

# Κεφάλαιο 3: Ανάλυση μιας μεταβλητής

---

## **Γενικά**

Στο Κεφάλαιο αυτό θα παρουσιάσουμε κάποιες μεθόδους της Περιγραφικής Στατιστικής και της Στατιστικής Συμπερασματολογίας που αφορούν στην ανάλυση μιας μεταβλητής. Η Περιγραφική Στατιστική (Descriptive Statistics) περιλαμβάνει όλες τις μεθόδους για τη συνόψιση, περιγραφή και παρουσίαση των δεδομένων. Περιλαμβάνει, δηλαδή, όλα τα απαραίτητα εργαλεία για την οργάνωση των δεδομένων και τη συνοπτική και αποτελεσματική παρουσίασή τους. Η Στατιστική Συμπερασματολογία (ή αλλιώς Επαγωγική Στατιστική) (Inductive Statistics ή Statistical Inference) περιλαμβάνει μεθόδους για την ανάλυση και την εξαγωγή συμπερασμάτων για τον συνολικό μας πληθυσμό, όταν αυτό που έχουμε στα χέρια μας είναι ένα δείγμα. Στην επόμενη παράγραφο θα αναλύσουμε κάποιες μεθόδους της Περιγραφικής Στατιστικής για την ανάλυση μίας ποσοτικής ή ποιοτικής μεταβλητής, και στη συνέχεια θα ασχοληθούμε με κάποιους ελέγχους της Στατιστικής Συμπερασματολογίας που αφορούν μια μεταβλητή.

## **Περιγραφική Στατιστική**

### **Ποιοτικές μεταβλητές**

Οι μέθοδοι οργάνωσης και παρουσίασης ποιοτικών δεδομένων συνοψίζονται στους πίνακες συχνοτήτων και σε γραφήματα που παράγονται από τη διαδικασία `Frequencies`.

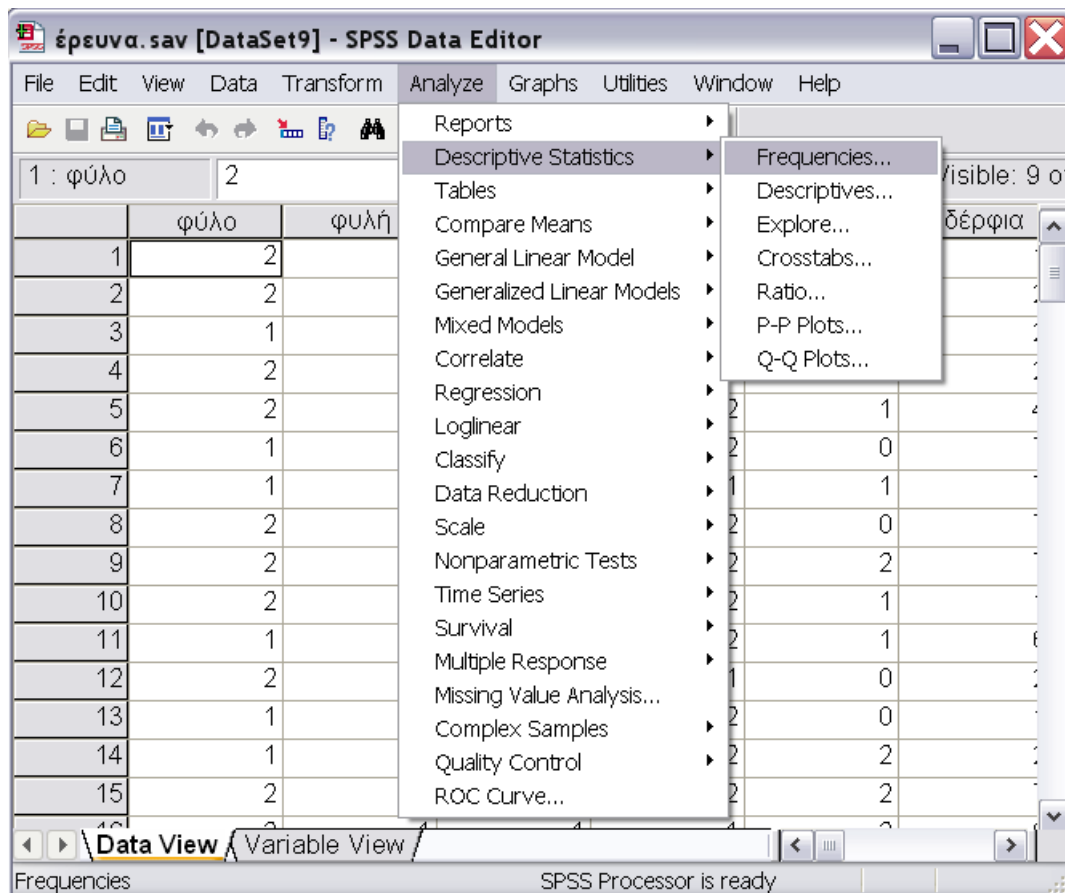
### **Η διαδικασία FREQUENCIES**

#### **Παράδειγμα 3.1**

Θα χρησιμοποιήσουμε τα δεδομένα του αρχείου `ερευνα.sav` και τη διαδικασία `Frequencies` για την ανάλυση της ποιοτικής μεταβλητής φυλή.

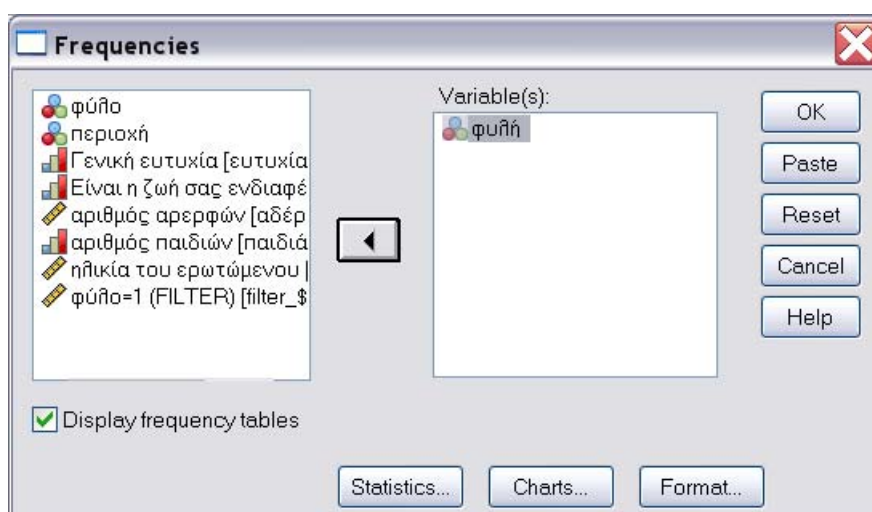
Από το κύριο μενού των εντολών επιλέγουμε διαδοχικά (Εικόνα 3.1):

Analyze → Descriptive Statistics → Frequencies ...



**Εικόνα 3.1** Η διαδικασία Frequencies

Επιλέγουμε τη μεταβλητή φυλή και τη μετακινούμε στο παράθυρο Variable(s) (Εικόνα 3.2).



**Εικόνα 3.2** Το παράθυρο διαλόγου Frequencies

Εξ ορισμού, η διαδικασία θα υπολογίσει μόνο τον πίνακα συχνοτήτων (Εικόνα 3.3):

### Statistics

φυλή

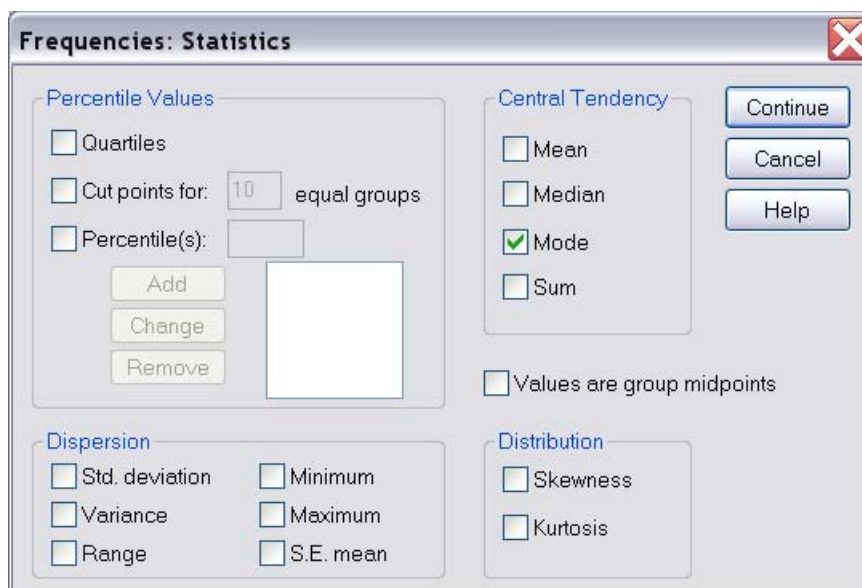
N	Valid	100
	Missing	0

### φυλή

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid λευκή	68	68,0	68,0	68,0
μαύρη	22	22,0	22,0	90,0
άλλη	10	10,0	10,0	100,0
Total	100	100,0	100,0	

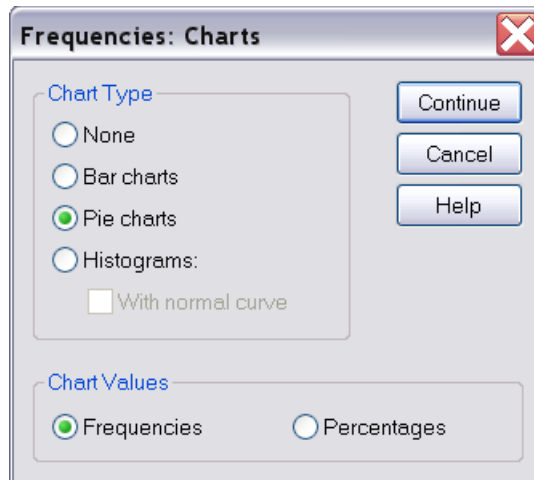
**Εικόνα 3.3** Πίνακας συχνοτήτων για τη μεταβλητή φυλή

