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 Ε1

Επιμέλεια: ΒΑΣΙΟΥ ΓΕΩΡΓΙΑ - ΠΑΠΑΘΑΝΑΣΟΠΟΥΛΟΥ ΧΡΥΣΑΝΘΗ

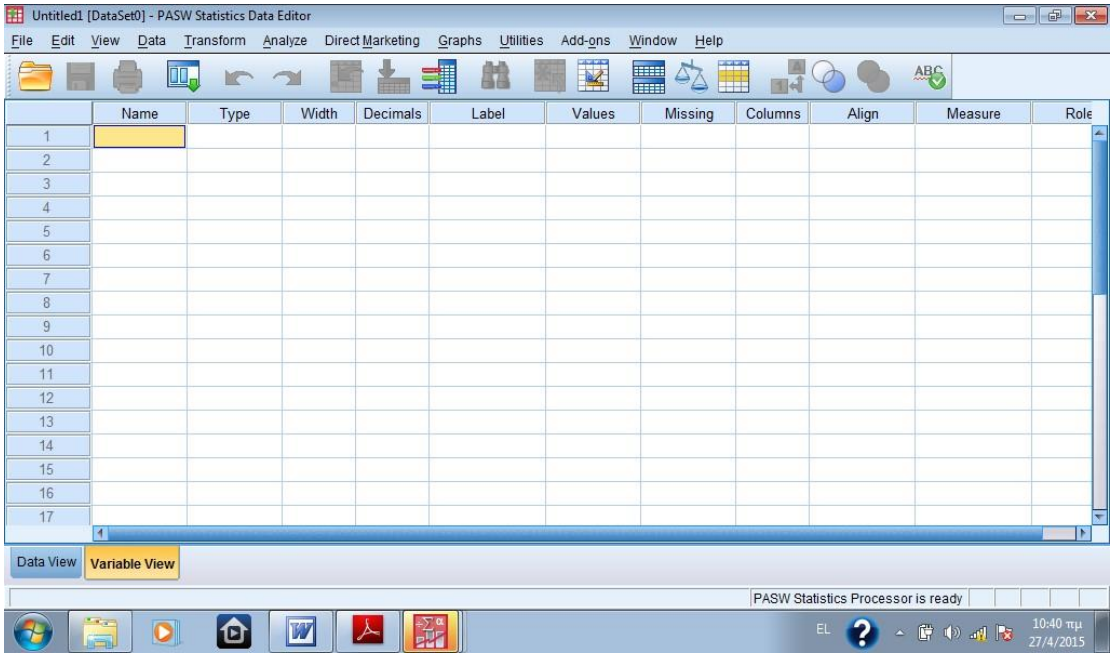
ΜΑΘΗΜΑΤΙΚΟΣ M.Sc ΜΑΘΗΜΑΤΙΚΟΣ M.Sc

Για τη συγγραφή των σημειώσεων βασιστήκαμε σε υλικό που μας παρείχαν οι πρώην συνεργάτες του Τμήματος, κα Αθανασία Μπουμπούλη και κος Κωνσταντίνος Κουνετάς, τους οποίους ευχαριστούμε θερμά.

ΠΕΡΙΕΧΌΜΕΝΑ

ΕΝΟΤΗΤΑ 1Η : ΤΟ ΠΕΡΙΒΑΛΛΟΝ ΤΟΥ SPSS (ΈΚΔΟΣΗ PASW 18)

Με την είσοδό μας στο SPSS βλέπουμε δύο φύλλα Excel, το Variable View και το Data View.



Στο Variable View εισάγουμε τις μεταβλητές του προβλήματός μας καθώς και τα χαρακτηριστικά της κάθε μεταβλητής.

- **Name**

Γράφουμε το όνομα της μεταβλητής έχοντας, όμως, τους παρακάτω περιορισμούς :

- Γράφουμε στα Λατινικά με μικρά ή κεφαλαία γράμματα.
- Το όνομα δεν πρέπει να περιλαμβάνει περισσότερους από 64 χαρακτήρες.
- Το όνομα πρέπει να αρχίζει με γράμμα ή με ένα από τα σύμβολα @ , #, $ και δεν πρέπει να τελειώνει με τελεία ή κάτω παύλα.
- Το όνομα μπορεί να περιέχει γράμματα, ψηφία ή τα σύμβολα @, #, _, $ αλλά όχι σημεία στίξης (πλην της τελείας), αστεράκια και κενά.

- Στο όνομα δεν πρέπει να περιλαμβάνονται οι λέξεις AND, NOT, BY, ALL, OR, TO, ADD, NE, EQ, LE, LT, GT, GE, και WITH.

- **Type**

Επιλέγουμε *numeric* όταν η μεταβλητή είναι ποσοτική και *string* όταν η μεταβλητή είναι ποιοτική.

- **Width**

Επιλέγουμε το πλήθος των χαρακτήρων των δεδομένων μας

- **Decimals**

Επιλέγουμε το πλήθος των δεκαδικών ψηφίων που έχουν τα δεδομένα μας αν αυτά είναι αριθμοί.

- **Label**

Γράφουμε μία επεξήγηση για το όνομα της μεταβλητής μας, η οποία θα εμφανίζεται και στην παρουσίαση των αποτελεσμάτων.

- **Values**

Αυτή την επιλογή τη χρησιμοποιούμε μόνο σε ποιοτικές μεταβλητές για να κωδικοποιήσουμε τις τιμές της.

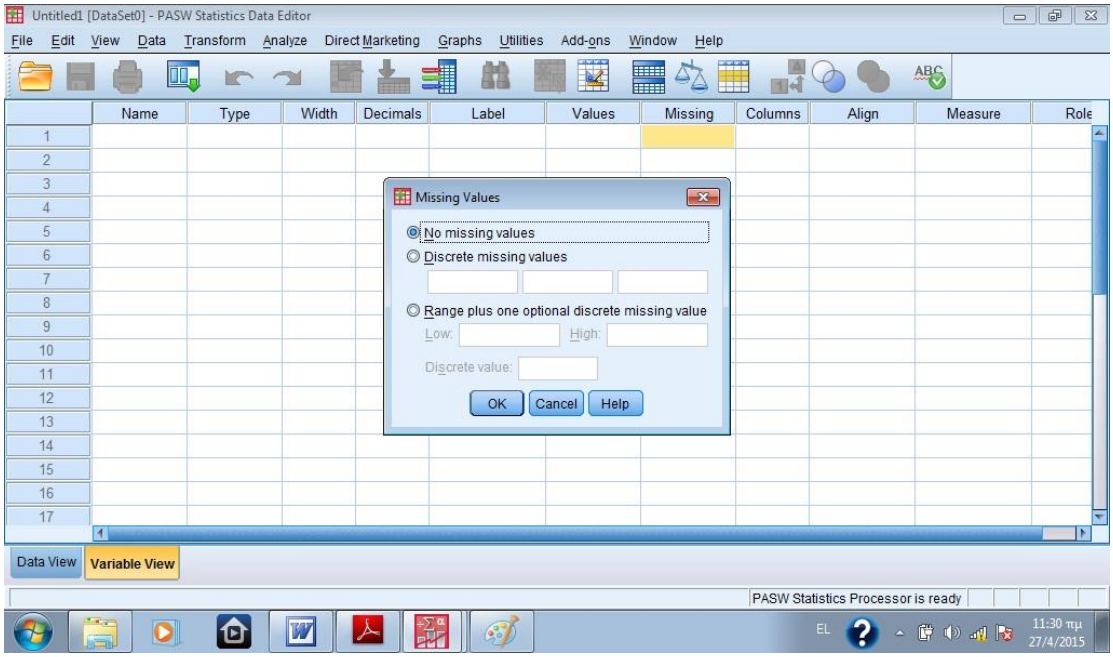
- **Missing**

Σε αυτή την φόρμα σημειώνουμε τις τιμές ή τους κωδικούς των τιμών τις οποίες δεν θέλουμε, για κάποιο λόγο, να συμπεριλάβουμε στην επεξεργασία των αποτελεσμάτων. Κλικάροντας πάνω στο πρώτο κελί που είναι κάτω από τη λέξη missing ανοίγει το παράθυρο Missing Values στο οποίο μπορούμε να επιλέξουμε :

- *No missing values* αν θέλουμε όλες οι τιμές να συμπεριληφθούν στην επεξεργασία των αποτελεσμάτων

- *Discrete missing values* αν θέλουμε να μη συμπεριλάβουμε τις τιμές που θα αναγράψουμε στα τρία παράθυρα που βλέπουμε από κάτω

- *Range plus one optional discrete missing value* αν θέλουμε να μη συμπεριλάβουμε όλες τις τμές από την τιμή που γράφω στο παράθυρο Low έως και την τιμή που γράφω στο παράθυρο High. Αν θέλουμε από αυτές τις τιμές να εξαιρέσουμε κάποια, τη γράφουμε στο παράθυρο Discrete value.



- **Columns**

Ρυθμίζουμε το πλήθος των χαρακτήρων που θέλουμε να χωρούν σε κάθε κελί της στήλης.

- **Align**

Επιλέγουμε τη στοίχιση που θέλουμε να έχουν τα δεδομένα μέσα στα κελιά.

- **Measure**

Σε αυτή τη φόρμα μπορούμε να επιλέξουμε :

- *Scale* αν η μεταβλητή μας είναι ποσοτική
- *Nominal* αν η μεταβλητή είναι ποιοτική με ονομαστικά δεδομένα
- *Ordinal* αν η μεταβλητή είναι ποιοτική με διατακτικά δεδομένα

Η γραμμή μενού περιλαμβάνει τις παρακάτω επιλογές :

- **File :** Μπορούμε να ανοίξουμε ένα αρχείο (νέο ή ήδη υπάρχον), να αποθηκεύσουμε, να εκτυπώσουμε κ.ο.κ.

- **Edit :** Μπορούμε να τροποποιήσουμε ή να αντιγράψουμε τμήματα του αρχείου δεδομένων.

- **View :** Μπορούμε να προσαρμόζουμε τα διάφορα στοιχεία του παραθύρου ανάλογα με τις επιλογές μας.

- **Data :** Μπορούμε να πραγματοποιήσουμε αλλαγές στα δεδομένα.

- **Transform :** Μπορούμε να πραγματοποιήσουμε αλλαγές στις μεταβλητές.

- **Analyze :** Πραγματοποιούμε τη στατιστική ανάλυση των δεδομένων.

- **Direct Marketing :** Περιέχει εφαρμογές για διαχείριση επιχειρησιακών δεδομένων.

- **Graphs :** Δημιουργούμε γραφικές παραστάσεις και διαγράμματα.

- **Utilities :** Πρόκειται για μια επιλογή γενικών χρήσεων. Για παράδειγμα, δίνονται πληροφορίες για μια μεταβλητή ή ένα αρχείο.

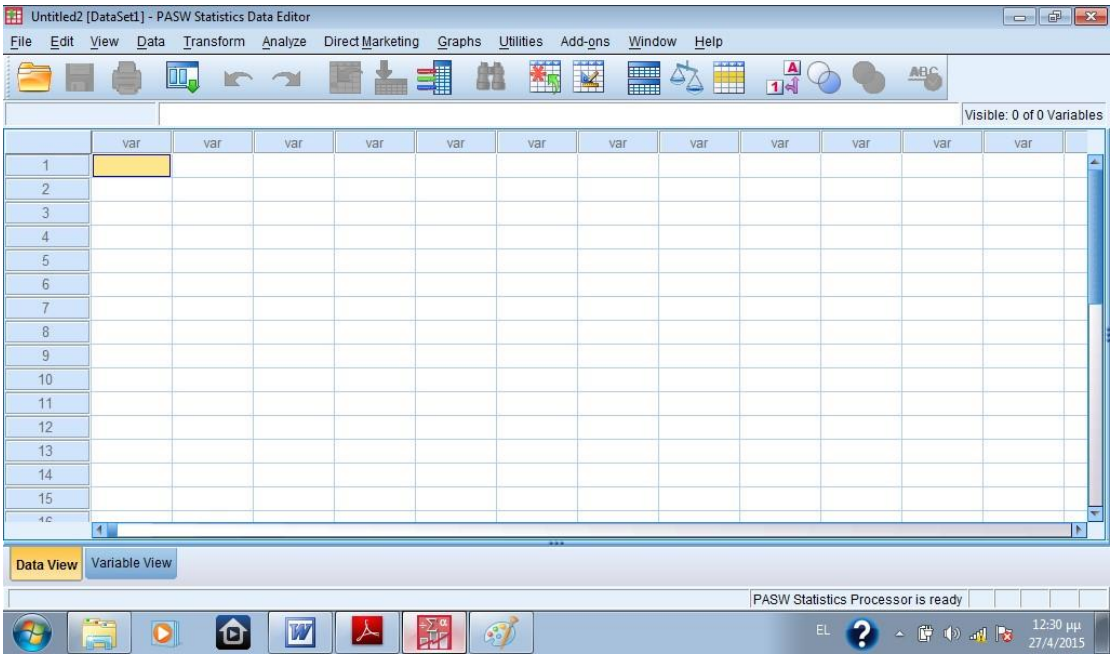
- **Add-ons :** Περιλαμβάνει πρόσθετες παροχές της IBM (εταιρείας-κατόχου του SPSS)

- **Window :** Μπορούμε να μεταβούμε σε κάποιο άλλο ενεργό παράθυρο.

- **Help :** Προσφέρει διάφορα είδη βοήθειας.

Κάτω από τη γραμμή μενού υπάρχει η γραμμή εργαλείων η οποία περιέχει με μορφή εικόνας ή σχήματος εντολές που ήδη βρίσκονται στη γραμμή μενού.

Στο δεύτερο φύλλο του Excel, στο Data View, εισάγουμε τα δεδομένα δηλαδή τις τιμές της κάθε μεταβλητής.


Untitled2 [DataSet1] - PASW Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 0 of 0 Variables
var
Data View
Variable View
PASW Statistics Processor is ready
EL
12:30 μμ
27/4/2015

ΠΑΡΑΓΡΑΦΟΣ 2.1 : ΕΙΣΑΓΩΓΗ ΜΕΤΑΒΛΗΤΩΝ ΚΑΙ ΔΕΔΟΜΕΝΩΝ: ΕΝΤΟΛΕΣ COMPUTE VARIABLE , RECODE

- **Compute Variable (Υπολογισμός τιμών μεταβλητής)**

Με αυτή την εντολή υπολογίζονται οι τιμές μιας νέας μεταβλητής (ή επαναϋπολογίζονται οι τιμές μιας ήδη υπάρχουσας) με βάση τους μετασχηματισμούς των τιμών άλλων μεταβλητών.

- **Recode into Different Variables (Επανακωδικοποίηση σε διαφορετική μεταβλητή)**

Η επανακωδικοποίηση σε διαφορετική μεταβλητή αφορά την αντικατάσταση των τιμών μιας μεταβλητής με άλλες, μόνο που τώρα οι νέες τιμές που δημιουργούνται, καταχωρούνται σε μια νέα μεταβλητή που ορίζει ο χρήστης, ενώ η αρχική μεταβλητή παραμένει αναλλοίωτη.

ΆΣΚΗΣΗ

Έστω ότι επιλέγεται ένα τυχαίο **δείγμα 35 παιδιών προσχολικής ηλικίας**. Για κάθε παιδί εξετάζεται **ο δείκτης νοημοσύνης του**, **το ύψος του**, **ο χρόνος σε δευτερόλεπτα που διανύει τα 100 μέτρα**, **η συμπεριφορά του** και **η οικονομική κατάσταση της οικογένειας του**. Τα δεδομένα παρουσιάζονται στον πίνακα που ακολουθεί όπου στη στήλη **Φύλο Α= Αγόρι, Θ= Κορίτσι**, στη στήλη **Διαγωγή Α= Κοσμιωτάτη και Β= Κοσμία**, στη στήλη **Οικονομική Κατάσταση Α=0-450, Β= 450600, Γ=600-900 και Δ= 900 ευρώ και άνω**.

1. Να καταχωρηθούν τα δεδομένα του πίνακα στο S.P.S.S..

Α/Α	ΦΥΛΟ	ΔΙΑΓΩΓΗ	ΟΙΚ. ΚΑΤΑΣΤΑΣΗ	IQ	ΥΨΟΣ	ΧΡΟΝΟΣ

1	Α	Β	Β	11	95	22
2	Θ	Α	Γ	90	98	25
3	Θ	Α	Γ	90	92	18
4	Θ	Α	Γ	90	104	19
5	Α	Α	Α	104	85	21
6	Α	Α	Β	72	96	20
7	Θ	Β	Β	105	89	21
8	Α	Α	Δ	93	103	22
9	Α	Α	Γ	99	110	18
10	Α	Α	Β	93	85	27
11	Α	Α	Β	84	94	30
12	Θ	Α	Α	95	98	21
13	Θ	Α	Γ	93	96	24
14	Α	Α	Δ	78	99	26
15	Θ	Α	Β	108	83	19
16	Θ	Β	Β	100	87	27
17	Α	Α	Α	81	85	25

18	Α	Α	Γ	77	97	24
19	Α	Α	Γ	67	96	23
20	Α	Α	Α	100	107	28
21	Α	Α	Β	104	102	29
22	Θ	Α	Δ	111	106	19
23	Θ	Α	Δ	122	95	20
24	Α	Α	Β	99	82	28
25	Θ	Α	Β	108	94	31
26	Θ	Α	Α	126	90	19
27	Α	Β	Α	90	90	23
28	Α	Α	Γ	110	96	32
29	Α	Α	Γ	117	87	27
30	Θ	Α	Α	119	97	24
31	Θ	Α	Δ	105	90	22
32	Α	Β	Β	100	107	18
33	Θ	Α	Α	75	92	25
34	Α	Α	Γ	96	98	30
35	Θ	Α	Δ	81	95	23

ΛΥΣΗ

Αρχικά πρέπει **να προσδιοριστούν οι μεταβλητές** που θα πρέπει να δημιουργηθούν, και **να γίνει η διάκριση μεταξύ ποσοτικών και ποιοτικών μεταβλητών** έτσι ώστε να καταχωρηθούν σωστά.

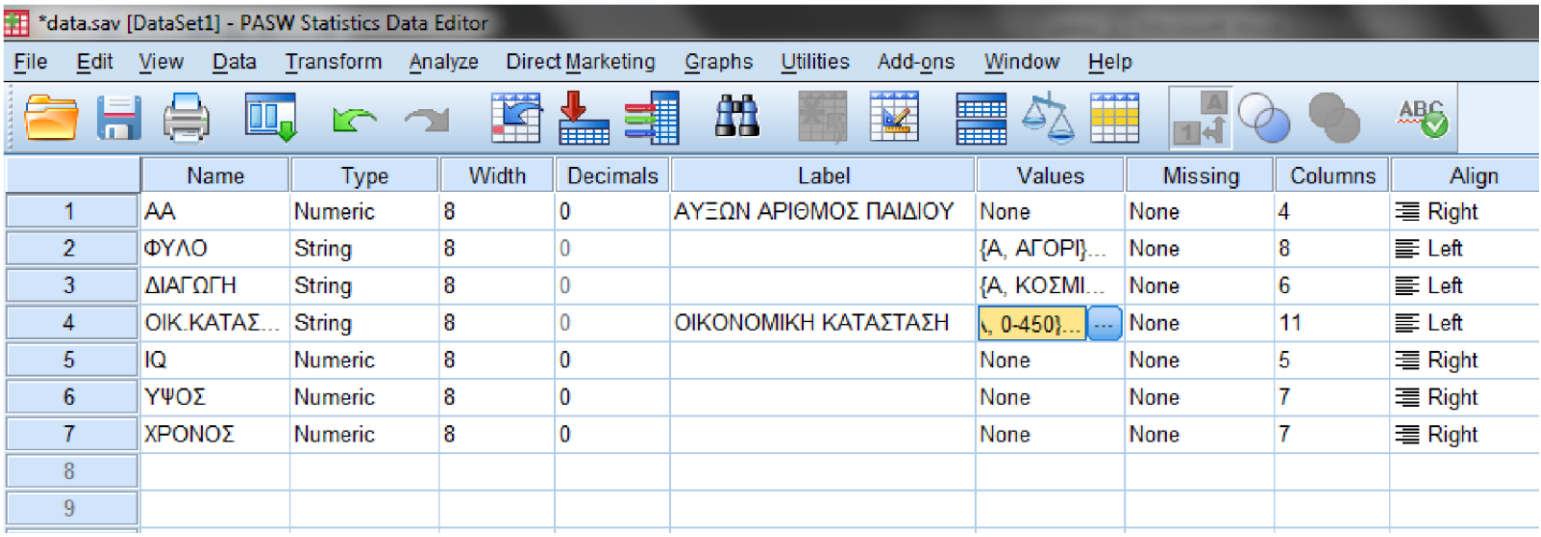
Μεταβλητές	Κατηγορία
Α/Α	**Ποσοτική**
ΦΥΛΟ	**Ποιοτική**
ΔΙΑΓΩΓΗ	**Ποιοτική**
ΟΙΚ. ΚΑΤΑΣΤΑΣΗ	**Ποιοτική**
IQ	**Ποσοτική**
ΥΨΟΣ	**Ποσοτική**
ΧΡΟΝΟΣ	**Ποσοτική**

Πίνακας 1

Αφού **οριστούν οι μεταβλητές στο φύλλο Variable View**, στην συνέχεια θα πρέπει να **καταχωρηθούν τα δεδομένα του παραπάνω πίνακα στο φύλλο Data View** του S.P.S.S..

Ορισμός των μεταβλητών στο φύλλο Variable View:

Στο παράθυρο Variable View και **στο πλαίσιο Name** εισάγουμε τα κωδικοποιημένα



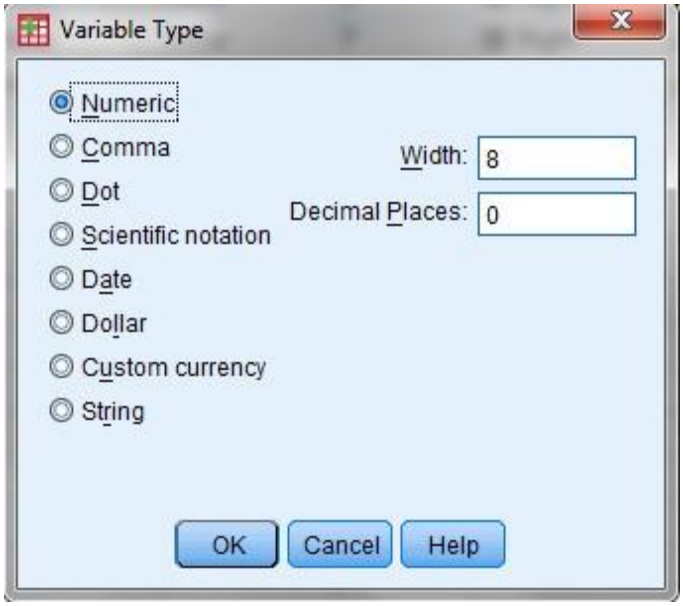
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align
1	ΑΑ	Numeric	8	0	ΑΥΞΩΝ ΑΡΙΘΜΟΣ ΠΑΙΔΙΟΥ	None	None	4	Right
2	ΦΥΛΟ	String	8	0		{Α, ΑΓΟΡΙ}...	None	8	Left
3	ΔΙΑΓΩΓΗ	String	8	0		{Α, ΚΟΣΜΙ...	None	6	Left
4	ΟΙΚ.ΚΑΤΑΣ...	String	8	0	ΟΙΚΟΝΟΜΙΚΗ ΚΑΤΑΣΤΑΣΗ	ι, 0-450}...	None	11	Left
5	IQ	Numeric	8	0		None	None	5	Right
6	ΥΨΟΣ	Numeric	8	0		None	None	7	Right
7	ΧΡΟΝΟΣ	Numeric	8	0		None	None	7	Right
8									
9									

ονόματα των μεταβλητών μας, έστω Sex, Time. Στο **πεδίο Label** δηλώνεται η πλήρης περιγραφή του ονόματος της μεταβλητής που βοηθά στην καλύτερη παρουσίαση των αποτελεσμάτων μας. Με αυτόν τον τρόπο δηλώνουμε την ονομασία που θα εμφανίζεται στους πίνακες των αποτελεσμάτων των αναλύσεων που θα ακολουθήσουν π.χ. Φύλο, Διαγωγή, Οικονομική Κατάσταση, Δείκτης Νοημοσύνης, Ύψος, Χρόνος σε δευτερόλεπτα.

Καθορισμός του τύπου της μεταβλητής (Variable Type):

Στο πλαίσιο Variable Type το λογισμικό με βάση τις τιμές που πληκτρολογούμε καθορίζει αυτόματα τον τύπο της μεταβλητής, έχοντας ως προεπιλογή να τις εμφανίζει αριθμητικές (numeric) με 2 δεκαδικά ψηφία (Decimals Places) και συνολικό μήκος (δηλώνεται στο πλαίσιο Width) 8 θέσεων. Για τον υπολογισμό του μήκους μίας μεταβλητής λαμβάνονται υπόψη το πρόσημο, το ακέραιο μέρος, η δεκαδική τελεία.

Αν τα δεδομένα είναι τέτοια που παραβιάζονται αυτές οι προεπιλογές πρέπει να τις τροποποιήσουμε κατάλληλα ανοίγοντας την παρακάτω καρτέλα:



Στην συγκεκριμένη άσκηση ορίζουμε τις μεταβλητές σύμφωνα με τις κατηγορίες του Πίνακα 1, και τροποποιούμε τον αριθμό των δεκαδικών ψηφίων, ορίζουμε Decimal Places 0), καθώς όλες οι ποσοτικές μεταβλητές της άσκησης δεν έχουν δεκαδικά ψηφία.

Στην εισαγωγή των ποιοτικών μεταβλητών επιλέγουμε Variable Type: String

Ετικέτες τιμών μιας μεταβλητής (Value Labels):

Για να μην ανατρέχουμε συνεχώς στην άσκηση (ή στο ερωτηματολόγιο) προκειμένου να θυμηθούμε τι σημαίνει για κάθε **ποιοτική μεταβλητή** ο κάθε κωδικός της, είναι χρήσιμο όλοι οι κωδικοί των μεταβλητών να καταγράφονται στο πλαίσιο **Values** του παραθύρου **Variable View**.

Επομένως, στο πεδίο αυτό ουσιαστικά εισάγουμε στο λογισμικό τις συμβάσεις τις οποίες κάναμε κατά την καταχώρηση των δεδομένων. Αυτό επιτυγχάνεται κλικάροντας το δεξί άκρο του κελιού, το οποίο σχηματίζεται από την μεταβλητή και τη στήλη Values.

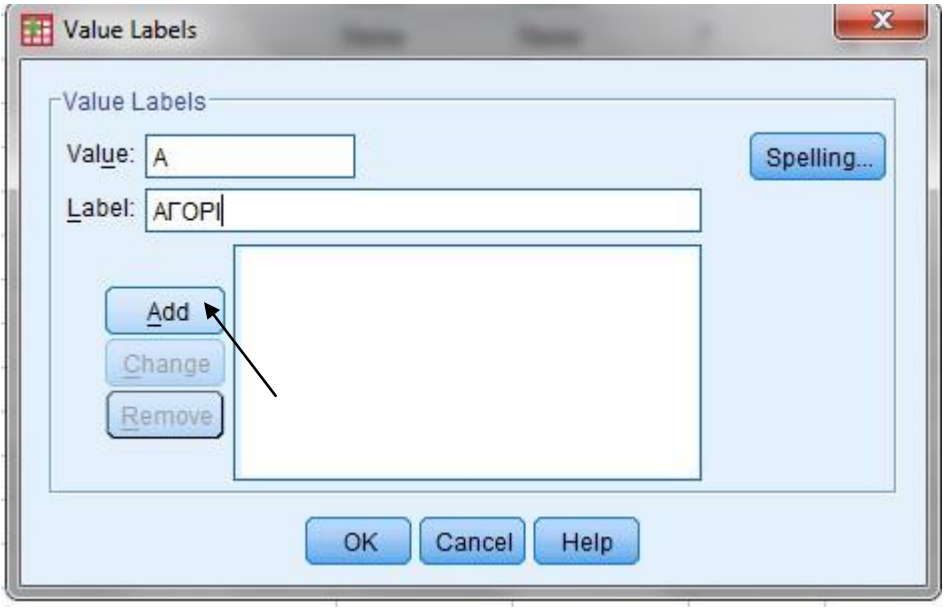
Στη συγκεκριμένη άσκηση θα εισάγουμε ετικέτες για τις τιμές των εξής μεταβλητών:

Για την μεταβλητή: **Φύλο, όπου Α= ΑΓΟΡΙ, Θ= ΚΟΡΙΤΣΙ**.

Για την μεταβλητή: **Διαγωγή, όπου Α= ΚΟΣΜΙΩΤΑΤΗ και Β= ΚΟΣΜΙΑ**,

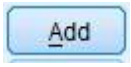
Για την μεταβλητή: **Οικονομική Κατάσταση, όπου Α=0-450, Β= 450-600, Γ=600-900 και Δ= 900 ευρώ και άνω**.

Η εισαγωγή των ετικετών γίνεται ως εξής:



Value: Α

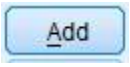
Label: ΑΓΟΡΙ

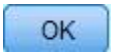


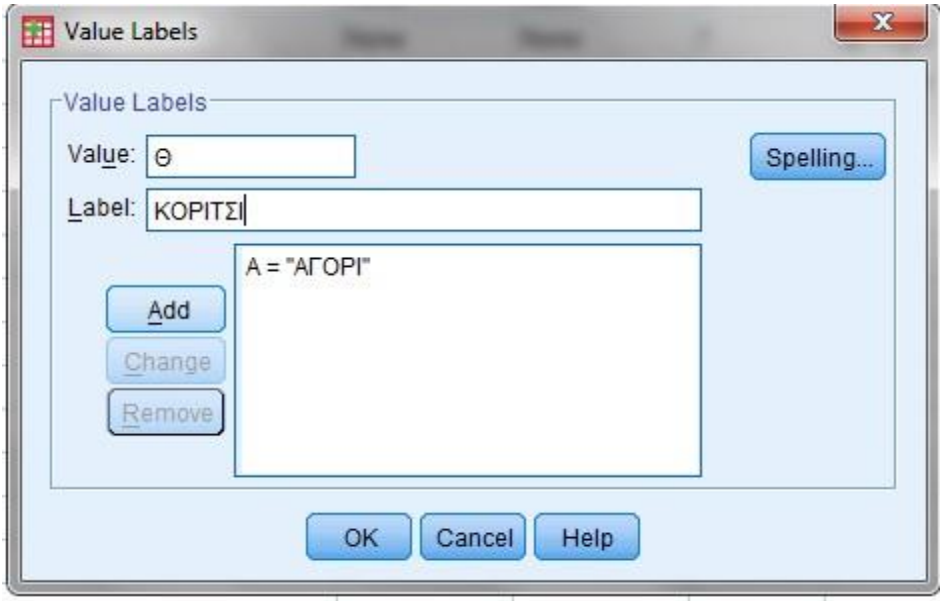
Στη συνέχεια,

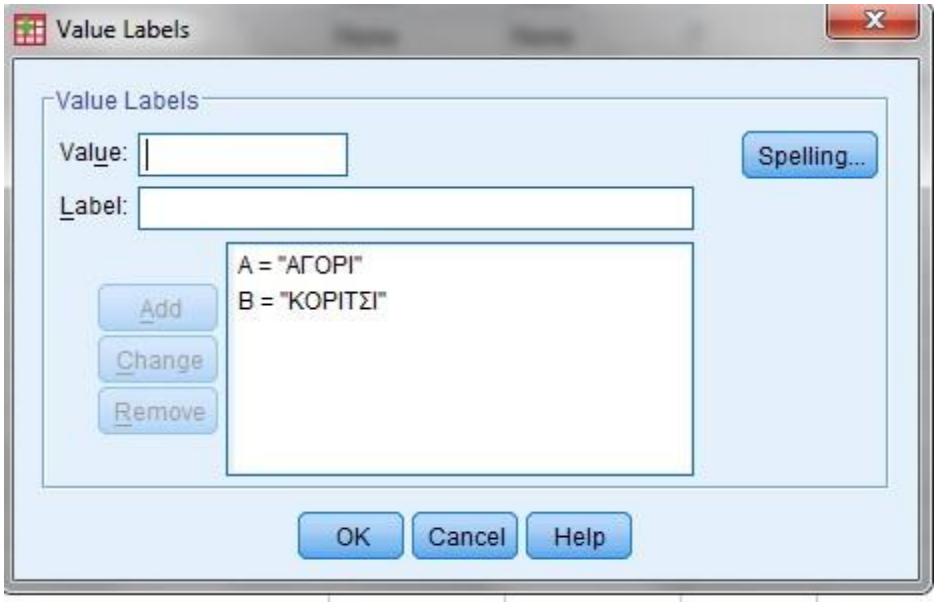
Value: Θ

Label: ΚΟΡΙΤΣΙ



και 





Ομοίως θα γίνει και η καταχώρηση των ετικετών για τις υπόλοιπες ποιοτικές μεταβλητές της άσκησης.

Δημιουργία Νέων Μεταβλητών – Μετασχηματισμός Δεδομένων (COMPUTE)

2. Για κάθε παιδί της άσκησης έχει καταγραφεί ο χρόνος σε δευτερόλεπτα που διανύει τα 100 μέτρα. Να δημιουργηθεί μία νέα μεταβλητή που θα μετρά τον προαναφερθέντα χρόνο με μονάδα μέτρησης το λεπτό.

ΛΥΣΗ

Αρχικά πρέπει να αποφασίσουμε τον μετασχηματισμό που θέλουμε να υλοποιήσουμε στα δεδομένα μας.

Συγκεκριμένα εδώ θέλουμε να μετασχηματίσουμε τις τιμές της μεταβλητής ΧΡΟΝΟΣ ως εξής: να τις εκφράσουμε σε λεπτά, δηλαδή: Π.χ.

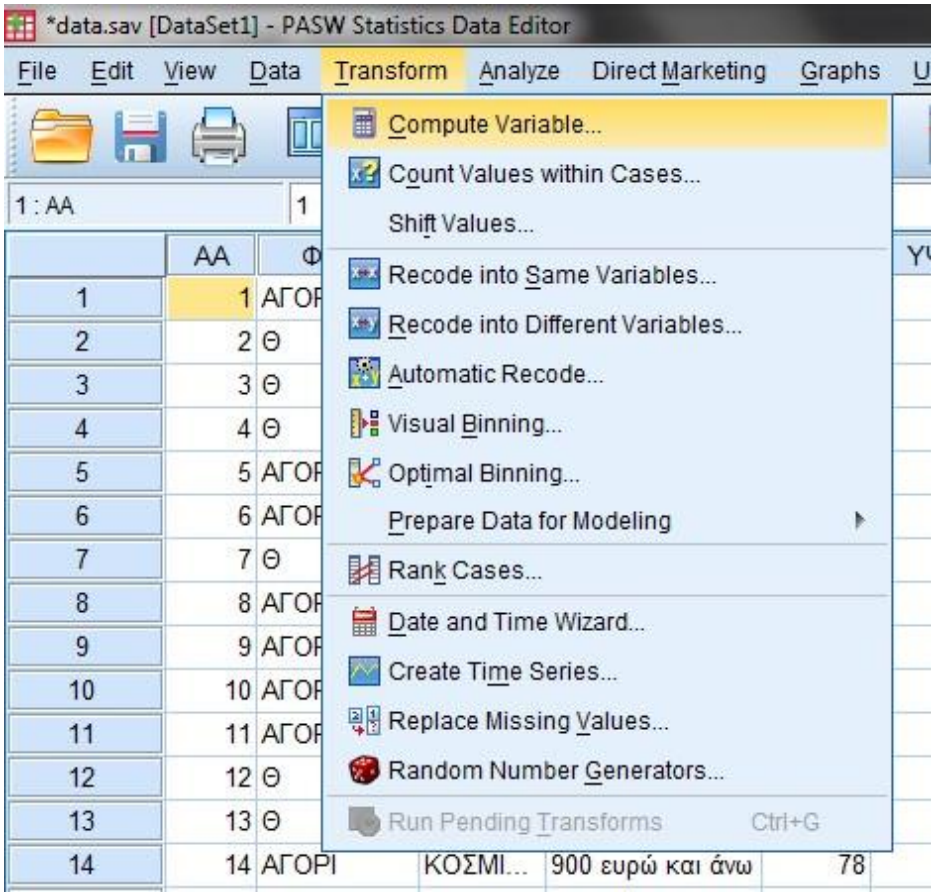
1 λεπτό 60 δευτερόλεπτα

Χ 22 δευτερόλεπτα

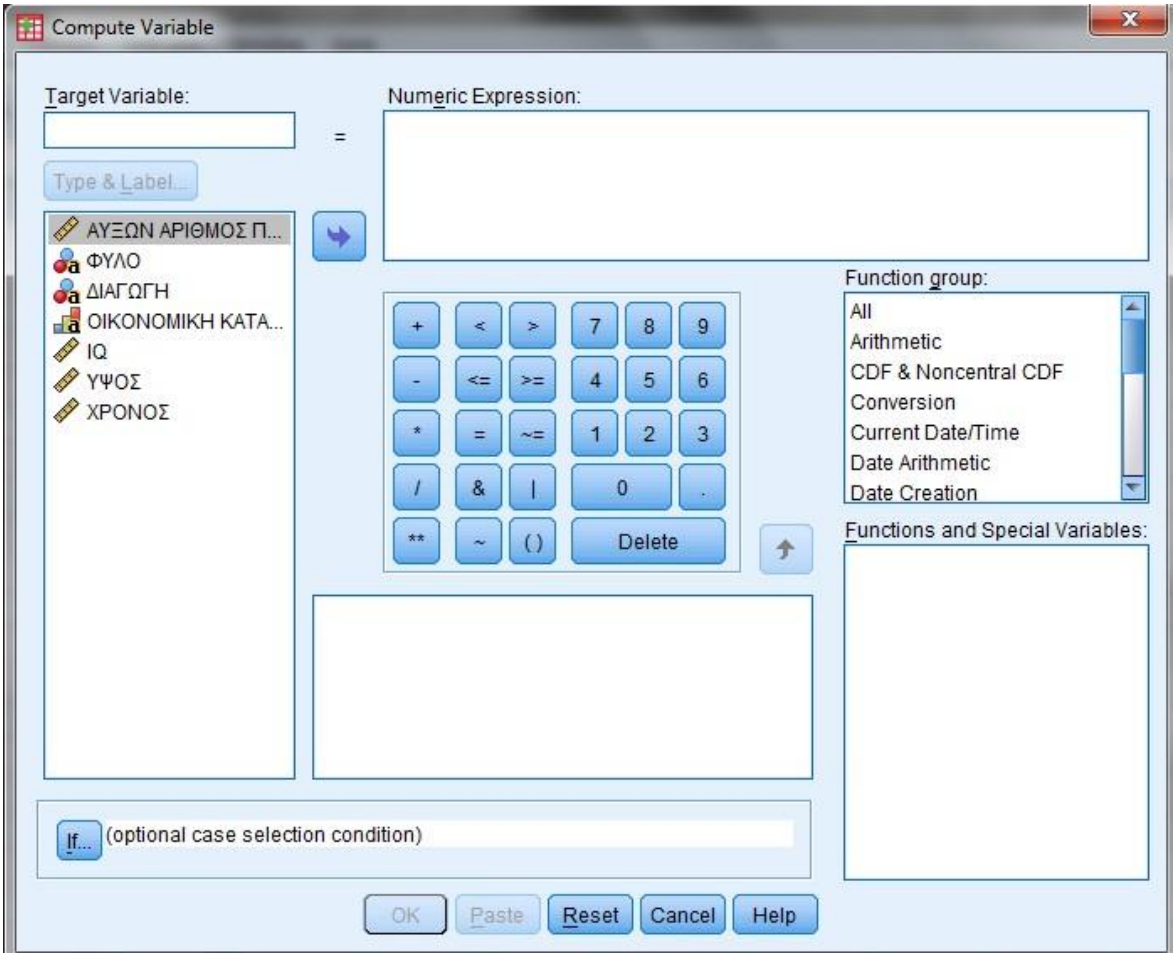
Χ= 22/60 λεπτά

Για την υλοποίηση του παραπάνω μετασχηματισμού ακολουθούμε τα ακόλουθα βήματα:

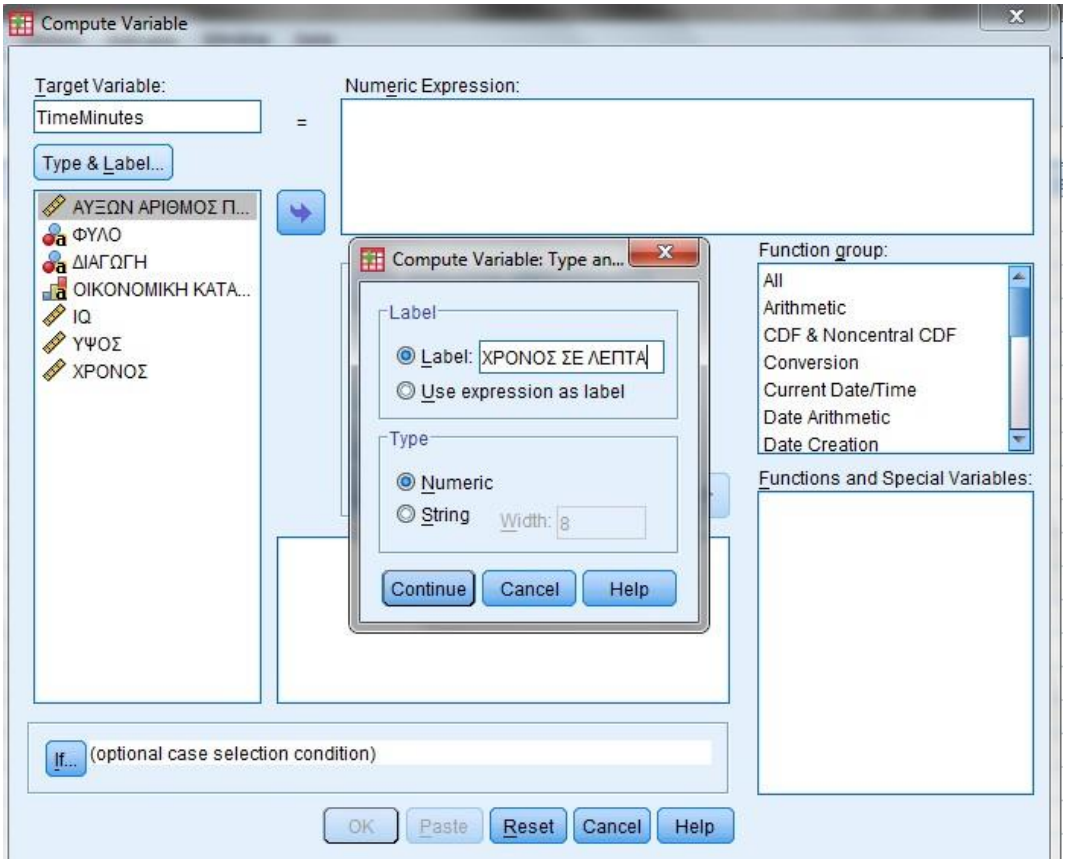
1. Από το βασικό μενού επιλέγουμε **Transform→Compute Variable**



2. Στο παράθυρο Compute Variable στο πλαίσιο **Target Variable** δηλώνουμε το όνομα της νέας μεταβλητής, έστω TimeMinutes.



