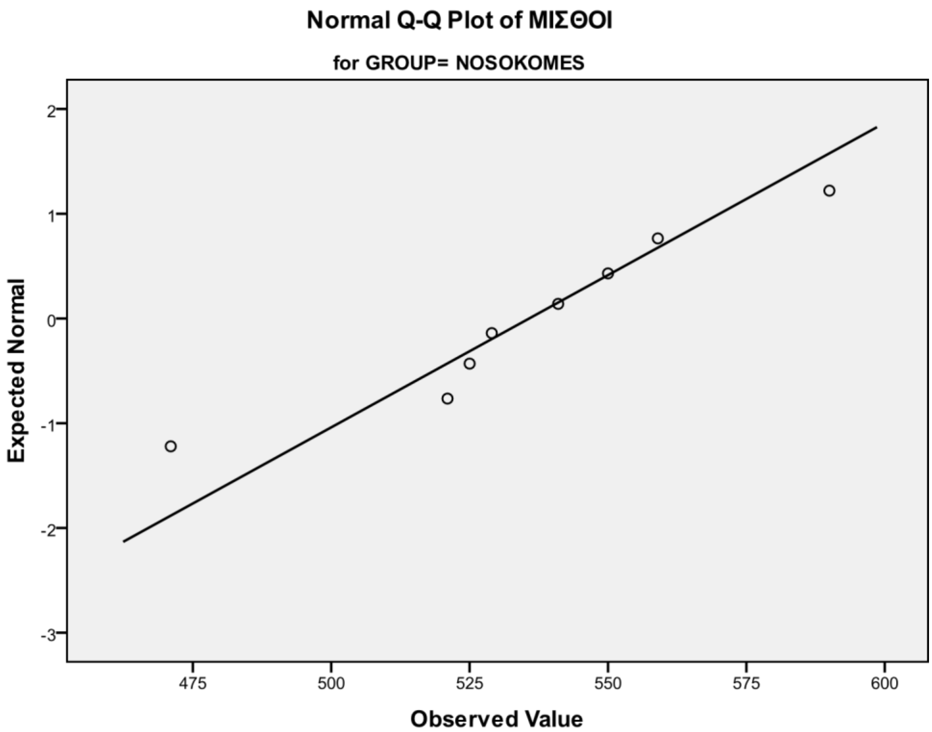
Tests of Normality

GROUP		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
ΜΙΣΘΟΙ	DASKALES	,159	12	,200*	,959	12	,774
	NOSOKOMES	,209	8	,200*	,958	8	,788

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

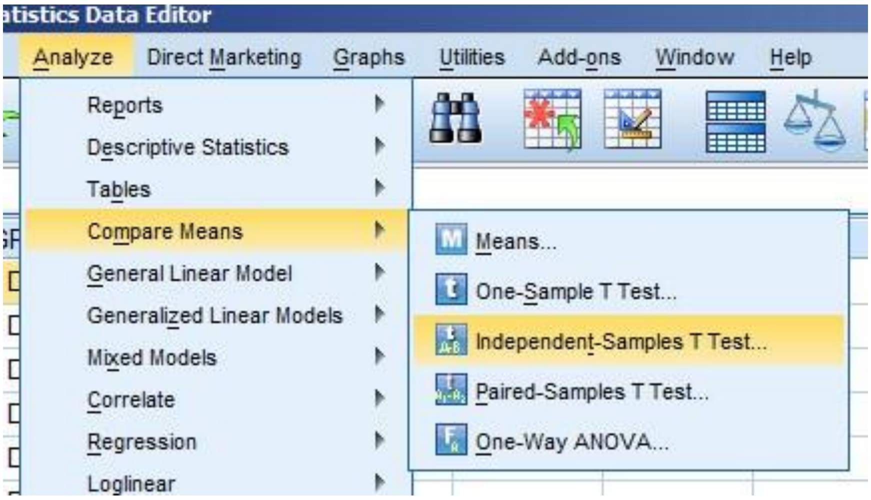




Από τον έλεγχο προκύπτει ότι τόσο ο μισθός των δασκάλων, όσο και ο μισθός των νοσοκόμων ακολουθούν κανονική κατανομή.

6. Έλεγχος t-test. Επιλέγουμε

Analyze → Compare Means → Independent Samples t-test

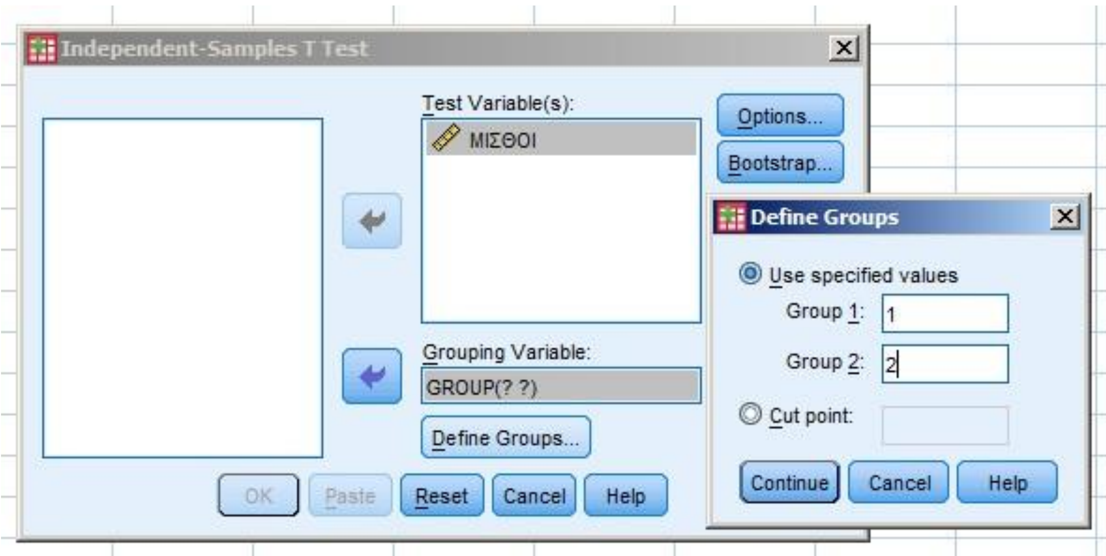


Και στη συνέχεια:

- Επιλέγουμε την μεταβλητή που παριστά το χαρακτηριστικό που μας ενδιαφέρει, **ΜΙΣΘΟΙ,** και τη μεταφέρουμε στο παράθυρο **TEST VARIABLE**.

  (<u>Παρατήρηση:</u> Ασφαλώς μπορούμε να επιλέξουμε ταυτόχρονα περισσότερες από μία μεταβλητές. Για την κάθε μία από αυτές θα πραγματοποιηθεί ένα ξεχωριστό ttest.)

- Στο παράθυρο **GROUPING VARIABLE** εισάγουμε τη μεταβλητή **GROUP** και στη συνέχεια πατάμε στην επιλογή **DEFINE GROUPS** και εμφανίζεται το παράθυρο Define Group

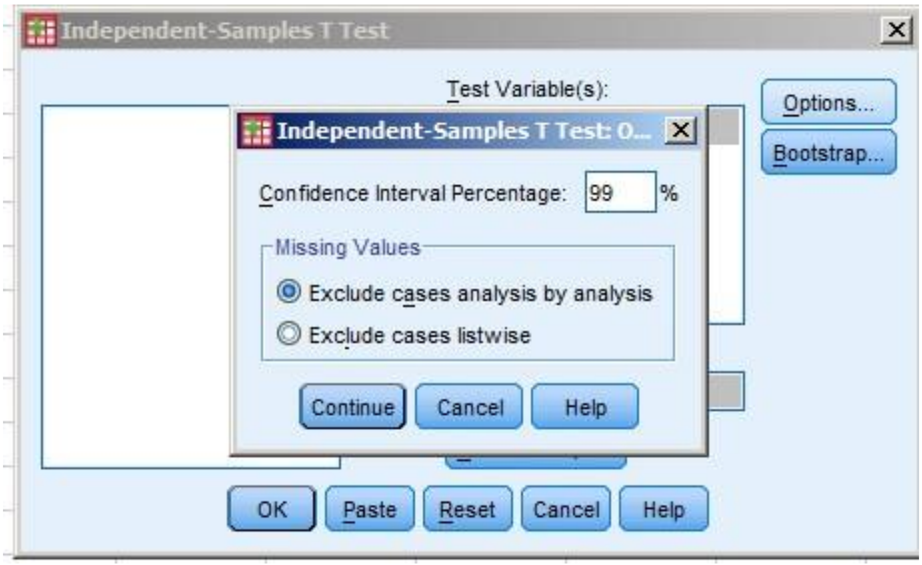


Στην επιλογή **used specified values** και στις θέσεις GROUP 1 και GROUP 2 θα βάλουμε τους ανάλογους κωδικούς που χρησιμοποιήσαμε και στην διαδικασία VALUE LABELS της μεταβλητής Group. Για τους μισθούς των Δασκάλων και των Νοσοκόμων οι κωδικοί είναι 1 και 2.

CUT POINT: Εναλλακτικά, μπορούμε να δηλώσουμε έναν αριθμό ο οποίος θα χωρίσει στα δύο τις τιμές της μεταβλητής που ορίζει τα δύο δείγματα. Τότε το πρώτο δείγμα σχηματίζεται από όλες τις περιπτώσεις που αντιστοιχούν σε τιμή μικρότερη του αριθμού που δηλώσαμε και το δεύτερο από τις υπόλοιπες.

- Με την επιλογή **OPTIONS** καθορίζουμε τον τρόπο χειρισμού των ελλειπουσών τιμών και προσδιορίζουμε το επίπεδο σημαντικότητας του διαστήματος εμπιστοσύνης που θα κατασκευαστεί.

Δηλώνουμε το επιθυμητό διάστημα εμπιστοσύνης 99% (γιατί στην εκφώνηση της άσκησης δίνεται ότι α = 0.01) και στη συνέχεια ΟΚ



- **CONFIDENCE INTERVAL:** Εξ ορισμού υπολογίζεται το 95% διάστημα εμπιστοσύνης για τη διαφορά $\mu_1 - \mu_2$ των μέσων τιμών των δύο πληθυσμών. Μπορούμε να δηλώσουμε οποιαδήποτε τιμή στο διάστημα [1, 99] υπολογίζοντας έτσι το αντίστοιχο Δ.Ε.
- **MISSING VALUES:** Το SPSS πραγματώνει χωριστά το κάθε test που ζητήθηκε χρησιμοποιώντας όλες τις περιπτώσεις που είναι έγκυρες για τις μεταβλητές που συμμετέχουν σ' αυτό (Exclude case analysis by analysis). Μπορούμε όμως να ζητήσουμε από το SPSS να χρησιμοποιήσει σε όλα τα test μόνο τις περιπτώσεις που είναι ταυτόχρονα έγκυρες για όλες τις μεταβλητές του καταλόγου Test Variables (Exlude cases listwise). Η όποια επιλογή έχει φυσικά νόημα μόνο στην περίπτωση που ο κατάλογος Test Variables περιλαμβάνει περισσότερες από μια μεταβλητές.