Η διαδικασία δεν θα υπολογίσει κανένα στατιστικό μέτρο. Ούτως ή άλλως σε αυτήν την περίπτωση μόνο η εύρεση της επικρατούσας τιμής έχει νόημα. Αν θέλουμε να υπολογιστεί, θα ενεργοποιήσουμε την επιλογή **Statistics...** (Εικόνα 3.2). Εκεί θα επιλέξουμε το **mode** (επικρατούσα τιμή) (Εικόνα 3.4) και θα πατήσουμε **Continue**:



**Εικόνα 3.4** Παράθυρο διαλόγου **Frequencies: Statistics**

Αν θέλουμε να κατασκευαστεί κάποιο γράφημα θα πρέπει να κάνουμε Κλικ στο **Charts...** (Εικόνα 3.2). Τότε, θα εμφανιστεί το παρακάτω παράθυρο διαλόγου (Εικόνα 3.5):

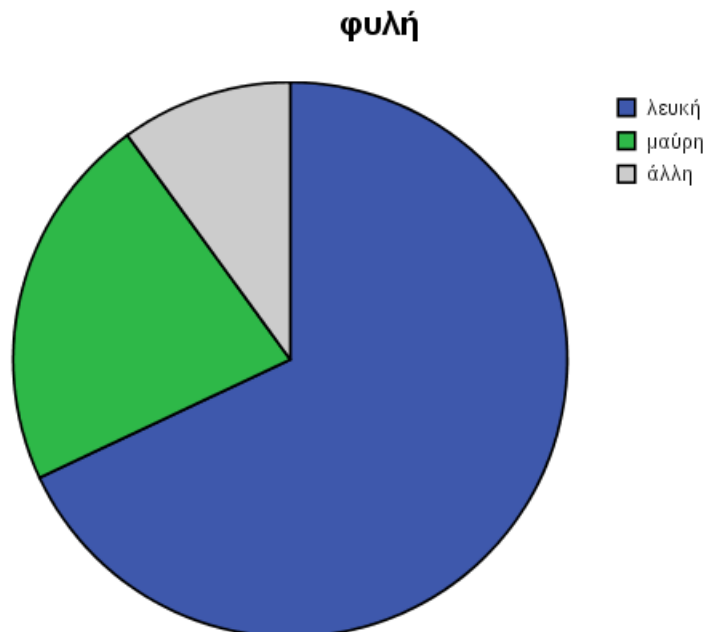


**Εικόνα 3.5** Παράθυρο Frequencies: Charts

Επιλέγουμε το Pie Charts, πατάμε Continue και OK στο παράθυρο Frequencies. Το αποτέλεσμα των επιλογών μας φαίνεται στην Εικόνα 3.6. Ο Πίνακας συχνοτήτων παραμένει ο ίδιος με τον Πίνακα της Εικόνας 3.3.

#### Statistics

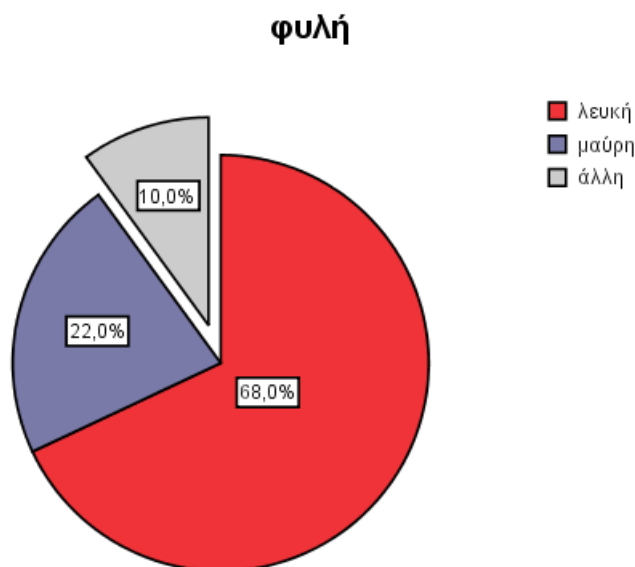
φυλή		
N	Valid	100
	Missing	0
Mode		1



**Εικόνα 3.6** Αποτέλεσμα της διαδικασίας Frequencies

Παρατηρούμε ότι η επικρατούσα τιμή είναι το 1 (που αντιστοιχεί στη λευκή φυλή), γεγονός που φυσικά συμβαδίζει με τα αποτελέσματα του πίνακα συχνοτήτων και το

κυκλικό διάγραμμα. Στη συνέχεια μπορούμε φυσικά να μορφοποιήσουμε το διάγραμμά μας, εφ' όσον το επιθυμούμε. Για παράδειγμα (Εικόνα 3.7):



**Εικόνα 3.7** Μορφοποιημένο κυκλικό διάγραμμα

Η αποκοπή του κυκλικού τομέα έγινε με χρήση των εντολών `Elements` → `Explode Slice`.

### Ποσοτικές μεταβλητές

Για τον υπολογισμό των αριθμητικών περιγραφικών μέτρων μιας ποσοτικής μεταβλητής χρησιμοποιούμε τις διαδικασίες `Descriptives`, `Frequencies` και `Explore`.

Κάποια από τα πιο συνηθισμένα στατιστικά περιγραφικά μέτρα που υπολογίζονται είναι:

1. Η **μέση τιμή** (mean) των παρατηρήσεων του δείγματος που είναι το άθροισμα όλων των παρατηρήσεων διαιρεμένο δια του συνολικού αριθμού τους. Η μέση τιμή των παρατηρήσεων συμβολίζεται με  $\bar{X}$  και ισούται με  $\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n}$ .
2. Η **διάμεσος** (median) του δείγματος. Αν οι παρατηρήσεις διαταχθούν κατά τάξη μεγέθους και το πλήθος των παρατηρήσεων είναι άρτιος αριθμός, τότε η διάμεσος ισούται με το ημίαθροισμα των δύο μεσαίων παρατηρήσεων. Αν είναι περιττός τότε είναι η μεσαία τιμή. Η διάμεσος ουσιαστικά χωρίζει τη διατεταγμένη σειρά των παρατηρήσεων σε δύο μέρη, έτσι ώστε τα μισά περίπου δεδομένα (το 50%) να βρίσκονται στην αριστερή πλευρά και τα άλλα μισά (το άλλο 50%) στη δεξιά.

3. Η **επικρατούσα τιμή** (mode), δηλαδή, η πιο συχνά εμφανιζόμενη τιμή στο δείγμα.

4. Η **τυπική απόκλιση** του δείγματος μας  $\left( s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \right)$  που

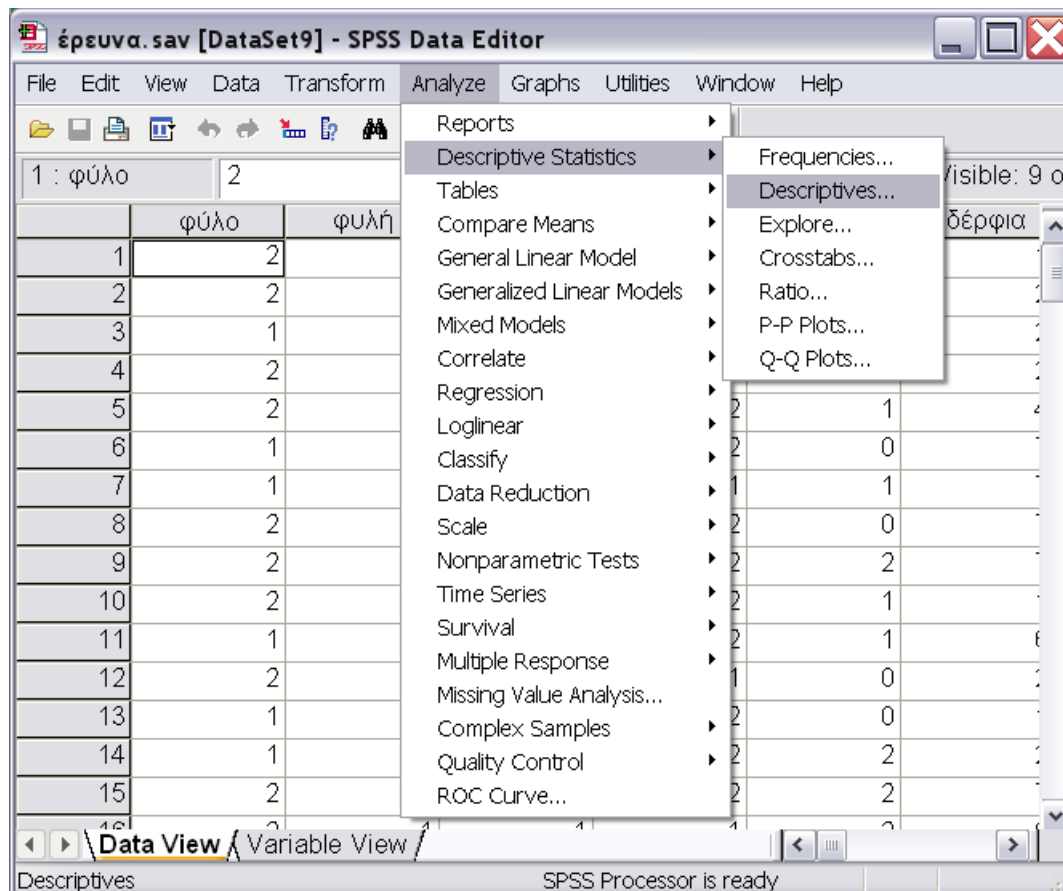
χρησιμοποιείται για να μετράει τη μεταβλητότητα του δείγματός μας, δηλαδή, το πόσο απέχουν οι παρατηρήσεις μας από τη μέση τιμή.

5. Η **μέγιστη** (maximum) και **ελάχιστη** (minimum) τιμή των παρατηρήσεων μας.