3. Έπειτα, από το πλαίσιο **Type & Label** οδηγούμαστε σε ένα παράθυρο όπου εκεί μπορούμε να ορίσουμε μια πιο λεπτομερή περιγραφή της μεταβλητής που θα δημιουργήσουμε π.χ. ΧΡΟΝΟΣ ΣΕ ΛΕΠΤΑ, καθώς και να ορίσουμε και τον τύπο της: Numeric.

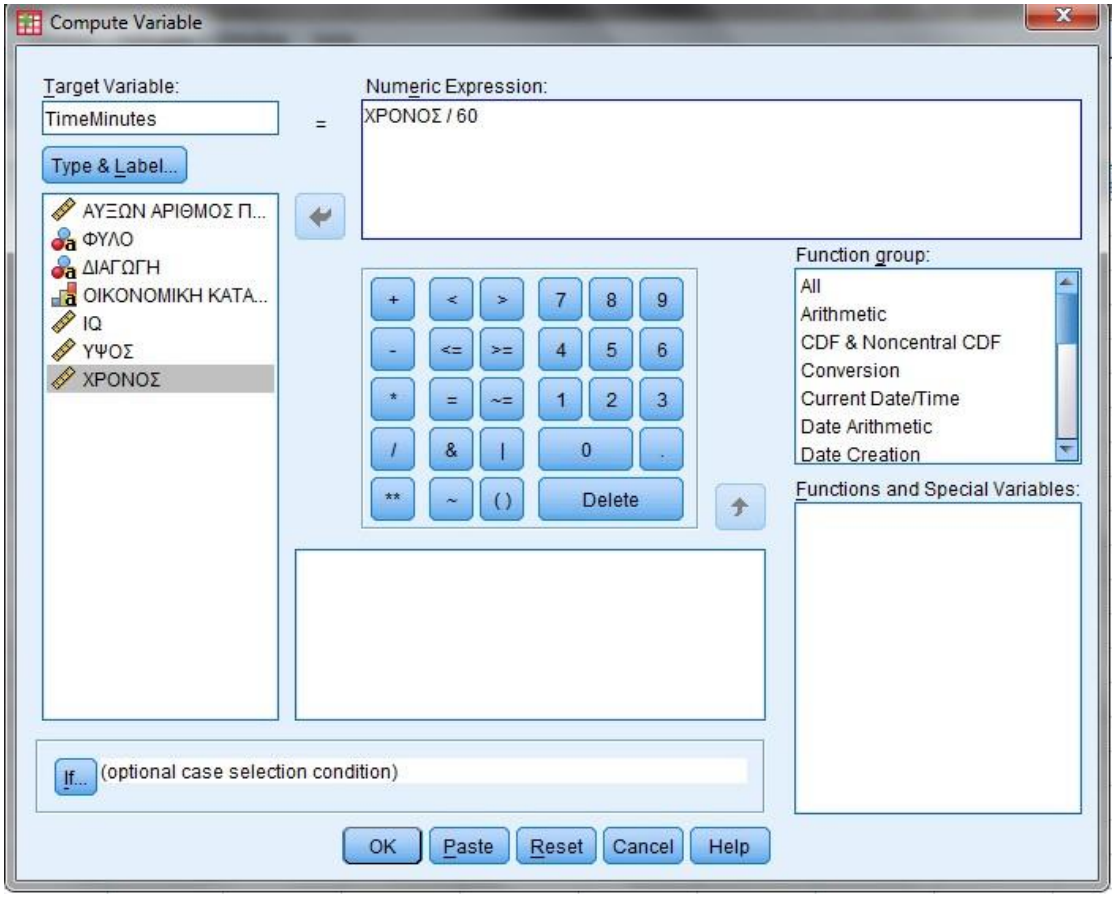


4. Στο πλαίσιο **Numeric Expression** σχηματίζουμε τον κατάλληλο μετασχηματισμό κάνοντας χρήση του Calculator Pad (αν χρειαστεί να χρησιμοποιήσουμε δεκαδικούς τότε κάνουμε χρήση της τελείας) και των συναρτήσεων που δίνονται στο πλαίσιο Function Group. Στο πλαίσιο αυτό μεταξύ άλλων έχουμε την ακόλουθη ομαδοποίηση των συναρτήσεων:

- **All**: δίνονται όλες οι συναρτήσεις σε αλφαβητική σειρά διάταξης.
- **Arithmetic**: δίνονται αριθμητικές συναρτήσεις όπως η απόλυτη τιμή (Abs), το
- συνημίτονο (Cos), το ημίτονο (Sin), ο δεκαδικός λογάριθμος (Lg10), ο φυσικός λογάριθμος (Ln), η τετραγωνική ρίζα (Sqrt) κ.ά.
- **CDF and Noncentral CDF**: δίνονται οι τιμές των αθροιστικών συναρτήσεων κατανομών (cdf=cumulative distribution function) και των μη κεντρικών αθροιστικών συναρτήσεων κατανομών ειδικών, γνωστών κατανομών όπως η διωνυμική, η εκθετική, η κανονική, η μη κεντρική t κατανομή κ.ά.
- **Inverse DF**: δίνει την τιμή της κατανομής για την οποία η αθροιστική συνάρτηση κατανομής είναι ίση με προκαθορισμένη πιθανότητα.

- **PDF and NonCentral PDF**: μας δίνει την τιμή της συνάρτησης πυκνότητας πιθανότητας ή της συνάρτησης πιθανότητας για γνωστές τιμές των παραμέτρων της κατανομής σε προκαθορισμένη τιμή.
- **Random Numbers**: δημιουργεί μία στήλη δεδομένων που αποτελούν ένα τυχαίο δείγμα από διάφορους πληθυσμούς π.χ. από έναν εκθετικό ή κανονικό πληθυσμό.
- **Statistical**: δίνονται στατιστικές συναρτήσεις όπως είναι η μέση τιμή, η τυπική απόκλιση, η διακύμανση, ο συντελεστής μεταβλητότητας κ.ά.

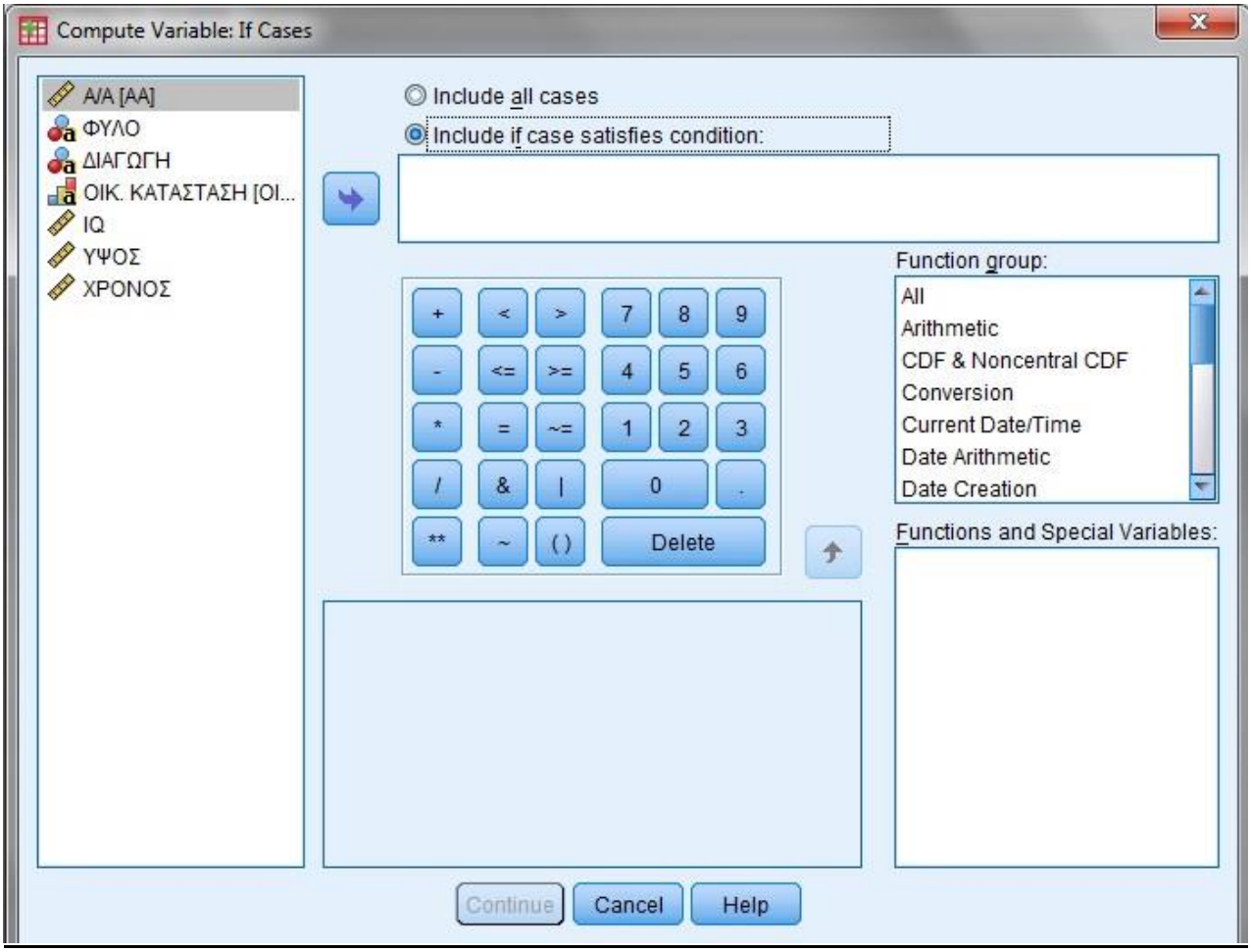
Για την συγκεκριμένη άσκηση έχουμε στο πεδίο Numeric Expression δημιουργούμε τον τύπο: ΧΡΟΝΟΣ/ 60.



Επιλέγουμε την μεταβλητή ΧΡΟΝΟΣ από το παράθυρο αριστερά στο οποίο εμφανίζονται όλες οι μεταβλητές και με το την μετακινούμε στο πλαίσιο Numeric Expression. Από το Calculator Pad / και 6 και 0 επιλέγουμε για να σχηματίσουμε τον τύπο που θα μετασχηματίσει τις τιμές τις μεταβλητής χρόνος και θα δώσει τις τιμές της νέας μεταβλητής.

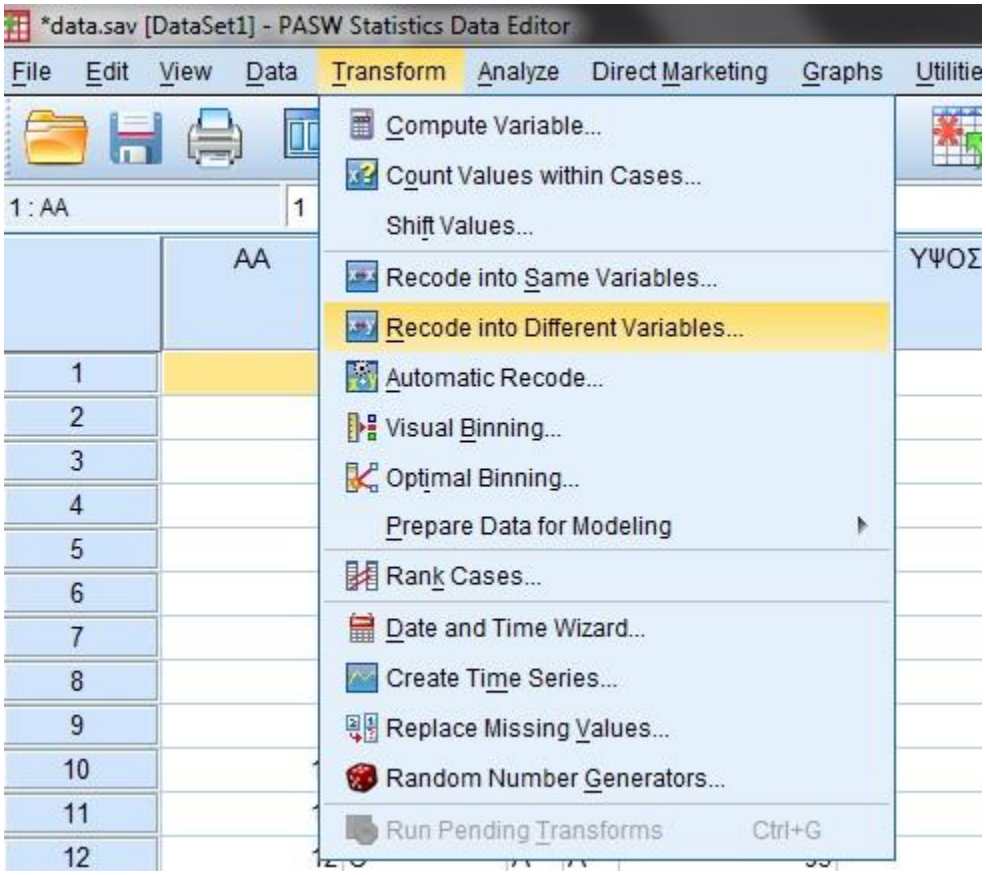
Χρήσιμες πληροφορίες για την εντολή Compute Variable:

Χρησιμοποιούμε την επιλογή **If…(optional case selection condition)** αν η ύπαρξη τιμών της νέας μεταβλητής εξαρτάται από την ικανοποίηση ή όχι μίας συνθήκης ή έκφρασης μίας άλλης μεταβλητής. Εφόσον θέλουμε να χρησιμοποιηθούν **μόνο** οι τιμές της μεταβλητής που ικανοποιούν κάποια συνθήκη, επιλέγουμε το πλαίσιο **Include if case satisfies condition**, και στο πλαίσιο που ακολουθεί σχηματίζεται η επιθυμητή συνθήκη. Αφού δηλωθεί η αναγκαία συνθήκη πατάμε Continue.



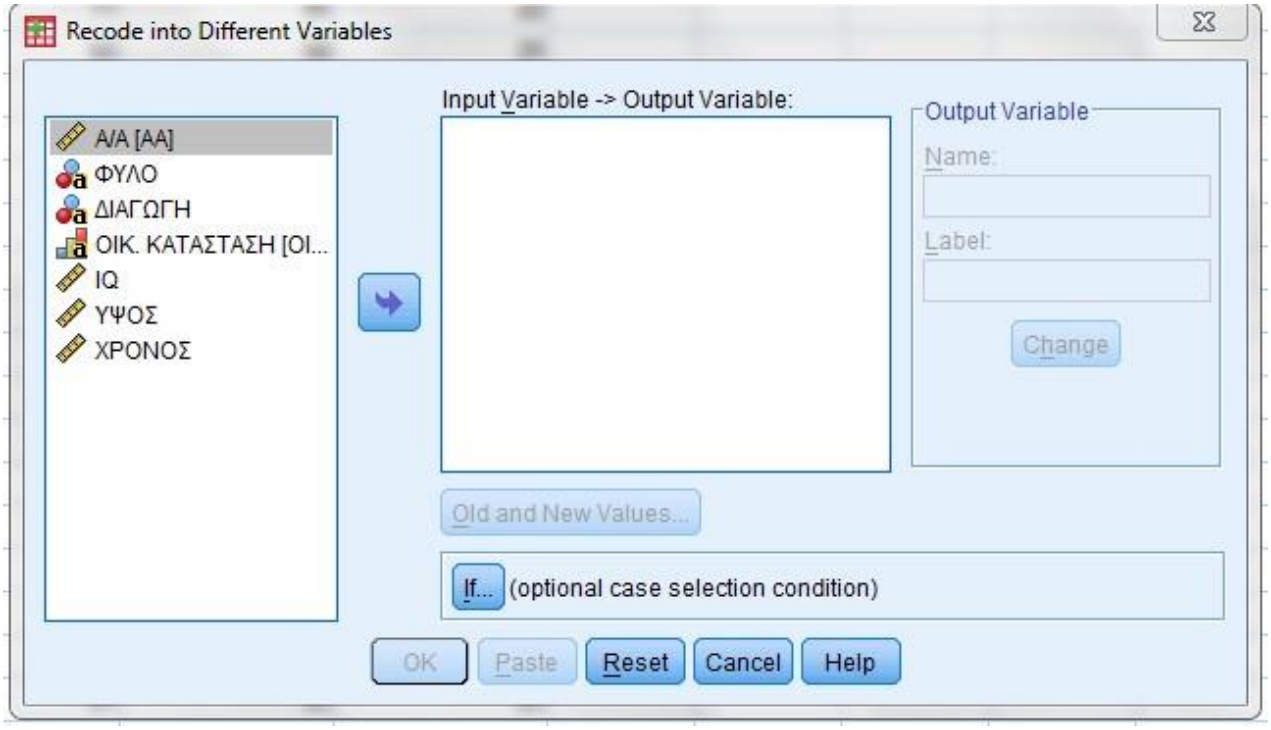
Επανακωδικοποίηση Μεταβλητών (RECODE)

Πολλές φορές εκτός από το μετασχηματισμό των δεδομένων μίας μεταβλητής, χρειάζεται να γίνει επανακωδικοποίηση όλης της μεταβλητής. Αυτό είναι προτιμότερο να γίνεται με τη διαδικασία **Recode→Into Different Variables**.



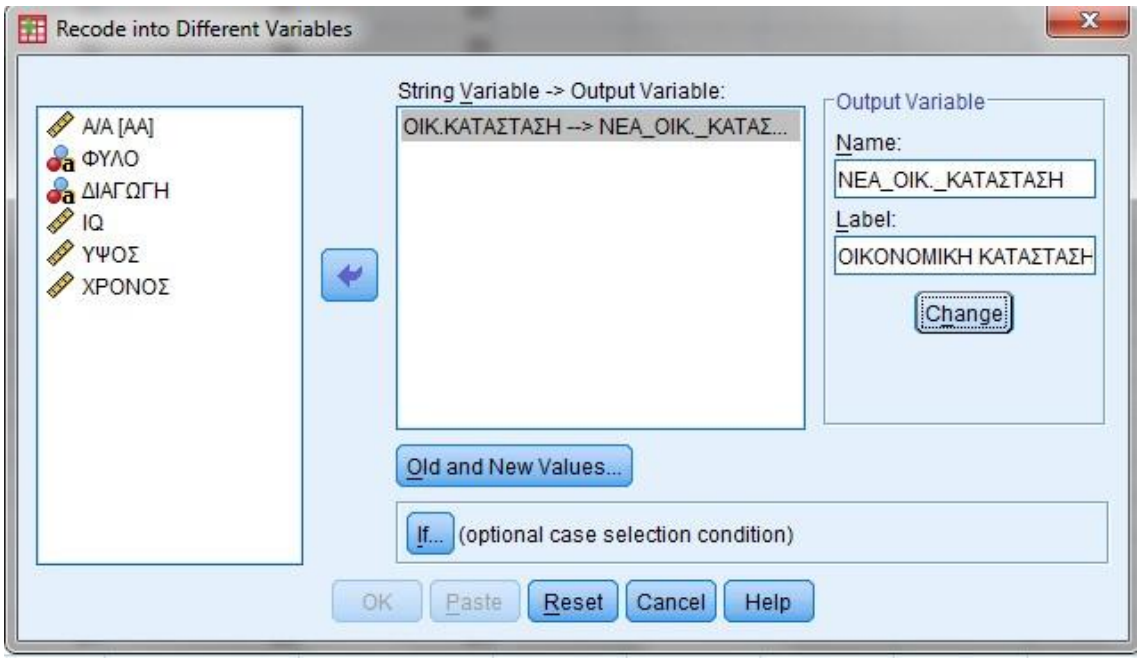
Για παράδειγμα, στα συγκεκριμένα δεδομένα να γίνει μία νέα κωδικοποίηση για τη μεταβλητή Οικονομική Κατάσταση, σύμφωνα με την οποία: όσοι ανήκουν στις κατηγορίες Α και Β θα ανήκουν σε μία νέα κατηγορία: 0 – 600 ευρώ, και όσοι στις Γ και Δ σε μία νέα κατηγορία: 600 ευρώ και άνω. Δηλαδή, όσων παιδιών οι οικογένειες έχουν εισόδημα από 0-600 ευρώ αποτελούν μία κατηγορία, ενώ τα υπόλοιπα μία άλλη.

Χρησιμοποιώντας την εντολή **Recode→Into Different Variables** ενεργοποιείται το παρακάτω παράθυρο διαλόγου:



Στο πλαίσιο **Input Variable→Output Variable** τοποθετούμε τη μεταβλητή που θέλουμε να επανακωδικοποιήσουμε (στην συγκεκριμένη άσκηση την μεταβλητή ΟΙΚ. ΚΑΤΑΣΤΑΣΗ).

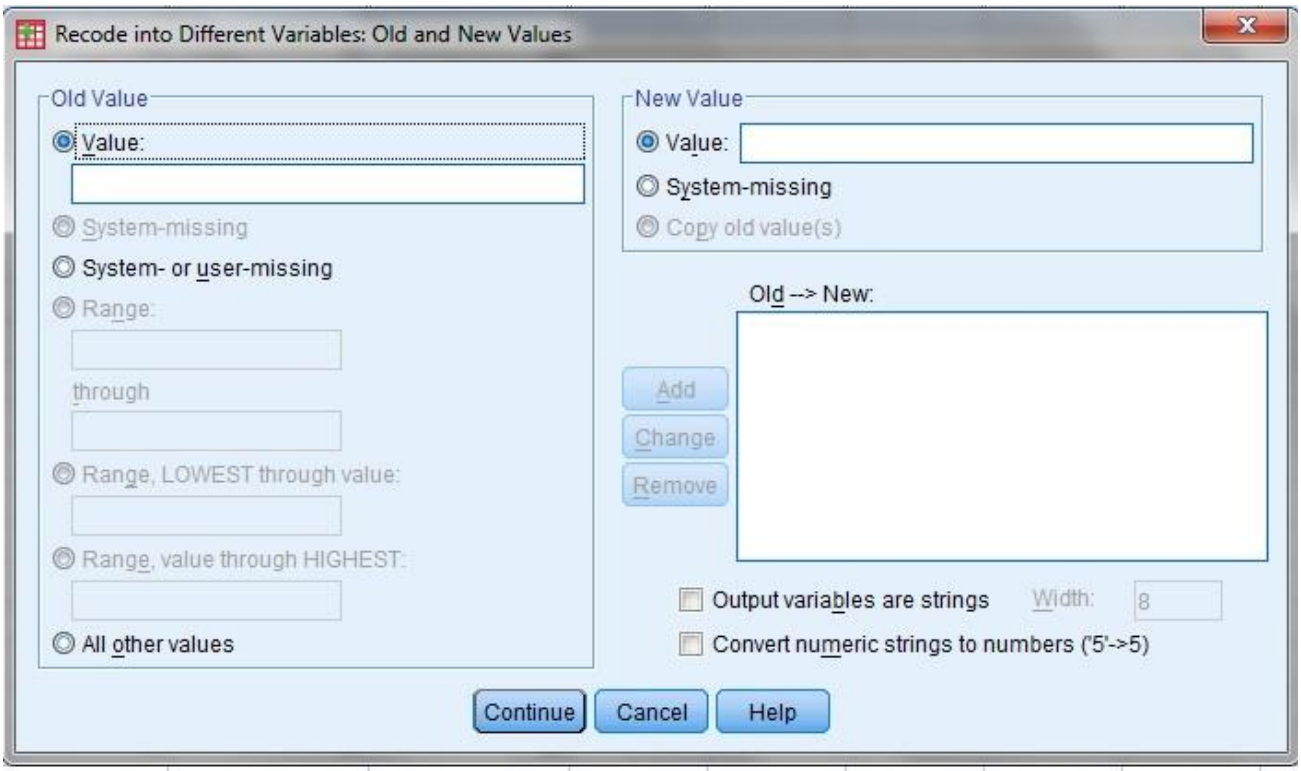
Στο **Output Variable**, στο πλαίσιο **Name**, δηλώνουμε το όνομα της νέας μεταβλητής που θέλουμε να δημιουργήσουμε π.χ. ΝΕΑ ΟΙΚ. ΚΑΤΑΣΤΑΣΗ, και αν θέλουμε να δηλώσουμε μία πλήρη περιγραφή της νέας μεταβλητής, αυτό το κάνουμε στο πλαίσιο **Output Variable Label**, π.χ. ΟΙΚΟΝΟΜΙΚΗ ΚΑΤΑΣΤΑΣΗ ΜΕΤΑ ΑΠΟ ΚΩΔΙΚΟΠΟΙΗΣΗ.



Η αλλαγή επιτυγχάνεται πατώντας το πλαίσιο Change.

Η επιθυμητή επανακωδικοποίηση πρέπει να δηλωθεί στο παράθυρο που προκύπτει πατώντας το Old and New Values...

Γενικά, από την επιλογή Old and New Values... οδηγούμαστε σε ένα παράθυρο διαλόγου που μας επιτρέπει την επανακωδικοποίηση μίας ποιοτικής μεταβλητής αλλά και την κωδικοποίηση μίας ποσοτικής σε ποιοτική.



Θα πρέπει να έχουμε όμως υπόψη δύο πολύ βασικούς κανόνες:

- Δεν θα πρέπει να υπάρχουν τομές στις κατηγορίες π.χ. τα διαστήματα [25,50], [50,75] δεν μπορούν να αποτελούν δύο νέες κατηγορίες μίας υπάρχουσας ποσοτικής μεταβλητής.
- Θα πρέπει να υπάρχουν κατάλληλες κατηγορίες για κάθε τιμή των αρχικών δεδομένων π.χ. οι κατηγορίες 25-49, 50-75 δε πρέπει να χρησιμοποιηθούν αν στα αρχικά δεδομένα υπάρχουν τιμές μεταξύ 49 και 50.

Στο αριστερό μέρος της εντολής (πλαίσιο **Old Value**) παρατηρούμε ότι υπάρχουν οι ακόλουθες επιλογές:

a) **Value**: μία παλιά τιμή και αντιστοιχίζεται σε μία νέα τιμή (η οποία δηλώνεται στο πλαίσιο Value του New Value).

b) **System missing:** δηλώνουμε πως θα επανακωδικοποιηθούν οι ελλιπείς τιμές-κενά κελιά.

c) **System-or user missing:** δηλώνουμε πως θα επανακωδικοποιηθούν οι ελλιπείς τιμές τόσο του συστήματος όσο και αυτές που έτσι κατοχυρώθηκαν από το χρήστη.

d) **Range:** δηλώνεται πως θα επανακωδικοποιηθεί ένα διάστημα τιμών, το κάτω και άνω άκρο του οποίου, δίνεται στα πλαίσια που ακολουθούν.

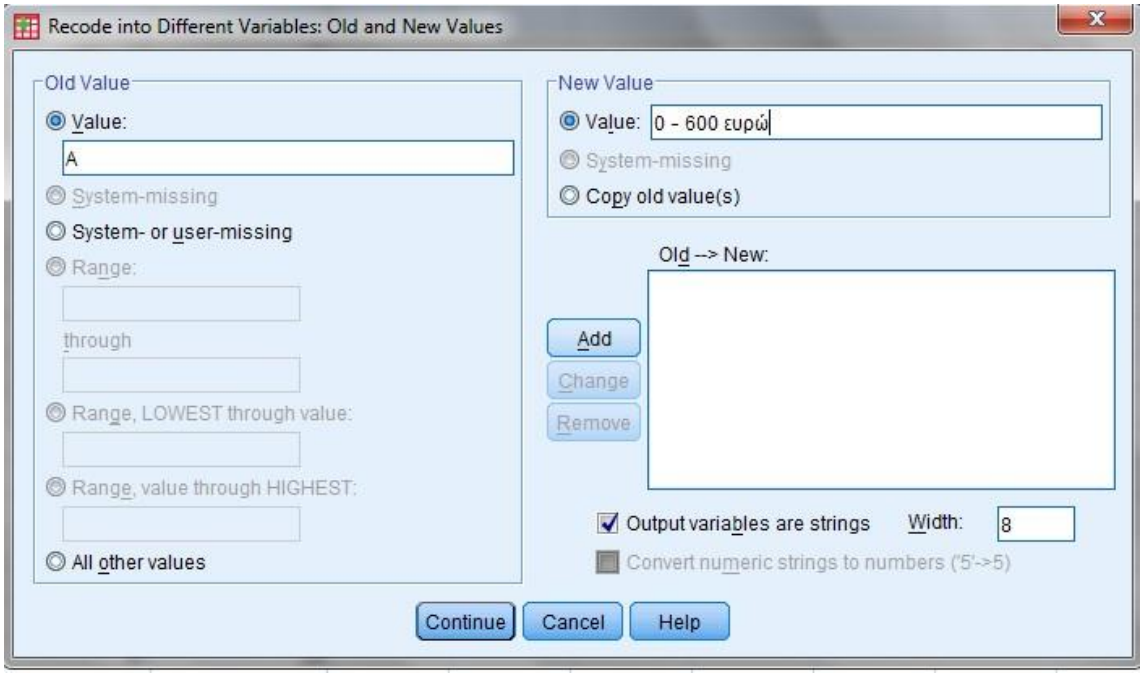
e) **Range, LOWEST through value:** δηλώνεται πως θα επανακωδικοποιηθούν οι τιμές από την μικρότερη ως αυτή που δηλώνεται στο πλαίσιο που ακολουθεί.

f) **Range, value through HIGHEST:** δηλώνεται πως θα επανακωδικοποιηθούν οι τιμές από αυτή που δηλώνεται στο πλαίσιο που ακολουθεί ως τη μεγαλύτερη.

g) **All other values:** δηλώνεται πως θα επανακωδικοποιηθούν όλες οι υπόλοιπες τιμές.

Έτσι για το παράδειγμά μας, θα πρέπει να δηλώσουμε ότι οι παλιές τιμές Α και Β αντιστοιχούν σε μία νέα κατηγορία με τιμή έστω 0 – 600 ευρώ, ενώ οι παλιές Γ και Δ σε μία νέα, με τιμή έστω >=600.

Τα παραπάνω θα πρέπει να δηλωθούν στο παρακάτω πλαίσιο:



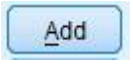
Αναλυτικά:

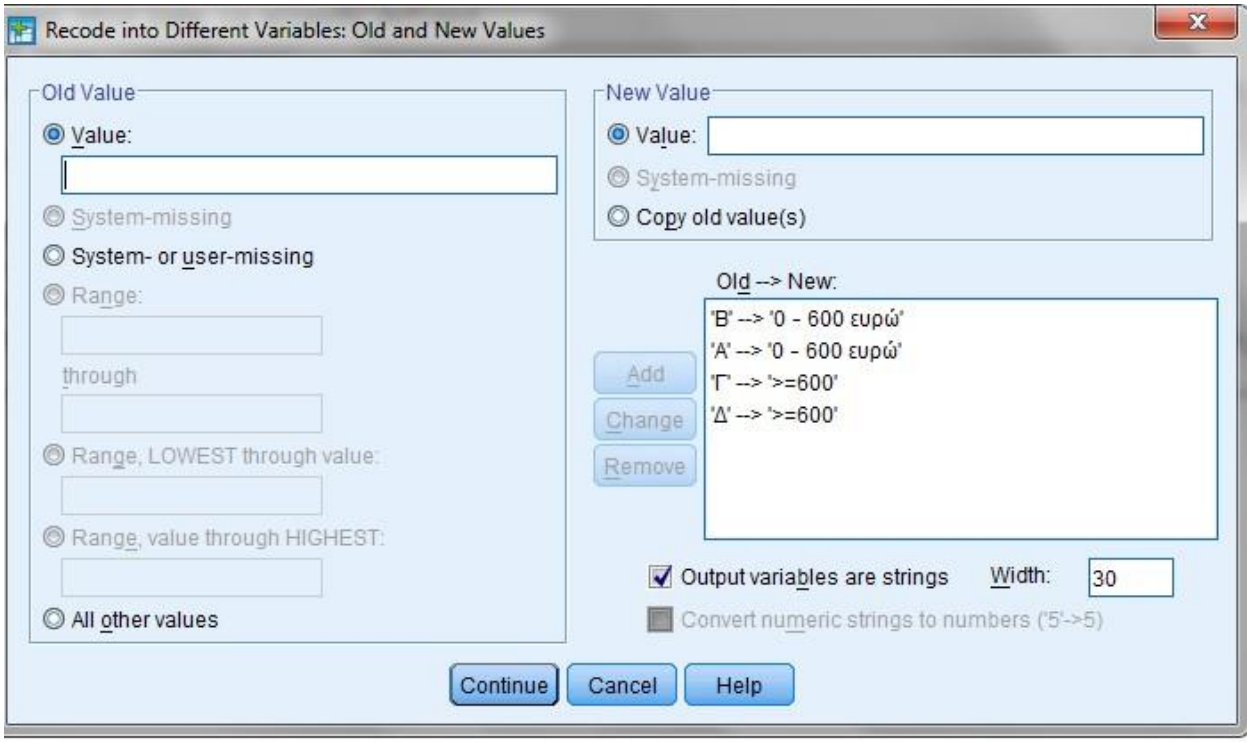
Αρχικά θα πρέπει να επιλεγεί  γιατί η νέα μεταβλητή που θα δημιουργηθεί μέσω της εντολής είναι ποιοτική.

Old Value: A, New Value: 0 – 600 ευρώ, 

Old Value: B, New Value: 0 – 600 ευρώ, 

Old Value: Γ, New Value: >=600, 

Old Value: Δ, New Value: >=600, 



Continue και OK

Μετά την εκτέλεση της εντολής στο φύλλο Data View δημιουργείται η μεταβλητή ΝΕΑ_ΟΙΚ_ΚΑΤΑΣΤΑΣΗ.

Γ	ΟΙΚ.ΚΑΤΑΣΤΑΣΗ	IQ	ΥΨΟΣ	ΧΡΟΝΟΣ	ΝΕΑ_ΟΙΚ_ΚΑΤΑΣΤΑΣΗ
	B	11	95	22	0 - 600 ευρώ
	Γ	90	98	25	>=600
	Γ	90	92	18	>=600
	Γ	90	104	19	>=600
	A	104	85	21	0 - 600 ευρώ
	B	72	96	20	0 - 600 ευρώ
	B	105	89	21	0 - 600 ευρώ
	Δ	93	103	22	>=600
	Γ	99	110	18	>=600
	B	93	85	27	0 - 600 ευρώ
	B	84	94	30	0 - 600 ευρώ
	A	95	98	21	0 - 600 ευρώ

ΠΑΡΑΓΡΑΦΟΣ 2.2 : ΠΙΝΑΚΑΣ ΚΑΤΑΝΟΜΗΣ ΣΥΧΝΟΤΗΤΩΝ ΓΙΑ ΠΟΣΟΤΙΚΗ ΔΙΑΚΡΙΤΗ ΜΕΤΑΒΛΗΤΗ. (Η ΔΙΑΔΙΚΑΣΙΑ ΕΙΝΑΙ Η ΙΔΙΑ ΚΑΙ ΓΙΑ ΠΟΙΟΤΙΚΗ ΜΕΤΑΒΛΗΤΗ)

Ποσοτική Διακριτή ονομάζεται η μεταβλητή που παίρνει τιμές μόνο ακέραιους αριθμούς.