Ερμηνεία αποτελεσμάτων:

Group Statistics

GROUP	N	**Mean**	Std. Deviation	Std. Error Mean

ΜΙΣΘΟΙ DASKALES	12	**526,75**	22,840	6,593
NOSOKOMES	8	**535,75**	34,404	12,164

Στον πίνακα **(Group Statistics)** εμφανίζονται:

- το πλήθος των στοιχείων των δύο δειγμάτων.
- Ο μέσος του κάθε δείγματος. **(Mean).**
- η τυπική απόκλιση αυτών **(Std. Deviation)**
- το τυπικό σφάλμα του μέσου **(Std. Error Mean)**

Από τον πίνακα **Group Statistics** μας ενδιαφέρουν μόνο οι μέσοι (mean) των δύο δειγμάτων.

Στον πίνακα (**INDEPENDENT SAMPLES TEST)** κάνουμε αρχικά τον έλεγχο για την **ισότητα των διακυμάνσεων - Levene's test for equality of variances.**

Η αρχική υπόθεση αυτού του τεστ είναι ότι οι Διακυμάνσεις (Variances) των δύο υποομάδων είναι ίσες (equal variances assumed) έναντι του γεγονότος ότι οι διασπορές είναι άνισες (equal variances not assumed).

$$H_0: \sigma_1^2 = \sigma_2^2 \qquad H_1: \sigma_1^2 \neq \sigma_2^2$$

Από το **Levene's Test for Equality of Variances** θα συγκρίνουμε το sig με το α, για να δούμε αν αποδεχόμαστε ή απορρίπτουμε την H_0. Έχουμε sig = 0.322 > α = 0.01. Έτσι αποδεχόμαστε την αρχική μας υπόθεση H_0 (ότι οι διασπορές είναι ίσες) και συνεχίζουμε με τον έλεγχο των μέσων.

Independent Samples Test

	Levene's Test for Equality of Variances		t-test for Equality of Means						
								99% Confidence Interval of the Difference	
	F	Sig.	t	df	Sig. (2tailed)	Mean Difference	Std. Error Difference	Lower	Upper
ΜΙΣΘΟΙ Equal variances assumed	1,035	,322	-,706	18	,489	-9,000	12,740	-45,672	27,672

	Equal variances not assumed			-,650	11,108	,529	-9,000	13,836	-51,888	33,888

Στη συνέχεια του ίδιου πίνακα έχουμε το τεστ για την **ισότητα των μέσων (t - test for equality of means).** Η αρχική υπόθεση του τεστ είναι ότι οι δύο υποομάδες έχουν τον ίδιο μέσο, έναντι του γεγονότος ότι οι δυο υποομάδες δεν έχουν τον ίδιο μέσο.

Παρατηρούμε, βέβαια, ότι υπάρχουν δύο τιμές sig. Εμείς θα επιλέξουμε τη μία από τις δύο με βάση την αποδοχή ή την απόρριψη της αρχικής υπόθεσης του προηγούμενου ελέγχου.

Στην περίπτωσή μας θα χρησιμοποιήσουμε το sig της πρώτης γραμμής.

Από τον πίνακα GROUP STATISTICS **ελέγχουμε αν οι μέσοι ικανοποιούν την ανισότητα της H_1** δηλαδή εάν: **$mean_Δ$ = 526,75 < $mean_N$ = 535,75**
που ισχύει, επομένως οι μέσοι ικανοποιούν την ανισότητα της H_1 και θα συγκρίνουμε την τιμή sig/2 =0 ,489/2 = 0,2445 με το επίπεδο σημαντικότητας α = 0,01.

Παρατηρούμε ότι sig/2 = 0,2445 > α = 0.01 Συνεπώς αποδεχόμαστε την H_0 και έτσι φαίνεται ότι ο μισθός των δασκάλων να μην είναι μικρότερος από αυτών των νοσοκόμων σε επίπεδο σημαντικότητας α = 0.01.

Στον ίδιο πίνακα δίνονται στη συνέχεια:

- η διαφορά των μέσων **(MEAN DIFFERENCE)**
- το τυπικό σφάλμα της διαφοράς **(STD. ERROR OF DIFFERENCE)** και
- τα όρια **(LOWER VALUE - UPPER VALUE)** του διαστήματος εμπιστοσύνης της διαφοράς των μέσων **(99% CONFIDENCE INTERVAL OF THE DIFFERENCE).**

Ασκήσεις

1. Μια διαφημιστική εταιρεία θέλει να συγκρίνει την διαφημιστική δαπάνη δύο επιχειρήσεων. Κατέγραψε λοιπόν την ετήσια διαφημιστική δαπάνη τους των τελευταίων 9 ετών. Αν οι πρώτες 9 παρατηρήσεις αφορούν την πρώτη επιχείρηση και οι άλλες 9 την δεύτερη, μπορούμε να ισχυριστούμε ότι η πρώτη έχει μεγαλύτερη διαφημιστική δαπάνη από την δεύτερη; (α=5%)

Α/Α	ΕΤΗΣΙΑ ΔΙΑΦΗΜΙΣΤΙΚΗ ΔΑΠΑΝΗ ΣΕ ΧΙΛΙΑΔΕΣ ΕΥΡΩ
1	29
2	32
3	29
4	25
5	34
6	40
7	27
8	31
9	32
10	37
11	32
12	35
13	28
14	41
15	44
16	35
17	34
18	32

Απάντηση :
Μετά τον έλεγχο διαπιστώνουμε ότι η διαφημιστική δαπάνη για τις δύο επιχειρήσεις θεωρείται ίδια.

2. Εξετάζονται δείγματα νερού ως προς το pH από 16 λίμνες. Αν οι πρώτες 8 παρατηρήσεις αφορούν λίμνες από την περιοχή Α και οι άλλες 8 λίμνες από την περιοχή Β, να εξεταστεί αν είναι δυνατόν το pH του νερού στην περιοχή Α να είναι μικρότερο από αυτό της περιοχής Β. (α=5%)

Α/Α	pH
1	6,9
2	6,2
3	6,3
4	5,9
5	6
6	7
7	6,5
8	6,6
9	7
10	6,9
11	6,7
12	7,1
13	6,8
14	7,1
15	7
16	7,2

Απάντηση :

Από τον έλεγχο προκύπτει ότι το pH του νερού στην περιοχή A είναι μικρότερο από το αντίστοιχο της περιοχής B.

5. ΕΛΕΓΧΟΣ ΥΠΟΘΕΣΕΩΝ ΓΙΑ ΤΟΥΣ ΜΕΣΟΥΣ –ΕΞΑΡΤΗΜΕΝΑ ΔΕΙΓΜΑΤΑ (Paired samples t-test)

Το κριτήριο Paired samples t-test χρησιμοποιείται όταν θέλουμε να συγκρίνουμε τους αριθμητικούς μέσους μ_1 και μ_2 δύο εξαρτημένων δειγμάτων.

Ε**ξαρτημένα είναι δύο δείγματα** όταν τα στοιχεία τους αναφέρονται στο ίδιο αντικείμενο (χαρακτηριστικό ατόμου), εξετάζουν την ίδια παράμετρο (μεταβλητή) αλλά διαφοροποιούνται ως προς ένα επιμέρους προσδιοριστικό στοιχείο (π.χ. χρονική στιγμή).

Είδη ελέγχου

$H_0 : \mu_1 = \mu_2$ $\quad H_1 : \mu_1 \neq \mu_2$ (αμφίπλευρος έλεγχος)

$H_0 : \mu_1 = \mu_2$ $\quad H_1 : \mu_1 > \mu_2$ (μονόπλευρος έλεγχος)

$H_0 : \mu_1 = \mu_2$ $\quad H_1 : \mu_1 < \mu_2$ (μονόπλευρος έλεγχος)

Για να μπορούμε να χρησιμοποιήσουμε το κριτήριο, πρέπει να ισχύουν τα παρακάτω :

- Και τα δύο δείγματα θα πρέπει να έχουν επιλεγεί τυχαία
- Και τα δύο δείγματα θα πρέπει να προέρχονται από κανονικά κατανεμημένους πληθυσμούς. Επειδή ο έλεγχος κανονικότητας που κάνουμε έχει κάποιες μικροδιαφορές σε σχέση με την κλασική διαδικασία Explore, θα τον δούμε αναλυτικά στη συνέχεια.

  Εναλλακτικά, επιτρέπεται η χρήση τους χωρίς έλεγχο, όταν τα μεγέθη των δειγμάτων είναι αρκετά μεγάλα (> 30).
- να γνωρίζουμε επίπεδο σημαντικότητας α που μας ενδιαφέρει

Τα βήματα που ακολουθούμε για την διαδικασία αυτή στο SPSS είναι τα παρακάτω :

1. Δημιουργούμε δύο μεταβλητές (μία για κάθε δείγμα) με βάση το προσδιοριστικό στοιχείο που διαφοροποιεί τα δύο δείγματα στο Variable View και εισάγουμε τα δεδομένα και για τα δύο δείγματα στο Data View.

2. Για τον έλεγχο κανονικότητας, επιλέγουμε:

   Analyze → Descriptive Statistics → Explore

   Στη συνέχεια

   - Στο Dependent List μεταφέρουμε και τις δύο μεταβλητές

     Επιλέγουμε στο Display το Plots.
   - Επιλέγουμε δεξιά το Plots και κλικάρουμε μόνο το Normality plots with tests.
3. Για τον έλεγχο t-test επιλέγουμε :

   Analyze → Compare Means → Paired Samples t-test
4. Στο Paired Variables μεταφέρουμε διαδοχικά και τις δύο μεταβλητές ώστε να δημιουργηθεί ζεύγος μεταβλητών. Πατάμε OK.

Αποτελέσματα στο Output :

● Πίνακας Paired Samples Statistics
Από αυτόν τον πίνακα μας ενδιαφέρουν μόνο οι μέσοι (mean) των δύο δειγμάτων.

● Πίνακας Paired Samples Test
Από αυτόν τον πίνακα μας ενδιαφέρει μόνο ο αριθμός sig

Συμπέρασμα:

Αρχικά θα κάνουμε έλεγχο συσχέτισης των δύο δειγμάτων.

Το συμπέρασμα του ελέγχου για τους μέσους, προκύπτει όπως και στην περίπτωση του One Sample t-test, χρησιμοποιώντας το κατάλληλο σε κάθε περίπτωση sig, δηλαδή:

● **Αμφίπλευρος έλεγχος**

- αν sig>α τότε αποδεχόμαστε την υπόθεση H_0 ➢ αν sig<α τότε απορρίπτουμε την υπόθεση H_0

● **Μονόπλευρος έλεγχος**

Αν οι δύο mean ικανοποιούν την ανισότητα της H_1 τότε ισχύουν τα εξής :

- αν $\frac{sig}{2} > \alpha$ τότε αποδεχόμαστε την υπόθεση H_0
- αν $\frac{sig}{2} < \alpha$ τότε απορρίπτουμε την υπόθεση H_0

Αν οι δύο mean δεν ικανοποιούν την ανισότητα της H_1 τότε ισχύουν τα εξής:

- αν $1 - \frac{sig}{2} > \alpha$ τότε αποδεχόμαστε την υπόθεση H_0
- αν $1 - \frac{sig}{2} < \alpha$ τότε απορρίπτουμε την υπόθεση H_0

Εφαρμογή της διαδικασίας Paired samples t-test

Η Υπηρεσία αστικών συγκοινωνιών μιας πόλης έκανε μια μελέτη για να διαπιστώσει αν ο φωτισμός στον δρόμο τη νύχτα συντελεί στην μείωση των αυτοκινητιστικών δυστυχημάτων.