### Η διαδικασία DESCRIPTIVES

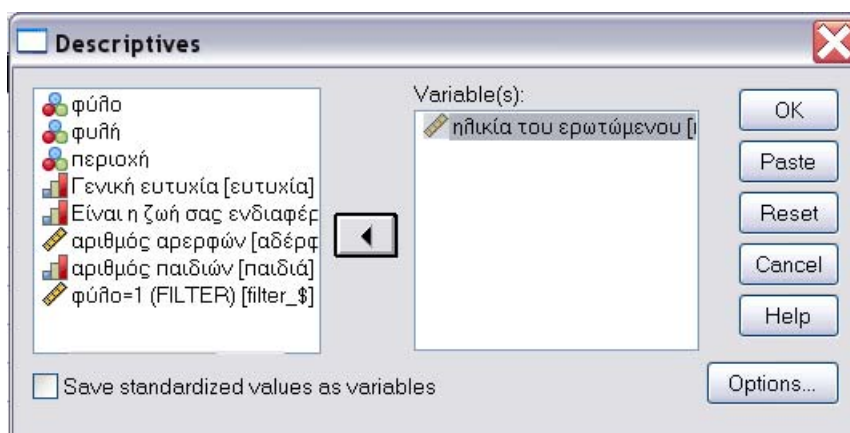
Η διαδικασία Descriptives βοηθά στον υπολογισμό των στατιστικών μέτρων μιας ή και περισσότερων ποσοτικών μεταβλητών. Για κάθε μεταβλητή, το εργαλείο αυτό υπολογίζει: τη μέση τιμή, το τυπικό σφάλμα, τη διάμεσο, την επικρατούσα τιμή, την τυπική απόκλιση, τη διακύμανση, την κύρτωση, τη λοξότητα, την ελάχιστη και τη μέγιστη τιμή, κ.ά. Από τη βασική γραμμή εντολών επιλέγουμε (Εικόνα 3.8):

Analyze → Descriptive Statistics → Descriptives...


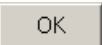


Εικόνα 3.8 Η διαδικασία Descriptives

Τότε, εμφανίζεται το παράθυρο διαλόγου *Descriptives*, όπως φαίνεται στην Εικόνα 3.9.



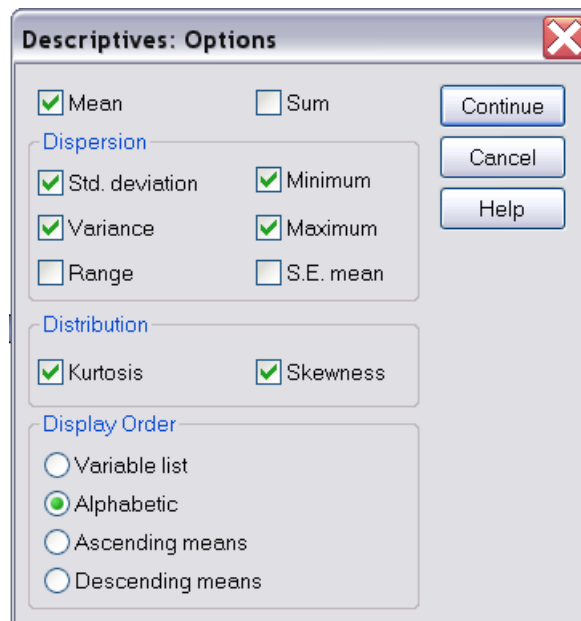
**Εικόνα 3.9** Το παράθυρο διαλόγου *Descriptives*

Επιλέγουμε τη μεταβλητή (ή τις μεταβλητές) που θέλουμε να περιγράψουμε και τη μετακινούμε στο παράθυρο *Variable(s)* με το αντίστοιχο βελάκι . Στη συνέχεια πατάμε . Η διαδικασία θα προχωρήσει στον υπολογισμό των στατιστικών μέτρων, όπως φαίνεται παρακάτω:

#### Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
ηλικία του ερωτώμενου	99	20	80	44,12	15,957
Valid N (listwise)	99				

Παρατηρούμε ότι η διαδικασία εξ ορισμού υπολογίζει τα εξής στατιστικά μέτρα: μέση τιμή, τυπική απόκλιση, μέγιστη και ελάχιστη τιμή των παρατηρήσεων. Αν θέλουμε να βρούμε και τις τιμές άλλων στατιστικών μέτρων, μπορούμε να το πραγματοποιήσουμε με την επιλογή *Options*, όπως επίσης και να καθορίσουμε τη σειρά εμφάνισης τους από το *Display Order* (Εικόνα 3.10).



**Εικόνα 3.10** Παράθυρο Descriptives: Options

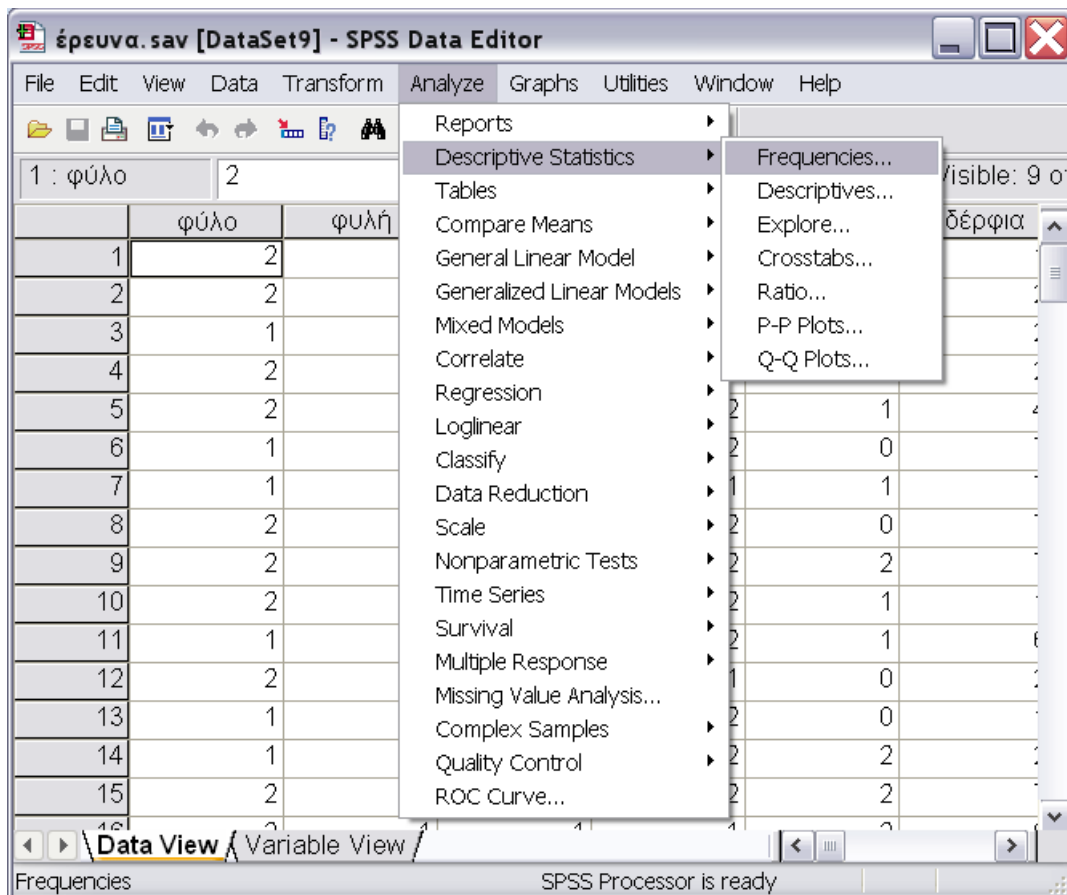
Η διαδικασία Descriptives είναι, εν γένει, μία φτωχή διαδικασία, η οποία περιορίζεται στην παρουσίαση των στατιστικών μέτρων μιας ή περισσότερων μεταβλητών σε έναν πίνακα. Αν θέλουμε μια πιο εξειδικευμένη παρουσίαση των δεδομένων μας αυτό μπορεί να γίνει με τη βοήθεια της διαδικασίας Frequencies.

### **Η διαδικασία FREQUENCIES**

Επιλέγουμε διαδοχικά από τη γραμμή των εντολών (Εικόνα 3.11):

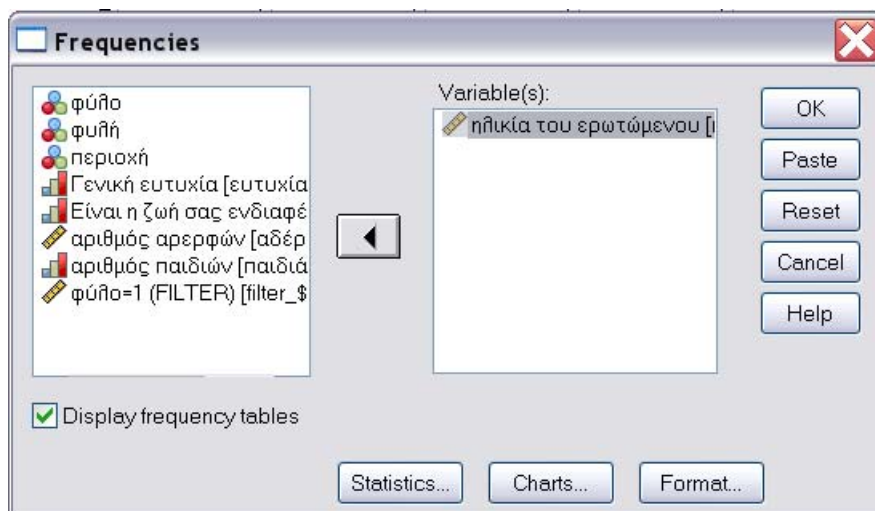
Analyze → Descriptive Statistics → Frequencies ...



