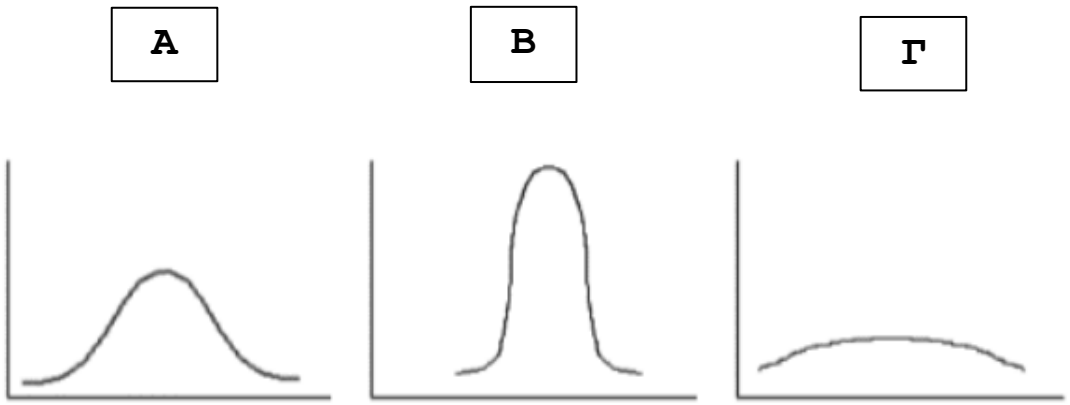
4ο ΕΡΓΑΣΤΗΡΙΟ

Σε αυτό το εργαστήριο θα εξηγήσουμε **την έννοια της κατανομής των τιμών** μιας ποσοτικής μεταβλητής, θα παρουσιάσουμε διαφορετικές μορφές (σχήματα) κατανομών και θα εξηγήσουμε τι είναι **η κανονική κατανομή.** Θα δούμε, επίσης, πώς ελέγχουμε τη μορφή μιας κατανομής χρησιμοποιώντας το πρόγραμμα SPSS και πώς μπορούμε να υπολογίσουμε τη διασπορά των τιμών μιας κανονικής κατανομής γύρω από το μέσο όρο εάν γνωρίζουμε την τυπική της απόκλιση. Τέλος, θα περιγράψουμε τη διαδικασία διερεύνησης της σχέσης (συνάφειας) μεταξύ δύο μεταβλητών στο δείγμα μιας έρευνας.

1. Μορφές κατανομών - Κανονική κατανομή

Κατανομή τιμών ονομάζουμε **το σχήμα της παράστασης** που προκύπτει όταν όλες οι τιμές μιας ποσοτικής μεταβλητής τοποθετηθούν σε μια γραφική παράσταση, όπου ο άξονας των x δείχνει τις τιμές της μεταβλητής και ο άξονας των y τις συχνότητες με τις οποίες εμφανίζεται η καθεμία. Υπάρχουν πολλές (θεωρητικά άπειρες) μορφές που μπορεί να έχει μια κατανομή τιμών. Εδώ θα επικεντρωθούμε σε μια συγκεκριμένη μορφή κατανομής, την κωδωνοειδή καμπύλη. Παρακάτω εμφανίζονται μερικοί τύποι κωδωνοειδών κατανομών.



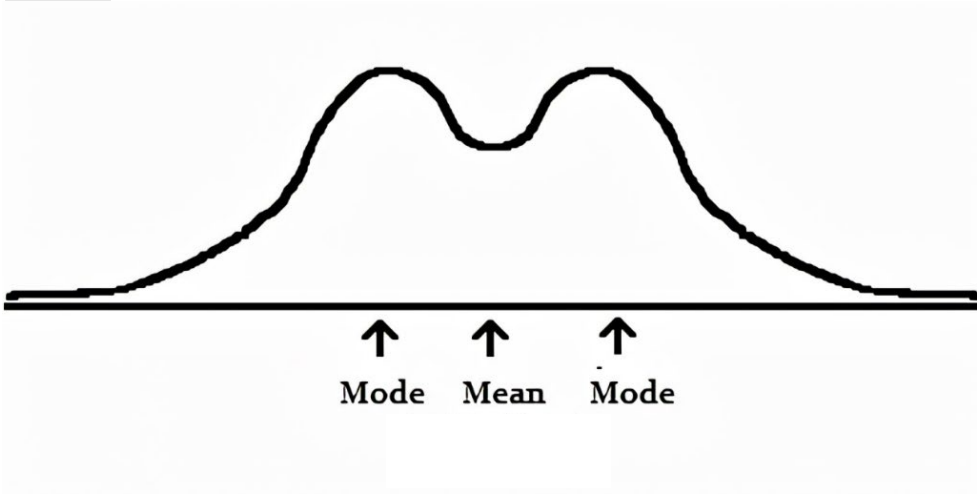
Οι παραπάνω καμπύλες είναι όλες συμμετρικές, καθώς, αν τις διπλώναμε στη μέση, τα δύο μέρη τους θα συνέπιπταν. Η καμπύλη Β χαρακτηρίζεται από τη συγκέντρωση των τιμών στο κέντρο της κατανομής και ονομάζεται **λεπτόκυρτη κατανομή**. Η καμπύλη Γ χαρακτηρίζεται από μικρό βαθμό συγκέντρωσης των τιμών γύρω από το κέντρο της κατανομής και ονομάζεται **πλατύκυρτη κατανομή**. Τέλος, η καμπύλη Α έχει την ιδανική μορφή της λεγόμενης κανονικής κατανομής και ονομάζεται **μεσόκυρτη κατανομή**[1]. Η **κανονική κατανομή** είναι η πιο σημαντική από όλες, καθώς

[1]Η κύρτωση χαρακτηρίζει την αιχμηρότητα της καμπύλης μιας κατανομής. Η λεπτόκυρτη κατανομή έχει μεγαλύτερη κύρτωση (μεγαλύτερη αιχμηρότητα καμπύλης) από την κανονική κατανομή, ενώ η πλατύκυρτη κατανομή έχει μικρότερη κύρτωση (μικρότερη αιχμηρότητα καμπύλης) από την κανονική κατανομή.

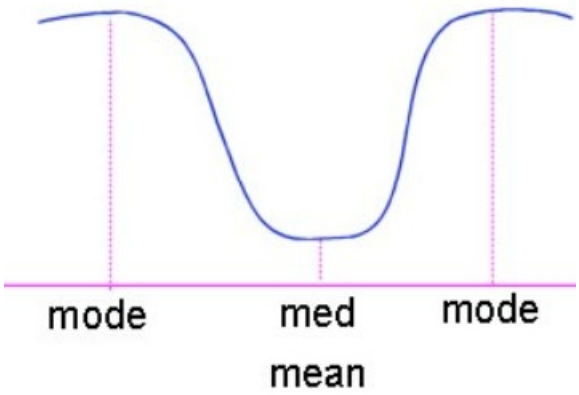
πολλές από τις τεχνικές που εφαρμόζονται στο πλαίσιο της επαγωγικής στατιστικής προϋποθέτουν ότι οι τιμές των ποσοτικών μεταβλητών σχηματίζουν κανονική κατανομή.

Αξίζει να σημειώσουμε ότι δεν είναι μόνο οι κωδωνοειδείς κατανομές συμμετρικές. Υπάρχουν και συμμετρικές μη κωδωνοειδείς κατανομές, όπως οι παρακάτω:

Δ



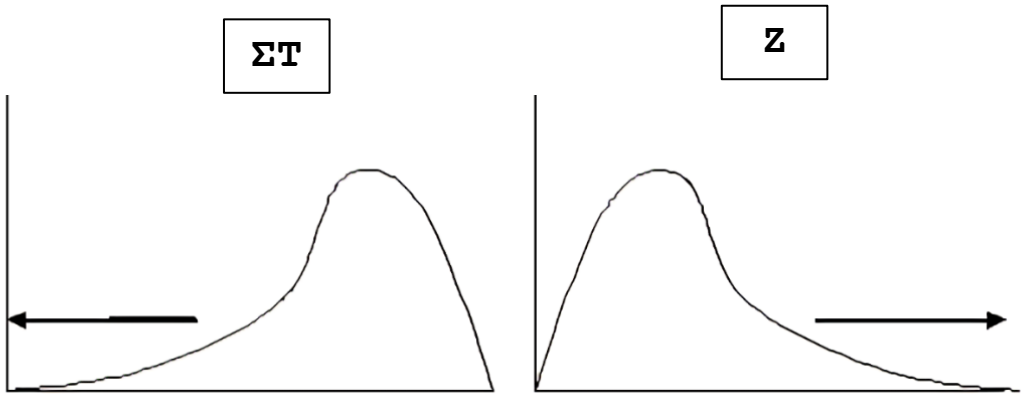
Ε



Οι κατανομές Δ και Ε ονομάζονται **δικόρυφες συμμετρικές** (bimodal symmetrical), καθώς έχουν δύο δεσπόζουσες τιμές (modes) κι αν τις διπλώσουμε στη μέση το ένα μέρος θα πέσει ακριβώς πάνω στο άλλο. Μάλιστα, η κατανομή Ε είναι γνωστή και ως **κατανομή U** (U-distribution).

Όταν μια κατανομή δεν είναι συμμετρική ονομάζεται ασύμμετρη ή λοξή (skewed). Όταν χαρακτηρίζουμε μια κατανομή ως **θετικά ασύμμετρη (positively skewed)** εννοούμε ότι παρουσιάζει επιμήκυνση (ουρά) προς το δεξιό άκρο του οριζόντιου

άξονα (που αντιστοιχεί στις μεγαλύτερες τιμές της κλίμακας μέτρησης). Το αντίθετο συμβαίνει με μια **αρνητικά ασύμμετρη κατανομή (negatively skewed)** η οποία παρουσιάζει επιμήκυνση (ουρά) προς το αριστερό άκρο του οριζόντιου άξονα (που αντιστοιχεί στις χαμηλότερες τιμές της κλίμακας μέτρησης). Οι κατανομές ΣΤ και Ζ που παρουσιάζονται παρακάτω είναι παραδείγματα αρνητικά ασύμμετρης και θετικά ασύμμετρης κατανομής, αντίστοιχα.