Ο ακόλουθος πίνακας δείχνει τον μέσο ετήσιο αριθμό δυστυχημάτων σε δεκατρία σημεία της πόλης ένα χρόνο πριν και ένα χρόνο μετά την εγκατάσταση του νυχτερινού φωτισμού.

ΘΕΣΗ	Α	Β	Γ	Δ	Ε	Ζ	Η	Θ	Ι	Κ	Λ	Μ	Ν
Δυστυχήματα Πριν	8	12	5	4	6	3	4	3	2	6	6	9	15

Δυστυχήματα Μετά	5	3	2	1	4	2	2	4	3	5	0	8	11

Είναι τα δεδομένα αυτά ισχυρή ένδειξη ότι ο νυχτερινός φωτισμός συντελεί στην μείωση των δυστυχημάτων; (α = 0,05)

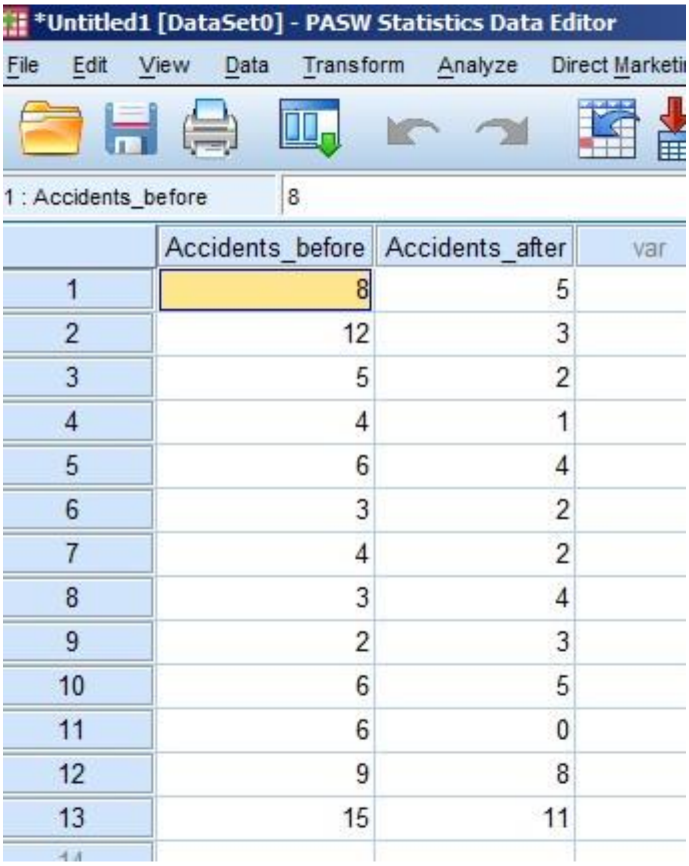
Υποθέστε ότι οι δύο μεταβλητές ακολουθούν κανονική κατανομή

Προτού ξεκινήσουμε την διαδικασία μέσω του προγράμματος SPSS θα πρέπει να έχουμε ξεκαθαρίσει ότι πρόκειται γα εξαρτημένα δείγματα, γιατί **στην περίπτωση των εξαρτημένων δειγμάτων θα πρέπει να δημιουργήσουμε δυο ξεχωριστές μεταβλητές σε αντίθεση με την μια που δημιουργούσαμε στα ανεξάρτητα δείγματα**.

Στην παρούσα άσκηση έχουμε 2 δείγματα δυστυχημάτων, με 13 παρατηρήσεις το κάθε δείγμα. Επειδή έχουμε το γεγονός ότι οι παρατηρήσεις μας αφορούν δείγματα τα οποία εμφανίζονται σε δυο διαφορετικές χρονικές στιγμές και εξετάζουν ατυχήματα πριν και μετά τον φωτισμό ενός δρόμου στα ίδια σημεία, τα θεωρούμε ως εξαρτημένα και δουλεύουμε με το Paired Samples t-test.

Τα βήματα που ακολουθούμε είναι:

1. Δημιουργούμε τις δύο μεταβλητές **Accidents_before** και **Accidents_after** οι οποίες και χαρακτηρίζουν τα δεδομένα της άσκησης μας, και εισάγουμε τις τιμές τους



	Accidents_before	Accidents_after	var
1	8	5	
2	12	3	
3	5	2	
4	4	1	
5	6	4	
6	3	2	
7	4	2	
8	3	4	
9	2	3	
10	6	5	
11	6	0	
12	9	8	
13	15	11	

2. Επειδή η άσκηση μας ρωτά για το αν ο φωτισμός όντως συντέλεσε στην μείωση των δυστυχημάτων, ο έλεγχος διαμορφώνεται ως εξής (πρόκειται για μονόπλευρο έλεγχο):

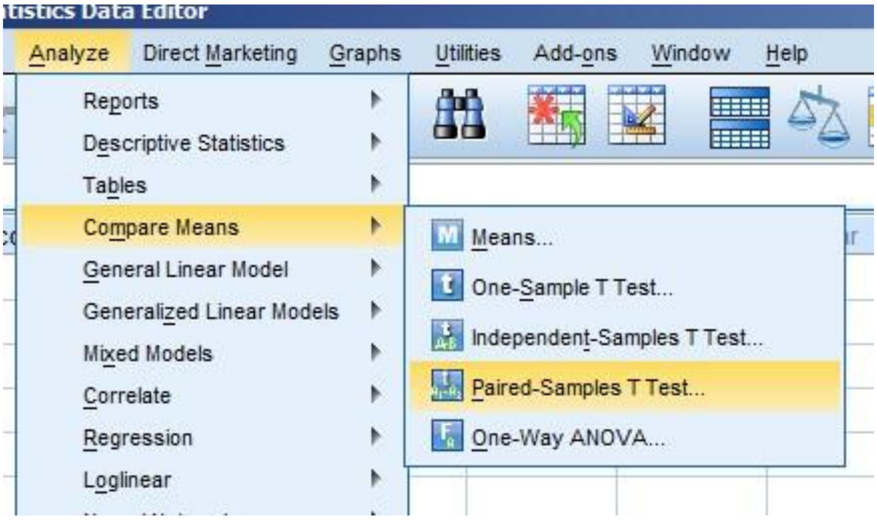
$$H_0 : \mu_\Pi = \mu_M \qquad H_1 : \mu_\Pi > \mu_M$$

3. Έλεγχος κανονικότητας

   Δεν είναι απαραίτητος (βλέπε εκφώνηση άσκησης)

4. Έλεγχος t-test. Επιλέγουμε

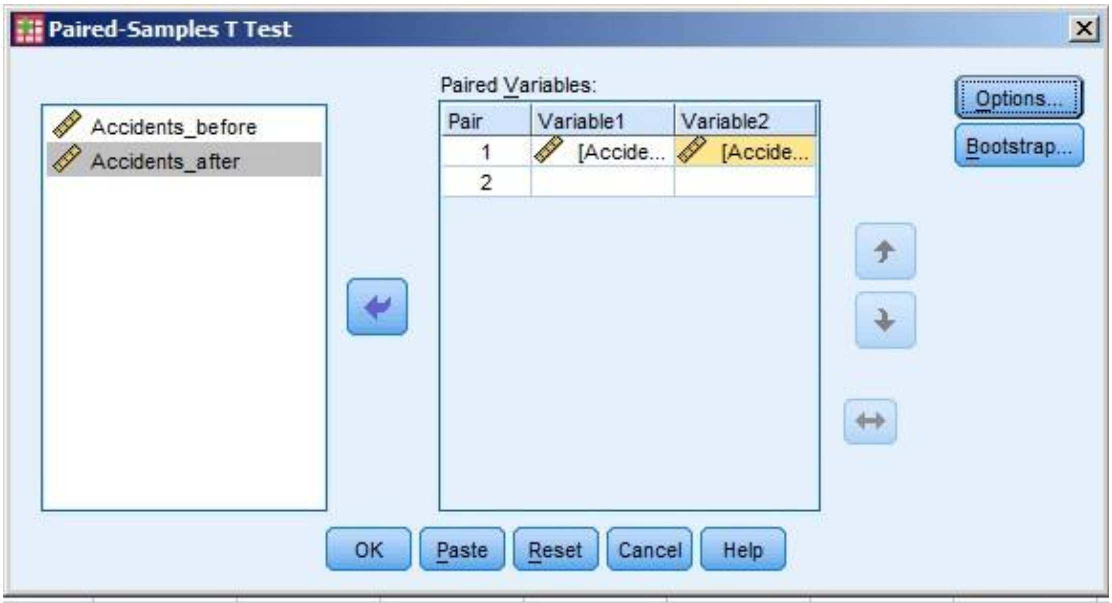
   Analyze → Compare Means → Paired Samples t-test



Και στη συνέχεια:

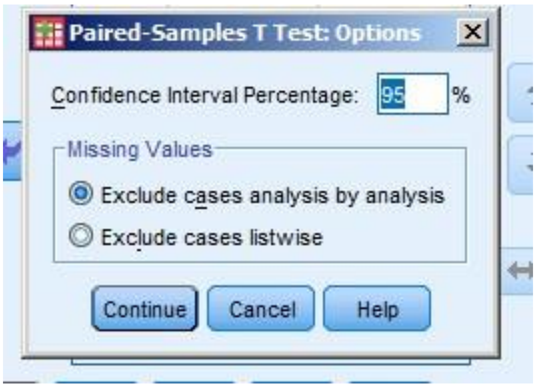
- Μετακινούμε την μία μεταβλητή**, Accidents_before,** στο πεδίο Variable 1, και την δεύτερη μεταβλητή, **Accidents_after,** στο πεδίο Variable 2.

  (<u>Παρατήρηση:</u> Μπορούμε να επιλέξουμε ταυτόχρονα περισσότερα από ένα ζεύγος μεταβλητών π.χ. pair 2 (επαναλαμβάνοντας την πιο πάνω διαδικασία). Για το καθένα ζεύγος θα πραγματωθεί ένα ξεχωριστό Paired Test.



- Με την επιλογή **OPTIONS** καθορίζουμε τον τρόπο χειρισμού των ελλειπουσών τιμών και προσδιορίζουμε το επίπεδο σημαντικότητας του διαστήματος εμπιστοσύνης που θα κατασκευαστεί.

Δηλώνουμε εδώ το επιθυμητό διάστημα εμπιστοσύνης 95%



- **CONFIDENCE INTERVAL:** Εξ ορισμού υπολογίζεται το 95% διάστημα εμπιστοσύνης για τη διαφορά $\mu_1 - \mu_2$ των μέσων τιμών των δύο πληθυσμών. Μπορούμε να δηλώσουμε οποιαδήποτε τιμή στο διάστημα [1, 99] υπολογίζοντας έτσι το αντίστοιχο Δ.Ε.
- **MISSING VALUES:** Το SPSS πραγματώνει χωριστά το κάθε test που ζητήθηκε χρησιμοποιώντας όλες τις περιπτώσεις που είναι έγκυρες για τις μεταβλητές που συμμετέχουν σ' αυτό (Exclude case analysis by analysis). Μπορούμε όμως να ζητήσουμε από το SPSS να χρησιμοποιήσει σε όλα τα test μόνο τις περιπτώσεις που είναι ταυτόχρονα έγκυρες για όλες τις μεταβλητές του καταλόγου Test Variables (Exlude cases listwise). Η όποια επιλογή έχει φυσικά νόημα μόνο στην περίπτωση που ο κατάλογος Test Variables περιλαμβάνει περισσότερες από μια μεταβλητές.