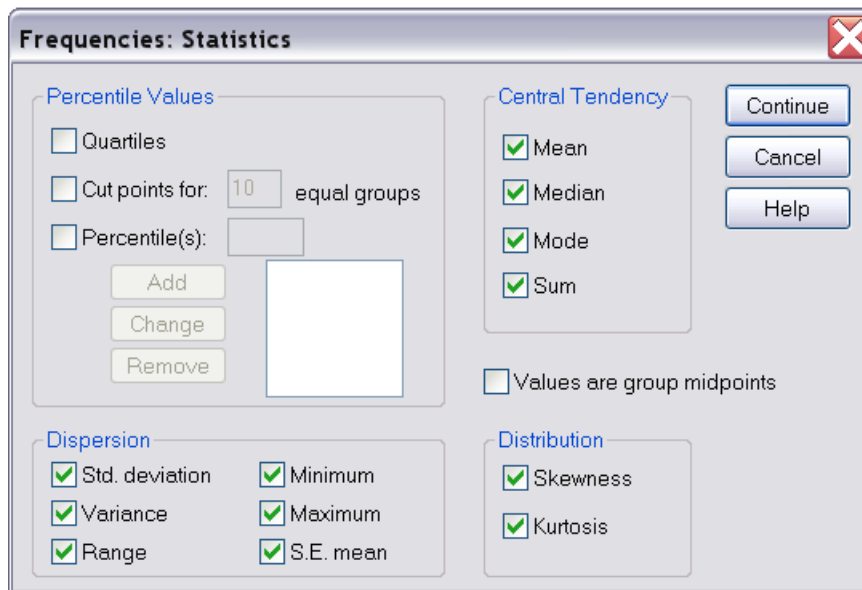


**Εικόνα 3.11** Επιλογή της διαδικασίας Frequencies

Επιλέγοντας στο παρακάτω παράθυρο (Εικόνα 3.12) το **Statistics...** έχουμε πολλές περισσότερες επιλογές (Εικόνα 3.13)

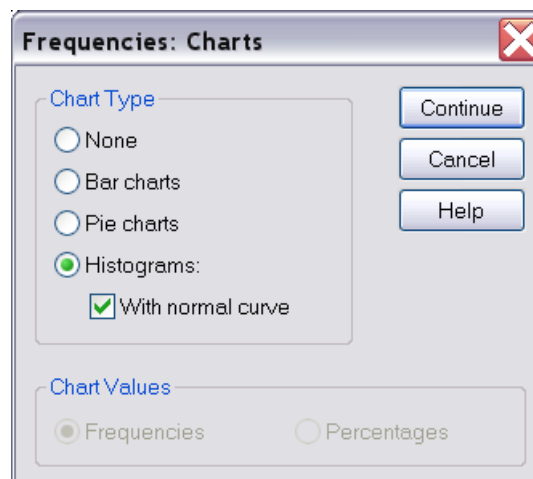


**Εικόνα 3.12** Παράθυρο διαλόγου Frequencies



**Εικόνα 3.13** Παράθυρο διαλόγου *Frequencies: Statistics*

Επιπλέον επιλογές έχουμε, αν κάνουμε Click στο **Charts...** (Εικόνα 3.14).  
Συνεχίζουμε με το **Continue**.



**Εικόνα 3.14.** Παράθυρο: *Frequencies: Charts*

Τα αποτελέσματα που θα πάρουμε στον *Output Viewer* φαίνονται στις Εικόνες 3.15 και 3.16.

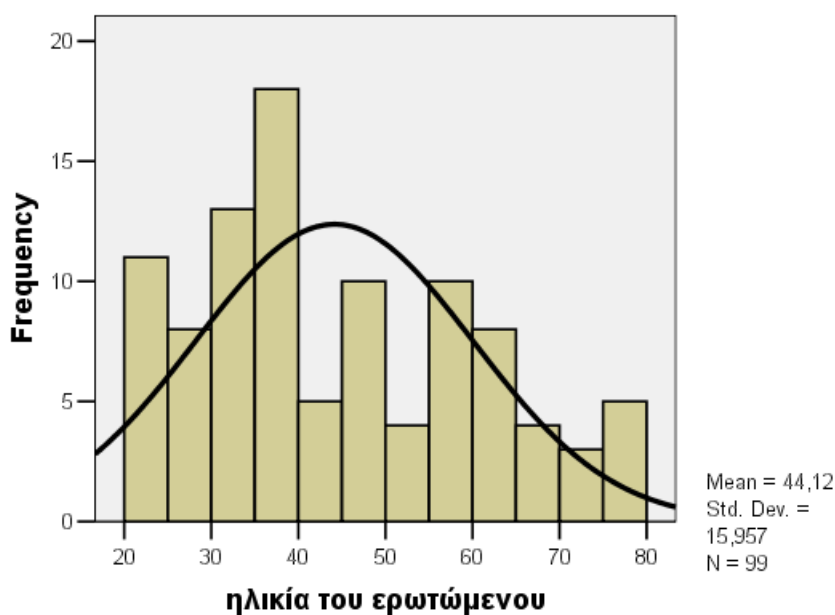
## Frequencies

### Statistics

ηλικία του ερωτώμενου		
N	Valid	99
	Missing	1
Mean		44,12
Std. Error of Mean		1,604
Median		39,00
Mode		35
Std. Deviation		15,957
Variance		254,618
Skewness		,435
Std. Error of Skewness		,243
Kurtosis		-,822
Std. Error of Kurtosis		,481
Range		60
Minimum		20
Maximum		80

**Εικόνα 3.15** Αριθμητικά περιγραφικά μέτρα για τη μεταβλητή ηλικία

### Histogram

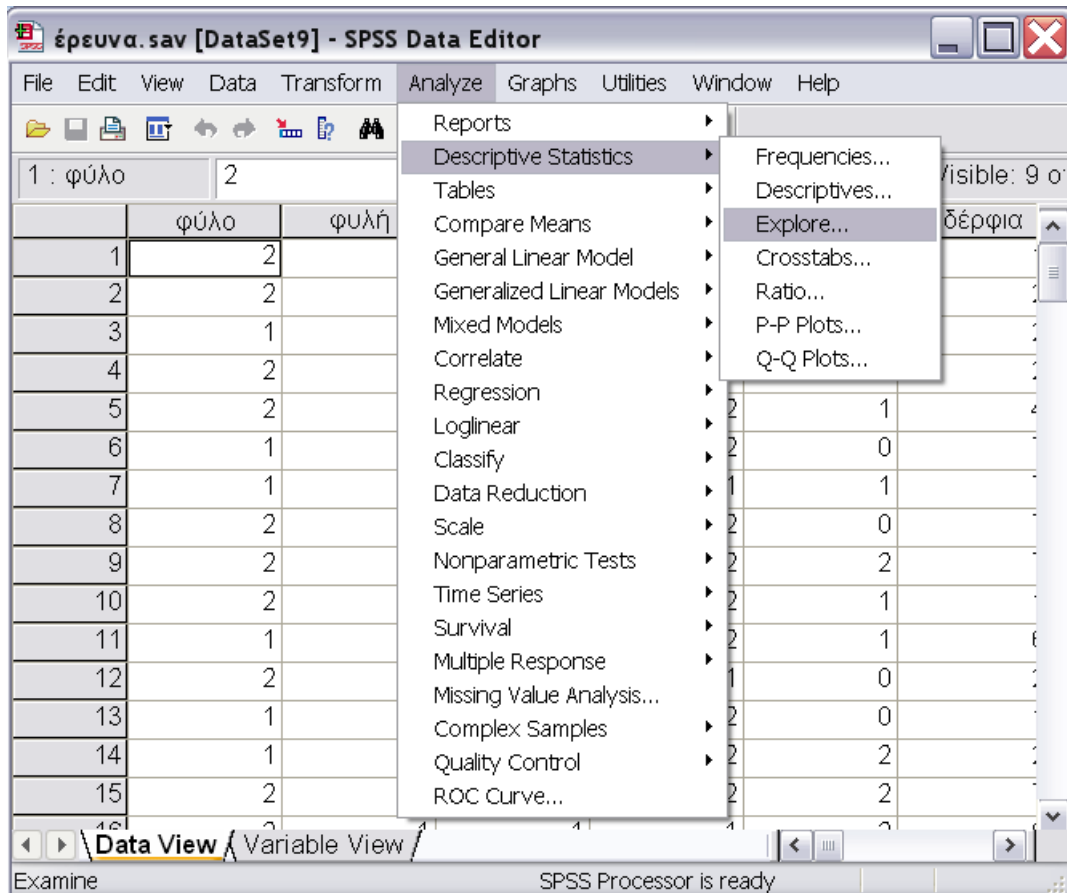


**Εικόνα 3.16** Ιστόγραμμα συχνοτήτων

### Η διαδικασία EXPLORE - Έλεγχος κανονικότητας

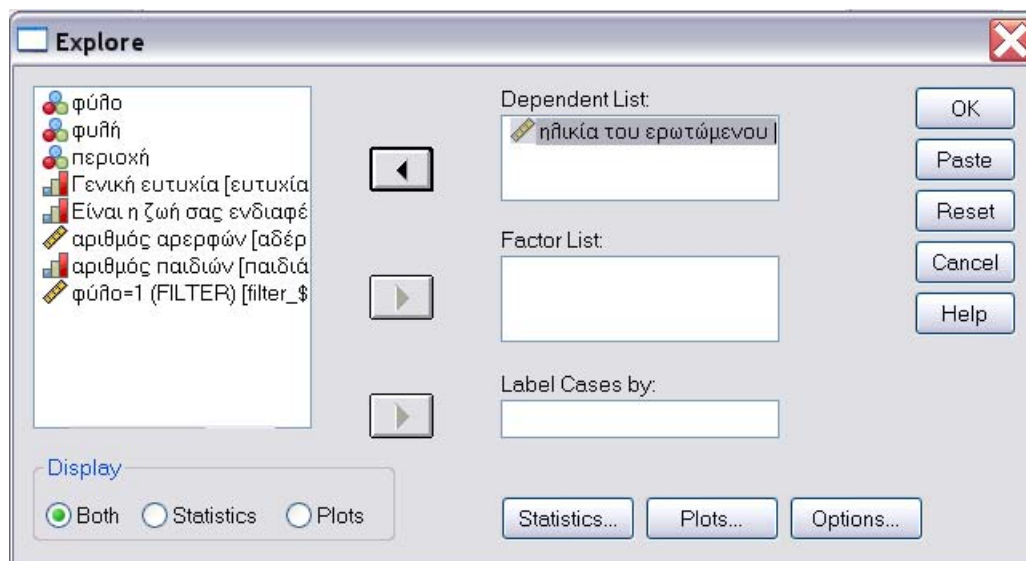
Με τη διαδικασία *Explore* μπορούμε να πετύχουμε την πιο πλούσια και πλήρη περιγραφή των παρατηρήσεων μιας ποσοτικής μεταβλητής (πλουσιότερης της διαδικασίας *Descriptives* και *Frequencies*). Θα χρησιμοποιήσουμε τα δεδομένα του αρχείου *έρευνα.sav*. Επιλέγουμε διαδοχικά από το μενού των εντολών (Εικόνα 3.17):

Analyze → Descriptive Statistics → Explore...

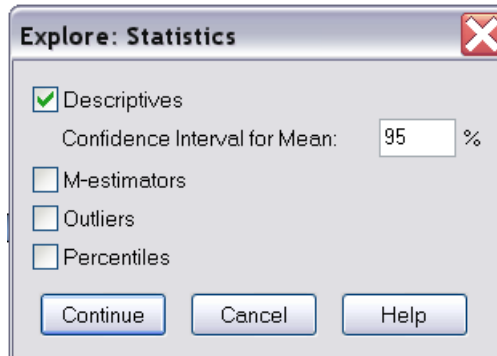


**Εικόνα 3.17** Η διαδικασία Explore

Εισάγουμε τη μεταβλητή ηλικία του ερωτώμενου στο πλαίσιο Dependent List (Εικόνα 3.18) και στη συνέχεια πατάμε το **Statistics...**, αν θέλουμε να πάρουμε και άλλες πληροφορίες εκτός από αυτές που δίνει εξ ορισμού η διαδικασία (Εικόνα 3.19).