ΑΣΚΗΣΗ

Ο παρακάτω πίνακας δείχνει τον αριθμό των αυτοκινήτων που απάντησαν πως διαθέτουν 50 οικογένειες:

0	1	0	2	3	2	0	0	1	0
1	1	0	1	1	1	0	2	0	1
2	1	1	1	1	0	1	0	0	1
1	1	0	1	0	3	2	0	0	0
1	1	0	2	3	1	1	2	0	1

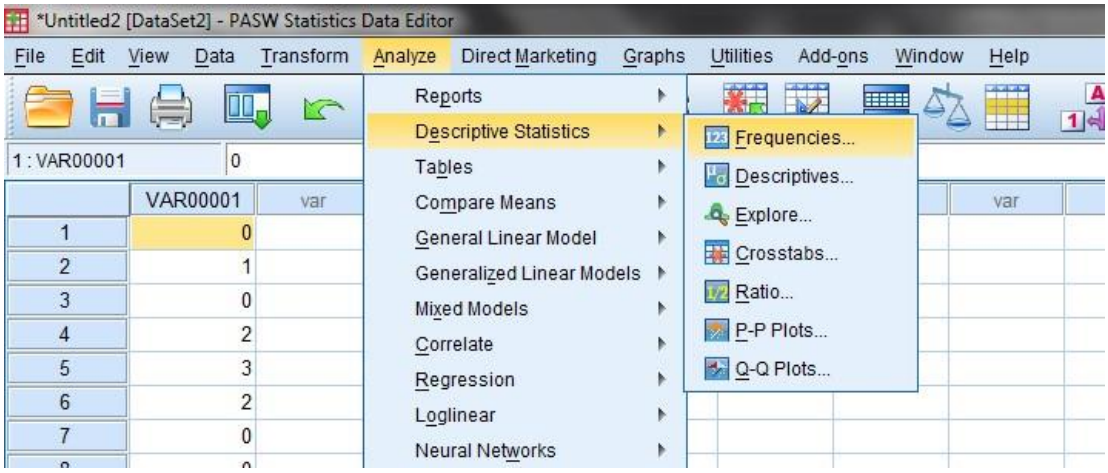
Για τα δεδομένα του παραπάνω πίνακα, να κατασκευαστούν:

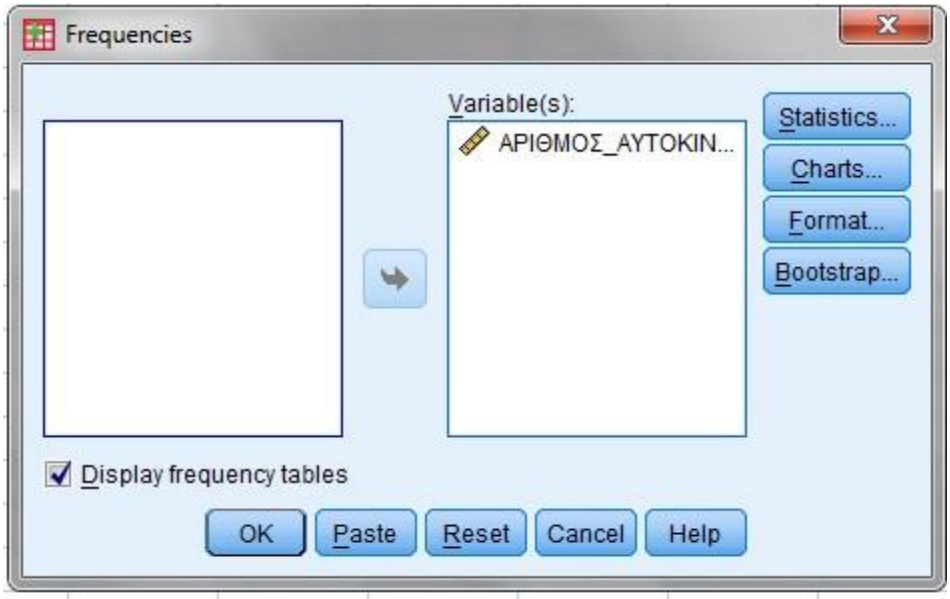
1. Ο πίνακας συχνοτήτων (απλών, σχετικών και αθροιστικών)
2. Τα κατάλληλα διαγράμματα

ΛΥΣΗ

Αρχικά δημιουργούμε την μεταβλητή **ΑΡΙΘΜΟΣ_ΑΥΤΟΚΙΝΗΤΩΝ** (Type: Numeric, Decimals: 0) και εισάγουμε τις τιμές του παραπάνω πίνακα. **Προσοχή:** Όλες οι τιμές θα πρέπει να περαστούν σε μία στήλη καθώς είναι οι διαφορετικές τιμές μίας μεταβλητής. Να θυμίσουμε ότι: κάθε στήλη στο SPSS αποτελεί και μία μεταβλητή και κάθε γραμμή αποτελεί την απάντηση του δείγματος στη μεταβλητή.

Η συνοπτική παρουσίαση των δεδομένων διακριτών ποσοτικών μεταβλητών γίνεται με την ακόλουθη διαδικασία: **Analyze→Descriptive Statistics→Frequencies.**





Στο νέο παράθυρο διαλόγου που προκύπτει επιλέγουμε την ή τις προς ανάλυση μεταβλητή/ μεταβλητές, και την μεταφέρουμε στο κουτί Variable(s). Έχοντας επιλέξει το πλαίσιο **Display frequency tables** θα παραχθούν σε νέο παράθυρο Output του SPSS οι ακόλουθοι πίνακες συχνοτήτων:

Statistics

ΑΡΙΘΜΟΣ_ΑΥΤΟΚΙΝΗΤΩΝ

N	50

	Valid	0
	Missing	

Ο πρώτος πίνακας Statistics μας δίνει των αριθμό των έγκυρων τιμών της μεταβλητής και τον αριθμό των ελλιπών τιμών της μεταβλητής. Από τον πρώτο πίνακα πληροφορούμαστε ότι οι διαθέσιμες δειγματικές τιμές είναι 50 (Valid=50) και δεν υπάρχουν ελλιπείς τιμές (Missing=0).

ΑΡΙΘΜΟΣ_ΑΥΤΟΚΙΝΗΤΩΝ

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	18	36,0	36,0	36,0
	1	22	44,0	44,0	80,0
	2	7	14,0	14,0	94,0
	3	3	6,0	6,0	100,0
	Total	50	100,0	100,0	

Στην κορυφή του δεύτερου πίνακα βλέπουμε το όνομα της μεταβλητής.

Στην 1η στήλη εμφανίζονται οι τιμές της μεταβλητής οι οποίες τοποθετούνται κατά φυσική αύξουσα τάξη μεγέθους, δηλ. από τη μικρότερη προς τη μεγαλύτερη.

Στην 2η στήλη (Frequency) παρουσιάζονται οι συχνότητες. Εκφράζει πόσες φορές εμφανίζεται στον συνολικό πληθυσμό/ ή δείγμα κάθε τιμή της μεταβλητής.

Στην 3η στήλη (Percent) παρουσιάζονται οι σχετικές συχνότητες. Δηλαδή, υπολογίζει τα ποσοστά διαιρώντας την τιμή Frequency με την τιμή Total Cases.

Στην 4η στήλη (Valid Percent) παρουσιάζονται οι έγκυρες σχετικές συχνότητες. Δηλαδή, υπολογίζει τα ποσοστά διαιρώντας την τιμή Frequency με την τιμή Valid Cases. Πολλές φορές στους πίνακες κατανομής συχνοτήτων εμφανίζεται η ένδειξη Missing Values. Με τον όρο αυτό αναφέρονται τα κελιά που δεν περιέχουν τιμές. Δηλ. αν σε κάποια ερώτηση δεν απαντήσουν κάποια άτομα, τα αντίστοιχα κελιά θα μείνουν κενά. Τα κελιά αυτά αποτελούν Missing Values. Οι τιμές Percent και Valid Percent είναι ίσες μόνο όταν δεν υπάρχουν Missing Values.

Στην 5η στήλη (Cumulative Percent) παρουσιάζονται οι σχετικές αθροιστικές συχνότητες. Προκύπτει ξεκινώντας με την πρώτη τιμή της στήλης Valid Percent και προσθέτοντας κάθε φορά την επόμενη.

Έτσι, για το συγκεκριμένο παράδειγμα, από τη στήλη Frequency (Στήλη Συχνοτήτων) προκύπτει ότι από τις 50 συνολικά οικογένειες 0, 1, 2 και 3 αυτοκίνητα διαθέτουν 18, 22, 7, 3 οικογένειες αντίστοιχα. Επιπλέον, από τη στήλη Percent (Στήλη Σχετικών Συχνοτήτων) έχουμε π.χ. ότι 1 αυτοκίνητο διαθέτουν 22 από τις 50 οικογένειες, δηλαδή ποσοστό 44,0%. Επισημαίνεται ότι το ποσοστό στη στήλη Percent υπολογίζεται στο σύνολο των ερωτηθέντων συμπεριλαμβανομένου και των πιθανών ελλιπών τιμών. Από την άλλη μεριά το ποσοστό στη στήλη Valid Percent υπολογίζεται στο σύνολο αυτών που έχουν απαντήσει. Εδώ προφανώς προκύπτει ισότητα καθώς δεν έχουμε ελλιπείς παρατηρήσεις. Τέλος, από τη στήλη Cumulative Percent (Στήλη Αθροιστικών Σχετικών Συχνοτήτων) προκύπτει για παράδειγμα ότι το 80,0% των οικογενειών έχουν το πολύ μέχρι 1 αυτοκίνητο. Η στήλη αυτή όπως γνωρίζουμε από τη θεωρία έχει νόημα μόνο για ποσοτικές μεταβλητές όπως αυτή του παραδείγματος, ή για διατάξιμες ποιοτικές μεταβλητές.

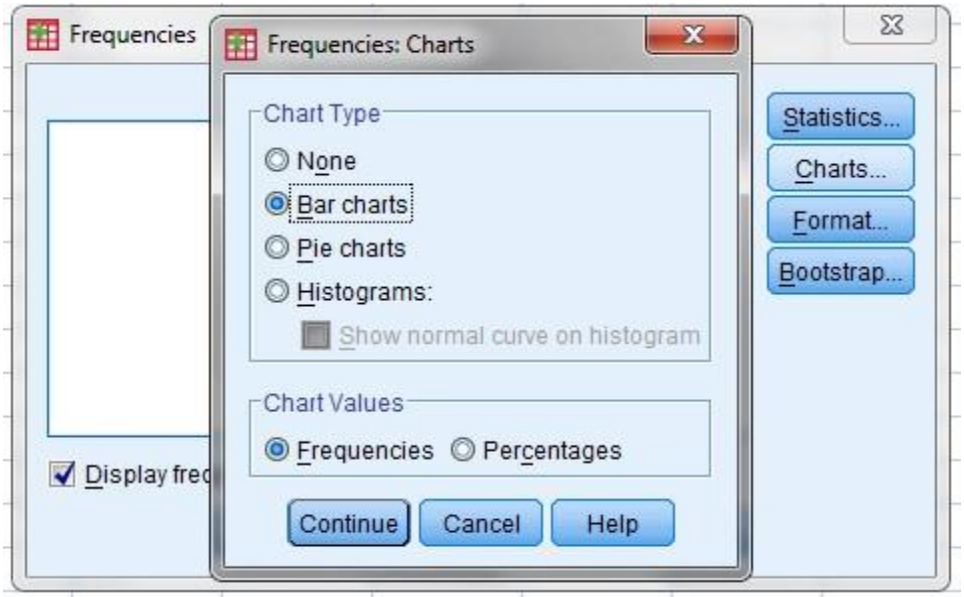
Από την επιλογή Charts μπορούμε να κατασκευάσουμε: Ραβδογράμματα (Bar charts), Κυκλικά Διαγράμματα (Pie charts). Τα ιστογράμματα (Histograms), όπως θα δούμε και στην επόμενη ενότητα, αφορούν την περίπτωση ποσοτικών μεταβλητών.

Δημιουργία γραφημάτων: Ραβδόγραμμα και κυκλικό διάγραμμα.

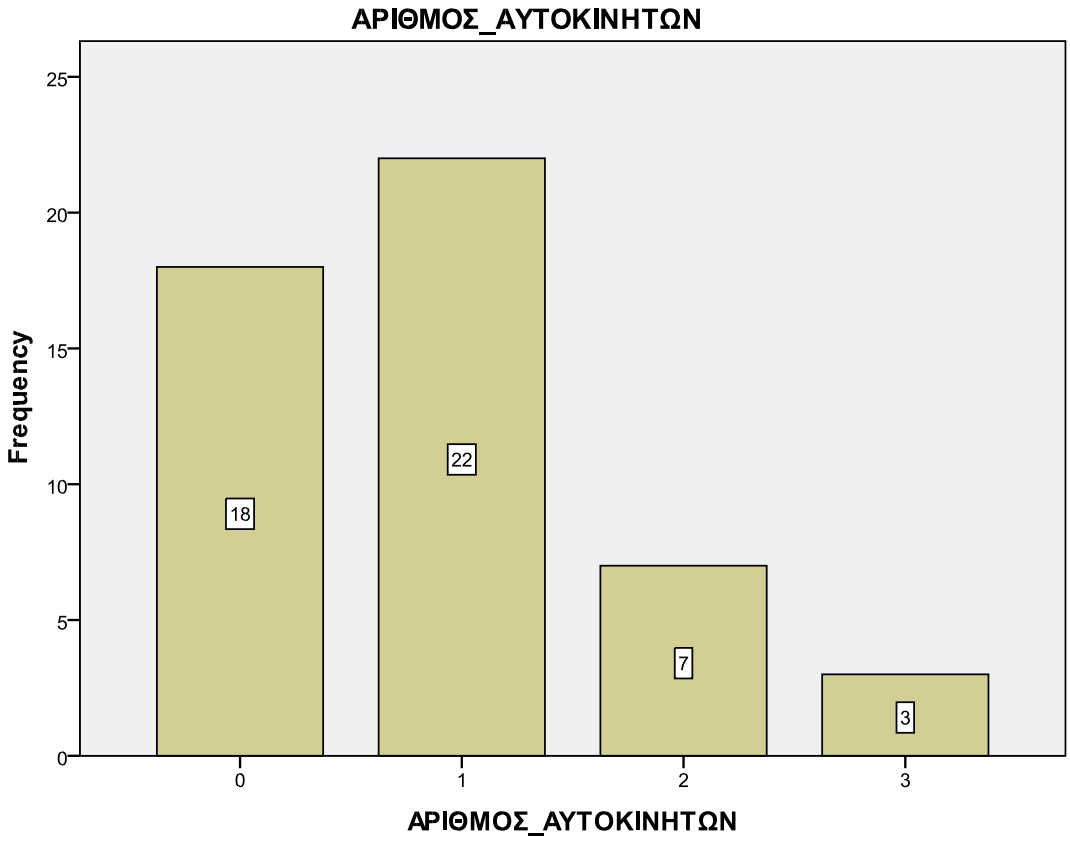
Οι γραφικές παραστάσεις που μπορούμε να χρησιμοποιήσουμε για να περιγράψουμε μια διακριτή μεταβλητή (με μικρό αριθμό τιμών) ή μία ποιοτική μεταβλητή είναι το ραβδόγραμμα και το κυκλικό διάγραμμα.

Και τα δύο αυτά διαγράμματα μπορούν να δημιουργηθούν από την εντολή που παρουσιάστηκε παραπάνω για την κατασκευή πίνακα συχνοτήτων πατώντας την επιλογή Charts... .

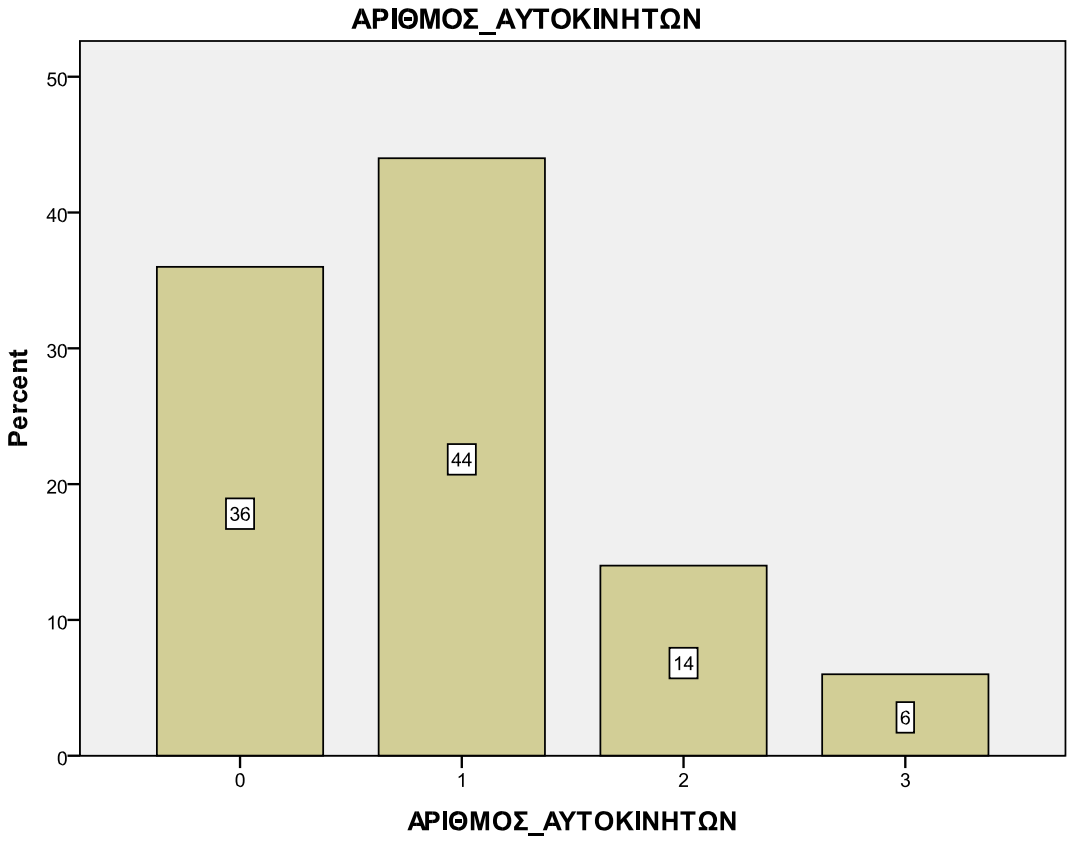
Δυστυχώς κάθε φορά έχουμε τη δυνατότητα μίας επιλογής μεταξύ του Bar Charts και Pie Charts. Επιλέγοντας π.χ. την κατασκευή ραβδογράμματος (ή κυκλικού διαγράμματος), ενεργοποιείται η επιλογή Chart Values από όπου επιλέγοντας Frequencies ή Percentages καθορίζουμε αν στον κατακόρυφο άξονα των υπό κατασκευή ραβδογραμμάτων ή κυκλικών διαγραμμάτων θα εμφανίζονται οι απόλυτες συχνότητες (Frequencies) ή οι σχετικές συχνότητες (Percentages), αντίστοιχα.



Το γράφημα που προκύπτει από τις επιλογές: Bar Chart και Frequencies είναι το ακόλουθο:



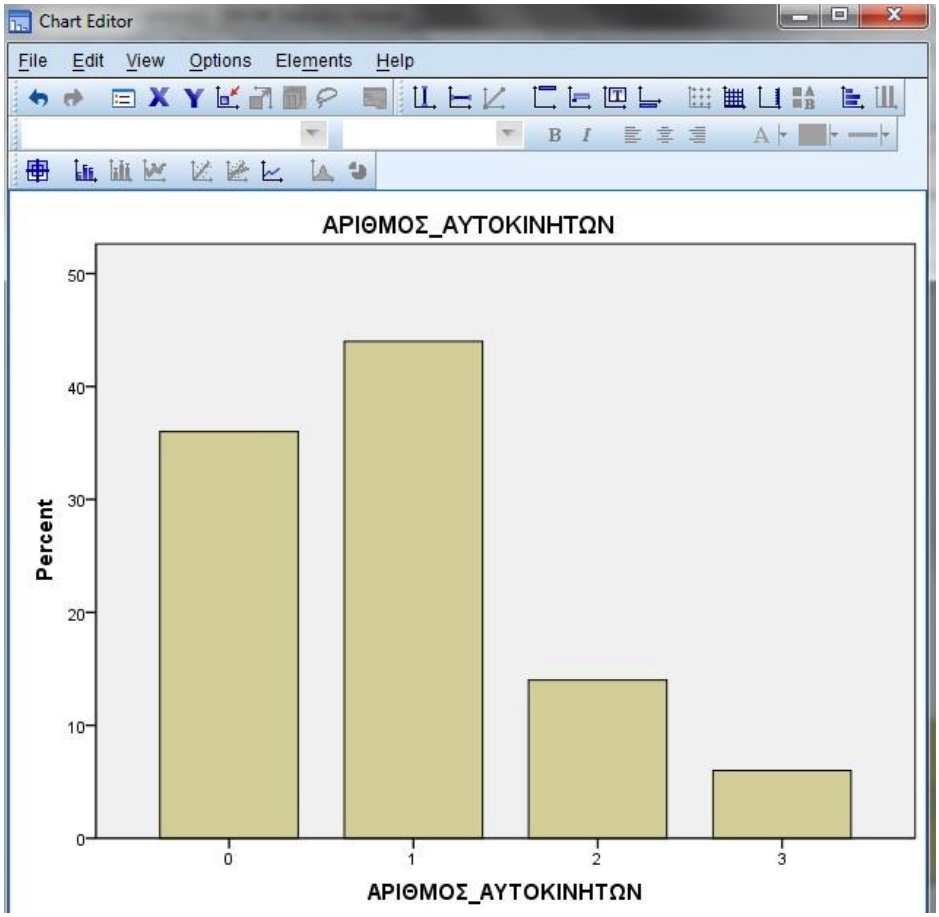
Το γράφημα που προκύπτει από τις επιλογές: Bar Chart και Percentages είναι το ακόλουθο:



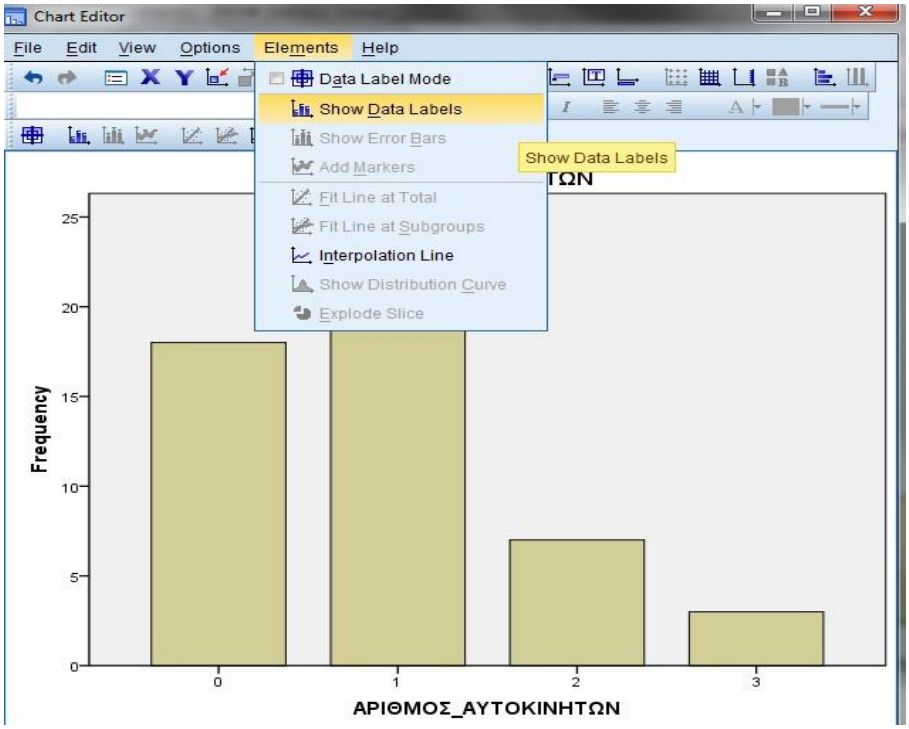
Παρατήρηση:

Αρχικά το γράφημα που δημιουργείται στο παράθυρο Output του Spss δεν εμφανίζονται τα labels στις μπάρες. Θα πρέπει να τα προσθέσουμε ως εξής:

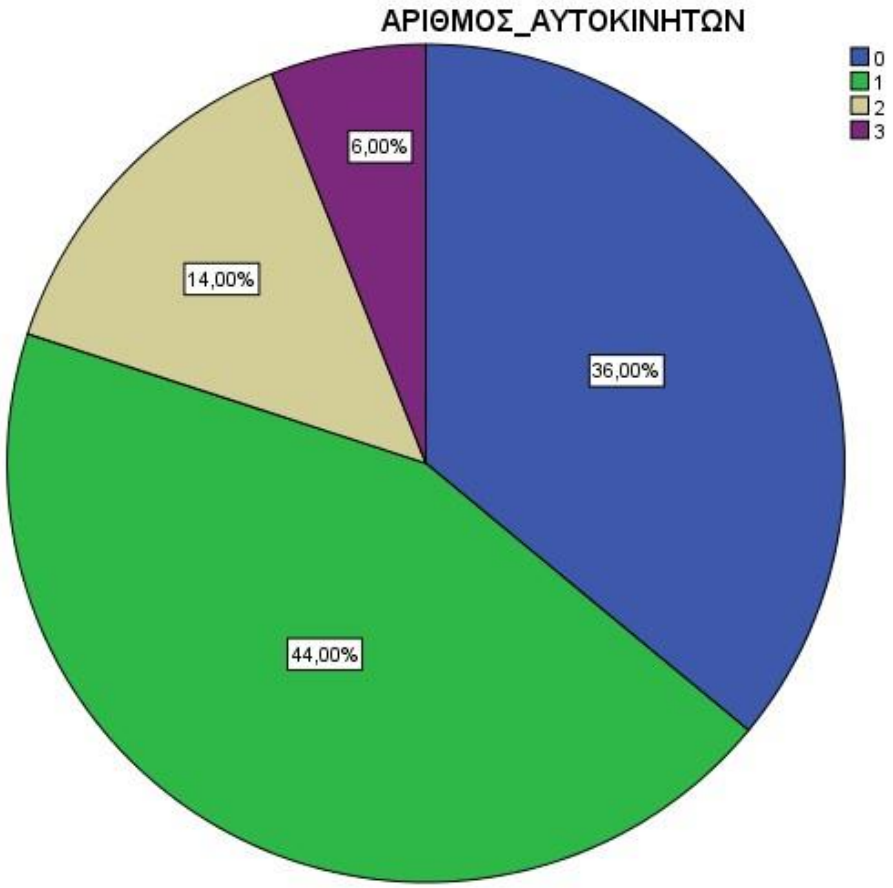
Πατάμε με το ποντίκι διπλό κλικ στο γράφημα για να γίνει ενεργό στο παράθυρο του chart editor που ανοίγει.



Στην συνέχεια από το μενού επιλέγουμε: Elements – show data labels



Το γράφημα που προκύπτει από τις επιλογές: Pie Charts και Percentages είναι το ακόλουθο:



Παρατήρηση:

Με τον ίδιο τρόπο που περιγράψαμε παραπάνω προσθέτουμε labels στους κυκλικούς τομείς του γραφήματος. Συνήθως στα γραφήματα Pie Charts επιλέγουμε να τοποθετούμε ως labels τα ποσοστά.

ΠΑΡΑΓΡΑΦΟΣ 2.3 : ΟΜΑΔΟΠΟΙΗΣΗ ΤΩΝ ΤΙΜΩΝ ΜΙΑΣ ΜΕΤΑΒΛΗΤΗΣ- ΠΙΝΑΚΑΣ ΣΥΧΝΟΤΗΤΩΝ ΓΙΑ ΠΟΣΟΤΙΚΗ ΣΥΝΕΧΗ ΜΕΤΑΒΛΗΤΗ

Ομαδοποίηση είναι μια διαδικασία με την οποία ομαδοποιούμε (χωρίζουμε σε ομάδες) τα δεδομένα μας. Τη διαδικασία αυτή την χρησιμοποιούμε για να έχουμε καλύτερη παρουσίαση των αποτελεσμάτων μιας στατιστικής ανάλυσης και των διαγραμμάτων, στη περίπτωση που έχουμε μεγάλο μέγεθος δείγματος (π.χ. 100 παρατηρήσεις) ή στη περίπτωση που έχουμε μεγάλη διαφοροποίηση (ποικιλία) στα δεδομένα μας (αυτό συμβαίνει όταν έχουμε ποσοτικές συνεχείς μεταβλητές).

Ποσοτική Συνεχής είναι η μεταβλητή που μπορεί να πάρει όλες τις τιμές που περιέχονται σε ένα διάστημα [α,β].

ΑΣΚΗΣΗ

Δίνονται τα αναστήματα σε cm μιας ομάδας 50 ατόμων.

155	156	159	160	162	162	164	166	166	167
167	167	167	167	168	168	169	171	171	171
171	172	172	172	172	172	172	172	173	174
176	176	177	177	177	178	178	178	179	179
181	181	181	182	182	183	184	185	186	189

Να κατασκευαστεί ο πίνακας κατανομής συχνοτήτων, το ιστόγραμμα και το αθροιστικό διάγραμμα.

ΛΥΣΗ

Αρχικά δημιουργούμε την μεταβλητή **ΎΨΟΣ** και εισάγουμε τα δεδομένα του παραπάνω πίνακα.

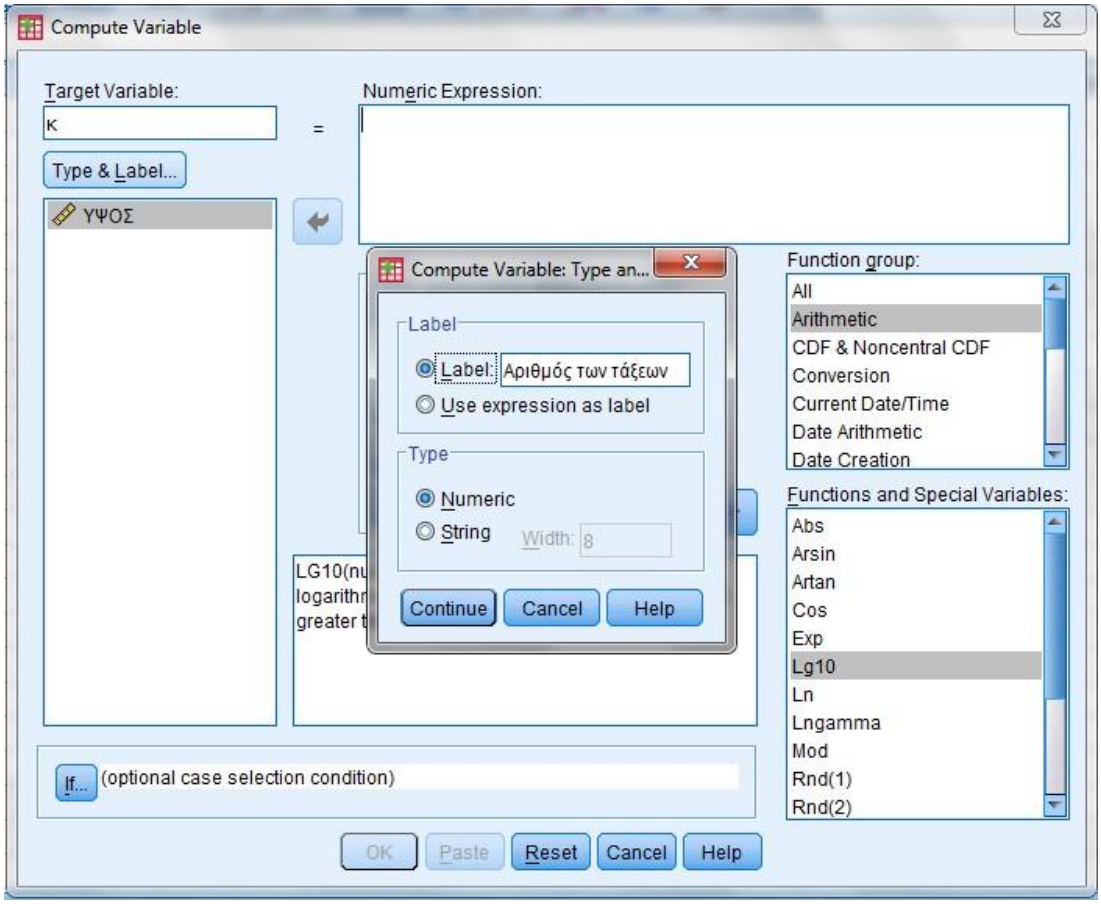
1) **Προσδιορίζουμε τον αριθμό των τάξεων (κ)** με τον κανόνα του Sturges: κ = 1 + 3,322logn όπου κ = αριθμός τάξεων και n = ο αριθμός

παρατηρήσεων

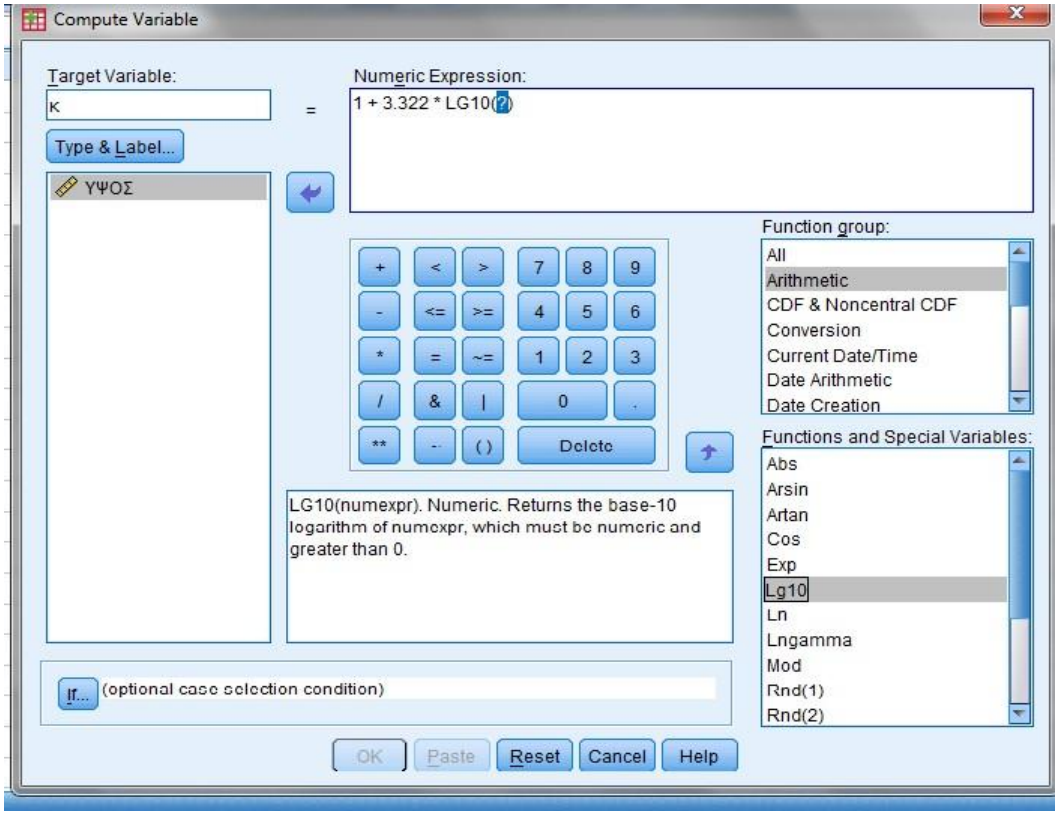
Ο υπολογισμός του κ μπορεί να γίνει μέσω του SPSS με την εξής διαδικασία:

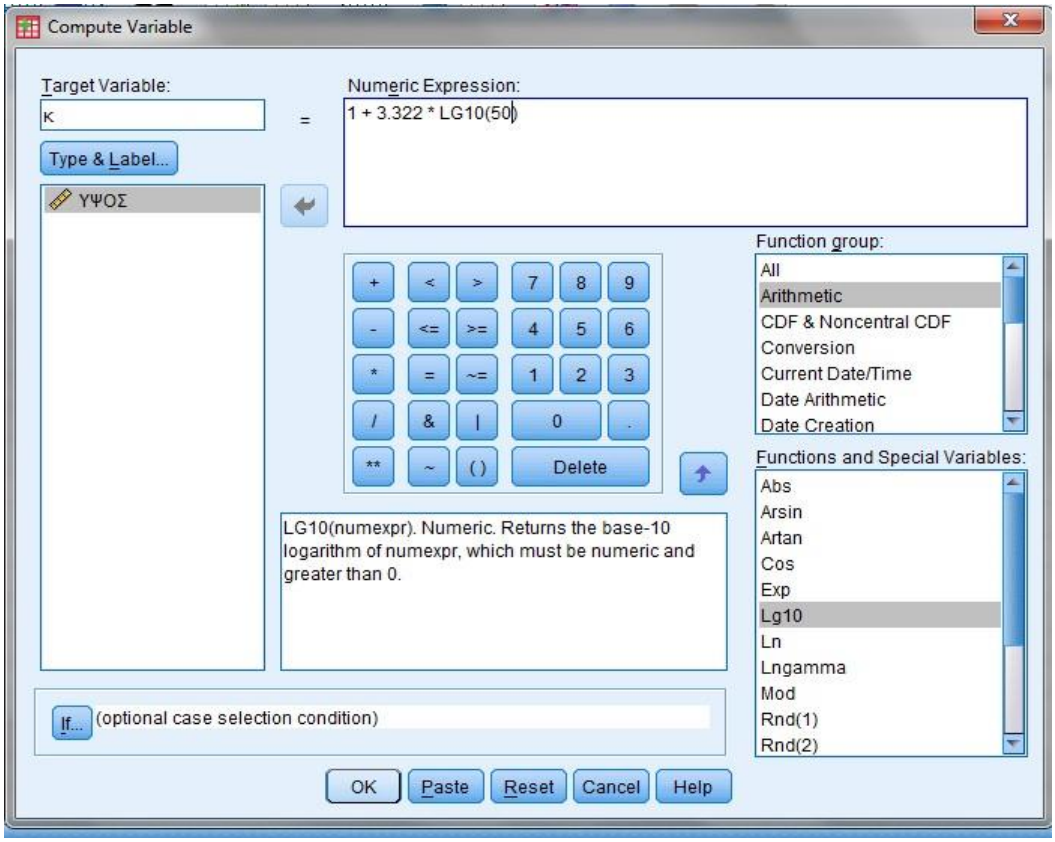
- Πατάμε στο μενού ***TRANSFORM - COMPUTE***

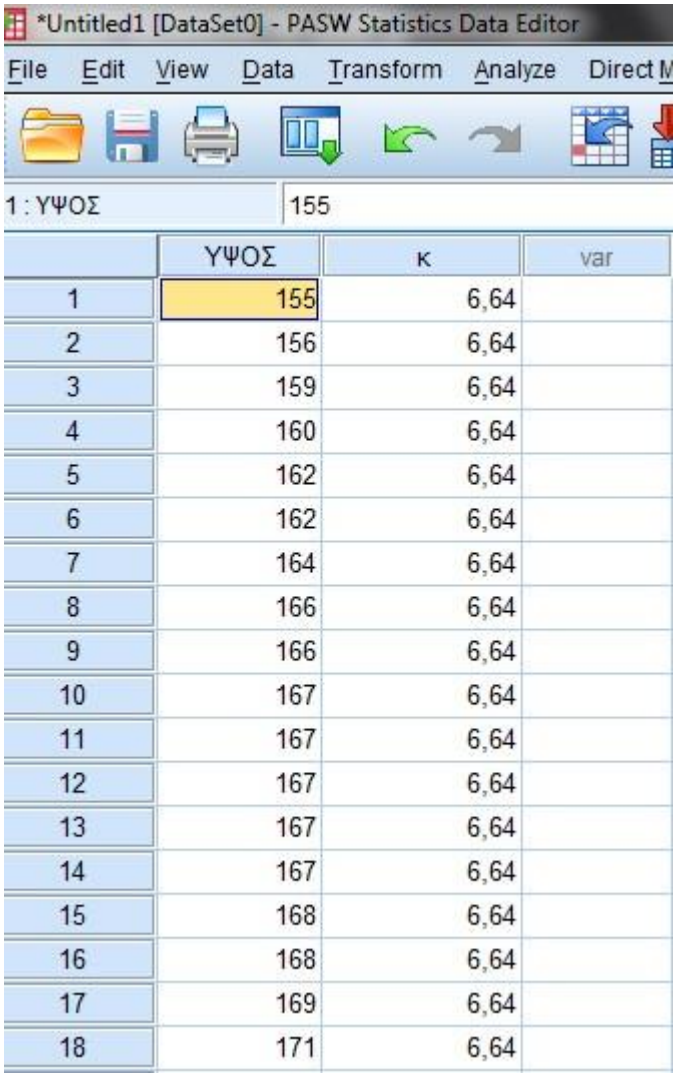
Αριστερά στο παράθυρο στην ένδειξη ***TARGET VARIABLE*** δίνουμε ως όνομα μεταβλητής το ***κ*** και επιλέγοντας την ένδειξη ***TYPE & LABEL*** ορίζουμε ότι το ***κ*** είναι ο αριθμός των τάξεων και είναι ***TYPE : NUMERIC*** (δηλ. παίρνει αριθμητικές τιμές).