Ερμηνεία αποτελεσμάτων:

Ο πίνακας **Paired Samples Statistics** περιέχει τα γνωστά βασικά στατιστικά μέτρα και για τα δύο δείγματα τα οποία εξετάζουμε.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Accidents_before	6,38	13	3,776	1,047
	Accidents_after	3,85	13	2,968	,823

Ο πίνακας **Paired Samples Correlations** μας δίνει:

- τον συντελεστή συσχέτισης των δύο μεταβλητών (Correlation coefficient)
- την significant value του αντιστοίχου ελέγχου.

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Accidents_before & Accidents_after	13	,697	,008

Από τον πίνακα Paired Samples Correlations τώρα θα πρέπει να ελέγξουμε την υπόθεση της συσχέτισης δηλαδή εάν υπάρχει συσχέτιση μεταξύ των δυο δειγμάτων ή δεν υπάρχει και τα δείγματα είναι ασυσχέτιστα όποτε δεν έχει και νόημα ο έλεγχος εξαρτημένων δειγμάτων.

Ο έλεγχος διαμορφώνεται ως εξής:

$$\mathbf{H_0: \rho = 0 \qquad H_1: \rho \neq 0}$$

δηλαδή ο συντελεστής συσχέτισης ισούται με το μηδέν και δεν υπάρχει συσχέτιση, έναντι του γεγονότος ότι ο συντελεστής συσχέτισης είναι διάφορος του μηδενός όποτε και υφίσταται η συσχέτιση.

Συγκρίνοντας την τιμή sig = 0.008 με το επίπεδο σημαντικότητας α= 0.05 παρατηρούμε ότι sig < α οπότε απορρίπτω την υπόθεση H_0 και θεωρώ ότι υπάρχει συσχέτιση ανάμεσα σε αυτές τις δυο μεταβλητές.

Το γεγονός αυτό επιβεβαιώνεται και από την τιμή του συντελεστή συσχέτισης Correlation 0.697 που δίνεται επίσης από τον πίνακα Paired Samples Correlations και υποδηλώνει μέτρια συσχέτιση μεταξύ αυτών των μεταβλητών.

Επομένως μπορούμε να προχωρήσουμε με τον έλεγχο των μέσων.

Paired Samples Test

		Paired Differences					t	df	Sig. (2tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Accidents_before - Accidents_after	2,538	2,727	,756	,891	4,186	3,356	12	,006

Ο πίνακας **PAIRED SAMPLES TEST** δίνει τις **διαφορές (paired differences):**

• των μέσων **(mean)**

• των τυπικών αποκλίσεων **(std. deviaiton)**

- των τυπικών σφαλμάτων των μέσων **(std. Error of mean)** και
- τα όρια **(Upper-Lower)** του διαστήματος εμπιστοσύνης της διαφοράς των μέσων **(95% Confidence Interval for Means)**

Προσοχή μεγάλη θα πρέπει να δοθεί στο γεγονός ότι **ο έλεγχος μας είναι μονόπλευρος**, οπότε θα πρέπει πρώτα να ελέγξουμε αν οι μέσοι ικανοποιούν την ανισότητα της H_1 δηλαδή εάν,

$$mean_{\Pi} = 6{,}38 > mean_{A} = 3{,}85$$

Αφού οι μέσοι ικανοποιούν την ανισότητα της H_1 θα συγκρίνουμε την τιμή sig/2 από τον πίνακα **Paired Samples Test** με το επίπεδο σημαντικότητας

$$sig/2 = 0{,}006/2 = 0{,}003 < \alpha = 0.05$$

Συνεπώς απορρίπτουμε την H_0 και έτσι φαίνεται ότι ο αριθμός των ατυχημάτων όντως μειώθηκε με την είσοδο του φωτισμού σε επίπεδο σημαντικότητας $\alpha = 0.05$.

Άσκηση

Ο παρακάτω πίνακας δείχνει τον αριθμό των ελαττωματικών προϊόντων που παράγονται από μια εταιρεία το πρωί και το απόγευμα σε διάστημα 4 ημερών. Να ελεγχθεί σε επίπεδο σημαντικότητας 5% αν παράγονται περισσότερα ελαττωματικά προϊόντα το πρωί.

ΗΜΕΡΕΣ	ΠΡΩΙ	ΑΠΟΓΕΥΜΑ
1	10	8
2	12	9
3	15	12
4	19	15

Απάντηση :
Από τον έλεγχο προκύπτει ότι το πρωί παράγονται περισσότερα ελαττωματικά προϊόντα.

Επαναληπτικές ασκήσεις στον έλεγχο υποθέσεων

1. Σε μια βιομηχανία θεωρείται ότι ο μέσος μισθός ισούται με 920 €. Για να το εξακριβώσουμε, πήραμε ένα δείγμα 25 εργαζομένων και καταγράψαμε τους μισθούς τους. Να ελεγχθεί αν ο μέσος μισθός υπερβαίνει τα 920 €.

Αν οι πρώτες 13 παρατηρήσεις αφορούν μισθούς γυναικών και οι υπόλοιπες μισθούς ανδρών, να εξετάσετε αν υπάρχει μισθολογική διαφορά ανάμεσα σε άνδρες και γυναίκες.(α=5%)

Α/Α	ΜΙΣΘΟΣ ΣΕ €
1	1020
2	850
3	1000
4	963
5	896
6	759
7	914
8	689
9	856
10	1120
11	985
12	741
13	698
14	859
15	926
16	990
17	900
18	1008
19	741
20	623
21	1258
22	950
23	989
24	963
25	1000

Απάντηση

Από τον έλεγχο συμπεραίνουμε ότι ο μέσος μισθός δεν υπερβαίνει αλλά θεωρείται ίσος με 920€, ενώ δεν υπάρχει μισθολογική διαφορά μεταξύ γυναικών και ανδρών.

2. Μια μελέτη διεξήχθη για να διερευνήσει αν η βρώμη και ο αραβόσιτος βοηθούν στη μείωση της χοληστερόλης. Επελέγησαν τυχαία 14 άνδρες με υψηλή χοληστερόλη και υποβλήθηκαν σε δύο δίαιτες. Η πρώτη δίαιτα περιλάμβανε πρωινό με βρώμη και η δεύτερη με αραβόσιτο. Μετά το τέλος κάθε δίαιτας μετρήθηκαν τα επίπεδα της LDL χοληστερόλης στο αίμα τους και τα αποτελέσματα δίνονται στον ακόλουθο πίνακα.

Να ελεγχθεί σε επίπεδο σημαντικότητας α=5% αν τα επίπεδα χοληστερόλης είναι χαμηλότερα μετά την πρώτη δίαιτα.

	LDL ΣΕ MMOL/LT	
ΑΤΟΜΟ	ΒΡΩΜΗ	ΑΡΑΒΟΣΙΤΟΣ
1	3,84	4,61
2	5,57	6,42
3	5,85	5,4
4	4,8	4,54
5	3,68	3,98
6	2,96	3,82
7	4,41	5,01
8	3,72	4,34
9	3,49	3,8
10	3,84	4,56
11	5,26	5,35
12	3,73	3,89
13	1,84	2,25
14	4,14	4,24

Απάντηση

Από τον έλεγχο προκύπτει ότι, πράγματι, η βρώμη βοηθάει περισσότερο στην ελάττωση των επιπέδων χοληστερόλης στο αίμα.

6. ΣΥΣΧΕΤΙΣΗ ΠΟΣΟΤΙΚΩΝ ΜΕΤΑΒΛΗΤΩΝ

Η **συσχέτιση** αναφέρεται στη διερεύνηση της σχέσης ανάμεσα σε δύο ποσοτικές μεταβλητές. Από τη μελέτη των χαρακτηριστικών που αντιπροσωπεύουν οι μεταβλητές, συνήθως διαχωρίζονται σε:

- Ανεξάρτητη μεταβλητή, δηλαδή η μεταβλητή της οποίας οι τιμές μεταβάλλονται χωρίς να επηρεάζονται από την εξέλιξη των τιμών της άλλης
- Εξαρτημένη μεταβλητή, δηλαδή η μεταβλητή της οποίας οι τιμές πιθανόν να μεταβάλλονται από την εξέλιξη των τιμών της άλλης.

Η σχέση μεταξύ των τιμών των δύο μεταβλητών, δηλαδή το **είδος** και ο **βαθμός** της συσχέτισης ελέγχονται αντίστοιχα στο SPSS μέσω:

- του **διαγράμματος διασποράς** του σμήνους των σημείων που αντιστοιχούν στις δύο μεταβλητές
- του **συντελεστή συσχέτιση Pearson**

Απαραίτητη προϋπόθεση είναι οι τιμές να προέρχονται από κανονικά κατανεμημένο πληθυσμό.

Τα βήματα που ακολουθούμε στο SPSS είναι τα παρακάτω :

1. Δημιουργούμε δύο μεταβλητές στο Variable View και εισάγουμε τα δεδομένα στο Data View.

Για το διάγραμμα διασποράς

2. Επιλέγουμε :

   Graphs → Legacy Dialogs → Scatter/Dot

3. Στο παράθυρο που ανοίγει επιλέγουμε Simple Scatter και Define.
4. Μεταφέρουμε την ανεξάρτητη μεταβλητή στον άξονα X και την εξαρτημένη μεταβλητή στον άξονα Y. Πατάμε OK.

Για τον συντελεστή Pearson

5. Επιλέγουμε :

Analyze → Correlate → Bivariate

6. Στο Variables μεταφέρουμε και τις δύο μεταβλητές μας.
7. Κλικάρουμε Pearson, στο Test of Significance το Two-tailed, και Flag significant correlations (συνήθως είναι επιλεγμένα). Πατάμε ΟΚ.

Αποτελέσματα στο Output :

● **Διάγραμμα Διασποράς (Graph)**. Στο διάγραμμα αυτό εμφανίζονται τα ζεύγη τιμών ως κυκλάκια και εξετάζουμε τη σχετική θέση τους:

- Αν υπάρχει μία ευθεία γραμμή γύρω από την οποία συγκεντρώνονται τα κυκλάκια, τότε η σχέση μεταξύ των δύο μεταβλητών είναι **γραμμική**.
- Αν υπάρχει καμπύλη (π.χ. παραβολή) γύρω από την οποία συγκεντρώνονται τα κυκλάκια, τότε η σχέση μεταξύ των δύο μεταβλητών καθορίζεται από τη μορφή της καμπύλης.
- Αν δεν υπάρχει καμπύλη γύρω από την οποία συγκεντρώνονται τα κυκλάκια, δηλαδή αυτά βρίσκονται τυχαία κατανεμημένα στο διάγραμμα, τότε οι δύο μεταβλητές είναι **ασυσχέτιστες**.

