3ο ΕΡΓΑΣΤΗΡΙΟ

Δείκτες διασποράς των τιμών ποσοτικών μεταβλητών

Στο 3ο κατά σειρά εργαστήριο θα παρουσιάσουμε πώς εξάγουμε **περιγραφικούς δείκτες διασποράς τιμών,** καθώς και **διαγράμματα διασποράς,** για καθεμία από τις μεταβλητές που υπάρχουν σε μια βάση δεδομένων, χρησιμοποιώντας το πρόγραμμα SPSS. **Οι δείκτες διασποράς εφαρμόζονται σε ποσοτικές μεταβλητές.** Κάποιοι (όπως το εύρος) εφαρμόζονται τόσο σε συνεχείς όσο και σε ασυνεχείς μεταβλητές, ενώ άλλοι (όπως το ενδοτεταρτημοριακό εύρος) εφαρμόζονται μόνο σε συνεχείς μεταβλητές ή σε ασυνεχείς μεταβλητές με πολλές όμως διαφορετικές τιμές ώστε να έχουν παρόμοιες ιδιότητες με τις συνεχείς μεταβλητές. Κάθε φορά που θα παρουσιάζουμε έναν δείκτη διασποράς θα επισημαίνουμε και το είδος των μεταβλητών στις οποίες εφαρμόζεται.

Στο προηγούμενο εργαστήριο παρουσιάσαμε τους περιγραφικούς δείκτες κεντρικής τάσης που μας δείχνουν **το «κέντρο» των τιμών μιας μεταβλητής** ή αλλιώς, το αριθμητικό σημείο γύρω από το οποίο συγκεντρώνονται οι τιμές της. Σε αυτό το εργαστήριο θα παρουσιάσουμε μια άλλη κατηγορία περιγραφικών δεικτών που μας δείχνουν **τη διακύμανση ή διασπορά των τιμών μιας μεταβλητής γύρω από το «κέντρο» της**, δηλαδή πόσο αποκλίνουν οι διαφορετικές τιμές της από τους δείκτες κεντρικής τάσης (πόσο κοντά ή πόσο μακριά βρίσκονται από αυτούς). Οι δείκτες αυτοί ονομάζονται **δείκτες διασποράς.**

Γιατί όμως είναι απαραίτητη η μέτρηση της διασποράς των τιμών μιας μεταβλητής;
Ας υποθέσουμε ότι ένας καθηγητής έδωσε το ίδιο τεστ μαθηματικών σε δύο ομάδες μαθητών (ομάδα Α και Β) και είχε τις ακόλουθες επιδόσεις σε εικοσαβάθμια κλίμακα:

Πίνακας 1 – Επιδόσεις μαθητών στο τεστ μαθηματικών

Ομάδα Α	Ομάδα Β
15	10
17	11
16	15
14	19
13	20

Ο καθηγητής θέλει να δει ποια ομάδα είχε την καλύτερη επίδοση. Παρόλο που οι επιδόσεις διαφέρουν ανάμεσα στις δύο ομάδες, ο Μέσος Όρος και η Διάμεσος είναι ακριβώς τα ίδια και στις δύο!! Πιο συγκεκριμένα, ο Μέσος Όρος ισούται με 15 και στις δύο ομάδες, ενώ η διάμεσος είναι επίσης 15. Παρατηρούμε, ωστόσο, ότι η ομάδα Β έχει περισσότερες ακραίες τιμές από την Α. Αυτό σημαίνει ότι η ομάδα Β είναι **πιο ανομοιογενής** σε σχέση με την ομάδα Α. Το γεγονός αυτό είναι εξίσου

σημαντικό και ενδιαφέρον με τη μέση επίδοση των μαθητών, καθώς π.χ. επηρεάζει τον απαιτούμενο βαθμό εξατομίκευσης της διδασκαλίας. Ο βαθμός στον οποίο οι τιμές μιας μεταβλητής αποκλίνουν από τους δείκτες κεντρικής τάσης φανερώνεται μέσω των δεικτών διασποράς που θα παρουσιάσουμε σε αυτό το μάθημα. Συγκεκριμένα, θα παρουσιάσουμε τους εξής δείκτες: το εύρος, το ενδοτεταρτημοριακό εύρος, τη μέση απόκλιση και την τυπική απόκλιση.

Α. Εύρος (range)

Το εύρος είναι ο απλούστερος δείκτης διασποράς και εφαρμόζεται σε όλες τις ποσοτικές μεταβλητές, συνεχείς και ασυνεχείς. Ορίζεται ως η διαφορά της μέγιστης και της ελάχιστης τιμής μιας μεταβλητής. Στο παραπάνω παράδειγμα, το εύρος των τιμών της πρώτης ομάδας είναι 4 (17-13), ενώ το εύρος των τιμών της δεύτερης ομάδας είναι 10 (20-10). **ΠΡΟΣΟΧΗ:** Το εύρος είναι μια συγκεκριμένη τιμή και όχι το διάστημα μεταξύ δυο τιμών! Είναι, επομένως, λάθος να πούμε ότι «το εύρος κυμαίνεται από 10 έως 20».