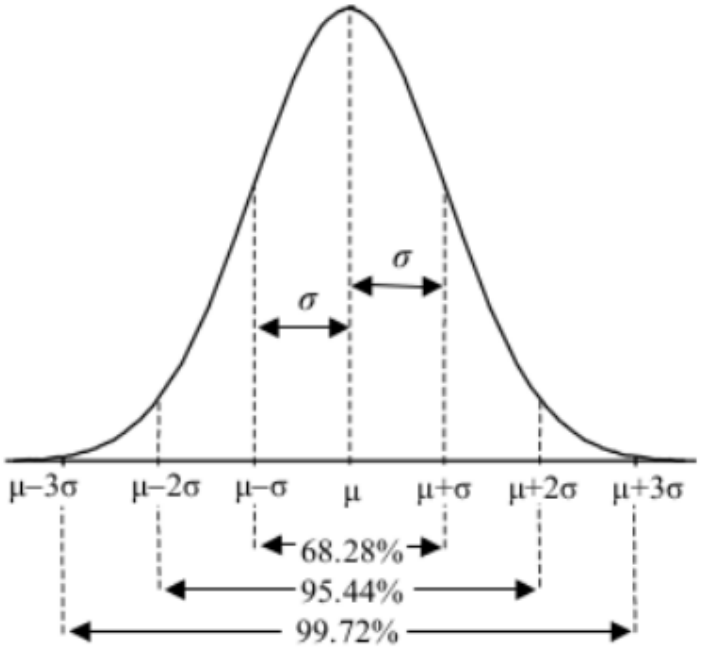
Με βάση τα παραπάνω, καταλαβαίνουμε ότι η μορφή μιας κατανομής καθορίζεται από τη συμμετρία ή την ασυμμετρία της (επομένως, από τη φορά της), από την κύρτωση της (διασπορά τιμών) και από τον αριθμό των δεσποζουσών τιμών της (modes).

1.1 Η κανονική κατανομή

Κανονική κατανομή είναι η μεσόκυρτη κωδωνοειδής κατανομή, γνωστή και ως **καμπύλη του Gauss**. Η κανονική κατανομή έχει τα ακόλουθα χαρακτηριστικά:

- Ο μέσος όρος, η διάμεσος και η δεσπόζουσα συμπίπτουν.
- Εκτείνεται συμμετρικά σε κάθε πλευρά του μέσου όρου.
- Έχει ρέουσα κλίση, της οποίας το πιο απότομο σημείο βρίσκεται σε απόσταση μιας τυπικής απόκλισης εκατέρωθεν του μέσου όρου.
- Σε απόσταση τριών τυπικών αποκλίσεων από το μέσο όρο η κλίση είναι σχεδόν οριζόντια, πολύ κοντά στο μηδέν.
- Το ποσοστό της περιοχής που περιλαμβάνεται μεταξύ του μέσου όρου και:
  - ± 1 τυπική απόκλιση είναι περίπου 68%
  - ± 2 τυπικές αποκλίσεις είναι περίπου 95%
  - ± 3 τυπικές αποκλίσεις είναι περίπου 99,7%

Στο παρακάτω σχήμα παρουσιάζονται τα βασικά χαρακτηριστικά της κανονικής κατανομής.



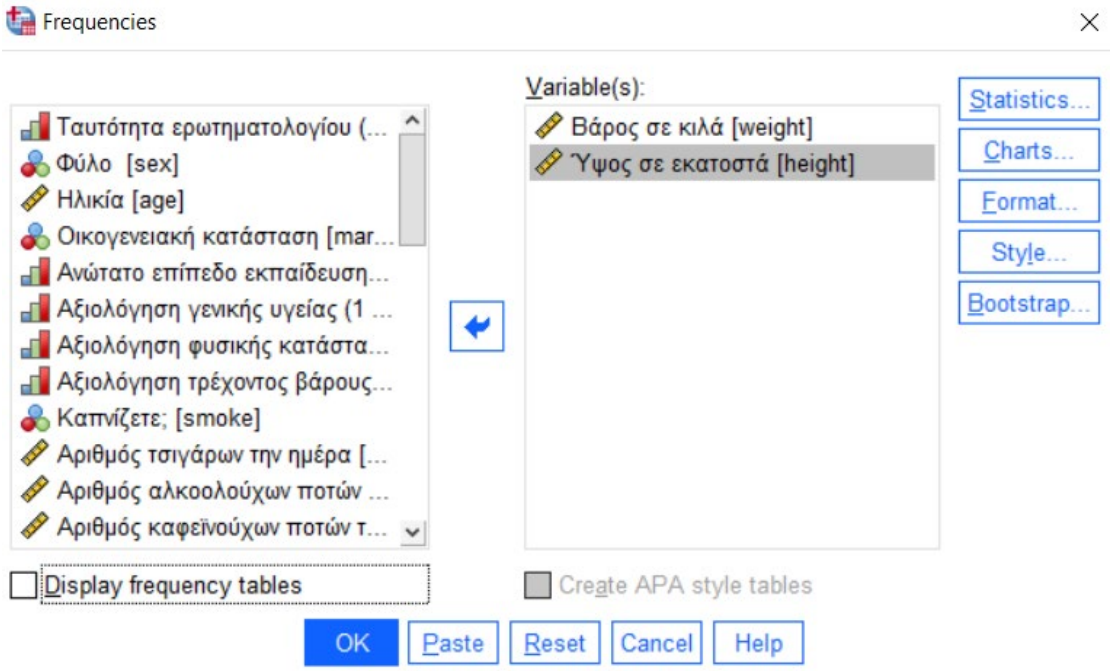
Στις έρευνες, οι κατανομές των τιμών δεν είναι απόλυτα «κανονικές». Παρόλα αυτά, γίνεται αποδεκτό ότι ακόμα κι αν μια κατανομή απομακρύνεται από το ιδανικό σχήμα της κανονικής κατανομής, αυτό δεν δημιουργεί απαραίτητα πρόβλημα στην ερμηνεία των δεδομένων αρκεί η απόκλιση αυτή να μην είναι πάρα πολύ μεγάλη. Έτσι δεν χρειάζεται να ανησυχούμε αν π.χ. η δεσπόζουσα, η διάμεσος ή ο μέσος όρος διαφέρουν κάπως μεταξύ τους, ειδικά αν το πλήθος των τιμών είναι μεγαλύτερο του 30. Βεβαίως πρέπει να τονίσουμε ότι η κρίση αυτή είναι σε μεγάλο βαθμό υποκειμενική, καθώς δεν υπάρχουν τρόποι αξιολόγησης του ανεκτού βαθμού απόκλισης από το «κανονικό» και επιπλέον, αυτό εξαρτάται από τον κλάδο της έρευνας (π.χ. παιδαγωγικά, διοίκηση, ιατρική, χρηματοοικονομικά, πολιτική επιστήμη, κ.λπ.).

1.2. Εύρεση της μορφής της κατανομής στο SPSS

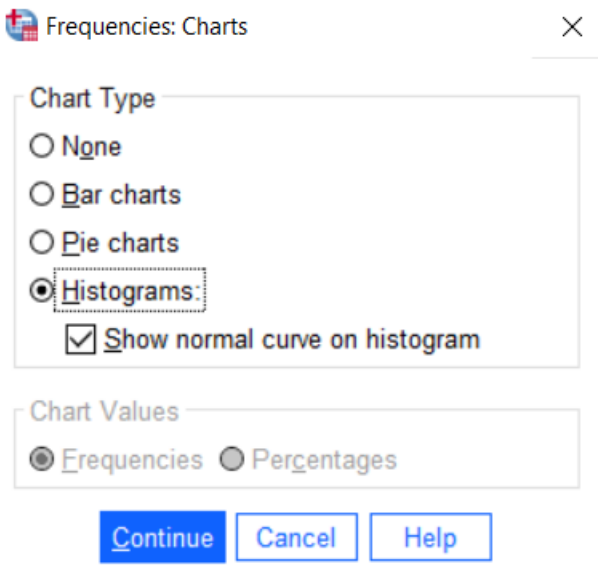
Ανοίγουμε το αρχείο sleep.sav (Εργαστήριο 4) που περιλαμβάνει δεδομένα από μία έρευνα αναφορικά με τα προβλήματα ύπνου που αντιμετωπίζουν οι εργαζόμενοι σε ένα πανεπιστήμιο της Αυστραλίας, τις αιτίες των προβλημάτων, καθώς και τις επιπτώσεις στη ζωή των ατόμων. Ας υποθέσουμε ότι θέλουμε να εξετάσουμε το σχήμα της κατανομής των τιμών ύψους και βάρους των συμμετεχόντων στην έρευνα. Αξίζει να σημειωθεί ότι **πολλά ανθρώπινα χαρακτηριστικά, καθώς και φυσικά φαινόμενα γενικότερα, προσεγγίζουν την κανονική κατανομή**. Ας δούμε αν αυτό επαληθεύεται στο συγκεκριμένο δείγμα συμμετεχόντων. Γενικά, όσο μεγαλύτερο είναι το δείγμα, τόσο περισσότερο προσιδιάζει η κατανομή των τιμών στην κανονική κατανομή.

Επιλέγουμε **Analyze→ Descriptive Statistics → Frequencies**

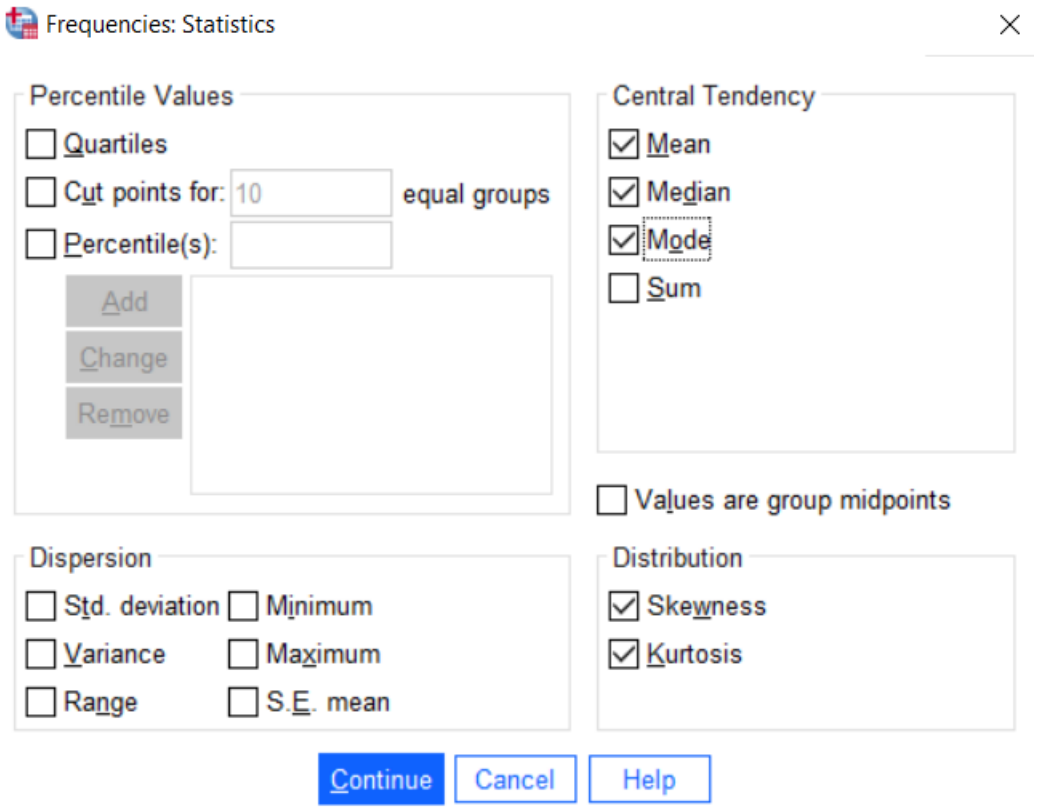
Στο παράθυρο διαλόγου που ανοίγει μεταφέρουμε τις μεταβλητές weight και height στο πλαίσιο Variable(s) και απενεργοποιούμε το Display frequency tables.