Στην περίπτωση που έχουμε γραμμική συσχέτιση των δύο μεταβλητών, μπορούμε σε ορισμένες περιπτώσεις να καθορίσουμε περαιτέρω το είδος της συσχέτισης:

- Αν ο συντελεστής διεύθυνσης της ευθείας γραμμής γύρω από την οποία συγκεντρώνονται τα κυκλάκια είναι θετικός, τότε έχουμε **θετική γραμμική συσχέτιση**
- Αν ο συντελεστής διεύθυνσης της ευθείας γραμμής γύρω από την οποία συγκεντρώνονται τα κυκλάκια είναι αρνητικός, τότε έχουμε **αρνητική γραμμική συσχέτιση**

● **Πίνακας Correlations.** Μας δίνει το μέτρο του βαθμού συσχέτισης, το οποίο είναι αξιόπιστο **μόνο** στην περίπτωση γραμμικής συσχέτισης, δηλαδή τον **συντελεστή συσχέτισης Pearson** ο οποίος εμφανίζεται στο άνω δεξιό τμήμα του πίνακα και παίρνει τιμές από -1 έως +1.

- Αν ο συντελεστής συσχέτισης είναι θετικός, τότε έχουμε **θετική γραμμική συσχέτιση**, δηλαδή όταν οι τιμές της μίας μεταβλητής αυξάνονται, αυξάνονται και της άλλης
- Αν ο συντελεστής συσχέτισης είναι αρνητικός, τότε έχουμε **αρνητική γραμμική συσχέτιση**, δηλαδή όταν οι τιμές της μίας μεταβλητής αυξάνονται, οι τιμές της άλλης μειώνονται.

Επιπλέον, όσο μεγαλύτερη είναι η απόλυτη τιμή του συντελεστή αυτού, τόσο πιο ισχυρή είναι η σχέση των δύο μεταβλητών καθώς και η δυνατότητα πρόβλεψης της εξαρτημένης μεταβλητής με βάση την ανεξάρτητη.

Εάν η απόλυτη τιμή του συντελεστή συσχέτισης Pearson βρίσκεται:

- στο διάστημα [0, 0.2], τότε η συσχέτιση χαρακτηρίζεται **ασήμαντη**,
- στο διάστημα [0.2, 0.4], τότε η συσχέτιση χαρακτηρίζεται **μέτρια**,

- στο διάστημα [0.4, 0.7], τότε η συσχέτιση χαρακτηρίζεται **σημαντική**,
- στο διάστημα [0.7, 1], τότε η συσχέτιση χαρακτηρίζεται **ισχυρή**.

Εφαρμογή της διαδικασίας

Στο παράδειγμα που ακολουθεί περιγράφεται αναλυτικά η διαδικασία ελέγχου του είδους και του βαθμού συσχέτισης μεταξύ δύο μεταβλητών.

Δίνεται η βαθμολογία 10 μαθητών στα Μαθηματικά και στη Φυσική. Να γίνει έλεγχος συσχέτισης.

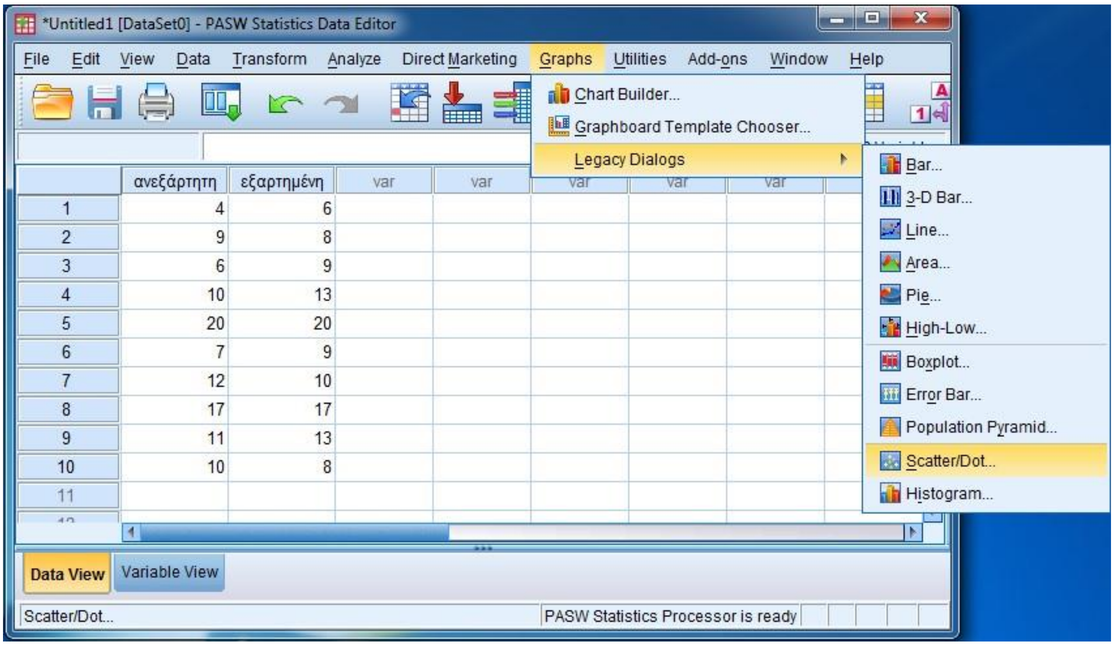
Μαθηματικά	4	9	6	10	20	7	12	17	11	10
Φυσική	6	8	9	13	20	9	10	17	13	8

Στο συγκεκριμένο παράδειγμα μπορούμε να θεωρήσουμε οποιαδήποτε από τις δύο μεταβλητές ως ανεξάρτητη (π.χ. βαθμούς στα Μαθηματικά) και την άλλη ως εξαρτημένη (π.χ. βαθμούς στη Φυσική).

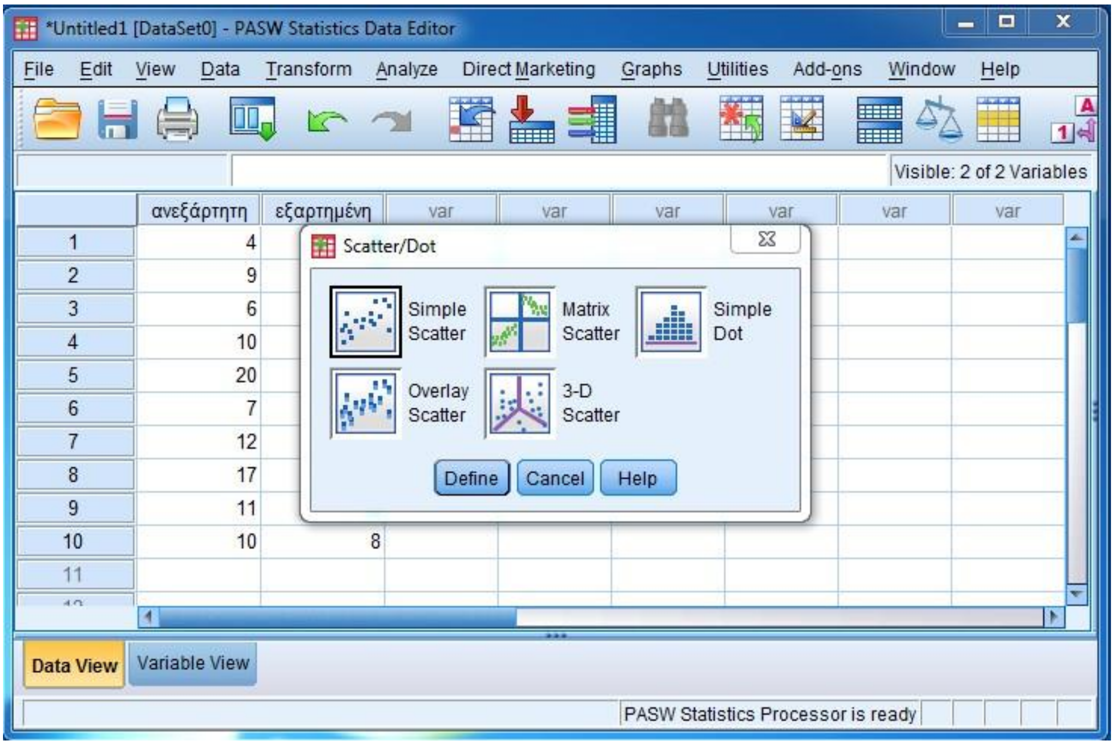
(Θα κάνουμε πρώτα έλεγχο κανονικότητας, ο οποίος εδώ δεν εμφανίζεται για λόγους συντομίας)

Τα βήματα που ακολουθούμε είναι:

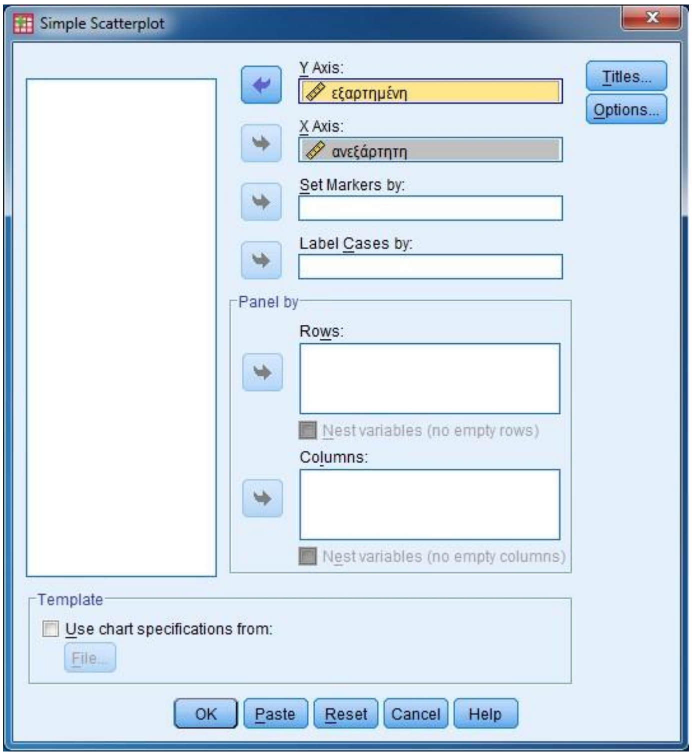
1. Από το κεντρικό παράθυρο διαλόγου επιλέγουμε:



