

3

Εισαγωγή στην ανάλυση μεταβλητών με το IBM SPSS Statistics

Στόχοι του κεφαλαίου

- ✓ Εξοικείωση με το περιβάλλον του SPSS
- ✓ Εξοικείωση με τις διαδικασίες περιγραφικής ανάλυσης μιας μεταβλητής
- ✓ Εξοικείωση με τη διαδικασία FREQUENCIES
- ✓ Εξοικείωση με τη διαδικασία DESCRIPTIVES
- ✓ Εξοικείωση με τη διαδικασία EXPLORE
- ✓ Εύρεση βασικών στατιστικών μέτρων
- ✓ Εισαγωγή στην έννοια του στατιστικού ελέγχου

3.1 Γενικά

Στο κεφάλαιο που ακολουθεί θα παρουσιάσουμε κάποιες βασικές μεθόδους Περιγραφικής Στατιστικής που αφορούν στην ανάλυση μιας μεταβλητής. Η Περιγραφική Στατιστική (Descriptive Statistics) περιλαμβάνει όλες τις μεθόδους για τη συνόψιση, περιγραφή και παρουσίαση των δεδομένων. Περιλαμβάνει, δηλαδή, όλες τις απαραίτητες τεχνικές για την οργάνωση των δεδομένων και τη συνοπτική και αποτελεσματική παρουσίασή τους. Στην επόμενη παράγραφο θα αναλύσουμε κάποιες μεθόδους της Περιγραφικής Στατιστικής για την ανάλυση μίας ποσοτικής ή ποιοτικής μεταβλητής. Θα ξεκινήσουμε με την ανάλυση μιας ποιοτικής μεταβλητής.

3.2 Η διαδικασία FREQUENCIES

3.2.1 Περιγραφική ανάλυση μιας ποιοτικής μεταβλητής

Οι μέθοδοι οργάνωσης και παρουσίασης ποιοτικών δεδομένων αφορούν κυρίως στους πίνακες συχνотήτων και σε γραφήματα που παράγονται από τη διαδικασία Frequencies. Θα χρησιμοποιήσουμε τα δεδομένα του Παραδείγματος 3.1 για να δείξουμε πώς χρησιμοποιείται αυτή η διαδικασία όταν θέλουμε να αναλύσουμε μία ποιοτική μεταβλητή.

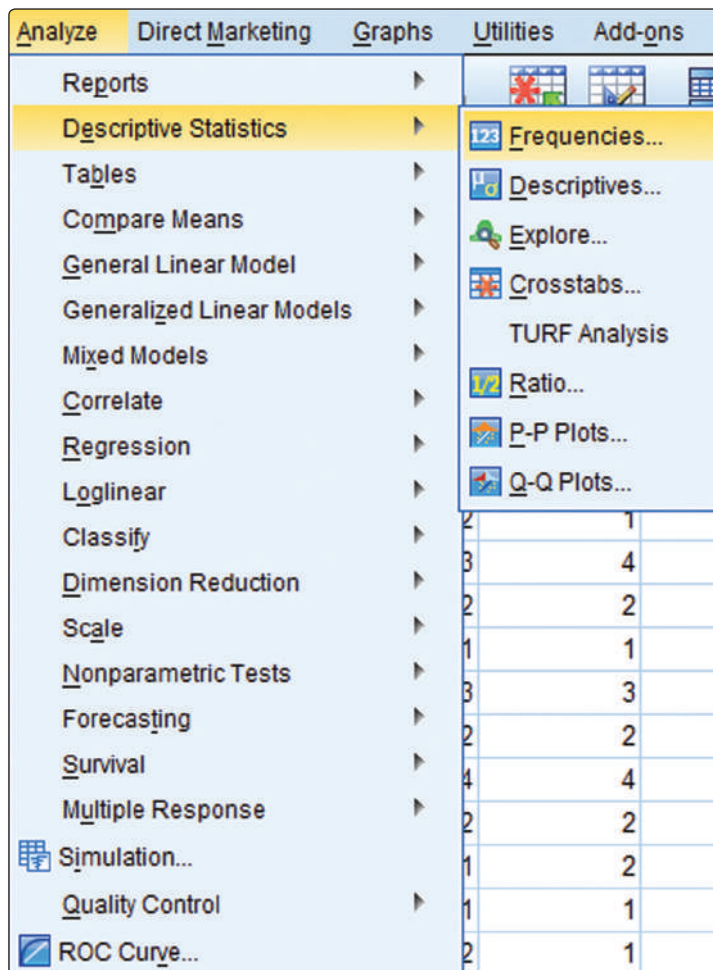


Παράδειγμα 3.1

Στο αρχείο Survey1.sav περιέχονται δεδομένα που αφορούν στο εκπαιδευτικό επίπεδο των ερωτώμενων. Θα χρησιμοποιήσουμε τα δεδομένα του αρχείου και τη διαδικασία Frequencies για να αναλύσουμε την ποιοτική μεταβλητή level of education.

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

Analyze → Descriptive Statistics → Frequencies...

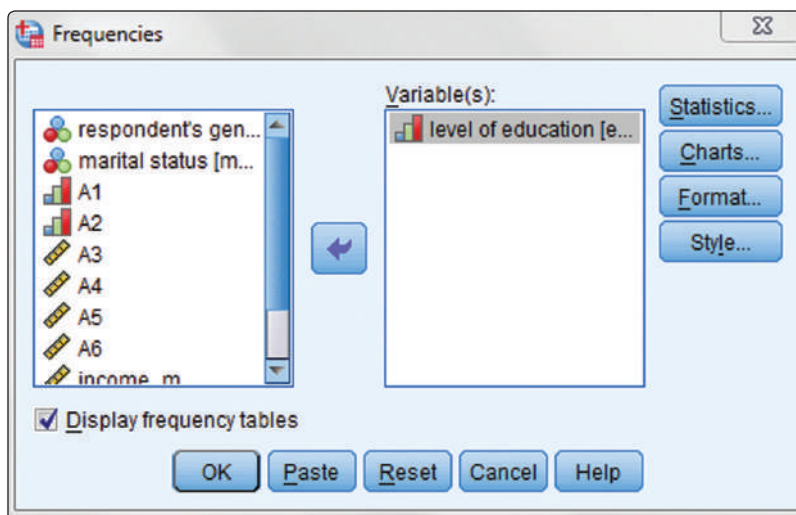


Εικόνα 3.1 Η διαδικασία Frequencies

Από τη λίστα των μεταβλητών που εμφανίζεται αριστερά στο παράθυρο διαλόγου Frequencies επιλέγουμε τη μεταβλητή level of education και τη μετακινούμε στο παράθυρο Variable(s) (Εικόνα 3.2). Στο ίδιο παράθυρο παρατηρούμε την ύπαρξη του πλήκτρου Statistics... που χρησιμοποιούμε εκτενώς, όταν θέλουμε να κάνουμε κάποιες αλλαγές στις προεπιλεγμένες ρυθμίσεις κατά την εκτέλεση των στατιστικών μας αναλύσεων. Το Charts... επιλέγεται όταν θέλουμε να παραχθεί κάποιο επιπλέον γράφημα.

Κεφάλαιο Τρίτο

Εισαγωγή στην ανάλυση μεταβλητών με το IBM SPSS Statistics



Εικόνα 3.2 Το παράθυρο διαλόγου 'Frequencies'

Εξ ορισμού, η διαδικασία θα υπολογίσει μόνο τον πίνακα συχνοτήτων και τα αποτελέσματα αυτής της διαδικασίας που θα εμφανιστούν στον Output Viewer φαίνονται στην Εικόνα 3.3.