Έπειτα πατώντας στο κουμπί Charts ανοίγει νέο παράθυρο διαλόγου όπου επιλέγουμε τα εξής: Histograms και Show normal curve on histogram. Πατάμε Continue.



Προτού πατήσουμε OK, κάνουμε κλικ στο κουμπί Statistics όπου ανοίγει νέο παράθυρο διαλόγου και επιλέγουμε τα εξής: Mean, Median, Mode, Skewness και Kurtosis. Πατάμε Continue και OK.



Αρχικά εμφανίζεται ο παρακάτω πίνακας με τους περιγραφικούς δείκτες κεντρικής τάσης (Mean, Median, Mode) και σχήματος κατανομής (skewness, kurtosis). Παρατηρούμε ότι και στις δύο μεταβλητές, ο μέσος όρος και η διάμεσος βρίσκονται πολύ κοντά μεταξύ τους. Ωστόσο, στην περίπτωση του ύψους η δεσπόζουσα διαφέρει από το μέσο όρο και τη διάμεσο κατά 5 περίπου εκατοστά. Θα μπορούσαμε να θεωρήσουμε, όμως, ότι αυτές οι αποκλίσεις δεν είναι τόσο σοβαρές.

Statistics

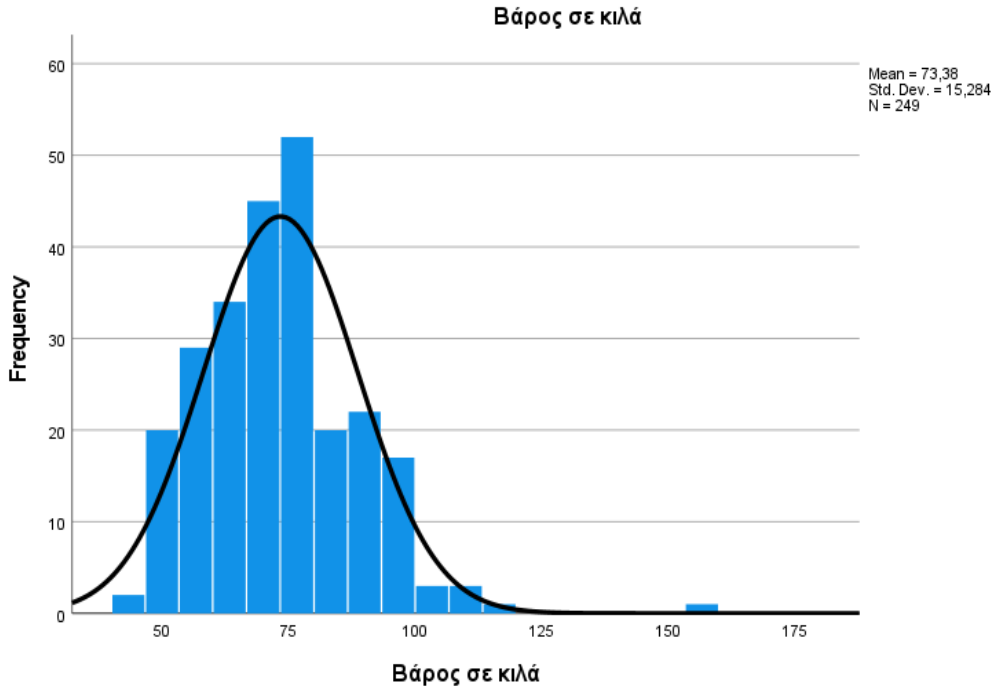
		Βάρος σε κιλά	Ύψος σε εκατοστά
N	Valid	249	246
	Missing	22	25
Mean		73,38	170,22
Median		72,00	170,00
Mode		75	165
Skewness		1,009	,355
Std. Error of Skewness		,154	,155
Kurtosis		3,472	-,610
Std. Error of Kurtosis		,307	,309

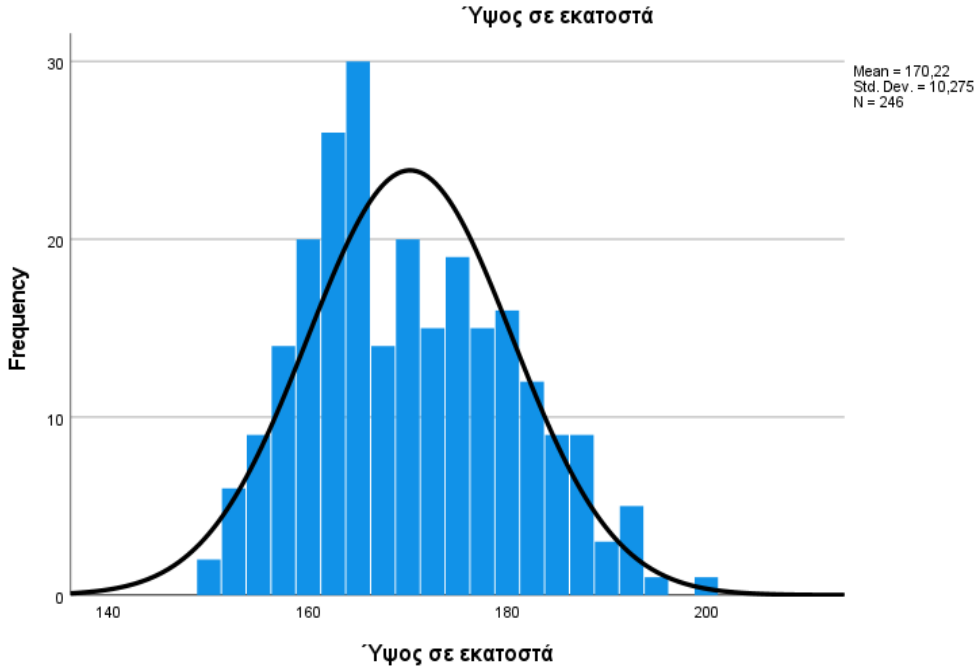
Στον ίδιο πίνακα υπάρχουν, επίσης, δύο άλλες τιμές: η τιμή skewness και η τιμή kurtosis. Η πρώτη μετρά την ασυμμετρία μιας κατανομής σε σχέση με την κανονική κα-

τανομή. Η κανονική κατανομή έχει τιμή ασυμμετρίας ίση με το 0 (είναι απόλυτα συμμετρική). Επομένως, οποιαδήποτε τιμή πάνω από το 0 δηλώνει θετική ασυμμετρία (positive skewness) ενώ οποιαδήποτε τιμή κάτω από το 0 δηλώνει αρνητική ασυμμετρία (negative skewness). Τόσο το βάρος, όσο και το ύψος, έχουν κάποια θετική ασυμμετρία (1 και 0,36, αντίστοιχα). Ωστόσο, η ασυμμετρία της μεταβλητής του ύψους είναι πολύ μικρή.

Η τιμή kurtosis δηλώνει την κυρτότητα (αιχμηρότητα) της καμπύλης κατανομής σε σχέση με την κανονική κατανομή . Η κανονική κατανομή έχει κυρτότητα ίση με το 3. Οποιαδήποτε τιμή πάνω από το 3 υποδηλώνει μια σχετικά λεπτόκυρτη κατανομή τιμών, ενώ οι τιμές πάνω από το 3 δηλώνουν σχετικά πλατύκυρτες κατανομές. Εδώ παρατηρούμε ότι η μεταβλητή του βάρους έχει σχετικά λεπτόκυρτη κατανομή (3,47) που όμως βρίσκεται πολύ κοντά στην τιμή 3 της κανονικής κατανομής, ενώ η μεταβλητή του ύψους έχει σχετικά πλατύκυρτη κατανομή τιμών (-0,61). Γενικά, όμως, αυτές οι αποκλίσεις - ιδιαίτερα εκείνη της μεταβλητής του βάρους - δεν θεωρούνται τόσο σοβαρές στις κοινωνικές έρευνες ώστε να δημιουργούν προβλήματα, είτε στην ερμηνεία της τυπικής απόκλισης, είτε στην περαιτέρω εφαρμογή τεχνικών επαγωγικής στατιστικής (παρόλο που στην ιατρική ή στα χρηματοοικονομικά, καθώς και στην περίπτωση σοβαρών διοικητικών αποφάσεων, μπορεί να έχουν αυξημένη σημασία).

Εκτός από τον πίνακα με τους στατιστικούς δείκτες, στο output εμφανίζονται και τα παρακάτω ιστογράμματα. Παρατηρούμε ότι κάθε ιστόγραμμα αντιπαραβάλλεται με το σχήμα της κανονικής κατανομής, ώστε ο ερευνητής να έχει μια άμεση οπτική εντύπωση σχετικά με το κατά πόσο το σχήμα της υπό μελέτης κατανομής προσιδιάζει ή αποκλίνει από το αντίστοιχο της κανονικής κατανομής. Παρατηρώντας το δύο ιστογράμματα βλέπουμε ότι η μεταβλητή του βάρους αποκλίνει περισσότερο σε ό,τι αφορά την ασυμμετρία (skewness) ενώ η μεταβλητή του ύψους αποκλίνει περισσότερο σε ό,τι αφορά την κύρτωση (kurtosis). Σε όλες τις περιπτώσεις, ωστόσο, οι αποκλίσεις δεν είναι τόσο σοβαρές και ίσως να εξαλείφονταν εάν είχαμε στη διάθεση μας μεγαλύτερο δείγμα ανθρώπων.