**Εικόνα 3.18** Παράθυρο διαλόγου Explore



**Εικόνα 3.19** Παράθυρο διαλόγου `Explore: Statistics`

Συνεχίζουμε με το `Continue`. Στη συνέχεια πατώντας το `Plots...` μπορούμε να επιλέξουμε και κάποιο άλλο γράφημα, π.χ. ιστόγραμμα ή να κάνουμε έναν έλεγχο προσαρμογής των δεδομένων μας στην κανονική κατανομή (Εικόνα 3.20). Συνεχίζουμε με το `Continue` και στο παράθυρο `Explore` πατάμε `OK`.



**Εικόνα 3.20** Παράθυρο διαλόγου `Explore: Plots`

Τα αποτελέσματα που θα πάρουμε στον `Output Viewer` φαίνονται παρακάτω στις Εικόνες 3.21, 3.22. Στο παράθυρο `Descriptives` εμφανίζεται εκτός των άλλων ένα 95% – διάστημα εμπιστοσύνης για τη μέση τιμή. Δηλαδή, είμαστε σίγουροι με πιθανότητα ίση με 0.95, ότι η μέση τιμή για τη μεταβλητή ηλικία του ερωτώμενου βρίσκεται στο διάστημα (40.94, 47.30). Δίνεται επίσης η 5% – ισοσταθμισμένη μέση τιμή. Γενικά, η  $n\%$  ισοσταθμισμένη μέση τιμή ( $n\%$  – trimmed mean), ορίζεται ως η μέση τιμή που υπολογίζεται όταν οι  $n\%$  μεγαλύτερες και οι  $n\%$  μικρότερες τιμές έχουν διαγραφεί. Εδώ  $n = 5$ . Η διαγραφή

των ακραίων τιμών έχει καλύτερο αποτέλεσμα, ειδικά όταν τα δεδομένα δεν προέρχονται από κανονική κατανομή.

## Explore

### Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
ηλικία του ερωτώμενου	99	99,0%	1	1,0%	100	100,0%

### Descriptives

			Statistic	Std. Error
ηλικία του ερωτώμενου	Mean		44,12	1,604
	95% Confidence Interval for Mean	Lower Bound	40,94	
		Upper Bound	47,30	
	5% Trimmed Mean		43,54	
	Median		39,00	
	Variance		254,618	
	Std. Deviation		15,957	
	Minimum		20	
	Maximum		80	
	Range		60	
	Interquartile Range		25	
	Skewness		,435	,243
	Kurtosis		-,822	,481

### Tests of Normality

	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ηλικία του ερωτώμενου	,137	99	,000	,949	99	,001

a. Lilliefors Significance Correction

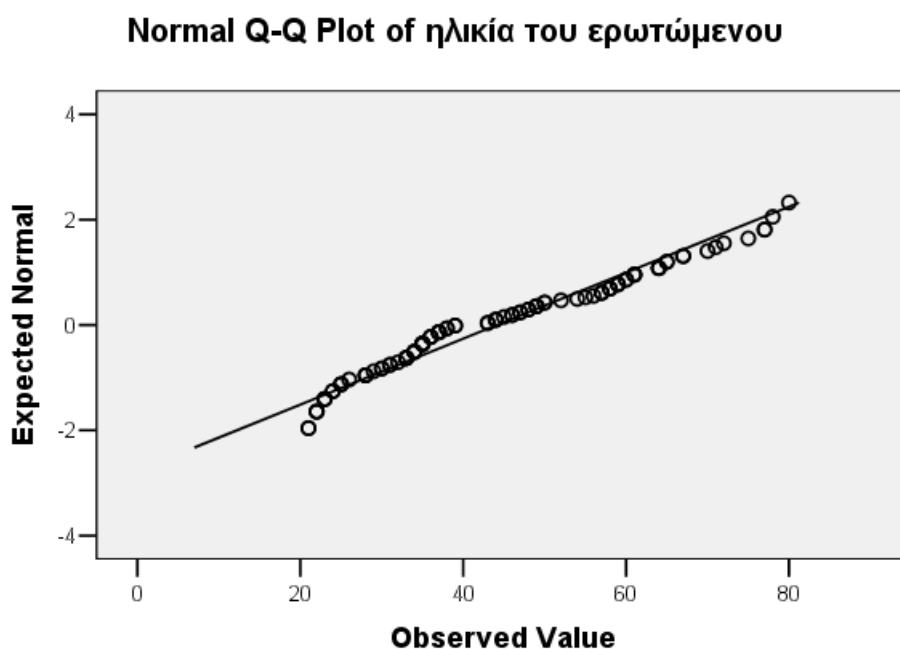
**Εικόνα 3.21** Αποτελέσματα της διαδικασίας Explore

Γίνεται επίσης έλεγχος προσαρμογής των δεδομένων μας στην κανονική κατανομή, όπως ζητήθηκε (Tests of Normality) (Εικόνα 3.21). Αν το δείγμα μας είναι μικρότερο του 50, τότε θα κοιτάξουμε το στατιστικό κριτήριο Shapiro-Wilk. Αν όχι, τότε θα κοιτάξουμε το στατιστικό κριτήριο Kolmogorov-Smirnov. Για να αποφασίσουμε αν ισχύει η μηδενική μας υπόθεση, δηλαδή ότι το δείγμα προέρχεται από πληθυσμό που ακολουθεί την κανονική κατανομή, εξετάζουμε την τιμή στη στήλη Sig. του

κριτηρίου Kolmogorov-Smirnov, δεδομένου ότι το δείγμα μας είναι μεγέθους μεγαλύτερου από 50. Αν αυτή είναι μικρότερη από το 0.05, τότε απορρίπτουμε τη μηδενική υπόθεση, αν όχι, τότε την αποδεχόμαστε. Εδώ, η τιμή είναι ίση με  $0.000 < 0.05$ , οπότε η μηδενική υπόθεση απορρίπτεται. Δεν μπορούμε, λοιπόν, να ισχυριστούμε με βεβαιότητα 95%, ότι το δείγμα προέρχεται από πληθυσμό που ακολουθεί την κανονική κατανομή.

Δίνεται επίσης ένα Q-Q plot (Εικόνα 3.22), το οποίο ερμηνεύεται με τον ίδιο τρόπο όπως και το P-P plot που περιγράφηκε στο Κεφάλαιο 2. Τα αντίστοιχα σημεία για τη μεταβλητή μας αποκλίνουν αισθητά από την ευθεία. Δίνονται επιπλέον ένα φυλλόγραμμα (stem and leaf plot) (Εικόνα 3.23) και ένα θηκόγραμμα (Εικόνα 3.24). Στο φυλλόγραμμα αναγράφονται όλες οι παρατηρήσεις που έχουμε, όλες οι τιμές δηλαδή της μεταβλητής ηλικία. Η στήλη που βρίσκεται αριστερά της κάθετης γραμμής είναι γνωστή ως **κορμός**, ενώ οι άλλοι αριθμοί δεξιά της γραμμής είναι τα **φύλλα**. Ο κορμός αντιπροσωπεύει τις «δεκάδες» και τα φύλλα τις «μονάδες». Για παράδειγμα, η πρώτη γραμμή που αρχίζει με το «2» έχει τις μονάδες «0», το «1» δύο φορές, το «2» 3 φορές κ.ο.κ. Επομένως, ξέρουμε ότι υπάρχουν οι παρατηρήσεις 20, 21, 21, 22, 22, 22, κ.ο.κ.

Αν επιθυμούμε να γίνουν και άλλα γραφήματα, πρέπει να ενεργοποιήσουμε την επιλογή Plots.



**Εικόνα 3.22** Q-Q plot

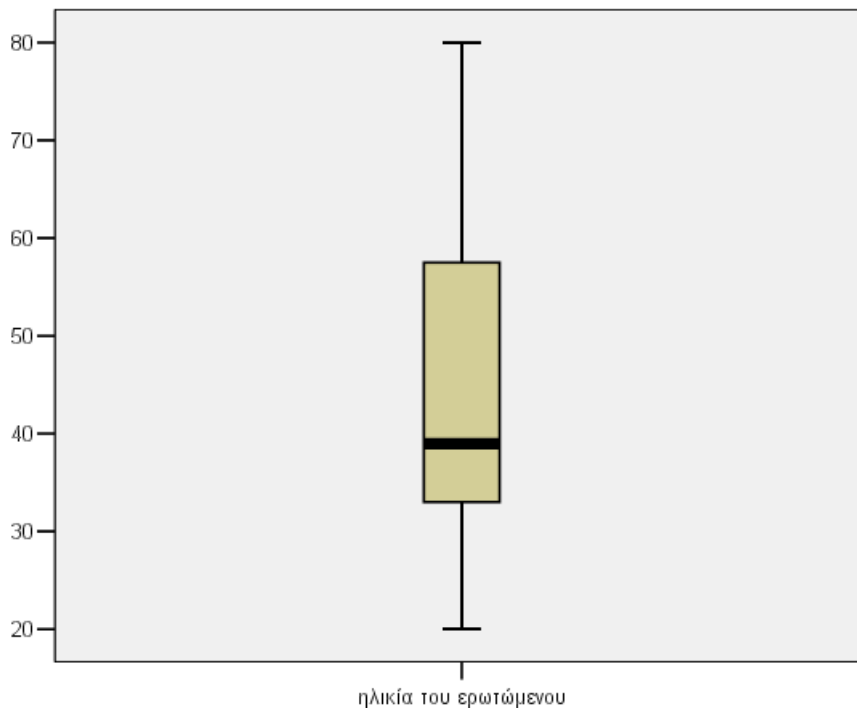
### Φυλλόγραμμα

ηλικία του ερωτώμενου Stem-and-Leaf Plot

Frequency	Stem & Leaf
11,00	2 . 01122233344
8,00	2 . 55568889
13,00	3 . 0011233334444
18,00	3 . 555555566677778899
5,00	4 . 33444
10,00	4 . 5667788999
4,00	5 . 0024
10,00	5 . 5677788999
8,00	6 . 00111444
4,00	6 . 5577
3,00	7 . 012
4,00	7 . 5778
1,00	8 . 0

Stem width: 10  
Each leaf: 1 case(s)

**Εικόνα 3.23** Φυλλόγραμμα



**Εικόνα 3.24** Θηκόγραμμα



## Στατιστική Συμπερασματολογία

Το SPSS ενσωματώνει στατιστικούς ελέγχους (tests) που σχετίζονται με την ανάλυση μιας μεταβλητής. Με τους στατιστικούς ελέγχους επιχειρούμε να διαπιστώσουμε, για παράδειγμα:

1. Αν η παρατηρούμενη συνάρτηση κατανομής μιας μεταβλητής συμπίπτει με κάποια από τις γνωστές θεωρητικές Poisson, Normal (Κανονική), κλπ.
2. Αν οι παρατηρήσεις μιας μεταβλητής είναι τυχαίες (Runs test).
3. Αν οι παρατηρούμενες συχνότητες των κατηγοριών μιας μεταβλητής απέχουν ή όχι από τις θεωρητικές αναμενόμενες μιας γνωστής κατανομής ( $\chi^2$  - test καλής προσαρμογής), κ.ά.