Statistics

level of education

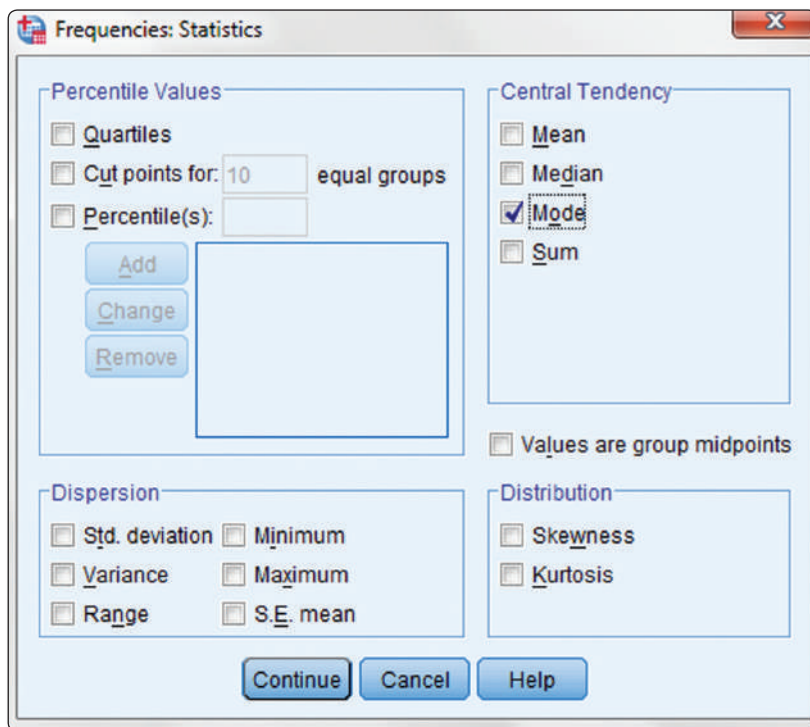
N	Valid	30
	Missing	0

level of education

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Low	10	33.3	33.3	33.3
	Medium	8	26.7	26.7	60.0
	High	12	40.0	40.0	100.0
	Total	30	100.0	100.0	

Εικόνα 3.3 Πίνακας συχνοτήτων για τη μεταβλητή level of education

Παρατηρούμε ότι έχουμε 30 παρατηρήσεις και καμία ελλείπουσα τιμή, ενώ 12/30 ερωτώμενους π.χ. έχουν υψηλό μορφωτικό επίπεδο (8/30 έχουν μεσαίο και 10/30 χαμηλό). Η διαδικασία δεν θα υπολογίσει κανένα στατιστικό μέτρο και ούτως ή άλλως σε αυτήν την περίπτωση μόνο η εύρεση της επικρατούσας τιμής θα είχε νόημα. Αν θέλουμε να υπολογιστεί, θα ενεργοποιήσουμε την επιλογή **Statistics...** στην Εικόνα 3.2. Εκεί θα επιλέξουμε το mode (επικρατούσα τιμή) (Εικόνα 3.4) και θα πατήσουμε Continue:

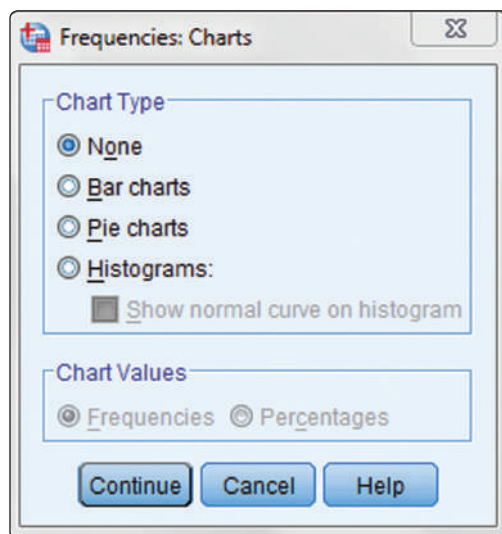


Εικόνα 3.4 Παράθυρο διαλόγου: 'Frequencies: Statistics'

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

Κεφάλαιο Τρίτο

Εισαγωγή στην ανάλυση μεταβλητών με το IBM SPSS Statistics



Εικόνα 3.5 Παράθυρο διαλόγου: 'Frequencies: Statistics'

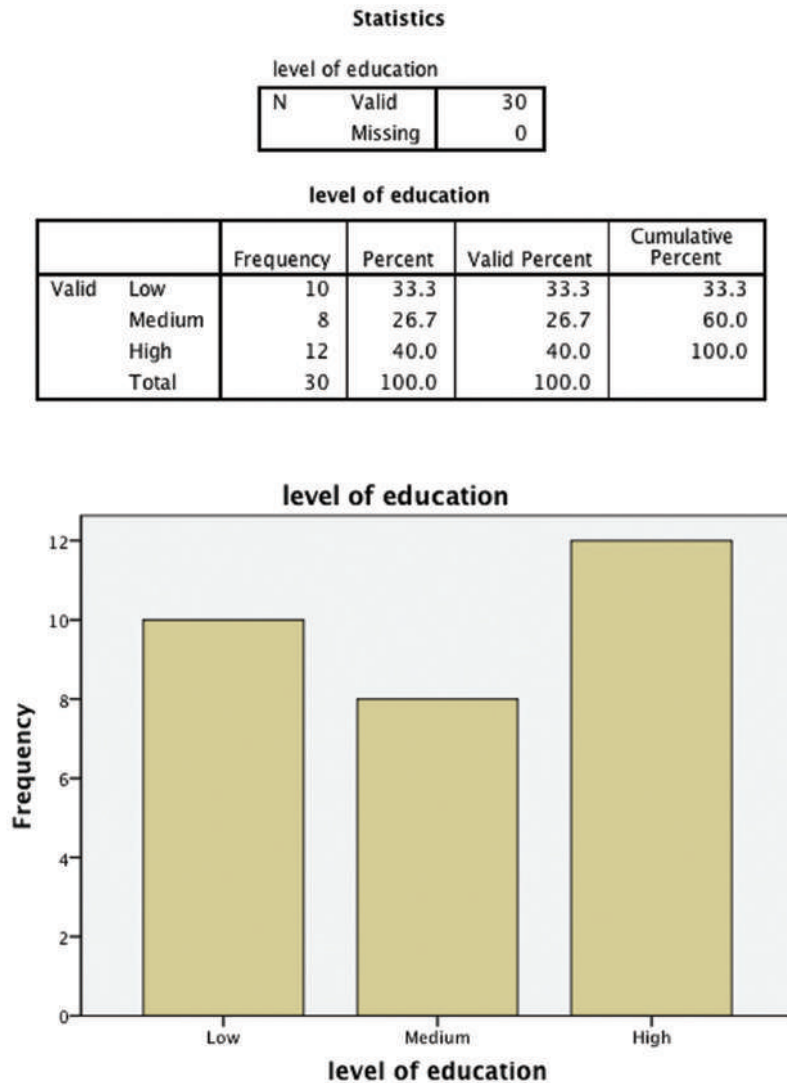
Επιλέγουμε το Bar Charts, πατάμε Continue και OK στο παράθυρο Frequencies. Το αποτέλεσμα των επιλογών μας φαίνεται στην Εικόνα 3.6. Ο Πίνακας συχνοτήτων παραμένει ο ίδιος με τον Πίνακα της Εικόνας 3.3. Παρατηρούμε ότι η επικρατούσα τιμή είναι το 3 (που αντιστοιχεί στο ανώτερο εκπαιδευτικό επίπεδο), γεγονός που φυσικά συμβαδίζει με τα αποτελέσματα του πίνακα συχνοτήτων και το ραβδόγραμμα.



Ερώτηση

- ✓ Πόσο είναι το ποσοστό των ερωτώμενων που έχει εκπαιδευτικό επίπεδο μεσαίο ή χαμηλό;
- ✓ Πόσο είναι το ποσοστό των ερωτώμενων που έχει εκπαιδευτικό επίπεδο υψηλό;

Η απάντηση στην πρώτη ερώτηση είναι 60% (φαίνεται από το αθροιστικό ποσοστό στην τελευταία στήλη) και η απάντηση στη δεύτερη είναι 40% (Πίνακας level of education, Εικόνα 3.6).