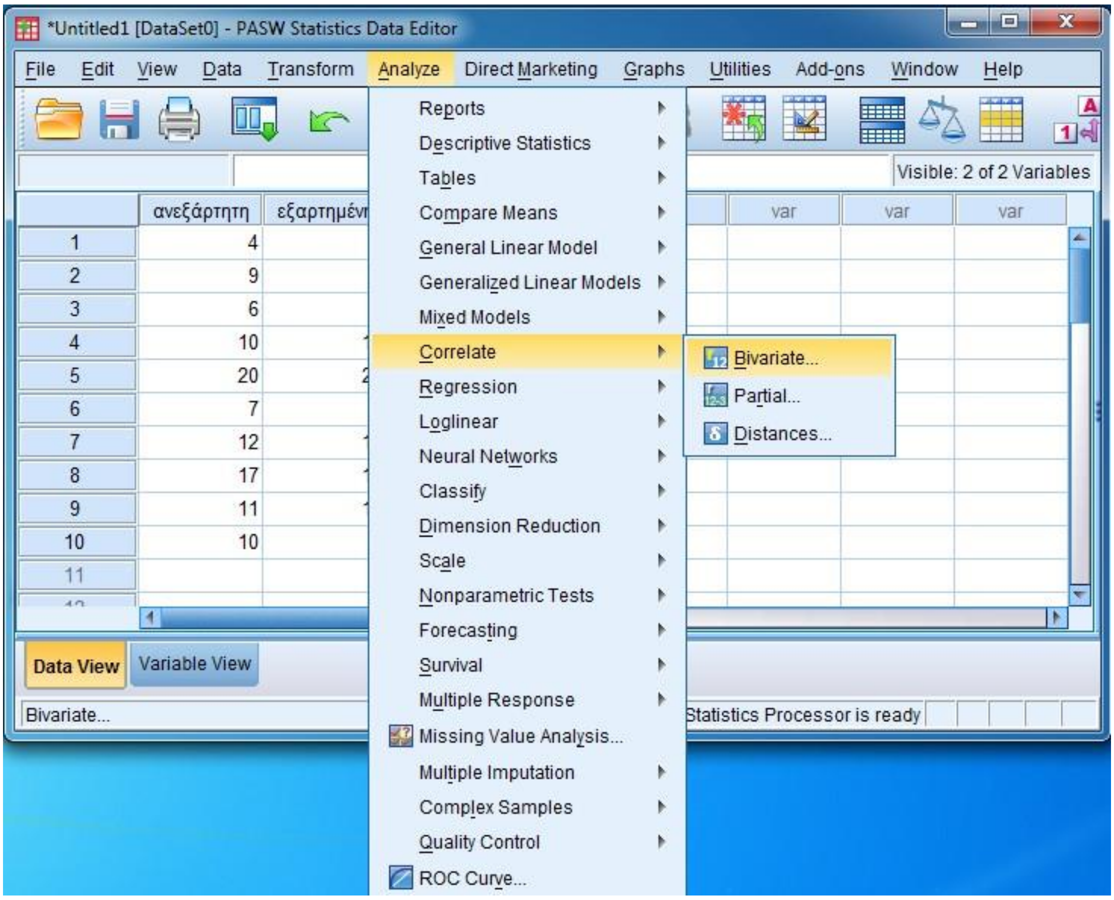
2. Στο παράθυρο διαλόγου που προκύπτει επιλέγουμε Simple Scatter και Define.



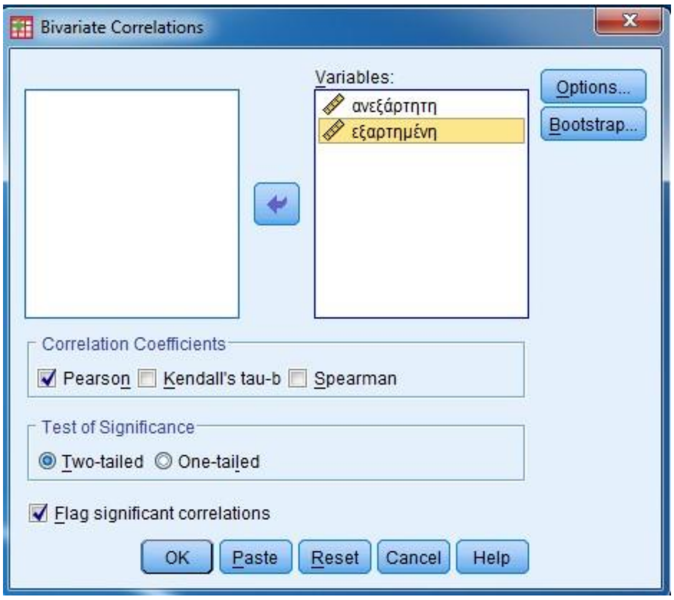
3. Μεταφέρουμε την ανεξάρτητη μεταβλητή στον άξονα X και την εξαρτημένη μεταβλητή στον άξονα Y. Πατάμε OK.



4. Από το κεντρικό παράθυρο διαλόγου επιλέγουμε:

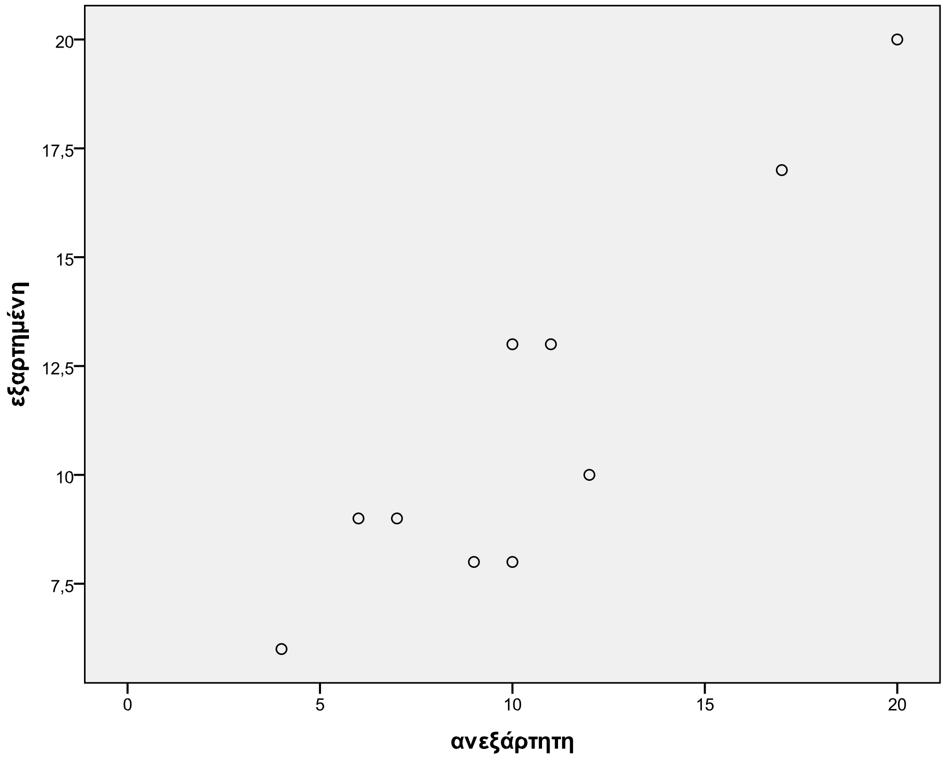


5. Στο Variables μεταφέρουμε και τις δύο μεταβλητές μας και πατάμε OK



Ερμηνεία αποτελεσμάτων:

Διάγραμμα Διασποράς



Οι δύο μεταβλητές έχουν θετική γραμμική συσχέτιση καθώς υπάρχει ευθεία (με θετικό συντελεστή διεύθυνσης) γύρω από την οποία φαίνεται να συγκεντρώνονται οι παρατηρήσεις.

Στον πίνακα Correlations μας δίνεται η τιμή του συντελεστή συσχέτισης Pearson, που είναι 0.916, άρα πράγματι έχουμε θετική γραμμική ισχυρή συσχέτιση.

Correlations

		ανεξάρτητη	εξαρτημένη
ανεξάρτητη	Pearson Correlation	1	,916**
	Sig. (2-tailed)		,000
	N	10	10
εξαρτημένη	Pearson Correlation	,916**	1
	Sig. (2-tailed)	,000	
	N	10	10

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

7. ΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Στην περίπτωση που έχουμε δύο μεταβλητές μεταξύ των οποίων υπάρχει γραμμική συσχέτιση (σημαντική ή ισχυρή) μπορούμε να προσδιορίσουμε την εξίσωση της βέλτιστης ευθείας γύρω από την οποία συγκεντρώνονται οι παρατηρήσεις, καθώς και μερικά επιπλέον στοιχεία για τη σύνδεσή τους και τη δυνατότητα χρήσης της ευθείας για πρόβλεψη των τιμών της εξαρτημένης μεταβλητής.

Τα βήματα που ακολουθούμε στο SPSS είναι τα παρακάτω :

1. Δημιουργούμε τις δύο μεταβλητές στο Variable View και εισάγουμε τα δεδομένα στο Data View.
2. Χαρακτηρίζουμε τις μεταβλητές ως «ανεξάρτητη» και «εξαρτημένη».
3. Κάνουμε έλεγχο συσχέτισης για να ελέγξουμε εάν οι μεταβλητές μας συνδέονται με γραμμική σχέση (σημαντική ή ισχυρή, για να μπορούμε να προχωρήσουμε τη διαδικασία).

   Στο διάγραμμα διασποράς μπορούμε να εμφανίσουμε την ευθεία παλινδρόμησης, ακολουθώντας τα βήματα:

   - Κάνουμε διπλό κλικ πάνω στο διάγραμμα διασποράς και ανοίγει το παράθυρο Chart Editor.
   - Στο Chart Editor κάνουμε διπλό κλικ σε ένα σημείο του διαγράμματος και όλα τα σημεία αλλάζουν χρώμα.
   - Επιλέγουμε

     Elements → Fit line at total.

   - Κλείνουμε το Chart Editor και βλέπουμε πάνω στο διάγραμμα διασποράς σχεδιασμένη την ευθεία παλινδρόμησης.
4. Επιλέγουμε

   Analyze → Regression → Linear
5. Στο παράθυρο Dependent μεταφέρουμε την εξαρτημένη μεταβλητή και στο Independent την ανεξάρτητη μεταβλητή.
6. Επιλέγουμε Statistics και κλικάρουμε Estimates, Model fit και Durbin-Watson. Πατάμε Continue.
7. Επιλέγουμε Plots. Στο Y μεταφέρουμε το ZRESID και στο X το ZPRED. Κλικάρουμε Histogram και Normal probability plot. Πατάμε Continue.
8. Επιλέγουμε Save και στο Predicted values κλικάρουμε Unstandardized ενώ στο Residuals κλικάρουμε Unstandardized και Standardized. Πατάμε Continue και OK.

Αποτελέσματα στο Output

Στη συνέχεια θα αναλύσουμε τα βασικά συμπεράσματα που προκύπτουν από το Output

● Πίνακας Model Summary

Στα στοιχεία αυτού του πίνακα υπάρχουν τα ακόλουθα βασικά αριθμητικά μέτρα:

- R, που είναι η απόλυτη τιμή του συντελεστή συσχέτισης Pearson και μας δείχνει το βαθμό συσχέτισης των δύο μεταβλητών
- R Square, που ισούται με το τετράγωνο του R και καλείται δείκτης προσδιορισμού. Ο δείκτης αυτός διαβάζεται ως ποσοστό και εκφράζει το ποσοστό μεταβλητότητας της εξαρτημένης μεταβλητής που οφείλεται στην ανεξάρτητη.

● Πίνακας Anova

Ο πίνακας αυτός κάνει ανάλυση διασποράς του μοντέλου και αρχικά μας ενδιαφέρει ο αριθμός Sig. Ειδικότερα:

- Αν sig μικρότερο από το επίπεδο σημαντικότητας τότε το μοντέλο της παλινδρόμησης είναι κατάλληλο για πρόβλεψη.
- Αν Sig μεγαλύτερο από το επίπεδο σημαντικότητας τότε το μοντέλο της παλινδρόμησης δεν είναι κατάλληλο για πρόβλεψη.