Δεξιά στο παράθυρο στην ένδειξη ***NUMERIC EXPRESSION*** πληκτρολογούμε τον κανόνα του Sturges: 1 + 3,322logn από το calculator για να μην γίνουν λάθη. Για να εισάγουμε τον λογάριθμο (δηλαδή το **LG10(?)** που βλέπουμε στον τύπο) το επιλέγουμε από το παράθυρο **Function group: Arithmetic**, και στο παράθυρο **Functions and Special Variables: Lg10.** Το ερωτηματικό το αντικαθιστούμε με τον αριθμό των παρατηρήσεων δηλ. το 50. Έπειτα ***O.K..***


Compute Variable
Target Variable:
κ
=
Numeric Expression:
1 + 3.322 * LG10(?)
Type & Label...
ΥΨΟΣ
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Date Creation
+ < > 7 8 9
- <= >= 4 5 6
* = ~= 1 2 3
/ & | 0 .
** ( ) Delete
Functions and Special Variables:
Abs
Arsin
Artan
Cos
Exp
Lg10
Ln
Lngamma
Mod
Rnd(1)
Rnd(2)
LG10(numexpr). Numeric. Returns the base-10 logarithm of numexpr, which must be numeric and greater than 0.
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help


Compute Variable
Target Variable:
κ
Type & Label...
ΥΨΟΣ
=
Numeric Expression:
1 + 3.322 * LG10(50)
+ < > 7 8 9
- <= >= 4 5 6
* = ~= 1 2 3
/ & | 0 .
** ~ ( ) Delete
LG10(numexpr). Numeric. Returns the base-10 logarithm of numexpr, which must be numeric and greater than 0.
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Date Creation
Functions and Special Variables:
Abs
Arsin
Artan
Cos
Exp
Lg10
Ln
Lngamma
Mod
Rnd(1)
Rnd(2)
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help



	ΥΨΟΣ	κ	var
1	155	6,64	
2	156	6,64	
3	159	6,64	
4	160	6,64	
5	162	6,64	
6	162	6,64	
7	164	6,64	
8	166	6,64	
9	166	6,64	
10	167	6,64	
11	167	6,64	
12	167	6,64	
13	167	6,64	
14	167	6,64	
15	168	6,64	
16	168	6,64	
17	169	6,64	
18	171	6,64	

Επιστρέφουμε στον ***DATA EDITOR*** όπου έχει δημιουργηθεί η μεταβλητή ***κ*** με τιμή 6.64. Άρα κ

≈ 7

2) Στη συνέχεια **υπολογίζουμε το πλάτος των τάξεων** από τον τύπο

δ = R / κ, όπου δ = πλάτος των

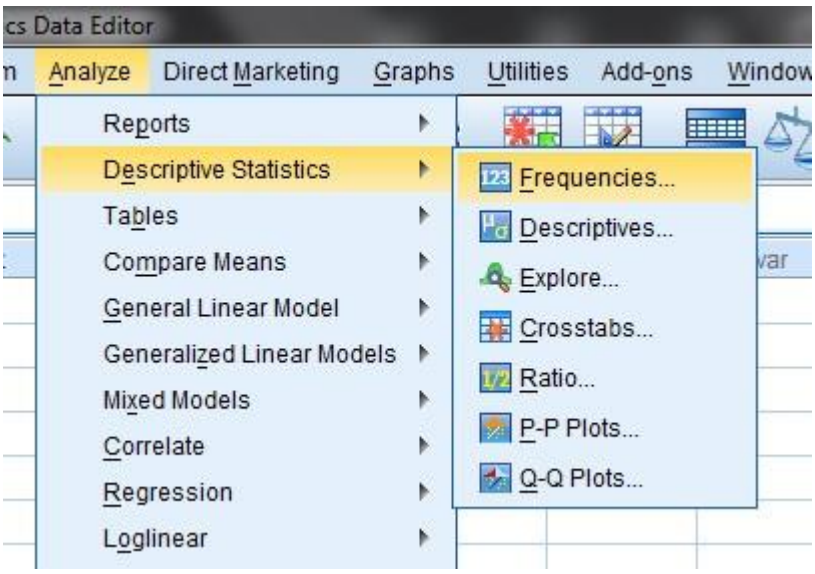
τάξεων,

R = εύρος μεταβολής (δηλαδή R = Max – Min), και κ

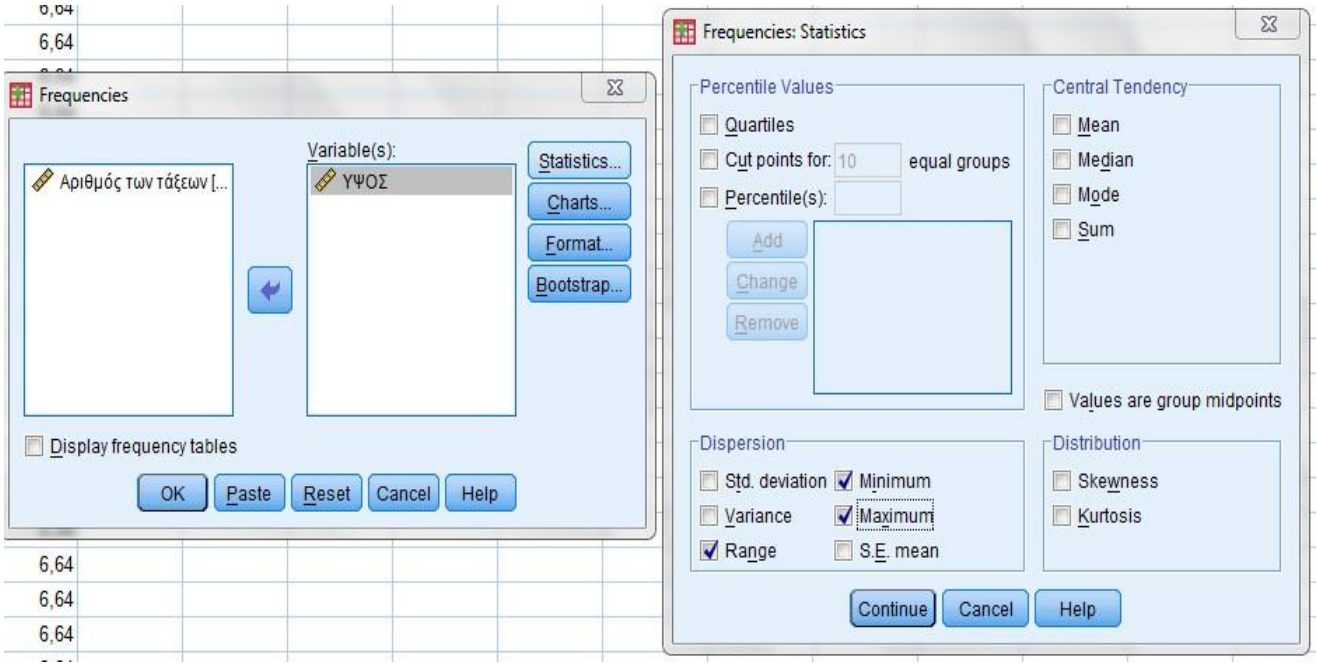
= αριθμός τάξεων

Το εύρος μεταβολής (RANGE) μπορούμε να το υπολογίσουμε από την εντολή

ANALYSE - DESCRIPTIVE STATISTICS - FREQUENCIES



Επιλέγουμε από το παράθυρο αριστερά την μεταβλητή **'ΥΨΟΣ'** και με πάτημα στο ***μαύρο βέλος*** αυτή μεταφέρεται στο παράθυρο δεξιά. Προσέχουμε να μην έχει τσεκαριστεί η ένδειξη ***DISPLAY FREQUENCIES TABLES*** για να μην δημιουργήσουμε πίνακα συχνοτήτων, επιλέγουμε την ένδειξη Statistics... και εμφανίζεται το παράθυρο ***Frequencies: Statistics*** από το οποίο επιλέγουμε τις ενδείξεις ***Range*** (εύρος μεταβολής)***, Minimum*** (η ελάχιστη τιμή των δεδομένων) και ***Maximum*** (η μέγιστη τιμή των δεδομένων). Στη συνέχεια πατάμε ***CONTINUE*** και ***O.K..***



Εμφανίζεται το ***OUTPUT*** στο οποίο περιέχεται ο παρακάτω πίνακας:

Statistics

ΥΨΟΣ

N	Valid	50
	Missing	0
Range		34
Minimum		155
Maximum		189

Οπότε δ = R / κ = 34 / 7 = 4,85 ≈ 5

3) Σύμφωνα με τα παραπάνω **θα δημιουργήσουμε 7 τάξεις με πλάτος 5**. Ξεκινάμε από την ελάχιστη τιμή το 155, και δημιουργούμε:

1η τάξη : 155 – 160

2η τάξη : 160 – 165

3η τάξη : 165 – 170

4η τάξη : 170 – 175

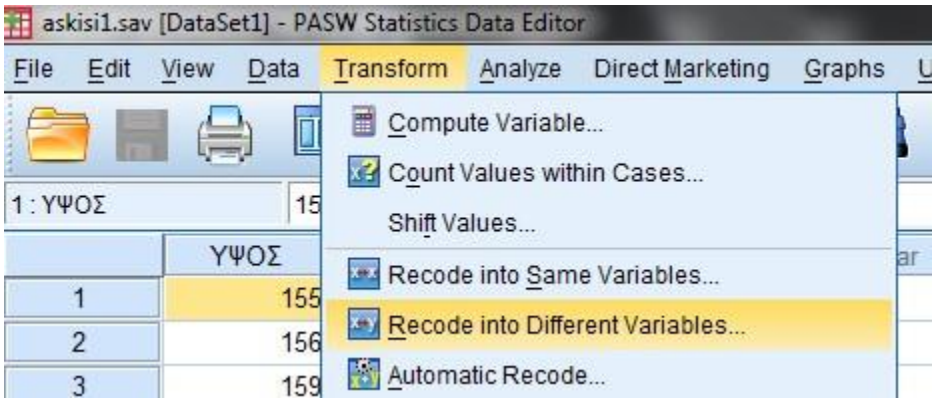
5η τάξη : 175 – 180

6η τάξη : 180 – 185

7η τάξη : 185 - 190

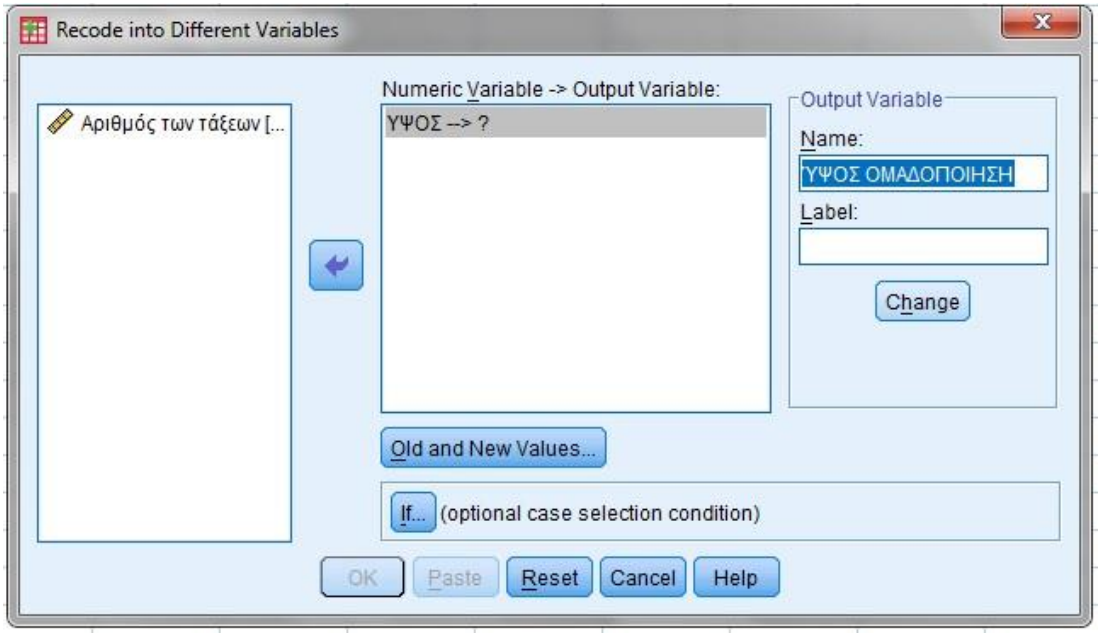
Οι τάξεις που δημιουργεί το SPSS είναι της μορφής ( ], δηλαδή κλειστές στο πάνω όριο οπότε π.χ. η τιμή 160 θα συμπεριληφθεί στην πρώτη τάξη. Προσέχουμε η ελάχιστη και η μέγιστη τιμή των δεδομένων να περιέχονται στις τάξεις που δημιουργήσαμε.

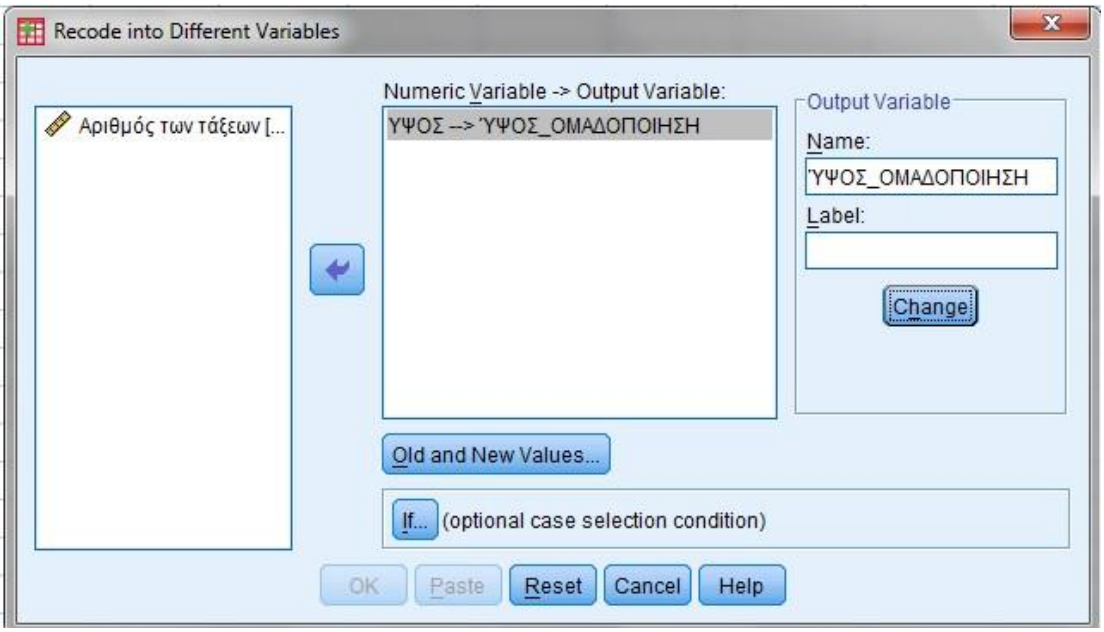
Πατάμε στο μενού **TRANSFORM - RECODE - INTO DIFFERENT VARIABLES.** Αν επιλέξουμε INTO SAME VARIABLES η ομαδοποίηση θα γίνει στην ίδια στήλη με αποτέλεσμα να χαθούν τα αρχικά δεδομένα (και σε περίπτωση που έχει γίνει κάποιο λάθος στην ομαδοποίηση δεν θα έχουμε διαθέσιμα τα αρχικά δεδομένα για να το διορθώσουμε). Επιλέγουμε INTO DIFFERENT VARIABLES επειδή θέλουμε η ομαδοποίηση να γίνει σε διαφορετική στήλη.

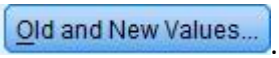


Επιλέγουμε από το παράθυρο αριστερά τη μεταβλητή **΄ΥΨΟΣ**΄ και με πάτημα στο μαύρο βέλος αυτή μεταφέρεται στο παράθυρο δεξιά. Στη θέση **NAME** γράφουμε το όνομα που θα δώσουμε στη νέα στήλη π.χ. **΄ΥΨΟΣ_ΟΜΑΔΟΠΟΙΗΣΗ΄** και πατώντας στην ένδειξη 

το όνομα που δώσαμε, πηγαίνει δίπλα στο όνομα της προηγούμενης μεταβλητής στη θέση του ερωτηματικού.

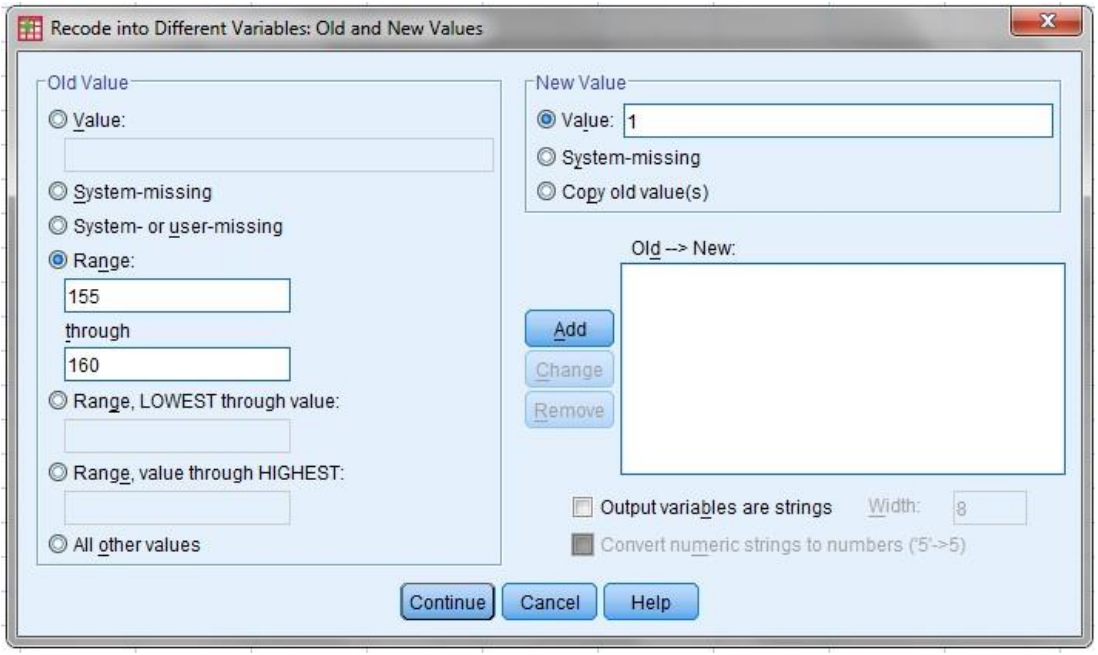




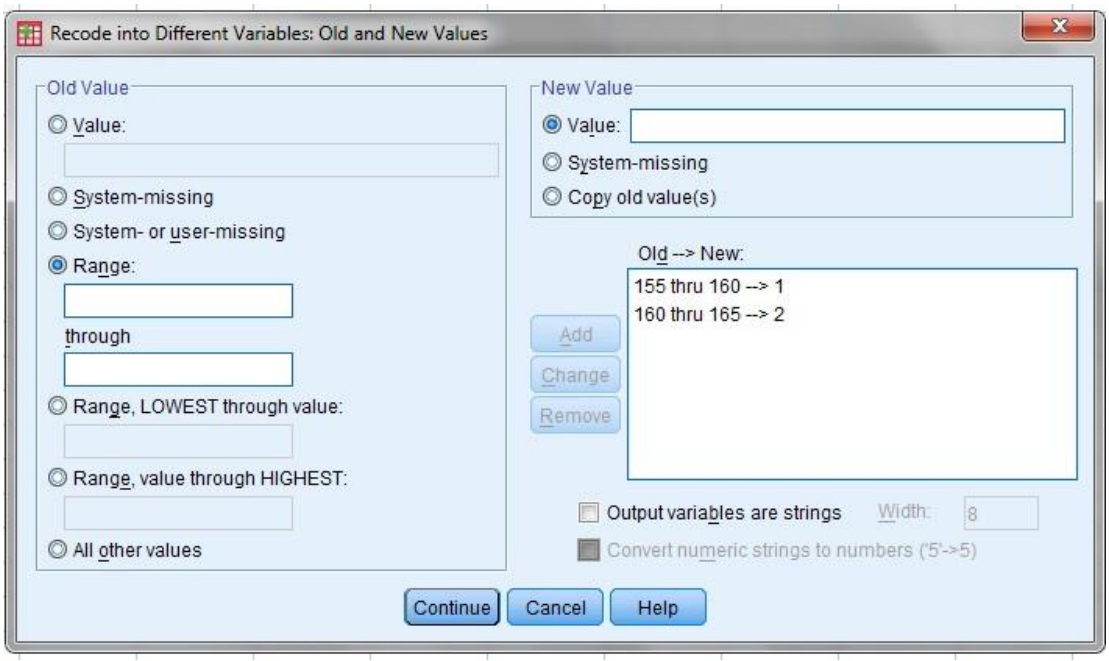
Στη συνέχεια πατάμε στην επιλογή 

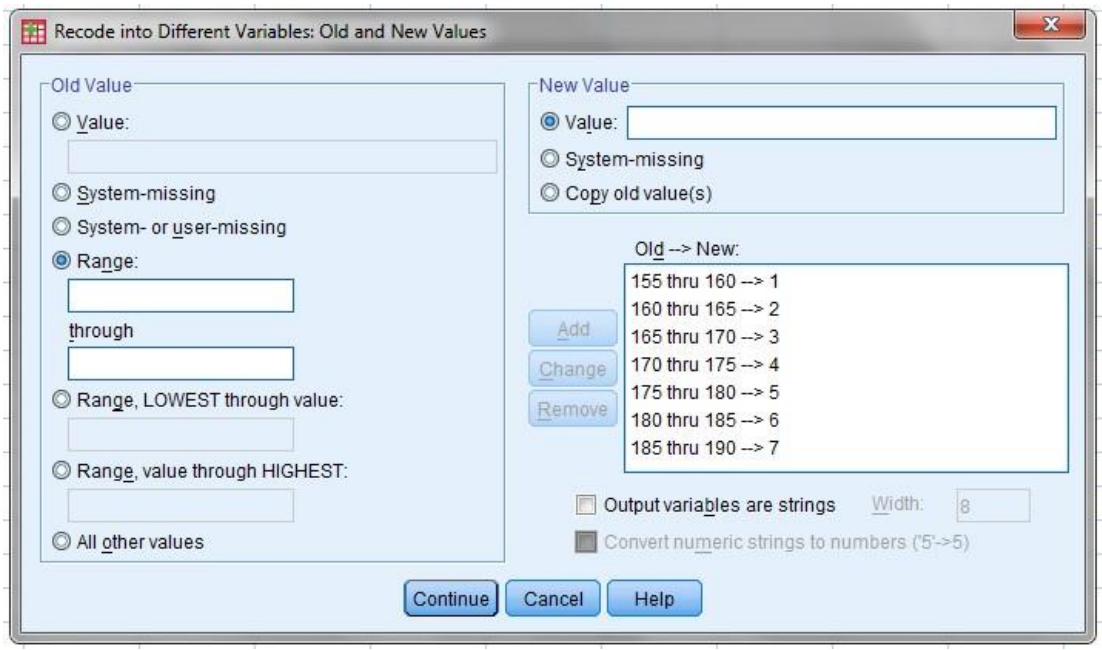
.

Πατάμε στην επιλογή **(OLD VALUE) RANGE** και στα δύο παράθυρα γράφουμε το κατώτερο και το ανώτερο όριο της πρώτης τάξης, δηλ. 155 through 160. Πατάμε δεξιά στην ένδειξη (NEW VALUE) VALUE και δίνουμε τον κωδικό ο οποίος αντικαθιστά την πρώτη τάξη. Δηλαδή, για την ομάδα 155 – 160 ο κωδικός είναι 1.



Στην συνέχεια πατάμε  για την εισαγωγή της τάξης και του κωδικού της στο παράθυρο.



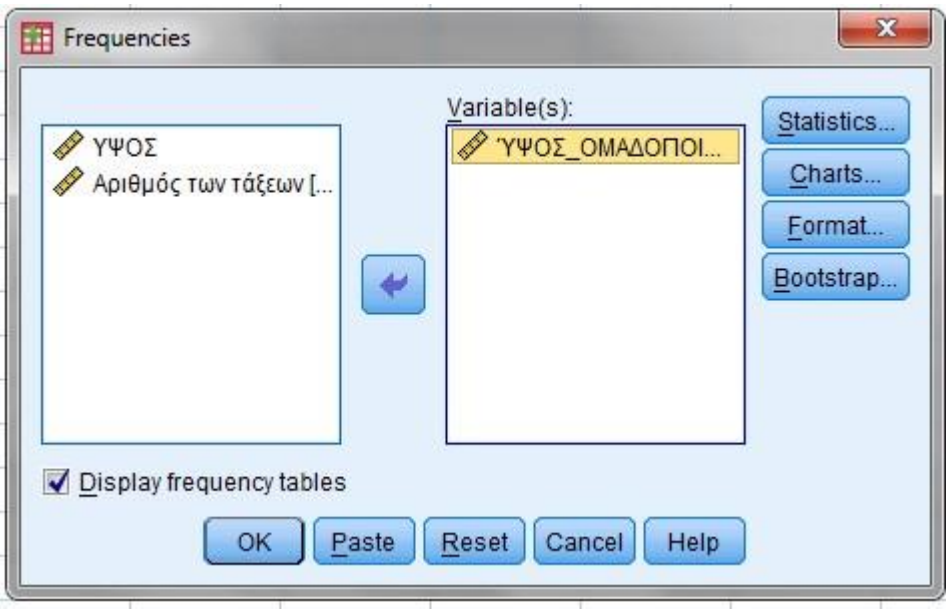


Συνεχίζουμε με τον τρόπο αυτό μέχρι να τελειώσουν όλες οι τάξεις. Στη συνέχεια CONTINUE – Ο.Κ.

Με την παραπάνω διαδικασία στο φύλλο DATA EDITOR δημιουργήσαμε μία νέα μεταβλητή την **΄ΥΨΟΣ_ΟΜΑΔΟΠΟΙΗΣΗ΄** στην οποία πλέον εμφανίζονται οι κωδικοί και όχι οι τιμές της μεταβλητής **΄ΥΨΟΣ**΄.

4) **Για να δημιουργήσουμε τον πίνακα κατανομής συχνοτήτων για τα ομαδοποιημένα δεδομένα,** δημιουργούμε τον πίνακα κατανομής συχνοτήτων της νέας μεταβλητής που δημιουργήσαμε, ακλουθώντας την γνωστή διαδικασία (προσοχή η μεταβλητή που μεταφέρεται δεξιά είναι η **΄ΥΨΟΣ_ΟΜΑΔΟΠΟΙΗΣΗ΄)**:

ANALYSE - DESCRIPTIVE STATISTICS – FREQUENCIES. Επιλέγουμε από το παράθυρο αριστερά την μεταβλητή **΄ΥΨΟΣ_ΟΜΑΔΟΠΟΙΗΣΗ΄** και με πάτημα στο *μαύρο βέλος* αυτή μεταφέρεται στο παράθυρο δεξιά. Τσεκάρουμε την ένδειξη **DISPLAY FREQUENCY TABLES** και στη συνέχεια **Ο.Κ.**



Στο **OUTPUT** δημιουργείται ο πίνακας συχνοτήτων με την παρακάτω μορφή:

Statistics

ΎΨΟΣ_ΟΜΑΔΟΠΟΙΗΣΗ

N	Valid	50
	Missing	0
Range		6,00
Minimum		1,00
Maximum		7,00

ΎΨΟΣ_ΟΜΑΔΟΠΟΙΗΣΗ

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1,00	4	8,0	8,0	8,0
	2,00	3	6,0	6,0	14,0
	3,00	10	20,0	20,0	34,0
	4,00	13	26,0	26,0	60,0
	5,00	10	20,0	20,0	80,0
	6,00	8	16,0	16,0	96,0
	7,00	2	4,0	4,0	100,0
	Total	50	100,0	100,0	

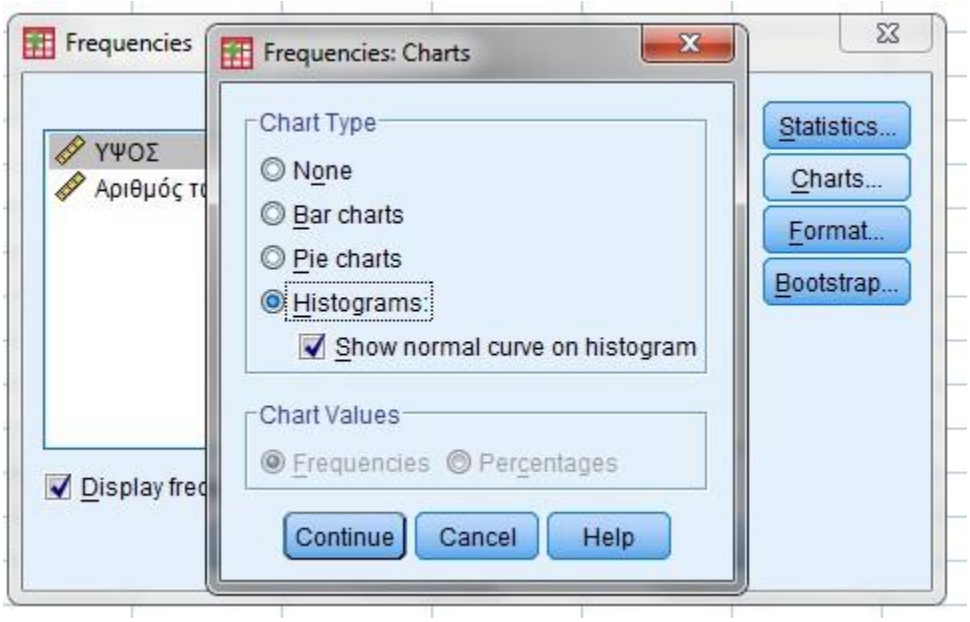
Σύμφωνα με τον παρακάτω πίνακα, 4 από τα 50 άτομα, δηλαδή σε ποσοστό 8,0% έχουν ύψος από 160 έως 165, 3 από τα 50 άτομα, δηλαδή σε ποσοστό 6,0% έχουν ύψος από 165 έως 170, κ.ο.κ..

Β) Για την **κατασκευή ιστογράμματος** ακολουθούμε την παρακάτω διαδικασία :

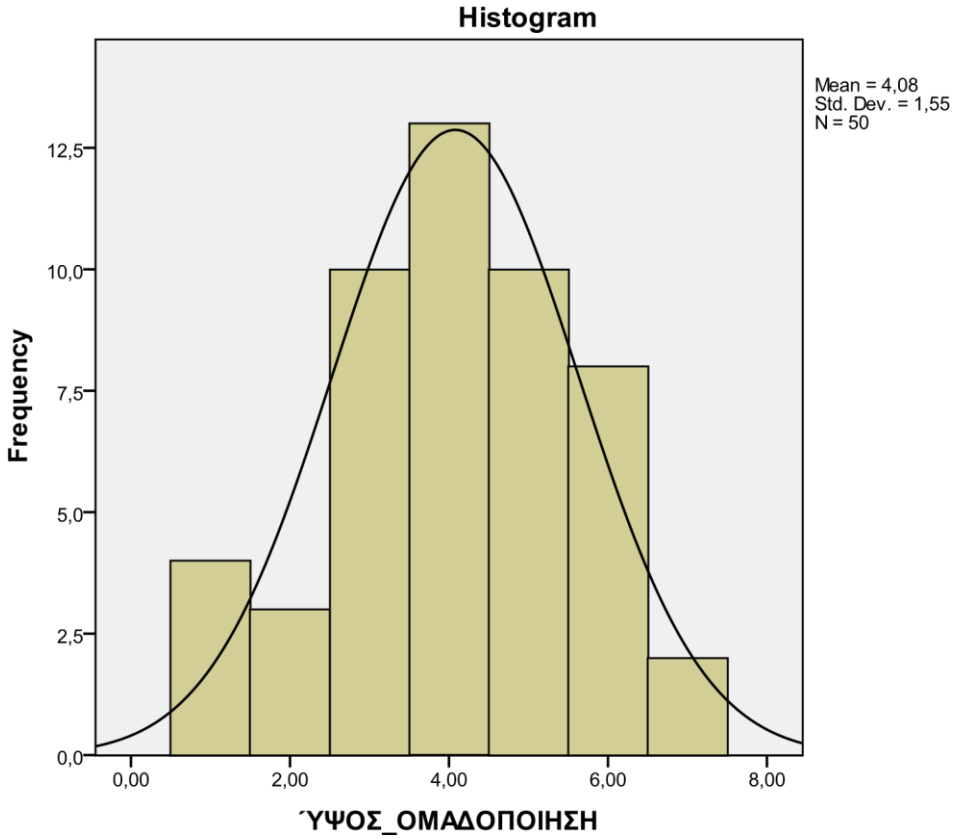
ANALYSE - DESCRIPTIVE STATISTICS - FREQUENCIES

Επιλέγουμε από το παράθυρο αριστερά την μεταβλητή **΄ΥΨΟΣ_ΟΜΑΔΟΠΟΙΗΣΗ΄** και με πάτημα στο **μαύρο βέλος** αυτή μεταφέρεται στο παράθυρο δεξιά.

Τσεκάρουμε την ένδειξη **CHARTS** και εμφανίζεται το παράθυρο **FREQUENCIES : CHARTS** από το οποίο επιλέγουμε **Histograms** και **Show normal curve on histogram** *για* να προστεθεί στο γράφημα η καμπύλη της κανονικής κατανομή.



Στη συνέχεια πατάμε **CONTINUE** και **Ο.Κ.** και εμφανίζεται στο **OUTPUT** το παρακάτω γράφημα:



Γ) Για την **κατασκευή του αθροιστικού διαγράμματος** πρέπει πρώτα να βρούμε τις κεντρικές τιμές κάθε τάξης. Η κεντρική τιμή είναι το ημιάθροισμα του κατώτερου και του ανώτερου ορίου της κάθε τάξης. Οπότε έχουμε :

1η τάξη : 155 – 160 κεντρική τιμή **157,5**

2η τάξη : 160 – 165 κεντρική τιμή **162,5**

3η τάξη : 165 – 170 κεντρική τιμή **167,5**

4η τάξη : 170 – 175 κεντρική τιμή **172,5**

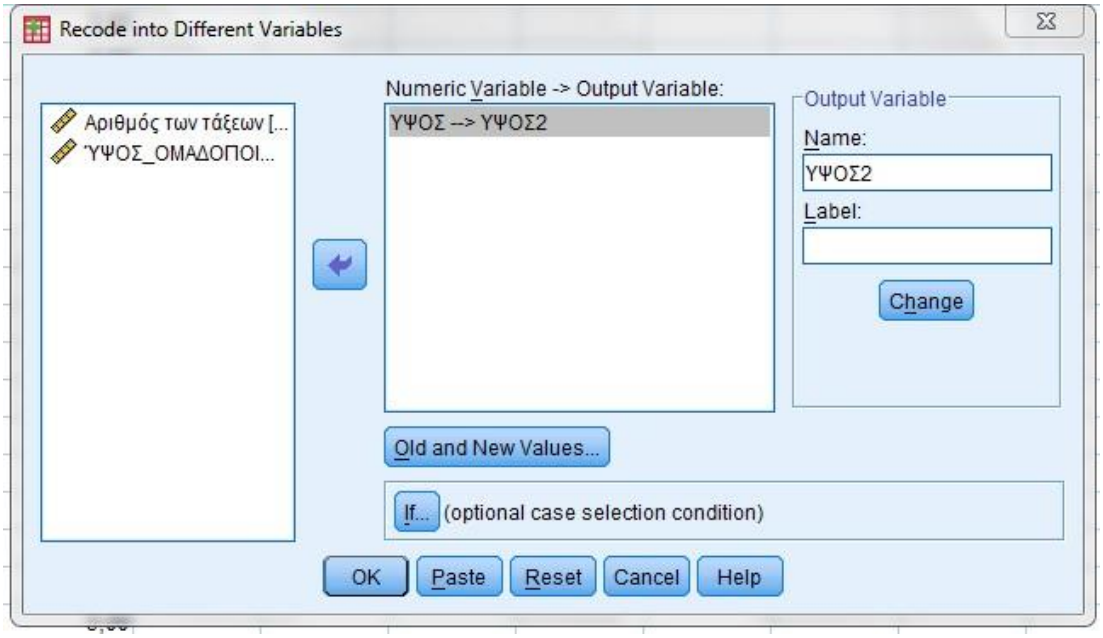
5η τάξη : 175 – 180 κεντρική τιμή **177,5**

6η τάξη : 180 – 185 κεντρική τιμή **182,5**

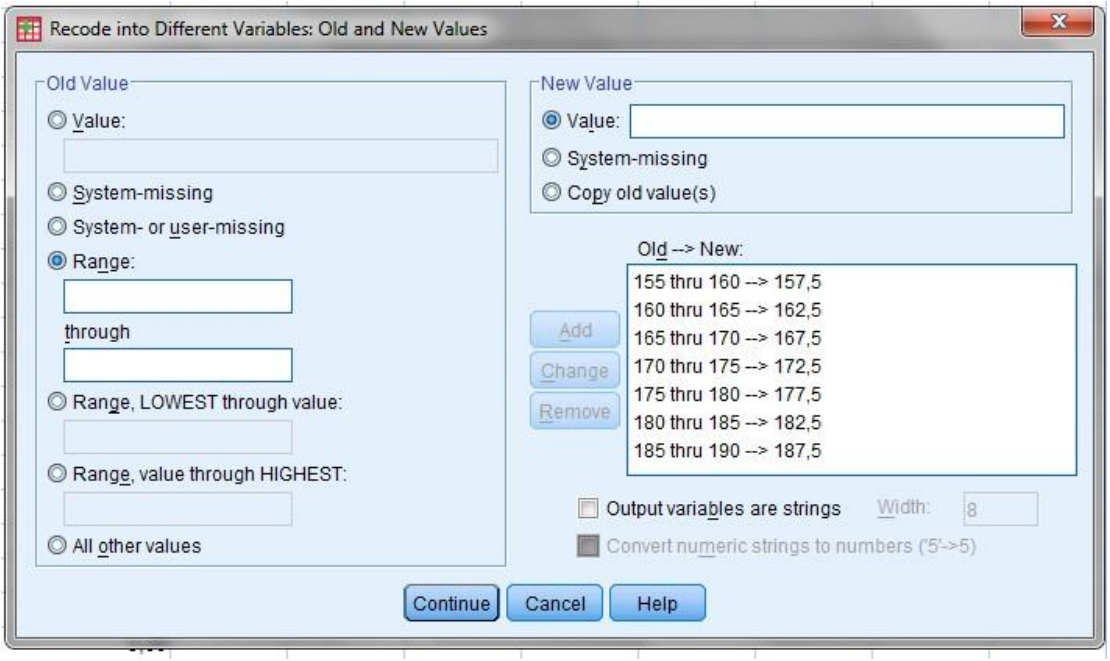
7η τάξη : 185 – 190 κεντρική τιμή **187,5**

Πατάμε στο μενού **TRASFORM - RECODE - INTO DIFFERENT VARIABLES.** Επιλέγουμε από το παράθυρο αριστερά τη μεταβλητή **΄ΥΨΟΣ΄** και με πάτημα στο μαύρο βέλος

αυτή μεταφέρεται στο παράθυρο δεξιά. Στη θέση **NAME** γράφουμε το όνομα που θα δώσουμε στη νέα στήλη **΄ΥΨΟΣ2**΄ και πατώντας στην ένδειξη **CHANGE** το όνομα που δώσαμε, πηγαίνει δίπλα στο όνομα της προηγούμενης μεταβλητής στη θέση του ερωτηματικού.