Β. Ενδοτεταρτημοριακό εύρος (interquartile range)

Ενδοτεταρτημοριακό εύρος είναι το εύρος του κεντρικού 50% των τιμών μιας μεταβλητής. Το υπόλοιπο 50% των τιμών (ανώτερο 25% και κατώτερο 25% των τιμών) δεν λαμβάνεται υπόψη κατά τον υπολογισμό του ενδοτεταρτημοριακού εύρους. Τα σημεία που χωρίζουν τις τιμές μιας μεταβλητής σε τέσσερα ίσα τμήματα ονομάζονται τεταρτημόρια (quartiles). Κάθε μεταβλητή έχει 3 τεταρτημόρια: Το 1^{ο} είναι το σημείο κάτω από το οποίο βρίσκεται το 25% των τιμών. Το 2^{ο} τεταρτημόριο είναι το σημείο κάτω από το οποίο βρίσκεται το 50% των τιμών (δηλαδή ταυτίζεται με τη Διάμεσο). Το 3ο τεταρτημόριο είναι το σημείο κάτω από το οποίο βρίσκεται το 75% των τιμών.

Το ενδοτεταρτημοριακό εύρος εφαρμόζεται συνήθως (αλλά όχι μόνο) σε συνεχείς μεταβλητές ή σε ασυνεχείς (διακριτές) μεταβλητές που όμως παίρνουν τόσες πολλές διαφορετικές τιμές ώστε να αποκτούν τις ιδιότητες των συνεχών μεταβλητών.

Γ. Μέση απόκλιση (mean deviation) και τυπική απόκλιση (standard deviation)

Είναι σημαντικό σε μια μεταβλητή να γνωρίζουμε, εκτός από το μέσο όρο των τιμών της, τις αποστάσεις των διαφόρων τιμών από τον εν λόγω μέσο όρο. Η διαφορά μιας συγκεκριμένης τιμής από το μέσο όρο ονομάζεται «απόκλιση» και συμβολίζεται με το λατινικό γράμμα **d.** Η απόκλιση μιας συγκεκριμένης τιμής υπολογίζεται με τον τύπο $\mathbf{di = x_i - \bar{x}}$ (όπου i είναι ο αύξων αριθμός του συμμετέχοντα που έδωσε τιμή $\mathbf{x_i}$ στο ερωτηματολόγιο του). Στο προηγούμενο παράδειγμα, η απόκλιση του αριθμού 13 που πήρε ένας μαθητής της Α ομάδας από το Μ.Ο. είναι **d = 13 – 15 = -2.** Παρατηρούμε ότι η απόκλιση έχει αρνητικό πρόσημο, δηλαδή ο συγκεκριμένος μαθητής έγραψε κάτω από το μέσο όρο της ομάδας του.

Με βάση τις αποκλίσεις **di** των διαφόρων τιμών μιας μεταβλητής, μπορούμε να υπολογίσουμε δύο διαφορετικούς δείκτες διασποράς: **τη μέση απόκλιση** και **την τυπική απόκλιση.**

ΜΕΣΗ ΑΠΟΚΛΙΣΗ (MD)

Η **μέση απόκλιση** είναι ο μέσος όρος όλων των επιμέρους αποκλίσεων αφού πρώτα αγνοήσουμε πιθανά αρνητικά πρόσημα (παίρνοντας δηλ. τις απόλυτες τιμές τους). Μας δείχνει δηλαδή πόσο απέχουν κατά μέσο όρο οι διάφορες τιμές μιας μεταβλητής από τον μέσο όρο της. Υπολογίζεται με τον τύπο:

$$MD = \frac{\sum |Xi - \bar{x}|}{N}$$

όπου **N** είναι το πλήθος των τιμών της μεταβλητής εκτός των missing values.

ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ (s)

Η **τυπική απόκλιση** είναι λίγο διαφορετική από τη μέση απόκλιση. Αντί για την «μέση απόσταση» μας δείχνει την «τυπική απόσταση» των τιμών μιας μεταβλητής από το μέσο όρο και υπολογίζεται ως εξής:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}}$$

- x_i είναι οι τιμές της κατανομής ($x_1, x_2\ x_3 \ldots x_N$)
- $\bar{x}$ είναι ο μέσος όρος των τιμών αυτών.
- Ν είναι το μέγεθος του δείγματος (πλήθος τιμών).