Επιπλέον, με κάποιες γραφικές παραστάσεις (πιθανοθεωρητικά γραφήματα P-P, Q-Q) επιχειρούμε να διαπιστώσουμε πόσο κοντά σε μία συγκεκριμένη κατανομή είναι τα δεδομένα που επεξεργαζόμαστε.

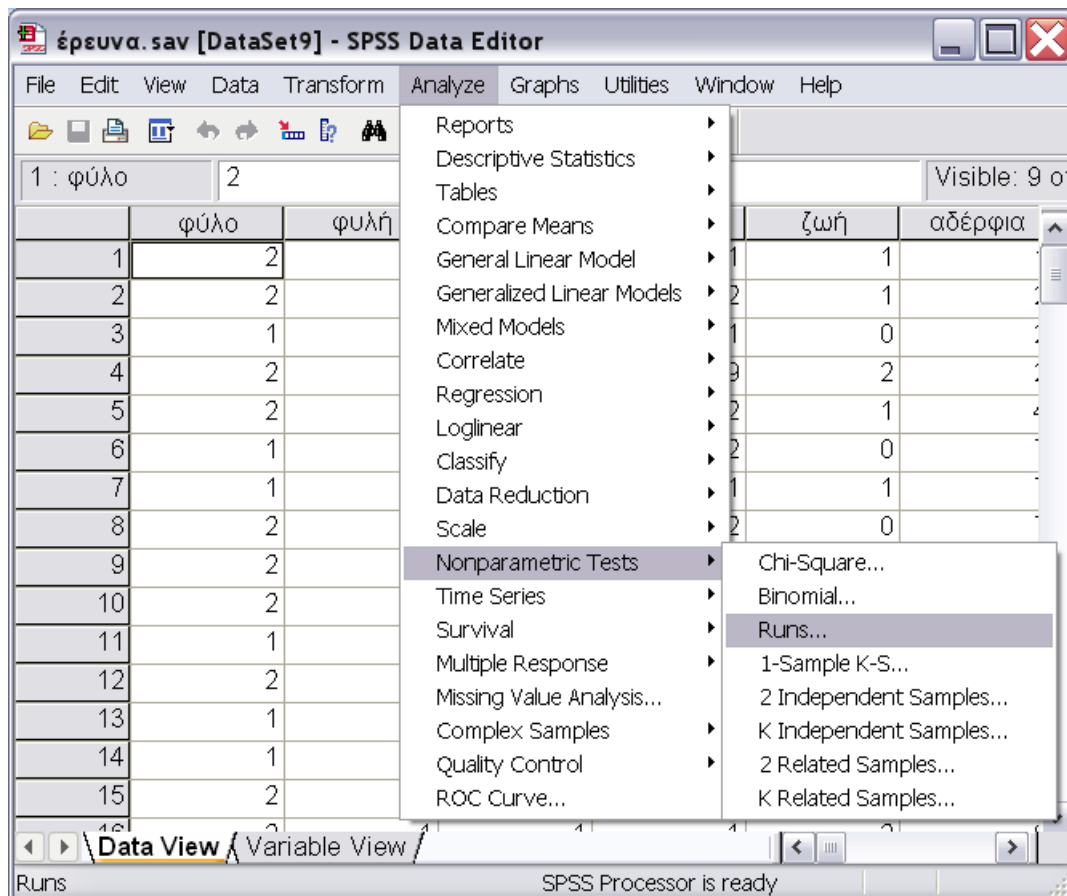
Θα ασχοληθούμε στη συνέχεια με δύο στατιστικούς ελέγχους: της τυχειότητας των παρατηρήσεων του δείγματος και τον έλεγχο υπόθεσης για τη μέση τιμή.

### Έλεγχος τυχειότητας του δείγματος (Runs test)

Πολλά στατιστικά τεστ υποθέτουν ότι οι παρατηρήσεις στο δείγμα είναι ανεξάρτητες, δηλαδή, η σειρά με την οποία συγκεντρώθηκαν τα δεδομένα είναι τυχαία. Αν η σειρά έχει σημασία, τότε το δείγμα δεν είναι τυχαίο και δεν μπορούμε να βγάλουμε ακριβή συμπεράσματα για τον πληθυσμό μας από αυτό. Γιαυτό είναι απαραίτητο να ελέγξουμε τα δεδομένα μας για τυχόν παραβίαση αυτής της υπόθεσης (της τυχειότητας του δείγματος). Για να είναι αξιόπιστα τα συμπεράσματά μας, θα πρέπει το δείγμα μας να είναι τυχαίο. Ο έλεγχος αυτός γίνεται με τη βοήθεια του στατιστικού τεστ Runs. Το τεστ Runs ελέγχει αν η σειρά εμφάνισης των τιμών μιας μεταβλητής είναι τυχαία. Το Run είναι μια ακολουθία τιμών που μοιάζουν και πιο συγκεκριμένα, μια ακολουθία τιμών που βρίσκονται προς την ίδια πλευρά ενός σημείου τομής (cut point). Ένα δείγμα με πάρα πολλά Runs ή με πολύ λίγα φαίνεται να μην είναι τυχαίο. Για παράδειγμα: Ας υποθέσουμε ότι ένα δείγμα 50 ατόμων ρωτάται σχετικά με το αν θα αγόραζε ή όχι ένα προϊόν. Η υποτιθέμενη τυχειότητα του δείγματος θα αμφισβητούνταν αν και τα 50 άτομα ήταν του ίδιου φύλου. Το Run test θα μπορούσε να χρησιμοποιηθεί για να διαπιστωθεί αν το δείγμα είναι όντως τυχαίο. Η υπόθεση  $H_0$ , η μηδενική υπόθεση δηλαδή σε αυτόν τον έλεγχο, είναι ότι η σειρά των παρατηρήσεων είναι τυχαία.

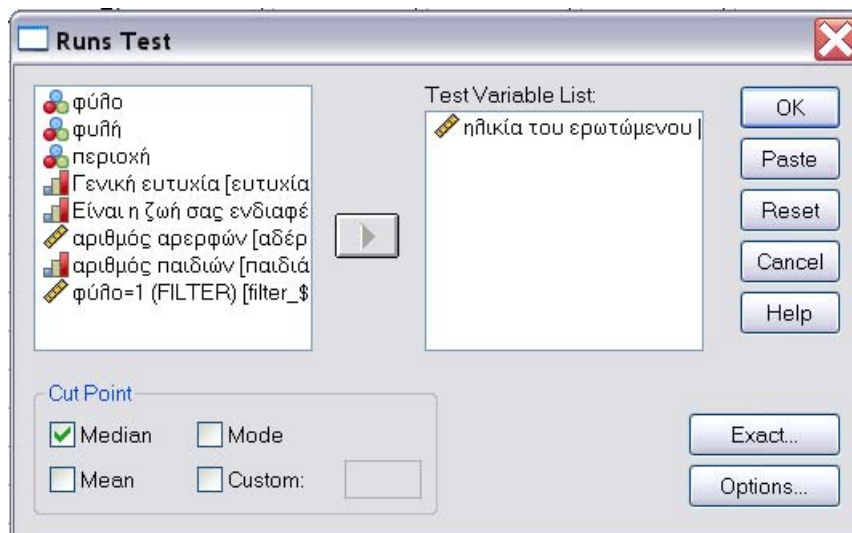
Από τη βασική γραμμή εντολών επιλέγουμε διαδοχικά (Εικόνα 3.25):

Statistics → Nonparametric Tests → Runs...




**Εικόνα 3.25** Τεστ Runs

Τότε, θα εμφανιστεί το παρακάτω παράθυρο διαλόγου (Εικόνα 3.26):



**Εικόνα 3.26** Παράθυρο διαλόγου Runs Test

Η διαδικασία αρχικά ταξινομεί κάθε τιμή της μεταβλητής, ανάλογα με το αν βρίσκεται πάνω ή κάτω από ένα σημείο τομής (cut point) (Εικόνα 3.27). Στη συνέχεια κάνει

τον έλεγχο για να επιβεβαιώσει ότι δεν υπάρχει κάποια τάξη στην ακολουθία που προέκυψε. Το σημείο τομής μπορεί να είναι ένα από τα παρακάτω στατιστικά μέτρα: η διάμεσος (προεπιλεγμένο), η μέση τιμή, η επικρατούσα τιμή ή κάποια άλλη συγκεκριμένη τιμή. Στο παράδειγμα αυτό χρησιμοποιούνται τα δεδομένα του έρευνα.sav. Από τον πίνακα των μεταβλητών επιλέγουμε μία ή περισσότερες μεταβλητές και τη μετακινούμε στο `Test Variable List` χρησιμοποιώντας το αντίστοιχο εικονίδιο . Στο συγκεκριμένο παράδειγμα, επιλέγουμε τη μεταβλητή ηλικία του ερωτώμενου, την τοποθετούμε στο `Test Variable List` και πατάμε OK. Ως σημείο τομής επιλέγουμε τη διάμεσο και θα πάρουμε ως αποτέλεσμα τον παρακάτω πίνακα (Εικόνα 3.27):

**Runs Test**

	ηλικία του ερωτώμενου	
Test Value(a)	39	→ 1
Cases < Test Value	48	→ 2
Cases >= Test Value	51	→ 3
Total Cases	99	→ 4
Number of Runs	45	→ 5
Z	-1,103	→ 6
Asymp. Sig. (2-tailed)	,270	→ 7

a Median

**Εικόνα 3.27** Αποτελέσματα του `Runs Test`

1. Το `test value` (τιμή ελέγχου) χρησιμοποιείται ως σημείο τομής (`cut point`) για να διχοτομήσει το δείγμα. Σε αυτό το παράδειγμα, το σημείο τομής είναι η διάμεσος, που είναι ίση με 39.
2. Από 99 περιπτώσεις, 48 έπεσαν κάτω από τη διάμεσο. Ας θεωρήσουμε αυτές τις περιπτώσεις ως «αρνητικές».
3. Οι υπόλοιπες 51 περιπτώσεις έπεσαν ακριβώς ή πάνω από τη διάμεσο. Ας θεωρήσουμε αυτές τις περιπτώσεις ως «θετικές».
4. Ο συνολικός αριθμός των περιπτώσεων είναι 99.
5. Το επόμενο στατιστικό είναι ένας μετρητής των παρατηρουμένων `Runs` στη μεταβλητή που ελέγχουμε. Όπως αναφέραμε και προηγουμένως το `Run` είναι μία ακολουθία των περιπτώσεων που βρίσκονται στην ίδια πλευρά του σημείου τομής. Ο συνολικός αριθμός των `Runs` είναι 45.