Κάθε ιστόγραμμα μας δίνει, επιπλέον, την τιμή του μέσου όρου και της τυπικής απόκλισης. Με βάση αυτά και όσα γνωρίζουμε για την κανονική κατανομή, μπορούμε να υπολογίσουμε τα παρακάτω:

- Το 68% των συμμετεχόντων έχει βάρος που εμπίπτει στο διάστημα 73,38 ± 15,28 κιλά.
- Το 95% των συμμετεχόντων έχει βάρος που εμπίπτει στο διάστημα 73,38 ± (2 x 15,28) κιλά.

- Το 99,7% των συμμετεχόντων έχει βάρος που εμπίπτει στο διάστημα 73,38 ± (3 x 15,28) κιλά.

Παρομοίως, όσον αφορά στο ύψος:

- Το 68% των συμμετεχόντων έχει ύψος που εμπίπτει στο διάστημα 170,22 ± 10,28 εκατοστά.
- Το 95% των συμμετεχόντων έχει ύψος που εμπίπτει στο διάστημα 170,22 ± (2 x 10,28) εκατοστά.
- Το 99,7% των συμμετεχόντων έχει ύψος που εμπίπτει στο διάστημα 170,22 ± (3 x 10,28) εκατοστά.

2. Ανάλυση της σχέσης (συνάφειας) δύο μεταβλητών

Μία στατιστική ανάλυση σπάνια περιορίζεται στη μελέτη μίας μεταβλητής. Συνήθως απαιτείται η μελέτη της σχέσης μεταξύ δύο ή και περισσότερων μεταβλητών. Η μελέτη αυτή ονομάζεται **ανάλυση συνάφειας.** Στην ενότητα αυτή θα δείξουμε πώς μπορούμε να εξετάσουμε τη σχέση (συνάφεια) μεταξύ δύο μεταβλητών χρησιμοποιώντας το πρόγραμμα SPSS. Η τεχνική που θα εφαρμόσουμε εξαρτάται από το είδος των μεταβλητών. Συγκεκριμένα, στο μάθημα αυτό θα ασχοληθούμε με την εύρεση πιθανών σχέσεων μεταξύ: α) δύο μεταβλητών που είναι είτε nominal είτε ordinal και β) δύο μεταβλητών όπου η μία είναι nominal ή ordinal και η άλλη είναι τύπου scale. Στο επόμενο εργαστήριο θα δούμε πώς εξετάζουμε τη συνάφεια όταν και οι δύο μεταβλητές είναι τύπου scale (δείκτες συσχέτισης).

2.1. Συνάφεια μεταξύ μεταβλητών τύπου nominal ή ordinal

Έχουμε ήδη δώσει ένα παράδειγμα διερεύνησης της συνάφειας μεταβλητών τύπου nominal ή ordinal στο 2^{ο} κατά σειρά εργαστήριο. Εκεί εξετάσαμε πώς η συχνότητα επίσκεψης στα καταστήματα τηλεπικοινωνίας διαφοροποιείται ανάλογα με το φύλο των συμμετεχόντων στην έρευνα (μεταξύ αντρών και γυναικών). Είχαμε καταλήξει στο συμπέρασμα ότι οι άντρες του δείγματος επισκέπτονται πιο συχνά τα καταστήματα τηλεπικοινωνίας από τις γυναίκες του δείγματος, **χωρίς να μπορούμε να γενικεύσουμε αυτή τη δήλωση στον ευρύτερη πληθυσμό πελατών της χώρας, καθώς δεν κάναμε κάποια ανάλυση επαγωγικής στατιστικής.**

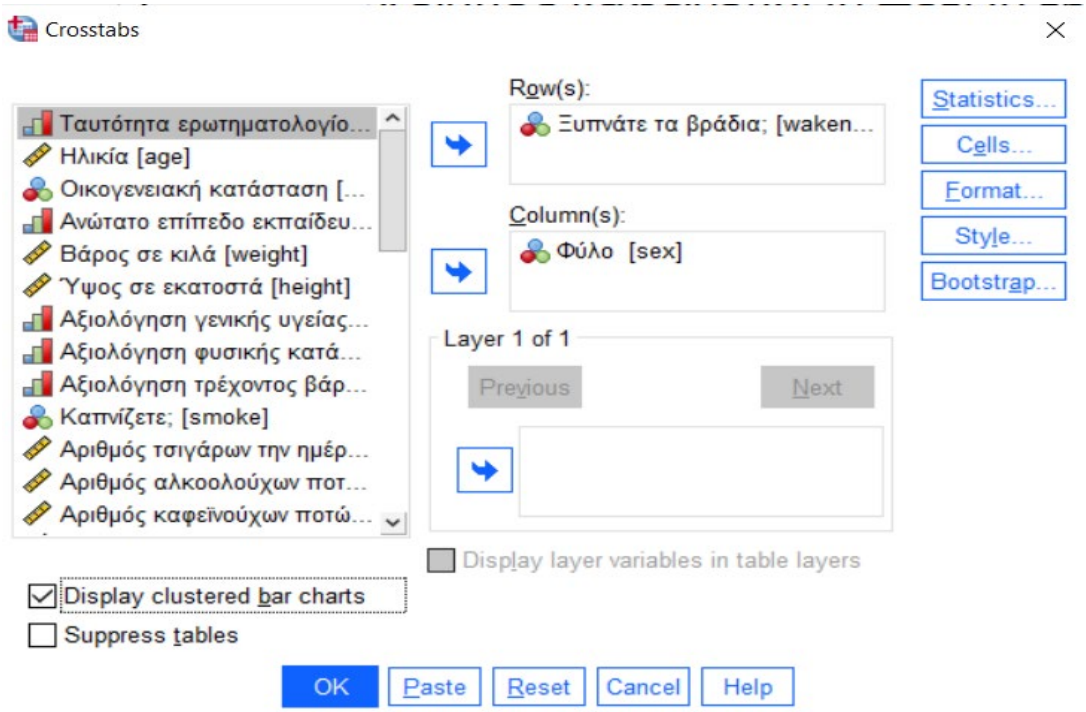
Γενικά, όταν εξετάζουμε τη συνάφεια μεταξύ δύο μεταβλητών, συνήθως ορίζουμε ποια από τις δύο είναι εκείνη που επηρεάζεται (εξαρτάται) από την άλλη (η οποία την επηρεάζει). Η μεταβλητή που επηρεάζεται ονομάζεται **εξαρτημένη μεταβλητή** (dependent variable) ενώ η μεταβλητή που επηρεάζει ονομάζεται **ανεξάρτητη μεταβλητή** (independent variable ή factor). Τα δημογραφικά χαρακτηριστικά συνήθως ορίζονται ως ανεξάρτητες μεταβλητές σε μια έρευνα οι οποίες επηρεάζουν τη συμπεριφορά/σκέψεις/στάσεις/συναισθήματα των ανθρώπων (εξαρτημένες μεταβλητές).

Ας ανοίξουμε το αρχείο sleep.sav που χρησιμοποιήσαμε και προηγουμένως. Ας υποθέσουμε ότι θέλουμε να εξετάσουμε τη σχέση ανάμεσα στις μεταβλητές «ξυπνάτε

τα βράδια;» και «φύλο» που είναι και οι δύο τύπου nominal. Για τους σκοπούς αυτής της ανάλυσης θα ορίσουμε τη μεταβλητή «ξυπνάτε τα βράδια;» ως εξαρτημένη και τη μεταβλητή «φύλο» ως ανεξάρτητη. Η διαδικασία αυτή υλοποιείται ως εξής:

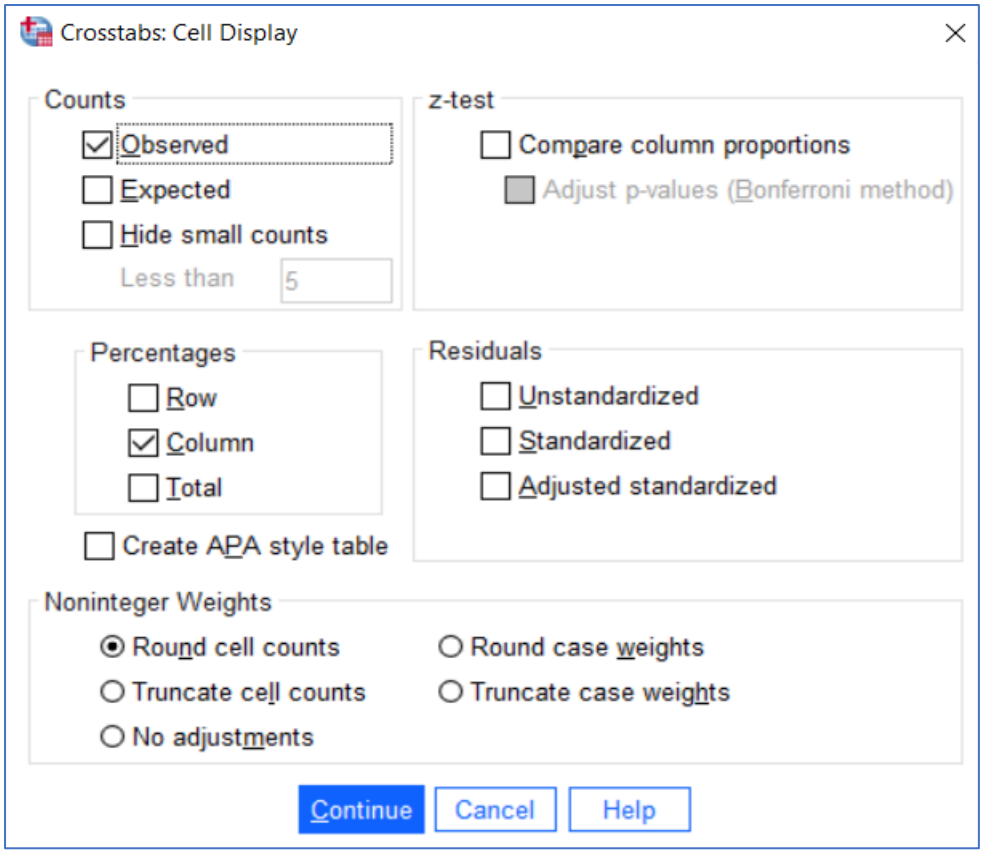
Analyze →Descriptive Statistics→ Crosstabs

Στο νέο παράθυρο διαλόγου που προκύπτει, στο πλαίσιο Row(s) μεταφέρω την εξαρτημένη μεταβλητή «ξυπνάτε τα βράδια;», ενώ στο πλαίσιο Column(s) μεταφέρω τη μεταβλητή «φύλο».[2] Επίσης επιλέγω Display clustered bar charts.



Στο κουτάκι Cells στο πλαίσιο Percentages επιλέγω Column. Έπειτα, Continue και OK.