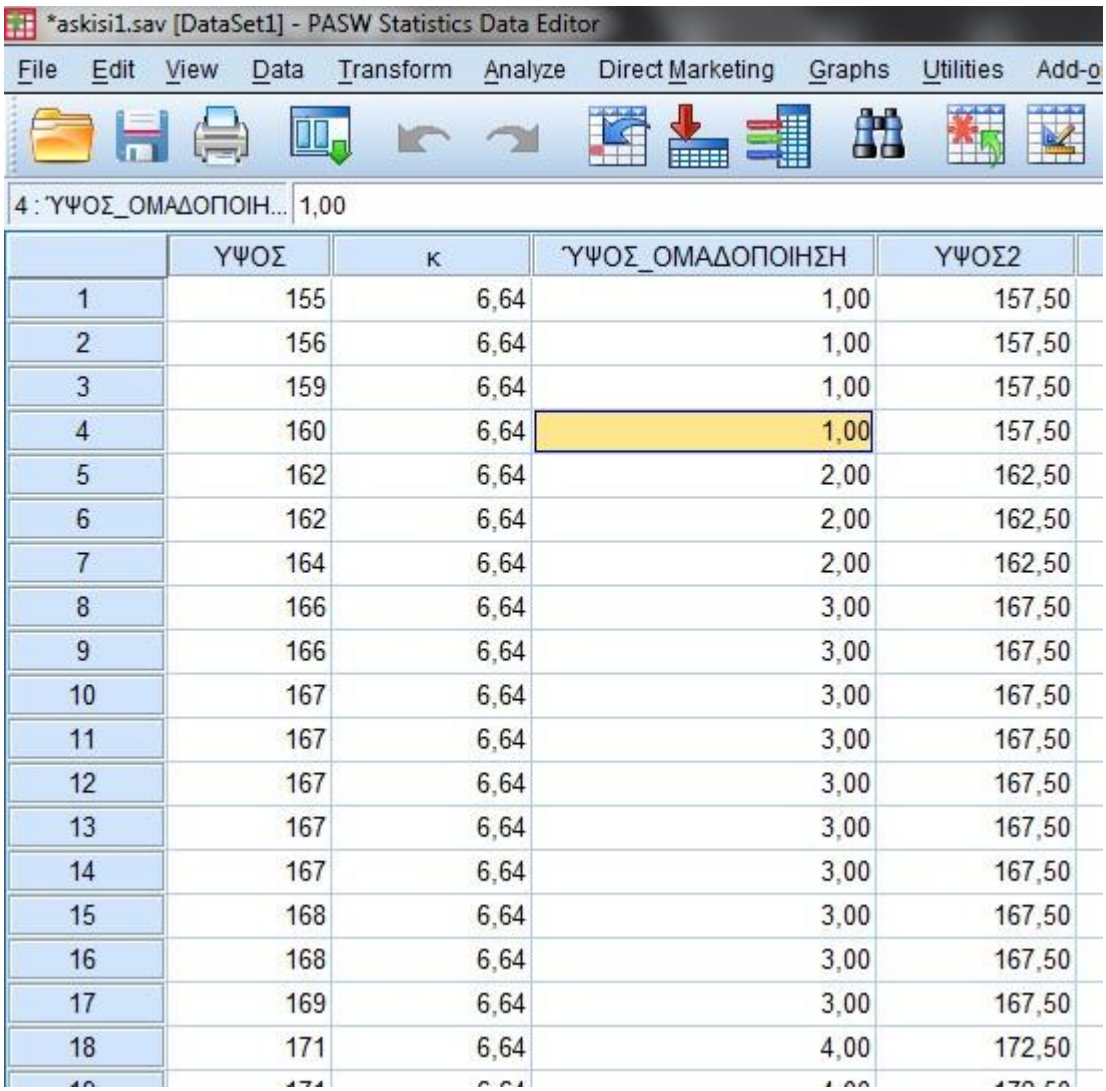


Στη συνέχεια πατάμε στην επιλογή **OLD AND NEW VALUES.** Πατάμε στην επιλογή **(OLD VALUE) RANGE** και στα δύο παράθυρα γράφουμε το κατώτερο και το ανώτερο όριο της πρώτης τάξης, δηλ. 155 through 160. Πατάμε δεξιά στην ένδειξη **(NEW VALUE) VALUE** και δίνουμε την κεντρική τιμή της τάξης. Δηλαδή για την ομάδα 155 – 160 η κεντρική τιμή είναι 157,5. Στην συνέχεια **ADD** για την εισαγωγή της τάξης και της κεντρικής τιμής της στο παράθυρο.



Συνεχίζουμε με τον τρόπο αυτό μέχρι να τελειώσουν όλες οι τάξεις. Στη συνέχεια **CONTINUE – O.K.**

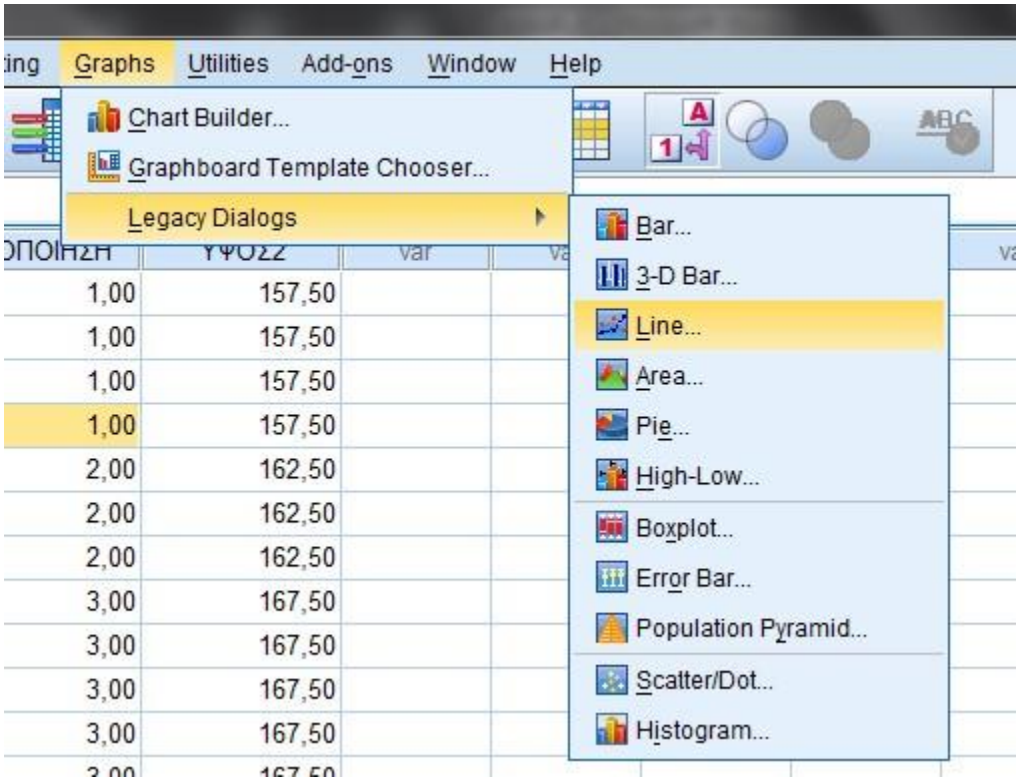
Με την παραπάνω διαδικασία στο φύλλο **DATA EDITOR** δημιουργήσαμε μία νέα μεταβλητή, την **ΎΨΟΣ2΄,** στην οποία πλέον εμφανίζονται οι κεντρικές τιμές των τάξεων, και όχι οι τιμές της μεταβλητής ΎΨΟΣ΄.

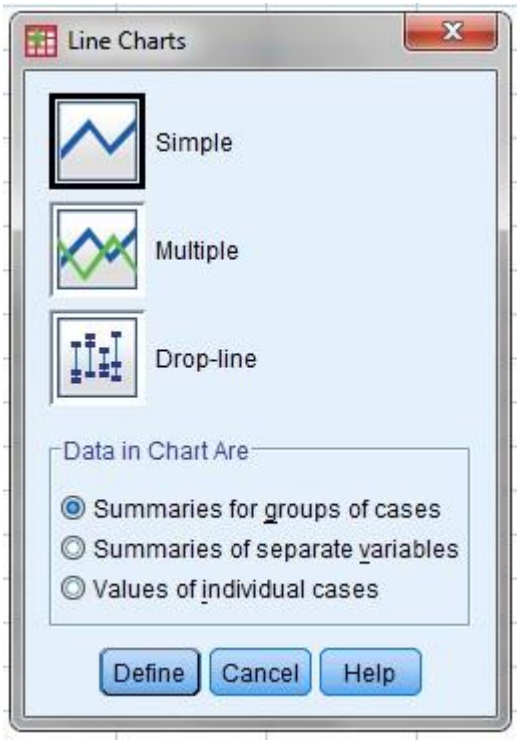


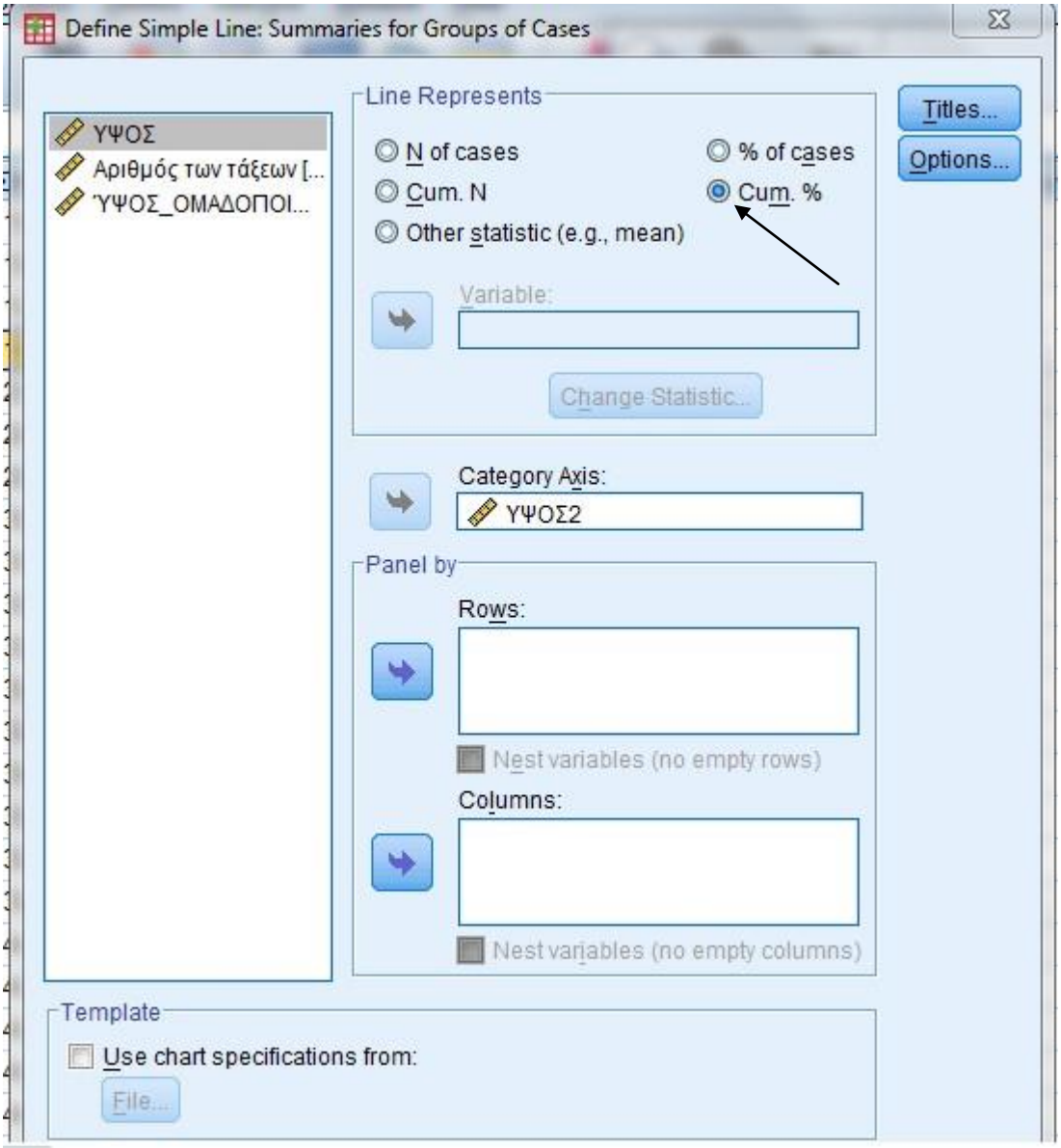
	ΥΨΟΣ	κ	ΎΨΟΣ_ΟΜΑΔΟΠΟΙΗΣΗ	ΥΨΟΣ2
1	155	6,64	1,00	157,50
2	156	6,64	1,00	157,50
3	159	6,64	1,00	157,50
4	160	6,64	1,00	157,50
5	162	6,64	2,00	162,50
6	162	6,64	2,00	162,50
7	164	6,64	2,00	162,50
8	166	6,64	3,00	167,50
9	166	6,64	3,00	167,50
10	167	6,64	3,00	167,50
11	167	6,64	3,00	167,50
12	167	6,64	3,00	167,50
13	167	6,64	3,00	167,50
14	167	6,64	3,00	167,50
15	168	6,64	3,00	167,50
16	168	6,64	3,00	167,50
17	169	6,64	3,00	167,50
18	171	6,64	4,00	172,50

Στην συνέχεια, για την κατασκευή του αθροιστικού διαγράμματος ακολουθούμε την εξής διαδικασία:

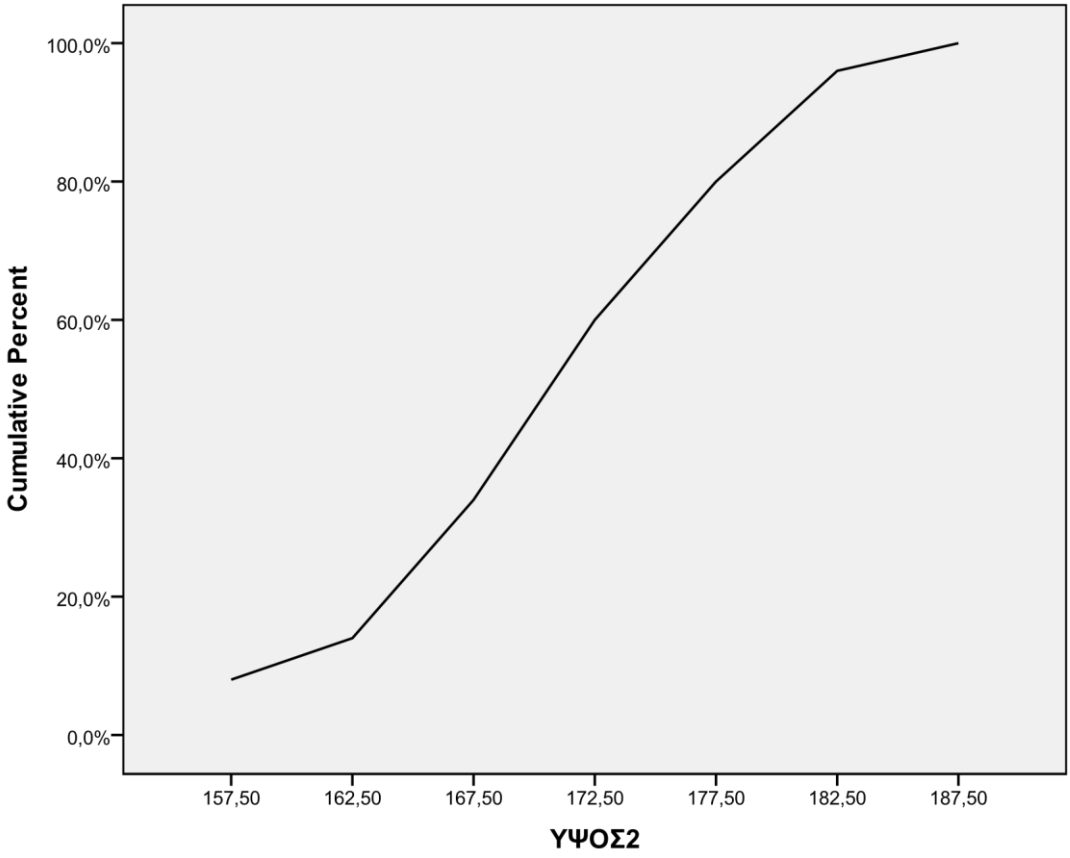
GRAPHS – LIGACY DIALOGS - LINE - SIMPLE - SUMMARIES FOR GROUP OF CASES - DEFINE - Category Axis: ΎΨΟΣ2΄ Line Represents: Cum % - O.K.


Graphs
Utilities
Add-ons
Window
Help
Chart Builder...
Graphboard Template Chooser...
Legacy Dialogs
Bar...
3-D Bar...
Line...
Area...
Pie...
High-Low...
Boxplot...
Error Bar...
Population Pyramid...
Scatter/Dot...
Histogram...
1,00 157,50
1,00 157,50
1,00 157,50
1,00 157,50
2,00 162,50
2,00 162,50
2,00 162,50
3,00 167,50
3,00 167,50
3,00 167,50
3,00 167,50


Line Charts
Simple
Multiple
Drop-line
Data in Chart Are
Summaries for groups of cases
Summaries of separate variables
Values of individual cases
Define
Cancel
Help



στο **OUTPUT** δημιουργείται το αθροιστικό διάγραμμα:



ΠΑΡΑΓΡΑΦΟΣ 2.4 : ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΜΕΤΡΑ ΘΕΣΗΣ ΚΑΙ ΜΕΤΑΒΛΗΤΟΤΗΤΑΣ - ΤΥΠΟΠΟΙΗΜΈΝΕΣ ΤΙΜΈΣ

Σε αυτή τη παράγραφο θα δούμε τα βασικά μέτρα θέσης και μεταβλητότητας καθώς και την ερμηνεία τους.

1. **Αριθμητικός μέσος** μ ή $\bar{x}$(mean): είναι ο μέσος όρος των παρατηρήσεων.

2. **Διάμεσος** M (median): το 50% των παρατηρήσεων είναι κάτω από την τιμή M και το υπόλοιπο 50% πάνω από αυτή.

3. **Επικρατούσα τιμή** M_0 (Mode): είναι η τιμή που εμφανίζεται με τη μεγαλύτερη συχνότητα στα δεδομένα μας.

4. **1ο τεταρτημόριο** Q_1 (Percentiles 25): το 25% των παρατηρήσεων είναι κάτω από την τιμή Q_1 και το υπόλοιπο 75% πάνω από αυτή.

3ο τεταρτημόριο Q_3 (Percentiles 75): το 75% των παρατηρήσεων είναι κάτω από την τιμή Q_3 και το υπόλοιπο 25% πάνω από αυτή.

5. **Ημιενδοτεταρτημοριακό εύρος**: ισούται με την ημιδιαφορά των δύο τεταρτημορίων, δηλαδή $Q = \frac{Q^3 - Q^1}{2}$.

6. **6ο δεκατημόριο** D_6 (Percentiles 60): το 60% των παρατηρήσεων είναι κάτω από την τιμή D_6 και το υπόλοιπο 40% πάνω από αυτή.

   7ο δεκατημόριο D_7 (Percentiles 70): το 70% των παρατηρήσεων είναι κάτω από την τιμή D_7 και το υπόλοιπο 30% είναι πάνω από αυτή.

7. **53ο εκατοστημόριο** P_{53} (Percentiles 53): το 53% των παρατηρήσεων είναι κάτω από την τιμή P_{53} και το υπόλοιπο 47% είναι πάνω από αυτή.

   95ο εκατοστημόριο P_{95} (Percentiles 95): το 95% των παρατηρήσεων είναι κάτω από την τιμή P_{95} και το υπόλοιπο 5% είναι πάνω από αυτή.

8. **Εύρος μεταβολής** R (Range): είναι η διαφορά της μικρότερης από τη μεγαλύτερη τιμή του δείγματός μας.

9. **Διακύμανση** σ^2 ή s^2 : δείχνει τη διασπορά των τιμών του πληθυσμού ή του δείγματος.

10. **Τυπική απόκλιση** σ ή s : ισούται με την τετραγωνική ρίζα της διακύμανσης.

11. **Συντελεστής μεταβλητότητας** CV: ισούται με το πηλίκο της τυπικής απόκλισης προς τον αριθμητικό μέσο.

12. **Συντελεστής ασυμμετρίας**: υπολογίζουμε το πηλίκο $\frac{\text{skewness}}{\text{std error of skewness}}$

    Αν το πηλίκο ανήκει στο διάστημα [-2,2] έχουμε συμμετρική κατανομή.

Αν το πηλίκο είναι μεγαλύτερο του 2 έχουμε θετική ασυμμετρία.

Αν το πηλίκο είναι μικρότερο του -2 έχουμε αρνητική ασυμμετρία.

13. **Συντελεστής κύρτωσης** : υπολογίζουμε το πηλίκο $\frac{\text{kurtosis}}{\text{std error of kurtosis}}$.

Αν το πηλίκο ανήκει στο διάστημα [-2,2] έχουμε μεσόκυρτη ή κανονική κατανομή.

Αν το πηλίκο είναι μεγαλύτερο του 2 έχουμε λεπτόκυρτη κατανομή.

Αν το πηλίκο είναι μικρότερο του -2 έχουμε πλατύκυρτη κατανομή.

Τυποποιημένη τιμή z ονομάζεται η απόσταση της παρατήρησης από τον αριθμητικό μέσο εκφρασμένη σε μονάδες τυπικής απόκλισης και δίνεται από τον τύπο $z = \frac{x - \bar{x}}{s}$, όπου x η τιμή της παρατήρησης.

ΑΣΚΗΣΗ

Η μέση θερμοκρασία σε μία πόλη κατά τον μήνα Απρίλιο είχε ως εξής:

15	16	15	18	20	18	17	19	16	18
17	16	15	20	19	20	22	22	17	20
18	18	17	18	17	20	18	16	21	21

Να υπολογισθούν και να δοθούν ερμηνείες στα παρακάτω ερωτήματα:

1. Αριθμητικός μέσος
2. Διάμεσος
3. $1^{ο}$ και $3^{ο}$ τεταρτημόριο

4. Επικρατούσα τιμή

5. Ημιενδοτεταρτομοριακό εύρος

6. Το 6^{O} και 7^{O} δεκατημόριο

7. Το 53^{o} και 90^{O} εκατοστημόριο

8. Το εύρος μεταβολής

9. Η διακύμανση

10. Η τυπική απόκλιση

11. Ο συντελεστής μεταβλητότητας

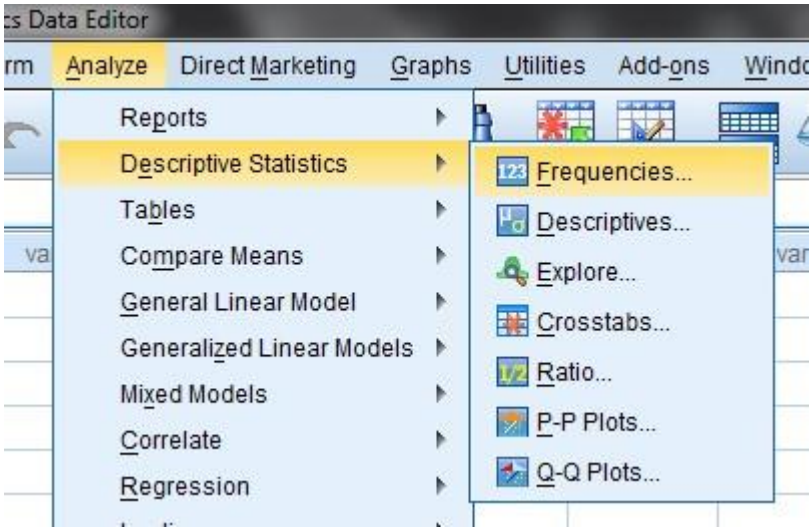
12. Ο συντελεστής ασυμμετρίας

13. Ο συντελεστής κύρτωσης

14. Οι τυποποιημένες τιμές της 18ης και 13ης παρατήρησης

ΛΥΣΗ

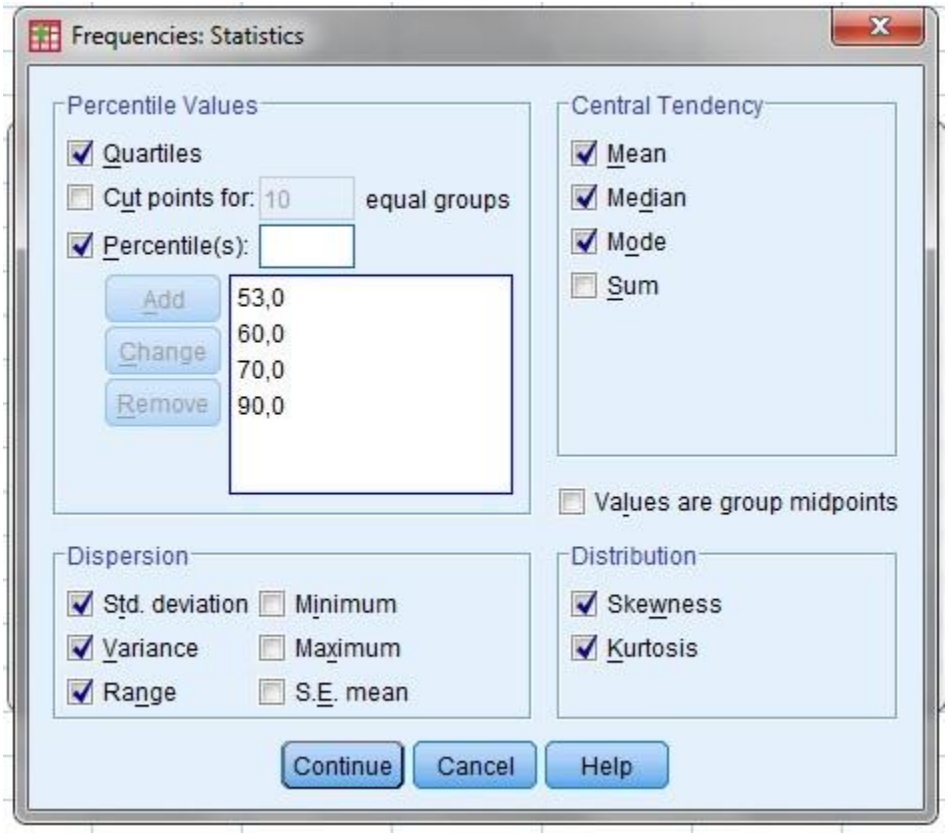
Δημιουργούμε την μεταβλητή **΄ΘΕΡΜΟΚΡΑΣΙΑ΄** και εισάγουμε τα δεδομένα. Για να υπολογίσουμε τα παραπάνω στατιστικά μέτρα θα πρέπει να ακολουθήσουμε την παραπάνω διαδικασία: **Analyze - Descriptive Statistics – Frequencies.**



Μεταφέρουμε με το  την μεταβλητή **΄ΘΕΡΜΟΚΡΑΣΙΑ΄** στο παράθυρο Variable(s) και πατάμε το πεδίο 

.

Στο παράθυρο ***Frequencies Statistics*** επιλέγουμε τα στατιστικά μέτρα που θέλουμε να υπολογιστούν:



1. Για τον αριθμητικό μέσο μ πατάμε **MEAN**

2. Για τη διάμεσο Μ πατάμε **MEDIAN**

3. Για το 1ο και 3ο τεταρτημόριο Q_1 και Q_3 πατάμε **QUARTILES**

4. Για την Επικρατούσα τιμή πατάμε **MODE**

5. Για το Ημιενδοτεταρτομοριακό εύρος έχουμε ήδη πατήσει **QUARTILES**, και υπολογίζεται με την εξής πράξη: $Q=\frac{Q_3-Q_1}{2}$

6. Για το $6^{Ο}$ και $7^{Ο}$ **δεκατημόριο** πατάμε **PERCENTILES** και γράφουμε στο παράθυρο το ποσοστό 60 και στη συνέχεια πατάμε **ADD,** μετά γράφουμε στο παράθυρο το ποσοστό 70 αντίστοιχα και στη συνέχεια πατάμε **ADD.**

7. Το $53^{ο}$ και $90^{Ο}$ **εκατοστημόριο** πατάμε **PERCENTILES** και γράφουμε στο παράθυρο το ποσοστό 53 και στη συνέχεια πατάμε **ADD,** μετά γράφουμε στο παράθυρο το ποσοστό 90 αντίστοιχα και στη συνέχεια πατάμε **ADD**.

8. Για το εύρος μεταβολής πατάμε **RANGE.**
9. Για τη διακύμανση πατάμε **VARIANCE.**

10. Για την τυπική απόκλιση πατάμε **STD. DEVIATION**

11.Για τον συντελεστής μεταβλητότητας $CV=\frac{S}{\mu}$ έχουμε πατήσει ήδη τις

ενδείξεις **STD. DEVIATION** και **MEAN,** οπότε κάνουμε την πράξη.

12. Για τον συντελεστή ασυμμετρίας πατάμε **SKEWNESS**

13. Για τον συντελεστή κύρτωσης πατάμε **KURTOSIS**

14. Οι τυποποιημένες τιμές της 18ης και 13ης παρατήρησης

Στη συνέχεια πατάμε **CONTINUE** και **Ο.Κ.** και εμφανίζεται το OUTPUT στο οποίο περιέχεται ο παρακάτω πίνακας με τα αποτελέσματα.

Statistics

N	Valid	30
	Missing	0
Mean		18,13
Median		18,00
Mode		18
Std. Deviation		2,030
Variance		4,120
Skewness		,257
Std. Error of Skewness		,427
Kurtosis		-,777
Std. Error of Kurtosis		,833
Range		7
Percentiles	25	16,75
	50	18,00
	53	18,00

60	18,00
70	19,70
75	20,00
90	21,00

ΣΧΟΛΙΑΣΜΟΣ:

1. Αριθμητικός μέσος μ=18,13. Δηλ. κατά μέσο όρο η θερμοκρασία στην συγκεκριμένη πόλη τον μήνα Απρίλιο ήταν 18,13 βαθμοί ^{0}C.
2. Διάμεσος Μ= 18. Δηλαδή το 50% των παρατηρήσεων είναι κάτω από 18 βαθμούς ^{0}C και το υπόλοιπο 50% πάνω από αυτή την θερμοκρασία (ή τις μισές μέρες η θερμοκρασία στην πόλη τον Απρίλιο ήταν κάτω από 18 βαθμούς ^{0}C και τις άλλες μισές ήταν πάνω από 18 βαθμούς ^{0}C).
3. Το πρώτο τεταρτημόριο Q_1 = Percentiles 25 = 16,75. Δηλαδή το 25% των παρατηρήσεων είναι κάτω από 16,75 βαθμούς ^{0}C και το υπόλοιπο 75% πάνω από αυτή την τιμή (ή στο 25% των ημερών η θερμοκρασία στην πόλη τον Απρίλιο ήταν κάτω από 16,75 βαθμούς ^{0}C και το 75% των ημερών ήταν πάνω από 16,75 βαθμούς ^{0}C).
4. Το τρίτο τεταρτημόριο Q_3 = Percentiles 75 = 20. Δηλαδή το 75% των παρατηρήσεων είναι κάτω από 20 βαθμούς ^{0}C και το υπόλοιπο 25% πάνω από αυτή την τιμή.
5. Επικρατούσα τιμή Mode=18. Δηλαδή η πιο συνήθης θερμοκρασία για τον μήνα Απρίλιο ήταν 18 βαθμοί ^{0}C.
6. Ημιενδοτεταρτομοριακό εύρος Q= $\frac{Q_3 - Q_1}{2} = \frac{20 - 16,75}{2} = 1,625$

7. 6^{O} δεκατημόριο D_6 = Percentiles 60 = 18. Δηλαδή το 60% των παρατηρήσεων είναι κάτω από 18 βαθμούς ^{0}C και το υπόλοιπο 40% είναι πάνω από αυτή την θερμοκρασία.
8. 7^{O} δεκατημόριο D_7 = Percentiles 70 = 19,70. Δηλαδή το 70% των παρατηρήσεων είναι κάτω από 19,70 βαθμούς ^{0}C και το υπόλοιπο 30% είναι πάνω από αυτή την τιμή.
9. 53^{O} εκατοστημόριο P_{53} = Percentiles 53 = 18. Δηλαδή το 53% των παρατηρήσεων είναι κάτω από 18 βαθμούς ^{0}C και το υπόλοιπο 47% είναι πάνω αυτή την θερμοκρασία.
10. 90^{O} εκατοστημόριο P_{90} = Percentiles 90 = 21. Δηλ. το 90% των παρατηρήσεων είναι κάτω από 21 βαθμούς ^{0}C και το υπόλοιπο 10% είναι πάνω από αυτή την τιμή.
11. Εύρος μεταβολής Range = 7. Δηλαδή η διαφορά μεταξύ μέγιστης και ελάχιστης θερμοκρασίας ήταν 7^{0}C.
12. Διακύμανση (ή διασπορά) σ^2 = 4,12. Ένα πρόβλημα στην ερμηνεία της διασποράς είναι ότι οι μονάδες μέτρησής της είναι οι μονάδα μέτρησης της μεταβλητής στο τετράγωνο. Για το λόγο αυτό ερμηνεύουμε την τυπική απόκλιση που έχει μονάδα μέτρησης ίδια με αυτή της μεταβλητής.
13. Τυπική απόκλιση σ = 2,03 ^{0}C, που δείχνει σχετικά μικρή μεταβλητότητα των τιμών της μεταβλητής.
14. Συντελεστής μεταβλητότητας CV=$\frac{2,03}{18,13}$ = 0,112 ή 11,2%
15. Συντελεστής ασυμμετρίας: υπολογίζουμε το πηλίκο Skewness/ Std. Error of Skewness= $\frac{0,257}{0,427} = 0,6$.

- Αν το πηλίκο ανήκει στο διάστημα [-2, 2] έχουμε συμμετρική κατανομή.
- Αν το πηλίκο είναι μεγαλύτερο από 2 έχουμε θετική ασυμμετρία.
- Αν το πηλίκο είναι μικρότερο από –2 έχουμε αρνητική ασυμμετρία.

Σύμφωνα με τα παραπάνω η συγκεκριμένη μεταβλητή έχει συμμετρική κατανομή.

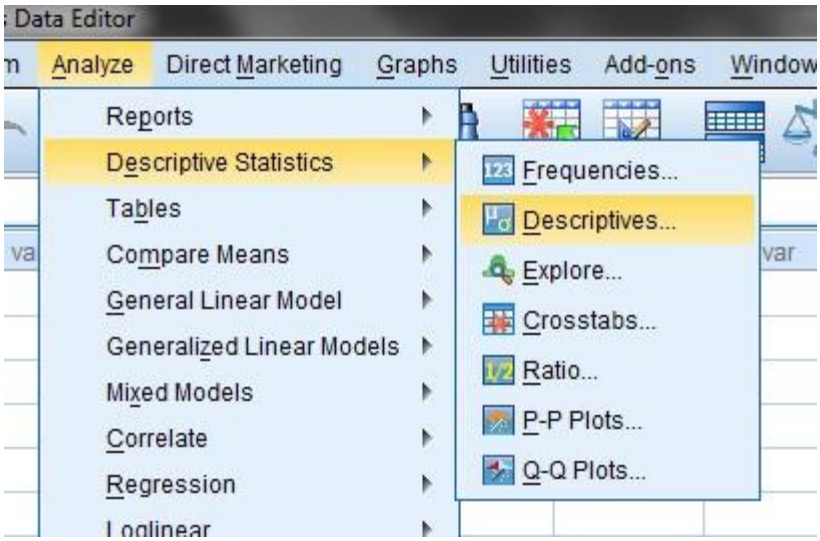
16. Συντελεστής κύρτωσης: Kurtosis **/** Std. Error of Kurtosis = $\frac{-0,777}{0,833} = -0,93$.

- Αν το πηλίκο ανήκει στο διάστημα [-2, 2] έχουμε μεσόκυρτη ή κανονική κατανομή.
- Αν το πηλίκο είναι μεγαλύτερο από 2 έχουμε λεπτόκυρτη κατανομή.
- Αν το πηλίκο είναι μικρότερο από -2 έχουμε πλατύκυρτη κατανομή.

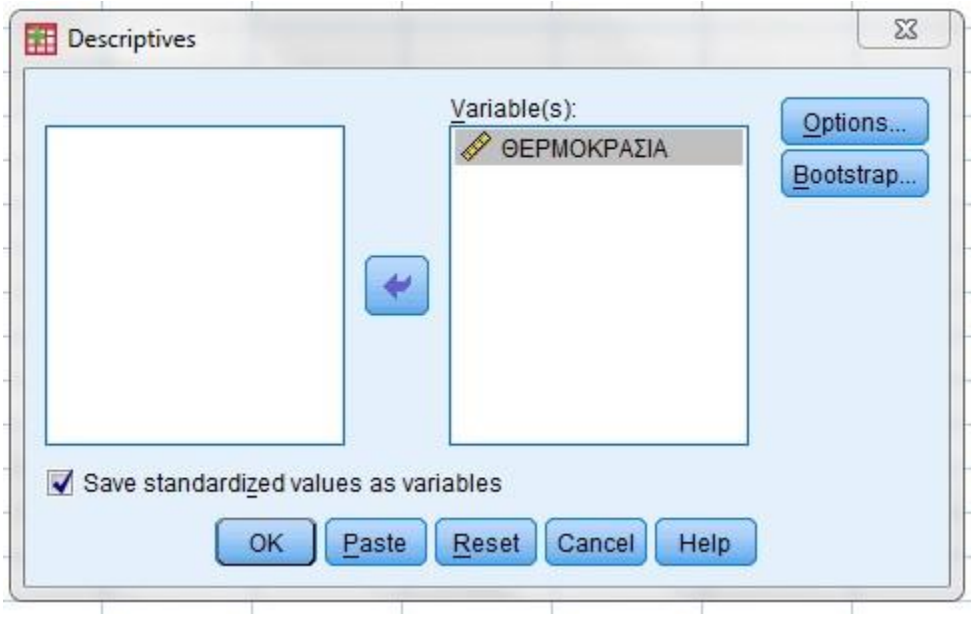
Σύμφωνα με τα παραπάνω έχουμε μεσόκυρτη ή κανονική κατανομή.

Υπολογισμός τυποποιημένων τιμών:

Για το υπολογισμό των τυποποιημένων τιμών θα πρέπει να ακολουθήσουμε την εξής διαδικασία: **ANALYSE - DESCRIPTIVE STATISTICS - DESCRIPTIVES**



Μεταφέρουμε με το  την μεταβλητή ΄**ΘΕΡΜΟΚΡΑΣΙΑ**΄ στο παράθυρο Variable(s) και πατάμε το πεδίο**,** και επιλέγουμε την ένδειξη **Save Standardized Values As Variables** και **Ο.Κ.**



Στο φύλλο **Data View** δημιουργείται μια νέα στήλη η ΖΘΕΡΜΟΚΡΑΣΙΑ η οποία περιέχει τις τυποποιημένες τιμές για κάθε τιμή της μεταβλητής.

	ΘΕΡΜΟΚΡΑΣΙΑ	ΖΘΕΡΜΟΚΡΑΣΙΑ
1	15	-1,54377
2	16	-1,05108
3	15	-1,54377
4	18	-,06569
5	20	,91969
6	18	-,06569
7	17	-,55838
8	19	,42700
9	16	-1,05108
10	18	-,06569
11	17	-,55838
12	16	-1,05108
13	15	-1,54377
14	20	,91969
15	19	,42700
16	20	,91969
17	22	1,90508
18	22	1,90508
19	17	-,55838
20	20	,91969
21	18	-,06569

Επομένως, Z_{18} = 1,90508 δηλαδή με μέση μηνιαία θερμοκρασία 18,13 βαθμούς ^{0}C η 18^{η} παρατήρηση (22^{O}) είναι 1,90508 τυπικές αποκλίσεις **πάνω από τον μέσο**. Z_{13} = -

1,54377 δηλ. η 13η παρατήρηση (15^{O}) είναι 1,54377 τυπικές αποκλίσεις **κάτω από τον αριθμητικό μέσο** (18,13^{O}).

Η ***συσχέτιση*** αναφέρεται στη διερεύνηση της σχέσης ανάμεσα σε μία μεταβλητή X (ανεξάρτητη μεταβλητή) και σε μία μεταβλητή Y (εξαρτημένη μεταβλητή).

Η σχέση μεταξύ των τιμών των δύο μεταβλητών μπορεί να απεικονιστεί με ένα ***διάγραμμα διασποράς*** του σμήνους των σημείων που αντιστοιχούν στις δύο μεταβλητές. Αν υπάρχει μία ευθεία γραμμή γύρω από την οποία μπορούν να προσαρμοστούν τα σημεία, τότε η σχέση μεταξύ των δύο μεταβλητών είναι γραμμική.

Το μέτρο που καθορίζει αυτή τη σχέση είναι ο ***συντελεστής συσχέτισης του Pearson***, ο οποίος παίρνει τιμές από -1 έως +1. Όσο μεγαλύτερη είναι η απόλυτη τιμή του συντελεστή αυτού, τόσο πιο ισχυρή είναι η σχέση των δύο μεταβλητών καθώς και η πρόβλεψη της μίας μεταβλητής με βάση την άλλη. Όταν ο συντελεστής παίρνει αρνητικές τιμές, τότε έχουμε αρνητική συσχέτιση, δηλαδή όταν οι τιμές της μίας μεταβλητής αυξάνονται, οι τιμές της άλλης μειώνονται. Όταν ο συντελεστής παίρνει θετικές τιμές, τότε έχουμε θετική συσχέτιση, δηλαδή όταν οι τιμές της μίας μεταβλητής αυξάνονται, αυξάνονται και της άλλης. Εδώ, πρέπει να επισημάνουμε ότι τέλειες συσχετίσεις στις οποίες η τιμή του συντελεστή είναι -1 ή +1 είναι αδύνατον να βρεθούν. Συσχετίσεις όπου η απόλυτη τιμή του συντελεστή συσχέτισης Pearson βρίσκεται στο διάστημα [0,0.2] χαρακτηρίζονται ως ασήμαντες, στο διάστημα (0.2,0.4] ως μέτριες, στο (0.4,0.7] ως σημαντικές και στο (0.7,1) ως ισχυρές.

