

1. Εισαγωγή

Στον στατιστικό έλεγχο υποθέσεων μας ενδιαφέρει να ελέγξουμε αν μία παράμετρος του πληθυσμού (για παράδειγμα η μέση τιμή ή η διακύμανση) ικανοποιεί μια υπόθεση ($H_0 \rightarrow$ μηδενική) έναντι μια εναλλακτικής υπόθεσης ($H_1 \rightarrow$ εναλλακτική).

Βασικές προϋποθέσεις για την χρήση κάποιων στατιστικών εργαλείων είναι:

1. Το δείγμα μας είναι τυχαίο. Συνήθως ο έλεγχος αυτός γίνεται από τον τρόπο λήψης του δείγματος, σε περίπτωση όμως που τα δεδομένα μας έχουν κάποια χρονολογική διάταξη, μπορεί να γίνει και με στατιστικό τεστ.
2. Δεν υπάρχουν ακραίες παρατηρήσεις. Ο έλεγχος γίνεται κυρίως γραφικά με την βοήθεια θηκογράμματος (boxplot).
3. Η κατανομή του πληθυσμού, από τον οποίο προήλθε το τυχαίο δείγμα είναι η κανονική κατανομή. (Για μέγεθος αρκετά μεγάλο μέγεθος δείγματος μπορούμε να παραλείψουμε αυτό τον έλεγχο (Κεντρικό Οριακό Θεώρημα)) Ο έλεγχος γίνεται με στατιστικά τεστ αλλά και γραφικά. Ο στατιστικός έλεγχος μπορεί να γίνει με το τεστ των Kolmogorov-Smirnov ή των Shapiro-Wilk. Ο γραφικός έλεγχος γίνεται με διάφορα γραφήματα όπως το Normal Q-Q Plot και το Detrended Normal Q-Q Plot .

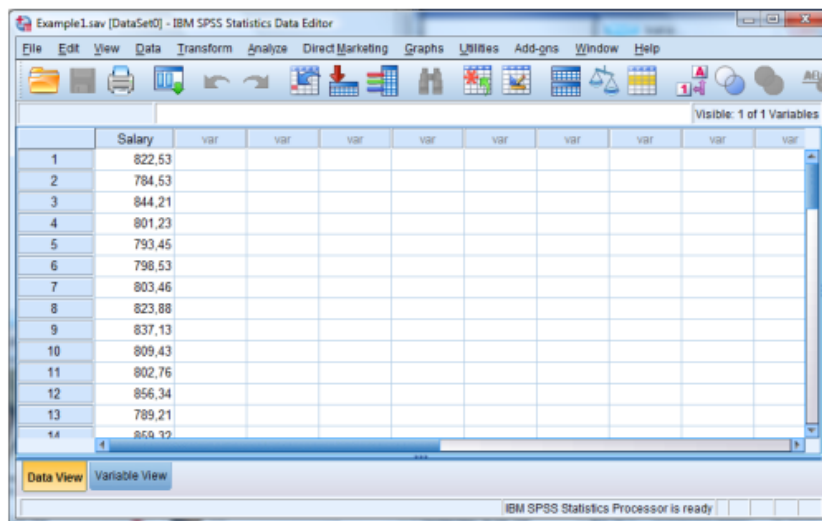
Παράδειγμα

Στον παρακάτω πίνακα φαίνονται οι μισθοί (σε χρηματικές μονάδες) 40 τυχαία επιλεγμένων υπαλλήλων που εργάζονται στον ιδιωτικό τομέα

822,53	784,53	844,21	801,23	793,45	798,53	803,46	823,88	837,13	809,43
802,76	856,34	789,21	859,32	807,67	798,12	800,56	840,43	787,13	815,43
819,32	806,24	784,12	844,44	825,53	779,23	824,12	813,56	817,43	825,51
802,22	816,34	803,72	804,12	835,37	849,43	834,29	826,21	827,10	828,78

Ελέγξτε αν τα προηγούμενα κριτήρια ικανοποιούνται για τα δεδομένα αυτά, ώστε στη συνέχεια να εφαρμοστούν διάφορα στατιστικά κριτήρια.

Αρχικά περνάμε όπως φαίνεται στο παρακάτω σχήμα, τα παραπάνω δεδομένα στο SPSS σε μια στήλη (μεταβλητή Salary) στην οποία έχουμε 40 περιπτώσεις (cases).



Καθορίζουμε επίσης και στο πεδίο Variable View και τα χαρακτηριστικά της ποσοτικής μεταβλητής (Salary).