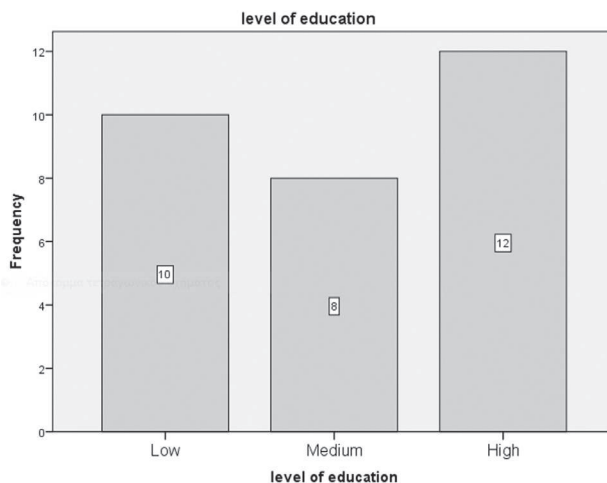


Εικόνα 3.6 Αποτέλεσμα της διαδικασίας Frequencies

Στη συνέχεια μπορούμε φυσικά να μορφοποιήσουμε το διάγραμμα μας, εφ' όσον το επιθυμούμε. Για παράδειγμα, μπορούμε να αλλάξουμε χρώμα και να προσθέσουμε Value Labels στις ράβδους (Εικόνα 3.7):

Κεφάλαιο Τρίτο

Εισαγωγή στην ανάλυση μεταβλητών με το IBM SPSS Statistics



Εικόνα 3.7 Μορφοποιημένο ραβδόγραμμα συχνοτήτων του μορφωτικού επιπέδου των ερωτώμενων

3.2.2 Περιγραφική ανάλυση μιας ποσοτικής μεταβλητής

Για τον υπολογισμό των αριθμητικών περιγραφικών μέτρων μιας ποσοτικής μεταβλητής χρησιμοποιούμε τις διαδικασίες Descriptives, Frequencies και Explore. Κάποια από τα πιο βασικά στατιστικά περιγραφικά μέτρα που μπορούν να υπολογιστούν είναι:

1. Η **μέση τιμή** (mean) των παρατηρήσεων του δείγματος που είναι το άθροισμα όλων των παρατηρήσεων διαιρεμένο δια του συνολικού αριθμού τους.
2. Η **διάμεσος** (median) του δείγματος. Η διάμεσος όπως γνωρίζουμε χωρίζει τη διατεταγμένη σειρά των παρατηρήσεων σε δύο μέρη, έτσι ώστε τα μισά περίπου δεδομένα να βρίσκονται στην αριστερή πλευρά και τα άλλα μισά στη δεξιά.
3. Η **επικρατούσα τιμή** (mode), δηλαδή, η πιο συχνά εμφανιζόμενη τιμή στο δείγμα.
4. Η **τυπική απόκλιση** του δείγματός μας που χρησιμοποιείται για να μετράει τη μεταβλητότητα του δείγματός μας.
5. Η **μέγιστη** (maximum) και **ελάχιστη** (minimum) τιμή των παρατηρήσεών μας.

3.3 Η διαδικασία DESCRIPTIVES

Η διαδικασία Descriptives βοηθά στον υπολογισμό των στατιστικών μέτρων μίας ή και περισσότερων ποσοτικών μεταβλητών. Για κάθε μεταβλητή, η διαδικασία αυτή μπορεί να υπολογίσει: τη μέση τιμή, το τυπικό σφάλμα, τη διάμεσο, την επικρατούσα τιμή, την τυπική απόκλιση, τη διακύμανση, την κύρτωση, τη λοξότητα, την ελάχιστη και τη μέγιστη τιμή, κ.ά.

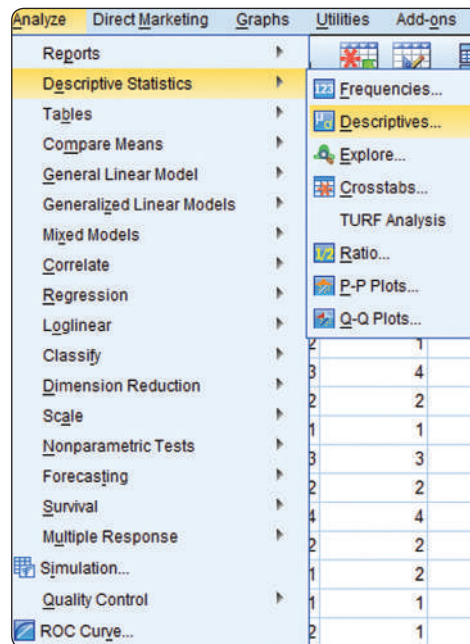


Παράδειγμα 3.2

Θα χρησιμοποιήσουμε τα δεδομένα του *Survey1.sav* και τη διαδικασία *Descriptives* για την ανάλυση της μεταβλητής που αφορά την ηλικία των ερωτώμενων.

Από τη βασική γραμμή εντολών επιλέγουμε (Εικόνα 3.8):

Analyze → Descriptive Statistics → Descriptives...

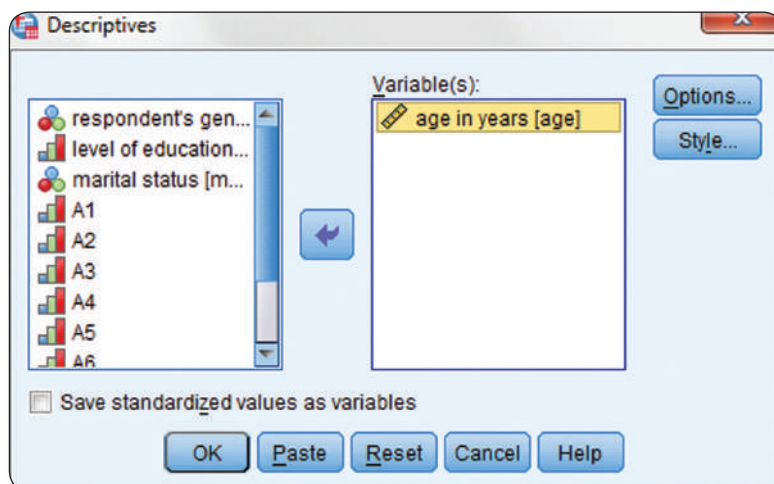


Εικόνα 3.8 διαδικασία Descriptives

Κεφάλαιο Τρίτο

Εισαγωγή στην ανάλυση μεταβλητών με το IBM SPSS Statistics

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



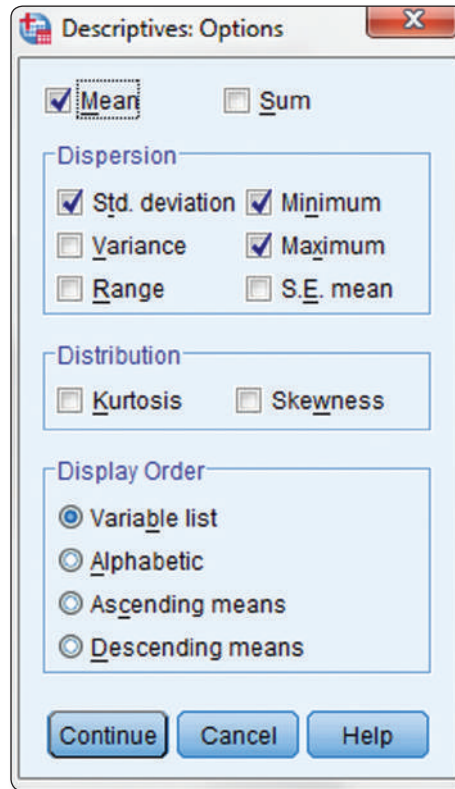
Εικόνα 3.9 Το παράθυρο διαλόγου: 'Descriptives'

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

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
age in years	30	18	71	39,77	18,343
Valid N (listwise)	30				

Εικόνα 3.10 Αποτελέσματα της διαδικασίας Descriptives

Παρατηρούμε ότι η διαδικασία εξ ορισμού υπολογίζει τα εξής στατιστικά μέτρα: μέση τιμή, τυπική απόκλιση, μέγιστη και ελάχιστη τιμή των παρατηρήσεων. Αν θέλουμε να βρούμε και τις τιμές άλλων στατιστικών μέτρων, μπορούμε να το πραγματοποιήσουμε με την επιλογή Options, όπως επίσης και να καθορίσουμε τη σειρά εμφάνισης τους από το Display Order (Εικόνα 3.11).