[2]Αυτή η επιλογή γίνεται συμβατικά. Θα μπορούσαμε να μεταφέρουμε την εξαρτημένη μεταβλητή στις στήλες (Columns) και την ανεξάρτητη μεταβλητή στις γραμμές (Rows). Σε μια τέτοια περίπτωση, όμως, στο επόμενο παράθυρο διαλόγου, στο πλαίσιο Percentages, θα έπρεπε να επιλέξουμε Row (αντί για Column) για να εξασφαλίσουμε ότι για κάθε φύλο ξεχωριστά, οι σχετικές συχνότητες (ποσοστά) έχουν υπολογιστεί επί του συνόλου των συμμετεχόντων του ίδιου φύλου και όχι επί του συνόλου όσων εμπίπτουν σε κάθε κατηγορία της εξαρτημένης μεταβλητής ή επί του συνόλου του δείγματος. Έτσι, αποφεύγουμε πιθανή λανθασμένη ερμηνεία σε περίπτωση που ένα φύλο έχει μεγαλύτερες συχνότητες από ένα άλλο λόγω υπεραντιπροσώπευσής του στο δείγμα, με αποτέλεσμα να επισκιάζει ή να αλλοιώνει οποιαδήποτε πραγματική σχέση συνάφειας μεταξύ εξαρτημένης και ανεξάρτητης μεταβλητής.



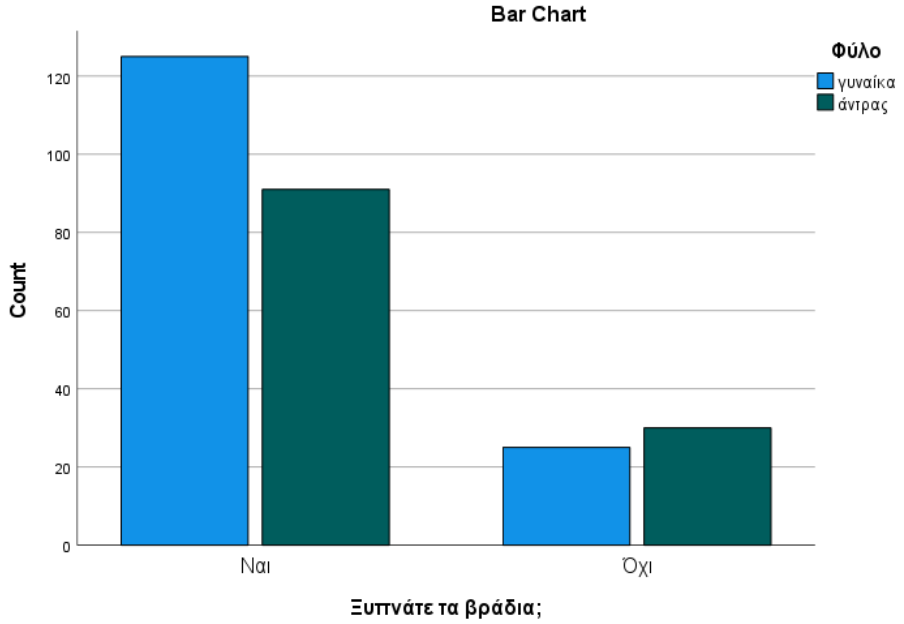
Στο output εμφανίζεται ο παρακάτω πίνακας συνάφειας που ονομάζεται και «πίνακας διπλής εισόδου» καθώς σε αυτόν εισέρχονται δύο διαφορετικές μεταβλητές.

Ξυπνάτε τα βράδια; * Φύλο Crosstabulation

			Φύλο		
			γυναίκα	άντρας	Total
Ξυπνάτε τα βράδια;	Ναι	Count	125	91	216
		% within Φύλο	83,3%	75,2%	79,7%
	Όχι	Count	25	30	55
		% within Φύλο	16,7%	24,8%	20,3%
Total		Count	150	121	271
		% within Φύλο	100,0%	100,0%	100,0%

Με βάση τον πίνακα, παρατηρούμε πως η συντριπτική πλειοψηφία των εργαζομένων στο πανεπιστήμιο (216 από τους 271 συμμετέχοντες, δηλαδή το 79,7%) ξυπνάει τα βράδια. Επιπλέον, παρατηρούμε κάποια διαφορά μεταξύ αντρών και γυναικών. Συγκεκριμένα, από το σύνολο των αντρών που απάντησαν σε αυτή την ερώτηση, το 75,2% ξυπνάνε τα βράδια, ενώ το αντίστοιχο ποσοστό των γυναικών είναι υψηλότερο, δηλαδή 83,3%. Υπάρχει δηλαδή κάποια συνάφεια μεταξύ φύλου και διακοπτόμενου ύπνου στο συγκεκριμένο δείγμα. Θα μπορούσε κάποιος να υποθέσει ότι οι γυναίκες είναι πιο επιρρεπείς σε προβλήματα ύπνου από ό,τι οι άντρες. Φυσικά, για να φτάσουμε σε αυτό το συμπέρασμα απαιτείται περαιτέρω στατιστική και θεωρητική διερεύνηση.

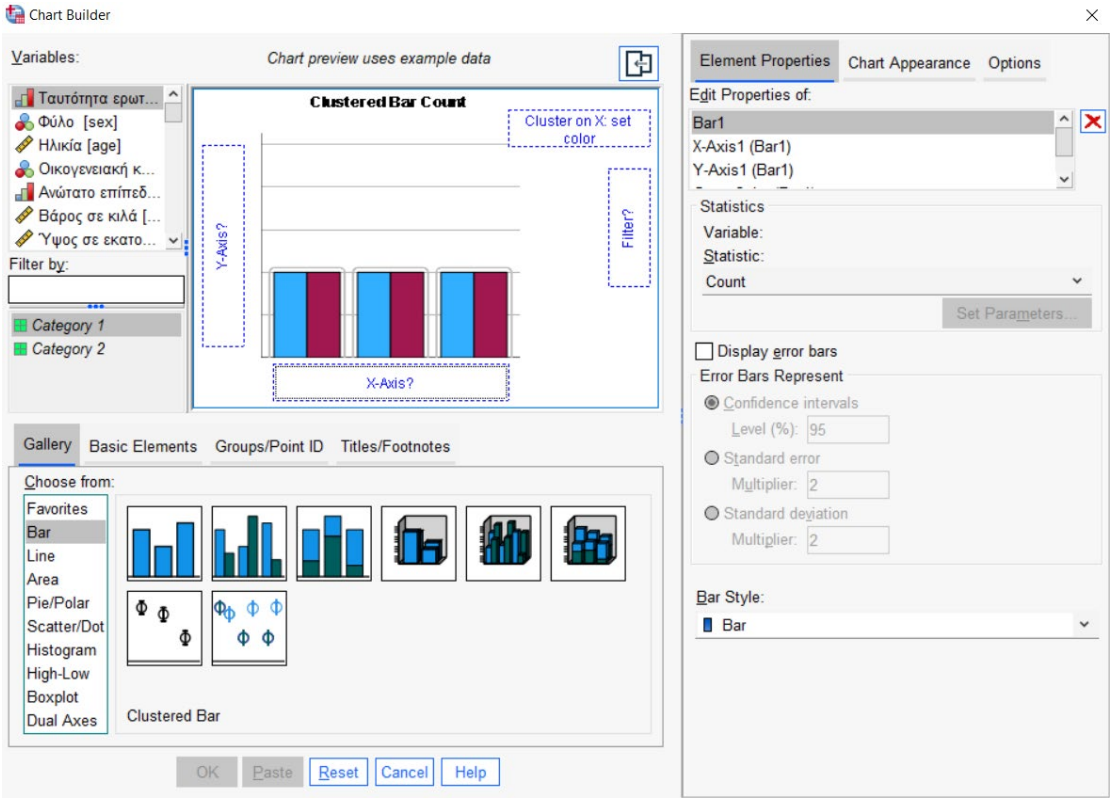
Στο output εμφανίζεται επίσης και το παρακάτω ομαδοποιημένο ραβδόγραμμα:



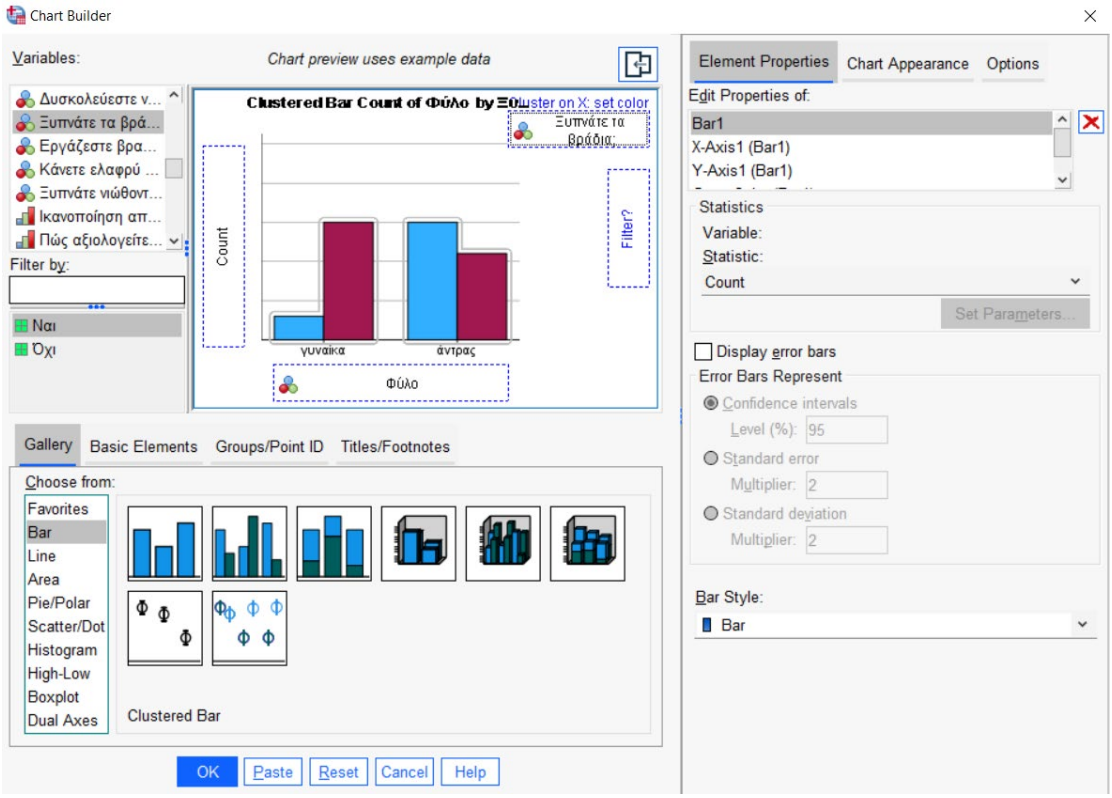
Κατά την ανάγνωση και ερμηνεία του παραπάνω ραβδογράμματος πρέπει να είμαστε προσεκτικοί, ειδικά όταν έχουμε υπεραντιπροσώπευση ενός φύλου στο δείγμα μας. Όπως αναφέρθηκε και προηγουμένως, σε μια τέτοια περίπτωση, το φύλο που υπεραντιπροσωπεύεται θα έχει μεγαλύτερες συχνότητες από το άλλο φύλο λόγω υπεραντιπροσώπευσης, με αποτέλεσμα να επισκιάζει ή να αλλοιώνει οποιαδήποτε πραγματική σχέση συνάφειας μεταξύ εξαρτημένης και ανεξάρτητης μεταβλητής.