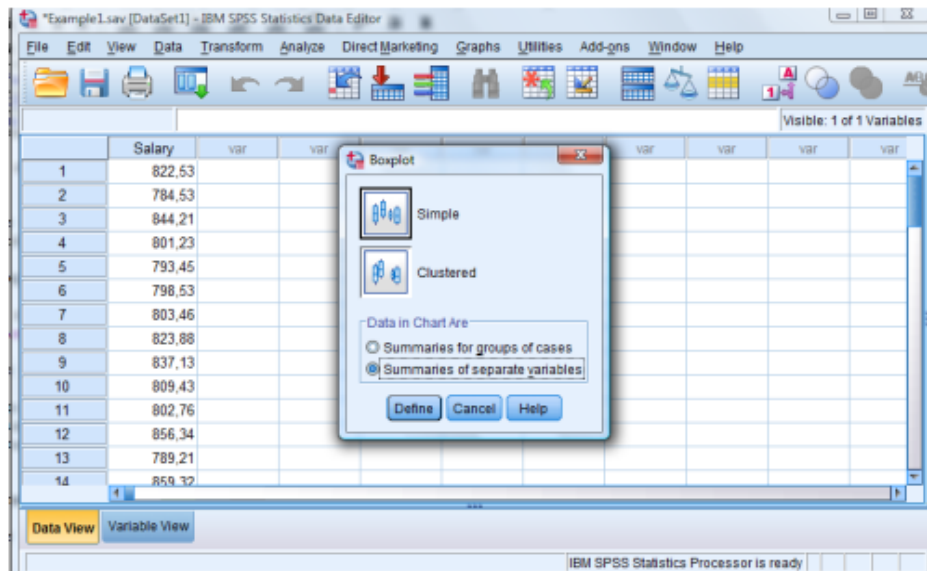
Στην συνέχεια ελέγχουμε αν ισχύουν οι προϋποθέσεις που αναφέρθηκαν προηγουμένως

1. Το δείγμα μας είναι τυχαίο. Αυτό προκύπτει από την εκφώνηση του παραδείγματος καθώς αναφέρεται ότι «...40 τυχαία επιλεγμένων υπαλλήλων...».

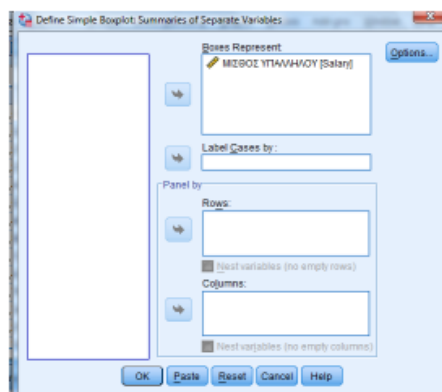
2. Δεν υπάρχουν ακραίες παρατηρήσεις. Στην συνέχεια ελέγχουμε αν υπάρχουν ακραίες παρατηρήσεις. Ο έλεγχος θα γίνει γραφικά με τη βοήθεια θηκογράμματος boxplot. Για να κατασκευάζουμε το θηκόγραμμα ακολουθούμε την παρακάτω διαδρομή.

Graphs/Legacy Dialogs/Boxplot

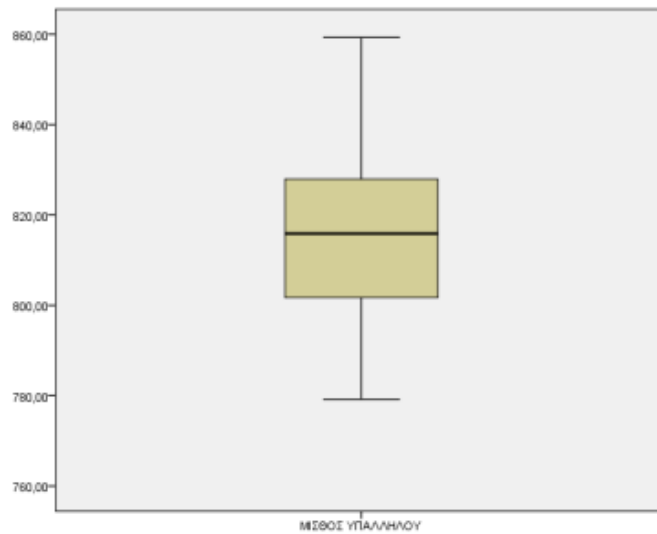
Στην συνέχεια στο παράθυρο που ακολουθεί επιλέγουμε Simple και Summaries of separate variables, όπως φαίνεται στο ακόλουθο σχήμα:



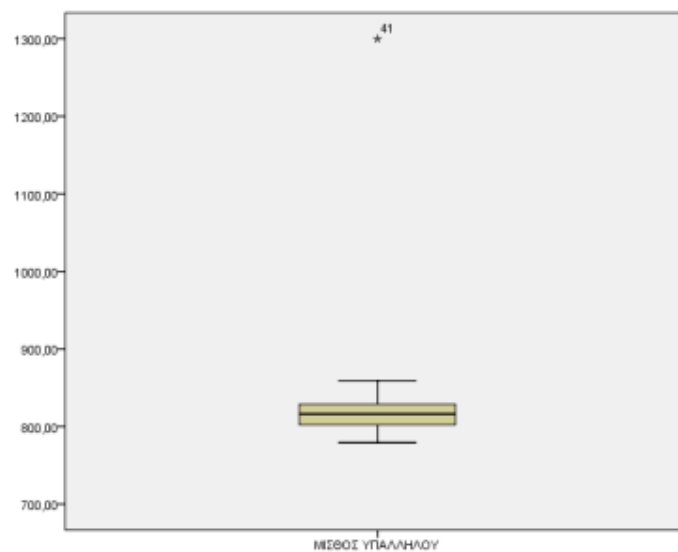
Τοποθετούμε την μεταβλητή Salary στο πλαίσιο Boxes Represented, όπως φαίνεται στο ακόλουθο σχήμα:



Μετά επιλέγουμε OK και προκύπτει το παρακάτω θηκόγραμμα (boxplot).



Παρατηρούμε ότι δεν υπάρχουν ακραίες παρατηρήσεις. Σε περίπτωση που είχαμε ακραίες παρατηρήσεις, αυτές θα σημειώνονταν με * και τον αύξοντα αριθμό της παρατήρησης δίπλα στο αστεράκι, όπως φαίνεται στο παρακάτω σχήμα:

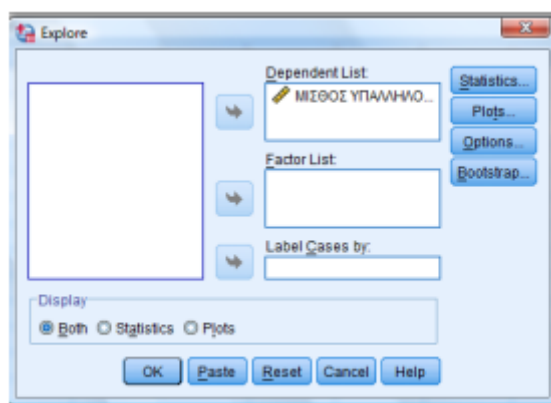


3. Η κατανομή του πληθυσμού, από τον οποίο προήλθε το τυχαίο δείγμα είναι η κανονική κατανομή. Όπως προαναφέρθηκε ο έλεγχος κανονικότητας του δείγματος μπορεί να γίνει γραφικά αλλά και με στατιστικά κριτήρια. Θα αναλύσουμε παρακάτω και τους δύο τρόπους. Ο έλεγχος με τα στατιστικά τεστ των Kolmogorov-Smirnov και Shapiro-Wilk γίνεται με τον παρακάτω τρόπο.

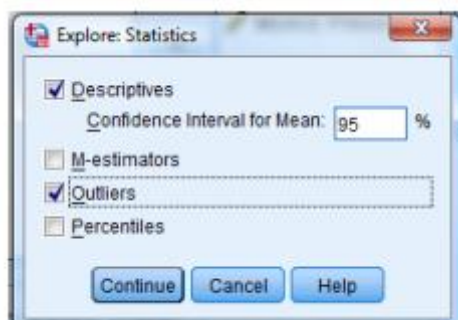
Ακολουθούμε την διαδρομή:

Analyze/Descriptive Statistics/Explore

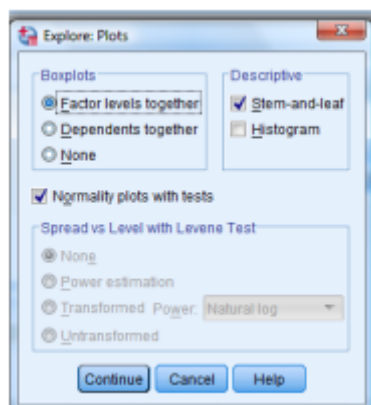
Το παράθυρο που προκύπτει είναι το εξής:



Σε αυτό το παράθυρο ελέγχουμε να έχει επιλεγθεί η ένδειξη Both, κάτω αριστερά, που σημαίνει ότι το SPSS θα παραθέσει και στατιστικούς δείκτες και γραφήματα που μας βοηθούν στον έλεγχο της κανονικότητας. Αφού μετακινήσουμε την μεταβλητή Salary δεξιά στο πλαίσιο Dependent List, προχώρουμε στις αναγκαίες ρυθμίσεις στα πλαίσια Statistics, Plots, Options. Στο πλαίσιο Statistics όπως φαίνεται στο επόμενο παράθυρο επιλέγουμε αν θέλουμε να δούμε τους περιγραφικούς δείκτες (τσεκάρουμε την ένδειξη Descriptives, αν θέλουμε διαφορετικό διάστημα εμπιστοσύνης, την ένδειξη Outliers αν θέλουμε να δούμε ακραίες τιμές) .