● Πίνακας Coefficients

Η κύρια στήλη του πίνακα αυτού είναι η **στήλη B (Unstandardized Coefficients)**, η οποία μας δίνει τους συντελεστές της ευθείας παλινδρόμησης:

- Στην πρώτη γραμμή βρίσκεται ο σταθερός όρος της ευθείας, ο οποίος εκφράζει την αναμενόμενη τιμή της εξαρτημένης μεταβλητής όταν η ανεξάρτητη είναι (ή τείνει να πάρει την τιμή) μηδέν.
- Στην δεύτερη γραμμή βρίσκεται ο συντελεστής διεύθυνσης της ευθείας, ο οποίος καλείται και συντελεστής παλινδρόμησης, ο οποίος εκφράζει την μεταβολή (αύξηση ή μείωση) της εξαρτημένης μεταβλητής όταν η ανεξάρτητη αυξηθεί κατά μία μονάδα.

Στην τελευταία στήλη του πίνακα οι τιμές Sig χρησιμοποιούνται αντίστοιχα για τον έλεγχο σημαντικότητας των συντελεστών της ευθείας. Ειδικότερα:

- Αν sig μικρότερο από το επίπεδο σημαντικότητας τότε ο αντίστοιχος συντελεστής είναι στατιστικά σημαντικός.
- Αν Sig μεγαλύτερο από το επίπεδο σημαντικότητας τότε ο αντίστοιχος συντελεστής δεν είναι στατιστικά σημαντικός.

● Πίνακας Residuals Statistics

Στον πίνακα αυτό βλέπουμε κυρίως τη μέγιστη και ελάχιστη απόκλιση των προβλεπόμενων τιμών (με βάση την ευθεία παλινδρόμησης) από τις παρατηρούμενες τιμές.

● Διαγράμματα Histogram, Normal P-P Plot και Scatterplot

Μας δίνουν πληροφορίες για την κανονικότητα.

● Data View

Στην οθόνη των δεδομένων έχουν προστεθεί τρεις στήλες. Από αυτές:

- η στήλη PRE_1 μας δίνει για κάθε παρατήρηση την προβλεπόμενη από το μοντέλο τιμή της εξαρτημένης μεταβλητής
- η στήλη RES_1 μας δίνει για κάθε παρατήρηση τη διαφορά της προβλεπόμενης από την πραγματική τιμή.

Εφαρμογή της διαδικασίας

Στο παράδειγμα που ακολουθεί περιγράφεται αναλυτικά η διαδικασία εύρεσης της ευθείας παλινδρόμησης και η ερμηνεία των αποτελεσμάτων.

Στον παρακάτω πίνακα δίνονται ο μισθός και η προϋπηρεσία 16 υπαλλήλων μιας εταιρείας.

ΕΤΗ ΠΡΟΥΠΗΡΕΣΙΑΣ	ΜΙΣΘΟΣ ΣΕ ΕΥΡΩ
6	800
8	830
12	850
15	900
20	950
5	750
7	780
8	760
7	770
10	810
13	820
1	580
3	600
15	900
16	920
18	930

I. Να ελεγχθεί η συσχέτιση των δύο μεταβλητών

II. Να δοθεί το κατάλληλο μοντέλο γραμμικής παλινδρόμησης και να ερμηνευτούν οι συντελεστές

III. Να γίνει έλεγχος σημαντικότητας των συντελεστών

IV. Το μοντέλο είναι κατάλληλο για πρόβλεψη;

V. Να ερμηνευθεί ο συντελεστής προσδιορισμού

VI. Να ερμηνευθεί η τέταρτη γραμμή της οθόνης δεδομένων

VII. Διατυπώστε ένα συμπέρασμα με βάση των πίνακα υπολοίπων

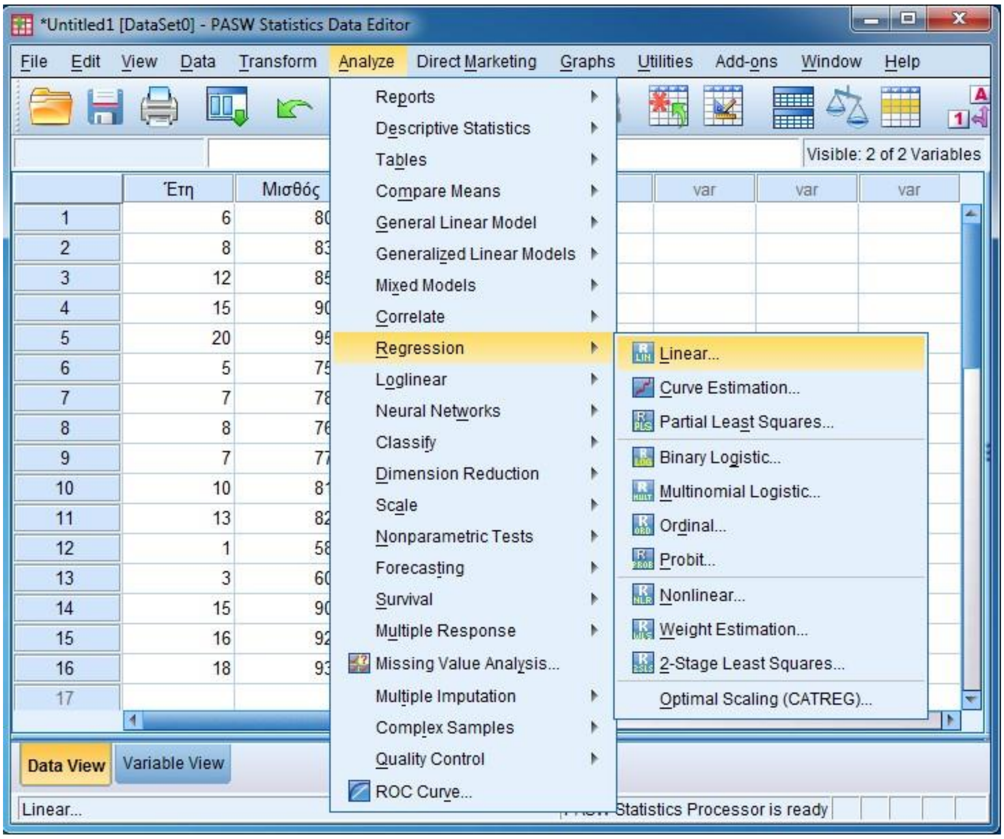
Χαρακτηρίζουμε αρχικά τις μεταβλητές μας:

- Ανεξάρτητη μεταβλητή είναι η προϋπηρεσία
- Εξαρτημένη μεταβλητή είναι ο μισθός

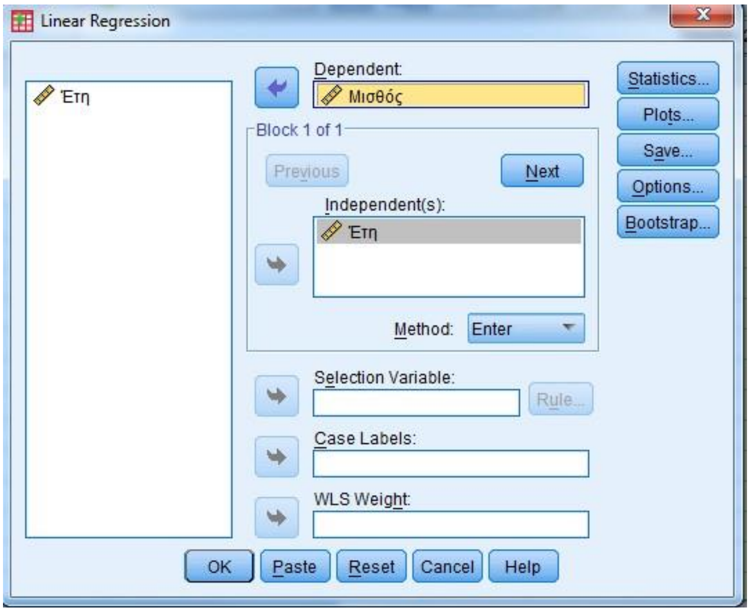
Για το ερώτημα (I) ακολουθούμε τα βήματα που έχουμε περιγράψει σε προηγούμενη ενότητα και βρίσκουμε ότι οι μεταβλητές μας είναι θετικά γραμμικά συσχετισμένες και ότι ο συντελεστής συσχέτισης Pearson είναι 0.926, άρα έχουμε ισχυρή συσχέτιση.

Τα βήματα που ακολουθούμε στη συνέχεια είναι:

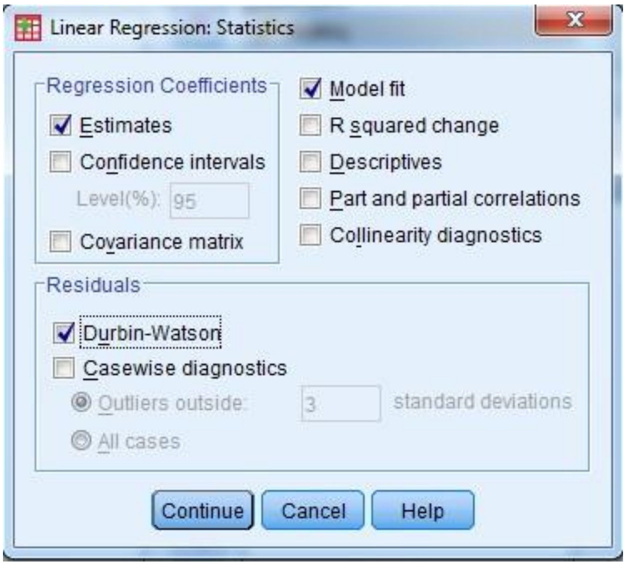
1. Από το κεντρικό παράθυρο διαλόγου επιλέγουμε:



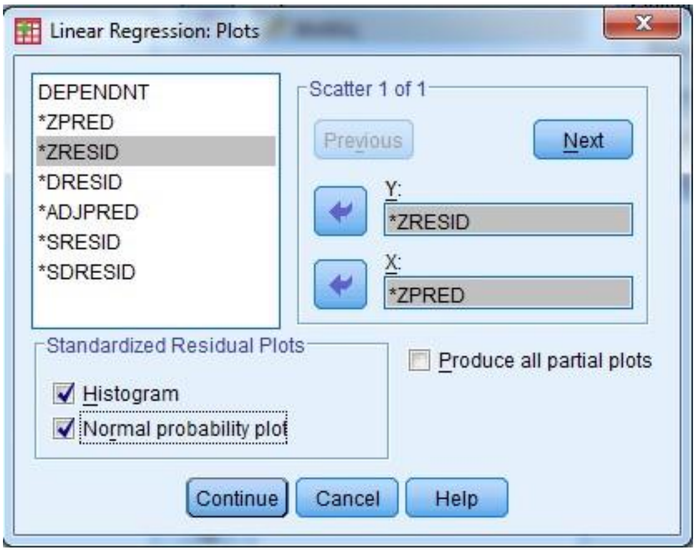
2. Στο παράθυρο διαλόγου που προκύπτει μεταφέρουμε στο πλαίσιο Dependent την εξαρτημένη μεταβλητή και στο Independent την ανεξάρτητη μεταβλητή.