Δεν χρειάζεται να θυμόμαστε τους παραπάνω τύπους για να τρέξουμε μια ανάλυση διασποράς στο SPSS. Απλά χρειάζεται να γνωρίζουμε ότι οι παραπάνω δείκτες φανερώνουν **πόσο ανομοιογενείς είναι οι τιμές** μιας μεταβλητής και τον βαθμό στον οποίο απέχουν από το μέσο όρο της.

Τόσο η μέση απόκλιση, όσο και τη τυπική απόκλιση, εκφράζονται στην ίδια κλίμακα μέτρησης με τα δεδομένα τα οποία περιγράφουν. Εάν δηλαδή μια μεταβλητή μετράει το βάρος σε κιλά, η τυπική απόκλιση των τιμών της θα εκφράζει κι αυτή βάρος σε κιλά. Το ίδιο θα ισχύει και για την μέση απόκλιση.

Η τυπική απόκλιση είναι ο πιο συχνά χρησιμοποιούμενος δείκτης διασποράς και προτιμάται σε σχέση με τη μέση απόκλιση. Εφαρμόζεται τόσο σε συνεχείς, όσο και σε ασυνεχείς, ποσοτικές μεταβλητές. Ωστόσο, η ερμηνεία της γίνεται περισσότερο αξιόπιστη όταν εφαρμόζεται σε συνεχείς ή ασυνεχείς μεταβλητές με πολλές διαφορετικές τιμές οι οποίες παίρνουν το σχήμα **κανονικής κατανομής** (κάτι που θα εξηγηθεί σε επόμενο μάθημα). Είναι δηλαδή ιδιαίτερα χρήσιμη όταν οι τιμές μιας μεταβλητής **κατανέμονται «κανονικά» γύρω από το μέσο όρο της** (τι ακριβώς σημαίνει αυτό, θα το εξηγήσουμε σε επόμενο μάθημα).

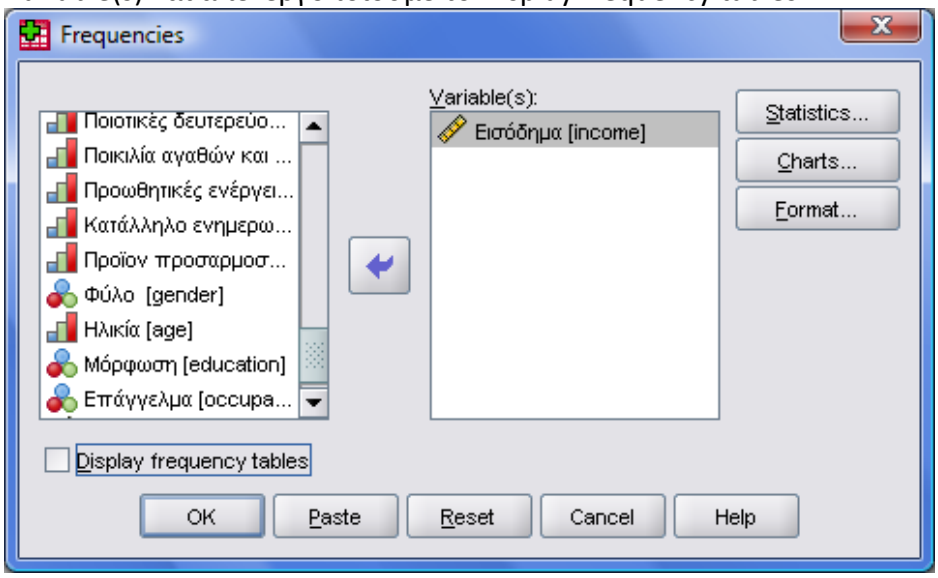
ΠΑΡΑΔΕΙΓΜΑ

Ανοίγουμε το αρχείο Store_spss_lab.sav που περιλαμβάνει δεδομένα από μια έρευνα ικανοποίησης πελατών τεσσάρων καταστημάτων τηλεπικοινωνίας. Ας υποθέσουμε ότι θέλουμε να περιγράψουμε το δείγμα της συγκεκριμένης έρευνας ως προς το

