ΚΕΦΑΛΑΙΟ 4

ΕΛΕΓΧΟΙ ΚΑΤΑΝΟΜΩΝ

Οι πληθυσμοί, ανεξάρτητα από το αν έχουν ίδιες θέσεις (ίσες μέσες τιμές) ή ίσες διασπορές, ενδέχεται να διαφέρουν πάρα πολύ ως προς άλλα χαρακτηριστικά τους. Έτσι, οι έλεγχοι υποθέσεων για τις παραμέτρους θέσης ή μεταβλητότητας ενός πληθυσμού (μέση τιμή, διάμεσος, ποσοστιαία σημεία, διασπορά) δεν αποκαλύπτουν διαφορές σε άλλα χαρακτηριστικά του πληθυσμού. Για να είμαστε, επομένως, σε θέση να ελέγξουμε υποθέσεις για την άγνωστη κατανομή πιθανότητας μιας τυχαίας μεταβλητής πάνω στην οποία έχουμε παρατηρήσεις (δείγμα), χρειαζόμαστε μιας άλλης μορφής υπόθεση. Αυτή θα είναι περιεκτικότερη, με την έννοια ότι θα πρέπει να αναφέρεται σε όλα τα ποσοστιαία σημεία ταυτόχρονα και όχι μόνο στην διάμεσο. Θα πρέπει, επίσης, να αναφέρεται σε όλες τις πιθανότητες ταυτόχρονα και όχι σε μία ή σε κάποιες μόνο από τις πιθανότητες. Με άλλά λόγια, χρειαζόμαστε υποθέσεις των οποίων ο έλεγχος θα απαντά στο ερώτημα: *αποτελούν οι παρατηρήσεις μας ένα δείγμα από κάποια συγκεκριμένη κατανομή;*

Υποθέσεις αυτής της μορφής μπορούν να ελεγχθούν με *ελέγχους καλής προσαρμογής (goodness-of-fit tests),* δηλαδή, με ελέγχους σχεδιασμένους να συγκρίνουν το δείγμα με τον τύπο του δείγματος που θα περιμέναμε να έχουμε από την κατανομή που υποθέτουμε, ώστε να μπορούμε να δούμε αν η υποτεθείσα συνάρτηση κατανομής προσαρμόζεται στα δεδομένα του δείγματος. Ο αρχαιότερος και γνωστότερος τέτοιος έλεγχος καλής προσαρμογής είναι ο έλεγχος χ^2 ο οποίος προτάθηκε από τον Pearson το 1900.

Αργότερα, ο Kolmogorov (1933, 1941) επενόησε έναν εναλλακτικό έλεγχο για τον ίδιο σκοπό και ο Smirnov (1939, 1948) τον επέκτεινε για την περίπτωση του ελέγχου της υπόθεσης ότι δύο ανεξάρτητα δείγματα μπορούν να υποτεθούν ότι προέρχονται από την ίδια κατανομή. Στα επόμενα, θα εξετάσουμε τους δύο αυτούς ελέγχους.

Εστω ότι X_1, X_2, ..., X_n είναι ένα τυχαίο δείγμα παρατηρήσεων πάνω σε μια τυχαία μεταβλητή Χ. Θα μπορούσαμε να ενδιαφερόμαστε να ελέγξουμε το ερώτημα κατά πόσον οι τιμές αυτές αποτελούν ένδειξη ότι η τυχαία μεταβλητή Χ ακολουθεί μία συγκεκριμένη κατανομή, π.χ. μία ομοιόμορφη κατανομή στο διάστημα μεταξύ 0 και 1 ή μεταξύ 0 και 10, ή μία κανονική κατανομή με μέση τιμή 20 και τυπική απόκλιση 2.7. Εναλλακτικά, θα μπορούσαμε να ενδιαφερόμαστε να εξετάσουμε ένα πολύ γενικότερο ερώτημα όπως: «*είναι εύλογο να υποθέσουμε ότι τα δεδομένα αυτά έχουν προέλθει από μια κανονική κατανομή με άγνωστη μέση τιμή και διασπορά;*» Ο έλεγχος χ^2, όπως θα δούμε, είναι ευέλικτος και επιτρέπει την εκτίμηση των αγνώστων παραμέτρων από τα δεδομένα. Ο έλεγχος του Kolmogorov, αντίθετα, απαιτεί κάποιες μεταβολές για να απαντήσει στο τελευταίο ερώτημα.

4.1 Ο χ^2 ΕΛΕΓΧΟΣ ΚΑΛΗΣ ΠΡΟΣΑΡΜΟΓΗΣ
(The χ^2 Goodness of Fit Test)

Ο αρχαιότερος και περισσότερο γνωστός έλεγχος καλής προσαρμογής είναι ο έλεγχος χ^2, ο οποίος προτάθηκε από τον Karl Pearson το 1900. Ο έλεγχος αυτός χρησιμοποιείται σε περιπτώσεις προβλημάτων στα οποία ενδιαφερόμεστε να εξετάσουμε αν τα

δεδομένα προέρχονται από μια ορισμένη κατανομή, δηλαδή, για τον έλεγχο υποθέσεων της μορφής

$$H_0: F_X(x) = F_0(x)$$

$$H_1: F_X(x) \neq F_0(x),$$

όπου $F_X(x) = P(X \leq x)$ είναι η συνάρτηση κατανομής της τυχαίας μεταβλητής X.

Τα δεδομένα είναι ένα τυχαίο δείγμα παρατηρήσεων μεγέθους n΄, οι οποίες είναι ταξινομημένες σε n κατηγορίες (κλάσεις) σύμφωνα με τον πίνακα που ακολουθεί:

Κλάση i	1	2	3	. . .	n	Σύνολο
O_i	O_1	O_2	O_3	. . .	O_n	$\sum_{i=1}^{n} O_i = n'$

Εδώ, O_i συμβολίζει τον αριθμό των παρατηρήσεων στην i κατηγορία. Έστω, για παράδειγμα, ότι έχουμε τις εξής παρατηρήσεις πάνω σε μια τυχαία μεταβλητή X:

1 3 1 5 1 0 4 3 6 2 7 5 0 3.

Ενας τρόπος κατάταξης των διακριτών αυτών δεδομένων είναι ο εξής:

	Κλάση (i)							
	0	1	2	3	4	5	6	7
Συχνότητα (O_i)	2	3	1	3	2	1	1	1

Ενας άλλος τρόπος κατάταξης των δεδομένων αυτών θα μπορούσε να εναι ο εξής:

	Κλάση (i)		
	1	2	3
	$X_j \le 1$	$2 \le X_j \le 5$	$X_j \ge 6$
Συχνότητα (O_i)	5	7	2

Στον παραπάνω πίνακα, X_j συμβολίζει την τυχούσα παρατήρηση πάνω στην τυχαία μεταβλητή X.

Η επιλογή του αριθμού και του εύρους των κλάσεων είναι στην κρίση του ερευνητή.

Εστω p_i^0 η πιθανότητα μιας τυχαίας παρατήρησης πάνω στην μεταβλητή X να ανήκει στην κατηγορία i, κάτω από την μηδενική υπόθεση ότι $F_0(x)$ είναι η συνάρτηση κατανομής της X. Δηλαδή,

$p_i^0 = P(\text{η παρατήρηση } X_j \text{ ανήκει στην κατηγορία i} \mid H_0)$ για κάθε i, j.

Τότε, ο αναμενόμενος αριθμός των παρατηρήσεων στην κατηγορία i, όταν η μηδενική υπόθεση είναι αληθής, ορίζεται από την σχέση

$$E_i = n' p_i^0 .$$

Η κατάλληλη στατιστική συνάρτηση για τον έλεγχο της υπόθεσης H_0 έναντι της H_1 είναι η

$$T = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i},$$

όπου n συμβολίζει τον αριθμό των κλάσεων στις οποίες είναι ταξινομημένες οι παρατηρήσεις. Η κατανομή της στατιστικής συνάρτησης T προσεγγίζεται, στις περισσότερες περιπτώσεις,

ικανοποιητικά από την κατανομή χ^2 με n-1 βαθμούς ελευθερίας. (Συμβολικά, $T \sim \chi^2_{n-1}$).

Από τον ορισμό της στατιστικής συνάρτησης T, είναι φανερό ότι οι μεγάλες τιμές της συνάρτησης αυτής αποτελούν ένδειξη εναντίον της μηδενικής υπόθεσης, αφού αντανακλούν μεγάλες αποκλίσεις μεταξύ των τιμών που παρατηρούνται στις διάφορες κατηγορίες και των τιμών που θα έπρεπε να παρατηρούνται στις αντίστοιχες κατηγορίες, αν η μηδενική υπόθεση ήταν αληθής. Επομένως, η κρίσιμη περιοχή του ελέγχου ορίζεται από την ανισότητα

$$T > \chi^2_{n-1,1-\alpha} \ ,$$

όπου $\chi^2_{n-1,1-\alpha}$ συμβολίζει το (1–α)-ποσοστιαίο σημείο της κατανομής χ^2_{n-1}.

Παρατήρηση: Η F_0 (x) ενδέχεται να έχει γνωστή παράμετρο, όπως, για παράδειγμα, στις υποθέσεις

$$H_0: F_X(x) = F_{Poisson(3)}(x)$$
$$H_0: F_X(x) = F_{N(3,5)}(x).$$

(Δηλαδή, η F_0 (x) καθορίζει μια συγκεκριμένη κατανομή). Ενδέχεται, όμως, η F_0 (x) να περιέχει άγνωστες παραμέτρους, όπως στις επόμενες δύο περιπτώσεις:

$$H_0: F_X(x) = F_{Poisson(\lambda)}(x)$$
$$H_0: F_X(x) = F_{N(\mu, \sigma^2)}(x).$$

(Η F_0 (x), δηλαδή, καθορίζει μια οικογένεια κατανομών). Σε αυτές τις περιπτώσεις, οι άγνωστες παράμετροι εκτιμώνται και το γεγονός αυτό αντανακλάται στους βαθμούς ελευθερίας της κατανομής της

στατιστικής συνάρτησης T. Συγκεκριμένα, η κατανομή της T δεν θα είναι η χ^2_{n-1} αλλά η χ^2_{n-1} - {αριθμός των εκτιμώμενων παραμέτρων}. Δηλαδή,

$$T = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{n-k-1},$$

όπου k συμβολίζει τον αριθμό των εκτιμώμενων παραμέτρων.

Παρατήρηση: Αν κάποιες από τις αναμενόμενες συχνότητες E_i είναι μικρές, η προσέγγιση της κατανομής της στατιστικής συνάρτησης T από την κατανομή χ^2 ενδέχεται να μην είναι ικανοποιητική. Πόσο μικρές μπορούν να είναι οι τιμές των E_i δεν είναι σαφές. Κατά τον Cochran (1952), καμία από τις τιμές E_i δεν πρέπει να είναι μικρότερη της μονάδας και, πάντως, το ποσοστό των τιμών E_i που δεν υπερβαίνουν το 5 δεν πρέπει να είναι μεγαλύτερο του 20%. Μεταγενέστερες μελέτες δείχνουν ότι αρκούν λιγότερο αυστηρές προϋποθέσεις. Για παράδειγμα, κατά τον Yarnold (1970), αν έχουμε περισσότερες από δύο κλάσεις, το ποσοστό των αναμενόμενων συχνοτήτων E_i που μπορούν να είναι μικρότερες του 5 μπορεί να είναι πολύ μεγαλύτερο. Συγκεκριμένα, αν ο αριθμός n των κλάσεων υπερβαίνει το 2, τότε ℓ από τις αναμενόμενες συχνότητες E_i μπορούν να είναι μικρότερες του 5, αρκεί να μην είναι μικρότερες του $5\ell/n$. Επομένως, 100q% από τις αναμενόμενες συχνότητες μπορούν να έχουν τιμή στο διάστημα [5q, 5), όπου $q = \ell/n$. Από το άλλο μέρος, κατά τον Slakter (1973), ο αριθμός m των θεωρούμενων κλάσεων μπορεί να υπερβαίνει τον αριθμό n των παρατηρήσεων και, κατά συνέπεια, η μέση τιμή των E_i μπορεί να είναι μικρότερη από την μονάδα. Στην πράξη, όταν υπάρχουν πολλές κλάσεις με αναμενόμενες συχνότητες

μικρότερες του 5, ορισμένοι ερευνητές επανακαθορίζουν τις κλάσεις με την συνένωση κάποιων από τις αρχικές κλάσεις, προκειμένου να οδηγηθούν σε κλάσεις με αναμενόμενες συχνότητες μεγαλύτερες ή ίσες του 5. Η επιλογή των συνενούμενων κλάσεων, όμως, είναι, εν γένει, αυθαίρετη και, συχνά, χωρίς νόημα.

Παράδειγμα 4.1.1: Ας θεωρήσουμε τυχαίο δείγμα 52 παρατηρήσεων πάνω στην ζήτηση X για ένα προϊόν (σε αριθμό μονάδων που παραγγέλλονται) ταξινομημένων ως εξής:

Κλάση i	≤ 1	2	3	4	5	6	7	≥ 8
Συχνότητα O_i	4	9	11	7	8	9	1	3

Για τον έλεγχο των υποθέσεων

H_0: $F_X(x) = F_{Poisson(\lambda)}(x)$

H_1: $F_X(x) \neq F_{Poisson(\lambda)}(x)$, για κάθε x = 0, 1, 2, ...

σε επίπεδο σημαντικότητας α=5%, κατασκευάζουμε τον πίνακα που ακολουθεί:

Κλάση i	X_j	O_i	p_i^0	$E_i = n' p_i^0$
1	≤ 1	4	0.0916	4.763
2	2	9	0.1465	7.618
3	3	11	0.1954	10.161
4	4	7	0.1954	10.161
5	5	8	0.1563	8.128
6	6	9	0.1042	5.418
7	7	1	0.0595	3.094
8	≥ 8	3	0.0511	2.657

$$n = \sum_{i=1}^{8} O_i = 52 .$$

Οι 8 κλάσεις που θεωρήθηκαν στον παραπάνω πίνακα επελέγησαν αυθαίρετα. Οι τιμές των πιθανοτήτων p_i^0, οι οποίες απαιτούνται για τον υπολογισμό των τιμών της τελευταίας στήλης του πίνακα, προκύπτουν από τις σχέσεις:

$$p_1^0 = P(X_j \le 1 | H_0) = P(X_{Poisson(\lambda)} \le 1)$$

. .
. .
. .

$$p_i^0 = P(X_j = i | H_0) = P(X_{Poisson(\lambda)} = i),\ i = 2, 3, ..., 7$$

$$p_8^0 = P(X_j \ge 8 | H_0) = P(X_{Poisson(\lambda)} \le 7) = 1 - \sum_{i=1}^{7} p_i^0 ,$$

όπου X_j συμβολίζει την τυχούσα παρατήρηση του δείγματος.

Είναι γνωστό ότι η παράμετρος λ είναι η μέση τιμή της κατανομής Poisson. Επομένως, η εκτιμήτρια της παραμέτρου λ είναι

$$\hat{\lambda} = \overline{X}, \quad \text{όπου} \quad \overline{X} = \frac{\sum_{i=1}^{n'} X_i}{n'} ,$$

αν οι n΄ παρατηρήσεις είναι αταξινόμητες, ή

$$\hat{\lambda} = \overline{X}_g, \text{ όπου } \overline{X}_g = \frac{\sum_{i=1}^{n} m_i O_i}{\sum_{i=1}^{n} O_i} = \frac{\sum_{i=1}^{n} m_i O_i}{n'},$$

αν οι n΄ παρατηρήσεις είναι ταξινομημένες σε n κατηγορίες με αντίστοιχα κεντρικά σημεία m_i , i = 1, 2, ..., n.

Για τα δεδομένα του παραδείγματός μας, η εκτίμηση της τιμής της παραμέτρου λ προκύπτει ίση με $\overline{x}$=4. Επομένως,

$$\hat{P}(X_{\text{Poisson}(4)} = x) = e^{-4}\frac{4^x}{x!}.$$

Επίσης, δεδομένου ότι n=8 και k=1, ισχύει ότι

$$T = \sum_{i=1}^{8} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{8-1-1} \equiv \chi^2_6 .$$

Η κρίσιμη περιοχή, επομένως, του ελέγχου μεγέθους 0.05 ορίζεται από την ανισότητα

$$T > \chi^2_{6,0.95} \equiv 12.59.$$

Η παρατηρούμενη τιμή, όμως, της στατιστικής συνάρτησης T είναι τ=5.256, η οποία δεν βρίσκεται μέσα στην κρίσιμη περιοχή. Επομένως, σε επίπεδο σημαντικότητας 5%, τα δεδομένα δεν παρέχουν ενδείξεις ότι η κατανομή της ζήτησης διαφέρει από την κατανομή Poisson. Το κρίσιμο επίπεδο του ελέγχου είναι

$$\hat{\alpha} = P(T \geq 5.256 \,|\, H_0) = 1 - P(T < 5.256 \,|\, H_0) = 1 - P(\chi^2_6 < 5.256)$$
$$= 0.5055$$

Παρατηρούμε ότι 2 από τις 8 θεωρηθείσες κλάσεις (ποσοστό ίσο με 25%) έχουν αναμενόμενες συχνότητες μικρότερες από 5. Αυτό είναι μέσα στα κατά Yarnold αποδεκτά όρια, αφού και οι δύο αυτές συχνότητες δεν είναι μικρότερες από 5.2/8=1.25. Στην περίπτωση, όμως, που υιοθετήσουμε την άποψη του Cochran, θα μπορούσαμε να

επανακαθορίσουμε τα όρια των κλάσεων, συνενώνοντας, για παράδειγμα, τις δύο τελευταίες κλάσεις, οι οποίες είναι και αυτές που μας «προβληματίζουν». Αυτό θα μας οδηγήσει σε 7 κλάσεις, με την τελευταία να αναφέρεται σε τιμές ≥ 7 και με παρατηρούμενη συχνότητα ίση με 1+3=4 και αντίστοιχη αναμενόμενη συχνότητα ίση με 3.094+2.657=5.751. Η ελεγχοσυνάρτησή μας τώρα, θα είναι η

$$T = \sum_{i=1}^{7} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{7-1-1} = \chi^2_5$$

και η κρίσιμη περιοχή του ελέγχου θα ορίζεται από την ανισότητα $T > \chi^2_{5,0.95} \equiv 11.071$. Η τιμή τ της ελεγχοσυνάρτησης T προκύπτει ίση με 4.328. Η προσέγγιση της τιμής του κρίσιμου επιπέδου του ελέγχου δεν διαφέρει πολύ από αυτή που είχαμε πριν από τον επανακαθορισμό των κλάσεων. Συγκεκριμένα,

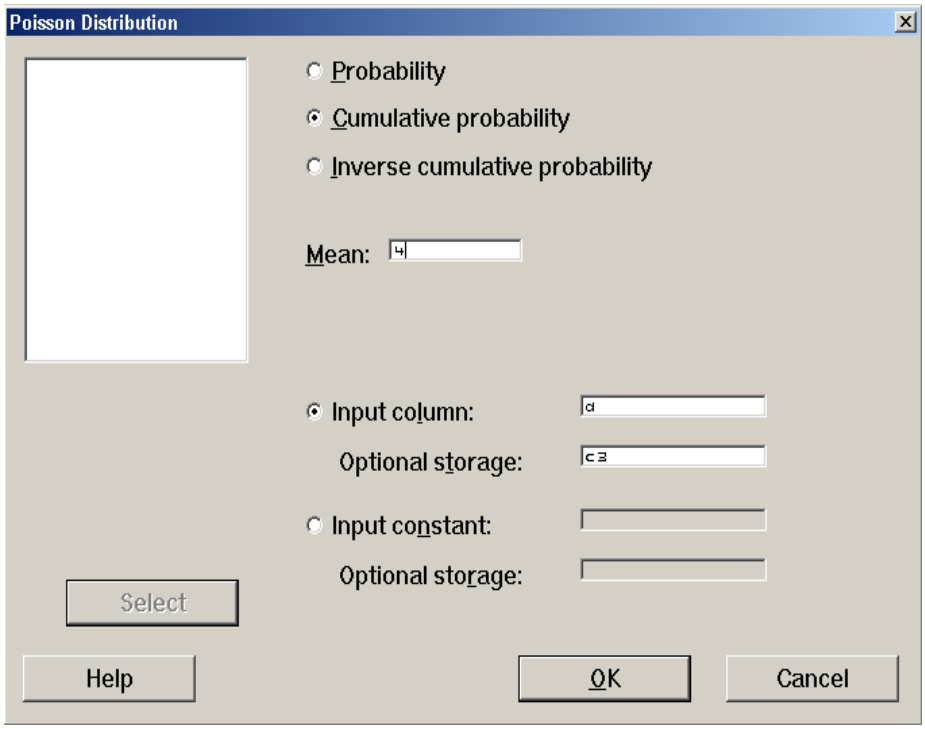
$$\hat{\alpha} = P(T \geq 4.328 \,|\, H_0) = 1 - P(\chi^2_5 < 4.328) = 0.5034 .$$

Λύση με το MINITAB: Ο έλεγχος χ^2 δεν παρέχεται απ' ευθείας από το MINITAB. Ο χρήστης πρέπει να διεξαγάγει την απαιτούμενη διαδικασία «χειροκίνητα» με το πρόγραμμα να τον βοηθάει μόνο στους υπολογισμούς.

Καταρχήν, υπολογίζονται οι πιθανότητες των κλάσεων κάτω από την H_0. Σε μία στήλη (π.χ. **C2**), καταχωρίζουμε τα άνω όρια όλων των κλάσεων εκτός από την τελευταία. Στην μεταβλητή αυτή δίνουμε το όνομα **d**. Αν υπολογίσουμε τις τιμές της συνάρτησης κατανομής της Poisson με λ=4 (εκτίμηση από το δείγμα) στα σημεία αυτά, η πιθανότητα της πρώτης κλάσης θα είναι η τιμή συνάρτησης κατανομής στο πρώτο άνω όριο, η πιθανότητα της δεύτερης κλάσης θα είναι η διαφορά της τιμής στο πρώτο όριο από αυτή του δευτέρου ορίου, κ.ο.κ.

Η πιθανότητα της τελευταίας κλάσης θα προκύψει με αφαίρεση του αθροίσματος των πιθανοτήτων όλων των προηγουμένων κλάσεων από την μονάδα.

Για να υπολογίσουμε την τιμή της συνάρτησης κατανομής στα σημεία της **d**, επιλέγουμε **Calc, Probability Distributions, Poisson** και οδηγούμεθα στο εξής πλαίσιο διαλόγου:



Για τον υπολογισμό της αθροιστικής συνάρτησης κατανομής, επιλέγουμε **Cumulative probability**. Στο πεδίο **Mean**, δηλώνουμε 4. Στο πεδίο **Input column**, δίνουμε το όνομα της στήλης που περιέχει τις τιμές στις οποίες θα υπολογισθεί η συνάρτηση κατανομής και, στο πεδίο **Optional storage**, δηλώνουμε την στήλη στην οποία θα καταχωρισθούν τα αποτελέσματα των υπολογισμών. Μετά τον υπολογισμό, το παράθυρο δεδομένων δείχνει ως εξής:

	C1	C2	C3
	x	d	f
	1	1	0.091578
	1	2	0.238103
	1	3	0.433470
	1	4	0.628837
	2	5	0.785130
	2	6	0.889326
	2	7	0.948866
	2		

Σημείωση: Για να εκτιμήσουμε την παράμετρο λ, καταχωρίζουμε το δείγμα σε μία στήλη (έστω **C1**) με όνομα **x** και επιλέγουμε **Stat, Basic Statistics, Display Descriptive Statistics**.

Στην στήλη **C4**, με το όνομα **p**, καταχωρίζουμε τις πιθανότητες των κλάσεων και, σε μία στήλη (έστω **C5**), με όνομα **o**, καταχωρίζουμε τις παρατηρούμενες συχνότητες O_i των κλάσεων. Επίσης, δημιουργούμε μία στήλη (έστω **C6**), με όνομα **e**, η οποία θα περιέχει τις αναμενόμενες συχνότητες E_i. Αυτό γίνεται από το πλαίσιο διαλόγου **Calc, Calculator** (βλέπε προηγούμενο παράδειγμα) πληκτρολογώντας **c4*52** στο πεδίο **Numeric Expression**.

Για τον υπολογισμό της τιμής της ελεγχοσυνάρτησης Τ επιλέγουμε **Calc**, **Calculator** και αποθηκεύουμε στην στήλη **C7** το αποτέλεσμα της εντολής **sum((o-e)**2/e)** του πεδίου **Numeric Expression**. Η τιμή που προκύπτει για την Τ είναι τ==5.258. Για τον υπολογισμό του κρίσιμου επιπέδου, υπολογίζουμε την τιμή της αθροιστικής συνάρτησης κατανομής της χ^2 με 6 βαθμούς ελευθερίας στο σημείο 5.258 και την αφαιρούμε από την μονάδα. Η προκύπτουσα

τιμή για το κρίσιμο επίπεδο (0.51) αποτελεί ένδειξη ότι η H_0 μπορεί να θεωρηθεί εύλογη.

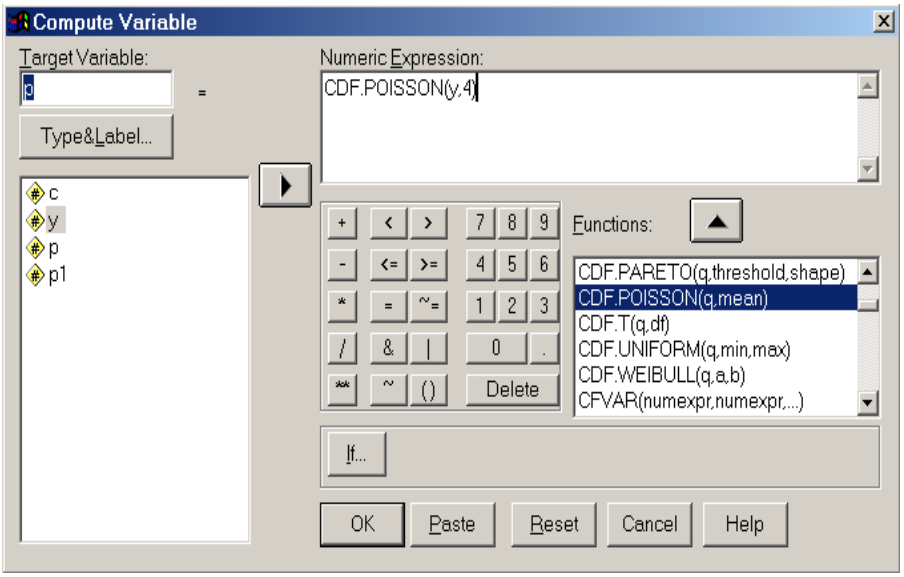
Λύση με το SPSS: Ο έλεγχος χ^2 δεν εκτελείται από το πακέτο SPSS αυτόματα και χωρίς κάποια προεργασία από την πλευρά του χρήστη. Για την διεξαγωγή του ελέγχου, απαιτείται ο ορισμός μιας μεταβλητής με τιμές τις παρατηρήσεις του δείγματος σε κωδικοποιημένη μορφή. Η κάθε τιμή της μεταβλητής αυτής θα δηλώνει σε ποια κλάση ανήκει η αντίστοιχη παρατήρηση του δείγματος.