**Εικόνα 3.11** Το παράθυρο διαλόγου: 'Descriptives: Options'

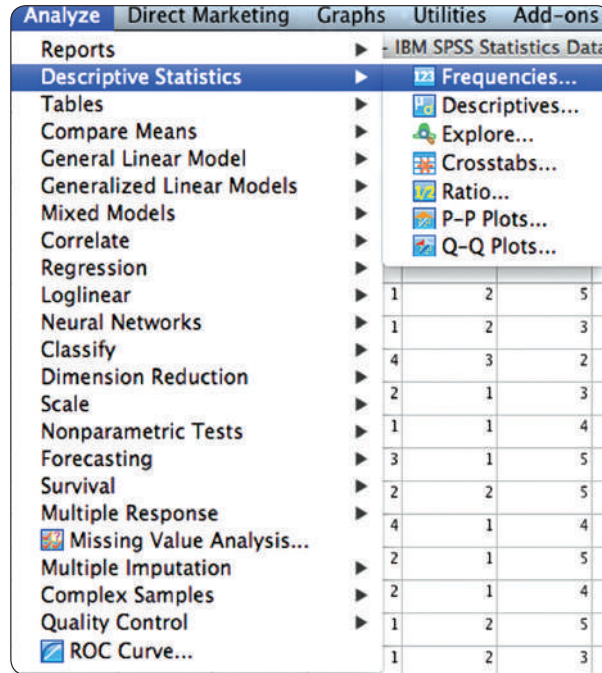
Η διαδικασία Descriptives είναι, εν γένει, μια φτωχή διαδικασία, η οποία περιορίζεται στην παρουσίαση των στατιστικών μέτρων μίας ή περισσότερων μεταβλητών σε έναν πίνακα. Αν θέλουμε μια πιο εξειδικευμένη παρουσίαση των δεδομένων μας αυτό μπορεί να γίνει με τη βοήθεια της διαδικασίας Frequencies.

Θα χρησιμοποιήσουμε τα δεδομένα του Survey1.sav και τη διαδικασία Frequencies για την ανάλυση της μεταβλητής age που αφορά στην ηλικία των ερωτώμενων. Επιλέγουμε διαδοχικά από τη γραμμή των εντολών (Εικόνα 3.12):

Analyze → Descriptive Statistics → Frequencies ...

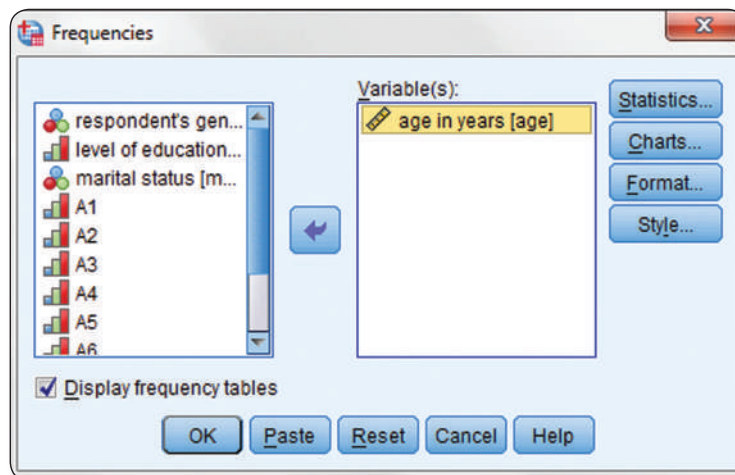
Κεφάλαιο Τρίτο

Εισαγωγή στην ανάλυση μεταβλητών με το IBM SPSS Statistics

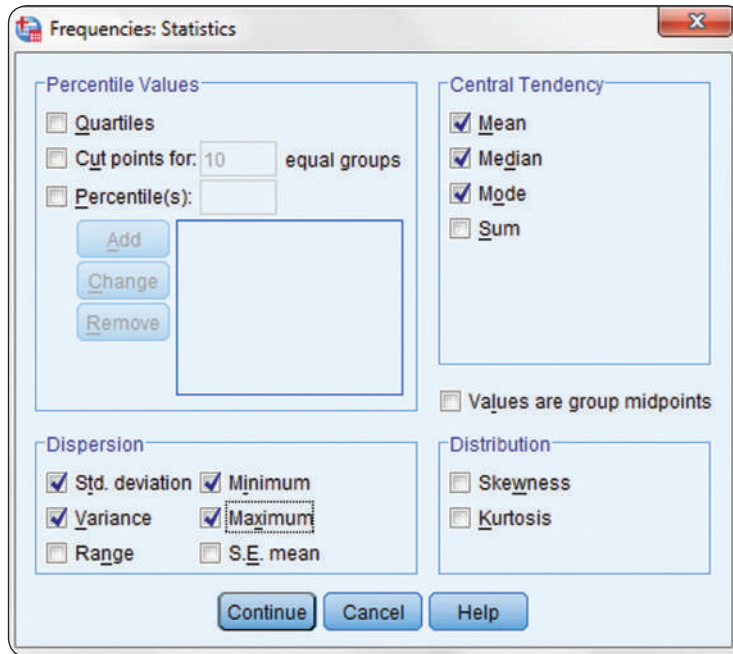


Εικόνα 3.12 Επιλογή της διαδικασίας Frequencies

Επιλέγοντας στο παρακάτω παράθυρο (Εικόνα 3.13) το Statistics έχουμε πολλές περισσότερες επιλογές (Εικόνα 3.14):

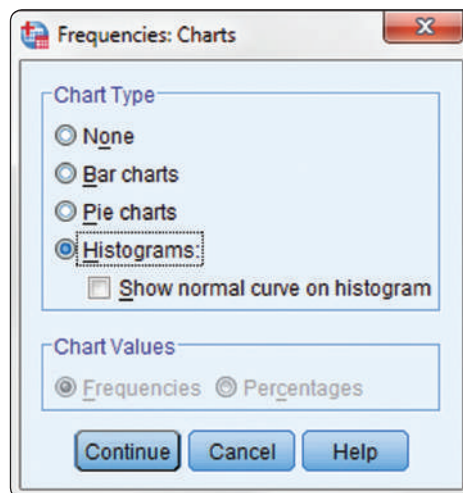


Εικόνα 3.13 Το παράθυρο διαλόγου: 'Frequencies'



Εικόνα 3.14 Το παράθυρο διαλόγου: 'Frequencies: Statistics'

Επιπλέον επιλογές έχουμε, αν κάνουμε Click στο Charts (Εικόνα 3.15). Συνεχίζουμε με το Continue.



Εικόνα 3.15 Το παράθυρο διαλόγου: 'Frequencies: Charts'

Κεφάλαιο Τρίτο

Εισαγωγή στην ανάλυση μεταβλητών με το IBM SPSS Statistics

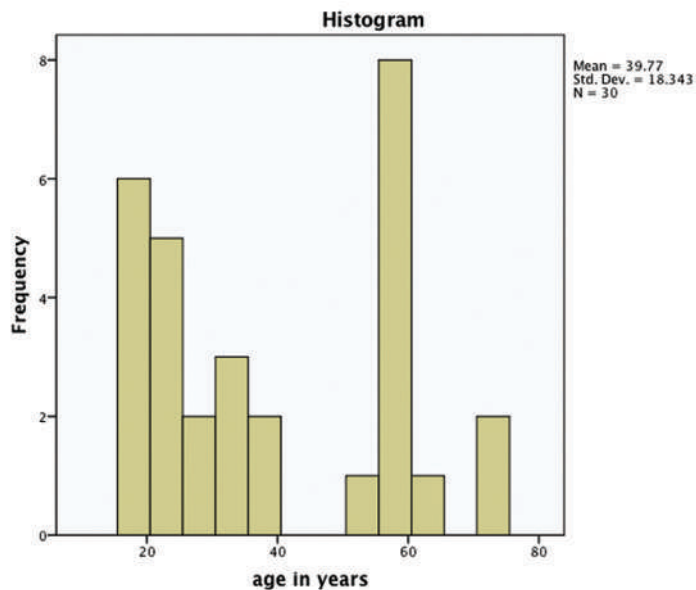
Τα αποτελέσματα που θα πάρουμε στον Output Viewer φαίνονται στις Εικόνες 3.16 και 3.17.