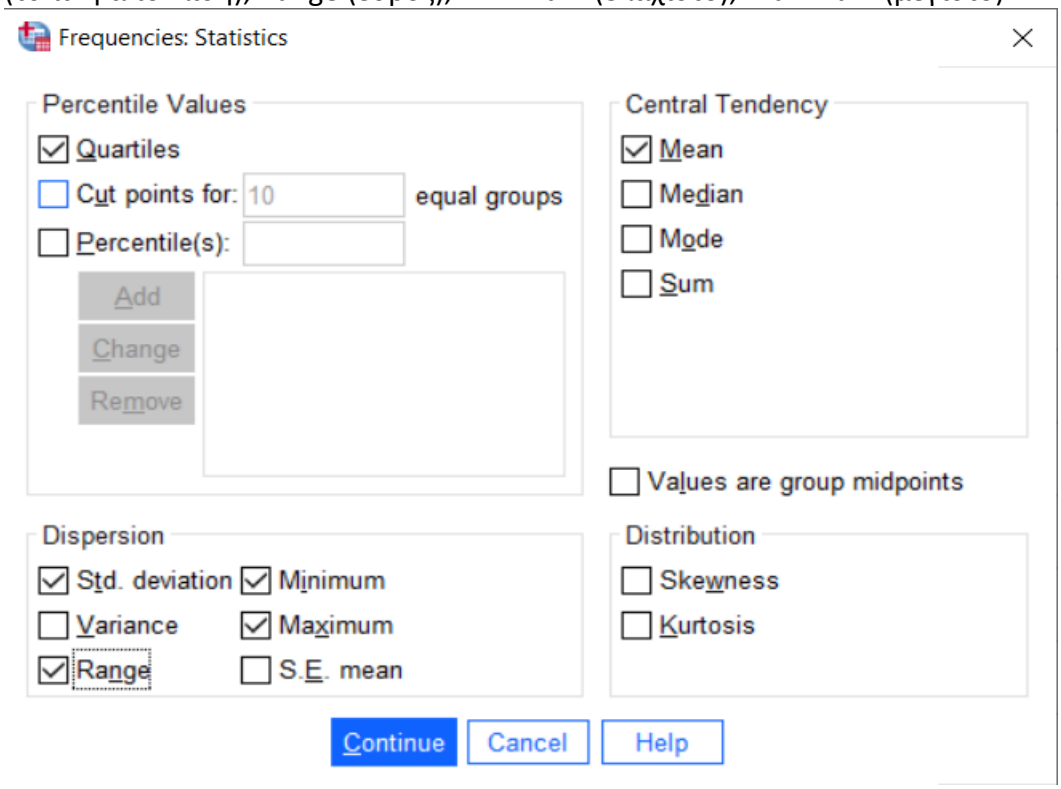
μηνιαίο εισόδημα και συγκεκριμένα, να δούμε πόσο ομοιογενείς ή ανομοιογενείς εισοδηματικά είναι οι πελάτες των καταστημάτων.

Για το σκοπό αυτό ακολουθούμε τα εξής βήματα:
Analyze→ Descriptive Statistics → Frequencies

Στο παράθυρο διαλόγου που ανοίγει μεταφέρουμε τη μεταβλητή income στο πλαίσιο Variable(s) και απενεργοποιούμε το Display frequency tables.



Έπειτα πατώντας στο κουμπί Statistics ανοίγει νέο παράθυρο διαλόγου όπου επιλέγουμε τα εξής: Quartiles (τεταρτημόρια), Mean (μέση τιμή), Std.deviation (τυπική απόκλιση), Range (εύρος), Minimum (ελάχιστο), Maximum (μέγιστο).



Έπειτα πατάμε Continue και OK.

Στο output εμφανίζεται ο παρακάτω πίνακας με τους περιγραφικούς στατιστικούς δείκτες που ζητήσαμε.

Statistics

Εισόδημα

N	Valid	200
	Missing	0
Mean		655,8250
Std. Deviation		474,66958
Range		2500,00
Minimum		,00
Maximum		2500,00
Percentiles	25	318,7500
	50	550,0000
	75	835,0000

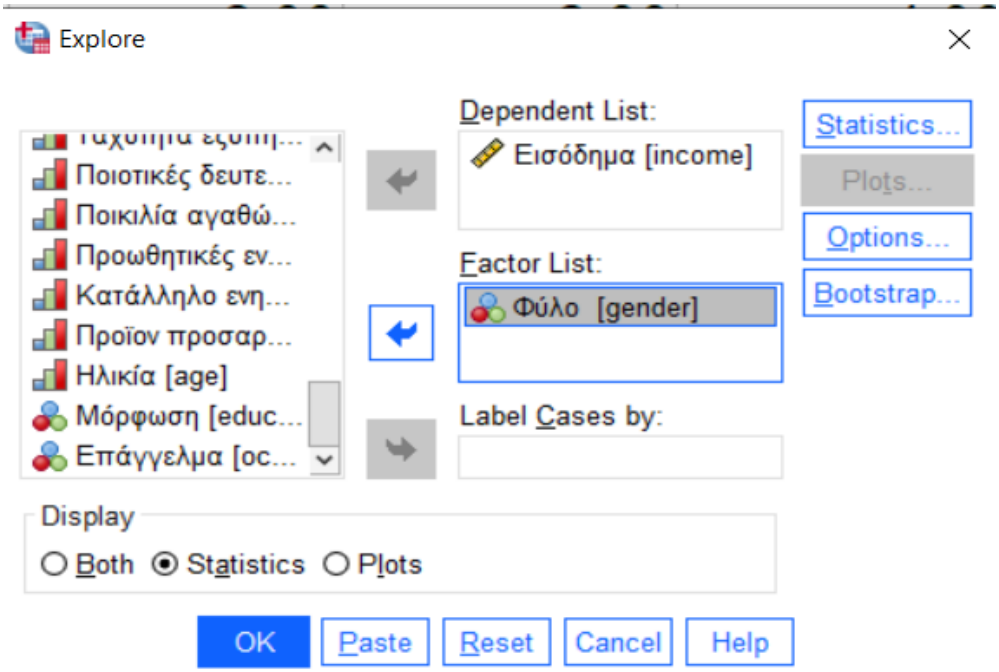
Παρατηρούμε ότι ο μέσος όρος του μηνιαίου εισοδήματος του δείγματος ισούται με 655,8€, ενώ η τυπική απόκλιση ισούται με 474,7€. Αυτό σημαίνει ότι οι περισσότεροι πελάτες στην έρευνα μας έχουν μηνιαίο εισόδημα ίσο με $655{,}8 \pm 474{,}7$ ευρώ. Δηλαδή το μηνιαίο εισόδημα των περισσότερων πελατών ανήκει στο διάστημα [181,1€, 1130,5€] αν θεωρήσουμε ότι οι τιμές του εισοδήματος έχουν κανονική κατανομή (κάτι που θα εξηγηθεί στο επόμενο μάθημα).