Τα δεδομένα του παραδείγματός μας είναι ήδη ταξινομημένα σε κλάσεις και κατά συνέπεια η εισαγωγή τους στο SPSS δεν χρειάζεται καμία προκαταρκτική επεξεργασία. Στην κάθε κλάση, αντιστοιχίζουμε μία κωδική τιμή. Στην πρώτη κλάση δίνουμε την τιμή 1, στη δεύτερη την τιμή 2, κ.ο.κ. Δημιουργούμε την μεταβλητή **c** και καταχωρίζουμε τις παρατηρήσεις του δείγματος κωδικοποιημένες. Δίνουμε, δηλαδή, 4 φορές την τιμή 1 (αφού στην πρώτη κλάση υπάρχουν τέσσερις παρατηρήσεις), 9 φορές την τιμή 2 και συνεχίζουμε με όμοιο τρόπο για όλες τις κλάσεις.

Το SPSS, εξετάζοντας την μεταβλητή **c**, θα υπολογίσει μόνο του τις παρατηρούμενες συχνότητες O_i. Χρειάζεται, όμως, να δηλώσουμε τις αναμενόμενες συχνότητες E_i τις οποίες θα πρέπει να υπολογίσουμε. Ο τρόπος υπολογισμού είναι ο ίδιος είτε η παρατηρούμενη μεταβλητή είναι διακριτή (όπως εδώ), είτε είναι συνεχής. Δημιουργούμε μία μεταβλητή (έστω **y**) η οποία περιέχει ώς τιμές τα άνω όρια όλων των κλάσεων εκτός από την τελευταία. Στο δικό μας παράδειγμα, όλες οι κλάσεις, εκτός από την πρώτη και την τελευταία, αποτελούνται από μία μόνο τιμή η οποία είναι και το άνω τους όριο ταυτόχρονα. Στην συνέχεια, υπολογίζουμε την τιμή της αθροιστικής συνάρτησης

κατανομής $F_0(.)$ στα σημεία της **y**. Η τιμή που αντιστοιχεί στο πρώτο άνω όριο είναι η πιθανότητα, κάτω από την H_0, να «πάρουμε» μια παρατήρηση μέσα στην πρώτη κλάση. Για να βρούμε την αντίστοιχη πιθανότητα για κάθε άλλη κλάση, αφαιρούμε από την τιμή της $F_0(.)$ που αντιστοιχεί στο άνω όριό της την αντίστοιχη τιμή της στο άνω όριο της προηγούμενης κλάσης.

Για να υπολογίσουμε τις τιμές της $F_0(.)$, επιλέγουμε **Transform, Compute** και οδηγούμεθα στο εξής πλαίσιο διαλόγου:



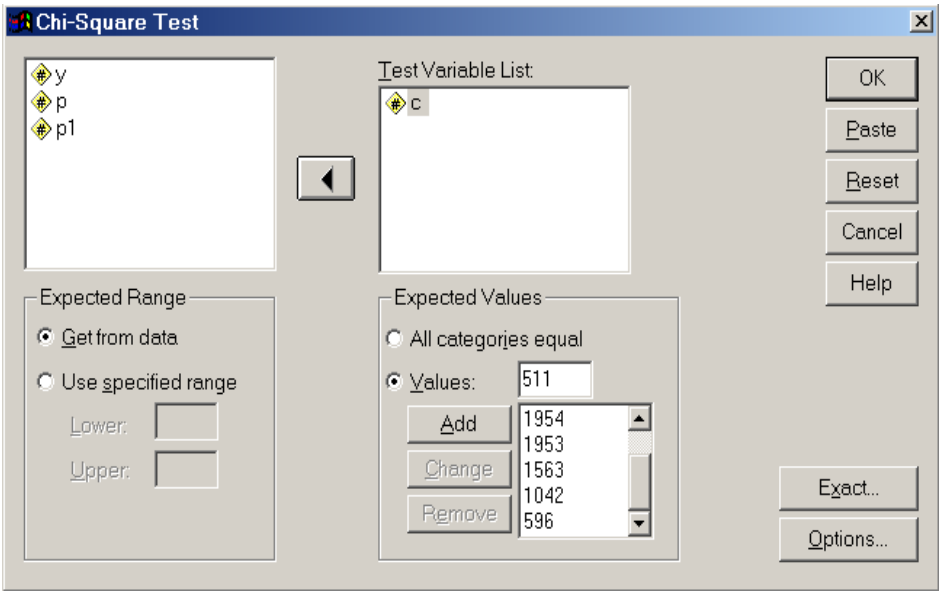
Στο πεδίο **Target Variable**, δίνουμε το όνομα μιας μεταβλητής που θα περιέχει τις τιμές της $F_0(x)$. Στο πεδίο **Numeric Expression** έχουμε δώσει μία έκφραση μέσω της οποία δίνεται η εντολή για τον υπολογισμό της συνάρτησης κατανομής της Poisson με $\lambda=4$ για κάθε σημείο y. Τα αποτελέσματα του υπολογισμού φαίνονται στον επόμενο πίνακα:

	c	y	p	p1
1	1	1	.0916	.0916
2	1	2	.2381	.1465
3	1	3	.4335	.1954
4	1	4	.6288	.1953
5	2	5	.7851	.1563
6	2	6	.8893	.1042
7	2	7	.9489	.0596
8	2	.	.	.0511
9	2	.	.	.

Στον παραπάνω πίνακα, η στήλη **p**, περιέχει τις τιμές της αθροιστικής συνάρτησης κατανομής της Poisson(λ=4) για κάθε τιμή της y. Η τελευταία περιέχει τα άνω όρια των κλάσεων εκτός από την τελευταία που εχει άνω όριο το άπειρο. Στην στήλη **p1** έχουμε καταχωρίσει τις πιθανότητες όλων των κλάσεων.

Σημείωση: Η τιμή λ=4 είναι η εκτίμηση του λ που προκύπτει από το δείγμα, δηλαδή από τον αριθμητικό μέσο του δείγματος που είναι εύκολο να υπολογισθεί από την επιλογή **Analyze**, **Descriptive Statistics**, **Descriptives**.

Για την διεξαγωγή του ελέγχου χ^2, επιλέγουμε **Analyze**, **Nonparametric Tests**, **Chi-Square** και οδηγούμεθα στο εξής πλαίσιο διαλόγου:



Στο πεδίο **Test Variable List**, δίνουμε **c**. Στο πλαίσιο **Expected Range**, επιλέγουμε **Get from data**. Έτσι, το πρόγραμμα θεωρεί ότι οι κωδικοί των κλάσεων μπορούν να πάρουν οποιαδήποτε ακέραιη τιμή μεταξύ της μικρότερης και της μεγαλύτερης από τις τιμές της **c**. Αν υπήρχαν και άλλες κλάσεις που δεν φαίνονταν στην **c** (επειδή καμμία τιμή του δείγματος δεν ανήκε σε αυτές), θα επιλέγαμε **Use specified range** και θα δίναμε το εύρος τιμών **των κωδικών**. Τότε, το πρόγραμμα θα γνώριζε ότι υπάρχουν κι άλλες κλάσεις με παρατηρούμενη συχνότητα μηδέν.

Στο πλαίσιο **Expected Values**, δηλώνουμε την θεωρητική πιθανότητα κάθε κλάσης, κάτω από την H_0. Επειδή οι κλάσεις δεν είναι ισοπίθανες, κάνουμε το εξής: Εκφράζουμε τις πιθανότητές τους ως κλάσματα ακεραίων με τον ίδιο παρονομαστή, τα οποία αθροίζουν στη μονάδα (π.χ. 3/10,4/10,1/10,1/10,1/10). Πρέπει να έχουμε τόσα κλάσματα όσες και οι κλάσεις. Μετά δηλώνουμε μόνο τους αριθμητές. Στο δικό μας παράδειγμα δίνουμε 916, 1465, 1954, 1953, 1563, 1042, 596, 511 (δηλαδή, τις τιμές της **p1** πολλαπλασιασμένες επί 10000 για

να γίνουν ακέραιοι). Η κάθε τιμή δίνεται με επιλογή **Values**, πληκτρολόγηση στο ενεργοποιούμενο πεδίο και πίεση του **Add**. Όταν τελειώσουμε, πιέζουμε **OK**. Αν εισαγάγουμε λάθος τιμή, μπορούμε να την διορθώσουμε (με επιλογή της, πληκτρολόγηση νέας τιμής και **Change**) ή να την διαγράψουμε (με επιλογή της και **remove**).

Τα αποτελέσματα του ελέγχου για το παράδειγμά μας έχουν την μορφή:

C

	Observed N	Expected N	Residual
1	4	4.8	-.8
2	9	7.6	1.4
3	11	10.2	.8
4	7	10.2	-3.2
5	8	8.1	-.1
6	9	5.4	3.6
7	1	3.1	-2.1
8	3	2.7	.3
Total	52		

Test Statistics

	C
Chi-Square[a]	5.258
df	7
Asymp. Sig.	.628

a 3 cells (37.5%) have expected frequencies less than 5. The minimum expected cell frequency is 2.7.

Στον πίνακα συχνοτήτων, βλέπουμε ότι το SPSS έχει υπολογίσει τα O_i και E_i για κάθε κλάση. Η τιμή της ελεγχοσυνάρτησης T είναι 5.258 (γραμμή **Chi-Square** του πίνακα **Test Statistics**), αλλά οι βαθμοί ελευθερίας (**df**) και το κρίσιμο επίπεδο (**Asymp. Sig.**) αναφέρονται στην περίπτωση που όλες οι παράμετροι της F_0 είναι γνωστές (το SPSS δεν γνωρίζει ότι εμείς εκτιμήσαμε μία παράμετρο από τα δεδομένα του δείγματος). Το κρίσιμο επίπεδο θα το υπολογίσουμε με την βοήθεια του SPSS από το μενού **Transform**, **Compute** και την συνάρτηση

CDF.CHISQ. Ο σωστός αριθμός των βαθμών ελευθερίας είναι 6 και η τιμή του κρίσιμου επιπέδου είναι 0.51. Συνεπώς, η υπόθεση ότι τα δεδομένα προέρχονται από μια κατανομή Poisson μπορεί να θεωρηθεί εύλογη.

Παρατηρούμε ότι το πρόγραμμα προειδοποιεί, σε υποσημείωση που ακολουθεί τον πίνακα συχνοτήτων, ότι τουλάχιστον 20% των κλάσεων έχουν αναμενόμενη συχνότητα μικρότερη του 5. Αν θέλουμε να προχωρήσουμε σε επανακαθορισμό των κλάσεων συνενώνοντας την έβδομη με την όγδοη κλάση, εργαζόμαστε ως εξής: Στην μεταβλητή **c**, αντικαθιστούμε τον κωδικό 8 με τον κωδικό 7, αφού τώρα έχουμε 7 κλάσεις και η νέα έβδομη θα περιλαμβάνει την αρχική έβδομη και αρχική όγδοη κλάση. Στο πλαίσιο διαλόγου του ελέγχου, θα πρέπει να διαγράψουμε από το πεδίο **Expected Values** την τιμή 511 (πιθανότητα της αρχικής όγδοης κλάσης) και να αντικαταστήσουμε την τιμή 596 με την τιμή 1107 (άθροισμα των τιμών 596 και 511, δηλαδή την πιθανότητα της νέας έβδομης κλάσης).

Λύση με το SAS: Το παράδειγμα δεν επιλύεται αυτόματα με το SAS. Μπορεί όμως να επιλυθεί χρησιμοποιώντας τις εντολές που ακολουθούν.

```
data poisson;
input x freq @@;
n=_N_;
prob=PDF('Poisson',x,4);
F=CDF('Poisson',x,4);
if n=1 then prob=f;
if n=8 then prob=1-CDF('Poisson',7,4);
expected=prob*52;
deviat=((freq-expected)**2)/expected;
cards;
1 4 2 9 3 11 4 7 5 8 6 9 7 1 8 3
;
run;
proc print;
run;
proc means sum;
var deviat;
run;
```

Στα παραπάνω, η εντολή **prob=PDF('Poisson',x,4);** δημιουργεί την μεταβλητή **prob** η οποία περιέχει τις τιμές $p_i^0=P(X_{Poisson(\lambda)}=i)$, i=2,3,…,7. Για τις τιμές p_1^0 και p_8^0, χρησιμοποιούνται αντίστοιχα οι σχέσεις $P(X_{Poisson(\lambda)} \le 1)$ και $P(X_{Poisson(\lambda)} \le 7)$, πράγμα το οποίο επιτυγχάνεται με τις εντολές

```
F=CDF('Poisson',x,4);
if n=1 then prob=f;
if n=8 then prob=1-CDF('Poisson',7,4);
```

Στην συνέχεια, υπολογίζονται οι αναμενόμενες συχνότητες, κάτω από την μηδενική υπόθεση με την εντολή **expected=prob*52;**, καθώς και οι ποσότητες $(O_i-E_i)^2/E_i$, οι οποίες αποθηκεύονται στην μεταβλητή **deviat** με την εντολή **deviat=((freq-expected)**2)/expected;**. Τέλος, το άθροισμα των τιμών αυτών, που είναι και η τιμή της ελεγχοσυνάρτησης, προκύπτει με τις εντολές

```
proc means sum;
var deviat;
run;
```

Το αποτέλεσμα της διαδικασίας αυτής δίδεται στον πίνακα που ακολουθεί.

OBS	X	FREQ	N	PROB	F	EXPECTED	DEVIAT
1	1	4	1	0.09158	0.09158	4.7621	0.12195
2	2	9	2	0.14653	0.23810	7.6193	0.25020
3	3	11	3	0.19537	0.43347	10.1591	0.06961
4	4	7	4	0.19537	0.62884	10.1591	0.98235
5	5	8	5	0.15629	0.78513	8.1273	0.00199
6	6	9	6	0.10420	0.88933	5.4182	2.36786
7	7	1	7	0.05954	0.94887	3.0961	1.41909
8	8	3	8	0.05113	0.97864	2.6589	0.04375

Analysis Variable : DEVIAT

Sum
5.2567903

Η τιμή 5.25679 είναι η τιμή της ελεγχοσυνάρτησης.

Παράδειγμα 4.1.2: Ας υποθέσουμε ότι η ημερήσια παραγωγή γάλακτος ενός αγελαδοτρόφου (σε λίτρα) για ένα τυχαίο δείγμα 40 ημερών συνοψίζεται στον πίνακα που ακολουθεί:

16.93	18.79	14.62	13.98	15.79	12.39	13.20	16.08
16.12	17.81	18.74	15.99	13.32	13.63	16.40	13.76
18.79	18.36	15.04	18.79	18.08	17.32	16.32	17.54
18.04	13.00	13.25	12.43	16.56	14.12	20.55	16.75
13.98	16.58	18.05	13.29	16.16	15.25	14.20	18.23

Αν Χ συμβολίζει την ημερήσια παραγωγή γάλακτος, να ελεγχθεί η υπόθεση

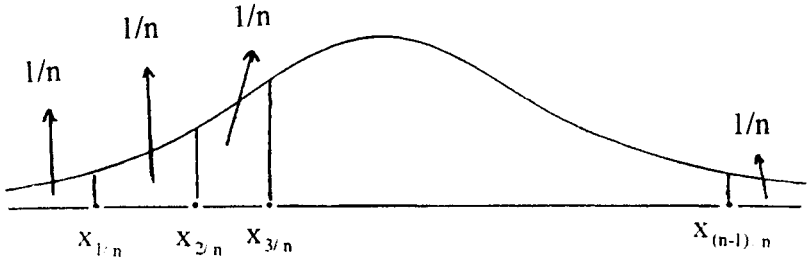
$$H_0 : F_X(x) = F_{N(\mu,\sigma^2)}(x), \quad \text{για κάθε} \quad -\infty < x < +\infty$$

σε επίπεδο σημαντικότητας α=1%.

Η γενική μεθοδολογία που προτιμάται στην συνεχή περίπτωση είναι η εξής:

Εστω ότι έχουμε n΄ παρατηρήσεις. Χωρίζουμε το διάστημα τιμών της τυχαίας μεταβλητής Χ σε n υποδιαστήματα (συνήθως όχι λιγότερα από 5), έτσι ώστε:

$$p_i^0 = P(\text{η X να ανήκει στο i υποδιάστημα}) = 1/n.$$



Εστω x_i το i-ποσοστιαίο σημείο της $N(\mu,\sigma^2)$. Τότε, επειδή $P(X \le x_i) = i$, και κατά συνέπεια, $P(X \le x_{i/n}) = i/n$, ισχύει ότι

$$P\left(Z \le \frac{x_{i/n} - \mu}{\sigma} \right) = i/n, \quad i = 1, 2, ..., n-1,$$

όπου Z συμβολίζει μια τυποποιημένη κανονική μεταβλητή. Τότε,

$$(x_{i/n} - \mu)/\sigma = z_{i/n},$$

όπου $z_{i/n}$ είναι το (i/n)-ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής και, επομένως,

$$x_{i/n} = \mu + \sigma z_{i/n}.$$

Αν οι παράμετροι μ και σ^2 είναι άγνωστες, η τιμή $x_{i/n}$ εκτιμάται από την στατιστική συνάρτηση

$$X^*_{i/n} = \begin{cases} \overline{X} + s^* z_{1/n}, & \text{αν } i/n \ge 0.5 \\ \overline{X} - s^* z_{1/n}, & \text{αν } i/n < 0.5, \end{cases}$$

όπου $\overline{X}$ και s^{*2} είναι οι αμερόληπτες εκτιμήσεις των παραμέτρων μ και σ^2 . Επομένως, μπορούμε να θεωρήσουμε κλάσεις της μορφής $[X^*_{(i-1)/n}, X^*_{i/n})$, $i = 1, 2, ..., n+1$, όπου X^*_0 και $X^*_{(n+1)/n}$ ορίζονται ίσα με $-\infty$ και $+\infty$ αντίστοιχα. Τότε, προκύπτει ο εξής πίνακας:

Κλάση i	O_i	p_i^0	$E_i = n' p_i^0$
$(-\infty, x^*_{1/n})$	O_1	1/n	n΄/n
$[x^*_{1/n}, x^*_{2/n})$	O_2	1/n	n΄/n
$[x^*_{2/n}, x^*_{3/n})$	O_3	1/n	n΄/n
.	.	.	.
.	.	.	.
.	.	.	.
$[x^*_{(n-1)/n}, +\infty)$	O_n	1/n	n΄/n

Εστω τώρα ότι στο συγκεκριμένο παράδειγμα αποφασίζουμε να ταξινομήσουμε τις παρατηρήσεις σε n = 8 κλάσεις, δηλαδή σε 8 υποδιαστήματα τέτοια ώστε

$$p_i^0 = P(\, X \in \text{στο i υποδιάστημα} \mid H_0\,) = 1/8 \equiv 0.125.$$

Από τα δεδομένα, προκύπτει ότι $\bar{x}=15.96$, $s^* =2.144$. Επομένως, έχουμε για το πάνω άκρο της πρώτης κλάσης ότι

$$x^*_{1/8} = x^*_{0.125} = 15.96 - 2.144\, z_{0.875} = 13.49.$$

Με όμοιο τρόπο υπολογίζουμε τα άκρα των υπόλοιπων κλάσεων:

$$x^*_{2/8} = x^*_{0.25} = \overline{X} - s^* z_{0.75} = 14.52, \quad x^*_{3/8} = x^*_{0.375} = \overline{X} - s^* z_{0.625} = 15.27,$$
$$x^*_{4/8} = x^*_{0.5} = \overline{X} + s^* z_{0.5} = 15.96, \quad x^*_{5/8} = x^*_{0.625} = \overline{X} + s^* z_{0.625} = 16.65,$$
$$x^*_{6/8} = x^*_{0.75} = \overline{X} + s^* z_{0.75} = 17.40, \quad x^*_{7/8} = x^*_{0.875} = \overline{X} + s^* z_{0.875} = 18.43,$$

Καταλήγουμε, επομένως, στον πίνακα που δίνεται στην συνέχεια. Οπως φαίνεται στον πίνακα αυτό, $E_i =5$, i = 1, 2, ..., 8. Άρα

$$T = \sum_{i=1}^{8} \frac{(O_i - 5)^2}{5} \sim \chi^2_{8-2-1}.$$

Η κρίσιμη περιοχή του ελέγχου μεγέθους 1% ορίζεται, επομένως, από την ανισότητα

$$T \geq \chi^2_{5,0.99} = 15.09.$$

Η παρατηρούμενη τιμή της T είναι 8.4. Άρα η μηδενική υπόθεση δεν απορρίπτεται σε επίπεδο σημαντικότητας 1%, δηλαδή, η κατανομή της ημερήσιας παραγωγής γάλακτος του συγκεκριμένου αγελαδοτρόφου δεν φαίνεται να διαφέρει στατιστικά σημαντικά από την κανονική κατανομή σε επίπεδο σημαντικότητας 1%.

Κλάση i	O_i	p_i^0	$E_i = 40\, p_i^0$
$(-\infty, 13.49)$	7	0.125	5
[13.49, 14.52)	6	0.125	5
[14.52, 15.27)	3	0.125	5
[15.27, 15.96)	1	0.125	5
[15.96, 16.65)	8	0.125	5
[16.65, 17.40)	3	0.125	5
[17.40, 18.43)	7	0.125	5
$[18.43, +\infty)$	5	0.125	5
	40		

Λύση με το MINITAB: Η διαδικασία διεξαγωγής του ελέγχου χ^2 είναι, όπως και στο προηγούμενο παράδειγμα, χρονοβόρα.

Καταρχήν, καταχωρίζουμε το δείγμα στην στήλη **C1** με όνομα **x**. Από το πλαίσιο **Stat**, **Basic Statistics**, **Display Descriptive Statistics** παίρνουμε τα εξής αποτελέσματα:

Descriptive Statistics

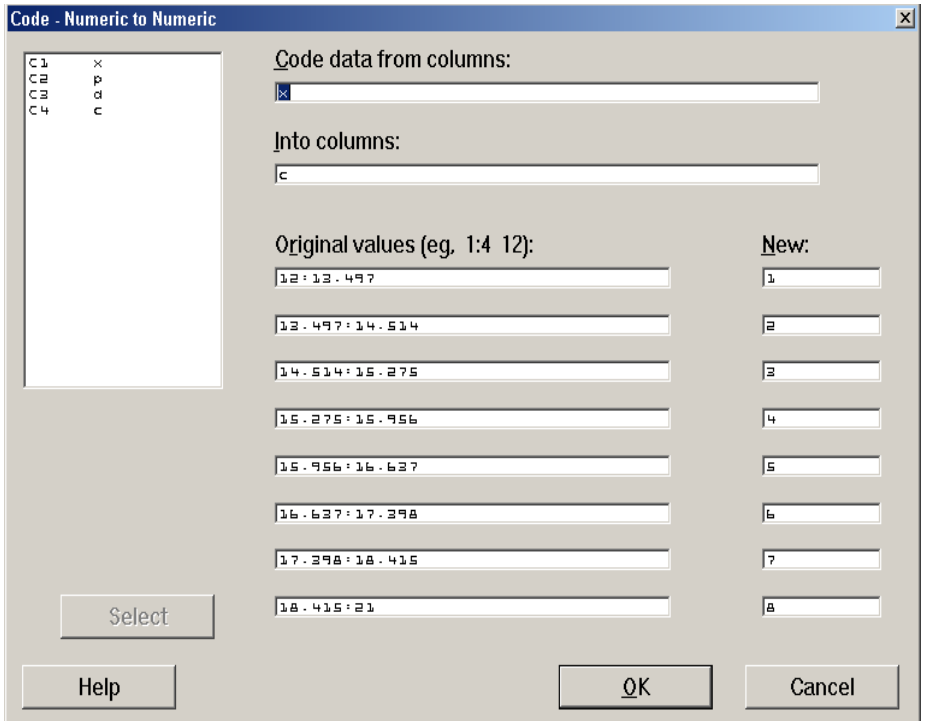
```
Variable       N       Mean     Median     TrMean
StDev    SE Mean
x             40     15.956     16.140     15.946
2.138      0.338

Variable       Minimum    Maximum         Q1         Q3
x               12.390     20.550     13.980     17.982
```

Στην συνέχεια, θα χρειασθούμε τις τιμές του μέσου και της τυπικής απόκλισης καθώς και την μικρότερη και μεγαλύτερη τιμή του δείγματος.

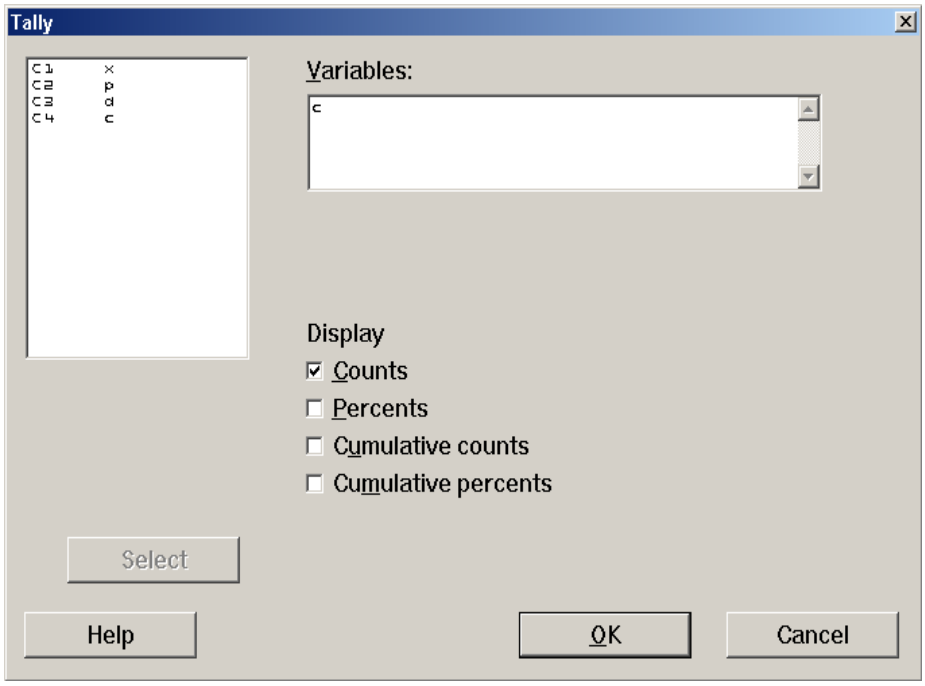
Στην στήλη **C2**, καταχωρίζουμε τις τιμές 0.125, 0.25, ..., 0.875 για να υπολογίσουμε τα αντίστοιχα ποσοστιαία σημεία της κανονικής κατανομής με μέση τιμή τον μέσο του δείγματος και τυπική απόκλιση την τυπική απόκλιση του δείγματος. Ο υπολογισμός γίνεται από το πλαίσιο **Calc, Probability Distributions, Normal** (αφήνεται στον αναγνώστη ως άσκηση). Τα αποτελέσματα καταχωρίζονται σε μία στήλη (έστω **C3**) με όνομα **d** και αποτελούν τα άνω όρια των κλάσεων στις οποίες θα ταξινομηθούν τα δεδομένα. Λείπουν μόνο το κάτω όριο της πρώτης κλάσης και το άνω της τελευταίας. Για τον λόγο αυτό, στην κορυφή της στήλης, δηλώνουμε μία τιμή μικρότερη από την μικρότερη τιμή του δείγματος και, στο τέλος της στήλης, να δηλώσουμε μία τιμή μεγαλύτερη από την μεγαλύτερη τιμή του δείγματος. Το παράθυρο των δεδομένων θα δείχνει ως εξής:

C1	C2	C3
x	p	d
16.93	0.125	12.0000
18.79	0.250	13.4966
14.62	0.375	14.5139
13.98	0.500	15.2747
15.79	0.625	15.9560
12.39	0.750	16.6373
13.20	0.875	17.3981
16.08		18.4154
16.12		21.0000
17.81		

Για να υπολογίσουμε τις παρατηρούμενες συχνότητες, κωδικοποιούμε τις παρατηρήσεις του δείγματος. Ο κωδικός κάθε παρατήρησης δείχνει σε ποια κλάση ανήκει η παρατήρηση. Για την κωδικοποίηση, επιλέγουμε **Manip, Code, Numeric to Numeric** και οδηγούμεθα στο εξής πλαίσιο διαλόγου:



Στο πεδίο **Code data from columns**, δηλώνουμε την στήλη που περιέχει τα αταξινόμητα δεδομένα. Στο πεδίο **Into columns**, δηλώνουμε την στήλη στην οποία θα καταχωρισθούν οι κωδικοποιημένες παρατηρήσεις του δείγματος. Στα πεδία **Original values** και **New**, δηλώνουμε τα όρια των κλάσεων και τους κωδικούς που αντιστοιχούν σε κάθε μία. Μόλις δημιουργηθεί η κωδικοποιημένη μεταβλητή, επιλέγουμε **Stat, Tables, Tally** οπότε προκύπτει το ακόλουθο πλαίσιο διαλόγου:



Στο πεδίο **Variables**, δηλώνουμε την μεταβλητή με τους κωδικούς και, στο πεδίο **Display**, επιλέγουμε **Counts** ώστε να υπολογισθεί ο αριθμός των εμφανίσεων κάθε κωδικού, δηλαδή η συχνότητα κάθε κλάσης. Τα αποτελέσματα που παίρνουμε έχουν την μορφή:

Summary Statistics for Discrete Variables

c	Count
1	7
2	6
3	3
4	1
5	8
6	3
7	7
8	5
N=	40

Ακολουθεί η καταχώριση αυτών των παρατηρούμενων συχνοτήτων, σε μία στήλη και των αναμενόμενων συχνοτήτων (που είναι όλες ίσες με 5) σε μία άλλη στήλη και, ο υπολογισμός της τιμής της

ελεγχοσυνάρτησης Τ. Αυτή προκύπτει ίση με 8.4. Η τιμή του κρίσιμου επιπέδου είναι 0.136. Κατά συνέπεια, η μηδενική υπόθεση μπορεί να θεωρηθεί εύλογη σε όλα τα συνήθη επίπεδα σημαντικότητας.