Statistics

age in years		
N	Valid	30
	Missing	0
Mean		39.77
Median		33.50
Mode		25 ^a
Std. Deviation		18.343
Variance		336.461
Minimum		18
Maximum		71

a. Multiple modes exist. The smallest value is shown

Εικόνα 3.16 Στατιστικά μέτρα για την μεταβλητή ηλικία



Εικόνα 3.17 Ιστόγραμμα συχνοτήτων της μεταβλητής ηλικία

3.4 Η διαδικασία EXPLORE

Με τη διαδικασία Explore μπορούμε να πετύχουμε την πιο πλούσια και πλήρη περιγραφή των παρατηρήσεων μιας ποσοτικής μεταβλητής (πλουσιότερη από αυτήν που πετυχαίνουμε με τη διαδικασία Descriptives και Frequencies).

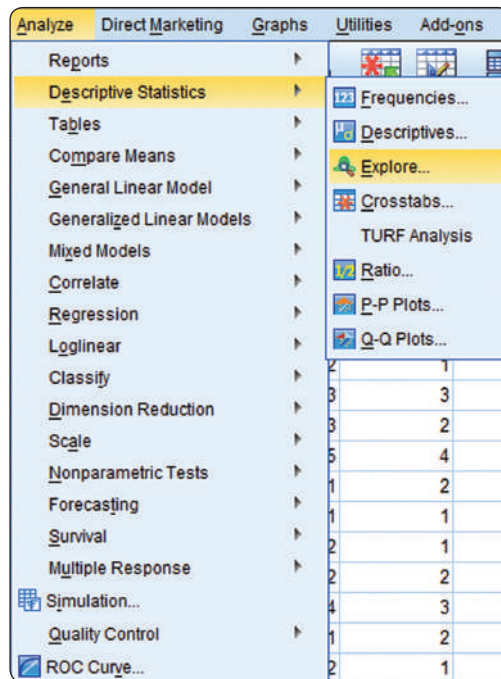


Παράδειγμα 3.3

Θα χρησιμοποιήσουμε τα δεδομένα του αρχείου Grades2.sav και τη διαδικασία Explore για την ανάλυση της ποσοτικής μεταβλητής που αφορά στη βαθμολογία του Απολυτηρίου Λυκείου των ερωτώμενων.

Επιλέγουμε διαδοχικά από το μενού των εντολών (Εικόνα 3.18):

Analyze → Descriptive Statistics → Explore...

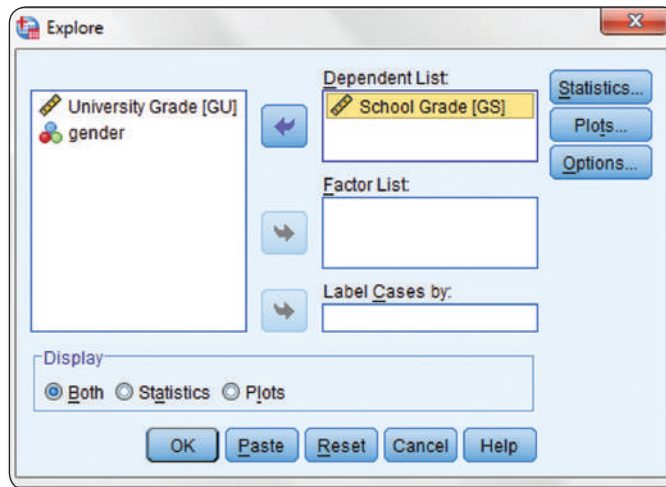


Εικόνα 3.18 Η διαδικασία Explore

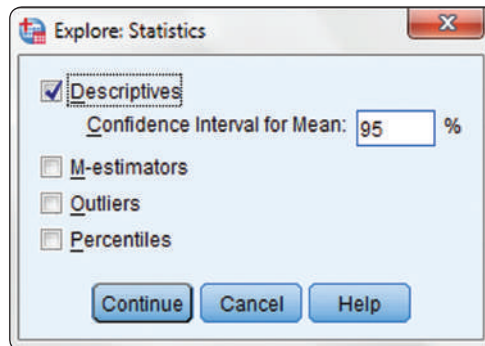
Κεφάλαιο Τρίτο

Εισαγωγή στην ανάλυση μεταβλητών με το IBM SPSS Statistics

Εισάγουμε τη μεταβλητή School Grade [GS] στο πλαίσιο Dependent List (Εικόνα 3.19) και στη συνέχεια πατάμε το Statistics, αν θέλουμε να πάρουμε και άλλες πληροφορίες εκτός από αυτές που δίνει εξ ορισμού η διαδικασία (Εικόνα 3.20).

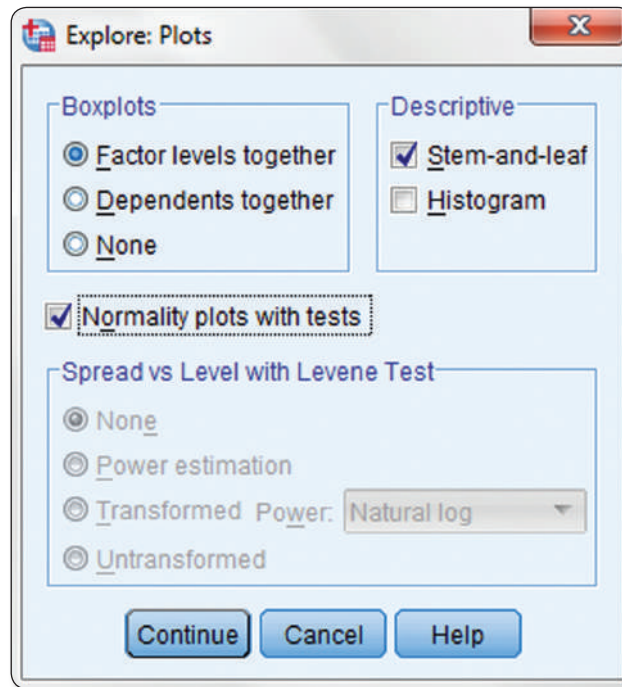


Εικόνα 3.19 Το παράθυρο διαλόγου: 'Explore'



Εικόνα 3.20 Το παράθυρο διαλόγου: 'Explore: Statistics'

Συνεχίζουμε με το Continue. Πατώντας το Plots μπορούμε να επιλέξουμε και κάποιο άλλο γράφημα, π.χ. ιστόγραμμα ή να κάνουμε έναν έλεγχο προσαρμογής των δεδομένων μας στην κανονική κατανομή (Εικόνα 3.21). Συνεχίζουμε με το Continue και στο παράθυρο διαλόγου Explore πατάμε OK.



Εικόνα 3.21 Το παράθυρο διαλόγου: 'Explore: Plots'

Τα αποτελέσματα που θα πάρουμε στον Output Viewer φαίνονται παρακάτω στις Εικόνες 3.22, 3.23. Στο παράθυρο Descriptives εμφανίζεται εκτός των άλλων ένα 95% - διάστημα εμπιστοσύνης για τη μέση τιμή. Δηλαδή, μπορούμε να πούμε με βεβαιότητα 95% ότι η πληθυσμιακή μέση τιμή για τη μεταβλητή ηλικία βρίσκεται στο διάστημα (40.94, 47.30). Δίνεται επίσης η 5% - ισοσταθμισμένη μέση τιμή.