Ο ίδιος πίνακας μας δείχνει ότι το εύρος των τιμών (range) της μεταβλητής του εισοδήματος είναι 2500€. Δηλαδή το μέγιστο και το ελάχιστο εισόδημα που κατέγραψε η έρευνα διαφέρουν μεταξύ τους 2500€. Τέλος, κοιτώντας τον παραπάνω πίνακα, παρατηρούμε ότι το 1^{ο} τεταρτημόριο της μεταβλητής του εισοδήματος έχει τιμή 318,8€, ενώ το 3^{ο} τεταρτημόριο έχει τιμή 835€. Επομένως, μπορούμε να υπολογίσουμε το ενδοτεταρτημοριακό εύρος ως εξής: 3^{ο} τεταρτημόριο – 1^{ο} τεταρτημόριο = 835€ - 318,8€ = 516,2€. Συγκρίνοντας το εύρος (range) και το ενδοτεταρτημοριακό εύρος (interquartile range) της μεταβλητής «εισόδημα» καταλαβαίνουμε ότι το δεύτερο είναι πολύ μικρότερο του πρώτου.

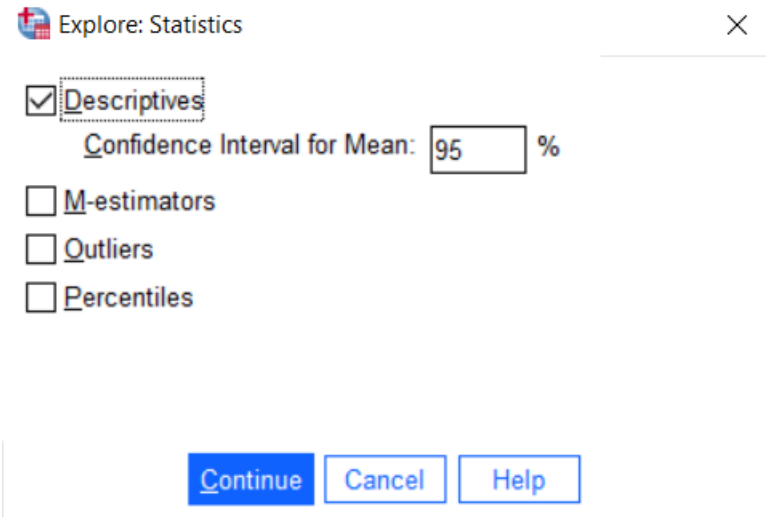
Πολλές φορές οι ερευνητές θέλουν να εξετάσουν τους περιγραφικούς στατιστικούς δείκτες μιας μεταβλητής scale με βάση τις τιμές μιας άλλης μεταβλητής τύπου ordinal ή nominal. Για παράδειγμα, ένας ερευνητής μπορεί να θέλει να δει αν το μηνιαίο εισόδημα των πελατών διαφοροποιείται μεταξύ ανδρών και γυναικών. Για να απαντήσουμε στο ερώτημα αυτό χρησιμοποιούμε την εντολή explore.

Συγκεκριμένα: Analyze→ Descriptive Statistics → Explore.

Στο παράθυρο διαλόγου που ανοίγει μεταφέρω τη μεταβλητή «Εισόδημα» στο πλαίσιο Dependent List ενώ στο πλαίσιο Factor List μεταφέρω τη μεταβλητή «Φύλο». Επίσης, στο πλαίσιο Display επιλέγω Statistics.



Στη συνέχεια, επιλέγω την εντολή Statistics και στο νέο πλαίσιο διαλόγου που εμφανίζεται επιλέγω Descriptives (είναι ήδη προεπιλεγμένο από το SPSS).