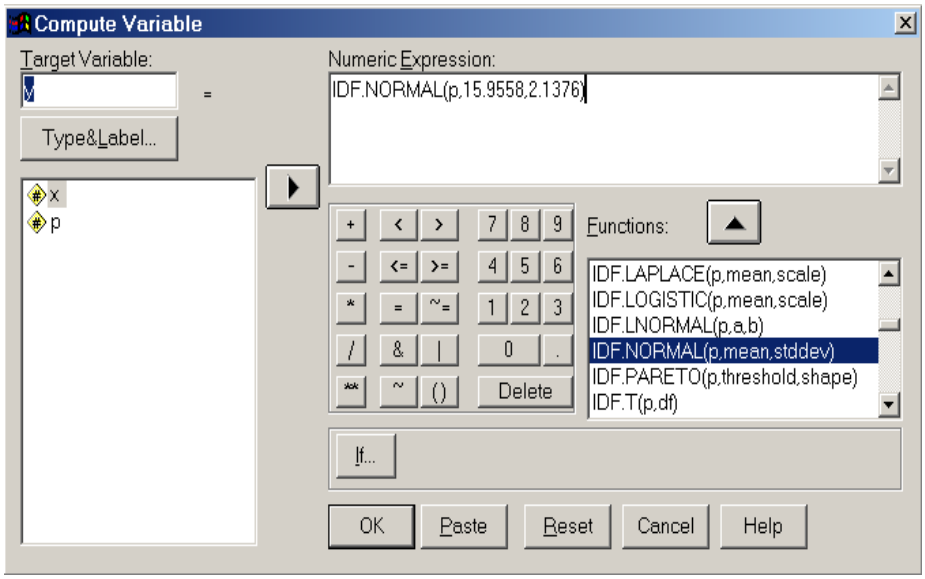
Λύση με το SPSS: Επειδή οι παρατηρήσεις του δείγματος δεν είναι ταξινομημένες, η διαδικασία θα πρέπει να ξεκινήσει με τον ορισμό κατάλληλων κλάσεων και των κωδικών τιμών τους. Για την αποφυγή σύγχυσης, καλό είναι η αύξουσα σειρά των κωδικών να αντιστοιχεί στην διάταξη των κλάσεων, δηλαδή ο μικρότερος κωδικός στην πρώτη κλάση, ο αμέσως μεγαλύτερος στη δεύτερη κ.ο.κ.

Όπως έχει αποφασισθεί, ο αριθμός των κλάσεων στις οποίες θα ταξινομηθούν τα δεδομένα του παραδείγματος είναι 8 ισοπίθανα διαστήματα τιμών κάτω από την κανονική κατανομή, με μέση τιμή και διακύμανση τις αντίστοιχες εκτιμήσεις από το δείγμα. Για τον υπολογισμό των ορίων των κλάσεων, πρέπει να γίνουν τα εξής:

- Να υπολογισθεί ο μέσος $\bar{x}$ και η τυπική απόκλιση s του δείγματος
- Να υπολογισθούν τα 0.125,0.25,0.375,...,0.875 ποσοστιαία σημεία της $N(\bar{x}, s^2)$ κατανομής.

Ας υποθέσουμε ότι το δείγμα έχει εισαχθεί στην μεταβλητή **x**. Με **Analyze**, **Descriptive Statistics**, **Descriptives**, υπολογίζουμε τις τιμές $\bar{x}$ και s οι οποίες προκύπτουν ίσες με 15.9558 και 2.1376, αντίστοιχα.

Κατόπιν, σε μία μεταβλητή (έστω **p**) καταχωρίζουμε τις τιμές 0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875 που δηλώνουν τα ποσοστιαία σημεία τα οποία θέλουμε να υπολογίσουμε. Προκειμένου να ζητήσουμε από το SPSS να αποθηκεύσει σε μια άλλη μεταβλητή (έστω **y**) τα εν λόγω ποσοστιαία σημεία, επιλέγουμε **Transform**, **Compute** και οδηγούμεθα στο εξής πλαίσιο διαλόγου

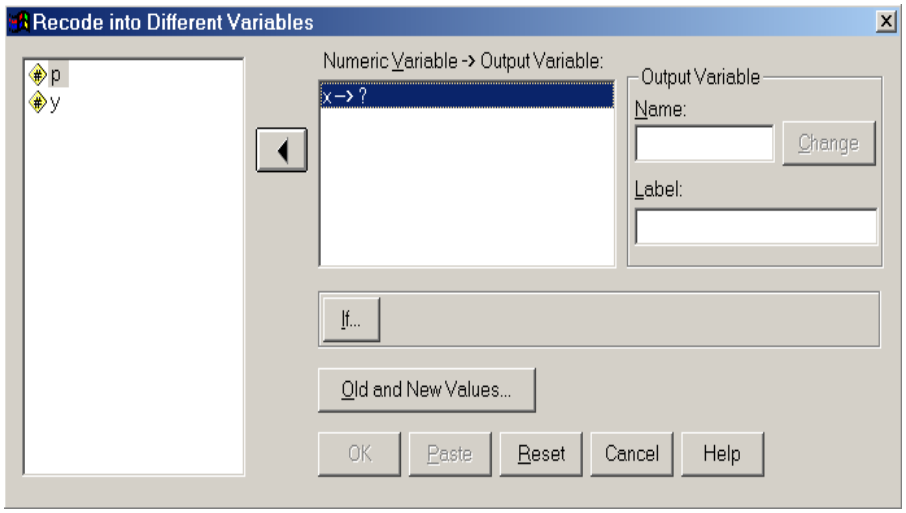


Στο πεδίο **Target Variable**, δίνουμε το όνομα της μεταβλητής στην οποία θα αποθηκευθούν τα ποσοστιαία σημεία. Στο πεδίο **Numeric Expression**, «δίνουμε» στο SPSS την εντολή να υπολογίσει τα ποσοστιαία σημεία της κανονικής κατανομής με μέση τιμή 15.9558 και τυπική απόκλιση 2.1376, για τις πιθανότητες που βρίσκονται στην μεταβλητή **p** και να τις αποθηκεύσει στην μεταβλητή **y**. Μετά από αυτή την διαδικασία το φύλλο δεδομένων δείχνει ως εξής:

	x	p	y
1	16.93	.125	13.50
2	18.79	.250	14.51
3	14.62	.375	15.27
4	13.98	.500	15.96
5	15.79	.625	16.64
6	12.39	.750	17.40
7	13.20	.875	18.41

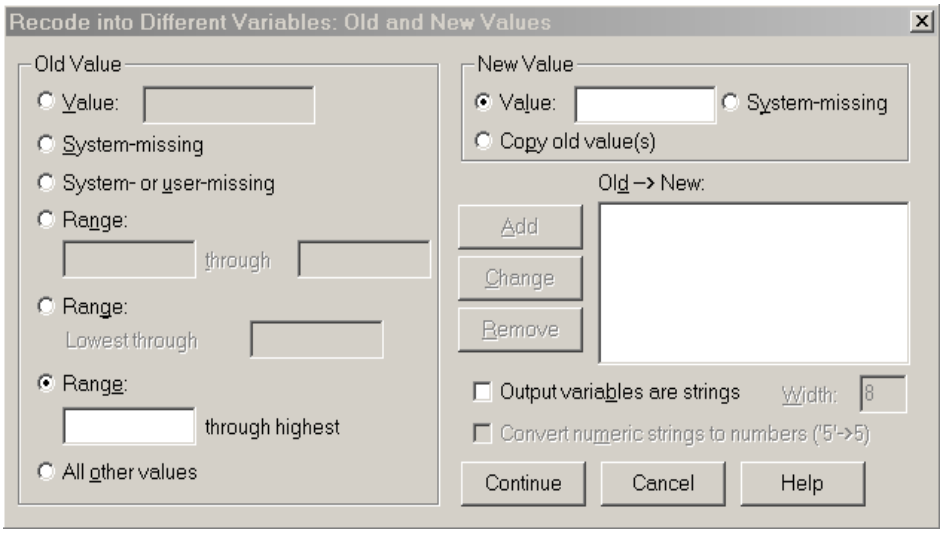
Το επόμενο βήμα είναι να δημιουργήσουμε την κωδική μεταβλητή που θα δηλώνει σε ποια κλάση ανήκει κάθε παρατήρηση της **x**.

Επιλέγοντας **Transform**, **Recode**, **Into Different Variables**, εμφανίζεται το εξής πλαίσιο διαλόγου:

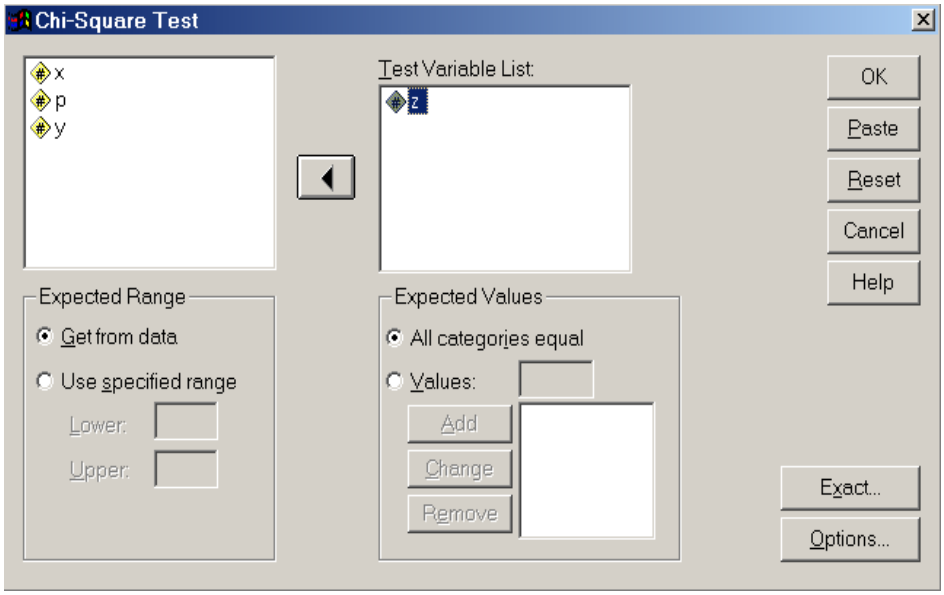


Επιλέγουμε την **x** και την στέλνουμε στο πεδίο **Numeric Variable – Output Variable**. Αμέσως ενεργοποιείται το πλαίσιο **Output Variable**, στο πεδίο **Name** του οποίου δίνουμε το όνομα που θέλουμε να έχει η κωδική μεταβλητή (έστω **z**). Τότε, ενεργοποιείται το πλήκτρο **Change**, το οποίο πιέζουμε.

Στην συνέχεια, πρέπει να δηλώσουμε στο SPSS ποιες είναι οι κωδικές τιμές και πώς θα γίνει η κωδικοποίηση. Πιέζοντας **Old and New Values**, προκύπτει το εξής πλαίσιο διαλόγου



Εδώ δηλώνονται οι τιμές της αρχικής μεταβλητής στο πλαίσιο **Old Value** και οι αντίστοιχοι κωδικοί στο πλαίσιο **New Value**. Για το δικό μας παράδειγμα, στο **Old Value** πρέπει να δίνουμε το εύρος τιμών του δείγματος στο οποίο αντιστοιχεί κάθε κωδικός. Για την πρώτη κλάση (που περιέχει τις τιμές που είναι μικρότερες του 13.5), επιλέγουμε **Range** (με την υποσημείωση **Lowest through**) και, στο πλαίσιο που ενεργοποιείται, δίνουμε 13.5. Στο πλαίσιο **New Value**, επιλέγουμε **Value** και δίνουμε 1. Μετά, πιέζουμε το πλήκτρο **Add** που έχει στο μεταξύ ενεργοποιηθεί και το ζεύγος εύρους τιμών και κωδικού εμφανίζεται στο πεδίο **Old → New**. Εργαζόμενοι με όμοιο τρόπο, εισάγουμε τις υπόλοιπες κλάσεις [13.5,14.51), [14.51,15.27), [15.27,15.96), [15.96,16.64), [16.64,17.4), [17.4,18.41), [τιμές μεγαλύτερες του 18.41) κωδικούς από το 2 έως το 8 και, εργαζόμενοι με όμοιο τρόπο, τις εισάγουμε στο πεδίο **Old → New**. Εν συνεχεία, επιλέγουμε **Analyze**, **Nonparametric Tests**, **Chi-Square** και οδηγούμεθα στο εξής πλαίσιο διαλόγου:



Στο πεδίο **Test Variable List**, δίνουμε **z**. Στο πλαίσιο **Expected Range**, επιλέγουμε **Get from data**. Στο πλαίσιο **Expected Values**, επιλέγουμε **All categories equal** (αφού οι κλάσεις είναι ισοπίθανες κάτω από την H_0) και πιέζουμε **OK**.

Τα αποτελέσματα δίνονται με την εξής μορφή:

Z

	Observed N	Expected N	Residual
1.00	7	5.0	2.0
2.00	6	5.0	1.0
3.00	3	5.0	-2.0
4.00	1	5.0	-4.0
5.00	8	5.0	3.0
6.00	3	5.0	-2.0
7.00	7	5.0	2.0
8.00	5	5.0	.0
Total	40		

Test Statistics

	Z
Chi-Square[a]	8.400
df	7
Asymp. Sig.	.299

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 5.0.

Κι εδώ πρέπει να υπολογίσουμε μόνοι μας την τιμή του κρίσιμου επιπέδου δεδομένου ότι οι βαθμοί ελευθερίας που χρησιμοποιεί το SPSS αντιστοιχούν σε πλήρως καθορισμένη κατανομή. Ο σωστός αριθμός των βαθμών ελευθερίας είναι 5 και το κρίσιμο επίπεδο που αντιστοιχεί στην τιμή τ=8.4 της Τ είναι 0.136. Κατά συνέπεια, η υπόθεση ότι τα δεδομένα προέρχονται από μία κανονική κατανομή μπορεί να θεωρηθεί εύλογη σε όλα τα συνήθη επίπεδα σημαντικότητας.

Λύση με το SAS: Για την διεξαγωγή του ελέγχου απαιτείται υπολογισμός των ορίων των 8 κλάσεων στις οποίες θα ταξινομήσουμε τις παρατηρήσεις. Εύκολα διαπιστώνεται ότι ο μέσος της υπό εξέταση μεταβλητής είναι 15.9558 και η τυπική απόκλιση 2.1376. Χρησιμοποιώντας τις εντολές

```
data milkpre;
input percen @@;
x=(probit(percen)*2.1376)+15.9558;
cards;
0.125 0.25 0.375 0.5 0.625 0.75 0.875
;
run;
proc print;
run;
```

παίρνουμε τα όρια των κλάσεων, αποθηκευμένα στην μεταβλητή **x**. Η εντολή **probit** δίδει τα σημεία $\mathbf{z_i}$, της τυποποιημένης κανονικής κατανομής, όπου i=0.125, 0.25,…0.875. Το αποτέλεσμα αυτής της διαδικασίας έχει την μορφή:

OBS	PERCEN	X
1	0.125	13.4968
2	0.250	14.5140
3	0.375	15.2747
4	0.500	15.9558
5	0.625	16.6369
6	0.750	17.3976
7	0.875	18.4148

Στην συνέχεια, μπορούμε να ταξινομήσουμε τις παρατηρήσεις μας στις 8 κλάσεις και να διεξαγάγουμε τον έλεγχο χ^2 με τις εξής εντολές.

```
data milk;
input x @@;
if x<13.5 then code=1;
if x>13.5 and x<=14.51 then code=2;
if x>14.51 and x<=15.27 then code=3;
if x>15.27 and x<=15.96 then code=4;
if x>15.96 and x<=16.64 then code=5;
if x>16.64 and x<=17.4 then code=6;
if x>17.4 and x<=18.41 then code=7;
if x>18.41 then code=8;
cards;
16.93 18.79 14.62 13.98 15.79 12.39 13.20 16.08
16.12 17.81 18.74 15.99 13.32 13.63 16.40 13.76
18.79 18.36 15.04 18.79 18.08 17.32 16.32 17.54
18.04 13.00 13.25 12.43 16.56 14.12 20.55 16.75
13.98 16.58 18.05 13.29 16.16 15.25 14.20 18.23
;
run;
proc print;
run;
proc freq;
tables code / chisq;
run;
```

Το αποτέλεσμα του ελέγχου έχει την μορφή:

CODE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	7	17.5	7	17.5
2	6	15.0	13	32.5
3	3	7.5	16	40.0
4	1	2.5	17	42.5
5	8	20.0	25	62.5
6	3	7.5	28	70.0
7	7	17.5	35	87.5
8	5	12.5	40	100.0

Chi-Square Test for Equal Proportions

Statistic = 8.400 DF = 7 Prob = 0.299

Η τιμή της ελεγχοσυνάρτησης είναι 8.4. Οπως και στην λύση του SPSS, πρέπει να υπολογίσουμε μόνοι μας την τιμή του κρίσιμου επιπέδου δεδομένου ότι οι βαθμοί ελευθερίας που χρησιμοποιεί το SAS αντιστοιχούν σε πλήρως καθορισμένη κατανομή. Ο σωστός αριθμός των βαθμών ελευθερίας είναι 5 και το κρίσιμο επίπεδο που αντιστοιχεί στην τιμή τ=8.4 της Τ είναι 0.136.

4.2 Ο ΕΛΕΓΧΟΣ KOLMOGOROV

(Kolmogorov Test)

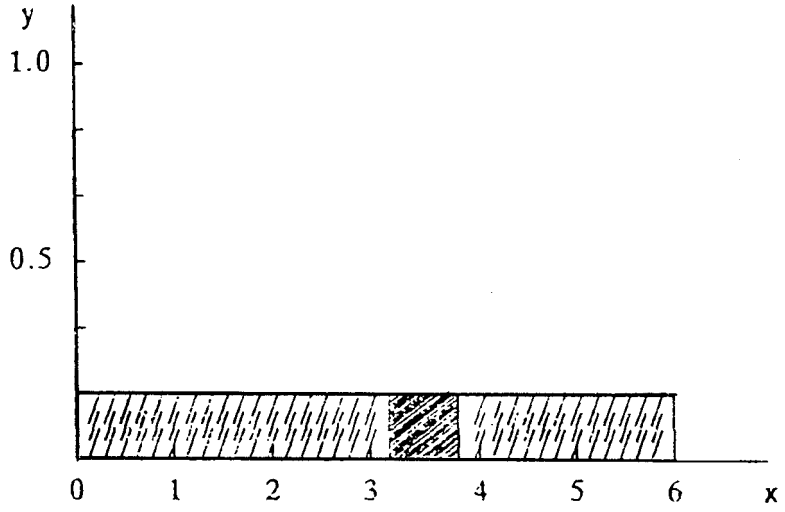
Ο έλεγχος αυτός είναι ίσως ο πιο χρήσιμος από τους ελέγχους καλής προσαρμογής, εν μέρει λόγω του ότι μας εφοδιάζει με μία εναλλακτική μέθοδο η οποία είναι σχεδιασμένη για διατεταγμένα δεδομένα (σε αντίθεση με τον έλεγχο χ^2 ο οποίος είναι σχεδιασμένος για δεδομένα σε ονομαστική κλίμακα) και, εν μέρει επειδή ο έλεγχος Kolmogorov στηρίζεται σε μια στατιστική συνάρτηση που δίνει την δυνατότητα ορισμού μιας *ζώνης εμπιστοσύνης (confidence band)* για την άγνωστη κατανομή, όπως θα δούμε αργότερα σε αυτό το κεφάλαιο.

Ας θεωρήσουμε το εξής πρόβλημα:

Εστω ότι έχουμε είκοσι τεμάχια νήματος μήκους ακριβώς έξι εκατοστών το καθένα. Το ένα από το δύο άκρα των νημάτων αυτών είναι στερεωμένο, ενώ στο άλλο άκρο εφαρμόζεται μία δύναμη μέχρι τα νήματα αυτά να κοπούν. Εάν το τεμάχιο του νήματος πάντοτε κόβεται στο ασθενέστερο σημείο του και το σημείο αυτό είναι εξίσου πιθανό να βρίσκεται οπουδήποτε κατά μήκος του τεμαχίου, τότε το σημείο θραύσης θα κατανέμεται ομοιόμορφα στο διάστημα (0,6), όπου η απόσταση μέχρι το σημείο θραύσης μετριέται σε εκατοστόμετρα από το στερεωμένο άκρο του νήματος. Τότε, η συνάρτηση πυκνότητας πιθανότητας της απόστασης X του σημείου θραύσης των νημάτων από το στερεωμένο άκρο τους, έχει την μορφή

$$f(x)=\begin{cases}0, x \le 0\\ 1/6, 0 < x \le 6\\ 0, x > 6\end{cases}$$

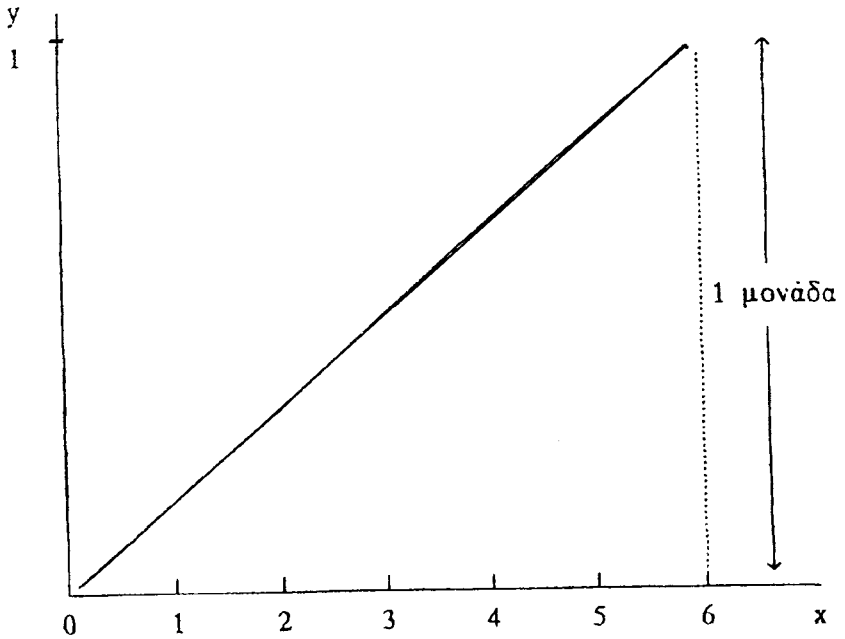
Το σχήμα 4.2.1 απεικονίζει την παραπάνω συνάρτηση πυκνότητας πιθανότητας



Σχήμα 4.2.1

Συνάρτηση πυκνότητας πιθανότητας της ομοιόμορφης κατανομής στο διάστημα (0,6)

Ο έλεγχος Kolmogorov δεν χρησιμοποιεί την συνάρτηση πυκνότητας πιθανότητας, αλλά την αντίστοιχη αθροιστική συνάρτηση κατανομής $F(x) = P(X \leq x)$, για κάθε x στο διάστημα (0,6). Η συνάρτηση αυτή απεικονίζεται στο σχήμα 4.2.2.



Σχήμα 4.2.2

Αθροιστική συνάρτηση κατανομής της ομοιόμορφης κατανομής στο διάστημα (0,6)

Για να ελεγχθεί η υπόθεση ότι οι 20 παρατηρήσεις που έχουμε αποτελούν ένα δείγμα από την ομοιόμορφη κατανομή στο διάστημα (0,6), θα πρέπει οι παρατηρήσεις αυτές να *συγκριθούν*, με κάποια έννοια, με την F(x). Ενας λογικός τρόπος σύγκρισης του τυχαίου δείγματος που έχουμε με την F(x) είναι μέσω της *εμπειρικής συνάρτησης κατανομής* S(x), η οποία, από τον ορισμό της, είναι το ποσοστό των παρατηρήσεων του δείγματος οι οποίες είναι το πολύ ίσες με την τιμή x, για κάθε x στο διάστημα $(-\infty, +\infty)$. Γενικότερα, δηλαδή, ο έλεγχος Kolmogorov συγκρίνει μία αθροιστική συνάρτηση κατανομής F(x) με την αντίστοιχη εμπειρική συνάρτηση κατανομής S(x) βασισμένη σε ένα δείγμα n παρατηρήσεων, όπου βέβαια

$$S(x) = \frac{\text{αριθμός τιμών του δείγματος που είναι το πολύ ίσες με } x}{n}.$$

Εάν οι δύο αυτές συναρτήσεις δεν συμφωνούν σε ικανοποιητικό βαθμό, τότε θα μπορούσαμε να απορρίψουμε την μηδενική υπόθεση και να συμπεράνουμε ότι η πραγματική, αλλά άγνωστη συνάρτηση κατανομής F(x) δεν ορίζεται από την μηδενική υπόθεση.