Η εγκυρότητα του συντελεστή του Pearson εξαρτάται από το αν ικανοποιούνται τα παρακάτω :

i) τα δεδομένα μας ακολουθούν την κανονική κατανομή, ii)

οι μεταβλητές συσχετίζονται γραμμικά.

Στην ***απλή γραμμική παλινδρόμηση*** πρέπει να βρούμε μια συνάρτηση, ή αλλιώς μια εξίσωση πρόβλεψης, με την οποία θα μπορούμε να προβλέψουμε τις τιμές της μεταβλητής Y βασιζόμενοι στις τιμές της μεταβλητής X. Όσο πιο ισχυρή είναι η συσχέτιση των δύο μεταβλητών τόσο πιο ακριβής θα είναι η πρόβλεψη. Η πρόβλεψη

αυτή θα γίνεται με τη βοήθεια της εξίσωσης πρόβλεψης που ονομάζεται ***εξίσωση παλινδρόμησης*** και είναι η εξίσωση μιας ευθείας.

Μπορούμε να βρούμε πολλές τέτοιες ευθείες, η ευθεία όμως που ταιριάζει καλύτερα στα δεδομένα μας είναι αυτή που βρίσκουμε με την μέθοδο ελαχίστων τετραγώνων. Η μέθοδος αυτή ελαχιστοποιεί το άθροισμα των τετραγώνων των αποκλίσεων από την ευθεία, όλων των σημείων του διαγράμματος διασποράς.

Η καλύτερη δυνατή ευθεία λέγεται ***ευθεία παλινδρόμησης και έχει εξίσωση*** $y = b_0 + b_1 \cdot x$ **,** όπου b_1 είναι ο συντελεστής διεύθυνσης της ευθείας και παριστάνει τη μεταβολή της εξαρτημένης μεταβλητής Υ όταν η ανεξάρτητη μεταβλητή Χ μεταβληθεί κατά μία μονάδα και b_0 είναι η τιμή της εξαρτημένης μεταβλητής Υ όταν η ανεξάρτητη μεταβλητή Χ γίνει μηδέν.

- Στις ασκήσεις που αφορούν συσχέτιση ή παλινδρόμηση δύο μεταβλητών, είναι απαραίτητο να μπορούμε να αναγνωρίσουμε ποια μεταβλητή είναι ανεξάρτητη και ποια εξαρτημένη προκειμένου να αποφασίσουμε ποια θα ονομάσουμε Χ και ποια Υ.

  Σε κάποιες ασκήσεις είναι ξεκάθαρο ποια μεταβλητή εξαρτάται από ποια (και οφείλουμε να το αναγνωρίσουμε αλλιώς η άσκηση δεν θα λυθεί σωστά) και σε κάποιες άλλες όχι. Σε αυτές τις ασκήσεις που δεν είναι ξεκάθαρη η εξάρτηση των μεταβλητών ή θα μας καθορίζει η εκφώνηση ποια μεταβλητή να θέσουμε ως ανεξάρτητη και ποια ως εξαρτημένη ή θα επιλέγουμε εμείς αυτό που θέλουμε.

ΑΣΚΗΣΗ 1

Ο παρακάτω πίνακας περιέχει τα δεδομένα 16 υπαλλήλων μιας μεγάλης πολυεθνικής εταιρίας σχετικά με τον μισθό του δεκαπενθημέρου και τα χρόνια υπηρεσίας στην

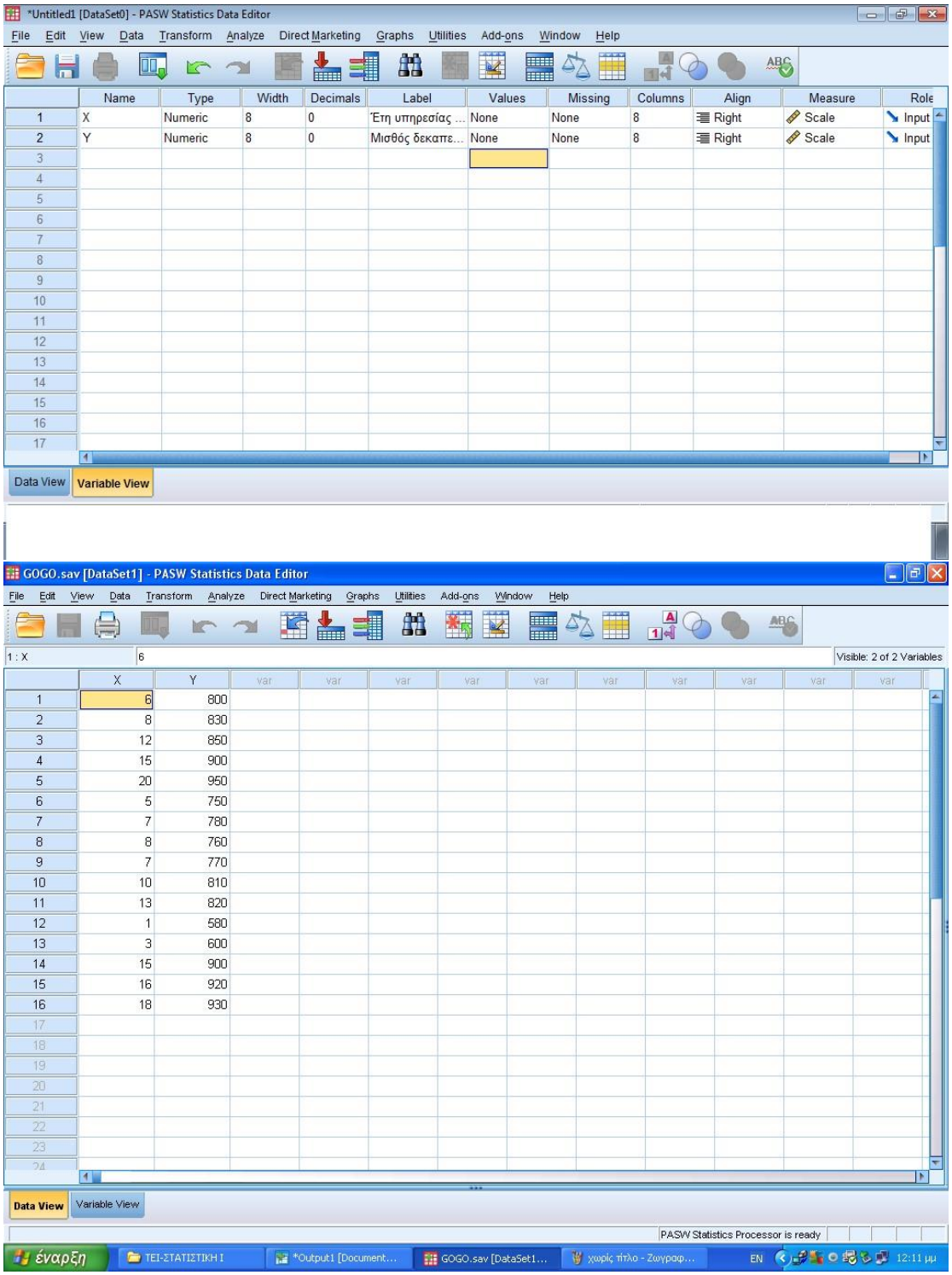
εταιρία. Να κατασκευαστεί το διάγραμμα διασποράς, να βρεθεί ο συντελεστής συσχέτισης Pearson και να σχολιασθούν τα αποτελέσματα.

Α/Α	ΜΙΣΘΟΣ	ΕΤΗ ΥΠΗΡΕΣΙΑΣ
1	800	6
2	830	8
3	850	12
4	900	15
5	950	20
6	750	5
7	780	7
8	760	8
9	770	7
10	810	10
11	820	13
12	580	1
13	600	3
14	900	15
15	920	16
16	930	18

ΛΥΣΗ

Στη συγκεκριμένη άσκηση είναι προφανές ότι ο μισθός εξαρτάται από τα έτη υπηρεσίας, οπότε θέτουμε σαν X τη μεταβλητή «έτη υπηρεσίας» και σαν Y τη μεταβλητή «μισθός δεκαπενθημέρου».

Εισάγουμε τις δύο μεταβλητές στο Variable View και τα δεδομένα στο Data View.

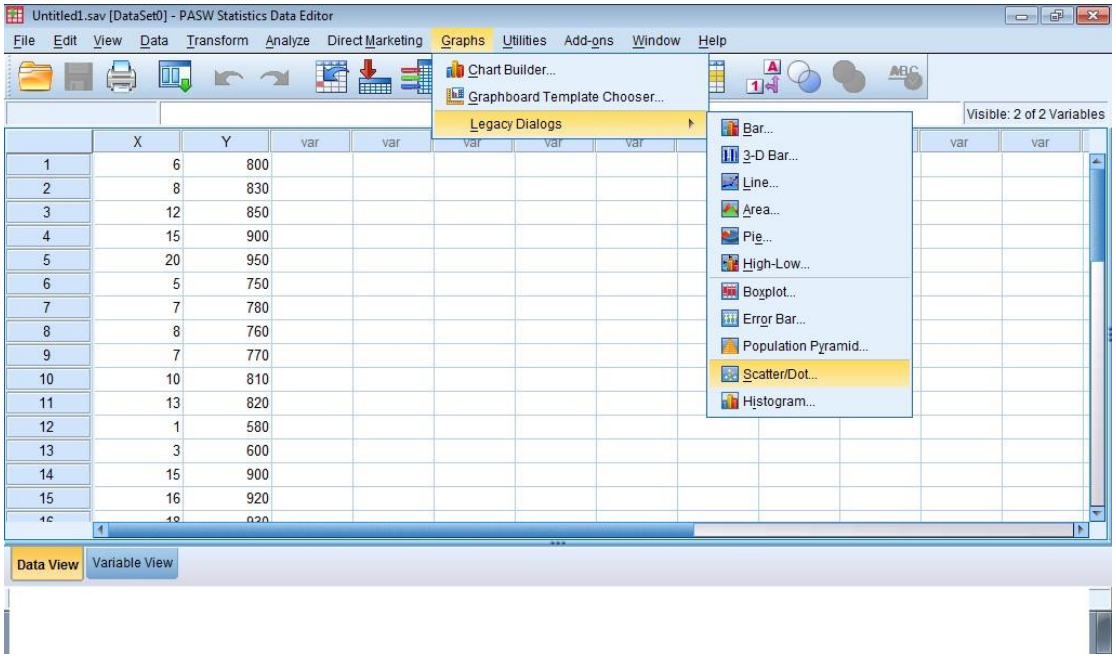


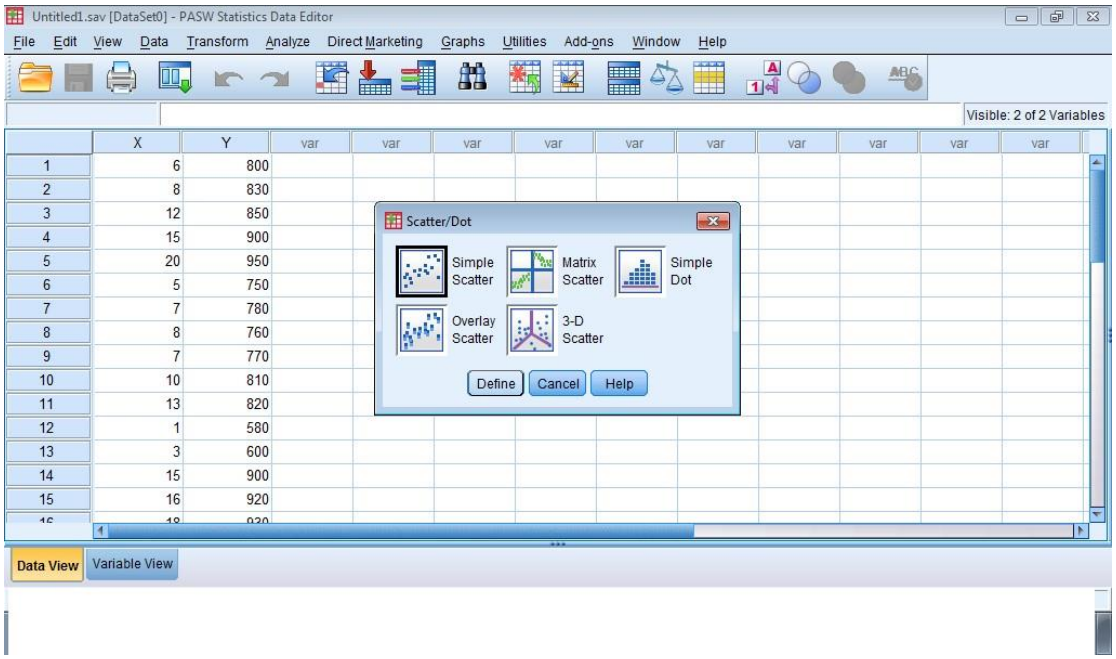
Για να κατασκευάσουμε το διάγραμμα διασποράς ακολουθούμε τα παρακάτω βήματα :

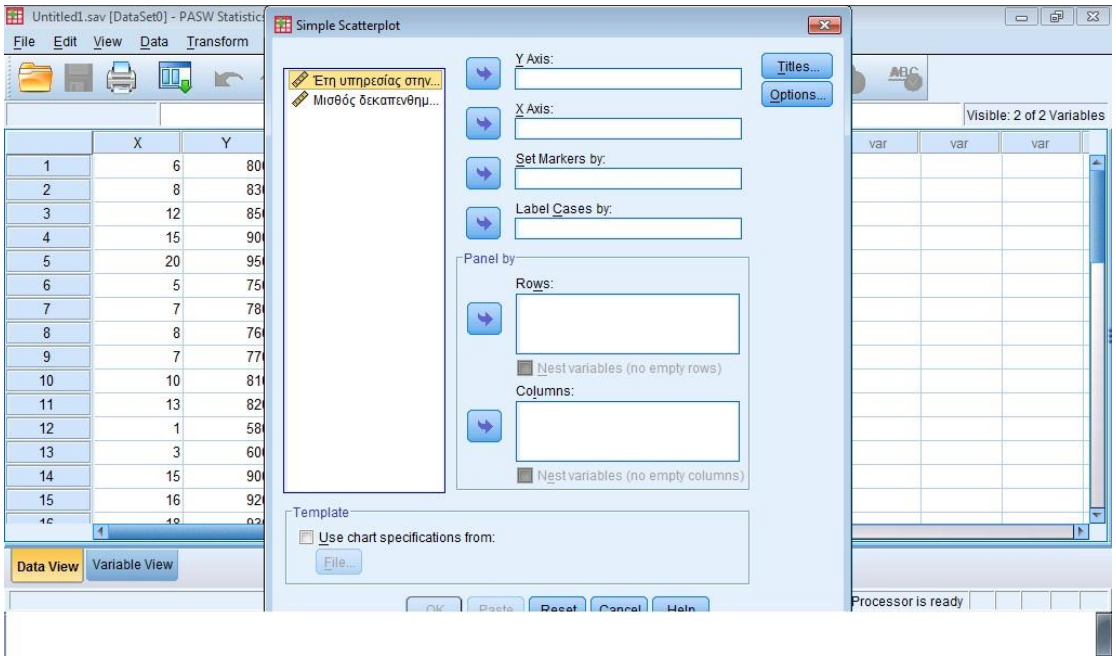
Graphs → Legacy Dialogs → Scatter/Dot → Simple Scatter → Define

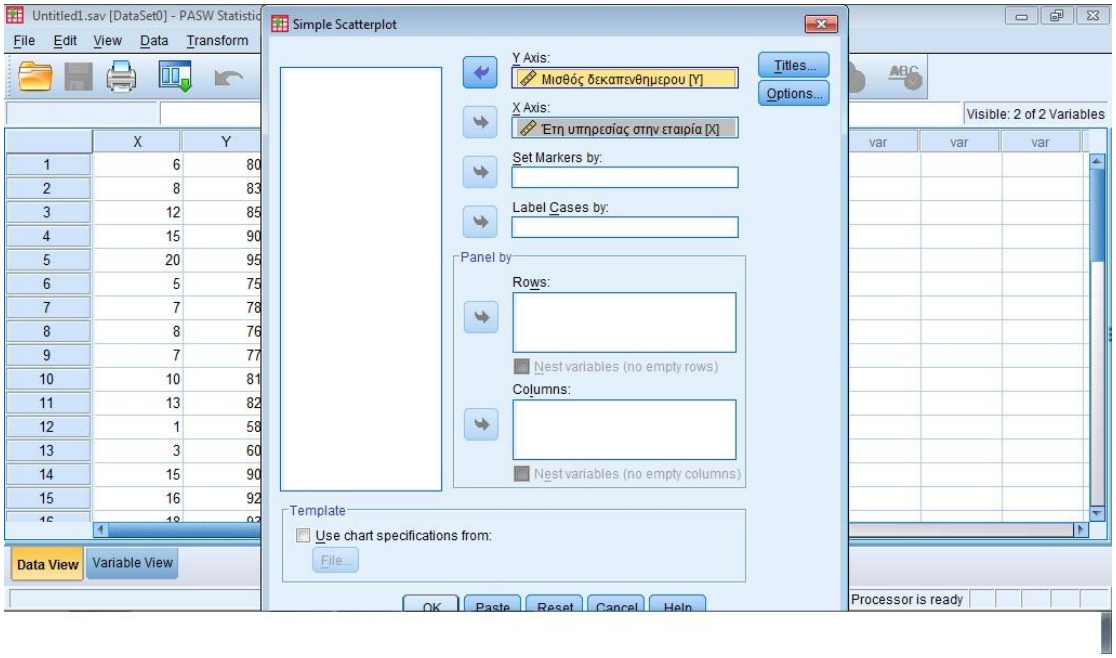
Μεταφέρουμε στο πλαίσιο Y axis τη μεταβλητή Y και στο X axis τη μεταβλητή X

Πατάμε ΟΚ

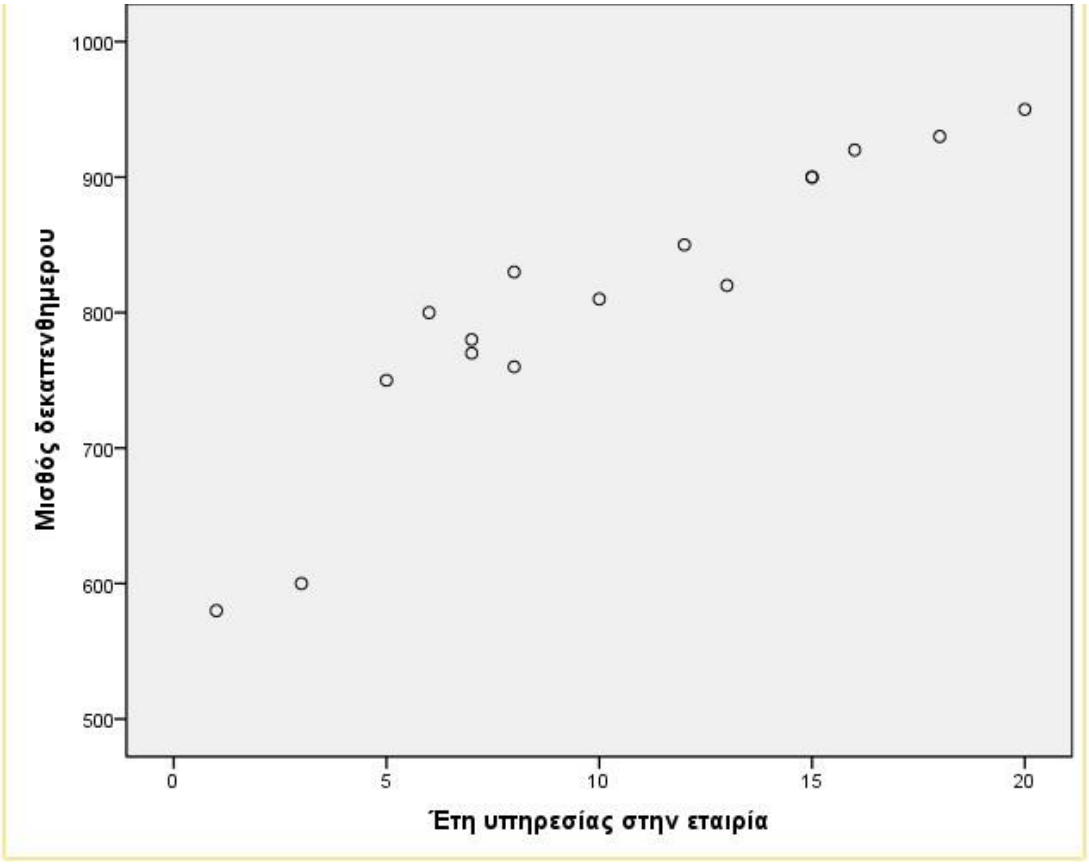





Untitled1.sav [DataSet0] - PASW Statistic
File Edit View Data Transform
Simple Scatterplot
Έτη υπηρεσίας στην...
Μισθός δεκαπενθημ...
Y Axis:
X Axis:
Set Markers by:
Label Cases by:
Titles...
Options...
Panel by
Rows:
Nest variables (no empty rows)
Columns:
Nest variables (no empty columns)
Template
Use chart specifications from:
File...
OK
Paste
Reset
Cancel
Help
Visible: 2 of 2 Variables
var
var
var
X
Y
1 6
2 8
3 12
4 15
5 20
6 5
7 7
8 8
9 7
10 10
11 13
12 1
13 3
14 15
15 16
Data View
Variable View
Processor is ready



Αφού ολοκληρώσουμε τη διαδικασία βλέπουμε το διάγραμμα διασποράς στο οποίο φαίνεται ότι η διασπορά των σημείων είναι τέτοια που μας επιτρέπει να συμπεράνουμε ότι αυτά είναι συγκεντρωμένα γύρω από μια νοητή ευθεία. Η κλίση της ευθείας είναι θετική, σύμφωνα και με τη φορά των σημείων, οπότε τελικά συμπεραίνουμε ότι έχουμε μια θετική γραμμική συσχέτιση.



Για να βρούμε τον συντελεστή συσχέτισης Pearson ακολουθούμε τα παρακάτω βήματα :

Analyze → Correlate → Bivariate

Στο πλαίσιο Variables μεταφέρουμε και τις δύο μεταβλητές

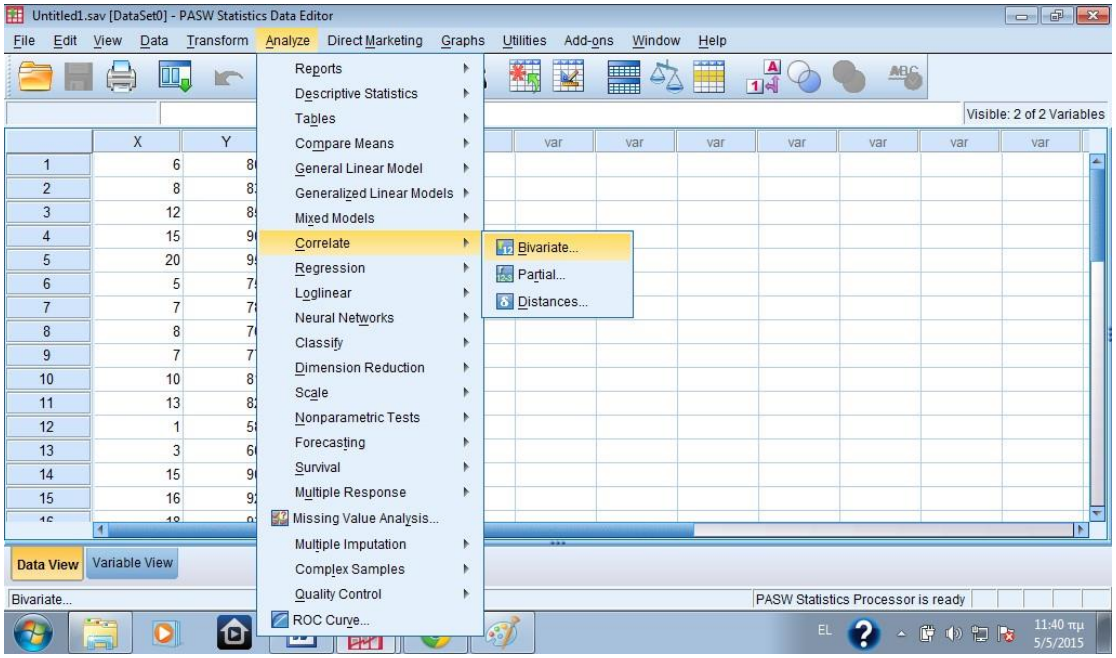
Στη φόρμα Correlation Coefficients επιλέγουμε Pearson

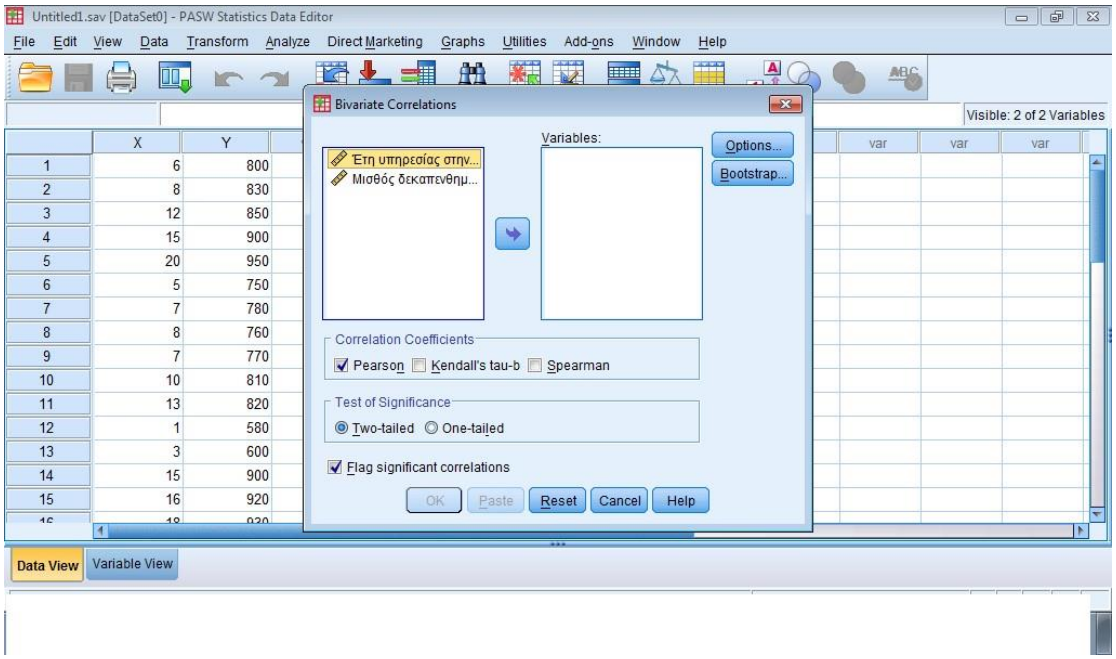
Στη φόρμα Test of Significance επιλέγουμε Two-tailed

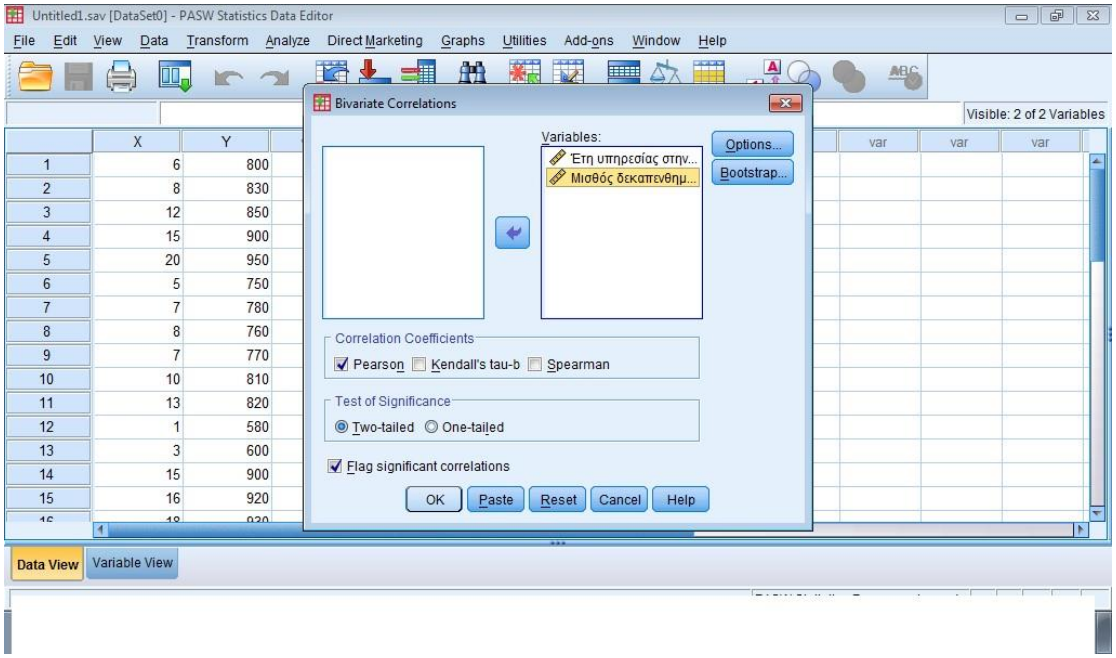
Επιλέγουμε το Flag significant correlations

Πατάμε ΟΚ

Στο Output εμφανίζεται ο πίνακας Correlations όπου βλέπουμε ότι ο συντελεστής συσχέτισης Pearson είναι 0,926. Αυτό σημαίνει ότι έχουμε μια ισχυρή θετική συσχέτιση, εφόσον ο συντελεστής είναι θετικός και η απόλυτη τιμή του ανήκει στο διάστημα [0.7 , 1].


Untitled1.sav [DataSet0] - PASW Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Quality Control
ROC Curve...
Bivariate...
Partial...
Distances...
Visible: 2 of 2 Variables
Data View
Variable View
Bivariate...
PASW Statistics Processor is ready


Untitled1.sav [DataSet0] - PASW Statistics Data Editor
Bivariate Correlations
Έτη υπηρεσίας στην...
Μισθός δεκαπενθημ...
Variables:
Options...
Bootstrap...
Correlation Coefficients
Pearson
Kendall's tau-b
Spearman
Test of Significance
Two-tailed
One-tailed
Flag significant correlations
OK
Paste
Reset
Cancel
Help
Visible: 2 of 2 Variables
Data View
Variable View



Correlations

Correlations

		Έτη υπηρεσίας στην εταιρία	Μισθός δεκαπενθημερου
Έτη υπηρεσίας στην εταιρία	Pearson Correlation	1	,926**
	Sig. (2-tailed)		,000
	N	16	16
Μισθός δεκαπενθημερου	Pearson Correlation	,926**	1
	Sig. (2-tailed)	,000	
	N	16	16

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

ΑΣΚΗΣΗ 2

Για τα δεδομένα της άσκησης 1 (γνωρίζοντας ότι ακολουθούν την κανονική κατανομή), να βρείτε τους συντελεστές της ευθείας παλινδρόμησης, τον δείκτη προσδιορισμού και αφού τους ερμηνεύσετε, να κατασκευάσετε την ευθεία παλινδρόμησης.

ΛΥΣΗ

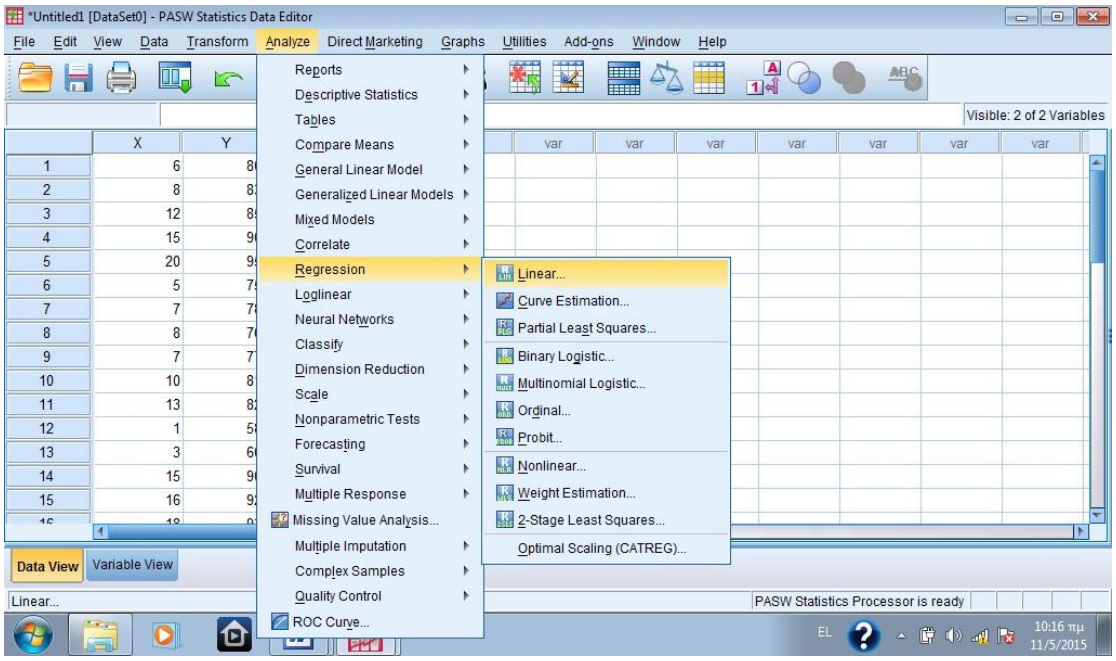
Επιλέγουμε :

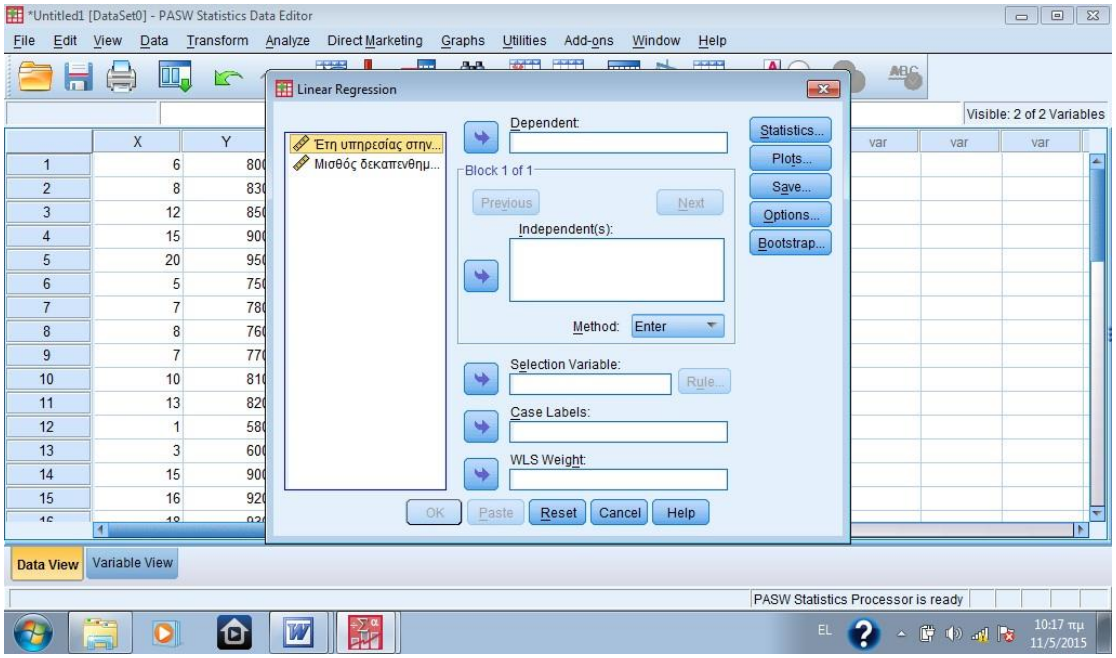
Analyze → Regression → Linear

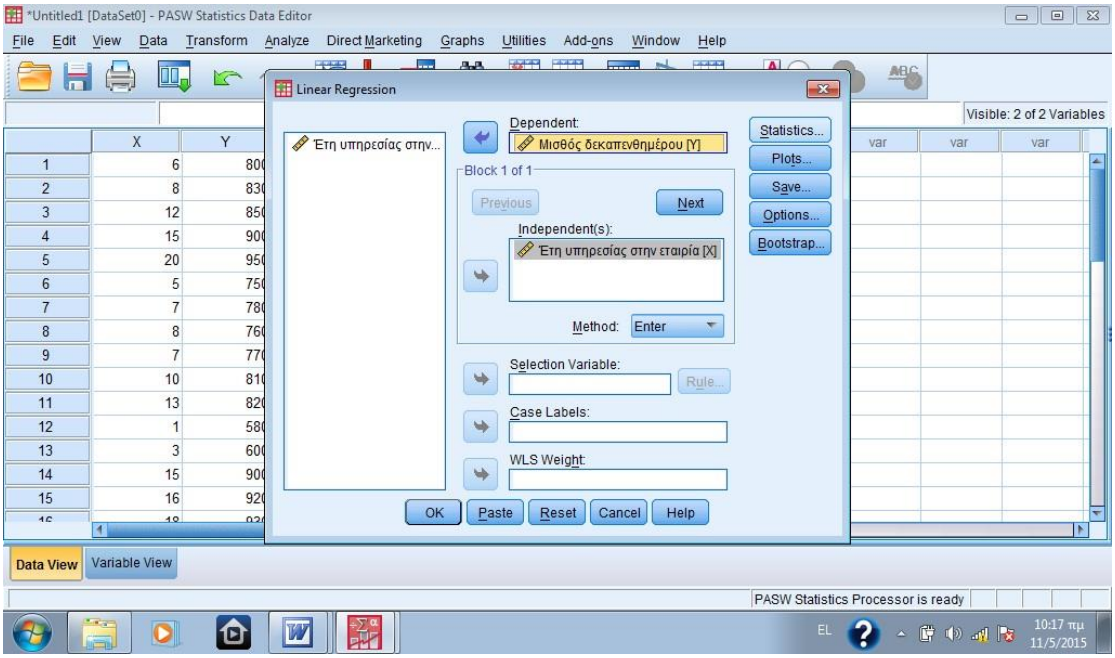
Στο παράθυρο Linear Regression μεταφέρουμε στο πλαίσιο Dependent τη μεταβλητή Y και στο πλαίσιο Independent τη μεταβλητή X.

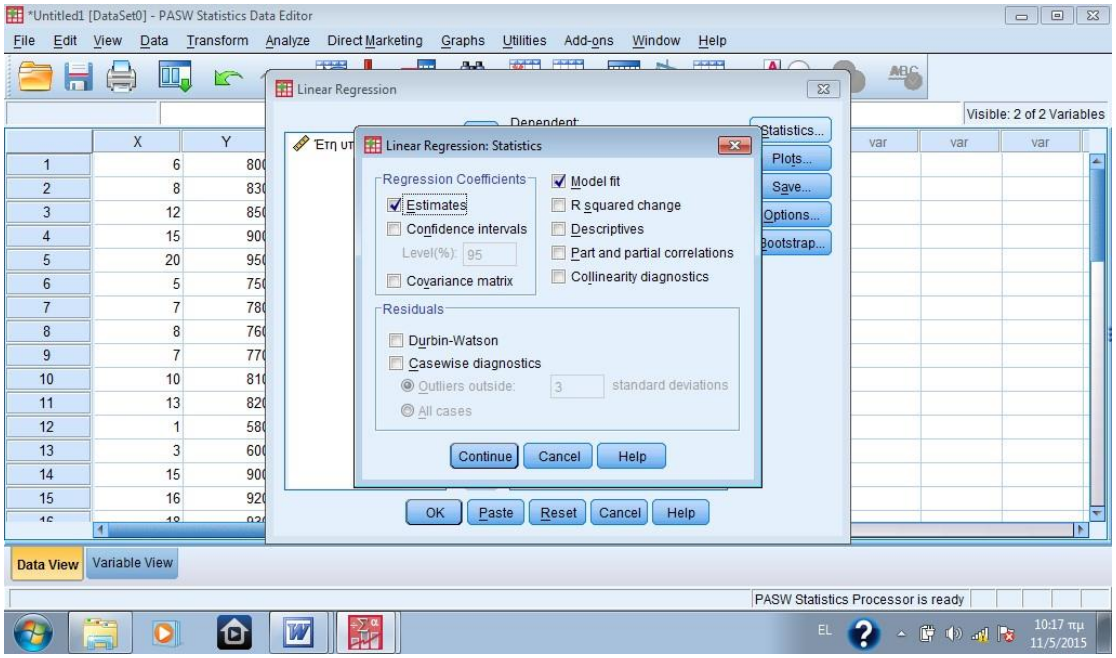
Επιλέγουμε Statistics και κλικάρουμε Estimates και Model Fit.