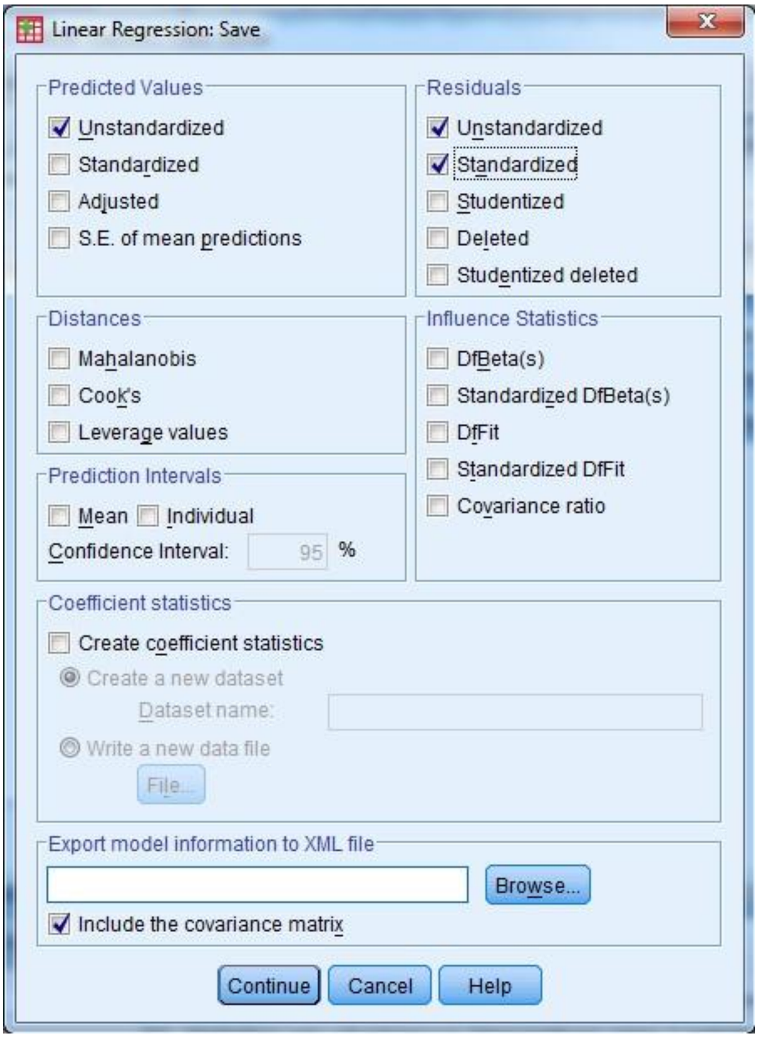
3. Επιλέγουμε Statistics και κλικάρουμε Estimates, Model fit και Durbin-Watson. Πατάμε Continue.



4. Επιλέγουμε Plots. Στο Y μεταφέρουμε το ZRESID και στο X το ZPRED. Κλικάρουμε Histogram και Normal probability plot. Πατάμε Continue.



5. Επιλέγουμε Save και στο Predicted values κλικάρουμε Unstandardized ενώ στο Residuals κλικάρουμε Unstandardized και Standardized. Πατάμε Continue και ΟΚ.


Linear Regression: Save
Predicted Values
Unstandardized
Standardized
Adjusted
S.E. of mean predictions
Residuals
Unstandardized
Standardized
Studentized
Deleted
Studentized deleted
Distances
Mahalanobis
Cook's
Leverage values
Influence Statistics
DfBeta(s)
Standardized DfBeta(s)
DfFit
Standardized DfFit
Covariance ratio
Prediction Intervals
Mean
Individual
Confidence Interval:
95
%
Coefficient statistics
Create coefficient statistics
Create a new dataset
Dataset name:
Write a new data file
File...
Export model information to XML file
Browse...
Include the covariance matrix
Continue
Cancel
Help

Ερμηνεία αποτελεσμάτων (με βάση τους πίνακες και τα ερωτήματα της άσκησης)

Ερωτήματα (II) και (III)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	626,650	22,455		27,907	,000
	Έτη	17,827	1,942	,926	9,181	,000

a. Dependent Variable: Μισθός

Από τον πίνακα Coefficients βρίσκουμε ότι η ευθεία γραμμικής παλινδρόμησης έχει εξίσωση

$$y= 17.827x + 626.650$$

όπου y είναι η εξαρτημένη μεταβλητή (μισθός) και x είναι η ανεξάρτητη μεταβλητή (έτη).

Ο συντελεστής παλινδρόμησης 17.827 εκφράζει ότι και κάθε επιπλέον χρόνο προϋπηρεσίας αναμένεται ο μισθός των εργαζομένων της εταιρείας να αυξάνεται κατά 17.827 ευρώ, κατά μέσο όρο.

Ο σταθερός συντελεστής 626.650 εκφράζει ότι με μηδενική προϋπηρεσία αναμένεται ένας υπάλληλος να έχει μισθό 626.65 ευρώ, κατά μέσο όρο.

Και οι δύο συντελεστές είναι στατιστικά σημαντικοί διότι Sig =0 (και για τους δύο).

Ερώτημα (IV)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	145867,770	1	145867,770	84,296	,000[a]
	Residual	24225,980	14	1730,427		
	Total	170093,750	15			

a. Predictors: (Constant), Έτη

b. Dependent Variable: Μισθός

Αφού ο δείκτης Sig του πίνακα ANOVA είναι μηδέν, το μοντέλο είναι κατάλληλο να χρησιμοποιηθεί για πρόβλεψη.

Ερώτημα (V)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,926 [a]	,858	,847	41,598	,911

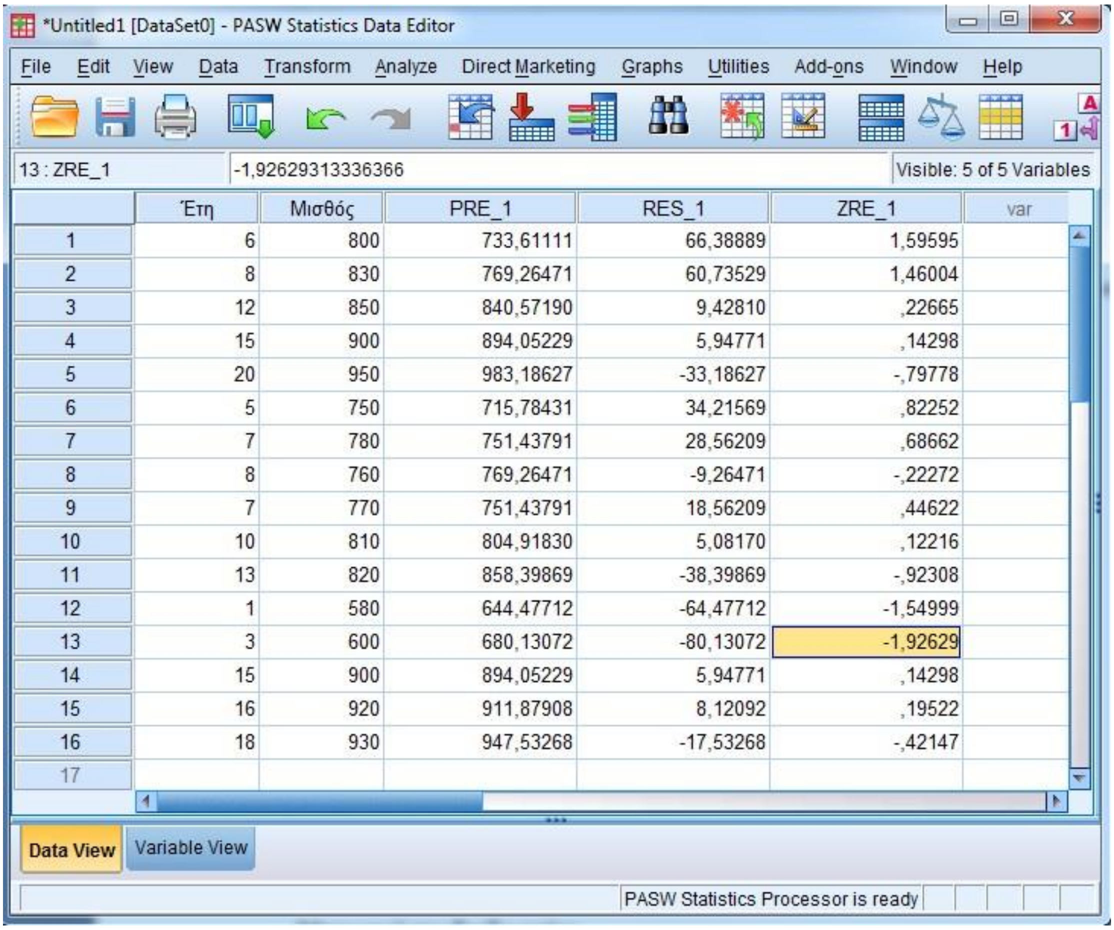
a. Predictors: (Constant), Έτη

b. Dependent Variable: Μισθός

Ο δείκτης προσδιορισμού είναι 85,8%, δηλαδή ο μισθός εξαρτάται κατά 85,8% από τα έτη προϋπηρεσίας.

Ερώτημα (VI)

Η οθόνη δεδομένων είναι



	Έτη	Μισθός	PRE_1	RES_1	ZRE_1	var
1	6	800	733,61111	66,38889	1,59595	
2	8	830	769,26471	60,73529	1,46004	
3	12	850	840,57190	9,42810	,22665	
4	15	900	894,05229	5,94771	,14298	
5	20	950	983,18627	-33,18627	-,79778	
6	5	750	715,78431	34,21569	,82252	
7	7	780	751,43791	28,56209	,68662	
8	8	760	769,26471	-9,26471	-,22272	
9	7	770	751,43791	18,56209	,44622	
10	10	810	804,91830	5,08170	,12216	
11	13	820	858,39869	-38,39869	-,92308	
12	1	580	644,47712	-64,47712	-1,54999	
13	3	600	680,13072	-80,13072	-1,92629	
14	15	900	894,05229	5,94771	,14298	
15	16	920	911,87908	8,12092	,19522	
16	18	930	947,53268	-17,53268	-,42147	
17						

Στην τέταρτη γραμμή βλέπουμε ότι ένας υπάλληλος με 15 έτη προϋπηρεσία έχει μισθό 900 ευρώ, ενώ με βάση το μοντέλο θα έπρεπε να έχει μισθό 894,05 ευρώ, δηλαδή αμείβεται κατά 5,95 ευρώ περισσότερα.

Ερώτημα (VII)

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	644,48	983,19	809,38	98,613	16
Residual	-80,131	66,389	,000	40,188	16
Std. Predicted Value	-1,672	1,763	,000	1,000	16
Std. Residual	-1,926	1,596	,000	,966	16

a. Dependent Variable: Μισθός

Ένα συμπέρασμα από τον πίνακα υπολοίπων είναι ότι ο χαμηλότερος μισθός που προβλέπει το μοντέλο (για τα δεδομένα μας) είναι 644,48 ευρώ, ενώ ο μέγιστος 983,19 ευρώ.

Άσκηση

1. Στον παρακάτω πίνακα φαίνεται η πορεία των κερδών μιας επιχείρησης στους 15 πρώτους μήνες λειτουργίας της. Να γίνει το διάγραμμα διασποράς και η ανάλυση παλινδρόμησης. Τι συμπεράσματα προκύπτουν;

Χ (ΜΗΝΑΣ)	Υ (ΚΕΡΔΗ ΣΕ ΧΙΛΙΑΔΕΣ ΕΥΡΩ)
1	35
2	32
3	42
4	31
5	28
6	20
7	17
8	15
9	10
10	12
11	9
12	7
13	8
14	11
15	8

Απάντηση :

Ευθεία παλινδρόμησης y=-2,361 x + 37,886