Για να αποφύγουμε μια τέτοια πιθανή παρερμηνεία, χρειάζεται να κατασκευάσουμε ραβδόγραμμα όπου στον οριζόντιο άξονα θα τοποθετήσουμε τις κατηγορίες της ανεξάρτητης μεταβλητής «Φύλο» (δηλαδή ΑΝΤΡΑΣ/ΓΥΝΑΙΚΑ) και σε κάθε μία από τις κατηγορίες αυτές θα αντιστοιχούν δύο ράβδοι, μία για κάθε κατηγορία της εξαρτημένης μεταβλητής «Ξυπνάτε τα βράδια;» (δηλαδή ΝΑΙ/ΟΧΙ). Χρειάζεται, επίσης, οι ράβδοι να εκφράζουν τα ποσοστά όσων ξυπνούν ή δεν ξυπνούν τα βράδια για κάθε φύλο ξεχωριστά (δηλαδή κάθε ζεύγος ράβδων να αθροίζει στο 100%). Με αυτόν τον τρόπο, εξασφαλίζουμε ότι τα αποτελέσματα δεν θα επηρεαστούν από τυχόν υπεραντιπροσώπευση του ενός ή του άλλου φύλου στο σύνολο του δείγματος. Για να το κάνουμε αυτό, επιλέγουμε από το κεντρικό μενού **Graphs →Chart Builder.**

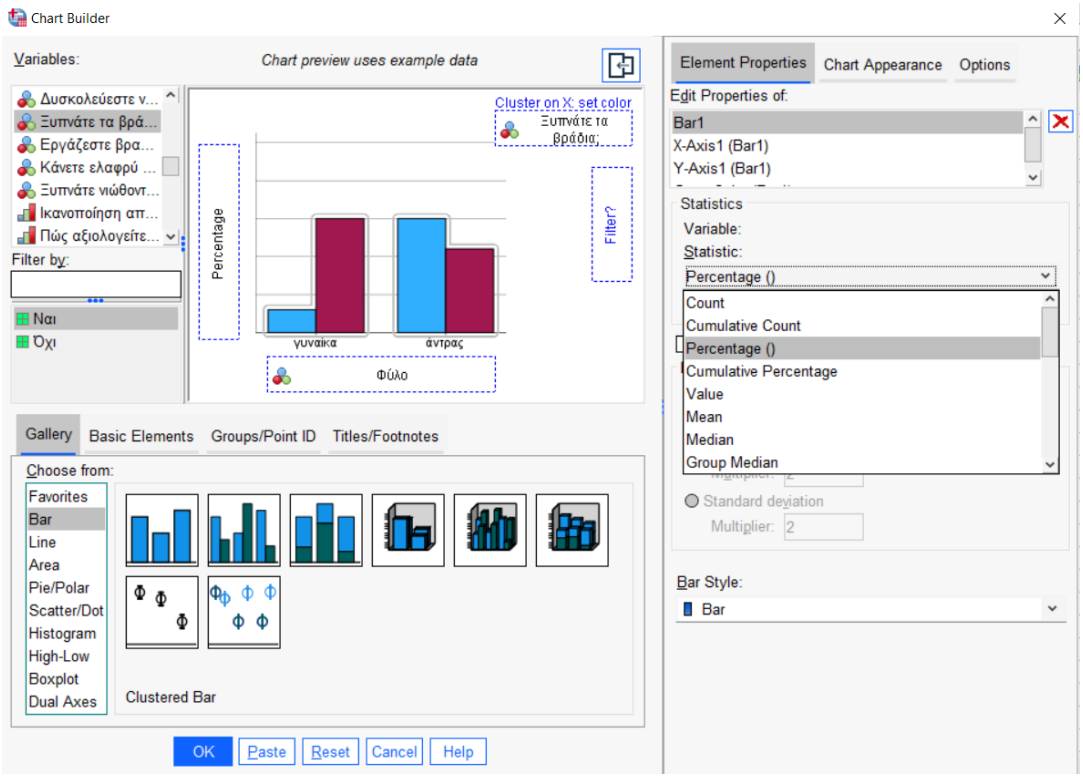
Στο πλαίσιο διαλόγου που εμφανίζεται επιλέγουμε Bar από τη λίστα Gallery και στη συνέχεια μεταφέρουμε το δεύτερο κατά σειρά εικονίδιο Clustered Bar στο πλαίσιο Chart Preview που βρίσκεται ακριβώς από πάνω.



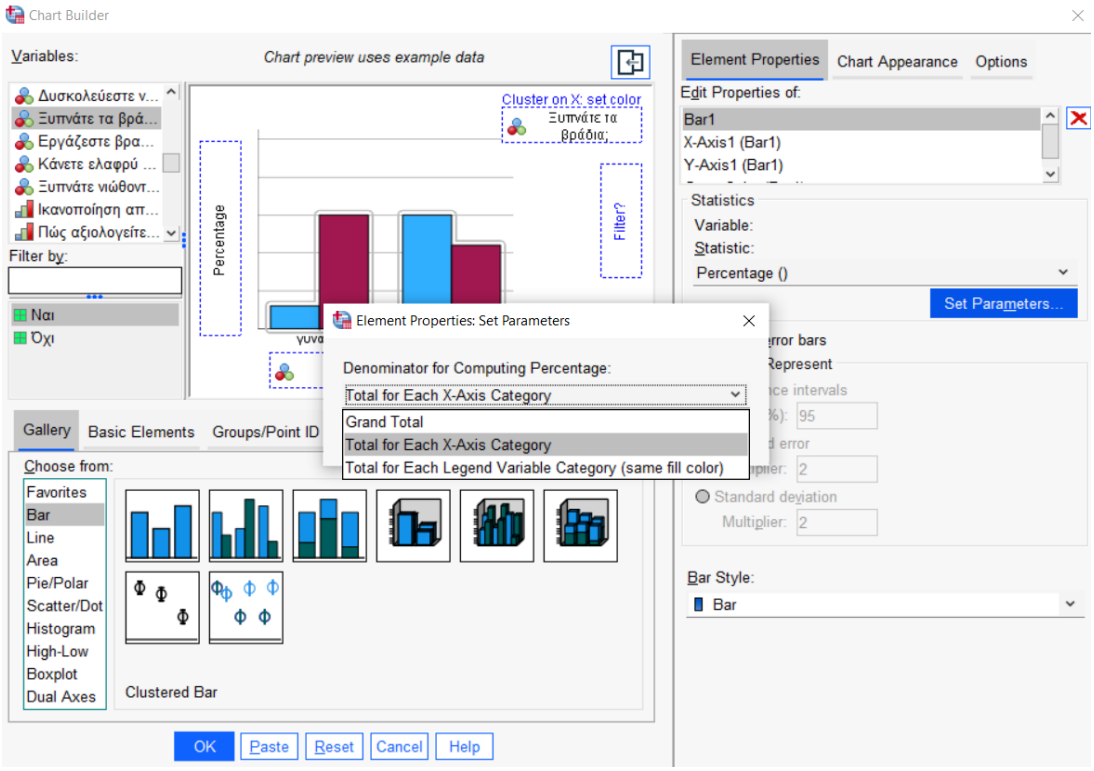
Κατόπιν μεταφέρουμε την ανεξάρτητη μεταβλητή «Φύλο» στη θέση X-Axis? και την εξαρτημένη μεταβλητή «Ξυπνάτε τα βράδια;» στη θέση Cluster on X: set color.



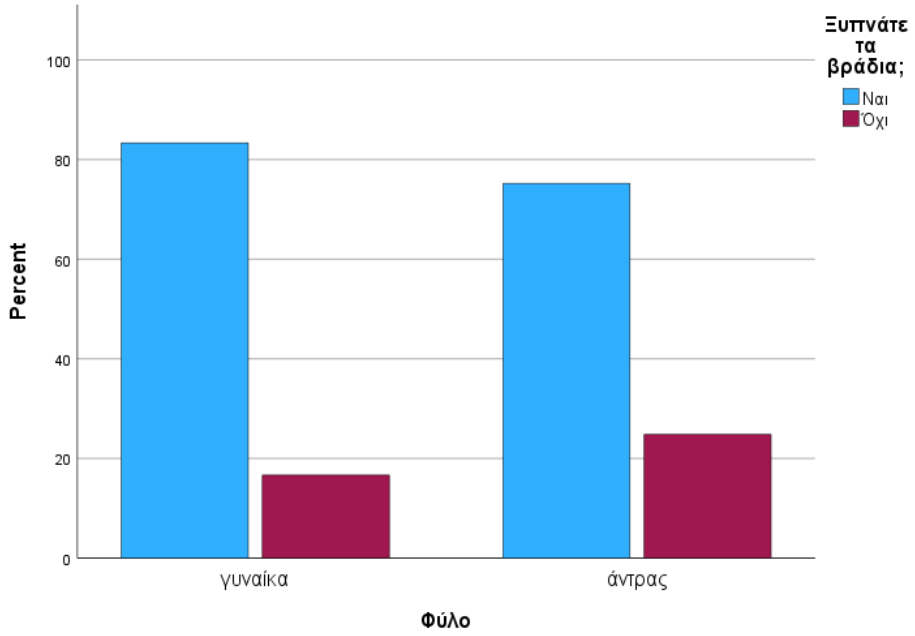
Στη συνέχεια, στο μενού Element Properties κάνουμε κλικ στο βελάκι που βρίσκεται δεξιά της επιλογής Count και στη λίστα επιλογών που εμφανίζεται επιλέγουμε Percentage.