Αλλά ποιά ελεγχοσυνάρτηση θα μπορούσαμε να χρησιμοποιήσουμε ως ένα μέτρο της εγγύτητας μεταξύ της S(x) και της F(x);

Ένα από τα απλούστερα μέτρα που θα μπορούσε να θεωρήσει κανείς είναι η μέγιστη κατακόρυφη απόσταση μεταξύ του γραφήματος της S(x) και αυτού της F(x). Αυτή πράγματι ήταν και η στατιστική συνάρτηση που προτάθηκε από τον Kolmogorov το 1933.

Η μεθοδολογία που ο Kolmogorov εισήγαγε μπορεί ευκολότερα να κατανοηθεί γραφικά στο πλαίσιο του προηγούμενου προβλήματος:

Παράδειγμα 4.2.1: Ας υποθέσουμε ότι οι αποστάσεις του σημείου θραύσης των 20 νημάτων είναι οι εξής:

0.6	0.8	1.1	1.2	1.4	1.7	1.8	1.9	2.2	2.4
2.5	2.9	3.1	3.4	3.4	3.9	4.4	4.9	5.2	5.9

(Για ευκολία οι αποστάσεις έχουν διαταχθεί κατά αύξουσα σειρά μεγέθους).

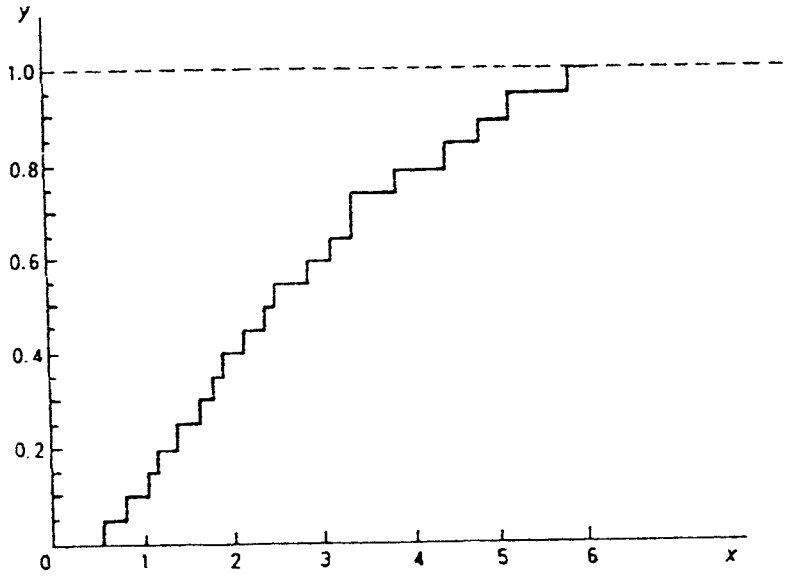
Για τον έλεγχο της υπόθεσης ότι η απόσταση του σημείου θραύσης από το στερεωμένο σημείο κατανέμεται ομοιόμορφα στο

διάστημα (0,6), θα πρέπει να προσδιορισθεί η κλιμακωτή συνάρτηση S(x):

Οταν x < 0.6 τότε S(x)=0. Στην τιμή 0.6 η συνάρτηση αυτή παρουσιάζει άλμα, οπότε παίρνει την τιμή 1/20=0.05, την οποία διατηρεί μέχρις ότου το x πάρει την τιμή 0.8, οπότε πηδά στην τιμή 0.10. Παρουσιάζει, δηλαδή, άλματα μεγέθους 0.05 σε κάθε σημείο θραύσης των νημάτων μέχρις ότου το x πάρει τη τιμή 3.4, οπότε η S(x) αυξάνει κατά 0.10, επειδή δύο θραύσεις παρατηρούνται στην τιμή x=3.4. Ο πίνακας που ακολουθεί παρουσιάζει την S(x) για κάθε x στο διάστημα $(-\infty, +\infty)$.

S(x)	για κάθε x στο διάστημα	S(x)	για κάθε x στο διάστημα
0.00	$-\infty < x < 0.6$	0.50	$2.4 \le x < 2.5$
0.15	$0.6 \le x < 0.8$	0.55	$2.5 \le x < 2.9$
0.10	$0.8 \le x < 1.1$	0.60	$2.9 \le x < 3.1$
0.15	$1.1 \le x < 1.2$	0.65	$3.1 \le x < 3.4$
0.20	$1.2 \le x < 1.4$	0.75	$3.4 \le x < 3.9$
0.25	$1.4 \le x < 1.7$	0.80	$3.9 \le x < 4.4$
0.30	$1.7 \le x < 1.8$	0.85	$4.4 \le x < 4.9$
0.35	$1.8 \le x < 1.9$	0.90	$4.9 \le x < 5.2$
0.40	$1.9 \le x < 2.2$	0.95	$5.2 \le x < 5.9$
0.45	$2.2 \le x < 2.4$	1.00	$5.9 \le x < +\infty$

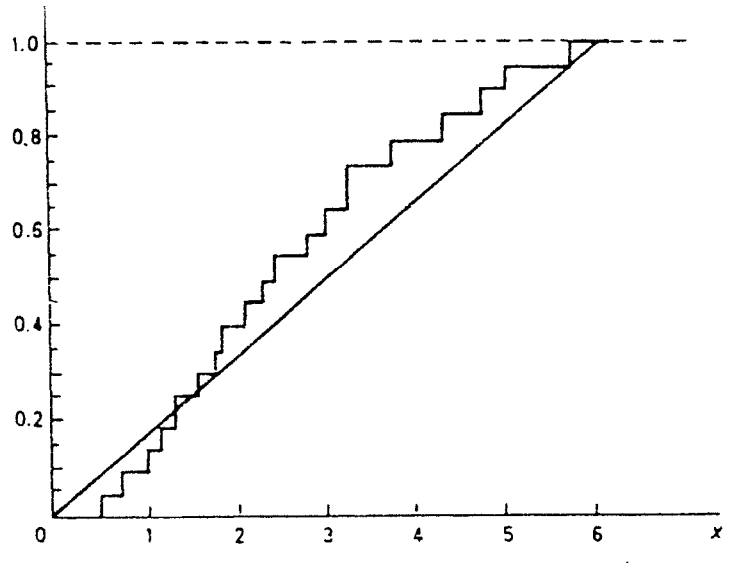
Η γραφική παράσταση της εμπειρικής συνάρτησης κατανομής δίνεται στο σχήμα 4.2.3.



Σχήμα 4.2.3

Εμπειρική συνάρτηση κατανομής των σημείων θραύσης των νημάτων

Η θεμελιώδης ιδέα του ελέγχου Kolmogorov είναι ότι η εμπειρική συνάρτηση κατανομής, ως εκτιμήτρια της αθροιστικής συνάρτησης κατανομής, δεν θα πρέπει να αποκλίνει σημαντικά από την τελευταία. Αναζητά, λοιπόν, ο έλεγχος αυτός την μέγιστη κατακόρυφη απόσταση μεταξύ των γραφημάτων της F(x) και της S(x). Αυτά απεικονίζονται στο σχήμα 4.2.4.



Σχήμα 4.2.4

Γραφική παράσταση των F(x) και S(x) των σημείων θραύσης των νημάτων

Από το σχήμα 4.2.4, είναι προφανές ότι η μέγιστη απόκλιση της S(x) από την F(x) παρατηρείται στο σημείο x=3.4. Ο γραφικός προσδιορισμός του σημείου στο οποίο παρατηρείται η μέγιστη απόκλιση της S(x) από την F(x) γίνεται με βάση τον πίνακα που ακολουθεί, ο οποίος περιέχει τις τιμές της διαφοράς F(x) – S(x). Ας σημειωθεί ότι, επειδή για το παράδειγμά μας,

$$f(x)=\begin{cases}0, & -\infty < x < 0\\ x/6, & 0 \le x < 6\\ 1, & 0 \le x < +\infty\end{cases}$$

και δοθέντος ότι η S(x) είναι σταθερή σε κάθε υποδιάστημα του $(-\infty, +\infty)$, η μέγιστη κατακόρυφη απόκλιση της S(x) από την F(x) σε κάθε υποδιάστημα παρατηρείται στο σημείο του διαστήματος

που αντιστοιχεί στο σημείο θραύσης, αφού στο σημείο αυτό η S(x) παρουσιάζει άλμα.

Πίνακας 4.2.1

Σύγκριση της F(x) με την S(x) των σημείων θραύσης των νημάτων

x_i	$F(x_i)$	$S(x_i)$	$F(x_i) - S(x_i)$	$F(x_i) - S(x_{i-1})$
0.6	0.10	0.05	0.05	0.10
0.8	0.13	0.10	0.03	0.08
1.1	0.18	0.15	0.03	0.08
1.2	0.20	0.20	0.00	0.05
1.4	0.23	0.25	–0.02	0.03
1.7	0.28	0.30	–0.02	0.03
1.8	0.30	0.35	–0.05	0.00
1.9	0.32	0.40	–0.08	–0.03
2.2	0.37	0.45	–0.08	–0.03
2.4	0.40	0.50	–0.10	–0.05
2.5	0.42	0.55	–0.13	–0.08
2.9	0.48	0.60	–0.12	–0.07
3.1	0.52	0.65	–0.13	–0.08
3.4	0.57	0.75	–0.18	–0.08
3.9	0.65	0.80	–0.15	–0.10
4.4	0.73	0.85	–0.12	–0.07
4.9	0.82	0.90	–0.08	–0.03
5.2	0.87	0.95	–0.08	–0.03
5.9	0.98	1.00	–0.02	–0.03

Από τον πίνακα 4.2.1, είναι προφανές ότι η μέγιστη κατ' απόλυτη τιμή απόκλιση της S(x) από την F(x) είναι 0.18 και παρατηρείται στο σημείο x=3.4.

Το επόμενο βήμα είναι να καθορισθεί η περιοχή απόρριψης της μηδενικής υπόθεσης ότι οι αποστάσεις των σημείων θραύσης από το στερεωμένο άκρο των νημάτων είναι ομοιόμορφα κατανεμημένες. Τότε, θα είμαστε σε θέση να συμπεράνουμε κατά πόσο μια μέγιστη απόκλιση μεγέθους 0.18 μπορεί να θεωρηθεί επαρκής ένδειξη υπέρ της απόρριψης της μηδενικής υπόθεσης σε κάποιο επίπεδο σημαντικότητας. Η μορφή της περιοχής απόρριψης εξαρτάται, βέβαια, από την μορφή της εναλλακτικής υπόθεσης. Ένας αμφίπλευρος έλεγχος θα ήταν κατάλληλος, αν η εναλλακτική μας υπόθεση ήταν ότι οι παρατηρήσεις μας έχουν προέλθει από κάποια άλλη κατανομή. Ομως, στην πράξη, θα μπορούσαμε να ενδιαφερόμαστε για μία εναλλακτική, η οποία θα υπαινισσόταν ότι το ποσοστό των φορών που το νήμα κόβεται σε απόσταση x από το στερεωμένο άκρο του νήματος είναι μεγαλύτερο (ή μικρότερο) από το αντίστοιχο ποσοστό της μηδενικής (ομοιόμορφης) κατανομής. Για παράδειγμα, εάν η δύναμη δεν ασκείται ομοιόμορφα σε όλο το μήκος του νήματος, αλλά είναι μεγαλύτερη προς το στερεωμένο άκρο του νήματος και μειώνεται σταθερά καθώς προχωράμε προς το άλλο άκρο του, τότε αυξάνεται η τάση να κόβεται το νήμα πιο κοντά στο στερεωμένο άκρο του. Με υψηλή την τιμή της πιθανότητας να κόβεται το νήμα προς το στερεωμένο άκρο του, είναι προφανές ότι η εμπειρική συνάρτηση κατανομής θα βρίσκεται παντού στο διάστημα $(-\infty, +\infty)$, υπεράνω της αθροιστικής συνάρτησης κατανομής F(x) της ομοιόμορφης κατανομής. Επομένως, μία μέγιστη απόκλιση με την S(x) μεγαλύτερη

από την F(x), εάν είναι στατιστικά σημαντική, θα συνηγορεί υπέρ της εναλλακτικής υπόθεσης.

Λύση με το MINITAB: Οπως έχουμε δει, το MINITAB υπολογίζει τις τάξεις μεγέθους των παρατηρήσεων μιας μεταβλητής, αλλά, σε αντίθεση με το SPSS, δίνει σε ισοβαθμούσες τιμές υποχρεωτικά τη μέση τιμή των τάξεων μεγέθους που θα είχαν αν δεν ταυτίζονταν. Αυτό όμως δεν προσφέρεται για περιπτώσεις όπως αυτές του παραδείγματός μας, όπου ζητάμε την εμπειρική συνάρτηση κατανομής, και, ως εκ τούτου, επιθυμούμε να αποδίδουμε στις ισοβαθμούσες τιμές την τάξη που θα είχε η μεγαλύτερη αν δεν ταυτίζονταν. (Ετσι, η τάξη μεγέθους κάθε παρατήρησης δείχνει πόσες παρατηρήσεις είναι μικρότερες ή ίσες από αυτήν). Άρα, θα πρέπει όλη η εργασία υπολογισμού της εμπειρικής συνάρτησης κατανομής να γίνει με το χέρι. Ετσι, δεν έχει νόημα να λυθεί το παράδειγμα με το MINITAB.

Λύση με το SPSS: Σε μία μεταβλητή (στην οποία έστω ότι δίνουμε το όνομα **x** και την ετικέτα **Απόσταση σημείου θραύσης**), καταχωρίζουμε το δείγμα. Οπως είναι ήδη γνωστό, το SPSS υπολογίζει, με την βοήθεια της εντολής **Transform**, **Rank Cases**, την εμπειρική συνάρτηση κατανομής ενός δείγματος. Αρκεί, στο πεδίο **Rank Types**, να επιλέξουμε **Fractional Ranks** για την τάξη μεγέθους και, στο πεδίο **Ties**, να επιλέξουμε **Rank assigned to Ties**, **High**. Το πρόγραμμα δίνει τη μεταβλητή **rx** (αφού **x** είναι το όνομα της μεταβλητής που περιέχει το δείγμα). Το φύλλο δεδομένων είναι:

	x	rx
1	.6	.0500
2	.8	.1000
3	1.1	.1500
4	1.2	.2000
5	1.4	.2500
6	1.7	.3000
7	1.8	.3500
8	1.9	.4000
9	2.2	.4500
10	2.4	.5000
11	2.5	.5500
12	2.9	.6000
13	3.1	.6500
14	3.4	.7500
15	3.4	.7500
16	3.9	.8000
17	4.4	.8500
18	4.9	.9000
19	5.2	.9500
20	5.9	1.0000

Για να διεξαγάγουμε τον έλεγχο Kolmogorov-Smirnov θα πρέπει, όπως είναι ήδη γνωστό να καταφύγουμε στο παράθυρο εντολών (**syntax window**), αφού μόνο έτσι μπορούμε να δηλώσουμε τις τιμές που θεωρούμε ότι έχουν οι παράμετροι της κατανομής κάτω από την μηδενική υπόθεση. Στο παράθυρο εντολών, δίνουμε την εντολή **npar tests k-s (uniform,0,6)=x.** Τα αποτελέσματα του ελέγχου είναι τα εξής:

One-Sample Kolmogorov-Smirnov Test

		Απόσταση σημείου θραύσης
N		20
Uniform Parameters[a,b]	Minimum	0

	Maximum	6
Most Extreme Differences	Absolute	.183
	Positive	.183
	Negative	-.100
Kolmogorov-Smirnov Z		.820
Asymp. Sig. (2-tailed)		.512

a Test distribution is Uniform.
b User-Specified

Η τιμή της στατιστικής συνάρτησης T (πεδίο **Most Extreme Difference: Absolute**) είναι 0.183. Οπως μπορεί εύκολα να διαπιστωθεί από τους πίνακες ποσοστιαίων σημείων της T, το κρίσιμο επίπεδο που αντιστοιχεί στην τιμή τ=0.183 είναι μεγαλύτερο από 0.2. Κατά συνέπεια, η μηδενική υπόθεση μπορεί να θεωρηθεί εύλογη σε όλα τα συνήθη επίπεδα σημαντικότητας. Η τιμή του κρίσιμου επιπέδου που δίνει το SPSS στην τελευταία γραμμή του πίνακα είναι ασυμπτωτική και επειδή το δείγμα μας είναι μικρό, δεν έχει έννοια να χρησιμοποιηθεί.

Στα επόμενα, διατυπώνεται με γενικότερο τρόπο η μορφή των προβλημάτων που αντιμετωπίζονται με τον έλεγχο Kolmogorov, η επιχειρηματολογία που αναπτύχθηκε για το είδος της στατιστικής συνάρτησης που απαιτείται για να *μετρηθεί* ο βαθμός της εγγύτητας των τιμών των συναρτήσεων F(x) και S(x) καθώς και η διεξαγωγή του ελέγχου.

Εστω λοιπόν X_1, X_2, ..., X_n ένα τυχαίο δείγμα μεγέθους n από κάποια άγνωστη κατανομή με αθροιστική συνάρτηση κατανομής F(x).

Εστω, επίσης, $F_0(x)$ μία σαφώς καθορισμένη συνάρτηση κατανομής. Οι υποθέσεις που ενδιαφέρει να ελεγχθούν διακρίνονται στις εξής τρεις κατηγορίες:

Α. (Αμφίπλευρη εναλλακτική υπόθεση)

H_0: $F(x) = F_0(x)$, για κάθε x στο $(-\infty, +\infty)$

H_1: $F(x) \neq F_0(x)$, για τουλάχιστον μια τιμή του x.

Β. (Μονόπλευρη εναλλακτική υπόθεση)

H_0: $F(x) \geq F_0(x)$, για κάθε x στο $(-\infty, +\infty)$

H_1: $F(x) < F_0(x)$, για τουλάχιστον μια τιμή του x.

Γ. (Μονόπλευρη εναλλακτική υπόθεση)

H_0: $F(x) \leq F_0(x)$, για κάθε x στο $(-\infty, +\infty)$

H_1: $F(x) > F_0(x)$, για τουλάχιστον μια τιμή του x.

Εστω S(x) η εμπειρική συνάρτηση κατανομής με βάση το τυχαίο δείγμα X_1, X_2, ..., X_n. Η στατιστική συνάρτηση ελέγχου ορίζεται διαφορετικά για τα τρία διαφορετικά σύνολα υποθέσεων Α, Β, και Γ.

Α. (Αμφίπλευρος έλεγχος): Η κατάλληλη στατιστική συνάρτηση ελέγχου για την περίπτωση αυτή είναι η μέγιστη κατακόρυφη απόσταση μεταξύ των συναρτήσεων S(x) και $F_0(x)$. Συμβολικά, γράφουμε

$$T = \sup_x \left| F_0(x) - S(x) \right|,$$

και διαβάζουμε "*η* T *είναι ίση με το supremum, για όλα τα* x *της απόλυτης τιμής της διαφοράς* $F_0(x) - S(x)$".

Β. (Μονόπλευρος έλεγχος): Είναι προφανές ότι, στην περίπτωση αυτή, η κατάλληλη στατιστική συνάρτηση ελέγχου είναι η μέγιστη κατακόρυφη απόσταση που έχει η $F_0(x)$ υπεράνω της S(x). Δηλαδή,

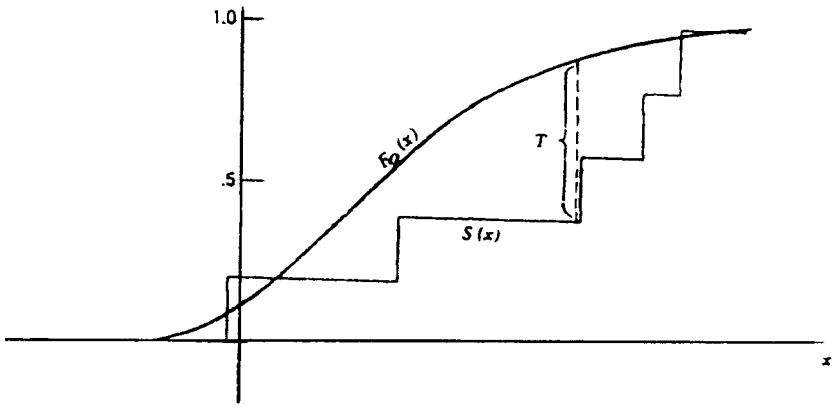
$$T^+ = \sup_{x\,:\,F_0(x) > S(x)} \left[F_0(x) - S(x)\right].$$

Η στατιστική αυτή συνάρτηση είναι παρόμοια με την T, με την διαφορά ότι, τώρα, θεωρούμε μόνο την μέγιστη διαφορά όταν η συνάρτηση $F_0(x)$ είναι υπεράνω της συνάρτησης S(x).

Γ. (Μονόπλευρος έλεγχος): Με ανάλογο τρόπο, στην περίπτωση αυτή, η κατάλληλη στατιστική συνάρτηση ελέγχου ορίζεται ως η μέγιστη κατακόρυφη απόσταση που έχει η S(x) υπεράνω της $F_0(x)$. Συμβολικά, γράφουμε

$$T^- = \sup_{x : F_0(x) < S(x)} [S(x) - F_0(x)].$$

και διαβάζουμε "T^- *είναι το supremum της διαφοράς* S(x) – F_0(x), *για όλα τα* x *για τα οποία η συνάρτηση* S(x) *είναι υπεράνω της συνάρτησης* F_0(x)".



Σχήμα 4.2.5

Η υποτεθείσα (μηδενική) συνάρτηση κατανομής F_0(x), η εμπειρική συνάρτηση κατανομής S(x) και η στατιστική συνάρτηση του Kolmogorov, T.

Είναι προφανές ότι, και στις τρεις περιπτώσεις, οι μεγάλες τιμές της στατιστικής συνάρτησης είναι αυτές που συνηγορούν υπέρ της απόρριψης της μηδενικής υπόθεσης αφού αντανακλούν χαμηλό βαθμό

εγγύτητας των τιμών των συναρτήσεων S(x) και $F_0(x)$. Η ακριβής κατανομή των στατιστικών συναρτήσεων T, T^+ και T^- έχει μελετηθεί από τον Kolmogorov. Τα (1–α)-ποσοστιαία σημεία των κατανομών αυτών που θα συμβολίζουμε με $w_{1-\alpha}$ περιέχονται στον πίνακα 10 του παραρτήματος, τόσο για τους μονόπλευρους ελέγχους όσο και για τον αμφίπλευρο έλεγχο. Ο πίνακας αυτός είναι ακριβής μόνο εάν η $F_0(x)$ είναι συνεχής κατανομή. Διαφορετικά, αυτά τα ποσοστιαία σημεία οδηγούν σε ένα συντηρητικό έλεγχο. Ο πίνακας 10 περιέχει ποσοστιαία σημεία για τους αμφίπλευρους ελέγχους σε επίπεδο σημαντικότητας α=0.20, 0.10, 0.05, 0.02 και 0.01, και για μονόπλευρους ελέγχους σε επίπεδο σημαντικότητας α=0.10, 0.05, 0.025, 0.01 και 0.005. Ο πίνακας είναι ακριβής για $n \leq 20$ στους αμφίπλευρους ελέγχους. Για τις περιπτώσεις που το μέγεθος του δείγματος n υπερβαίνει το 20 ($n > 20$), όπως και στην περίπτωση των μονόπλευρων ελέγχων, οι πίνακες δίνουν ικανοποιητικές προσεγγίσεις, οι οποίες, στις περισσότερες περιπτώσεις, συμπίπτουν με τις ακριβείς τιμές. Η προσέγγιση για $n > 40$ εξαρτάται από την ασυμπτωτική κατανομή της στατιστικής συνάρτησης και δεν είναι πολύ ακριβής παρά μόνο στις περιπτώσεις που το n είναι πολύ μεγάλο.

Παρατήρηση: Είναι προφανές ότι $T=\max(T^+, T^-)$.

Επιστρέφοντας στο παράδειγμα με τα σημεία θραύσης των νημάτων και υποθέτοντας ότι ενδιαφερόμαστε για μία αμφίπλευρη εναλλακτική υπόθεση (περίπτωση Α), θα έχουμε ότι η παρατηρούμενη τιμή της Τ είναι τ=0.18. Η τιμή αυτή δεν υπερβαίνει το 0.95-ποσοστιαίο σημείο της κατανομής της Τ που αντιστοιχεί στον αμφίπλευρο έλεγχο για n=20, αφού $w_{0.95}=0.294$. Επομένως, σε επίπεδο

σημαντικότητας 5%, η μηδενική υπόθεση δεν απορρίπτεται. Εάν, από την άλλη μεριά, ενδιαφερόμαστε να ελέγξουμε τις υποθέσεις της περίπτωσης Γ σε επίπεδο σημαντικότητας α=0.05, θα είχαμε ότι η τιμή 0.18 της στατιστικής συνάρτησης T^- δεν υπερβαίνει και πάλι την τιμή του αντίστοιχου ποσοστιαίου σημείου, όπως αυτό προκύπτει από τον πίνακα 10 για τον μονόπλευρο έλεγχο σε επίπεδο σημαντικότητας 5% ($w_{0.95}$=0.265). Επομένως, και στις δύο περιπτώσεις οι ενδείξεις που έχουμε από το δείγμα δεν είναι επαρκείς για να οδηγήσουν σε απόρριψη της μηδενικής υπόθεσης.

Το παρατηρούμενο επίπεδο σημαντικότητας είναι, από τον πίνακα 10 του παραρτήματος για την περίπτωση του αμφίπλευρου ελέγχου, ίση με

$$P(T \geq 0.18 \mid H_0) = 1 - P(T < 0.18) > 1 - P(T < 0.232) = 1 - 0.80 = 0.20.$$

Αντίστοιχα, το παρατηρούμενο επίπεδο σημαντικότητας για την περίπτωση του μονόπλευρου ελέγχου είναι

$$\hat{\alpha} = P(T^- \geq 0.18 \mid H_0) = 1 - P(T^- \leq 0.18)$$
$$> 1 - P(T^- \leq 0.232) = 1 - 0.90 = 0.10.$$

Ο έλεγχος Kolmogorov συχνά προτιμάται από τον έλεγχο χ^2 ως έλεγχος καλής προσαρμογής, όταν το μέγεθος του δείγματος είναι μικρό. Ο έλεγχος Kolmogorov είναι ακριβής ακόμη και για μικρά δείγματα, ενώ ο έλεγχος χ^2 υποθέτει ότι ο αριθμός των παρατηρήσεων είναι αρκετά μεγάλος, ώστε η κατανομή χ^2 να παρέχει μία ικανοποιητική προσέγγιση της κατανομής της στατιστικής συνάρτησης. Υπάρχουν πολλές αντιφάσεις όσο αφορά το ποιός έλεγχος είναι περισσότερο ισχυρός, αλλά η γενική εντύπωση φαίνεται να είναι ότι ο έλεγχος Kolmogorov είναι ενδεχομένως περισσότερο ισχυρός από τον έλεγχο χ^2 στις περισσότερες περιπτώσεις. Ενας προβληματισμός για τον έλεγχο Kolmogorov

αναφέρεται στο κατά πόσο αυτός αγνοεί πληροφορίες με το να χρησιμοποιεί μόνο την διαφορά με το μεγαλύτερο μέγεθος σε αντίθεση με τους ελέγχους οι οποίοι λαβαίνουν υπόψη τους όλες τις διαφορές. Τα πλεονεκτήματα των τελευταίων δεν είναι τόσα πολλά γιατί η τιμή της συνάρτησης S(x) εξαρτάται σε κάθε στάδιο από το πόσες παρατηρήσεις είναι μικρότερες από την συγκεκριμένη τιμή x. Επομένως, οι συγκρίσεις γίνονται με βάση τις συσσωρευμένες ενδείξεις μέχρι το συγκεκριμένο στάδιο.