Γενικά, η $n\%$ - ισοσταθμισμένη μέση τιμή ($n\%$ - trimmed mean), ορίζεται ως η μέση τιμή που υπολογίζεται όταν οι $n\%$ μεγαλύτερες και οι $n\%$ μικρότερες τιμές έχουν διαγραφεί. Στο SPSS το προεπιλεγμένο είναι το $n=5$.

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

Κεφάλαιο Τρίτο

Εισαγωγή στην ανάλυση μεταβλητών με το IBM SPSS Statistics

1. Αν η παρατηρούμενη συνάρτηση κατανομής μιας μεταβλητής συμπίπτει με κάποια από τις γνωστές θεωρητικές Poisson, Normal (Κανονική), κ.λπ.
2. Αν η σειρά εμφάνισης δύο τιμών μιας μεταβλητής είναι τυχαία (Runs test).
3. Αν οι παρατηρούμενες συχνότητες των κατηγοριών μιας μεταβλητής απέχουν ή όχι από τις θεωρητικές αναμενόμενες μιας γνωστής κατανομής (χ^2 -test καλής προσαρμογής), κ.ά.

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

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
School Grade	60	100.0%	0	0.0%	60	100.0%

Descriptives

			Statistic	Std. Error
School Grade	Mean		18.413	.1077
	95% Confidence Interval for Mean	Lower Bound	18.198	
		Upper Bound	18.629	
	5% Trimmed Mean		18.424	
	Median		18.400	
	Variance		.696	
	Std. Deviation		.8345	
	Minimum		16.3	
	Maximum		20.0	
	Range		3.7	
	Interquartile Range		1.2	
	Skewness		-.180	.309
	Kurtosis		-.440	.608

Tests of Normality

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
School Grade	.084	60	.200*	.982	60	.524

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

a. Lilliefors Significance Correction

Εικόνα 3.22 Αποτελέσματα της διαδικασίας Explore

Ένας τέτοιος έλεγχος προσαρμογής των δεδομένων μας στην κανονική κατανομή (Tests of Normality), ζητήθηκε εδώ μέσω του Normality tests with plots. Στο σημείο αυτό είναι σημαντικό να αναφέρουμε κάποιες βασικές πληροφορίες σχετικά με τους στατιστικούς ελέγχους, όπως αυτοί εκτελούνται από το SPSS, και που είναι απαραίτητες για τα επόμενα κεφάλαια. Όπως είναι γνωστό σε κάθε στατιστικό έλεγχο έχουμε μία σημαντική υπόθεση, τη μηδενική υπόθεση H_0 και το ενδιαφέρον μας εστιάζεται στο αν μπορούμε, με βάση τα δεδομένα μας να αποδεχτούμε ότι η υπόθεση αυτή ισχύει (ή να ισχυριστούμε ότι δεν έχουμε αρκετά στοιχεία για να την απορρίψουμε). Μία, λοιπόν, σημαντική τιμή που χρησιμοποιούμε στους στατιστικούς ελέγχους είναι το p -value ενός ελέγχου. Γενικά ισχύει ότι αν το p -value ενός ελέγχου είναι μικρότερο από το επίπεδο σημαντικότητας α , τότε απορρίπτουμε την υπόθεση H_0 , ενώ αν είναι μεγαλύτερο την αποδεχόμαστε.

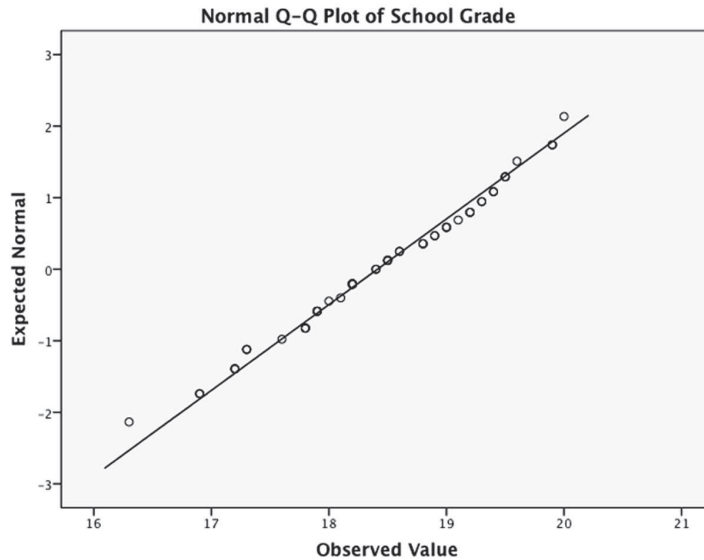
Στη συγκεκριμένη περίπτωση, αν το δείγμα μας είναι μικρότερο του 50, τότε θα κοιτάξουμε στην Εικόνα 3.22 τον στατιστικό έλεγχο Shapiro-Wilk. Αν όχι, τότε θα κοιτάξουμε το στατιστικό κριτήριο Kolmogorov-Smirnov. Για να αποφασίσουμε αν ισχύει η μηδενική μας υπόθεση, δηλαδή ότι το δείγμα προέρχεται από πληθυσμό που ακολουθεί την κανονική κατανομή, εξετάζουμε την τιμή στη στήλη Sig. του κριτηρίου Kolmogorov-Smirnov, δεδομένου ότι το δείγμα μας είναι μεγέθους μεγαλύτερου από 50. Αν αυτή είναι μικρότερη από το 0.05, τότε απορρίπτουμε τη μηδενική υπόθεση, αν όχι, τότε την αποδεχόμαστε. Εδώ, η τιμή είναι ίση με $0.200 > 0.05$, οπότε αποδεχόμαστε τη μηδενική υπόθεση. Μπορούμε, λοιπόν, να ισχυριστούμε με βεβαιότητα 95%, ότι το δείγμα προέρχεται από πληθυσμό που ακολουθεί την κανονική κατανομή.

Δίνεται επίσης ένα Q-Q plot (Εικόνα 3.23), το οποίο ερμηνεύεται με τον ίδιο τρόπο όπως και το P-P plot που περιγράφηκε στο Κεφάλαιο 2. Τα αντίστοιχα σημεία για τη μεταβλητή μας δεν αποκλίνουν αισθητά από την ευθεία. Δίνονται επιπλέον ένα φυλλόγραμμα (stem and leaf plot) (Εικόνα 3.24) και ένα θηκόγραμμα (Εικόνα 3.25). Στο φυλλόγραμμα αναγράφονται όλες οι παρατηρήσεις που έχουμε, όλες οι τιμές δηλαδή της μεταβλητής ηλικία. Η στήλη που βρίσκεται αριστερά της κάθετης γραμμής είναι γνωστή ως κορμός, ενώ οι άλλοι αριθμοί δεξιά της γραμμής είναι τα φύλλα. Ο κορμός αντιπροσωπεύει τις 'δεκάδες' και τα φύλλα τις 'μονάδες'. Για παράδειγμα, η πρώτη γραμμή που αρχίζει με το '16' έχει ως φύλλα το '3'. Επομένως, ξέρουμε ότι η 1^η παρατήρηση είναι το 16.3.