Πατάω continue και έπειτα OK. Εμφανίζεται στο output ο παρακάτω πίνακας. Παρατηρούμε στον πίνακα πως εμφανίζονται ξεχωριστοί περιγραφικοί στατιστικοί δείκτες για το κάθε φύλο.

Descriptives

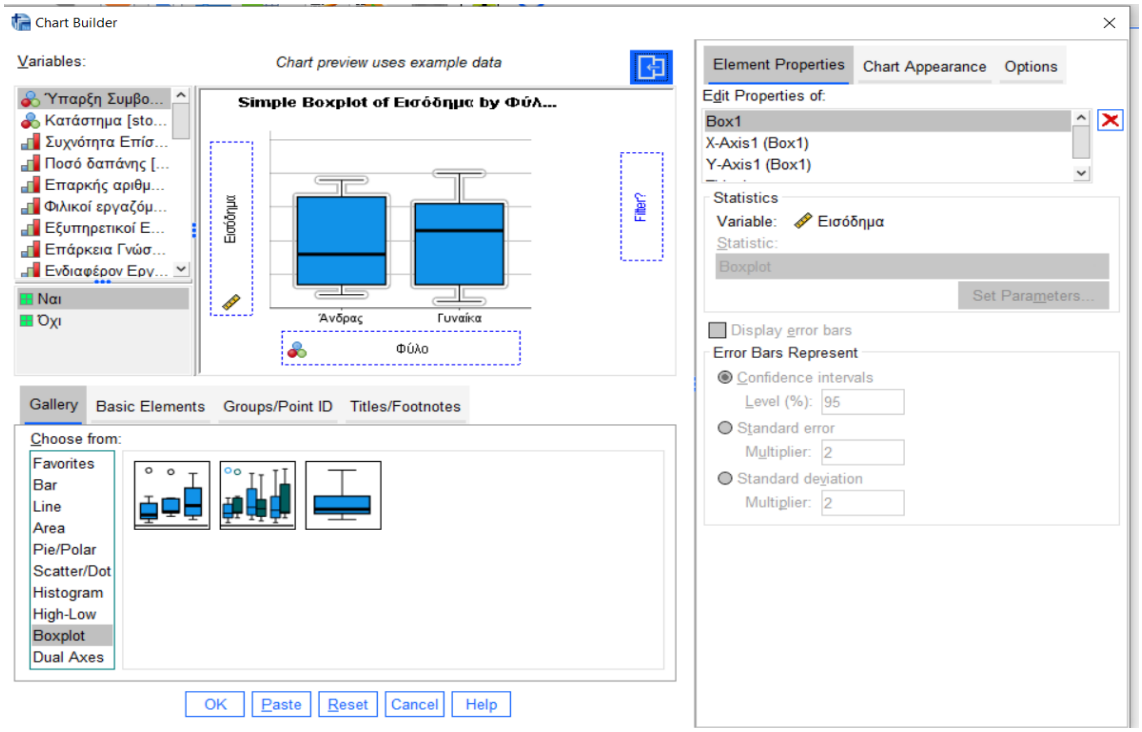
	Φύλο			Statistic	Std. Error
Εισόδημα	Άνδρας	Mean		693,8211	53,09356
		95% Confidence Interval for Mean	Lower Bound	588,4025	
			Upper Bound	799,2396	
		5% Trimmed Mean		654,4298	
		Median		570,0000	
		Variance		267798,000	
		Std. Deviation		517,49203	
		Minimum		9,00	
		Maximum		2500,00	
		Range		2491,00	
		Interquartile Range		652,00	
		Skewness		1,085	,247
		Kurtosis		1,092	,490
	Γυναίκα	Mean		621,4476	42,15213
		95% Confidence Interval for Mean	Lower Bound	537,8584	
			Upper Bound	705,0369	
		5% Trimmed Mean		591,7963	
		Median		535,0000	
		Variance		186564,211	
		Std. Deviation		431,93079	
		Minimum		,00	
		Maximum		1989,00	
		Range		1989,00	
		Interquartile Range		518,00	
		Skewness		,977	,236
		Kurtosis		,778	,467

Για παράδειγμα, η μέση τιμή (mean) του μηνιαίου εισοδήματος για τους άνδρες είναι 693,82€, ενώ για τις γυναίκες είναι 621,44€. Άρα οι γυναίκες πληρώνονται κατά μέσο όρο λιγότερο από τους άντρες. Επίσης, παρατηρούμε ότι η τυπική απόκλιση του εισοδήματος των αντρών είναι 517,49€, ενώ η τυπική απόκλιση του εισοδήματος των γυναικών είναι 431,93€. Με άλλα λόγια, **οι γυναίκες συνιστούν μια περισσότερο ομοιογενή ομάδα όσον αφορά στο εισόδημα τους συγκριτικά με τους άντρες** μεταξύ των οποίων παρατηρούνται μεγαλύτερες εισοδηματικές ανισότητες. Αντίστοιχες συγκρίσεις μεταξύ αντρών και γυναικών μπορούν να γίνουν και για τους υπόλοιπους περιγραφικούς δείκτες, όπως Διάμεσος, Μέγιστη και Ελάχιστη Τιμή, Εύρος και Ενδοτεταρτημοριακό εύρος.