Παράδειγμα 4.2.2: Εστω ότι από μία κατανομή με συνάρτηση κατανομής F(x) παίρνουμε το εξής τυχαίο δείγμα:

0.621 0.503 0.203 0.477 0.710

0.581 0.329 0.480 0.554 0.382.

Να ελεγχθεί η μηδενική υπόθεση ότι το δείγμα έχει προέλθει από την ομοιόμορφη κατανομή στο διάστημα (0,1) έναντι της εναλλακτικής ότι τα δείγμα έχει προέλθει από μία κατανομή της οποίας η αθροιστική συνάρτηση κατανομής, για μία τουλάχιστον τιμή του x, είναι μικρότερη από την αντίστοιχη τιμή της αθροιστικής συνάρτησης κατανομής της ομοιόμορφης κατανομής στο διάστημα (0,1).

Λύση: Είναι προφανές ότι η κατάλληλη στατιστική συνάρτηση για τον έλεγχο αυτών των υποθέσεων είναι η

$$T^{+} = \sup_{x:F_0(x)>S(x)} \left[F_0(x) - S(x)\right].$$

Εδώ,

$$F_0(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \le x < 1 \\ 1, & x \ge 1. \end{cases}$$

Για τον προσδιορισμό της συνάρτησης S(x) διατάσσουμε πρώτα το δείγμα κατά αύξουσα σειρά μεγέθους:

0.203	0.329	0.382	0.477	0.480
0.503	0.554	0.581	0.621	0.710.

Επομένως,

$$S(x) = \begin{cases} 0, & x < 0.203 \\ 0.1, & 0.203 \le x < 0.329 \\ 0.2, & 0.329 \le x < 0.382 \\ 0.3, & 0.382 \le x < 0.477 \\ 0.4, & 0.477 \le x < 0.480 \\ 0.5, & 0.480 \le x < 0.503 \\ 0.6, & 0.503 \le x < 0.554 \\ 0.7, & 0.554 \le x < 0.581 \\ 0.8, & 0.581 \le x < 0.621 \\ 0.9, & 0.621 \le x < 0.710 \\ 1, & x \ge 0.710. \end{cases}$$

Έχουμε, κατά συνέπεια, ότι

$$\begin{aligned} T^+ &= \sup_x \left[F_0(x) - S(x)\right] \\ &= \left[F_0(0.329) - S(0.329)\right] \\ &= 0.2289. \end{aligned}$$

Οπως προκύπτει από τον πίνακα 14 του παραρτήματος, $\hat{\alpha} > 0.10$. Επομένως, η μηδενική υπόθεση δεν απορρίπτεται σε οποιοδήποτε επίπεδο σημαντικότητας είναι μικρότερο από το $\hat{\alpha}$.

Εάν υποθέσουμε ότι η εναλλακτική υπόθεσή μας ήταν αμφίπλευρη, εάν, δηλαδή, υποθέσουμε ότι είχαμε να ελέγξουμε την υπόθεση ότι το δείγμα μας έχει προέλθει από την ομοιόμορφη κατανομή στο διάστημα (0,1) έναντι της εναλλακτικής ότι η κατανομή από την οποία προήλθε το δείγμα δεν είναι η ομοιόμορφη στο διάστημα (0,1), τότε η κατάλληλη στατιστική συνάρτηση θα ήταν η

$$T = \sup_x |F_0(x) - S(x)|.$$

Στην περίπτωση αυτή, θα είχαμε

$$T = \sup_x |F_0(x) - S(x)| = |F_0(0.710) - S(0.710)|$$

$$= |0.710 - 1| = 0.290.$$

Επομένως, σε επίπεδο σημαντικότητας 0.05, δεν θα απορρίπταμε την μηδενική υπόθεση αφού, όπως προκύπτει από τον πίνακα 14 του παραρτήματος,

$$\hat{\alpha} = P(T \geq 0.290 \mid H_0) = 1 - P(T < 0.290)$$
$$> 1 - P(T < 0.323) = 1 - 0.80 = 0.20.$$

Πράγματι, το 0.95-ποσοστιαίο σημείο της στατιστικής συνάρτησης T, όπως αυτό δίνεται στον πίνακα 14 του παραρτήματος είναι $w_{0.95} = 0.409$.

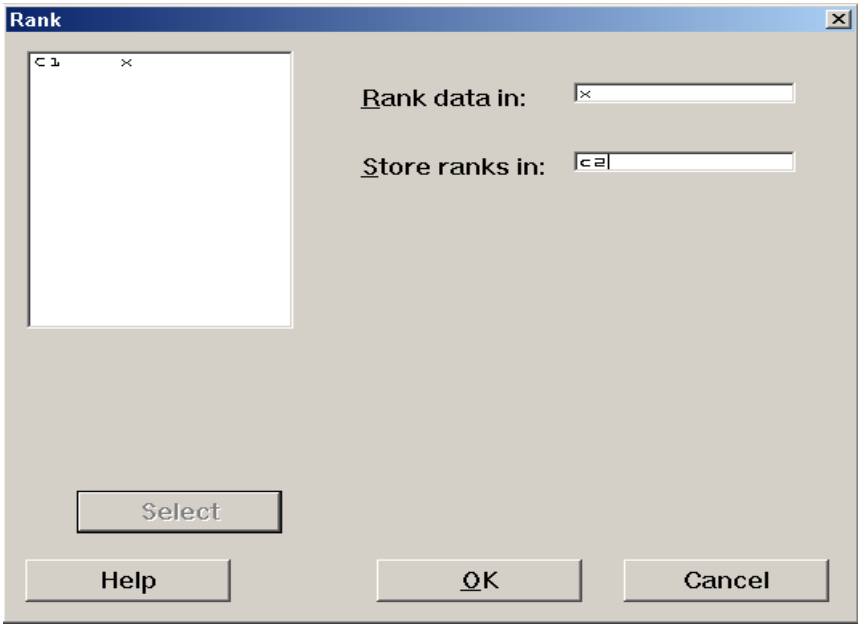
Λύση με το MINITAB: Η απ' ευθείας διεξαγωγή του ελέγχου Kolmogorov-Smirnov για ένα δείγμα είναι δυνατή με το MINITAB μόνο στις περιπτώσεις όπου, κάτω από την H_0, το δείγμα προέρχεται από την κανονική κατανομή. Το παράδειγμα, επομένως, θα πρέπει να λυθεί έμμεσα. Συγκεκριμένα, θα υπολογίσουμε, για κάθε παρατήρηση του δείγματος, την τιμή της εμπειρικής και της αθροιστικής συνάρτησης κατανομής κάτω από την μηδενική υπόθεση και θα προσδιορίσουμε την μέγιστη απόλυτη διαφορά τους. Για τον υπολογισμό της εμπειρικής συνάρτησης κατανομής ταξινομούμε το δείγμα κατ' αύξουσα τάξη μεγέθους, υπολογίζουμε τις τάξεις μεγέθους των παρατηρήσεων και τις διαιρούμε με το μέγεθος του δείγματος.

Παρατήρηση: Επειδή θέλουμε να υπολογίσουμε συνάρτηση κατανομής, σε περίπτωση ισοβαθμιών, δεν αποδίδουμε στις ισοβαθμούσες τιμές τη μέση τιμή των τάξεων μεγέθους που αυτές θα

είχαν αν δεν ισοβαθμούσαν, αλλά την τάξη μεγέθους που θα είχε η μεγαλύτερη από αυτές αν δεν ταυτίζονταν. (Ετσι, η τάξη μεγέθους κάθε παρατήρησης δείχνει πόσες παρατηρήσεις είναι μικρότερες ή ίσες από συτήν).

Οπως αναφέρθηκε στο προηγούμενο παράδειγμα, το MINITAB υποχρεωτικά αντιστοιχίζει την μέση τάξη μεγέθους σε ισοβαθμούσες παρατηρήσεις και, επομένως, στην περίπτωση ισοβαθμιών, ο υπολογισμός των τάξεων μεγέθους πρέπει να γίνει «με το χέρι». Μία τέτοια διαδικασία είναι χρονοβόρα και μπορεί να οδηγήσει σε λάθη, ιδιαίτερα στην περίπτωση μεγάλων δειγμάτων. Στην περίπτωση του παραδείγματος δεν υπάρχουν ισοβαθμίες τιμών.

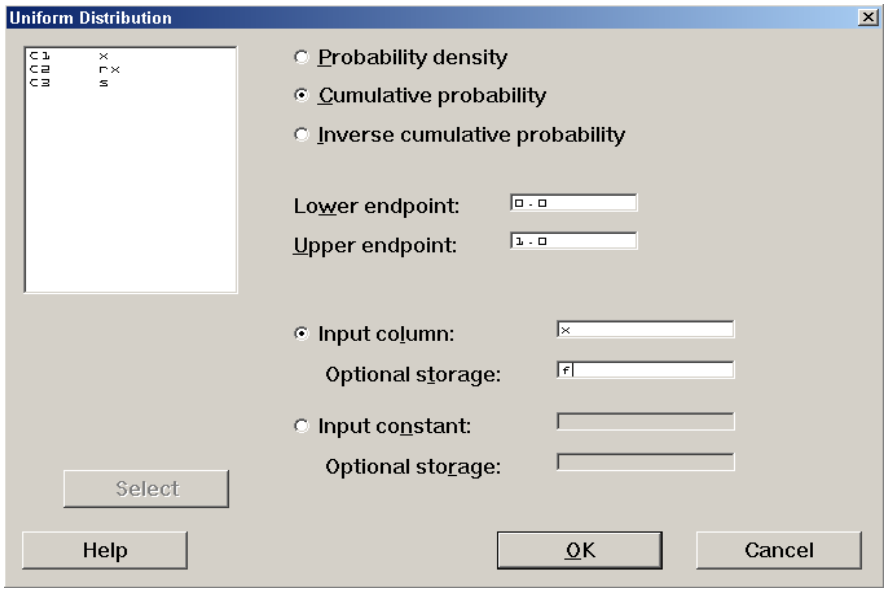
Καταχωρίζουμε το δείγμα σε μία μεταβλητή με όνομα **x** και το ταξινομούμε κατ' αύξουσα τάξη μεγέθους από την επιλογή **Sort**. Σε μία άλλη στήλη, με όνομα **rx**, καταχωρίζουμε τις τάξεις μεγέθους των παρατηρήσεων. Αυτό γίνεται με την επιλογή **Manip, Rank**, η οποία οδηγεί στο ακόλουθο πλαίσιο διαλόγου:



Στο πεδίο **Rank data in**, δηλώνουμε την μεταβλητή της οποίας τις τάξεις μεγέθους θα υπολογίσουμε και στο πεδίο **Store ranks in**, δηλώνουμε μία στήλη, στην οποία αυτές θα καταχωρισθούν. Στην συνέχεια μπορούμε, από το φύλλο δεδομένων, να δώσουμε όνομα στην στήλη αυτή.

Διαιρώντας τις τάξεις μεγέθους με το μέγεθος του δείγματος, παίρνουμε τις τιμές της εμπειρικής συνάρτησης κατανομής S(.) που αντιστοιχούν στις παρατηρήσεις του δείγματος. Τις τιμές αυτές καταχωρίζουμε στην μεταβλητή **s**. Οι αντίστοιχες τιμές της αθροιστικής συνάρτησης κατανομής που δηλώνει η μηδενική υπόθεση ($F_0(.)$) υπολογίζονται ως εξής:

Επιλέγουμε **Calc, Probability Distributions, Uniform** και οδηγούμεθα στο εξής πλαίσιο διαλόγου:



Επιλέγουμε **Cumulative probability** και, στα πεδία **Lower endpoint** και **Upper endpoint**, δηλώνουμε τις παραμέτρους της ομοιόμορφης κατανομής, της οποίας θέλουμε να υπολογίσουμε την συνάρτηση

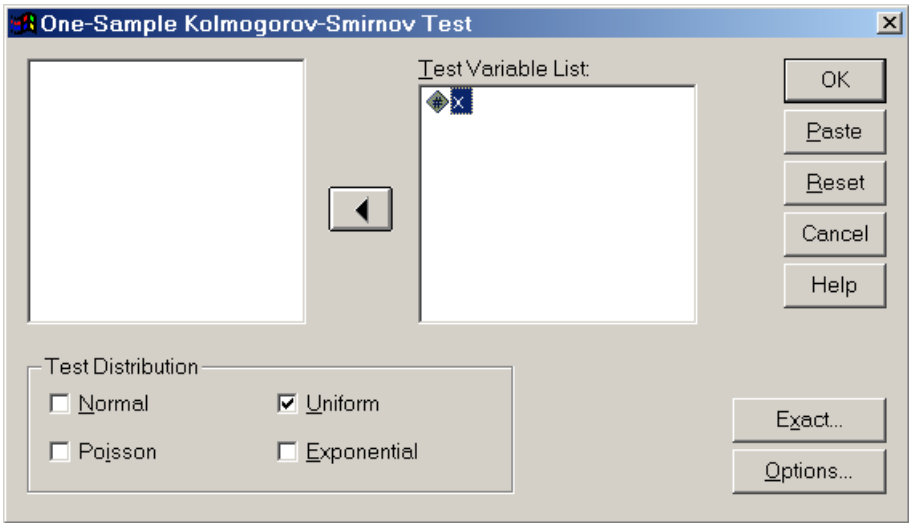
κατανομής. Στο πεδίο **Input column**, δηλώνουμε την μεταβλητή **x** στης τιμές της οποίας θέλουμε να υπολογίσουμε την συνάρτηση κατανομής $F_0(.)$ και, στο πεδίο **Optional storage**, δηλώνουμε μία μεταβλητή (έστω **f**) στην οποία θα αποθηκευθούν οι τιμές της συνάρτησης $F_0(.)$.

Επειδή πρέπει να υπολογίσουμε τις διαφορές της μορφής $|F_0(x_i) - S(x_i)|$ και $|F_0(x_i) - S(x_{i-1})|$, δημιουργούμε μία ακόμη στήλη (έστω **s1**) για τις τιμές της $S(x_{i-1})$. Αυτό επιτυγχάνεται θέτοντας την πρώτη τιμή της στήλη αυτής ίση με 0 και κάτω από αυτή να καταχωρίσουμε τις n–1 (9 στην περίπτωση μας) πρώτες τιμές της στήλης s. Σε δύο στήλες, έστω **d1** και **d2**, καταχωρίζουμε τις απόλυτες τιμές των διαφορών **f–s** και **f–s1**, αντίστοιχα και υπολογίζουμε την μέγιστη τιμή τους. Αυτή είναι η τιμή της ελεγχοσυνάρτησης T. Μετά τον υπολογισμό των διαφορών αυτών (από την επιλογή **Calc**, **Calculator** και πληκτρολογώντας **ABSO(f–s)** και **ABSO(f–s1)**) στο πεδίο **Numeric Expression**, το φύλλο δεδομένων θα δείχνει ως εξής:

x	rx	s	s1	f	d1	d2
0.203	1	0.1	0.0	0.203	0.103	0.203
0.329	2	0.2	0.1	0.329	0.129	0.229
0.382	3	0.3	0.2	0.382	0.082	0.182
0.477	4	0.4	0.3	0.477	0.077	0.177
0.480	5	0.5	0.4	0.480	0.020	0.080
0.503	6	0.6	0.5	0.503	0.097	0.003
0.554	7	0.7	0.6	0.554	0.146	0.046
0.581	8	0.8	0.7	0.581	0.219	0.119
0.621	9	0.9	0.8	0.621	0.279	0.179
0.710	10	1.0	0.9	0.710	0.290	0.190

Επομένως, η τιμή της ελεγχοσυνάρτησης τ είναι τ=0.29. Σύγκριση τςη τιμής αυτής με την κρίσιμη τιμή $w_{0.95}$=0.409 δείχνει ότι η μηδενική μπορεί να θεωρηθεί εύλογη.

Λύση με το SPSS: Για την διεξαγωγή του ελέγχου Kolmogorov-Smirnov για ένα δείγμα με το SPSS εργαζόμαστε ως εξής: Καταχωρίζουμε το δείγμα μας σε μία μεταβλητή (έστω **x**) και επιλέγουμε **Analyze, Nonparametric Tests, 1-Sample K-S** οδηγούμενοι στο παρακάτω πλαίσιο διαλόγου:



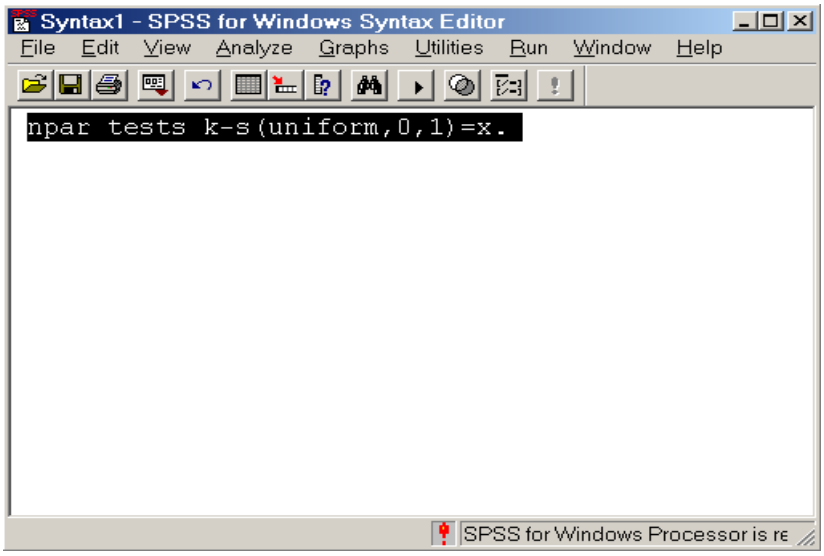
Στο πεδίο **Test Variable List**, δηλώνουμε το όνομα της μεταβλητής που περιέχει το δείγμα (**x**). Στο πεδίο **Test Distribution**, επιλέγουμε την κατανομή που ισχύει κάτω από την H_0 ($F_0(.)$). Δεν υπάρχει τρόπος να δηλώσουμε τις παραμέτρους της κατανομής αυτής. Επομένως, ως τιμές των παραμέτρων της ομοιόμορφης κατανομής της μηδενικής υπόθεσης του παραδείγματος θα θεωρηθούν η μικρότερη και η μεγαλύτερη τιμή του δείγματος αντί των τιμών 0 και 1. Τα αποτελέσματα που δίνει ο έλεγχος είναι τα εξής:

One-Sample Kolmogorov-Smirnov Test

		X
N		10
Uniform Parameters[a,b]	Minimum	.203
	Maximum	.710
Most Extreme Differences	Absolute	.240
	Positive	.100
	Negative	-.240
Kolmogorov-Smirnov Z		.760
Asymp. Sig. (2-tailed)		.610
Exact Sig. (2-tailed)		.533
Point Probability		.000

a Test distribution is Uniform.
b Calculated from data.

Οι διαφορές που δίνει το πρόγραμμα (**Absolute**=T, **Positive**=T^+, **Negative**=T^-) αντιστοιχούν σε σύγκριση των τιμών της εμπειρικής κατανομής με ομοιόμορφη κατανομή διαφορετική από αυτήν που δηλώνει η H_0 (δηλ. την F_0). Συνεπώς, δεν μπορούν να ληφθούν υπόψη. Ο τρόπος με τον οποίο μπορούμε να δηλώσουμε την σωστή ομοιόμορφη κατανομή είναι να καταφύγουμε στο παράθυρο εντολών (**syntax window**), το οποίο καλείται με την επιλογή **File, New, Syntax**. (Οι πλήρεις δυνατότητες του SPSS είναι διαθέσιμες μόνο από το **syntax window**). Για να διεξαγάγουμε τον σωστό έλεγχο, στο παράθυρο εντολών πληκτρολογούμε την εντολή **Npar tests k-s (uniform,0,1)=x**. Αυτή ζητάει από το SPSS να διεξαγάγει έλεγχο Kolmogorov-Smirnov προκειμένου να διαπιστώσει αν το δείγμα που είναι καταχωρισμένο στην μεταβλητή **x** μπορεί να θεωρηθεί ότι προέρχεται από την ομοιόμορφη κατανομή στο διάστημα (0, 1). Στην συνέχεια, επιλέγουμε την εντολή αυτή και πιέζουμε το πλήκτρο εκτέλεσης  (**Run Current**) όπως φαίνεται στην παρακάτω εικόνα:



Ο πίνακας που ακολουθεί περιέχει τώρα τα σωστά αποτελέσματα.

One-Sample Kolmogorov-Smirnov Test

		X
N		10
Uniform Parameters[a,b]	Minimum	0
	Maximum	1
Most Extreme Differences	Absolute	.290
	Positive	.290
	Negative	-.229
Kolmogorov-Smirnov Z		.917
Asymp. Sig. (2-tailed)		.370

a Test distribution is Uniform.
b User-Specified

Παρατηρούμε ότι η τιμή της ελεγχοσυνάρτησης T είναι 0.29 (πεδίο **Absolute**). Επίσης 0.29 είναι και η τιμή της ελεγχοσυνάρτησης T^+ (πεδίο **Positive**), ενώ –0.229 είναι η τιμή της ελεγχοσυνάρτησης T^- (πεδίο **Negative**). Για την ακρίβεια, στο πεδίο **Negative**, εμφανίζεται η αρνητική τιμή αυτού που στην θεωρία έχουμε συμβολίσει T^-. Το πεδίο T^- δίνει την τιμή της ελεγχοσυνάρτησης $\sqrt{n}\,T$ (πεδίο **Kolmogorov-Smirnov Z**) και το ασυμπτωτικό αμφίπλευρο κρίσιμο επίπεδο που είναι ίσο με 0.37. Επειδή στο συγκεκριμένο παράδειγμα $T=T^+$ και

επειδή T^+ είναι η ελεγχοσυνάρτηση που μας ενδιαφέρει, το μονόπλευρο κρίσιμο επίπεδο είναι 0.37/2=0.185.

Σημείωση: Κάθε εντολή στο **syntax window** πρέπει να τερματίζεται με τελεία.

Λύση με το SAS: Οπως και στο MINITAB, το SAS δεν παρέχει τον έλεγχο Kolmogorov-Smirnov ενός δείγματος για τον έλεγχο της υπόθεσης ότι το δείγμα προέρχεται από την ομοιόμορφη κατανομή. Έτσι, μπορούμε έμμεσα να διεξαγάγουμε τον έλεγχο χρησιμοποιώντας τις παρακάτω εντολές.

```
data uniform;
input x @@;
cards;
0.621 0.503 0.203 0.477 0.710
0.581 0.329 0.480 0.554 0.382
;
run;
proc rank data=uniform
        out=ranks
        ties=high
        fraction;
var x;
ranks r;
run;
data all; merge uniform(in=aa)ranks(in=bb);
if aa and bb;
data all;
set all;
r2=r-0.1;
f=cdf('uniform',x,0,1);
diff1=abs(f-r);
diff2=abs(f-r2);
run;
proc print; run;
proc means max;
 var diff1 diff2;
run;
```

Με τις εντολές **proc rank data=uniform out=ranks ties=high fraction;** υπολογίζουμε τις τάξεις μεγέθους που θα αποθηκευθούν στο σύνολο δεδομένων **ranks**. Με την υποεντολή **ties=high**, το πακέτο αποδίδει την μέγιστη τάξη μεγέθους σε περιπτώσεις ισοβαθμιών. (Αν δεν χρησιμοποιηθεί η υποεντολή αυτή, το πακέτο αποδίδει την μέση τάξη μεγέθους στις ισοβαθμίες). Με την υποεντολή **fraction**, οι τάξεις μεγέθους διαιρούνται με το μέγεθος του δείγματος, έτσι ώστε οι τάξεις

μεγέθους που προκύπτουν να μπορούν να θεωρηθούν τιμές από την εμπειρική αθροιστική συνάρτηση κατανομής. Οι τιμές αυτές έχουν καταχωρισθεί στην μεταβλητή με όνομα **r**. Επίσης, ορίζουμε την μεταβλητή **r2** ως την μεταβλητή **r** μειωμένη κατά 1/n.

Με την εντολή **f=cdf('uniform',x,0,1);** υπολογίζουμε τις θεωρητικές τιμές της αθροιστικής συνάρτησης της ομοιόμορφης κατανομής στο διάστημα (0,1), με τυχαίες μεταβλητές τις τιμές του δείγματος. Τέλος, με τις εντολές **diff1=abs(f-r); diff2=abs(f-r2);** υπολογίζουμε τις απόλυτες διαφορές μεταξύ των θεωρητικών και των εμπειρικών τιμών, ενώ με την εντολή **proc means max;** το πακέτο δίνει τις μέγιστες απόλυτες διαφορές. Τα αποτελέσματα είναι τα εξής:

The SAS System

OBS	X	R	R2	F	DIFF1	DIFF2
1	0.621	0.9	0.8	0.621	0.279	0.179
2	0.503	0.6	0.5	0.503	0.097	0.003
3	0.203	0.1	0.0	0.203	0.103	0.203
4	0.477	0.4	0.3	0.477	0.077	0.177
5	0.710	1.0	0.9	0.710	0.290	0.190
6	0.581	0.8	0.7	0.581	0.219	0.119
7	0.329	0.2	0.1	0.329	0.129	0.229
8	0.480	0.5	0.4	0.480	0.020	0.080
9	0.554	0.7	0.6	0.554	0.146	0.046
10	0.382	0.3	0.2	0.382	0.082	0.182

The SAS System

Variable	Maximum
DIFF1	0.2900000
DIFF2	0.2290000

Προφανώς, T=0.29.

4.2.1 Ζώνη Εμπιστοσύνης για την Συνάρτηση Κατανομής του Πληθυσμού

Ενα από τα πιο σημαντικά χαρακτηριστικά της ελεγχοσυνάρτησης του αμφίπλευρου ελέγχου Kolmogorov είναι ότι το (1–α)-ποσοστιαίο σημείο της κατανομής της μπορεί να χρησιμοποιηθεί

για να σχηματισθεί μια *ζώνη εμπιστοσύνης (confidence band)* για την πραγματική άγνωστη συνάρτηση κατανομής.