Κεφάλαιο Τρίτο

Εισαγωγή στην ανάλυση μεταβλητών με το IBM SPSS Statistics

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



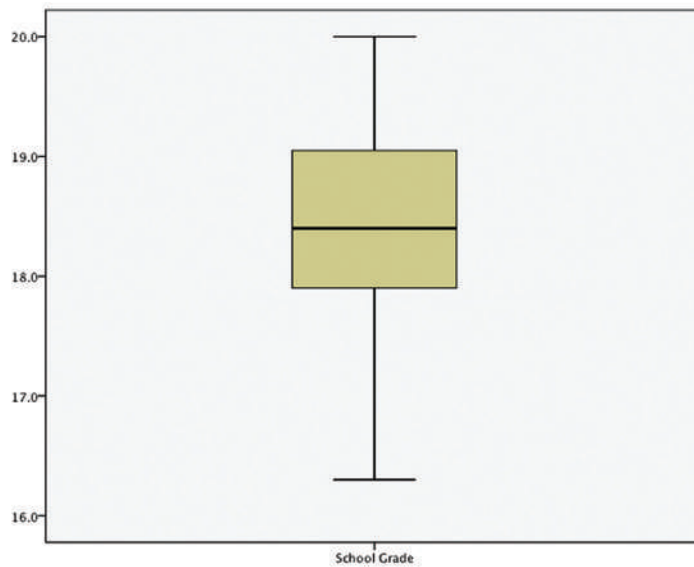
Εικόνα 3.23 Q-Q plot

School Grade

School Grade Stem-and-Leaf Plot

Frequency	Stem &	Leaf
1.00	16 .	3
2.00	16 .	99
6.00	17 .	222333
10.00	17 .	6888899999
12.00	18 .	01222222244
11.00	18 .	55556688899
11.00	19 .	00012223344
6.00	19 .	555699
1.00	20 .	0
Stem width: 1.0		
Each leaf: 1 case(s)		

Εικόνα 3.24 Φυλλόγραμμα της μεταβλητής School Grade



Εικόνα 3.25 Θηκόγραμμα της μεταβλητής School Grade

Θα ασχοληθούμε στη συνέχεια με τον στατιστικό έλεγχο της τυχειότητας της εμφάνισης των παρατηρήσεων του δείγματος.

3.5 Έλεγχος τυχαιότητας της εμφάνισης των παρατηρήσεων του δείγματος (Runs test)

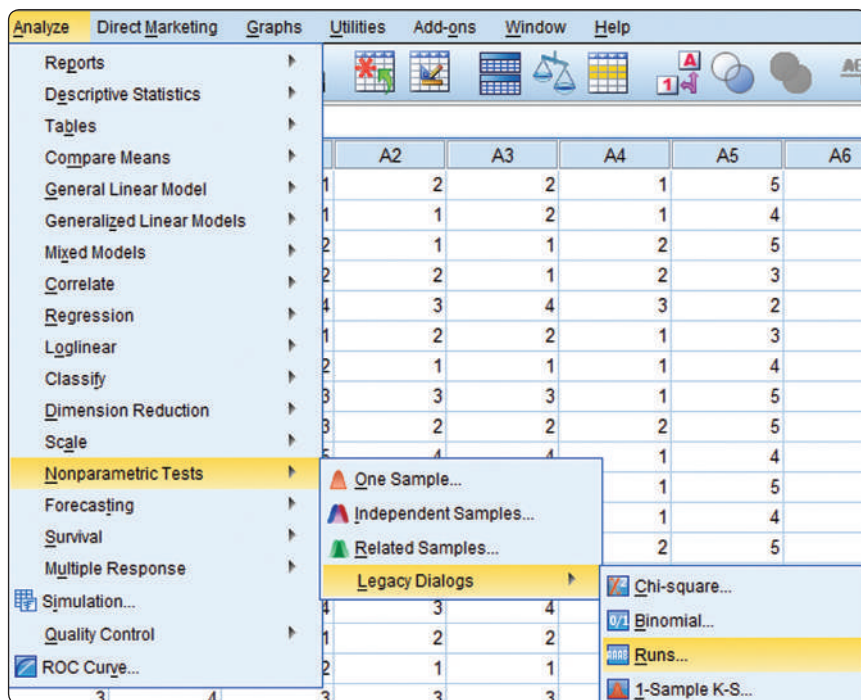
Πολλοί στατιστικοί έλεγχοι υποθέτουν ότι η σειρά εμφάνισης των παρατηρήσεων στο δείγμα είναι τυχαία, δηλαδή, η σειρά με την οποία συγκεντρώθηκαν τα δεδομένα είναι τυχαία. Αν η σειρά έχει σημασία, τότε το δείγμα δεν είναι τυχαίο και δεν μπορούμε να βγάλουμε ακριβή συμπεράσματα για τον πληθυσμό μας από αυτό. Γιαυτό είναι απαραίτητο να ελέγξουμε τα δεδομένα μας για τυχόν παραβίαση αυτής της υπόθεσης. Για να είναι αξιόπιστα τα συμπεράσματά μας, θα πρέπει το δείγμα μας να είναι τυχαίο. Ο έλεγχος αυτός γίνεται με τη βοήθεια του στατιστικού ελέγχου Runs. Το τεστ Runs ελέγχει αν η σειρά εμφάνισης των τιμών μιας μεταβλητής είναι τυχαία. Το Run είναι μια ακολουθία τιμών που μοιάζουν και πιο συγκεκριμένα, μια ακολουθία τιμών που βρίσκονται προς την ίδια πλευρά ενός σημείου τομής (cut point). Ένα δείγμα με πάρα πολλά Runs ή με πολύ λίγα φαίνεται να μην είναι τυχαίο.

Για παράδειγμα: Ας υποθέσουμε ότι ένα δείγμα 50 ατόμων ρωτάται σχετικά με το αν θα αγόραζε ή όχι ένα προϊόν. Η υποτιθέμενη τυχαιότητα του δείγματος θα αμφισβητούνταν αν και τα 50 άτομα ήταν του ίδιου φύλου. Το Run test θα μπορούσε να χρησιμοποιηθεί για να διαπιστωθεί αν το δείγμα είναι όντως τυχαίο.

Η υπόθεση H_0 , η μηδενική υπόθεση δηλαδή σε αυτόν τον έλεγχο, είναι ότι η σειρά των παρατηρήσεων είναι τυχαία.

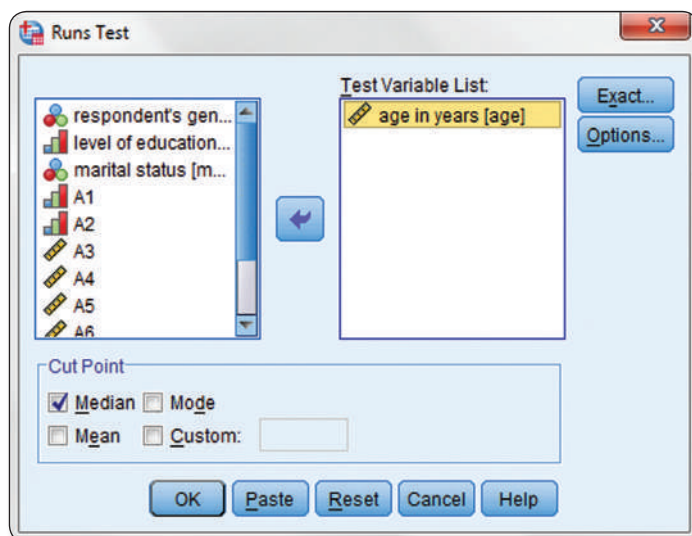
Από τη βασική γραμμή εντολών επιλέγουμε διαδοχικά (Εικόνα 3.26):

Statistics → Nonparametric Tests → Runs...



Εικόνα 3.26 Ο στατιστικός έλεγχος Runs

Τότε, θα εμφανιστεί το παρακάτω παράθυρο διαλόγου (Εικόνα 3.27):



Εικόνα 3.27 Το παράθυρο διαλόγου: 'Runs Test'

Κεφάλαιο Τρίτο

Εισαγωγή στην ανάλυση μεταβλητών με το IBM SPSS Statistics

Η διαδικασία αρχικά ταξινομεί κάθε τιμή της μεταβλητής, ανάλογα με το αν βρίσκεται πάνω ή κάτω από ένα σημείο τομής (Εικόνα 3.28). Στη συνέχεια κάνει τον έλεγχο για να επιβεβαιώσει ότι δεν υπάρχει κάποια τάξη στην ακολουθία που προέκυψε. Το σημείο τομής μπορεί να είναι ένα από τα παρακάτω στατιστικά μέτρα: η διάμεσος (προεπιλεγμένο), η μέση τιμή, η επικρατούσα τιμή ή κάποια άλλη συγκεκριμένη τιμή. Στο παράδειγμα αυτό χρησιμοποιούνται τα δεδομένα του `ereuna.sav`. Από τον πίνακα των μεταβλητών επιλέγουμε μία ή περισσότερες μεταβλητές και τη μετακινούμε στο Test Variable List χρησιμοποιώντας το αντίστοιχο εικονίδιο . Στο συγκεκριμένο παράδειγμα, επιλέγουμε τη μεταβλητή ηλικία του ερωτώμενου (`age`), την τοποθετούμε στο Test Variable List και πατάμε OK. Ως σημείο τομής επιλέγουμε τη διάμεσο και θα πάρουμε ως αποτέλεσμα τον παρακάτω πίνακα (Εικόνα 3.28):