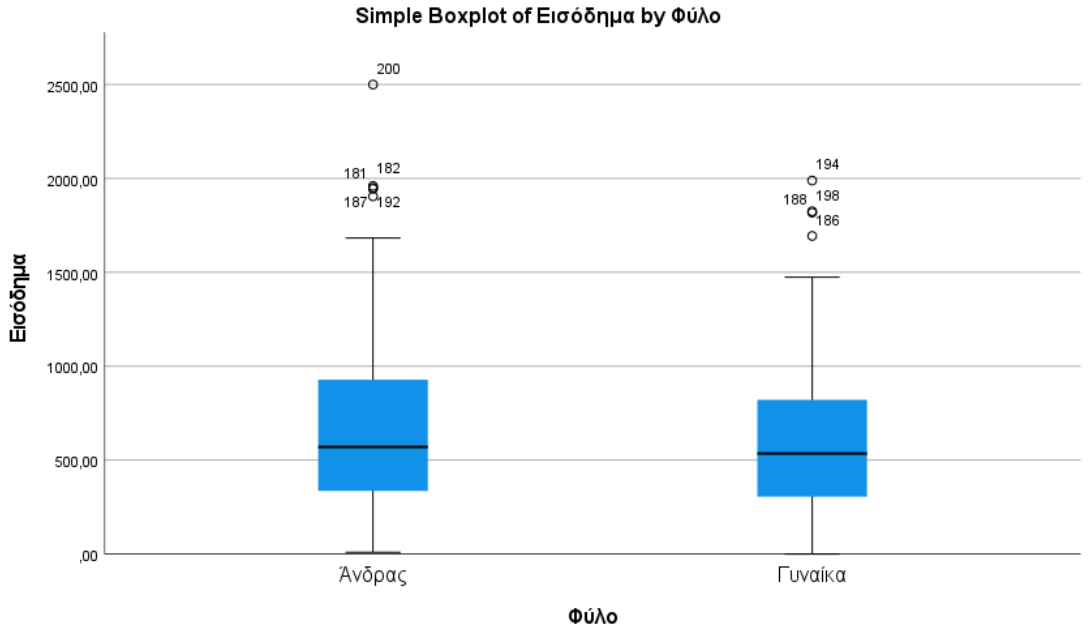
Μπορούμε να συγκρίνουμε τις κατανομές εισοδήματος αντρών και γυναικών κατασκευάζοντας ένα διάγραμμα που είναι γνωστό ως **«Διάγραμμα κουτιού»**

(boxplot). Για να το κάνουμε αυτό χρησιμοποιούμε την εντολή Graphs στο κυρίως μενού του SPSS. Συγκεκριμένα: Graphs→ Chart Builder.

Στο παράθυρο διαλόγου που εμφανίζεται επιλέγουμε Boxplot κάτω από το Gallery και στη συνέχεια επιλέγουμε το Simple Boxplot κάνοντας διπλό κλικ πάνω του. Στο chart preview μεταφέρουμε τη μεταβλητή Εισόδημα στο Y-Axis και τη μεταβλητή Φύλο στο X-Axis.