Ας θυμηθούμε ότι, για τον προσδιορισμό ενός διαστήματος εμπιστοσύνης για κάποια άγνωστη παράμετρο, χρησιμοποιούμε ένα τυχαίο δείγμα και με βάση αυτό το δείγμα υπολογίζουμε ένα άνω όριο και ένα κάτω όριο, τα οποία, με ένα δεδομένο συντελεστή εμπιστοσύνης, 1–α, παρέχουν ένα φάσμα εύλογων τιμών για την άγνωστη παράμετρο το οποίο ορίζεται από το μεταξύ τους διάστημα. Ακολουθώντας την ίδια λογική, θα μπορούσαμε με βάση ένα τυχαίο δείγμα από κάποιο πληθυσμό, του οποίου η συνάρτηση κατανομής είναι τελείως άγνωστη, να κατασκευάσουμε μία ζώνη εμπιστοσύνης, για την οποία ο βαθμός εμπιστοσύνης μας ότι περιέχει την άγνωστη συνάρτηση κατανομής είναι 1–α. Με απλά λόγια, δοθείσης της συνάρτησης S(x) για τα συγκεκριμένα δεδομένα, αναζητούμε μία περιοχή γύρω από την S(x) για την οποία, με την συνήθη ορολογία των διαστημάτων εμπιστοσύνης, ο συντελεστής εμπιστοσύνης μας ότι θα περιέχει την πραγματική αλλά άγνωστη συνάρτηση κατανομής F(x) εξ ολοκλήρου μετά στα όριά της είναι 1–α. Ο προσδιορισμός των ορίων γίνεται συνήθως γραφικά σε κατάλληλες κατακόρυφες αποστάσεις υπεράνω και κάτω από το γράφημα της S(x) με τον μόνο περιορισμό ότι τα όρια δεν επιτρέπουν στο γράφημα της F(x) να υπερβαίνει την έκταση από το 0 μέχρι το 1.

Εστω το τυχαίο δείγμα $X_1, X_2, ..., X_n$ μεγέθους n από κάποια άγνωστη κατανομή με συνάρτηση κατανομής F(x). Για να κατασκευάσουμε μία ζώνη εμπιστοσύνης για την άγνωστη συνάρτηση κατανομής F(x) πρέπει πρώτα να κατασκευάσουμε το γράφημα της εμπειρικής συνάρτησης κατανομής S(x), η οποία αντιστοιχεί στο δοθέν τυχαίο δείγμα. Στην συνέχεια, απαιτείται να προσδιορισθεί το

(1–α)-ποσοστιαίο σημείο της στατιστικής συνάρτησης Kolmogorov από τον πίνακα 14 του παραρτήματος για τον αμφίπλευρο έλεγχο (εάν μία αμφίπλευρη ζώνη εμπιστοσύνης είναι επιθυμητή) και για το κατάλληλο μέγεθος δείγματος n. Έστω ότι $w_{1-\alpha}$ είναι αυτό το ποσοστιαίο σημείο. Κατασκευάζουμε ένα γράφημα υπεράνω της S(x) σε απόσταση $w_{1-\alpha}$ και ονομάζουμε αυτό γράφημα U(x). Στην συνέχεια, κατασκευάζουμε ένα δεύτερο γράφημα κάτω από την S(x) σε απόσταση $w_{1-\alpha}$ και ονομάζουμε αυτό το δεύτερο γράφημα L(x). Τότε, τα γραφήματα των U(x) και L(x) αποτελούν το άνω και κάτω σύνορο, αντίστοιχα, μιας ζώνης εμπιστοσύνης για την οποία ο συντελεστής εμπιστοσύνης μας ότι περιέχει την άγνωστη συνάρτηση κατανομής F(x) εξ ολοκλήρου μεταξύ των συνόρων της είναι 1–α.

Είναι προφανές ότι δεν υπάρχει λόγος να παρασταθεί γραφικά το τμήμα της U(x) που αντιστοιχεί σε τιμές μεγαλύτερες του 1 ακόμα και στις περιπτώσεις που η $S(x) + w_{1-\alpha}$ υπερβαίνει την τιμή 1, αφού είναι γνωστό ότι μία συνάρτηση κατανομής δεν υπερβαίνει ποτέ τη μονάδα. Για τον ίδιο λόγο, η συνάρτηση L(x) δεν θα πρέπει να εκτείνεται κάτω από τον οριζόντιο άξονα. Τα σύνορα της ζώνης εμπιστοσύνης, επομένως, ορίζονται ως εξής:

$$L(x) = \begin{cases} S(x) - w_{1-\alpha}, & \text{αν } S(x) - w_{1-\alpha} \geq 0 \\ 0, & \text{αν } S(x) - w_{1-\alpha} < 0, \end{cases}$$

$$U(x) = \begin{cases} S(x) + w_{1-\alpha}, & \text{αν } S(x) + w_{1-\alpha} \leq 1 \\ 1, & \text{αν } S(x) + w_{1-\alpha} > 1. \end{cases}$$

Τότε, η ζώνη εμπιστοσύνης που ορίζεται από τα σύνορα L(x) και U(x) περιέχει εξ ολοκλήρου την άγνωστη συνάρτηση κατανομής F(x) για όλες τις τιμές του x με συντελεστή εμπιστοσύνης τουλάχιστον ίσο με 1–α. Συμβολικά, προσδίδοντας μία ευρύτερη ερμηνεία στον όρο πιθανότητα,

$$P(L(x) \le F(x) \le U(x), \text{ για κάθε } x) \ge 1-\alpha .$$

Η ισότητα ισχύει μόνο στην περίπτωση που η κατανομή είναι συνεχής. Δηλαδή για να είναι ο συντελεστής εμπιστοσύνης ακριβής, οι τυχαίες μεταβλητές είναι διακριτές, τότε η ζώνη εμπιστοσύνης είναι συντηρητική, δηλαδή, ο πραγματικός, αλλά άγνωστος συντελεστής εμπιστοσύνης είναι μεγαλύτερος από αυτόν που προσδιορίζουμε.

Παράδειγμα 4.2.3: Ας υποθέσουμε ότι θέλουμε να κατασκευάσουμε μία 90% ζώνη εμπιστοσύνης για μια άγνωστη συνάρτηση κατανομής F(x). Για τον σκοπό αυτό, έστω ότι έχει επιλεγεί από τον συγκεκριμένο πληθυσμό ένα δείγμα μεγέθους 20. Τα αποτελέσματα, διατεταγμένα κατά αύξουσα σειρά μεγέθους, δίνονται στον πίνακα που ακολουθεί:

16.7	17.4	18.1	18.2	18.8	19.3	22.4	22.4	24.0	24.7
25.9	27.0	25.1	35.8	36.5	37.6	39.8	42.1	43.2	46.2

Λύση: Το 0.90-ποσοστιαίο σημείο της ασυμπτωτικής κατανομής της στατιστικής συνάρτησης T προκύπτει από τον πίνακα 14 του παραρτήματος ότι είναι ίσο με 0.265, όταν το μέγεθος του δείγματος είναι n=20. Επομένως, η ζητούμενη ζώνη εμπιστοσύνης ορίζεται μεταξύ των συνόρων S(x)±0.265, με τον περιορισμό ότι η ζώνη είναι μεταξύ 0 και 1. Το σχήμα 4.2.6 απεικονίζει τις συναρτήσεις S(x), U(x) και L(x). Η ερμηνεία του γραφήματος αυτού είναι η εξής: Το συμπέρασμα *"η* F(x) *περιέχεται εξ ολοκλήρου μεταξύ των* U(x) *και*

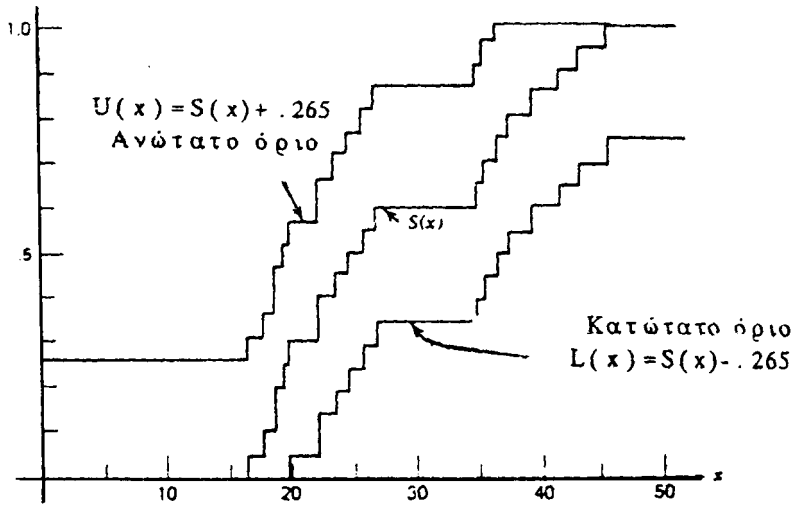
L(x)*”* ισχύει με συντελεστή εμπιστοσύνης τουλάχιστον ίσο με 0.90. Συμβολικά,

$$P(L(x) \le F(x) \le U(x) \ge 0.90,$$

όπου

$$L(x) = \begin{cases} S(x) - 0.265, & \text{αν } S(x) - 0.265 \ge 0 \\ 0, & \text{διαφορετικά} \end{cases}$$

$$U(x) = \begin{cases} S(x) + 0.265, & \text{αν } S(x) + 0.265 \le 1 \\ 1, & \text{διαφορετικά.} \end{cases}$$



Σχήμα 4.2.6

Ζώνη εμπιστοσύνης για την συνάρτηση κατανομής F(x)

Λύση με το MINITAB: Καταχωρίζουμε το δείγμα σε μία μεταβλητή με όνομα **x**. Για την κατασκευή της ζώνης εμπιστοσύνης, χρειάζεται να υπολογισθούν οι τιμές της εμπειρικής συνάρτησης κατανομής S(.) στα σημεία που αντιστοιχούν στις παρατηρήσεις του δείγματος. Για τον σκοπό αυτό, καταχωρίζουμε το δείγμα σε μία στήλη, διατάσσουμε τις

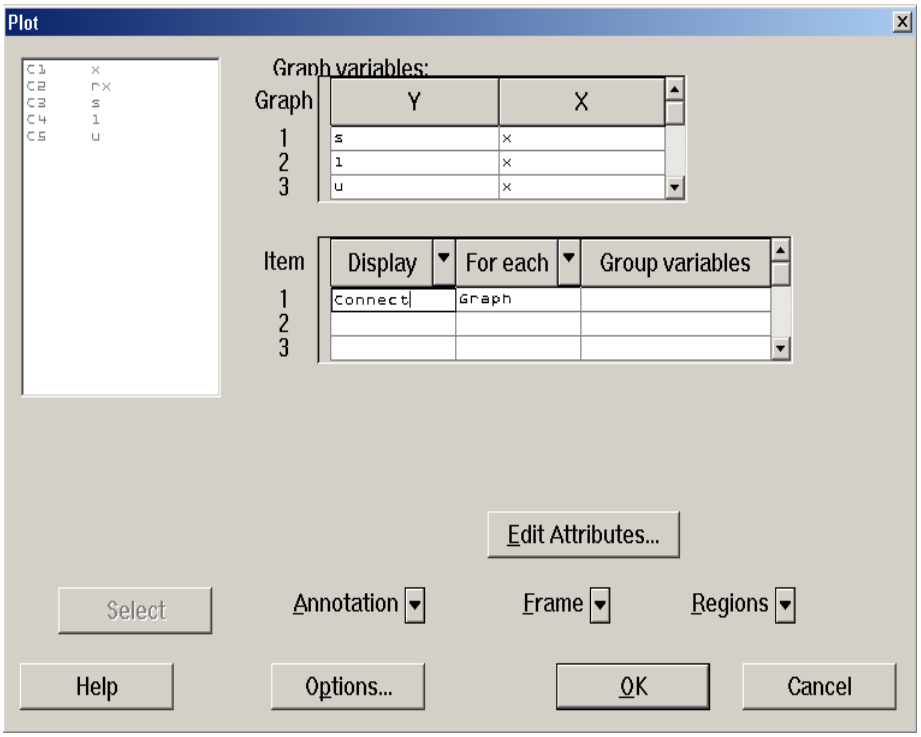
τιμές του κατ' αύξουσα τάξη μεγέθους και, σε μία άλλη στήλη, καταχωρίζουμε τις τάξεις μεγέθους τους. Επειδή υπάρχουν ισοβαθμίες, δεν μπορούμε να χρησιμοποιήσουμε την εντολή **Manip, Rank** αφού, για τον υπολογισμό των τιμών της S(.) οι τάξεις μεγέθους των ισοβαθμουσών τιμών πρέπει να ορίζονται ίσες με την τάξη μεγέθους που θα είχε η μεγαλύτερη από αυτές αν δεν ισοβαθμούσαν. Υπολογίζουμε, επομένως, τις τάξεις μεγέθους «με το χέρι» και τις καταχωρίζουμε σε μία μεταβλητή (έστω **rx**). Στην συνέχεια, διαιρώντας τις τάξεις μεγέθους με το μέγεθος του δείγματος, παίρνουμε τις τιμές της εμπειρικής συνάρτησης κατανομής S(.) στα σημεία που αντιστοιχούν στις παρατηρήσεις του δείγματος. Το φύλλο δεδομένων δείχνει ως εξής:

x	rx	s
25.9	12	0.60
27.0	13	0.65
35.8	14	0.70
36.5	15	0.75
37.6	16	0.80
39.8	17	0.85
42.1	18	0.90
43.2	19	0.95
46.2	20	1.00

(Προσέξτε την τάξη μεγέθους της τιμής 41). Στην συνέχεια, δημιουργούμε δύο ακόμη στήλες, **l** και **u**, στις οποίες καταχωρίζουμε το κάτω και το άνω όριο της 95% ζώνης εμπιστοσύνης της συνάρτησης κατανομής του πληθυσμού, από τον οποίο προέρχονται τα δεδομένα. Για τον σκοπό αυτό, χρησιμοποιούμε το 0.95 ποσοστημόριο της κατανομής της ελεγχοσυνάρτησης T του ελέγχου Kolmogorov-Smirnov για ένα δείγμα μεγέθους 20. Η τιμή του ποσοστημόριου είναι

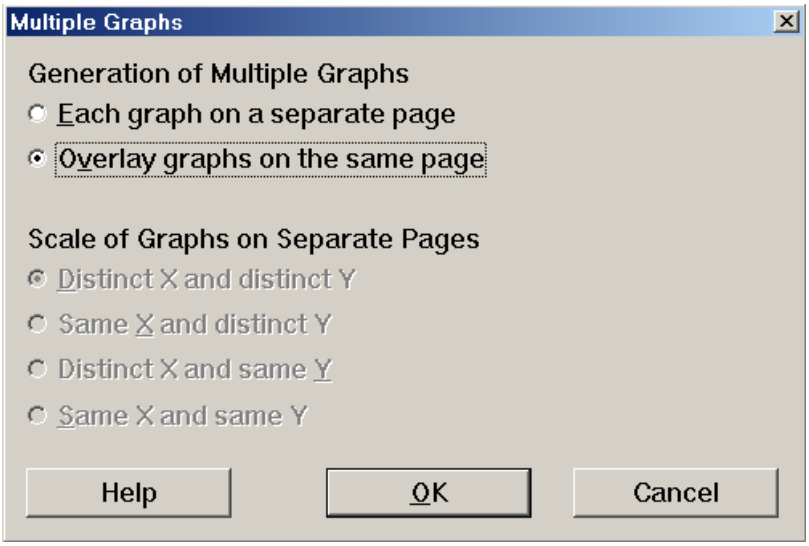
0.265. Η δημιουργία των στηλών αυτών γίνεται μέσω της επιλογής **Calc**, **Calculator** και πληκτρολογώντας **rmax(0,s-0.265)** και **rmin(1,s+0.265)** στο πεδίο **Numeric Expression**.

Απομένει να δούμε το γράφημα της ζώνης εμπιστοσύνης. Χρειάζεται, επομένως, να κατασκευάσουμε ένα γράφημα των μεταβλητών **s**, **l**, και **u** ως προς την μεταβλητή **x**. Για τον λόγο αυτό, επιλέγουμε **Graph**, **Plot** και οδηγούμεθα στο παρακάτω πλαίσιο διαλόγου:

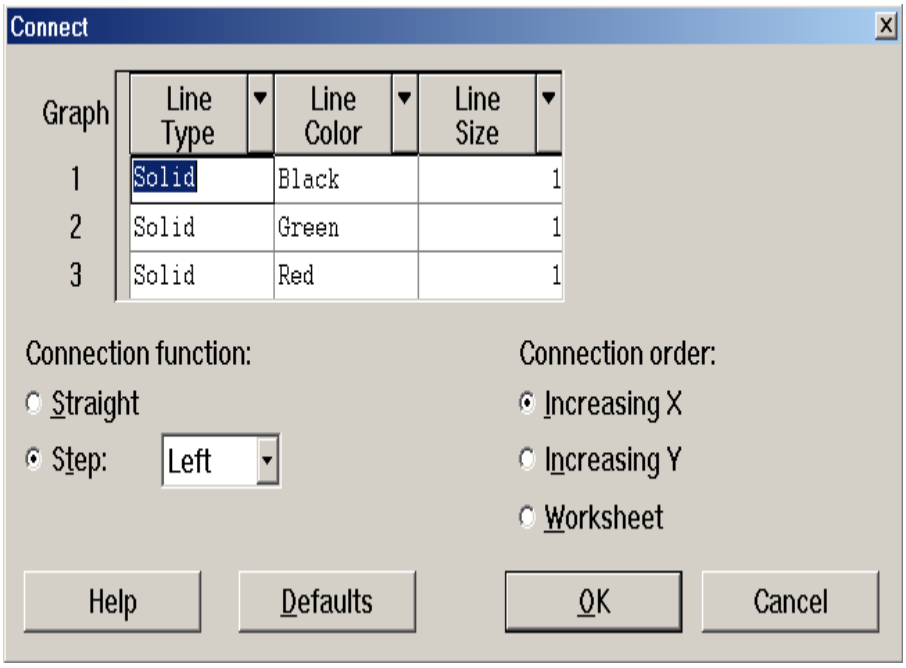


Στο πεδίο **Graph variables**, έχουμε την δυνατότητα να ορίσουμε περισσότερα του ενός διαγράμματα δηλώνοντας τα ονόματα των μεταβλητών που αντιστοιχούν στους άξονες X και Y του καθενός. Για τις ανάγκες του παραδείγματός μας, δηλώνουμε τα ζεύγη (**s x**, **l x** και **u**, **x**). Το MINITAB θα κατασκευάσει ένα ξεχωριστό γράφημα για κάθε

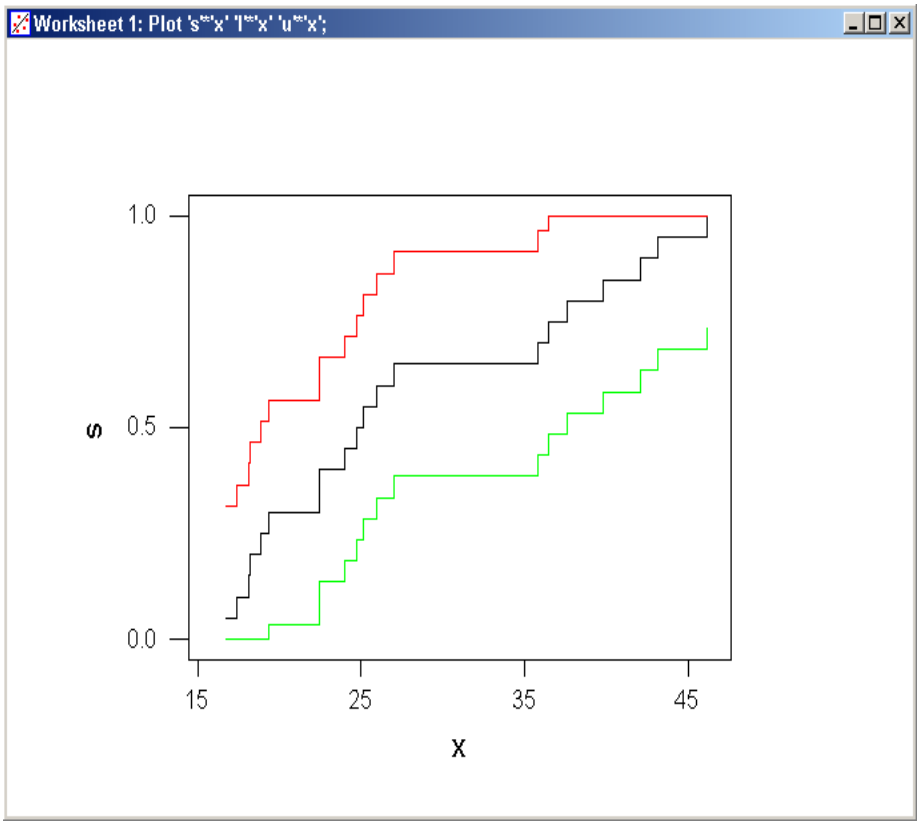
ζεύγος μεταβλητών. Για να έχουμε και τα τρία γραφήματα που θέλουμε σε ένα κοινό σύστημα αξόνων, επιλέγουμε **Frame, Multiple Graphs** και οδηγούμεθα στο πλαίσιο που ακολουθεί:



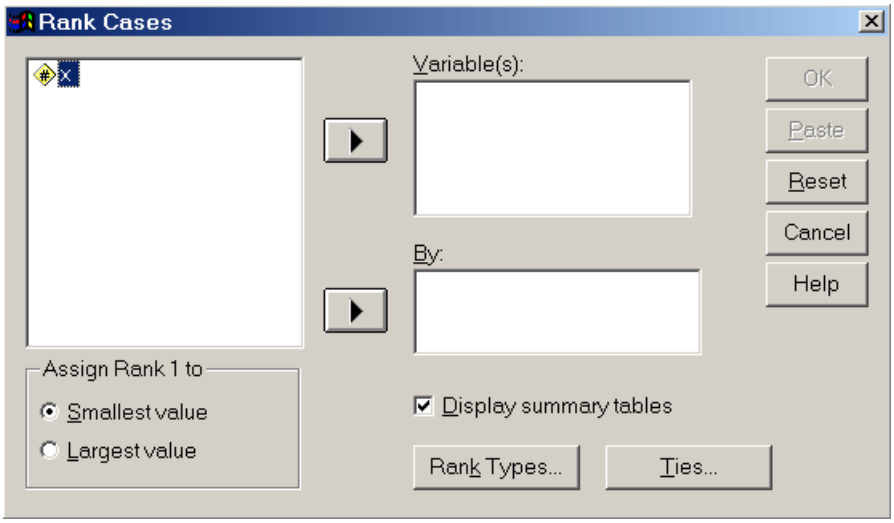
Επιλέγουμε **Overlay graphs on the same page**, και πιέζουμε **OK**. Στο πεδίο **Item** του επανεμφανιζόμενου προηγηθέντος πλαισίου (**Plot**), πιέζουμε το βέλος κοντά στην ένδειξη **Display** και επιλέγουμε **Connect**, ώστε τα σημεία στα γραφήματα να ενωθούν με ευθείες γραμμές. Τέλος, πιέζοντας το πλήκτρο **Edit Attributes**, προκύπτει το εξής πλαίσιο διαλόγου:



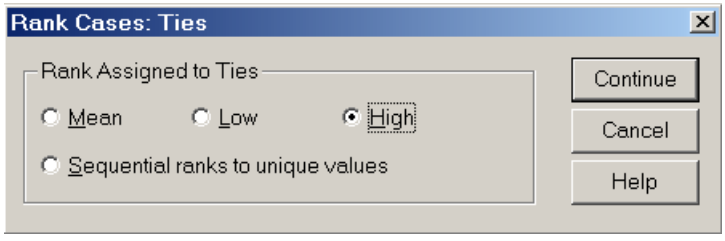
Μέσω αυτού του πλαισίου, μπορούμε να καθορίσουμε με τι είδους γραμμές θα ενώνονται τα σημεία. Συγκεκριμένα, επιλέγοντας τα σχετικά κελλιά του πεδίου **Graph** και πιέζοντας το βέλος δίπλα από τις ενδείξεις **Line Type**, **Line Color** και **Line Size**, καθορίζουμε τον τύπο της γραμμής που θα χρησιμοποιηθεί γαι το γράφημα (συμπαγής ή διακεκκομένη), καθώς επίσης το χρώμα και το μέγεθός της για κάθε ζεύγος μεταβλητών. Στο εικονιζόμενο πλαίσιο, έχει επιλεγεί διαφορετικό χρώμα για κάθε ζεύγος. Στο πεδίο **Connection function**, επιλέγουμε **Step** και **Left** ώστε το κάθε διάγραμμα να έχει βαθμωτή μορφή και να παρουσιάζει άλμα σε κάθε τιμή που αντιστοιχεί σε παρατήρηση του δείγματος. Στο πεδίο **Connection order**, επιλέγουμε **Increasing X**. Πιέζοντας **OK** και ξανά **OK** στο επανεμφανιζόμενο πλαίσιο διαλόγου **Plot**, προκύπτει το γράφημα της εμπειρικής συνάρτησης κατανομής και της ζώνης εμπιστοσύνης της:



Λύση με το SPSS: Το SPSS δεν υπολογίζει την εμπειρική συνάρτηση κατανομής αυτομάτως. Ο υπολογισμός των τιμών της μπορεί να γίνει με έμμεσο τρόπο. Ως πρώτο βήμα, διατάσσουμε τις παρατηρήσεις του δείγματος κατ' αύξουσα τάξη μεγέθους με χρήση του **Data, Sort Cases** κατά τα ήδη γνωστά και καταχωρίζουμε το προκύπτον διατεταγμένο δείγμα στην μεταβλητή **x**. Στην συνέχεια, σε κάθε παρατήρηση του δείγματος, αντιστοιχίζουμε την τάξη μεγέθους της: Επιλέγοντας **Transform, Rank Cases,** προκύπτει το ακόλουθο πλαίσιο διαλόγου:



Στο πεδίο **Variable(s)** του πλαισίου αυτού, δηλώνουμε την μεταβλητή, τις τάξεις μεγέθους της οποίας θέλουμε να υπολογίσουμε (δηλαδή την μεταβλητή **x**). Στο πεδίο **Assign Rank 1 to,** δηλώνουμε αν η τάξη μεγέθους 1 θα υποδηλώνει την μικρότερη ή την μεγαλύτερη παρατήρηση. Για να έχουμε διάταξη κατ' αύξουσα σειρά, επιλέγουμε **Smallest value**. Στην συνέχεια, πιέζοντας το πλήκτρο **Ties**, εμφανίζεται το πλαίσιο που εικονίζεται παρακάτω, μέσω του οποίου μπορούμε να επιλέξουμε τον τρόπο με τον οποίο θα ορισθούν οι τάξεις των ισοβαθμουσών τιμών.

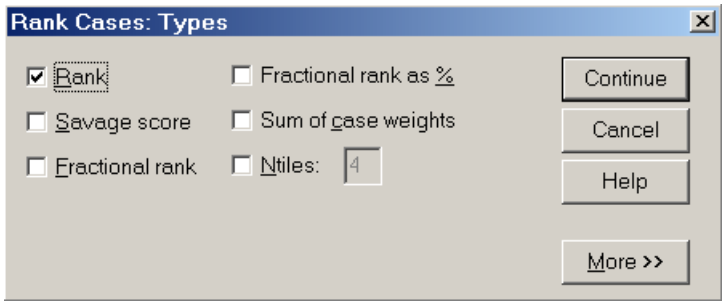


(Για παράδειγμα, όλες ίσες με την μέση τιμή των τάξεων μεγέθους που θα είχαν οι ισοβαθμούσες τιμές αν δεν ισοβαθμούσαν (επιλογή **Mean**), όλες ίσες με την μικρότερη από τις τάξεις μεγέθους που θα είχαν αν δεν ισοβαθμούσαν (επιλογή **Low**), όλες ίσες με την μεγαλύτερη από τις τάξεις μεγέθους που θα είχαν. (επιλογή **High**) ή όλες ίσες με την τάξη που θα είχαν αν ήταν μοναδικές τιμές (επιλογή **Sequential ranks to unique values**)). Τα παραπάνω γίνονται ευκολότερα αντιληπτά με το παράδειγμα που δίνει η βοήθεια (**Help**) του SPSS:

Value	Mean	Low	High	Sequential
10	1	1	1	1
15	3	2	4	2
15	3	2	4	2
15	3	2	4	2
16	5	5	5	3
20	6	6	6	4

Δεδομένου ότι ο υπολογισμός των τιμών της εμπειρικής συνάρτησης κατανομής S(.) συνίσταται στην διαίρεση κάθε τάξης μεγέθους με το μέγεθος του δείγματος, επιλέγουμε **High**. (Ετσι, η τάξη μεγέθους κάθε παρατήρησης δείχνει πόσες παρατηρήσεις είναι μικρότερες ή ίσες αυτής).