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

a Median

Εικόνα 3.28 Αποτελέσματα του Runs Test

1. Το test value (τιμή ελέγχου) χρησιμοποιείται ως σημείο τομής (cut point) για να διχοτομήσει το δείγμα. Σε αυτό το παράδειγμα, το σημείο τομής είναι η διάμεσος, που είναι ίση με 39.
2. Από τις 99 περιπτώσεις, 48 έπεσαν κάτω από τη διάμεσο. Ας θεωρήσουμε αυτές τις περιπτώσεις ως «αρνητικές».
3. Οι υπόλοιπες 51 περιπτώσεις έπεσαν ακριβώς ή πάνω από τη διάμεσο. Ας θεωρήσουμε αυτές τις περιπτώσεις ως «θετικές».
4. Ο συνολικός αριθμός των περιπτώσεων είναι 99.

5. Το επόμενο στατιστικό είναι ένας μετρητής των παρατηρουμένων Runs στη μεταβλητή που ελέγχουμε. Όπως αναφέραμε και προηγουμένως το Run είναι μία ακολουθία των περιπτώσεων που βρίσκονται στην ίδια πλευρά του σημείου τομής. Ο συνολικός αριθμός των Runs είναι 45.

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

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

1	32
1	35
0	26
0	25
5	59

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

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

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

Κεφάλαιο Τρίτο

Εισαγωγή στην ανάλυση μεταβλητών με το IBM SPSS Statistics

6. Το Z στατιστικό είναι ίσο -1.103.
7. Με την επιλογή που κάναμε για το σημείο τομής (τη διάμεσο) αποδεχόμαστε την υπόθεση H_0 (της τυχαιότητας των παρατηρήσεων), γιατί $0.270 > 0.05$. Δηλαδή, μπορούμε να ισχυριστούμε ότι η σειρά των παρατηρήσεων πάνω και κάτω από τη διάμεσο είναι τυχαία. Γενικά, όμως, μπορούμε να πούμε ότι τα αποτελέσματα του ελέγχου εξαρτώνται από την επιλογή του σημείου τομής.

Ασκήσεις

Τρίτου κεφαλαίου

Άσκηση 1^η

Δίνεται το βάρος και το ύψος δεκαέξι παιδιών ηλικίας 9 έως 11 ετών.

Πίνακας 3.1 Βάρος, ύψος και ηλικία παιδιών ηλικίας 9 έως 11 ($N=16$)

Βάρος	Ύψος	Ηλικία
51,50	150	9,0
49,00	152	10,5
49,50	149	10,0
50,50	151	11,0
51,00	157	10,5
52,50	148	9,5
50,00	149	9,0
51,00	152	10,5
49,50	151	9,5
56,00	150	9,0
54,50	153	10,5
52,00	155	9,5
53,50	148	11,0
46,50	147	9,0
44,00	150	9,0
47,00	151	9,5

Πληκτρολογήστε τα δεδομένα και χρησιμοποιήστε τη διαδικασία Descriptives

για να υπολογίσετε τα στατιστικά μέτρα των ποσοτικών μεταβλητών: Βάρος, Ύψος και Ηλικία. Χρησιμοποιώντας τα ίδια δεδομένα δώστε μια πλήρη περιγραφή των τριών αυτών μεταβλητών με τη διαδικασία Frequencies. Να σώσετε τα δεδομένα σας ως Exercise3_1.sav.

Άσκηση 2^η

Ένα δείγμα 24 ανθρώπων υποβάλλεται σε ένα τεστ μέτρησης του δείκτη της ανθρώπινης νοημοσύνης και τα αποτελέσματα είναι τα εξής:

Πίνακας 3.2 Αποτελέσματα μέτρησης δείκτη νοημοσύνης (N=24)

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

- Υπολογίστε τα μέτρα κεντρικής τάσης και τα μέτρα μεταβλητότητας για τη μεταβλητή IQ.
- Υπολογίστε ένα διάστημα εμπιστοσύνης για το μέσο IQ του πληθυσμού από τον οποίο προέρχεται το δείγμα.
- Μπορούμε να ισχυριστούμε ότι η μεταβλητή IQ ακολουθεί την κανονική κατανομή;
- Ερμηνεύστε το Q-Q διάγραμμα που θα εμφανιστεί στον Output Viewer.
- Ερμηνεύστε το θηκόγραμμα (boxplot) που θα εμφανιστεί στον Output Viewer (διάμεσος, ποσοστιαία σημεία, ελάχιστη και μέγιστη παρατηρούμενη τιμή και ακραίες τιμές).
- Μπορείτε να ισχυριστείτε ότι η σειρά εμφάνισης των παρατηρήσεων είναι τυχαία;
- Να σώσετε τα δεδομένα σας ως Exercise3_2.sav.

Άσκηση 3^η

Οι ακόλουθες τιμές (Πίνακας 3.3) αποτελούν επιδόσεις φοιτητών στα μαθηματικά (math), που πήραν μέρος σε ένα τεστ με σκοπό την εκτίμηση της μέσης απόδοσής τους.

Πίνακας 3.3 Επιδόσεις φοιτητών στα μαθηματικά (N=30)

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

Δώστε μια πλήρη ανάλυση της μεταβλητής math. Να σώσετε τα δεδομένα σας ως Exercise3_3.sav.

Άσκηση 4^η

Ερευνάτε τη διαθεσιμότητα κατοικίας για οικογένειες με χαμηλό εισόδημα με παιδιά στην περιοχή σας. Για να πάρετε μια ιδέα της τιμής πώλησης των σπιτιών, κοιτάζετε σε ένα μεσιτικό γραφείο και σημειώνετε τις τιμές για σπίτια με τρία υπνοδωμάτια. Στον παρακάτω πίνακα αναφέρονται αυτές οι τιμές. Υπολογίστε τη μέση τιμή ενός σπιτιού με τρεις κρεβατοκάμαρες για την περιοχή αυτή και δώστε ένα διάστημα εμπιστοσύνης για την απάντησή σας. Να σώσετε τα δεδομένα σας ως Exercise3_4.sav.

Πίνακας 3.4 Τιμές πώλησης ακινήτων (N=18)

175.900	188.500	179.950	181.000	190.900	172.250
172.500	169.950	183.600	191.450	177.500	198.400
178.250	200.000	185.900	199.000	145.200	187.000

Άσκηση 5^η

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

Πίνακας 3.5 Εβδομαδιαίες διατροφικές δαπάνες πενήντα οικογενειών μιας περιοχής (N=50)

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

Να επιβεβαιωθεί η τυχαιότητα εμφάνισης των παρατηρήσεων και η κανονικότητα του δείγματος. Να σώσετε τα δεδομένα σας ως Exercise3_5.sav.

Άσκηση 6^η

Δίνονται τα ημερομίσθια (σε ευρώ) που αντιστοιχούν σε ένα δείγμα 50 αποφοίτων Τμημάτων Κοινωνικής Πολιτικής.

Πίνακας 3.6 Ημερομίσθια (σε ευρώ) πενήντα αποφοίτων Τμημάτων Κοινωνικής Πολιτικής (N=50)

26	29	35	31	29	38	21	32	31	31
33	27	32	29	34	31	29	31	29	30
31	27	28	31	32	29	30	37	30	30
28	33	30	30	31	28	32	32	30	30
29	28	29	32	28	29	29	32	32	31

Να ελεγχθεί η τυχαιότητα εμφάνισης των παρατηρήσεων και η κανονικότητα του δείγματος. Να σώσετε τα δεδομένα σας ως Exercise3_6.sav.