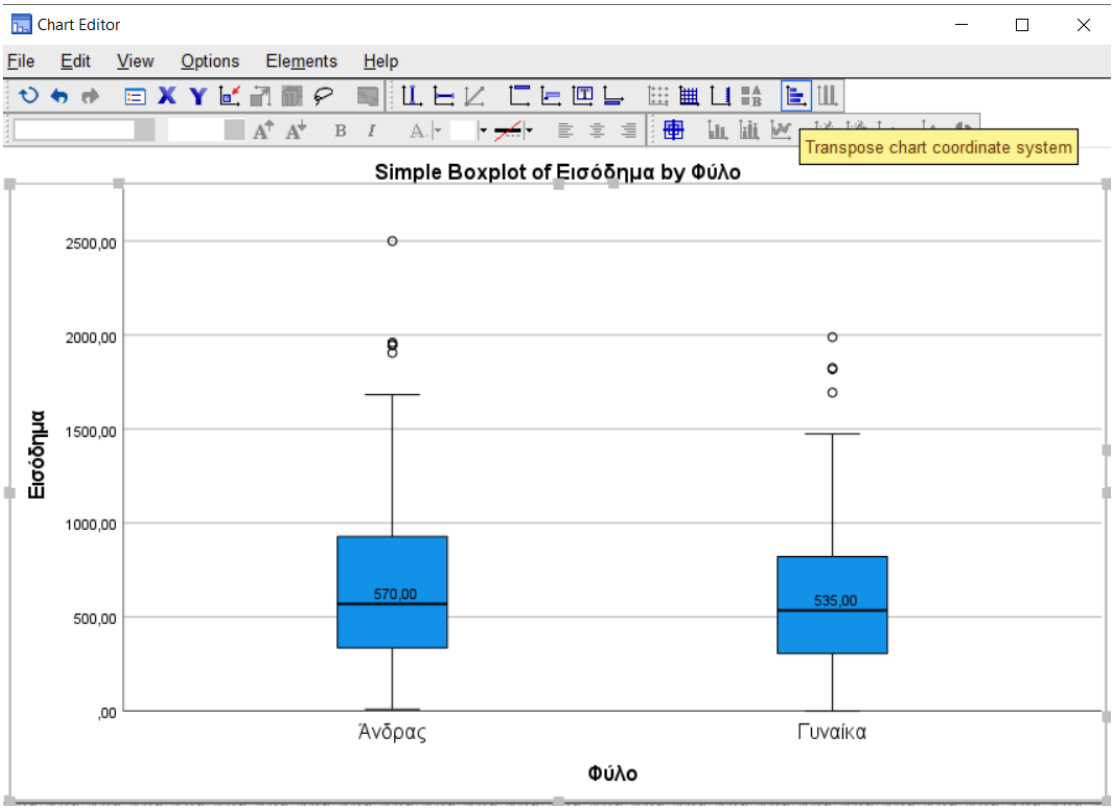


Πατάμε OK και εμφανίζονται τα παρακάτω διαγράμματα.



Εμφανίζονται δύο διαγράμματα, ένα για τους άντρες και ένα για τις γυναίκες. Σε κάθε διάγραμμα, το μπλε κουτί αντιστοιχεί στο κεντρικό 50% της αντίστοιχης κατανομής

εισοδήματος (δηλαδή το ύψος του κουτιού αντιστοιχεί στο ενδοτεταρτημοριακό εύρος). Η μεσαία μαύρη γραμμή είναι η Διάμεσος κάθε κατανομής. Εάν κάνουμε διπλό κλικ πάνω της, στη συνέχεια δεξί κλικ και επιλέξουμε Show data labels, θα εμφανιστεί η ακριβής τιμή της σε κάθε διάγραμμα, όπως φαίνεται παρακάτω.



Παρατηρούμε ότι η Διάμεσος του εισοδήματος των αντρών είναι 570€, ενώ η Διάμεσος των γυναικών είναι χαμηλότερη (535€). Η πάνω πλευρά κάθε κουτιού αντιστοιχεί στο 3ο τεταρτημόριο (Q3) και η κάτω πλευρά στο 1ο τεταρτημόριο (Q1) της αντίστοιχης κατανομής τιμών. Παρατηρούμε ότι τόσο το 1ο όσο και το 3ο τεταρτημόριο της κατανομής του εισοδήματος των αντρών βρίσκονται ψηλότερα από τα αντίστοιχα τεταρτημόρια του εισοδήματος των γυναικών. Η κάτω άκρη του «μουστακιού» κάθε κουτιού ισούται με Q1 – 1,5 x Ενδοτεταρτημοριακό εύρος ή με την ελάχιστη τιμή εάν αυτή είναι μεγαλύτερη. Η πάνω άκρη του μουστακιού ισούται με Q3 + 1,5 x Ενδοτεταρτημοριακό εύρος ή με την μέγιστη τιμή εάν αυτή είναι μικρότερη.

Κοιτώντας τα δύο διαγράμματα κουτιού για άντρες και γυναίκες ξεχωριστά, παρατηρούμε ότι το κεντρικό 50% των τιμών του εισοδήματος των αντρών δεν κατανέμεται τόσο συμμετρικά, κάτι που φαίνεται από τη θέση της Διαμέσου στο μπλε κουτί. Υπάρχει μια ασυμμετρία προς τα πάνω, δηλαδή οι μεγάλες τιμές απέχουν πολύ από τη Διάμεσο η οποία βρίσκεται πιο κοντά στις χαμηλές τιμές εισοδήματος. Με άλλα λόγια, υπάρχει μια μειοψηφία σχετικά εύπορων συμμετεχόντων που το εισόδημα τους διαφέρει δυσανάλογα από των υπολοίπων. Η παρατηρούμενη ασυμμετρία είναι μικρότερη στην ομάδα των γυναικών. Τέλος, υπάρχουν τρεις ακραίες (υψηλές) τιμές τόσο στους άντρες όσο και στις γυναίκες, με εκείνες των αντρών να ξεπερνούν τις αντίστοιχες των γυναικών.