Πιέζοντας **Continue** και **Rank Types** στο επανεμφανιζόμενο αρχικό πλαίσιο διαλόγου, οδηγούμεθα στο εξής πλαίσιο:

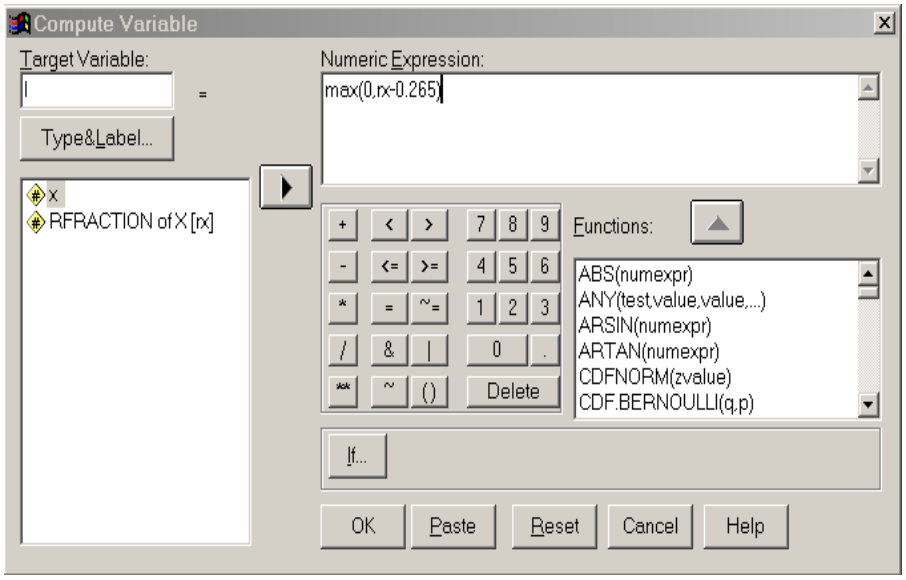


Με την επιλογή **Rank,** υπολογίζονται οι τάξεις μεγέθους των παρατηρήσεων του δείγματος. Με την επιλογή **Fractional rank,** όμως, υπολογίζονται οι τάξεις μεγέθους, οι οποίες στην συνέχεια διαιρούνται με το μέγεθος του δείγματος. Επομένως, επιλέγοντας **Fractional rank**, μπορεί να προκύψει η εμπειρική συνάρτηση κατανομής S(.).

Με την ολοκλήρωση της διαδικασίας, δημιουργείται η μεταβλητή **rx** (το SPSS της δίνει αυτόματα όνομα ίδιο με αυτό της αρχικής μεταβλητής (**x**) με την προσθήκη του γράμματος r). Στην συνέχεια, δημιουργούμε δύο άλλες μεταβλητές (έστω **u** και **l**) στις οποίες θα αποθηκευθούν οι τιμές του άνω και κάτω ορίου της 95% ζώνης εμπιστοσύνης αντίστοιχα. Στα επόμενα, δίνεται αναλυτική περιγραφή του τρόπου υπολογισμού των τιμών του κάτω ορίου όπως επίσης και συνοπτική περιγραφή του τρόπου υπολογισμού των τιμών του άνω ορίου.

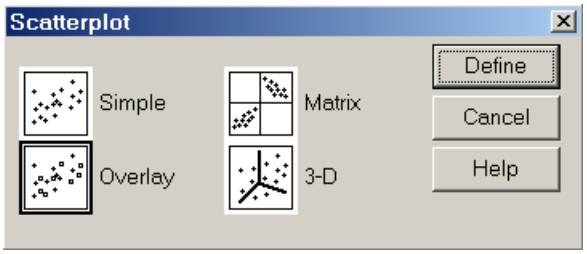
Σε κάθε δυνατή τιμή x όπου η εμπειρική συνάρτηση κατανομής παίρνει την τιμή S(x), το κάτω όριο έχει τιμή $L(x)=\max(0,S(x)-w_{0.95})$, ενώ το άνω όριο έχει τιμή $U(x)=\min(1,S(x)+w_{0.95})$. Προφανώς, ο υπολογισμός των τιμών αυτών είναι αναγκαίος μόνο σε σημεία που αντιστοιχούν σε παρατηρήσεις του δείγματος. Από τους πίνακες της κατανομής της ελεγχοσυνάρτησης T, προκύπτει ότι, για μέγεθος δείγματος n=20, η τιμή του $w_{0.95}$ είναι 0.265. Για να δημιουργήσουμε

την **l**, επιλέγουμε **Transform**, **Compute** και πληκτρολογούμε τα επανεμφανιζόμενα στο εικονιζόμενο πλαίσιο διαλόγου που ακολουθεί:

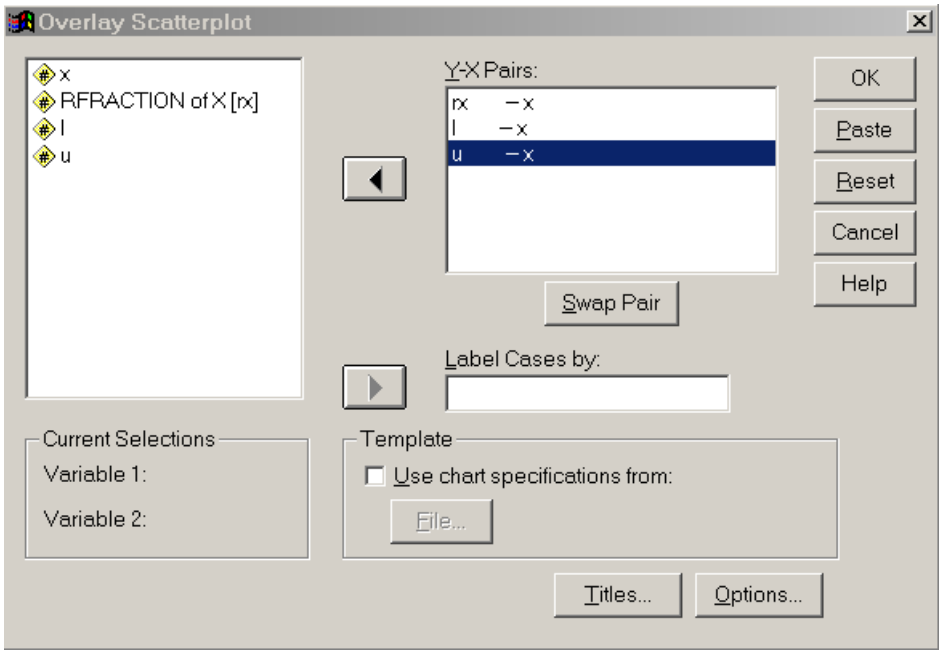


Η αντίστοιχη έκφραση (**Numeric Expression**) για την δημιουργία της **u** είναι **min(1,rx+0.265)**. Απομένει η κατασκευή του γραφήματος της εμπειρικής συνάρτησης κατανομής S(.) και των ορίων της ζώνης εμπιστοσύνης.

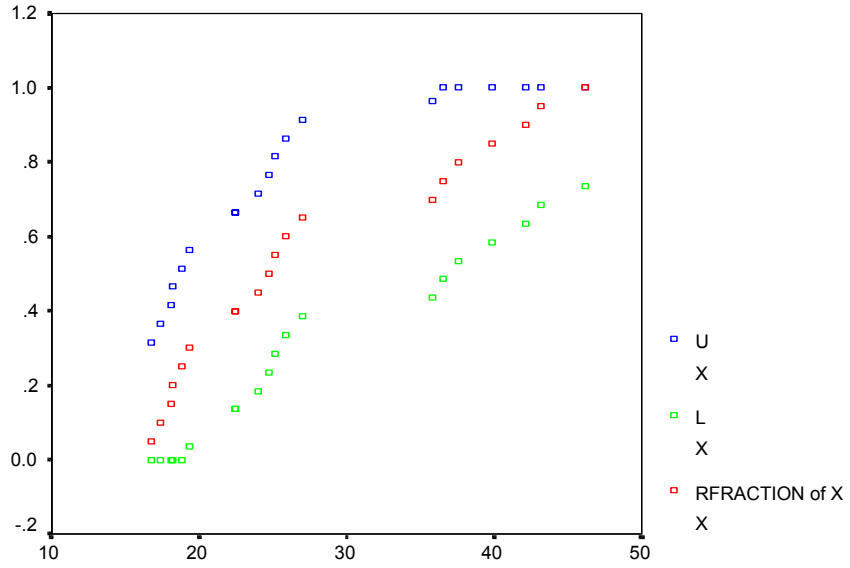
Ως πρώτο βήμα, επιλέγουμε **Graphs**, **Scatter** και οδηγούμεθα στο εξής πλαίσιο διαλόγου:



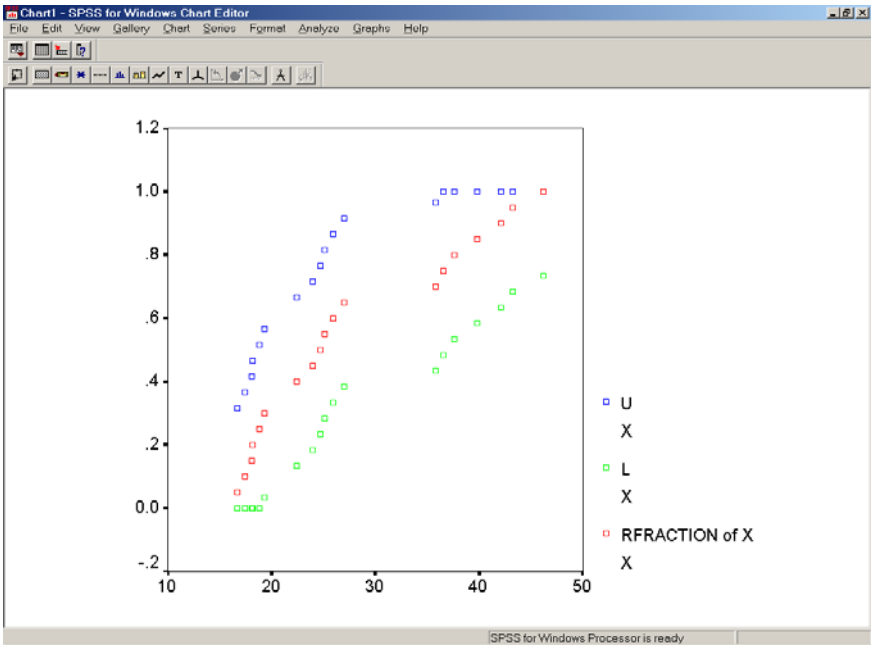
Επιλέγοντας **overlay** (επειδή, στην ουσία, έχουμε τρία χωριστά γραφήματα) και, στην συνέχεια, πιέζοντας **Define,** προκύπτει το ακόλουθο πλαίσιο:



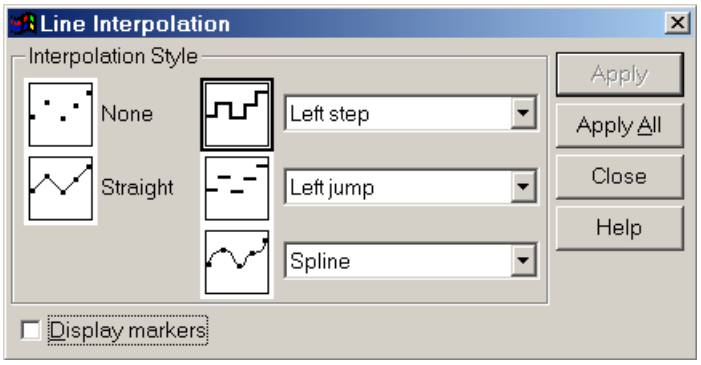
Στο πεδίο **Y-X Pairs,** δηλώνουμε όλα τα ζεύγη μεταβλητών που θέλουμε να παραστήσουμε γραφικά, δηλαδή, κάθε μία από τις **rx**, **l** και **u** με την **x**). Ενώ επιλέγουμε ένα ζεύγος, πρέπει να έχουμε πατημένο το πλήκτρο **Ctrl**. Αν τα μέλη κάποιου ζεύγους εμφανίζονται με σειρά αντίστροφη από αυτή που θέλουμε (δηλαδή με την **x** στον Y άξονα), πιέζουμε το πλήκτρο **Swap Pair**. Πιέζοντας **OK**, προκύπτει το ακόλουθο γράφημα:



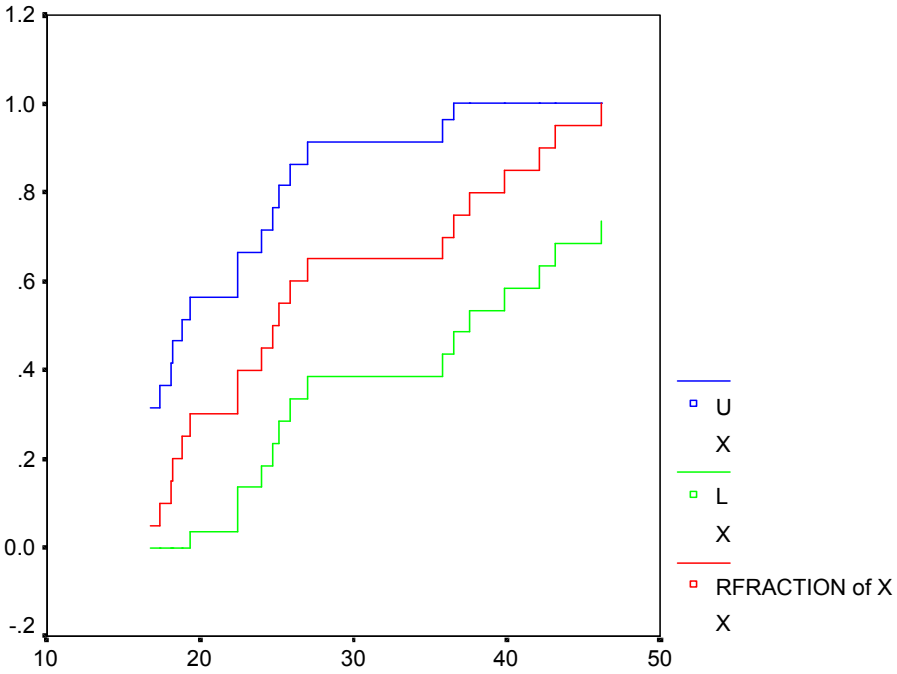
Το γράφημα αυτό δείχνει μόνο τα σημεία στα οποία οι συναρτήσεις L(.), S(.) και U(.) παρουσιάζουν πήδημα. Για να το φέρουμε στην κατάλληλη μορφή, κάνουμε διπλό κλικ πάνω στο γράφημα και μεταφερόμεθα στον **Chart editor**:



Ο **chart editor** είναι προσβάσιμος από οποιοδήποτε γράφημα του SPSS και μας επιτρέπει επεμβάσεις πάνω στα γραφήματα που δεν είναι εφικτές από το συνηθισμένο σύστημα μενού του προγράμματος. Για να αλλάξουμε τη μορφή των γραμμών, επιλέγουμε κατά σειρά **Format**, **Interpolation** και οδηγούμεθα στο εξής πλαίσιο διαλόγου:



Στο πεδίο **Interpolation Style,** επιλέγουμε **Left step** για να αποκτήσει το γράφημα την γνωστή κλιμακωτή μορφή που έχουν τα γραφήματα των συναρτήσεων κατανομής. Ακυρώνουμε την επιλογή **Display markers** για να μη φαίνονται τα τετραγωνίδια που παριστάνουν τα σημεία του γραφήματος διασποράς (διαγράμματος σημείων) και πιέζουμε, **Apply All** και, στην συνέχεια, **Close**. Βγαίνοντας από τον **chart editor,** οδηγούμεθα στο ζητούμενο γράφημα των ζωνών εμπιστοσύνης που τώρα έχει την κατάλληλη μορφή:



Λύση με το SAS: Το SAS δεν υπολογίζει αυτόματα την εμπειρική συνάρτηση κατανομής. Για τον υπολογισμό της, καταχωρίζουμε τα δεδομένα σε ένα αρχείο του SAS και θα κατασκευάσουμε τις τάξεις μεγέθους.

```
Data empiric;
input x @@;
cards;
16.7 17.4 18.1 18.2 18.8 19.3 22.4 22.4 24.0 24.7
25.9 27.0 25.1 35.8 36.5 37.6 39.8 42.1 43.2 46.2
;
run;
proc rank ties=high out=rxfile fraction;
var x;
ranks rx;
run;
proc print;
run;
```

Η διαδικασία **proc rank** δίδει στα δεδομένα τις τάξεις μεγέθους ενώ η υποεντολή **ties=high** δίδει στις ισοβαθμούσες τιμές την μεγαλύτερη

από τις τάξεις μεγέθους που θα είχαν αν δεν ισοβαθμούσαν. Επίσης, η υποεντολή **fraction** διαιρεί τις τάξεις μεγέθους με το μέγεθος του δείγματος. Οι τάξεις μεγέθους που προκύπτουν αποθηκεύονται στην μεταβλητή με όνομα **rx** που έχει αποθηκευτεί στο αρχείο **rxfile**. Το αποτέλεσμα των εντολών αυτών φαίνεται στον πίνακα που ακολουθεί.

OBS	X	RX
1	16.7	0.05
2	17.4	0.10
3	18.1	0.15
4	18.2	0.20
5	18.8	0.25
6	19.3	0.30
7	22.4	0.40
8	22.4	0.40
9	24.0	0.45
10	24.7	0.50
11	25.9	0.60
12	27.0	0.65
13	25.1	0.55
14	35.8	0.70
15	36.5	0.75
16	37.6	0.80
17	39.8	0.85
18	42.1	0.90
19	43.2	0.95
20	46.2	1.00

Στην συνέχεια, κατασκευάζουμε την 90% ζώνη εμπιστοσύνης με τις εντολές:

```
data rxfile;;
set rxfile;
lo=rx-0.265;
hi=rx+0.265;
l=max(0,lo);
h=min(1,hi);
run;
proc sort; by rx;

data all; merge empiric rxfile;
proc print;
run;
```

και, με την ενοποίηση των αρχείων που έχουν δημιουργηθεί (εντολή **merge**), έτσι ώστε να πάρουμε σε ένα κοινό αρχείο όλες τις μεταβλητές που έχουμε κατασκευάσει, προκύπτει το εξής αρχείο:

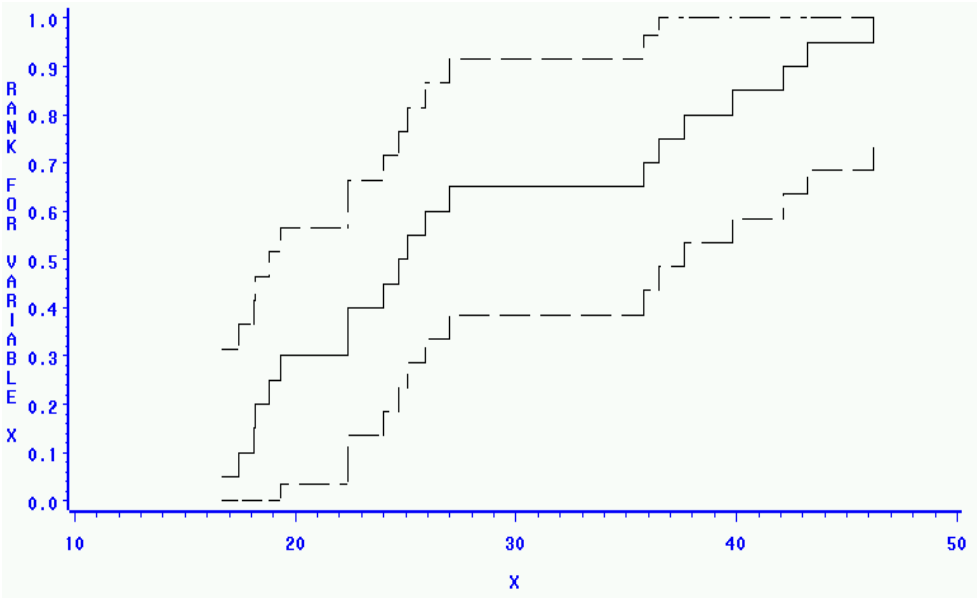
OBS	X	RX	LO	HI	L	H
1	16.7	0.05	-0.215	0.315	0.000	0.315
2	17.4	0.10	-0.165	0.365	0.000	0.365
3	18.1	0.15	-0.115	0.415	0.000	0.415
4	18.2	0.20	-0.065	0.465	0.000	0.465
5	18.8	0.25	-0.015	0.515	0.000	0.515
6	19.3	0.30	0.035	0.565	0.035	0.565
7	22.4	0.40	0.135	0.665	0.135	0.665
8	22.4	0.40	0.135	0.665	0.135	0.665
9	24.0	0.45	0.185	0.715	0.185	0.715
10	24.7	0.50	0.235	0.765	0.235	0.765
11	25.1	0.55	0.285	0.815	0.285	0.815
12	25.9	0.60	0.335	0.865	0.335	0.865
13	27.0	0.65	0.385	0.915	0.385	0.915
14	35.8	0.70	0.435	0.965	0.435	0.965
15	36.5	0.75	0.485	1.015	0.485	1.000
16	37.6	0.80	0.535	1.065	0.535	1.000
17	39.8	0.85	0.585	1.115	0.585	1.000
18	42.1	0.90	0.635	1.165	0.635	1.000
19	43.2	0.95	0.685	1.215	0.685	1.000
20	46.2	1.00	0.735	1.265	0.735	1.000

Η ζώνη εμπιστοσύνης που κατασκευάσθηκε μπορεί να παρουσιασθεί και γραφικά, χρησιμοποιώντας τις εντολές που ακολουθούν:

```
goptions reset=(axis, legend, pattern, symbol, title, footnote) norotate
         hpos=0 vpos=0 htext= ftext= ctext= target= gaccess= gsfmode= ;
goptions device=WIN  ctext=blue
         graphrc interpol=join;
symbol1 c=DEFAULT
    i=STEPSLJ
    v=NONE
    cv=RED
    ;
symbol2 c=DEFAULT
    i=STEPSLJ
    l=2

    v=NONE
```

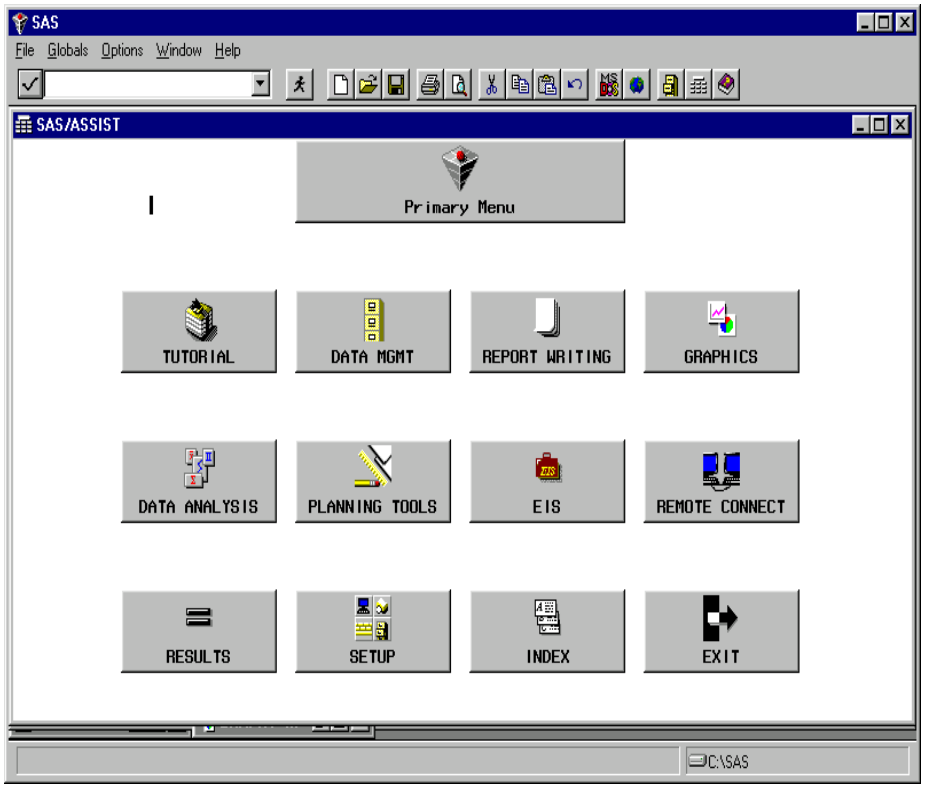
```
    ;
symbol3 c=DEFAULT
    i=STEPSLJ
    l=2
    v=NONE
    ;

axis1
   color=blue
   width=2.0
   ;
axis2
   color=blue
   width=2.0
   ;
axis3
   color=blue
   width=2.0
   ;
proc gplot data=WORK.ALL;
   plot (RX L H ) * X / overlay
      haxis=axis1
      vaxis=axis2
      ;
run;
 quit;
```

Οι εντολές αυτές οδηγούν στο ζητούμενο γράφημα, όπου η F(.) εμφανίζεται να περιέχεται εξ ολοκλήρου στην 90% ζώνη εμπιστοσύνης. (Στο γράφημα, τα όρια της ζώνης αυτής σημειώνονται με διακεκομμένες γραμμές).



Είναι προφανές, ότι το πρόγραμμα, που πρέπει να γραφεί και οι εντολές που πρέπει να πληκτρολογηθούν προκειμένου να κατασκευασθεί το ζητούμενο γράφημα, απαιτούν πολύ χρόνο και κόπο. Για τον λόγο αυτό, το SAS παρέχει ένα βοηθητικό πρόγραμμα που λειτουργεί με μενού και έτσι η κατασκευή των γραφημάτων είναι δυνατή χωρίς την χρήση εντολών:

Με βάση το βοηθητικό αυτό πρόγραμμα, αφού έχουμε εκτελέσει τις εντολές που δημιούργησαν τα αρχεία με τις απαραίτητες μεταβλητές, επιλέγουμε από το κεντρικό μενού **Globals** και **SAS/ASSIST**. Θα εμφανισθεί το παρακάτω μενού:



Στη συνέχεια, επιλέγουμε **GRAPHICS** και **high resolution** (για γραφήματα υψηλής ευκρίνειας). Από το μενού που εμφανίζεται, επιλέγουμε **Plots** και στην συνέχεια, **Multiple plots per axes**. Δίδοντας τις κατάλληλες μεταβλητές στα πεδία που εμφανίζονται, επιλέγουμε **Locals** και **Run** για να πάρουμε το ζητούμενο γράφημα.