παιδιά	ηλικία
2	61
1	32
1	35
0	26
0	25

Για παράδειγμα: αν κοιτάξουμε τα δεδομένα μας, παρατηρούμε ότι η πρώτη περίπτωση βρίσκεται πάνω από τη διάμεσο. Αυτή η ακολουθία που αποτελείται από την πρώτη παρατήρηση είναι το πρώτο Run, δεδομένου ότι η επόμενη τιμή, το 32, είναι μικρότερη από το 39.

1	32
1	35
0	26
0	25
5	59

Το δεύτερο Run ξεκινάει από τη 2<sup>η</sup> περίπτωση, στην οποία η ηλικία είναι ίση με 32 (και σταματάει στο 25 γιατί η επόμενη μέτρηση είναι ίση με 59 που είναι μεγαλύτερο από τη διάμεσο).

5	59
3	46
4	99
3	57
2	64
0	72
5	67
0	33

Το τρίτο Run ξεκινάει από την περίπτωση 6, στην οποία η ηλικία είναι μεγαλύτερη από τη διάμεσο και σταματάει στο 67, γιατί η επόμενη παρατήρηση (το 33) είναι μικρότερη της διαμέσου. Η διαδικασία αυτή συνεχίζεται έως ότου καλύψουμε και τις 99 περιπτώσεις.

6. Το Z στατιστικό είναι ίσο -1.103.
7. Με την επιλογή που κάναμε για το σημείο τομής (τη διάμεσο) αποδεχόμαστε την υπόθεση  $H_0$  (της τυχαιότητας των παρατηρήσεων), γιατί  $0.270 > 0.05$ . Δηλαδή, μπορούμε να ισχυριστούμε ότι η σειρά των παρατηρήσεων πάνω και

κάτω από τη διάμεσο είναι τυχαία. Γενικά, όμως, μπορούμε να πούμε ότι τα αποτελέσματα του τεστ εξαρτώνται από την επιλογή του σημείου τομής.

### Έλεγχος υπόθεσης για τη μέση τιμή (One - Sample T test)

Γενικά, το One-Sample T test μπορεί να χρησιμοποιηθεί κάθε φορά που θέλουμε να ελέγξουμε τη μέση τιμή του δείγματος έναντι μίας συγκεκριμένης τιμής ελέγχου.

#### Προϋποθέσεις:

1. Όπως και σε όλα τα T test, υποθέτουμε ότι τα δεδομένα προέρχονται από πληθυσμό που ακολουθεί την κανονική κατανομή, και
2. Το δείγμα έχει επιλεγεί τυχαία από τον πληθυσμό μας.

Επιπλέον προσοχή πρέπει να δίνεται στις ακραίες τιμές, γιατί επηρεάζουν πολύ τη δειγματική μέση τιμή. Τα θηκοκράμματα γενικά προσφέρουν μεγάλη βοήθεια ως προς αυτό το θέμα.

Πιο συγκεκριμένα, ο έλεγχος αυτός (One-Sample T test):

- Ελέγχει τη διαφορά που υπάρχει μεταξύ της μέσης τιμής του δείγματος και μιας γνωστής υποθετικής τιμής.
- Παράγει έναν πίνακα περιγραφικής στατιστικής (Descriptives) για κάθε ελεγχόμενη μεταβλητή.

Για την περιγραφή αυτής της διαδικασίας θα χρησιμοποιήσουμε τα δεδομένα από το αρχείο ΠΣ\_ΣΕΠ\_05.sav. Στο αρχείο αυτό περιέχονται οι βαθμολογίες των φοιτητών που έδωσαν το μάθημα της Περιγραφικής Στατιστικής, το Σεπτέμβριο του 2005. Θέλουμε να εξετάσουμε αν οι βαθμολογίες του δείγματος των φοιτητών προέρχονται από έναν πληθυσμό με μέση τιμή ίση με 5.

Ελέγχουμε αρχικά την κανονικότητα μέσω της διαδικασίας Explore:

Analyze → Descriptive Statistics → Explore...

Στο Plot επιλέγουμε το Normality plots with test και τα αποτελέσματα που παίρνουμε είναι τα εξής:

#### Tests of Normality

	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
εξετάσεις Σεπτ. 2005	,111	36	,200(*)	,940	36	,052

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

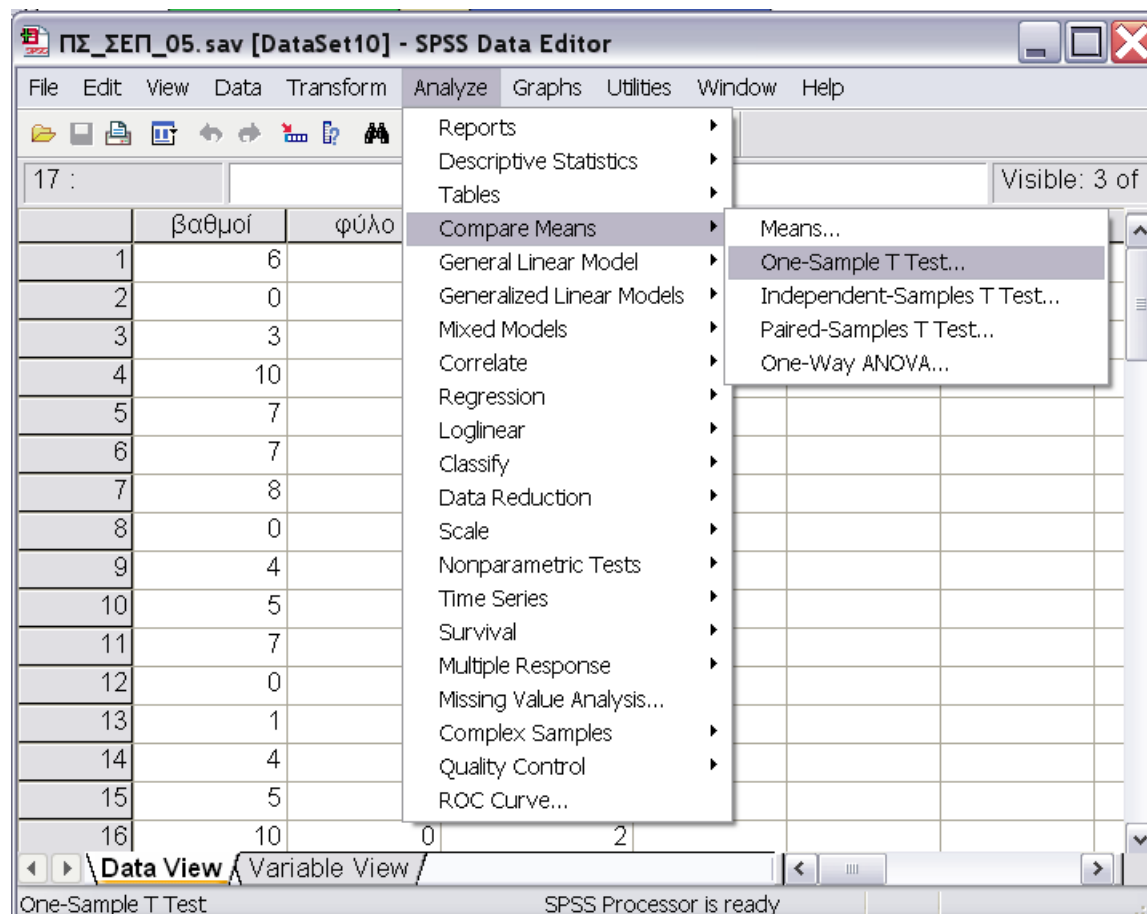
a Lilliefors Significance Correction

Το δείγμα μας έχει μέγεθος ίσο με 36, οπότε κοιτάζουμε το στατιστικό κριτήριο Shapiro-Wilk και την αντίστοιχη τιμή στη στήλη Sig. Αυτή είναι ίση με  $0.052 > 0.05$ ,

οπότε μπορούμε να ισχυριστούμε ότι τα δεδομένα προέρχονται από πληθυσμό που ακολουθεί την κανονική κατανομή.

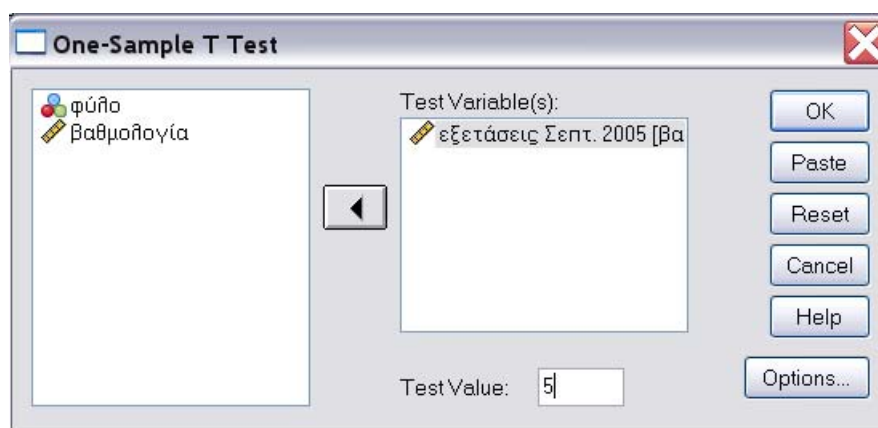
Στη συνέχεια επιλέγουμε από τη βασική γραμμή των εντολών:

Analyze → Compare Means → One-Sample T Test...