Στη συνέχεια επιλέγουμε Continue και επιστρέφουμε στο προηγούμενο παράθυρο. Επιλέγοντας το πλαίσιο Plots προκύπτει το παρακάτω παράθυρο



Στο παραπάνω παράθυρο επιλέγουμε οπωσδήποτε την ένδειξη Normality plots with tests για να προκύψουν οι πινάκες και τα γραφήματα που σχετίζονται με τα

στατιστικά κριτήρια για τον έλεγχο της κανονικότητας. Στο πλαίσιο Options το SPSS μας δίνει τη δυνατότητα να καθορίσουμε την πολιτική διαχείρισης των ελλειπουσών τιμών (missing values). Στην περίπτωση μας δεν έχει σημασία τι θα επιλέξουμε καθώς δεν έχουμε ελλείπουσες τιμές. Στην συνέχεια επιλέγουμε OK και προκύπτουν στο SPSS τα γραφήματα και οι πίνακες που μας χρειάζονται για τον έλεγχο της κανονικότητας.

Test of Normality

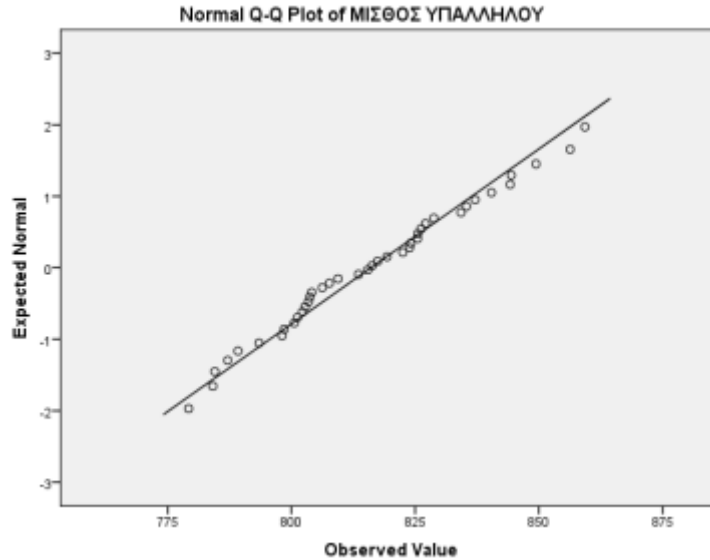
Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ΜΙΣΘΟΣ ΥΠΑΛΛΗΛΟΥ	,096	40	,200 [*]	,978	40	,623

*. This is a lower bound of the true significance.
a. Lilliefors Significance Correction

Παρατηρούμε ότι $\text{Sig.}=0,20=20\%$ (p-value) για το στατιστικό κριτήριο Kolmogorov-Smirnov και το $\text{Sig.}=0,623=62,3\%$ (p-value) για το στατιστικό κριτήριο των Shapiro-Wilk. (Για μέγεθος δείγματος μικρότερο του 50 ($n \leq 50$) επιλέγουμε το κριτήριο των Shapiro-Wilk, ενώ για μέγεθος δείγματος μεγαλύτερου του 50 ($n > 50$) επιλέγουμε το κριτήριο των Kolmogorov-Smirnov.) Αφού λοιπόν, για το στατιστικό τεστ των Shapiro-Wilk $\text{Sig.}=62,3\% > 5\%$ (το όριο που θέσαμε για να κρίνουμε την μηδενική μας υπόθεση), συμπεραίνουμε ότι δεν μπορούμε να απορρίψουμε την μηδενική μας υπόθεση. Δηλαδή, η κατανομή του πληθυσμού από τον οποίο προέρχεται το δείγμα μας είναι, προσεγγιστικά κανονική. Ο έλεγχος της κανονικότητας του δείγματος μπορεί να γίνει και γραφικά με τα διαγράμματα Normal Q-Q plot και Detrended Normal Q-Q Plot. Τα διαγράμματα που αφορούν την κανονικότητα προκύπτουν με την ίδια διαδικασία όπως προκύπτουν και οι πίνακες που αφορούν τα στατιστικά τεστ των Kolmogorov-Smirnov και Shapiro-Wilk.