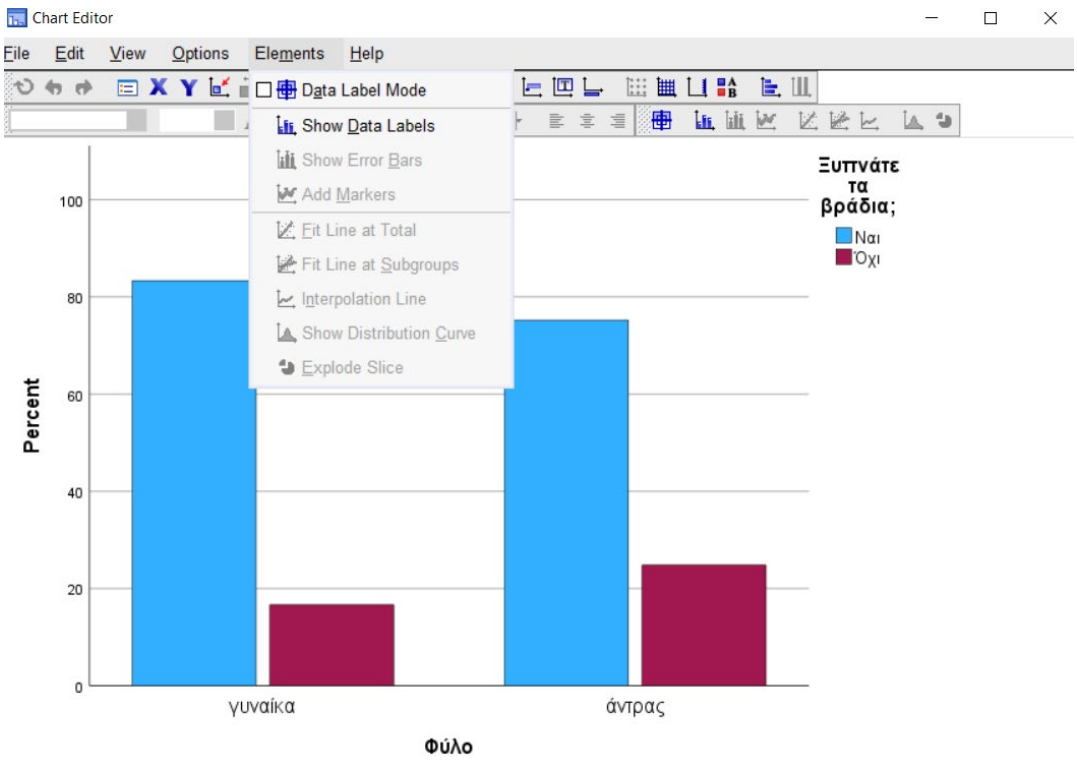
Κατόπιν πατάμε το κουμπί Set Parameters και στο νέο παράθυρο διαλόγου που εμφανίζεται επιλέγουμε Total for Each X-Axis Category.

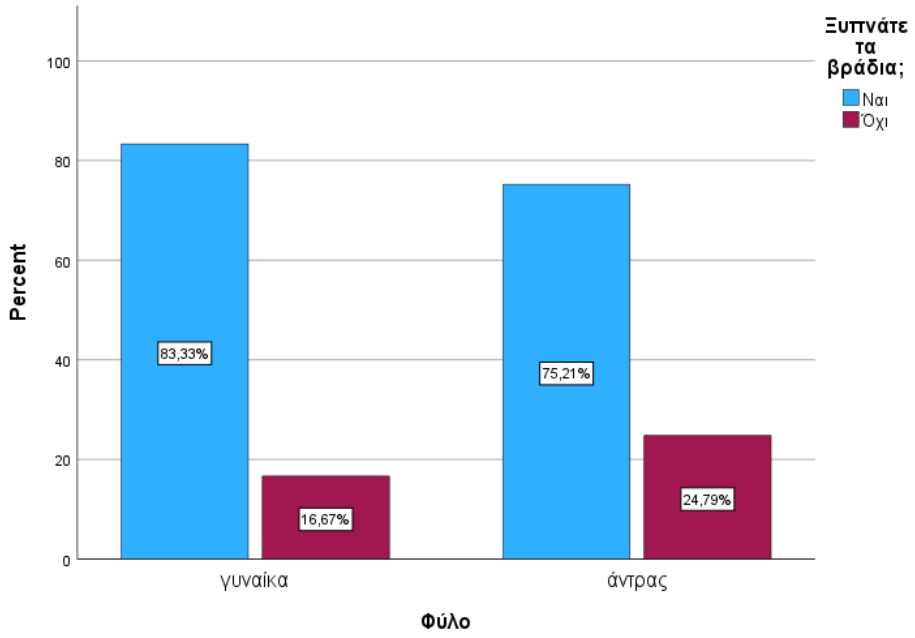


Τέλος, πατάμε Continue και OK και εμφανίζεται το παρακάτω ομαδοποιημένο ραβδόγραμμα.



Κάνοντας διπλό κλικ επάνω στις ράβδους εμφανίζεται το πλαίσιο διαλόγου Chart editor, όπου εάν επιλέξουμε στο μενού Elements την εντολή Show data labels θα εμφανιστούν τα ποσοστά (σχετικές συχνότητες) σε κάθε ράβδο.





Παρατηρούμε, λοιπόν, ότι στο σύνολο των γυναικών, το 83% περίπου ξυπνάει τα βράδια, ενώ στο σύνολο των αντρών το αντίστοιχο ποσοστό είναι μικρότερο, δηλαδή 75% περίπου.

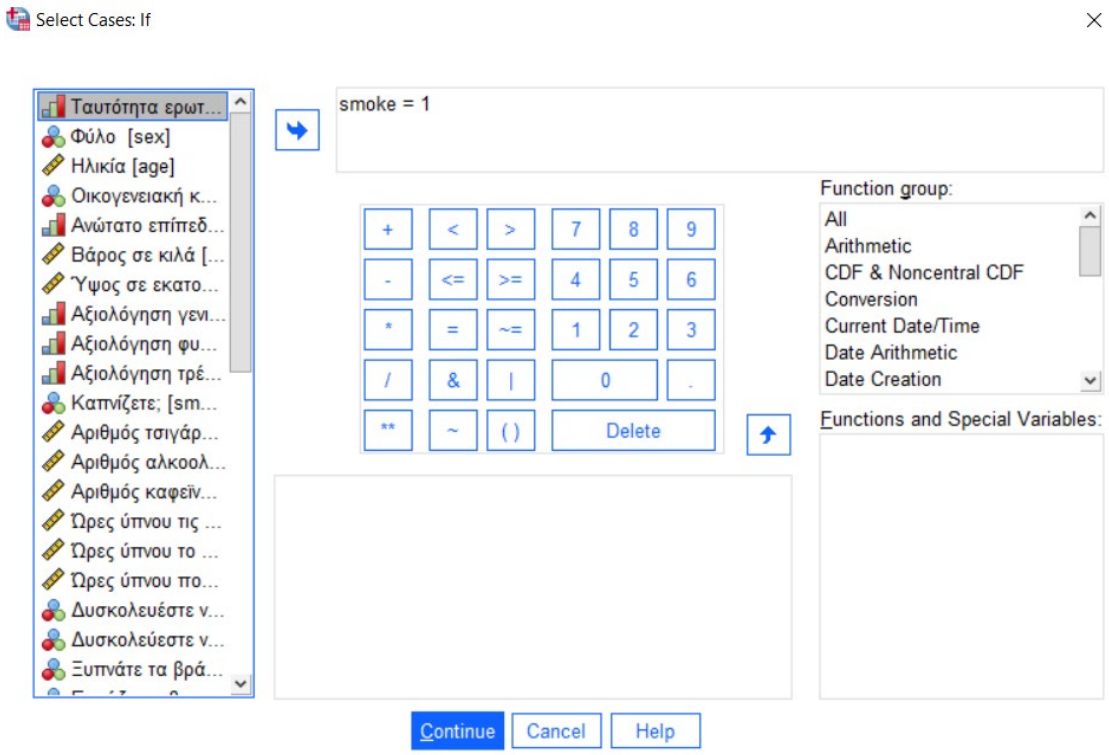
2.2. Συνάφεια μεταξύ δύο μεταβλητών nominal/ordinal και scale

Στο 3ο κατά σειρά εργαστήριο δώσαμε ένα παράδειγμα διερεύνησης της συνάφειας δύο μεταβλητών όπου η μία είναι τύπου nominal ή ordinal και η άλλη τύπου scale. Εκεί εξετάσαμε πώς το μηνιαίο εισόδημα των πελατών τεσσάρων καταστημάτων τηλεπικοινωνίας στην Ελλάδα διαφοροποιείται ανάλογα με το φύλο τους. Είχαμε καταλήξει στο συμπέρασμα ότι οι άντρες του δείγματος της έρευνας αυτής έχουν υψηλότερα και πιο ανομοιογενή εισοδήματα από τις γυναίκες, **χωρίς να μπορούμε να γενικεύσουμε αυτή τη δήλωση στον ευρύτερη πληθυσμό πελατών της χώρας, καθώς δεν κάναμε κάποια ανάλυση επαγωγικής στατιστικής.**

Εδώ θα δώσουμε ένα άλλο παράδειγμα χρησιμοποιώντας κάποιες εντολές του SPSS που δεν αναφέρθηκαν στο εργαστήριο 3. Ας ανοίξουμε το αρχείο sleep.sav που χρησιμοποιήσαμε και προηγουμένως. Ας υποθέσουμε ότι θέλουμε να εξετάσουμε τη σχέση ανάμεσα στις μεταβλητές «αριθμός τσιγάρων την ημέρα» και «φύλο» όπου η πρώτη είναι μεταβλητή τύπου scale και η δεύτερη μεταβλητή τύπου nominal. Με άλλα λόγια, θέλουμε να δούμε εάν μεταξύ των καπνιστών του δείγματος, ο αριθμός τσιγάρων που καπνίζουν την ημέρα εξαρτάται από το φύλο τους. Για τους σκοπούς αυτής της ανάλυσης, θα ορίσουμε τη μεταβλητή «αριθμός τσιγάρων την ημέρα» ως εξαρτημένη μεταβλητή και τη μεταβλητή «φύλο» ως ανεξάρτητη μεταβλητή. Η διαδικασία υλοποιείται ως εξής:

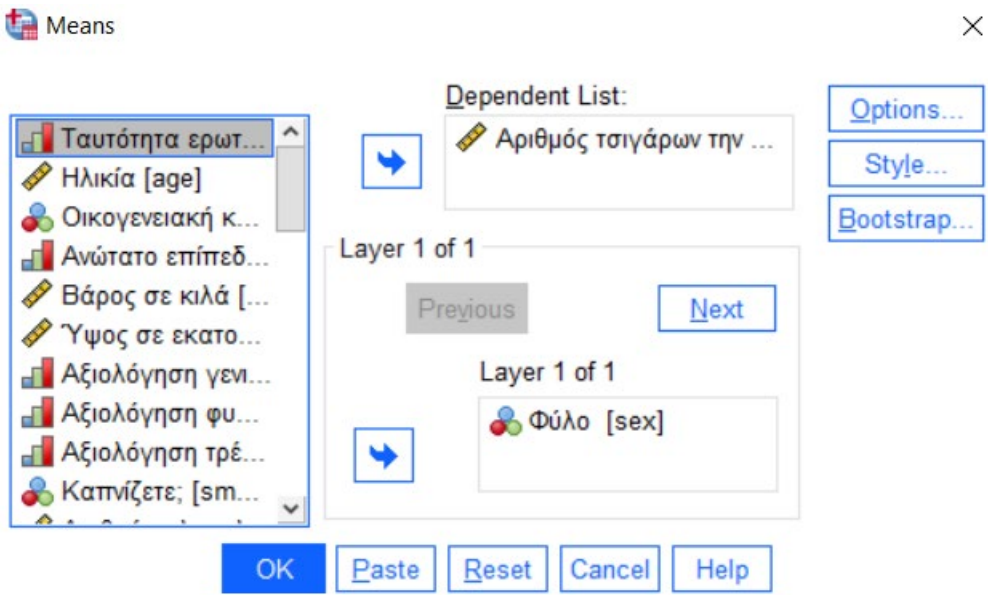
Αρχικά πρέπει να περιορίσουμε την ανάλυση μας στην υποομάδα των καπνιστών χρησιμοποιώντας την εντολή Select Cases.