Πατάμε Continue και OK.




*Untitled1 [DataSet0] - PASW Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Linear Regression
Έτη υπηρεσίας στην...
Μισθός δεκαπενθημ...
Dependent:
Block 1 of 1
Previous
Next
Independent(s):
Method: Enter
Selection Variable:
Rule...
Case Labels:
WLS Weight:
Statistics...
Plots...
Save...
Options...
Bootstrap...
OK
Paste
Reset
Cancel
Help
Visible: 2 of 2 Variables
Data View
Variable View
PASW Statistics Processor is ready


*Untitled1 [DataSet0] - PASW Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Linear Regression
Έτη υπηρεσίας στην...
Dependent:
Μισθός δεκαπενθημέρου [Y]
Block 1 of 1
Previous
Next
Independent(s):
Έτη υπηρεσίας στην εταιρία [X]
Method: Enter
Selection Variable:
Rule...
Case Labels:
WLS Weight:
Statistics...
Plots...
Save...
Options...
Bootstrap...
OK
Paste
Reset
Cancel
Help
Visible: 2 of 2 Variables
Data View
Variable View
PASW Statistics Processor is ready



Στο Output μας ενδιαφέρουν δύο πίνακες : ο πίνακας Model Summary και ο πίνακας Coefficients.

Model Summary

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

a. Predictors: (Constant), X

Coefficients[a]

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

a. Dependent Variable: Y

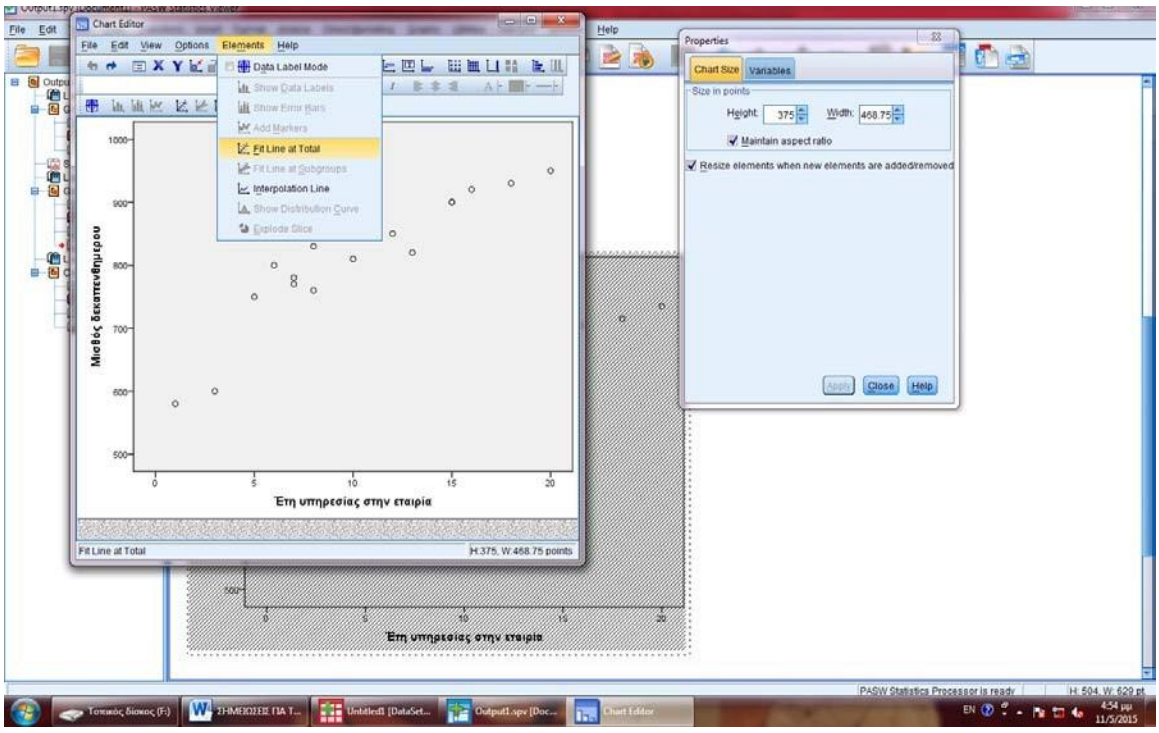
Στον πίνακα Model Summary βλέπουμε το R=0,926 που είναι η απόλυτη τιμή του συντελεστή συσχέτισης και το R Square= R^2=0,858 που είναι ο δείκτης προσδιορισμού. Ο δείκτης προσδιορισμού δηλώνει το ποσοστό της μεταβλητότητας της εξαρτημένης μεταβλητής Y που οφείλεται στην ανεξάρτητη μεταβλητή X. Δηλαδή,

στο συγκεκριμένο παράδειγμα, ο αριθμός 0,858 σημαίνει ότι η μεταβλητότητα του μισθού εξαρτάται σε ποσοστό 85,8% από τα έτη υπηρεσίας.

Στον πίνακα Coefficients βλέπουμε τους συντελεστές b_1,b_0 της ευθείας παλινδρόμησης $y = b_0 + b_1 \cdot x$. Στη συγκεκριμένη περίπτωση οι συντελεστές είναι b_0 $= 626{,}650$ και $b_1 = 17{,}827$ οπότε η εξίσωση της ευθείας είναι $y = 626{,}650 + 17{,}827 \cdot x$. Ο συντελεστής b_0 μας δείχνει ότι αν ένας υπάλληλος είναι νέος και δεν έχει έτη υπηρεσίας στην εταιρία, τότε ο μισθός του είναι 626,650€, ενώ ο συντελεστής b_1 μας δείχνει ότι αν η υπηρεσία μεταβληθεί κατά ένα έτος, τότε ο μισθός μεταβάλλεται κατά 17,827€.

Για να κατασκευάσουμε την ευθεία παλινδρόμησης κάνουμε διπλό κλικ πάνω στο διάγραμμα διασποράς (το οποίο κατασκευάστηκε στην προηγούμενη άσκηση) και ανοίγει ο Chart Editor. Επιλέγουμε Elements → Fit Line at Total.

Κλείνουμε τον Chart Editor.



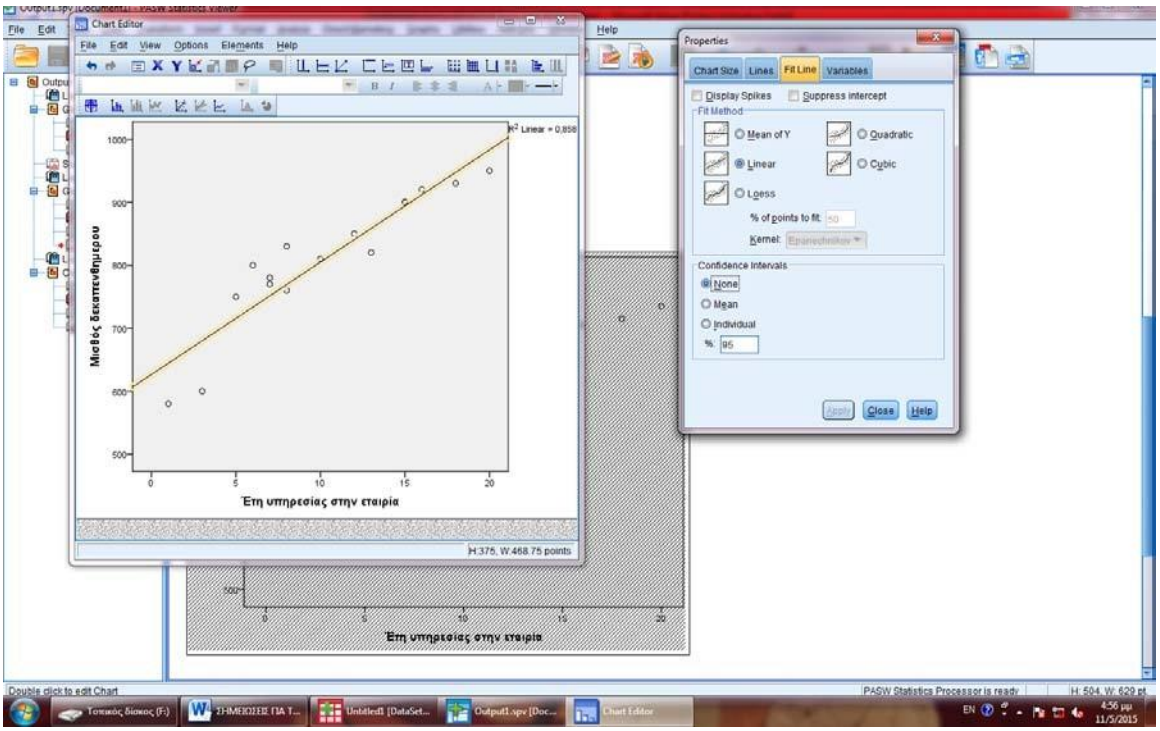
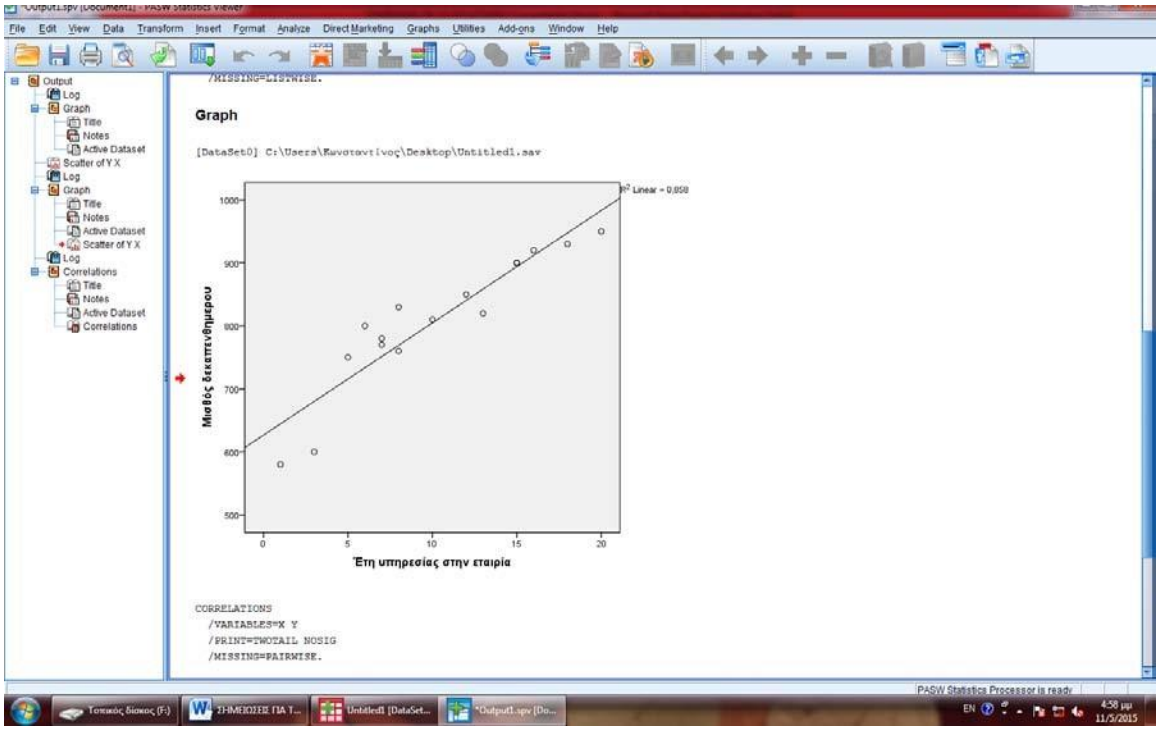


Chart Editor
Properties
Chart Size
Lines
Fit Line
Variables
Display Spikes
Suppress intercept
Fit Method
Mean of Y
Linear
Loess
Quadratic
Cubic
% of points to fit:
Kernel:
Confidence Intervals
None
Mean
Individual
%:
Apply
Close
Help
R² Linear = 0,858
Μισθός δεκαπενθημέρου
Έτη υπηρεσίας στην εταιρία
H:375, W:468,75 points
Double click to edit Chart
PASW Statistics Processor is ready


Output
Log
Graph
Title
Notes
Active Dataset
Scatter of Y X
Correlations
Graph
Μισθός δεκαπενθημέρου
Έτη υπηρεσίας στην εταιρία
CORRELATIONS
/VARIABLES=X Y
/PRINT=TWOTAIL NOSIG
/MISSING=PAIRWISE.
PASW Statistics Processor is ready

ΠΑΡΑΓΡΑΦΟΣ 2.6 : ΟΙ ΠΙΝΑΚΕΣ ΣΥΝΑΦΕΙΑΣ ΚΑΙ Η ΕΥΡΕΣΗ ΠΙΘΑΝΟΤΗΤΩΝ ΜΕ ΤΗ ΧΡΗΣΗ ΑΥΤΩΝ

- Τομή ενδεχομένων

  Τομή δύο ενδεχομένων Α και Β (συμβολίζεται $A \cap B$) είναι το ενδεχόμενο που συμβαίνει όταν τα δύο ενδεχόμενα Α και Β συμβούν ταυτόχρονα. Η πιθανότητα της τομής δύο ενδεχομένων ονομάζεται *συνδυασμένη πιθανότητα (join probability).*

- Ένωση ενδεχομένων

  Ένας άλλος τρόπος συνδυασμού ενδεχομένων είναι η ένωση.

  Ένωση δύο ενδεχομένων Α ή Β (συμβολίζεται $A \cup B$) είναι το ενδεχόμενο που συμβαίνει όταν συμβεί ένα από τα ενδεχόμενα Α ή Β ή και τα δύο.

- Ολική πιθανότητα

Η *ολική πιθανότητα (marginal probability)* υπολογίζεται αθροίζοντας μια γραμμή ή μια στήλη του πίνακα ο οποίος περιέχει τις συνδυασμένες πιθανότητες κάποιων ενδεχομένων.

(Τέτοιους πίνακες θα δούμε στην επίλυση των ασκήσεων)

- Δεσμευμένη πιθανότητα

  - Συχνά αυτό που μας ενδιαφέρει είναι πώς σχετίζονται μεταξύ τους δύο ενδεχόμενα και ιδιαίτερα ποιες είναι οι πιθανότητες να συμβεί ένα ενδεχόμενο αν γνωρίζουμε ότι κάποιο άλλο έχει συμβεί. Η πιθανότητα αυτή ονομάζεται *δεσμευμένη πιθανότητα (conditional probability)* ή αλλιώς πιθανότητα *υπό συνθήκη,* συμβολίζεται ως $P(B_1 | A_1)$ και διαβάζεται «πιθανότητα του B_1 δεδομένου του A_1.

  - Η πιθανότητα ενός ενδεχομένου Α δεδομένου ενός ενδεχομένου Β είναι : $P(A|B) = \frac{P(B \cap A)}{P(B)}$

    Όμοια, η πιθανότητα ενός ενδεχομένου Β δεδομένου του Α είναι :

    $$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

- Ανεξάρτητα ενδεχόμενα

  Ένας από τους στόχους της δεσμευμένης πιθανότητας είναι η εξάρτηση ενός ενδεχομένου από ένα άλλο. Για τον λόγο αυτό εισάγεται η έννοια των *ανεξάρτητων ενδεχομένων (independent events).*

  Δύο ενδεχόμενα Α και Β είναι ανεξάρτητα μεταξύ τους αν είναι :

  $P(A|B) = P(A)$ ή $P(B|A) = P(B)$

  Με απλά λόγια, δύο ενδεχόμενα είναι ανεξάρτητα μεταξύ τους αν η πιθανότητα του ενός δεν επηρεάζεται από την παρουσία του άλλου.

Ο ***πίνακας συνάφειας (contingency table) ή πίνακας διπλής εισόδου (crosstabulation)*** χρησιμοποιείται σαν εργαλείο για την περιγραφή της σχέσης μεταξύ δύο ονομαστικών μεταβλητών.

ΑΣΚΗΣΗ

Στον παρακάτω πίνακα παρουσιάζονται τα δεδομένα για 200 φοιτητές ανάλογα με το φύλο και το κάπνισμα. Το ερώτημα είναι αν το φύλο και το κάπνισμα είναι ανεξάρτητα.

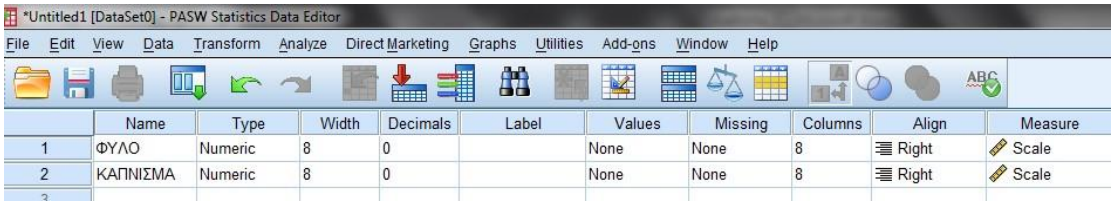
	ΚΑΠΝΙΣΜΑ		
ΦΥΛΟ	ΚΑΠΝΙΣΤΕΣ	ΜΗ ΚΑΠΝΙΣΤΕΣ	**ΣΥΝΟΛΟ**
ΑΓΟΡΙΑ	60	50	**110**
ΚΟΡΙΤΣΙΑ	40	50	**90**
ΣΥΝΟΛΟ	**100**	**100**	**200**

ΛΥΣΗ

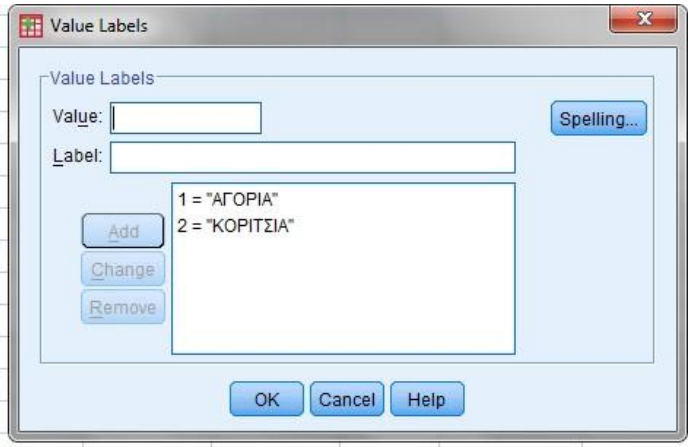
Όταν τα δεδομένα του δείγματος (εδώ 200 άτομα) δίνονται με την μορφή του πίνακα συχνοτήτων τότε το πρώτο βήμα είναι η εισαγωγή των δεδομένων στο SPSS.

<u>Μεθοδολογία για την εισαγωγή των δεδομένων στο SPSS:</u>

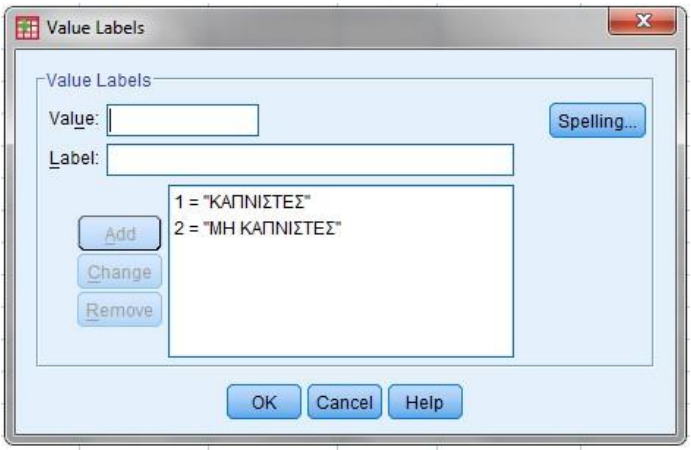
Βήμα 1ο: Δημιουργούμε τις μεταβλητές **ΦΥΛΟ και ΚΑΠΝΙΣΜΑ (Numeric)** με τα εξής labels:



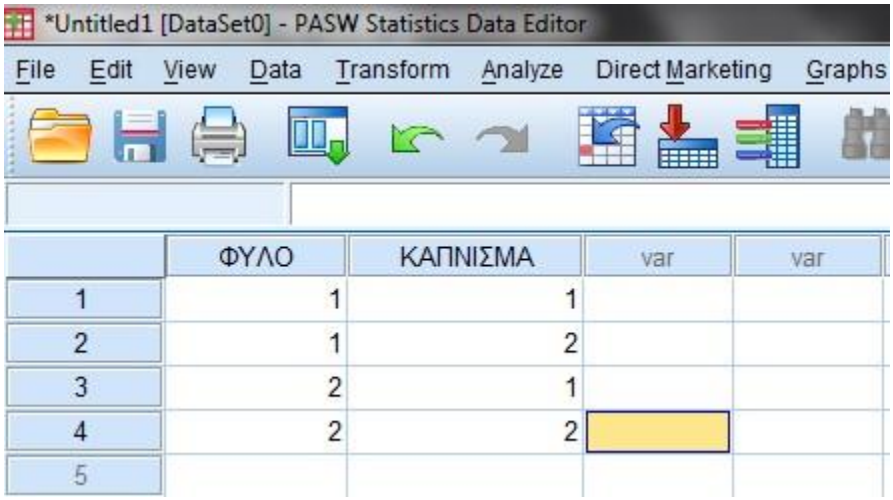
Για την μεταβλητή **ΦΥΛΟ**:



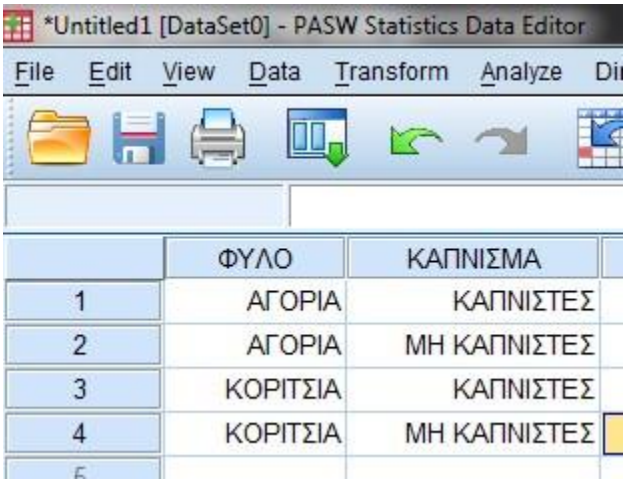
Και για την μεταβλητή **ΚΑΠΝΙΣΜΑ:**



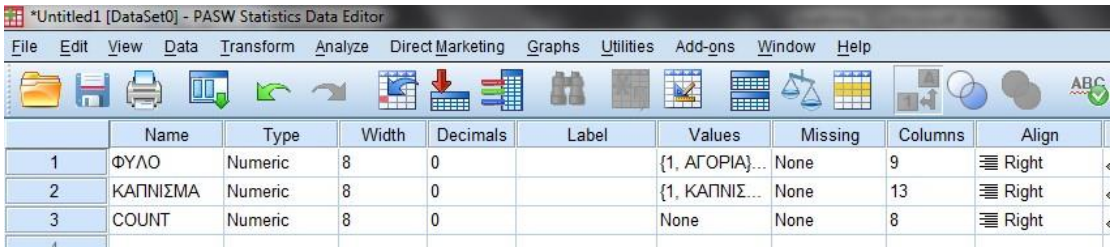
Και στο φύλο **Data View** εισάγουμε ως τιμές για τις μεταβλητές αυτές, όλους τους δυνατούς συνδυασμούς των τιμών τους:

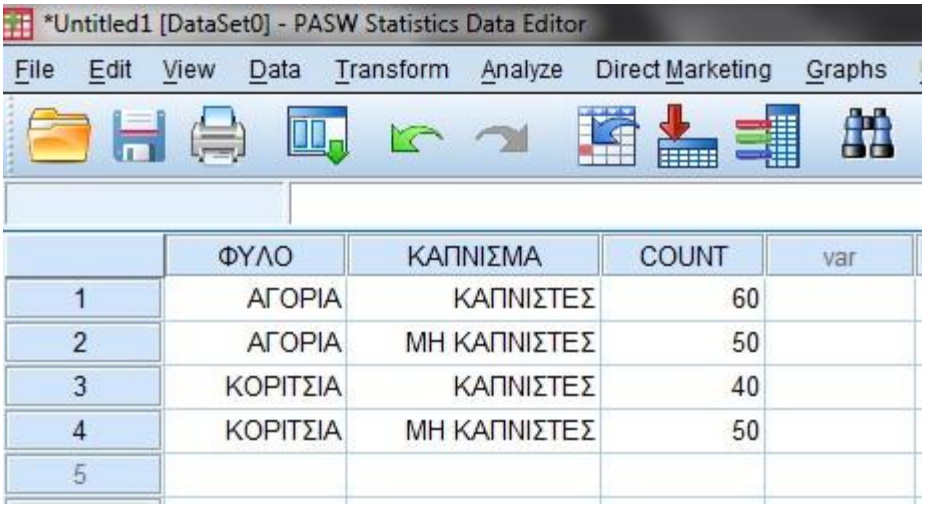


ή αν επιλέξουμε με το εικονίδιο  βλέπουμε τα labels των μεταβλητών αυτών:



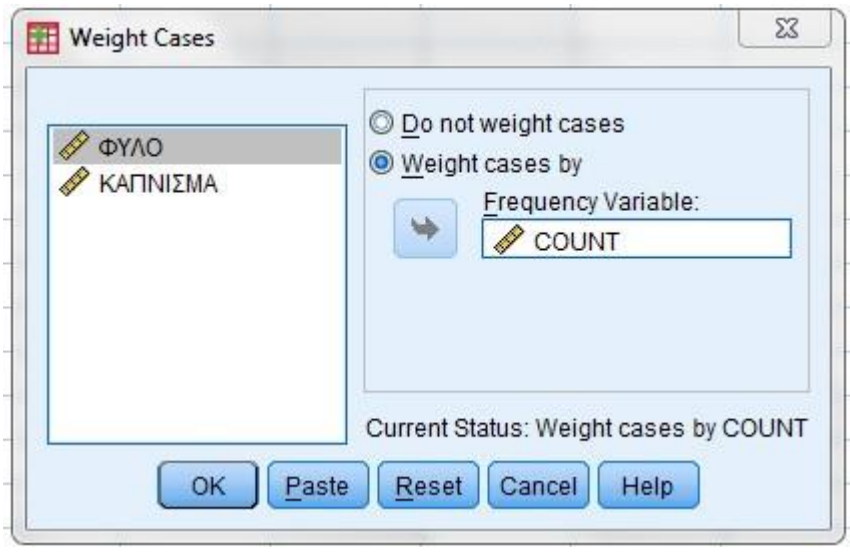
Βήμα 2ο: Δημιουργούμε την μεταβλητή **COUNT (Numeric)** στην οποία για τιμές εισάγουμε την συχνότητα εμφάνισης των συνδυασμών των τιμών, των μεταβλητών **ΦΥΛΟ και ΚΑΠΝΙΣΜΑ:**



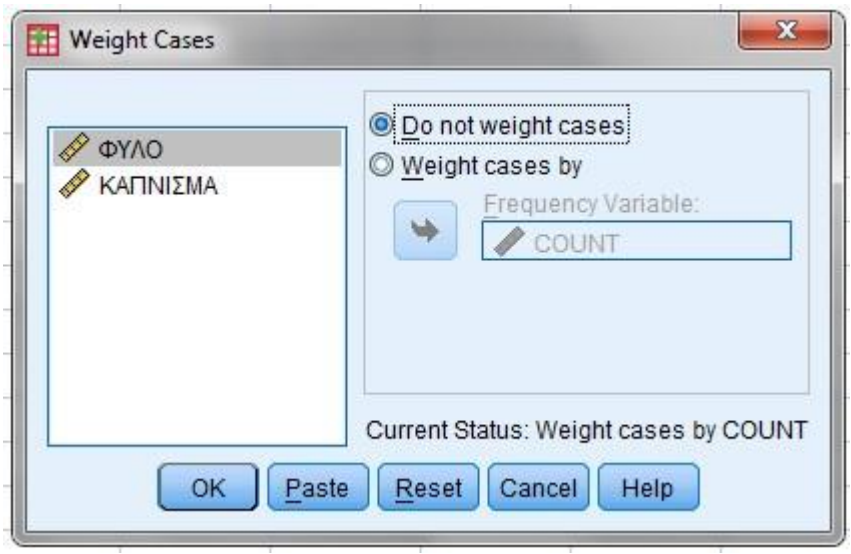


	ΦΥΛΟ	ΚΑΠΝΙΣΜΑ	COUNT	var
1	ΑΓΟΡΙΑ	ΚΑΠΝΙΣΤΕΣ	60	
2	ΑΓΟΡΙΑ	ΜΗ ΚΑΠΝΙΣΤΕΣ	50	
3	ΚΟΡΙΤΣΙΑ	ΚΑΠΝΙΣΤΕΣ	40	
4	ΚΟΡΙΤΣΙΑ	ΜΗ ΚΑΠΝΙΣΤΕΣ	50	
5				

Βήμα 3ο: Εντολή Weight Cases. Η εντολή **Weight Cases** σταθμίζει τις παρατηρήσεις ενός αρχείου με βάση τις τιμές μίας μεταβλητής συχνοτήτων (count variable). Η στάθμιση γίνεται με προσομοιωμένες επαναλήψεις των παρατηρήσεων του αρχείου, σύμφωνα με τις τιμές της μεταβλητής στάθμισης. Δηλαδή οι τιμές της μεταβλητής στάθμισης υποδεικνύουν τον αριθμό των επαναλήψεων που θα προκύψουν για κάθε παρατήρηση του αρχικού αρχείου μετά τη στάθμισή του. **Η στάθμιση γίνεται μέσω της εντολής Data - Weight Cases:**



Η στάθμιση παραμένει ενεργή μέχρι να ακυρωθεί, μέσω της επιλογής **Do not weight cases**:



Με τα τη εκτέλεση της εντολής **Weight Cases** δεν έχουν προστεθεί cases στο φύλλο των δεδομένων. Οι επαναλήψεις έχουν υλοποιηθεί εσωτερικά. Για να καταλάβουμε πλήρως πως ακριβώς δουλεύει η εντολή μπορούμε να δημιουργήσουμε τους πίνακες συχνοτήτων των μεταβλητών **ΦΥΛΟ και ΚΑΠΝΙΣΜΑ:**

ΦΥΛΟ

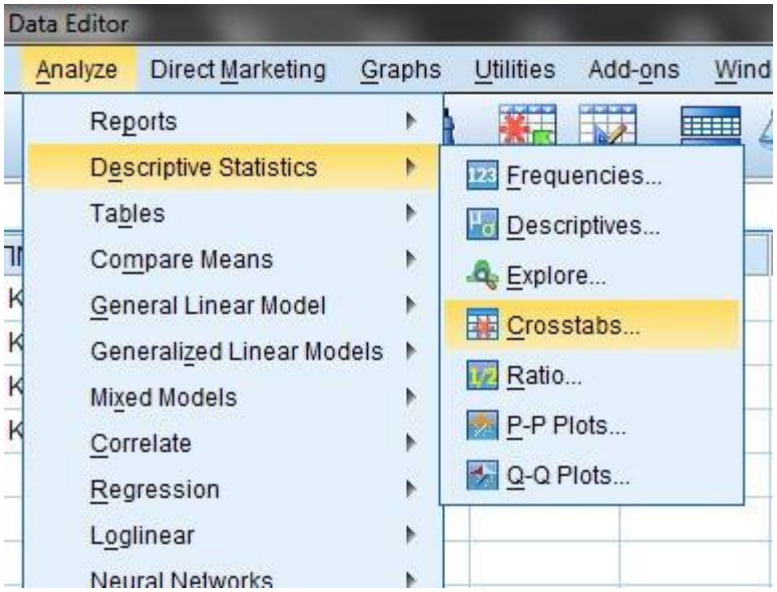
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	ΑΓΟΡΙΑ	110	55,0	55,0	55,0
	ΚΟΡΙΤΣΙΑ	90	45,0	45,0	100,0
	Total	200	100,0	100,0	

ΚΑΠΝΙΣΜΑ

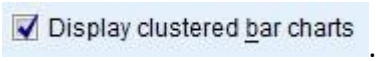
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	ΚΑΠΝΙΣΤΕΣ	100	50,0	50,0	50,0
	ΜΗ ΚΑΠΝΙΣΤΕΣ	100	50,0	50,0	100,0
	Total	200	100,0	100,0	

Από τους οποίους βλέπουμε πως ο αριθμός τους δείγματος είναι N=200, από τους οποίους τα ΑΓΟΡΙΑ είναι 110 και τα ΚΟΡΙΤΣΙΑ είναι 90. Η κατανομή ανάλογα με το Κάπνισμα είναι 100 ΚΑΠΝΙΣΤΕΣ και 100 ΜΗ ΚΑΠΝΙΣΤΕΣ.

Στην συνέχεια για την διερεύνηση της σχέσης μεταξύ των δύο μεταβλητών δημιουργούμε τον πίνακα συνάφειας (crosstabulation) μέσω της εντολής:



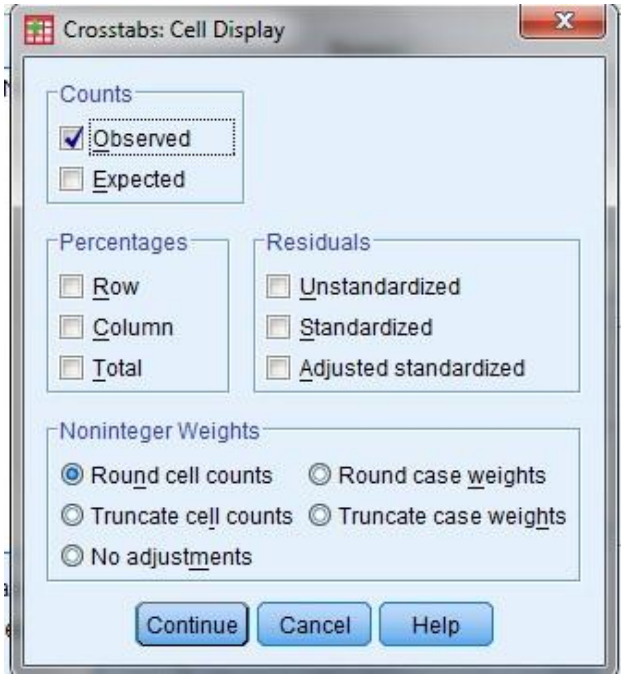
Στο παράθυρο **Row(s)** τοποθετούμε την μεταβλητή η οποία και θέλουμε να αποτελεί τις γραμμές του πίνακα μας, και αντίστοιχα στο παράθυρο **Column(s)** τοποθετούμε την μεταβλητή που θέλουμε να αποτελεί τις στήλες. Επίσης για την δημιουργία του γραφήματος των δυο μεταβλητών επιλέγουμε:



.

Επιλέγοντας 

εμφανίζεται το εξής μενού:

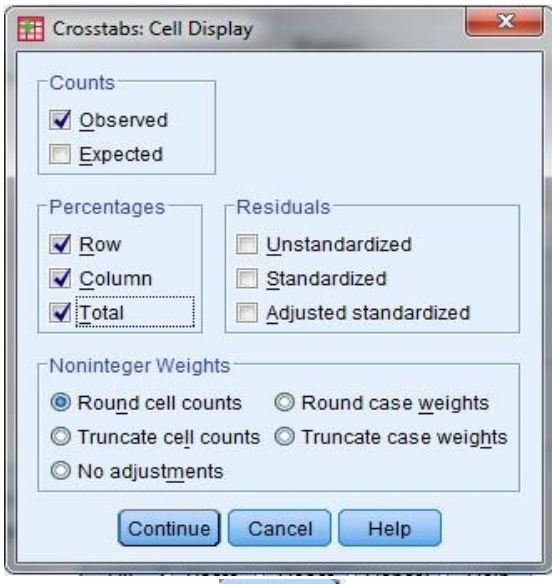


Από την ένδειξη **Percentages** μπορούμε να επιλέξουμε:

• **ROW**, αν θέλουμε να εμφανίζονται τα ποσοστά των γραμμών, • **Column**, αν θέλουμε να εμφανίζονται τα ποσοστά των στηλών, και

• **Total**, αν θέλουμε να εμφανίζονται τα συνολικά ποσοστά.

Επιλέγουμε την εμφάνιση και των τριών ποσοστών:



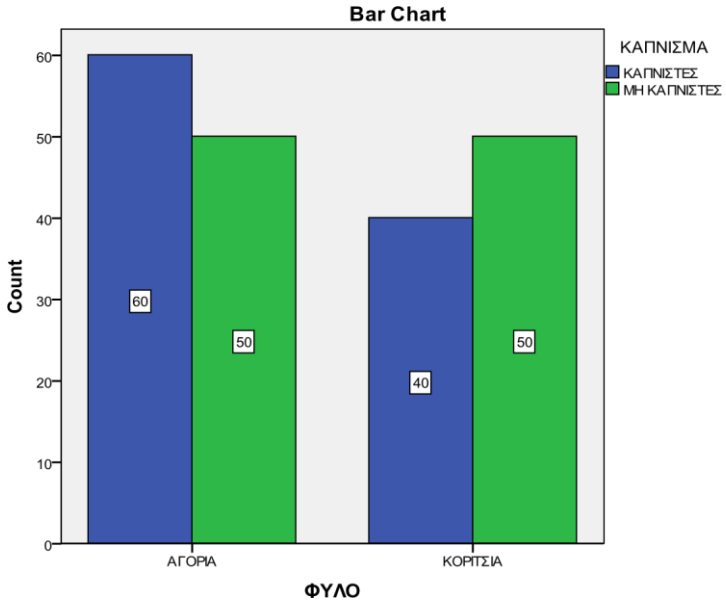
Με την επιλογή Continue επιστέφουμε στο παράθυρο της εντολής Crosstabs, και με OK δημιουργείται στο Output ο πίνακας με τα αποτελέσματα των επιλογών μας:

ΦΥΛΟ * ΚΑΠΝΙΣΜΑ Crosstabulation

			ΚΑΠΝΙΣΜΑ		
			ΚΑΠΝΙΣΤΕΣ	ΜΗ ΚΑΠΝΙΣΤΕΣ	Total
ΦΥΛΟ	ΑΓΟΡΙΑ				
		Count	60	50	110
		% within ΦΥΛΟ	54,5%	45,5%	100,0%
		% within ΚΑΠΝΙΣΜΑ	60,0%	50,0%	55,0%
		% of Total	30,0%	25,0%	55,0%
	ΚΟΡΙΤΣΙΑ				
		Count	40	50	90
		% within ΦΥΛΟ	44,4%	55,6%	100,0%
		% within ΚΑΠΝΙΣΜΑ	40,0%	50,0%	45,0%
		% of Total	20,0%	25,0%	45,0%

Total	Count	100	100	200
	% within ΦΥΛΟ	50,0%	50,0%	100,0%
	% within ΚΑΠΝΙΣΜΑ	100,0%	100,0%	100,0%
	% of Total	50,0%	50,0%	100,0%

Και το εξής γράφημα:



Ερμηνεία Row percent (% within ΦΥΛΟ):

- Από τα Αγόρια το 54,5% είναι καπνιστές και το 45,5% είναι μη καπνιστές.
- Από τα Κορίτσια το 44,4% είναι καπνιστές και το 55,6% είναι μη καπνιστές.