Το διάγραμμα Normal Q-Q Plot

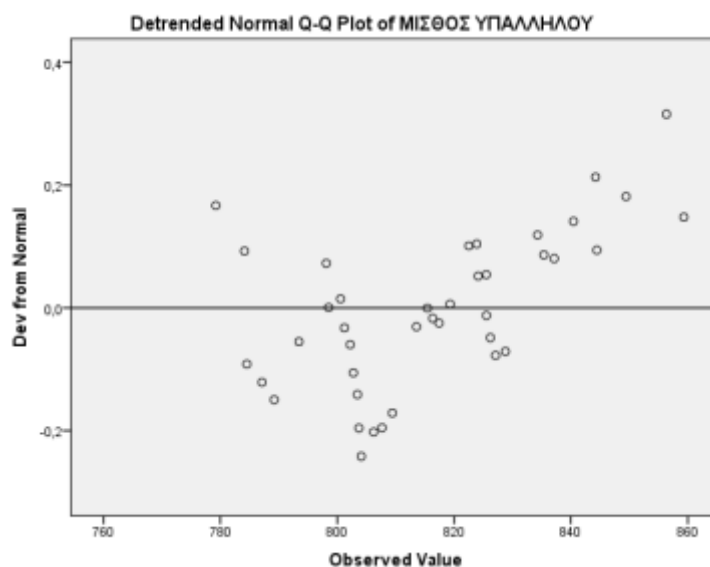
Το διάγραμμα Normal Q-Q Plot που προκύπτει για το παράδειγμα μας είναι το παρακάτω.



Στον άξονα των x στο παραπάνω διάγραμμα βρίσκονται οι παρατηρούμενες τιμές του μισθού των υπαλλήλων και στον άξονα των y βρίσκονται οι αναμενόμενες τιμές αυτού. Σε μια ιδανική κατάσταση για την κανονικότητα, όλα τα κυκλάκια θα βρίσκονται πάνω στην διχοτόμο των αξόνων. Αυτό όμως είναι σχεδόν απίθανο, οπότε μας ενδιαφέρει τα κυκλάκια να βρίσκονται πολύ κοντά στην διχοτόμο των αξόνων. Σε μια τέτοια περίπτωση μπορούμε να πούμε ότι το δείγμα ακολουθεί προσεγγιστικά την κανονική κατανομή. Στο συγκεκριμένο διάγραμμα δεν έχουμε μεγάλες αποκλίσεις οπότε μπορούμε να πούμε ότι το δείγμα προέρχεται από πληθυσμό που ακολουθεί την κανονική κατανομή.

Το διάγραμμα Detrended Normal Q-Q Plot

Το διάγραμμα Detrended Normal Q-Q Plot που προκύπτει για το παράδειγμα μας είναι το παρακάτω.



Σε αυτό το διάγραμμα στον άξονα των x βρίσκονται πάλι οι παρατηρούμενες τιμές για την μεταβλητή Salary. Στον άξονα των y βρίσκονται όμως τα αντίστοιχα ποσοστιαία σημεία μιας τυπικής κανονικής κατανομής, η οποία κατασκευάζεται με βάση το μέγεθος του δείγματος. Αυτό που κοιτάμε σε ένα τέτοιο διάγραμμα είναι τα κυκλάκια να είναι τυχαία κατανεμημένα, μέσα σε μια οριζόντια λωρίδα που σχηματίζεται με άξονα την ευθεία που διέρχεται από το μηδέν του κατακόρυφου άξονα, και διάμετρο αρκετά μεγάλη ώστε να περιλαμβάνει τα περισσότερα κυκλάκια. Αν ισχύει κάτι τέτοιο δεν έχουμε πρόβλημα κανονικότητας. Αντίθετα, αν τα κυκλάκια δεν είναι τυχαία κατανεμημένα, δηλαδή υπάρχουν πρότυπα (π.χ. $y = ax + b$, $y = ax^2 + bx + c$) ή συσσωρεύσεις (δηλαδή σε μερικά σημεία υπάρχει μεγάλη πυκνότητα και σε άλλα όχι) τότε υπάρχει πρόβλημα με την κανονικότητα του δείγματος. Στο παράδειγμα μας (παραπάνω σχήμα) βλέπουμε ότι τα κυκλάκια είναι τυχαία κατανεμημένα γύρω από το άξονα που διέρχεται από το μηδέν στον κατακόρυφο άξονα και επομένως δεν έχουμε πρόβλημα κανονικότητας.