Data → Select cases → If condition is satisfied όπου ορίζουμε τη μεταβλητή smoke=1.

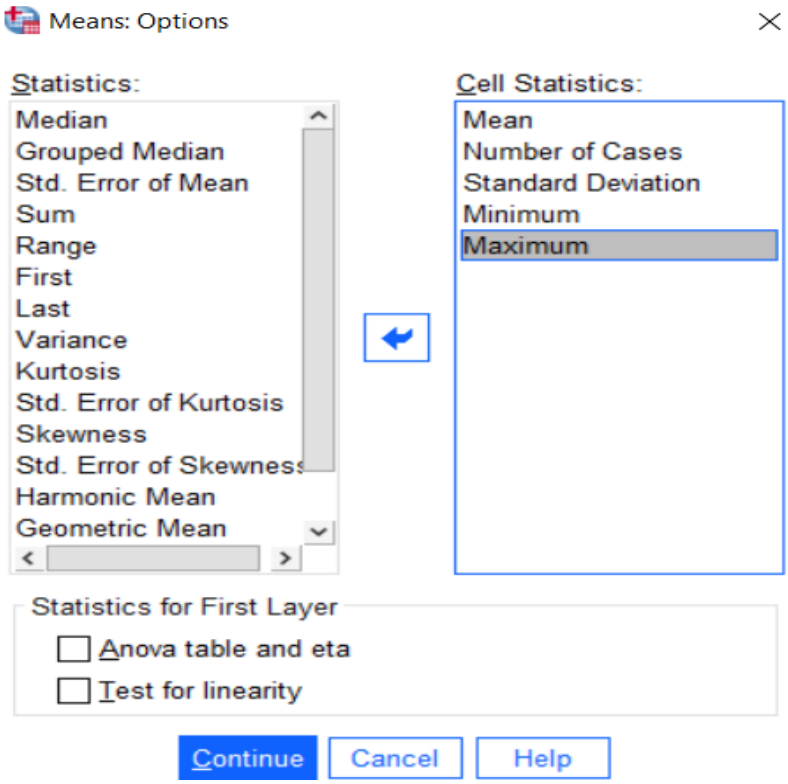


Πατάμε continue και OK.

Στη συνέχεια επιλέγουμε **Analyze → Compare Means → Means.** Στο πλαίσιο Dependent List τοποθετούμε τη μεταβλητή «αριθμός τσιγάρων την ημέρα» και στο πλαίσιο Layer 1 of 1 τοποθετούμε τη μεταβλητή «φύλο».



Κάνουμε κλικ στην επιλογή Options και επιβεβαιώνουμε ότι οι δείκτες Μέσος όρος (Mean) και τυπική απόκλιση (Standard Deviation) έχουν επιλεγεί και μεταφερθεί στο κουτί Cell statistics. Παρατηρούμε ότι αυτό όντως συμβαίνει. Μεταφέρουμε επιπλέον τους δείκτες Maximum και Minimum και στη συνέχεια, πατάμε continue και OK.



Εμφανίζεται ο παρακάτω πίνακας όπου βλέπουμε το μέσο όρο, την τυπική απόκλιση, τη μέγιστη και την ελάχιστη τιμή του αριθμού τσιγάρων που καπνίζουν οι γυναίκες και οι άντρες καπνιστές του δείγματος, ξεχωριστά. Παρατηρούμε ότι οι γυναίκες καπνίστριες (18 συνολικά) καπνίζουν κατά μέσο όρο 3 περίπου τσιγάρα περισσότερα την ημέρα από τους άντρες καπνιστές (15 άντρες συνολικά). Επίσης, υπάρχει μεγαλύτερη ανομοιογένεια στην ποσότητα των τσιγάρων που καπνίζουν οι γυναίκες από ό,τι οι άντρες (Τ.Α. = 17,42 και Τ.Α. =10,25, αντίστοιχα) ενώ η μέγιστη τιμή που καταγράφτηκε στις γυναίκες ήταν 78 τσιγάρα την ημέρα και στους άντρες 32.

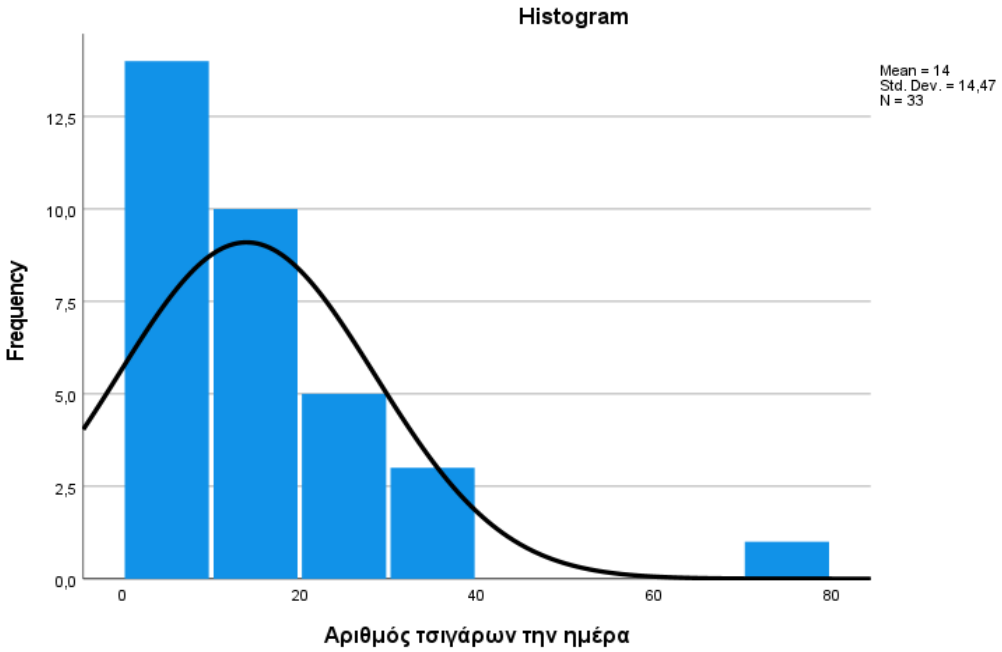
Report

Αριθμός τσιγάρων την ημέρα

Φύλο	Mean	N	Std. Deviation	Minimum	Maximum
γυναίκα	15,33	18	17,422	2	78
άντρας	12,40	15	10,246	1	32
Total	14,00	33	14,470	1	78

Μπορεί κάποιος να αναρωτηθεί εδώ εάν η μέγιστη τιμή των γυναικών είναι μια εξαιρετικά ακραία τιμή που επηρεάζει όλα τα άλλα αποτελέσματα. Δηλαδή μπορεί να αντιστοιχεί σε μια ειδική περίπτωση γυναίκας που είναι βαριά καπνίστρια χωρίς αυτό

να αντιπροσωπεύει τις υπόλοιπες γυναίκες γενικά. Αν πάρουμε το ιστόγραμμα της μεταβλητής «αριθμός τσιγάρων την ημέρα» καταλαβαίνουμε πως όντως κάτι τέτοιο συμβαίνει. Υπάρχει δηλαδή μια γυναίκα που καπνίζει πάρα πολλά τσιγάρα την ημέρα και απέχει κατά πολύ από τους υπόλοιπους καπνιστές.



Αν εξαιρέσουμε αυτή την περίπτωση από το δείγμα των καπνιστών και τρέξουμε ξανά την ανάλυση Compare Means θα πάρουμε τον παρακάτω πίνακα:

Report

Αριθμός τσιγάρων την ημέρα

Φύλο	Mean	N	Std. Deviation	Minimum	Maximum
γυναίκα	11,65	17	7,913	2	30
άντρας	12,40	15	10,246	1	32
Total	12,00	32	8,937	1	32

Παρατηρούμε τώρα ότι οι γυναίκες-καπνίστριες του δείγματος δεν διαφέρουν αξιοσημείωτα από τους άντρες-καπνιστές σε ό,τι αφορά τον μέσο αριθμό των τσιγάρων που καπνίζουν την ημέρα, οπότε καταλήγουμε στο συμπέρασμα πως δεν υπάρχει συνάφεια στο δείγμα μας μεταξύ των μεταβλητών «φύλο» και «αριθμός τσιγάρων την ημέρα». Το μόνο που θα μπορούσαμε να πούμε είναι πως οι άντρες συνιστούν μια περισσότερο ανομοιογενή ομάδα καπνιστών από τις γυναίκες συγκρίνοντας τις τυπικές αποκλίσεις των δύο. Βλέπουμε δηλαδή πόσο απαραίτητο είναι να ελέγχουμε τη βάση δεδομένων μας για ακραίες τιμές καθώς αυτές μπορούν να αλλοιώσουν υπερβολικά τα αποτελέσματα και τα συμπεράσματά μας!

Τέλος, επειδή η κατανομή των τιμών της μεταβλητής «αριθμός τσιγάρων την ημέρα» έχει μεγάλη θετική ασυμμετρία δεν μπορούμε να εφαρμόσουμε τον τρόπο υπολογισμού διαστημάτων στα οποία εμπίπτει η πλειοψηφία των καπνιστών του δείγματος που εφαρμόζεται στην περίπτωση συμμετρικών κανονικών κατανομών.