Ερμηνεία Column percent (% within ΚΑΠΝΙΣΜΑ):

- Από τους Καπνιστές το 60,0% είναι αγόρια και το 40,0% είναι κορίτσια.
- Από τους Μη Καπνιστές το 50,0% είναι αγόρια και το 50,0% είναι κορίτσια.

Ερμηνεία Total percent (% within total):

Από τα 200 άτομα του δείγματος:

- Το 30% είναι Αγόρια καπνιστές,
- Το 25% είναι Αγόρια Μη καπνιστές,
- Το 30% είναι Κορίτσια καπνιστές,
- Το 30% είναι Κορίτσια Μη καπνιστές.

Επισήμανση: Για την ερμηνεία των ποσοστών προσέχω αυτά να αθροίζουν στο 100%.

ΦΥΛΟ * ΚΑΠΝΙΣΜΑ Crosstabulation

		ΚΑΠΝΙΣΜΑ		
		ΚΑΠΝΙΣΤΕΣ	ΜΗ ΚΑΠΝΙΣΤΕΣ	Total
ΦΥΛΟ ΑΓΟΡΙΑ	Count	60	50	110
	% within ΦΥΛΟ	54,5%	45,5%	100,0%
	% within ΚΑΠΝΙΣΜΑ	60,0%	50,0%	55,0%
	% of Total	30,0%	25,0%	55,0%

ΚΟΡΙΤΣΙΑ	Count	40	50	90
	% within ΦΥΛΟ	44,4%	55,6%	100,0%
	% within ΚΑΠΝΙΣΜΑ	40,0%	50,0%	45,0%
	% of Total	20,0%	25,0%	45,0%
Total	Count	100	100	200
	% within ΦΥΛΟ	50,0%	50,0%	100,0%
	% within ΚΑΠΝΙΣΜΑ	100,0%	100,0%	100,0%
	% of Total	50,0%	50,0%	100,0%

Πίνακας συνάφειας και πιθανότητες:

$$P(A) = \frac{\text{ευνοϊκές περιπτώσεις}}{\text{δυνατές περιπτώσεις}}$$

Α. Οριακές πιθανότητες:

ΦΥΛΟ * ΚΑΠΝΙΣΜΑ Crosstabulation

	ΚΑΠΝΙΣΜΑ	Total

			ΚΑΠΝΙΣΤΕΣ	ΜΗ ΚΑΠΝΙΣΤΕΣ	
ΦΥΛΟ	ΑΓΟΡΙΑ	Count	60	50	110
		% within ΦΥΛΟ	54,5%	45,5%	100,0%
		% within ΚΑΠΝΙΣΜΑ	60,0%	50,0%	55,0%
		% of Total	30,0%	25,0%	55,0%
	ΚΟΡΙΤΣΙΑ	Count	40	50	90
		% within ΦΥΛΟ	44,4%	55,6%	100,0%
		% within ΚΑΠΝΙΣΜΑ	40,0%	50,0%	45,0%
		% of Total	20,0%	25,0%	45,0%
Total		Count	100	100	200
		% within ΦΥΛΟ	50,0%	50,0%	100,0%
		% within ΚΑΠΝΙΣΜΑ	100,0%	100,0%	100,0%
		% of Total	50,0%	50,0%	100,0%

1. Ποια η πιθανότητα αν επιλέξω τυχαία ένα άτομο από το δείγμα, αυτό να είναι αγόρι (ανεξάρτητα αν είναι καπνιστής ή όχι);

P(Αγόρι)= $\frac{110}{200} = 0,55$ ή 55% = πιθανότητα % of Total στο Total της γραμμής ΑΓΟΡΙΑ

2. Ποια η πιθανότητα αν επιλέξω τυχαία ένα άτομο από το δείγμα, αυτό να είναι κορίτσι (ανεξάρτητα αν είναι καπνιστής ή όχι);

P(Κορίτσι)= $\frac{90}{200} = 0,45$ ή 45% = πιθανότητα % of Total στο Total της γραμμής ΚΟΡΙΤΣΙΑ

3. Ποια η πιθανότητα αν επιλέξω τυχαία ένα άτομο από το δείγμα, αυτό να είναι Καπνιστής (ανεξάρτητα αν είναι Αγόρι ή Κορίτσι);

P(Καπνιστής)= $\frac{100}{200} = 0,50$ ή 50% = πιθανότητα % of Total στο τέλος της στήλης ΚΑΠΝΙΣΤΕΣ

4. Ποια η πιθανότητα αν επιλέξω τυχαία ένα άτομο από το δείγμα, αυτό να είναι Μη Καπνιστής (ανεξάρτητα αν είναι Αγόρι ή Κορίτσι);

P(Καπνιστής)= $\frac{100}{200} = 0,50$ ή 50% = πιθανότητα % of Total στο τέλος της στήλης ΜΗ ΚΑΠΝΙΣΤΕΣ

Β. Από κοινού πιθανότητες:

Στον ίδια πίνακα είναι οι πιθανότητες που βρίσκονται στο κύριο σώμα του:

ΦΥΛΟ * ΚΑΠΝΙΣΜΑ Crosstabulation

			ΚΑΠΝΙΣΜΑ		
			ΚΑΠΝΙΣΤΕΣ	ΜΗ ΚΑΠΝΙΣΤΕΣ	Total
ΦΥΛΟ	ΑΓΟΡΙΑ	Count	60	50	110
		% within ΦΥΛΟ	54,5%	45,5%	100,0%

		% within ΚΑΠΝΙΣΜΑ	60,0%	50,0%	55,0%
		% of Total	30,0%	25,0%	55,0%
	ΚΟΡΙΤΣΙΑ	Count	40	50	90
		% within ΦΥΛΟ	44,4%	55,6%	100,0%
		% within ΚΑΠΝΙΣΜΑ	40,0%	50,0%	45,0%
		% of Total	20,0%	25,0%	45,0%
Total		Count	100	100	200
		% within ΦΥΛΟ	50,0%	50,0%	100,0%
		% within ΚΑΠΝΙΣΜΑ	100,0%	100,0%	100,0%
		% of Total	50,0%	50,0%	100,0%

1. Ποια η πιθανότητα αν επιλέξω τυχαία ένα άτομο από το δείγμα, αυτό να είναι αγόρι **και** καπνιστής;

Ρ(Αγόρι∩ Καπνιστής)= $\frac{60}{200} = 0,30$ ή 30% = πιθανότητα % of Total μέσα στο συγκεκριμένο κελί, δηλαδή στο κελί ΑΓΟΡΙΑ ΚΑΠΝΙΣΤΕΣ.

2. Ποια η πιθανότητα αν επιλέξω τυχαία ένα άτομο από το δείγμα, αυτό να είναι αγόρι **και** Μη καπνιστής;

Ρ(Αγόρι∩Μη Καπνιστής)= $\frac{50}{200} = 0,25$ ή 25% = πιθανότητα % of Total μέσα στο συγκεκριμένο κελί, δηλαδή στο κελί ΑΓΟΡΙΑ ΜΗ ΚΑΠΝΙΣΤΕΣ.

3. Ποια η πιθανότητα αν επιλέξω τυχαία ένα άτομο από το δείγμα, αυτό να είναι κορίτσι **και** καπνιστής;

Ρ(ΚορίτσιΚαπνιστής)= $\frac{40}{200} = 0,20$ ή 20% = πιθανότητα % of Total μέσα στο συγκεκριμένο κελί, δηλαδή στο κελί ΚΟΡΙΤΣΙΑ ΚΑΠΝΙΣΤΕΣ.

4. Ποια η πιθανότητα αν επιλέξω τυχαία ένα άτομο από το δείγμα, αυτό να είναι κορίτσι **και** Μη καπνιστής;

Ρ(ΚορίτσιΜη Καπνιστής)= $\frac{50}{200} = 0,25$ ή 25% = πιθανότητα % of Total μέσα στο συγκεκριμένο κελί, δηλαδή στο κελί ΚΟΡΙΤΣΙΑ ΜΗ ΚΑΠΝΙΣΤΕΣ.

Γ. Δεσμευμένες πιθανότητες:

Στον ίδια πίνακα:

ΦΥΛΟ * ΚΑΠΝΙΣΜΑ Crosstabulation

			ΚΑΠΝΙΣΜΑ		
			ΚΑΠΝΙΣΤΕΣ	ΜΗ ΚΑΠΝΙΣΤΕΣ	Total
ΦΥΛΟ	ΑΓΟΡΙΑ	Count	60	50	110
		% within ΦΥΛΟ	54,5%	45,5%	100,0%
		% within ΚΑΠΝΙΣΜΑ	60,0%	50,0%	55,0%
		% of Total	30,0%	25,0%	55,0%

	ΚΟΡΙΤΣΙΑ			
		Count	40	50	90
		% within ΦΥΛΟ	44,4%	55,6%	100,0%
		% within ΚΑΠΝΙΣΜΑ	40,0%	50,0%	45,0%
		% of Total	20,0%	25,0%	45,0%
Total		Count	100	100	200
		% within ΦΥΛΟ	50,0%	50,0%	100,0%
		% within ΚΑΠΝΙΣΜΑ	100,0%	100,0%	100,0%
		% of Total	50,0%	50,0%	100,0%

1. Ποια η πιθανότητα ένας φοιτητής που είναι αγόρι (δηλαδή δεδομένου ότι είναι αγόρι) να είναι Καπνιστής;

P(Καπνιστής|Αγόρι)= $\frac{60}{110} = 0{,}545$ ή 54,5% = πιθανότητα % within ΦΥΛΟ μέσα στο κελί ΑΓΟΡΙΑ ΚΑΠΝΙΣΤΕΣ.

2. Ποια η πιθανότητα ένας φοιτητής που είναι αγόρι να είναι Μη Καπνιστής;

P(Μη Καπνιστής|Αγόρι)= $\frac{50}{110} = 0{,}455$ ή 45,5% = πιθανότητα % within ΦΥΛΟ μέσα στο κελί ΑΓΟΡΙΑ ΜΗ ΚΑΠΝΙΣΤΕΣ.

3. Ποια η πιθανότητα ένας φοιτητής που είναι Κορίτσι να είναι Καπνιστής;

P(Καπνιστής|Κορίτσι)= $\frac{40}{90} = 0,444$ ή 44,4% = πιθανότητα % within ΦΥΛΟ μέσα στο κελί ΚΟΡΙΤΣΙΑ ΚΑΠΝΙΣΤΕΣ.

4. Ποια η πιθανότητα ένας φοιτητής που είναι Κορίτσι να είναι Μη Καπνιστής;

P(Μη Καπνιστής|Κορίτσι)= $\frac{50}{90} = 0,556$ ή 55,6% = πιθανότητα % within ΦΥΛΟ μέσα στο κελί ΚΟΡΙΤΣΙΑ ΜΗ ΚΑΠΝΙΣΤΕΣ.

5. Ποια η πιθανότητα ένας φοιτητής που είναι καπνιστής (δηλαδή δεδομένου ότι είναι καπνιστής) να είναι Αγόρι;

P(Αγόρι|Καπνιστής)= $\frac{60}{100} = 0,60$ ή 60,0% = πιθανότητα % within ΚΑΠΝΙΣΜΑ μέσα στο κελί ΑΓΟΡΙΑ ΚΑΠΝΙΣΤΕΣ.

6. Ποια η πιθανότητα ένας φοιτητής που είναι καπνιστής (δηλαδή δεδομένου ότι είναι καπνιστής) να είναι Κορίτσι;

P(Κορίτσι|Καπνιστής)= $\frac{40}{100} = 0,40$ ή 40% = πιθανότητα % within ΚΑΠΝΙΣΜΑ μέσα στο κελί ΚΟΡΙΤΣΙΑ ΚΑΠΝΙΣΤΕΣ.

7. Ποια η πιθανότητα ένας φοιτητής που είναι Μη καπνιστής (δηλαδή δεδομένου ότι είναι Μη καπνιστής) να είναι Αγόρι;

P(Αγόρι|Μη Καπνιστής)= $\frac{50}{100} = 0,50$ ή 50,0% = πιθανότητα % within ΚΑΠΝΙΣΜΑ μέσα στο κελί ΑΓΟΡΙΑ ΜΗ ΚΑΠΝΙΣΤΕΣ.

8. Ποια η πιθανότητα ένας φοιτητής που είναι Μη καπνιστής (δηλαδή δεδομένου ότι είναι Μη καπνιστής) να είναι Κορίτσι;

P(Κορίτσι|Μη Καπνιστής)= $\frac{50}{100} = 0,50$ ή 50,0% = πιθανότητα % within ΚΑΠΝΙΣΜΑ μέσα στο κελί ΚΟΡΙΤΣΙΑ ΜΗ ΚΑΠΝΙΣΤΕΣ.

Θα ασχοληθούμε με την διαδικασία εύρεσης πιθανοτήτων στο SPSS. Συγκεκριμένα, θα έχουμε ένα πείραμα το οποίο θα εκτελείται ν φορές και στο οποίο ορίζουμε εμείς ένα αποτέλεσμα σαν επιτυχία και ένα άλλο σαν αποτυχία. Θα ζητάμε πιθανότητες της μορφής : $P(X \leq a)$, $P(X \geq a)$, $P(a \leq X \leq \beta)$ όπου Χ είναι μια μεταβλητή που ακολουθεί μια τυχαία κατανομή και το ενδεχόμενο του οποίου θα ψάχνουμε την πιθανότητα θα αντιστοιχεί στο πλήθος των επιτυχιών που θα έχουμε στις ν επαναλήψεις του πειράματος. Οι κατανομές με τις οποίες θα ασχοληθούμε είναι η Διωνυμική (Binomial), η Poisson, η Κανονική (Normal) και η Ομοιόμορφη (Uniform). Οι δύο πρώτες κατανομές αφορούν διακριτές μεταβλητές ενώ οι άλλες δύο αφορούν συνεχείς μεταβλητές.

Για να βρούμε αυτές τις πιθανότητες στο SPSS υπάρχουν κάποιες ιδιαιτερότητες :

i) Θέλουμε να χρησιμοποιήσουμε το SPSS σαν «κομπιουτεράκι» αλλά αυτό δεν γίνεται γιατί το SPSS χρειάζεται μία μεταβλητή και κάποια δεδομένα. Για τον λόγο αυτό πάμε στο Data View και στο πρώτο κελί γράφουμε έναν τυχαίο αριθμό. Αμέσως το SPSS δημιουργεί μία δική του μεταβλητή με το δεδομένο που του δώσαμε και αυτό μας επιτρέπει να προχωρήσουμε στην διαδικασία εύρεσης μιας πιθανότητας. ii) Το SPSS μπορεί να υπολογίσει μόνο πιθανότητες της μορφής $P(X \leq a)$, γι αυτό αν θέλουμε να βρούμε πιθανότητες άλλης μορφής, χρησιμοποιούμε τις ιδιότητες των πιθανοτήτων που είναι γνωστές από τη θεωρία.

Διαδικασία στο SPSS

Επιλέγουμε Transform → Compute

Στο Target variable γράφω την λέξη probability (πιθανότητα).

Στο Function group κλικάρουμε το CDF and Noncentral CDF και στο Function and special variables κλικάρουμε την κατανομή που θέλουμε δηλαδή CDF BINOM ή CDF POISSON ή CDF NORMAL και την μεταφέρουμε στο πεδίο Numeric Expression όπου

έχουμε την μορφή CDFBINOM(?,?,?) ή CDFPOISSON(?,?) ή CDFNORMAL(?,?,?). Στην πρώτη κατανομή στην θέση των ερωτηματικών γράφουμε (quant,n,p), στην δεύτερη (quant,mean) και στην τρίτη (quant,mean,stddeviation), δηλαδή τους αριθμούς που μας δίνει η άσκηση και αφορούν τα παρακάτω :

quant : η πιθανότητα που θέλουμε να βρούμε (π.χ. στην $P(X \leq a)$, τον αριθμό α)

n : αριθμός επαναλήψεων πειράματος p : πιθανότητα να έχουμε επιτυχία σε

κάθε επανάληψη mean : αριθμητικός μέσος stddeviation : τυπική απόκλιση

Πατάμε OK.

Αποτελέσματα

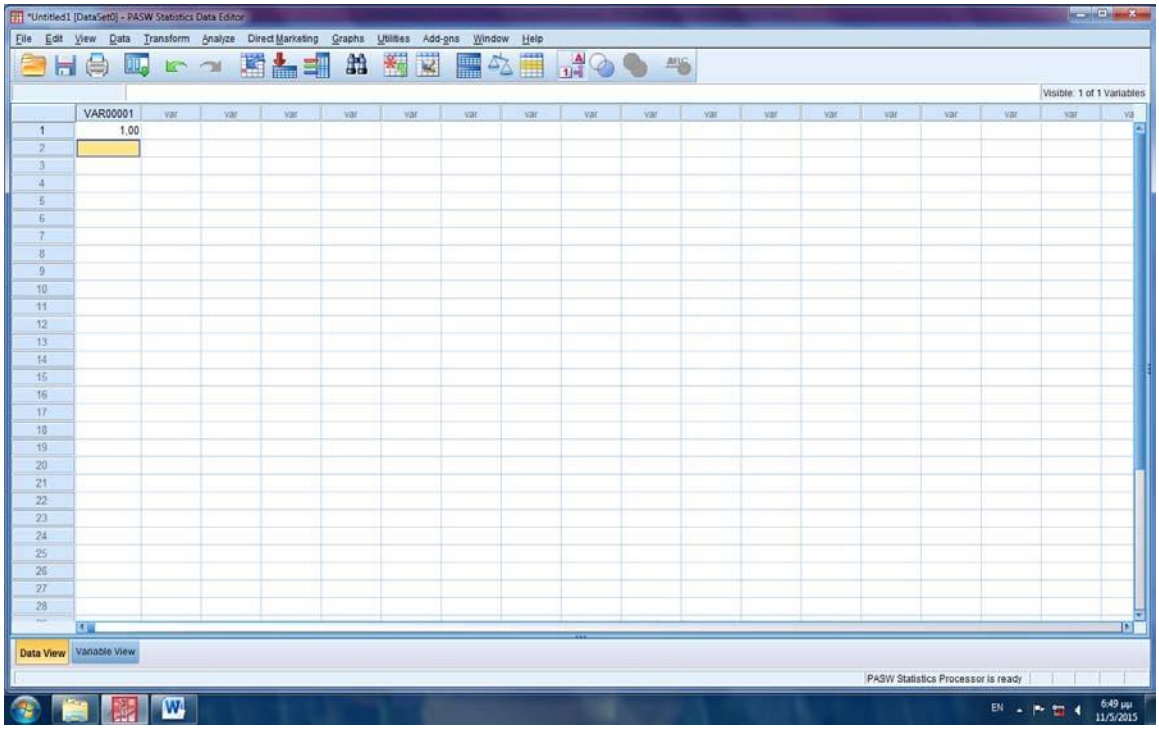
Στο Data View έχει προστεθεί μία στήλη με το όνομα probability. Ο αριθμός που βλέπουμε στην στήλη αυτή είναι η πιθανότητα που ζητήσαμε από το SPSS να μας βρει.

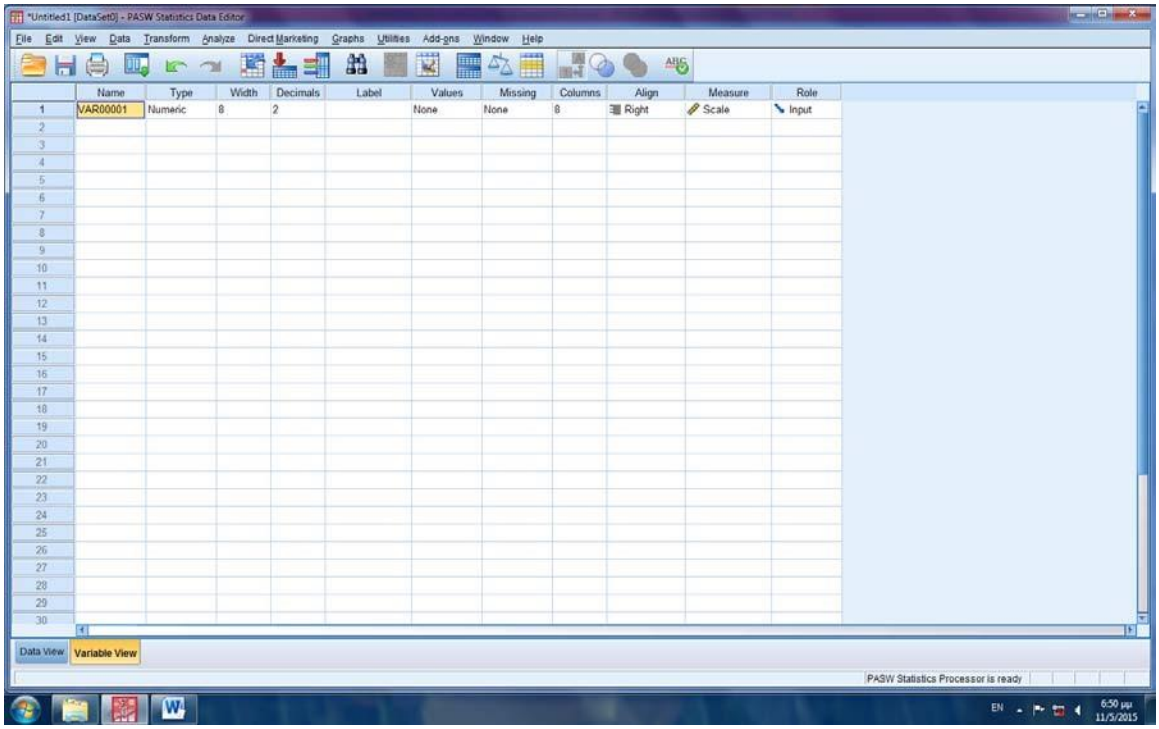
ΑΣΚΗΣΗ

Έστω μια τυχαία μεταβλητή X η οποία ακολουθεί τη διωνυμική κατανομή με πλήθος επαναλήψεων του πειράματος η=8 και πιθανότητα επιτυχίας σε κάθε επανάληψη p=0,33. Να υπολογιστούν οι πιθανότητες $P(X \leq 5)$ και $P(X \geq 6)$.

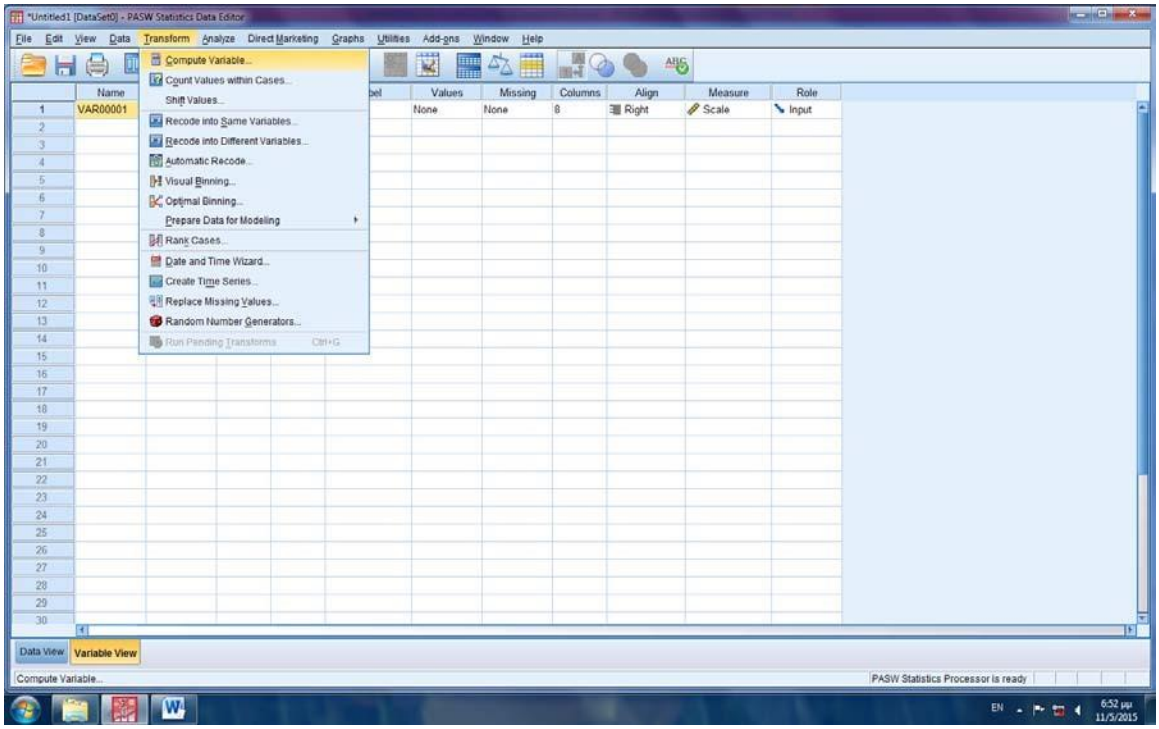
ΛΥΣΗ

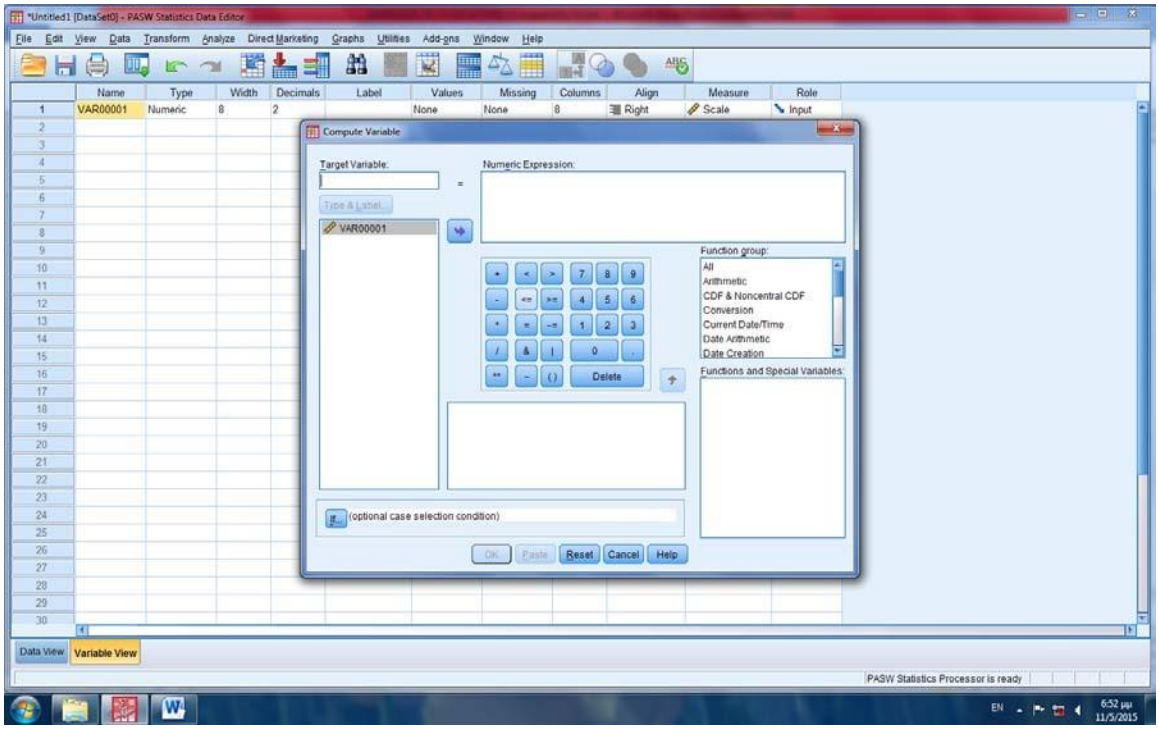
Πάμε στο Data View και στο πρώτο κελί γράφουμε έναν τυχαίο αριθμό. Αμέσως το SPSS δημιουργεί μία δική του μεταβλητή με το δεδομένο που του δώσαμε.





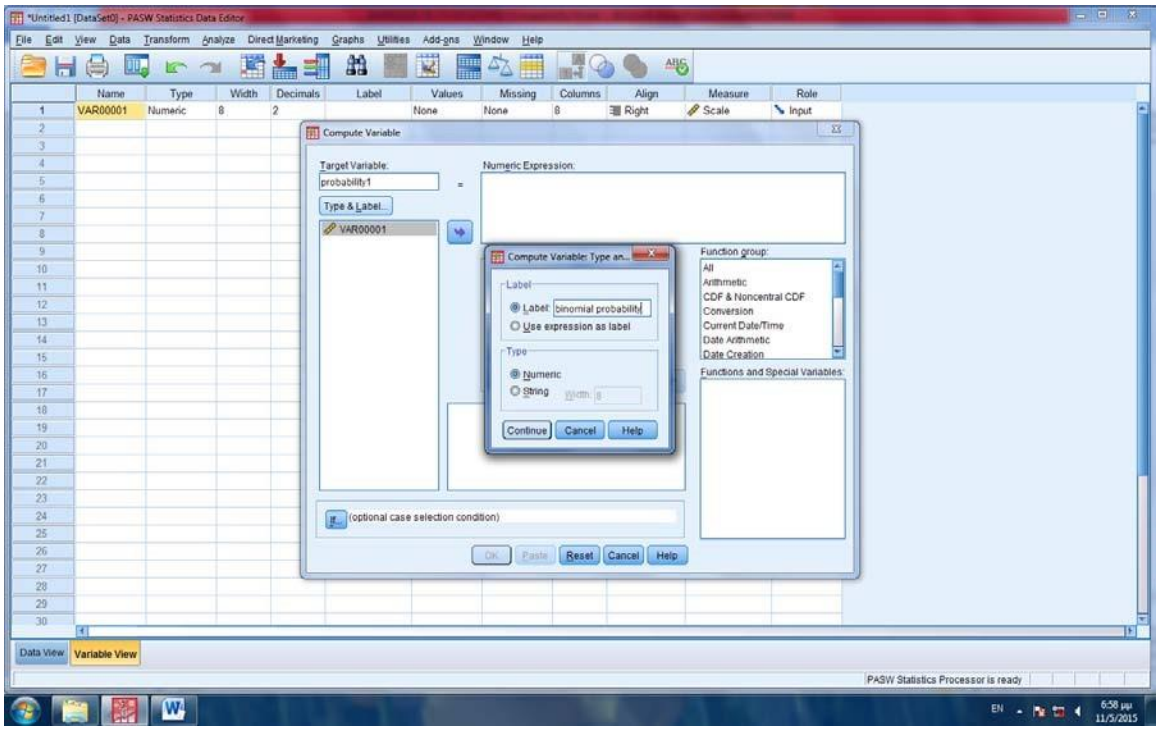
Στη συνέχεια επιλέγουμε Transform → Compute Variable



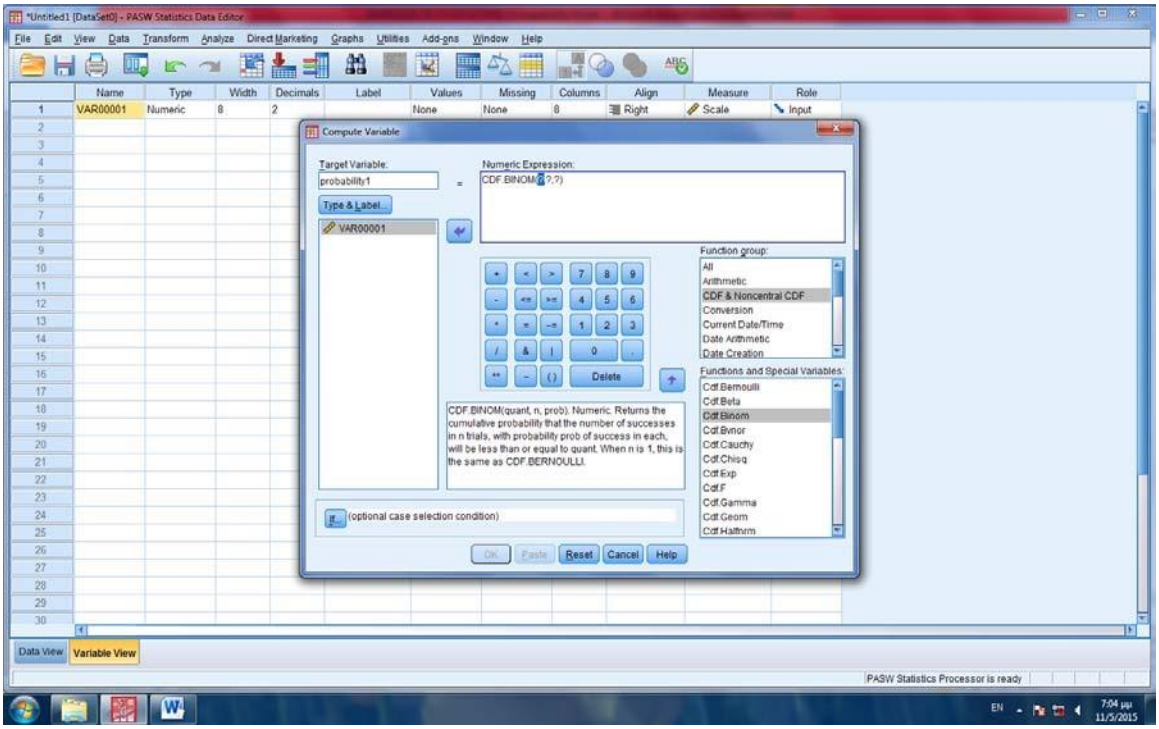


Στο παράθυρο που ανοίγει στο πλαίσιο Target Variable γράφουμε το όνομα της πιθανότητας που θέλουμε να υπολογίσουμε, για παράδειγμα probability1 και στο Type & Label επιλέγουμε Numeric και γράφουμε στο Label binomial probability ή ότι άλλο θέλουμε σαν επεξήγηση της πιυθανότητας που ζητάμε.

Πατάμε Continue.



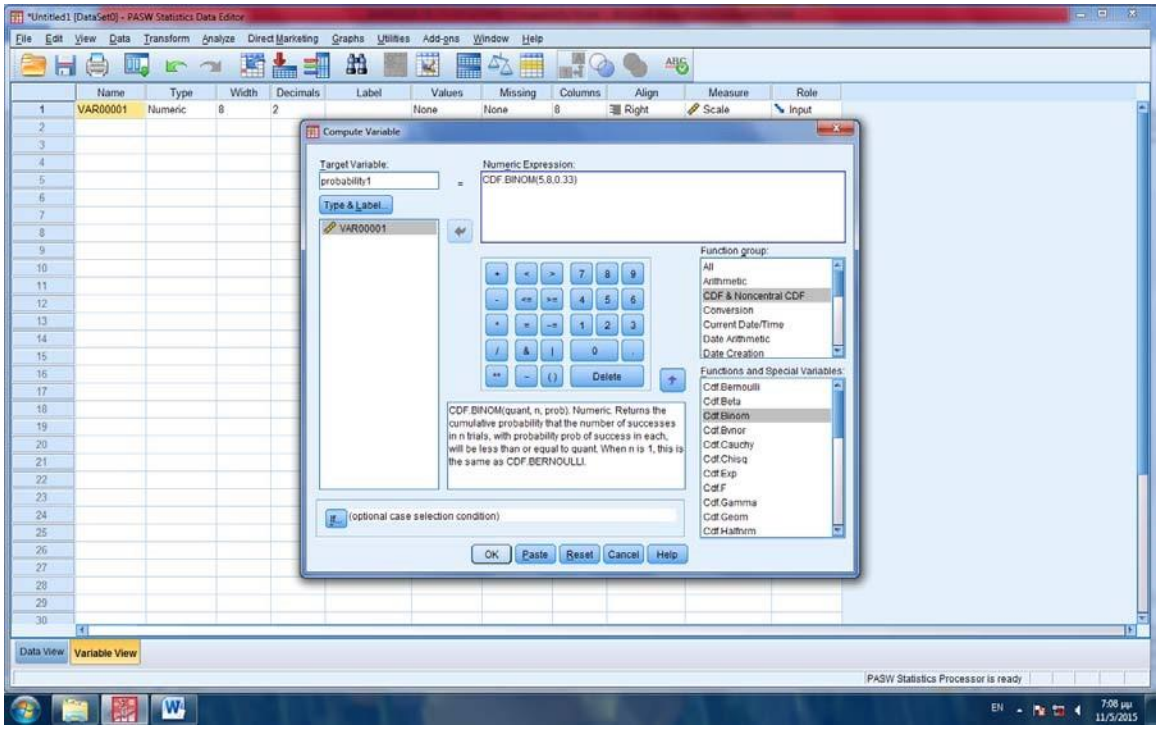
Από τη λίστα Function group επιλέγουμε CDF & Noncentral CDF και από τη λίστα Functions and Special Variables επιλέγουμε CDF Binom και τη μεταφέρουμε με το βελάκι στο Numeric Expression.

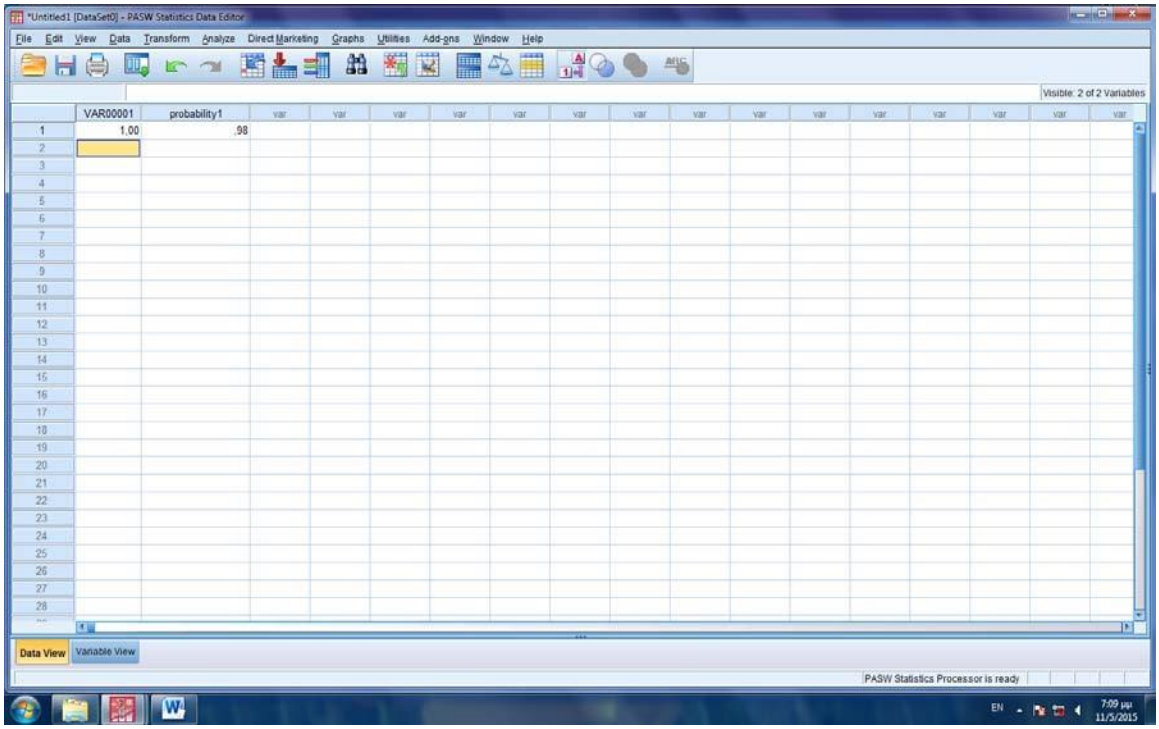


Στη θέση του πρώτου ερωτηματικού γράφουμε την τιμή 5 γιατί θέλουμε την πιθανότητα $P(X \leq 5)$, στη θέση του δεύτερου ερωτηματικού γράφουμε την τιμή του η που είναι 8 και στη θέση του τρίτου ερωτηματικού γράφουμε την τιμή του p που είναι 0.33.

Πατάμε ΟΚ.

Βλέπουμε στο Data View ότι έχει βρει την πιθανότητα η οποία είναι $P(X \leq 5)$ =0,98.





Για την άλλη πιθανότητα χρησιμοποιούμε τις ιδιότητες των πιθανοτήτων και βρίσκουμε :

$$P(X \geq 6) = 1 - P(X \prec 6) = 1 - P(X \leq 5) = 1 - 0{,}98 = 0{,}02$$