**Εικόνα 3.28** Έλεγχος υπόθεσης για τη μέση τιμή

Επιλέγουμε το βαθμός ως τη μεταβλητή ελέγχου και πληκτρολογούμε το 5 ως την τιμή ελέγχου στο Test Value και πατάμε OK.



**Εικόνα 3.29** Το παράθυρο διαλόγου One-Sample T Test

Εξ ορισμού υπολογίζεται το 95% – διάστημα εμπιστοσύνης για τη διαφορά της δοσμένης τιμής ελέγχου (εδώ η τιμή αυτή είναι ίση με 5) από τη μέση τιμή του δείγματος. Αν θέλουμε να υπολογίσουμε ένα διαφορετικό διάστημα εμπιστοσύνης, μπορούμε να επιλέξουμε το Options και εκεί να διαλέξουμε π.χ. το 90 ως confidence interval percentage. Πατάμε το Continue και στη συνέχεια OK στο παράθυρο διαλόγου του τεστ.

Ο πίνακας One-Sample Statistics εμφανίζει το μέγεθος του δείγματος, τη μέση τιμή, την τυπική απόκλιση και το τυπικό σφάλμα για το δείγμα μας.

**One-Sample Statistics**

	N	Mean	Std. Deviation	Std. Error Mean
εξετάσεις Σεπτ. 2005	36	4,25	3,074	,512

**Εικόνα 3.30** Ο πίνακας One-Sample Statistics

**One-Sample Test**

	Test Value = 5					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
εξετάσεις Σεπτ. 2005	-1,464	35	,152	-,750	-1,79	,29

**Εικόνα 3.31** Ο πίνακας One-Sample Test

Ο πίνακας One-Sample Test παρουσιάζει τα αποτελέσματα του T test που κάναμε.

- Η τρίτη στήλη (Sig. (2-tailed)) χρησιμοποιείται για την απόρριψη ή όχι της υπόθεσης  $H_0$ . Εδώ η τιμή αυτή είναι ίση με  $0.152 > 0.05$ , που σημαίνει ότι αποδεχόμαστε τη μηδενική υπόθεση. Μπορούμε, δηλαδή, να ισχυριστούμε με βεβαιότητα 95%, ότι το δείγμα προέρχεται από πληθυσμό με μέση τιμή ίση με 5.
- Η μέση διαφορά στην 4<sup>η</sup> στήλη (Mean Difference) υπολογίζεται αν από τη μέση τιμή του δείγματος αφαιρέσουμε την τιμή ελέγχου (εδώ το 5).
- Το 95% – διάστημα εμπιστοσύνης για τη διαφορά, μας δίνει μία εκτίμηση για τα όρια μεταξύ των οποίων βρίσκεται η πραγματική διαφορά. Το γεγονός ότι το διάστημα εμπιστοσύνης έχει το κάτω άκρο κάτω από το μηδέν και το πάνω άκρο πάνω από το μηδέν συνηγορεί υπέρ της αποδοχής της μηδενικής

υπόθεσης, μπορούμε να ισχυριστούμε, δηλαδή, ότι το δείγμα προέρχεται από πληθυσμό με μέση τιμή ίση με 5.



## Ασκήσεις

1. Πληκτρολογήστε τα παρακάτω δεδομένα:

	baros	ypsos	hlikia
1	51.50	150.00	9.00
2	49.00	152.00	10.50
3	49.50	149.00	10.00
4	50.50	151.00	11.00
5	51.00	157.00	10.50
6	52.50	148.00	9.50
7	50.00	149.00	9.00
8	51.00	152.00	10.50
9	49.50	151.00	11.00
10	56.00	150.00	11.00
11	54.50	153.00	10.50
12	52.00	155.00	9.50
13	53.50	148.00	11.00
14	46.50	147.00	9.00
15	44.00	150.00	9.00
16	47.00	151.00	9.50

Χρησιμοποιήστε τη διαδικασία `Descriptives` για να υπολογίσετε τα στατιστικά μέτρα των ποσοτικών μεταβλητών: Βάρος, Ύψος και Ηλικία.

2. Χρησιμοποιώντας τα δεδομένα της Άσκησης 1 δώστε μια πλήρη περιγραφή των τριών μεταβλητών χρησιμοποιώντας τη διαδικασία `Frequencies`.
3. Ένα δείγμα 24 ανθρώπων υποβάλλεται σε ένα τεστ μέτρησης του δείκτη της ανθρώπινης νοημοσύνης και τα αποτελέσματα είναι τα εξής:

96	125	105	130	92	76
85	124	128	95	101	112
115	130	88	98	108	118
100	101	99	87	89	123

- Υπολογίστε τα μέτρα κεντρικής τάσης και τα μέτρα μεταβλητότητας για τη μεταβλητή `IQ`.

- Υπολογίστε ένα 95% – διάστημα εμπιστοσύνης για το μέσο IQ του πληθυσμού από τον οποίο προέρχεται το δείγμα.
  - Μπορούμε να ισχυριστούμε ότι η μεταβλητή IQ ακολουθεί την κανονική κατανομή;
  - Ερμηνεύστε το Q-Q διάγραμμα που θα εμφανιστεί στον Output Viewer.
  - Ερμηνεύστε το θηκόγραμμα (boxplot) που θα εμφανιστεί στον Output Viewer (διάμεσος, ποσοστιαία σημεία, ελάχιστη και μέγιστη παρατηρημένη τιμή και ακραίες τιμές).
  - Μπορείτε να ισχυριστείτε ότι οι παρατηρήσεις αυτές προέρχονται από ένα τυχαίο δείγμα;
4. Οι ακόλουθες τιμές αποτελούν επιδόσεις φοιτητών στα μαθηματικά (*math*), που πήραν μέρος σε ένα τεστ με σκοπό την εκτίμηση της μέσης απόδοσης τους:

70	73	57	84	73	70
52	53	68	78	57	67
54	94	99	100	78	69
70	81	89	74	56	65
45	54	78	98	72	63

Δώστε μια πλήρη ανάλυση της μεταβλητής *math*.

5. Πήραμε τυχαία 20 φοιτητές και είχαν βάρη σε kg: 53, 69, 62, 78, 81, 55, 66, 62, 74, 60, 65, 80, 78, 56, 75, 72, 65, 69, 82, 85. Αν υποθέσουμε ότι το βάρος των φοιτητών ακολουθεί κανονική κατανομή μπορούμε να ισχυριστούμε ότι το δείγμα προέρχεται από πληθυσμό με μέσο βάρος 68 κιλά; Μπορούμε να ισχυριστούμε το ίδιο για μέσο βάρος 58 κιλά;
6. Μία δίαιτα εφαρμόστηκε επί μία εβδομάδα σε 27 ποντικούς και το βάρος σε gr που κέρδισε κάθε ποντικός ήταν:

79.1 81.0 77.3 79.1 80.0 79.1 79.1 77.3 80.2  
 78.8 82.7 78.9 75.4 87.5 78.3 77.2 67.8 79.0  
 68.9 90.0 87.5 78.1 81.9 65.9 63.2 74.8 81.3

Υποθέτοντας ότι το βάρος που κερδίστηκε ακολουθεί κανονική κατανομή μπορούμε να ισχυριστούμε ότι το δείγμα προέρχεται από πληθυσμό με μέση αύξηση βάρους ίση με 78 gr;

7. Μπορούμε να ισχυριστούμε ότι το δείγμα της Άσκησης 3 προέρχεται από πληθυσμό με μέσο IQ ίσο με 105;
8. Ερευνάτε τη διαθεσιμότητα κατοικίας για οικογένειες με χαμηλό εισόδημα με παιδιά στην περιοχή σας. Για να πάρετε μια ιδέα της τιμής πώλησης των σπιτιών, κοιτάζετε σε ένα μεσιτικό γραφείο και σημειώνετε τις τιμές για σπίτια με τρία υπνοδωμάτια. Στον παρακάτω πίνακα αναφέρονται αυτές οι τιμές. Υπολογίστε τη μέση τιμή ενός σπιτιού με τρεις κρεβατοκάμαρες για την περιοχή αυτή και δώστε ένα 95% –διάστημα εμπιστοσύνης για την απάντησή σας. Πιστεύετε ότι αυτό το δείγμα των τιμών είναι αντιπροσωπευτικό;

175900	188500	179950	181000	190900	172250
172500	169950	183600	191450	177500	198400
178250	200000	185900	199000	145200	187000

9. Οι εβδομαδιαίες δαπάνες για διατροφή του πληθυσμού των οικογενειών μιας περιοχής, ακολουθούν κανονική κατανομή. Ένα τυχαίο δείγμα 50 οικογενειών από τον πληθυσμό, έδωσε τα παρακάτω αποτελέσματα για τις εβδομαδιαίες δαπάνες τους για διατροφή (σε ευρώ):

50	40	30	60	80	45	55	36	67	90
100	55	45	70	60	98	100	54	47	65
30	75	45	35	30	45	35	85	65	45
38	40	65	90	95	63	54	87	98	51
45	70	50	40	65	45	50	75	85	90

Να επιβεβαιωθεί η τυχαιότητα και η κανονικότητα του δείγματος. Μπορούμε να ισχυριστούμε ότι η μέση εβδομαδιαία δαπάνη για διατροφή του πληθυσμού της περιοχής είναι ίση 60 ευρώ;

10. Δίνονται τα ημερομίσθια (σε ευρώ) που αντιστοιχούν σε ένα δείγμα 50 Κοινωνιολόγων.

28	27	31	30	29	31	28	32	31	31
31	27	32	29	34	31	29	31	29	30
30	27	28	31	32	29	30	31	30	30
29	33	30	30	31	28	32	30	30	30
29	28	29	32	28	29	29	30	29	31

Τα δεδομένα σε εθνικό επίπεδο υποστηρίζουν ότι το μέσο ημερομίσθιο γι' αυτήν την ειδικότητα είναι ίσο με 30. Να ελεγχθεί η τυχαιότητα και η κανονικότητα του δείγματος. Είναι το μέσο ημερομίσθιο του δείγματος σημαντικά διαφορετικό απ' ότι το μέσο ημερομίσθιο σε εθνικό επίπεδο;