4.3 ΕΛΕΓΧΟΙ ΚΑΛΗΣ ΠΡΟΣΑΡΜΟΓΗΣ ΓΙΑ ΟΙΚΟΓΕΝΕΙΕΣ ΚΑΤΑΝΟΜΩΝ

Ο έλεγχος καλής προσαρμογής του Kolmogorov είναι ένας "καλός" έλεγχος για τον έλεγχο της υπόθεσης ότι ένα τυχαίο δείγμα προέρχεται από μία συγκεκριμένη κατανομή. Ο έλεγχος

Kolmogorov καλύπτει μόνο τις περιπτώσεις στις οποίες η υποτεθείσα συνάρτηση κατανομής είναι εξ ολοκλήρου ορισμένη, δηλαδή, όταν δεν υπάρχουν άγνωστες παράμετροι που πρέπει να εκτιμηθούν με βάση το δείγμα. Διαφορετικά, ο έλεγχος γίνεται συντηρητικός. Αντίθετα, ο έλεγχος καλής προσαρμογής χ^2 είναι ευέλικτος και επιτρέπει την εκτίμηση ορισμένων παραμέτρων με βάση τα δεδομένα (Ενας βαθμός ελευθερίας απλώς αφαιρείται για κάθε παράμετρο που εκτιμάται). Όμως, ο έλεγχος χ^2 απαιτεί την ομαδοποίηση των δεδομένων και μια τέτοια ομαδοποίηση είναι συχνά αυθαίρετη. Επιπλέον, η κατανομή της στατιστικής συνάρτησης είναι μόνο κατά προσέγγιση γνωστή και, μερικές φορές, η ισχύς του ελέγχου δεν είναι πολύ καλή. Για τους λόγους αυτούς, έχουν μελετηθεί άλλοι έλεγχοι καλής προσαρμογής, κυρίως για κατανομές που συχνά χρησιμοποιούνται σε σχέση με πρακτικές εφαρμογές.

Στην βιβλιογραφία, έχουν μελετηθεί αρκετές παραλλαγές του ελέγχου Kolmogorov, οι οποίες επιτρέπουν την χρήση του σε περιπτώσεις όπου παράμετροι εκτιμώνται από τα δεδομένα. Στην πραγματικότητα, η στατιστική συνάρτηση παραμένει μεν η ίδια, αλλά η κατανομή της είναι διαφορετική. Για τον προσδιορισμό, επομένως ποσοστιαίων σημείων και κρίσιμων τιμών, απαιτούνται διαφορετικοί πίνακες. Οι πίνακες αυτοί δεν είναι οι ίδιοι για όλες τις κατανομές, αλλά αλλάζουν ανάλογα με τη μορφή της μηδενικής κατανομής.

Μία τέτοια παραλλαγή του ελέγχου Kolmogorov είναι αυτή για τον έλεγχο της σύνθετης υπόθεσης της κανονικότητας, δηλαδή, της υπόθεσης ότι ο πληθυσμός ανήκει στην οικογένεια των κανονικών κατανομών, χωρίς να προσδιορίζεται η μέση τιμή ή η διασπορά της κανονικής κατανομής. Ο έλεγχος αυτός μελετήθηκε για πρώτη φορά

από τον Lilliefors το 1967. Για το λόγο αυτό ο έλεγχος αυτός είναι γνωστός ως *έλεγχος κανονικότητας του Lilliefors*.

4.3.1 Ο Ελεγχος Κανονικότητας του Lillefors

Εστω X_1, X_2, ..., X_n ένα δείγμα μεγέθους n από κάποιο πληθυσμό με άγνωστη συνάρτηση κατανομής F(x). Να ελεγχθεί η υπόθεση

H_0: Το τυχαίο δείγμα προέρχεται από την κανονική κατανομή με άγνωστη μέση τιμή μ και άγνωστη διασπορά σ^2.

έναντι της εναλλακτικής

H_1: Το τυχαίο δείγμα προέρχεται από μία μη κανονική κατανομή.

Οι υποθέσεις αυτές μπορούν να ελεγχθούν με την χρήση της συνήθους αμφίπλευρης ελεγχοσυνάρτησης του Kolmogorov, η οποία ορίζεται ως η μέγιστη κατακόρυφη απόσταση μεταξύ της εμπειρικής συνάρτησης κατανομής των X_i και της συνάρτησης κατανομής της κανονικής κατανομής με μέση τιμή ίση με τον μέσο του δείγματος και τυπική απόκλιση ίση με την αμερόληπτη εκτίμησή της μέσω του δείγματος. Με άλλα λόγια, ως στατιστική συνάρτηση ελέγχου, μπορεί να χρησιμοποιηθεί η συνάρτηση T του αμφίπλευρου ελέγχου Kolmogorov για τον έλεγχο της μηδενικής υπόθεσης ότι η άγνωστη κατανομή του πληθυσμού είναι η κανονική με μέση τιμή ίση με $\overline{x}$ και τυπική απόκλιση ίση με s^*, όπου $\overline{x}$ είναι η παρατηρηθείσα τιμή του μέσου του δείγματος, όπως αυτός ορίζεται από την σχέση

$$\overline{X} = \sum_{i=1}^{n} X_i / n$$

και s^* είναι η τιμή της συνάρτησης

$$S^* = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2},$$

που χρησιμοποιείται ως αμερόληπτη εκτιμήτρια της σ. Ισοδύναμα, μπορούμε να υπολογίσουμε τις *τυποποιημένες* τιμές $Z_1, Z_2, ..., Z_n$ του δείγματος $X_1, X_2, ..., X_n$ που ορίζονται από την σχέση

$$Z_i = \frac{X_i - \overline{X}}{S^*}, \quad i = 1, 2, ..., n.$$

Τότε, οι αρχικές υποθέσεις μας είναι ισοδύναμες με τις υποθέσεις

H_0^*: Το τυχαίο δείγμα $Z_1, Z_2, ..., Z_n$ προέρχεται από την τυποποιημένη κανονική κατανομή.

H_1^*: Το τυχαίο δείγμα $Z_1, Z_2, ..., Z_n$ δεν προέρχεται από την τυποποιημένη κανονική κατανομή.

Η κατάλληλη στατιστική συνάρτηση ελέγχου, στην περίπτωση αυτή, θα είναι η μέγιστη κατακόρυφη απόκλιση της εμπειρικής συνάρτησης κατανομής $S^*(z)$ του τυποποιημένου δείγματος από την συνάρτηση κατανομής $F_0^*(z)$ της τυποποιημένης κανονικής κατανομής. Δηλαδή, η ελεγχοσυνάρτηση του Lilliefors ορίζεται από τη σχέση

$$T_1 = \sup_z \left|F_0^*(z) - S^*(z)\right| .$$

Οπως και στην περίπτωση της ελεγχοσυνάρτησης T του Kolmogorov, η αναλυτική μορφή της συνάρτησης κατανομής της στατιστικής συνάρτησης T_1 του Lilliefors είναι δύσκολο να προσδιορισθεί. Έτσι, ο Lilliefors μελέτησε και πινακοποίησε την ασυμπτωτική κατανομή της στατιστικής συνάρτησης T_1. Ο πίνακας

15 του παραρτήματος περιέχει τα κρίσιμα σημεία της κατανομής αυτής.

Προφανώς και στην περίπτωση αυτού του ελέγχου, είναι οι μεγάλες τιμές της στατιστικής συνάρτησης T, οι οποίες συνηγορούν υπέρ της απόρριψης της μηδενικής υπόθεσης, αφού τέτοιες τιμές αντανακλούν χαμηλό βαθμό εγγύτητας της εμπειρικής συνάρτησης κατανομής του τυποποιημένου δείγματος προς την συνάρτηση κατανομής της τυποποιημένης κανονικής κατανομής.

Παράδειγμα 4.3.1: Από τις ταφόπετρες του νεκροταφείου του Badenscallie του Wester Ross της Σκωτίας, έγινε μία καταγραφή των ηλικιών θανάτου των αρρένων, οι οποίοι ανήκαν σε τέσσερις εξέχουσες Σκωτικές οικογένειες (clans) της περιοχής αυτής. Από το σύνολο των 117 ηλικιών θανάτου που καταγράφηκαν, επελέγη ένα τυχαίο δείγμα 30 ηλικιών. Οι ηλικίες του δείγματος αυτού, διατεταγμένες κατά αύξουσα σειρά μεγέθους, ήταν οι εξής:

11 13 14 22 29 30 41 41 52 55 56 59 65 65 66
74 74 75 77 81 82 82 82 82 83 85 85 87 87 88.

Είναι εύλογο να υποθέσουμε ότι οι ηλικίες θανάτου κατανέμονται κανονικά;

Λύση: Από τα δεδομένα, υπολογίζουμε την τιμή του δειγματικού μέσου και της συνήθους αμερόληπτης εκτιμήτριας της τυπικής απόκλισης του πληθυσμού. Συγκεκριμένα, βρίσκουμε ότι $\bar{x} = 61.43$ και $s^* = 25.04$. Τότε, τυποποιώντας τις τιμές του δείγματος, οδηγούμεθα στον πίνακα 4.3.1.

Πίνακας 4.3.1

Έλεγχος κανονικότητας Lilliefors για τις ηλικίες θανάτου στο Badenscallie

x	z	$F^*(z)$	$S^*(z)$	$F^*(z_i) - S^*(z_i)$	$F^*(z_i) - S^*(z_{i-1})$
11	–2.014	0.022	0.033	–0.011	0.022
13	–1.934	0.026	0.067	–0.044	–0.007
14	–1.894	0.029	0.100	–0.071	–0.038
22	–1.575	0.058	0.133	–0.075	–0.042
29	–1..295	0.098	0.167	–0.069	–0.035
30	–1.255	0.105	0.200	–0.095	–0.062
41(2)	–0.816	0.207	0.267	–0.060	0.007
52	–0.377	0.353	0.300	0.053	0.086
55	–0.257	0.399	0.333	0.066	0.099
56	–0.217	0.414	0.367	0.047	0.081
59	–0.097	0.461	0.400	0.061	0.094
65(2)	0.142	0.556	0.467	0.089	0.156
66	0.183	0.572	0.500	0.072	0.105
74(2)	0.502	0.692	0.567	0.125	0.192
75	0.542	0.706	0.600	0.106	0.139
77	0.622	0.733	0.633	0.100	0.133
81	0.781	0.782	0.667	0.115	0.149
82(4)	0.821	0.794	0.800	–0.006	0.127
83	0.861	0.805	0.833	–0.028	–0.005
85(2)	0.942	0.827	0.900	–0.073	–0.006
87(2)	1.021	0.846	0.967	–0.121	–0.054
88	1.061	0.856	1.000	–0.144	–0.111

Σε παρένθεση δεξιά από την τιμή των παρατηρήσεων που εμφανίζονται στο δείγμα περισσότερες από μία φορά, σημειώνεται η συχνότητά τους.

Η τρίτη στήλη του πίνακα αυτού περιέχει τις τιμές της συνάρτησης κατανομής της τυποποιημένης κανονικής κατανομής στα σημεία z, τα οποία αντιστοιχούν στις διάφορες ηλικίες θανάτου και περιέχονται στη δεύτερη στήλη. Η τέταρτη στήλη περιέχει τις

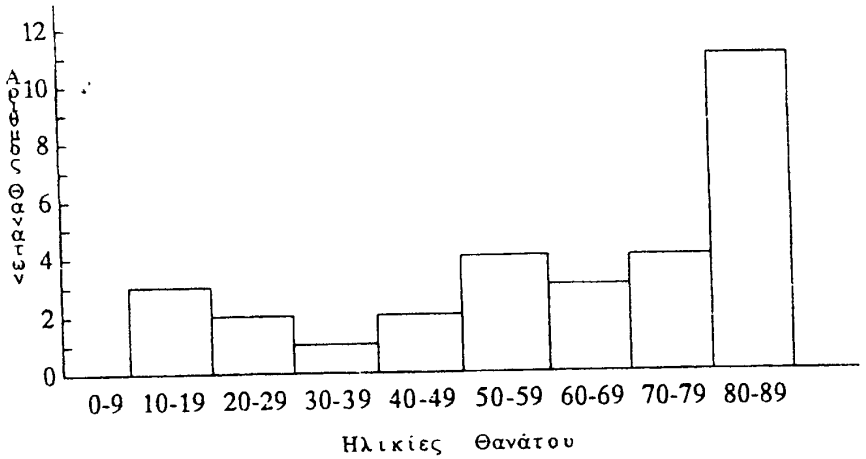
αντίστοιχες τιμές της εμπειρικής συνάρτησης κατανομής στα σημεία που αντιστοιχούν στις διάφορες ηλικίες θανάτου. Από την πέμπτη στήλη, είναι προφανές ότι η μέγιστη διαφορά έχει μέγεθος 0.192 και παρατηρείται όταν x = 74 (ή, ισοδύναμα, όταν z = 0.502). Από τον πίνακα 15 του παραρτήματος, προκύπτει ότι το παρατηρούμενο επίπεδο σημαντικότητας είναι

$$\hat{\alpha} = P(T_1 \geq 0.192 \mid H_0) = 1 - P(T_1 < 0.192 \mid H_0)$$

$$< 1 - P(T_1 < 0.187) = 1\text{-}0.99 = 0.01.$$

Είναι, δηλαδή, η τιμή του $\hat{\alpha}$ μικρότερη από 1%. Επομένως, η μέγιστη κατακόρυφη απόσταση μεταξύ της εμπειρικής συνάρτησης κατανομής του τυποποιημένου δείγματος και της συνάρτησης κατανομής της τυποποιημένης κανονικής είναι στατιστικά σημαντική σε επίπεδο σημαντικότητας 1% ή μεγαλύτερο. Κατά συνέπεια, η μηδενική υπόθεση απορρίπτεται σε όλα τα συνήθη επίπεδα σημαντικότητας.

Παρατήρηση: Το αποτέλεσμα του ελέγχου δεν είναι διαφορετικό από ό,τι θα περίμενε κανείς κοιτάζοντας λίγο πιο προσεκτικά τα δεδομένα. Πράγματι, μία ματιά στα δεδομένα δείχνει ότι μερικοί από τους άρρενες πέθαναν πολύ νέοι και ότι η κατανομή είναι κάπως στρεβλή (ασύμμετρη). Ένας μεγάλος αριθμός θανάτων, από την άλλη μεριά, συνέβη μετά την ηλικία των 80. Το σχήμα 4.3.1 απεικονίζει τα δεδομένα του δείγματος με μορφή ιστογράμματος με μήκος κλάσης 10.



Σχήμα 4.3.1

Ιστόγραμμα ηλικιών θανάτου στο Badenscallie

Λύση με το MINITAB: Η διεξαγωγή του ελέγχου του Lilliefors είναι δυνατή με το MINITAB μέσω της διεξαγωγής του ελέγχου Kolmogorov-Smirnov με βάση τις τυποποιημένες τιμές του δείγματος. Στην περίπτωση όμως που στο δείγμα υπάρχουν αρκετές τιμές που εμφανίζονται περισσότερες από μία φορά (όπως εδώ οι τιμές 41, 65, 74, 82, 85 και 87), το MINITAB δεν ενδείκνυται γαι την διεξαγωγή ελέγχων που απαιτούν υπολογισμό της τιμής της εμπειρικής συνάρτησης κατανομής S(.). Ο λόγος είναι ότι το πακέτο αντιστοιχίζει σε κάθε μία από τις μη διακεκριμένες τιμές την μέση τιμή των τάξεων (μεγέθους) που αυτές θα είχαν αν δεν ήταν διακεκριμένες και όχι την τάξη που θα είχε η μεγαλύτερη αν δεν ισοβαθμούσαν, όπως απαιτεί ο ορισμός της εμπειρικής συνάρτησης κατανομής.

Λύση με το SPSS: Η διεξαγωγή του ελέγχου κανονικότητας του Lilliefors συνίσταται στην διεξαγωγή του ελέγχου Kolmogorov-Smirnov για την υπό έλεγχο (κανονική) κατανομή με παραμέτρους τις

εκτιμήσεις των παραμέτρων με βάση το δείγμα. Συνεπώς, καταχωρίζουμε το δείγμα σε μία μεταβλητή (έστω **x**) και κάνουμε τον έλεγχο δηλώνοντας **Normal** στο πεδίο **Test Distribution**. Τα αποτελέσματα που παίρνουμε είναι:

One-Sample Kolmogorov-Smirnov Test

		X
N		30
Normal Parameters[a,b]	Mean	61.43
	Std. Deviation	25.04
Most Extreme Differences	Absolute	.192
	Positive	.144
	Negative	-.192
Kolmogorov-Smirnov Z		1.052
Asymp. Sig. (2-tailed)		.218
Exact Sig. (2-tailed)		.192
Point Probability		.000

a Test distribution is Normal.
b Calculated from data.

Η τιμή της ελεγχοσυνάρτησης T_1 δίνεται στην γραμμή **Most Extreme Differences, Absolute**. Αυτή είναι τ_1=0.192. Σημειώνεται ότι η τιμή του κρίσιμου επιπέδου που δίνει το SPSS (γραμμή **Exact Sig. (2-tailed)**) αντιστοιχεί σε αυτήν που δίνουν οι πίνακες του ελέγχου Kolmogorov-Smirnov. Επομένως, από τον παραπάνω πίνακα, χρειάζεται να κρατήσουμε μόνο την τιμή της ελεγχοσυνάρτησης T_1 και να την συγκρίνουμε, στην συνέχεια, με τα ποσοστιαία σημεία της κατανομής της από τον σχετικό πινάκα του παραρτήματος για τον έλεγχο Lilliefors.

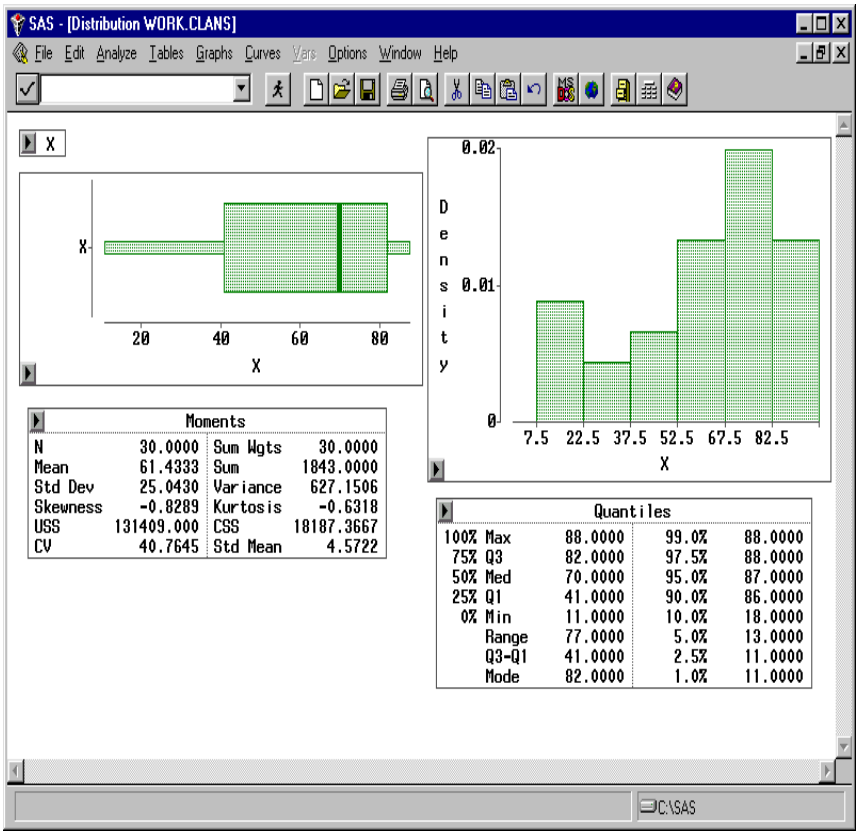
Λύση με το SAS: Οπως και τα πακέτα MINITAB και SPSS, το SAS δεν παρέχει τον έλεγχο Lilliefors, αλλά η τιμή της στατιστικής συνάρτησης ελέγχου μπορεί να υπολογισθεί μέσω του ελέγχου Kolmogorov-Smirnov για την υπό έλεγχο (κανονική) κατανομή, με

παραμέτρους (μέση τιμή και τυπική απόκλιση) τις εκτιμήσεις τους με βάση το δείγμα.

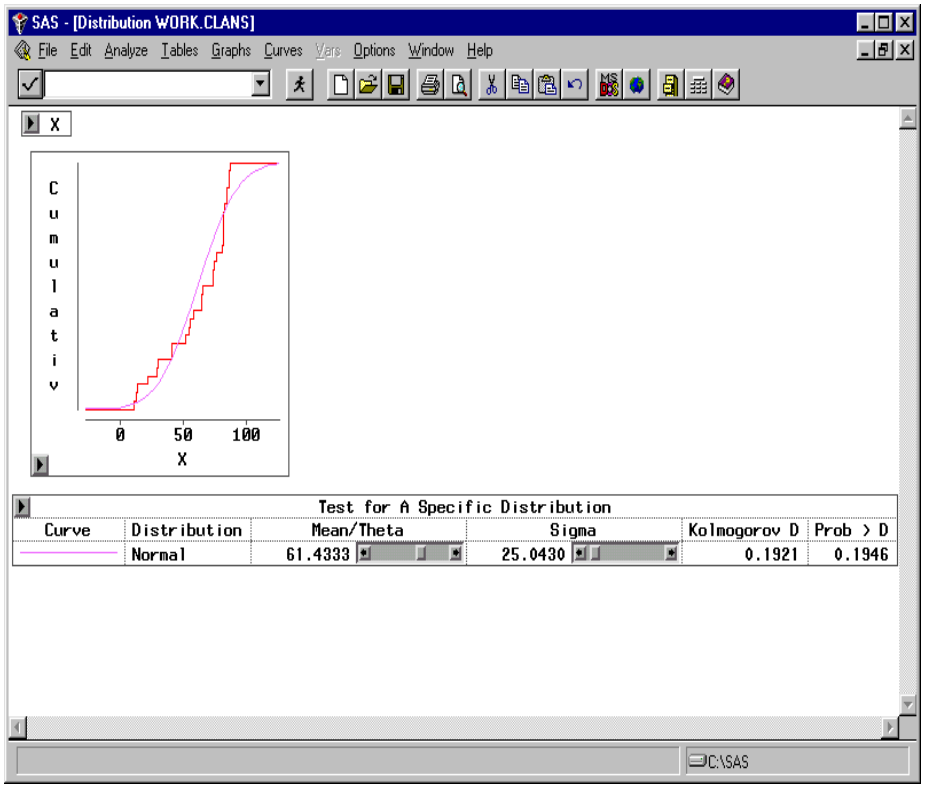
Εισάγουμε τα δεδομένα σε ένα αρχείο του SAS:

```
data clans;
input x @@;
cards;
11 13 14 22 29 30 41 41 52 55 56 59 65 65 66
74 74 75 77 81 82 82 82 82 83 85 85 87 87 88
;
run;
```

Ο ζητούμενος έλεγχος μπορεί να διεξαχθεί από το κεντρικό μενού. Επιλέγοντας κατά σειρά **Globals**, **Analyze** και **Interactive Data Analysis**, εμφανίζεται ένας κατάλογος με όλα τα διαθέσιμα αρχεία δεδομένων. Τα αρχεία τα οποία ο χρήστης εισάγει με εντολές, βρίσκονται στον κατάλογο **Work**. Από τον κατάλογο αυτό, επιλέγουμε (με διπλό πάτημα του πλήκτρου του ποντικιού) το όνομα του αρχείου που μας ενδιαφέρει (στο συγκεκριμένο παράδειγμα, το αρχείο **clans**). Στην συνέχεια, από το κεντρικό μενού επιλέγουμε **Analyze** και **Distribution (Y)**. Στο μενού που εμφανίζεται επιλέγουμε την μεταβλητή που μας ενδιαφέρει να μελετήσουμε (εδώ ονομάζεται **x**) και το πρόγραμμα δίδει γράφημα πλαισίου-απολήξεων, ιστόγραμμα καθώς και περιγραφικά μέτρα της μεταβλητής, όπως φαίνεται στο πλαίσιο που ακολουθεί.



Στην συνέχεια, για να προχωρήσουμε στον έλεγχο για κανονικότητα, επιλέγουμε, από το κεντρικό μενού, **Curves** (η επιλογή αυτή ενεργοποιείται μετά τον υπολογισμό των περιγραφικών μέτρων και την κατασκευή των γραφημάτων από το πακέτο) και **Test for a Specific Distribution**. Στον πίνακα που εμφανίζεται, δίνουμε τις τιμές της δειγματικής μέσης τιμής και τυπικής απόκλισης που υπολογίσθηκαν στα προηγούμενα. Το αποτέλεσμα δίνεται στο πλαίσιο που ακολουθεί.



Η τιμή της ελεγχοσυνάρτησης υπολογίζεται ίση με 0.1921. Βέβαια, το παρατηρούμενο επίπεδο σημαντικότητας (που εμφανίζεται στο πεδίο **Prob>D**) αντιστοιχεί στον έλεγχο Kolmogorov-Smirnov και, κατά συνέπεια, η τιμή 0.1921 θα πρέπει να συγκριθεί με το κατάλληλο ποσοστιαίο σημείο από τον πίνακα του ελέγχου Lilliefors, προκειμένου να αποφανθούμε για το εύλογο της υπόθεσης κανονικότητας της υπό εξέταση μεταβλητής.

Αξίζει επίσης να σημειωθεί ότι ο χρήστης μπορεί να επέμβει μεταβάλλοντας (με απλή μετακίνηση της μπάρας) την μέση τιμή και την τυπική απόκλιση της κανονικής κατανομής, παρατηρώντας τις μεταβολές του γραφήματος, τις μεταβολές στις αποκλίσεις που παρατηρούνται μεταξύ της εμπειρικής κατανομής και της θεωρητικής

κατανομής που δίδεται, και τις μεταβολές των αποτελεσμάτων του ελέγχου κανονικότητας, με βάση των παραμέτρων που δίδονται κάθε φορά.

Παρατήρηση: Συνήθως, η τιμή της ελεγχοσυνάρτησης του Lilliefors προσδιορίζεται γραφικά θεωρώντας τις αποστάσεις των γραφημάτων της εμπειρικής συνάρτησης κατανομής και της μηδενικής συνάρτησης κατανομής στα σημεία $z_i = (x_i - \bar{x})/s^*$, i = 1, 2, ..., n. Προσδιορίζεται, δηλαδή, ως η μέγιστη απόσταση μεταξύ των σημείων $(z_i , F_0^*(z_i))$ και $(z_i , S^*(z_i))$.

Παράδειγμα 4.3.2: Πενήντα διψήφιοι αριθμοί επελέγησαν τυχαία από ένα τηλεφωνικό κατάλογο. Οι αριθμοί, κατά αύξουσα σειρά μεγέθους, είναι οι εξής:

23 23 24 27 29 31 32 33 33 35
36 37 40 42 43 43 44 45 48 48
54 54 56 57 57 58 58 58 58 59
61 61 62 63 64 65 66 68 68 70
73 73 74 75 77 81 87 89 93 97

Να ελεγχθεί η υπόθεση ότι οι αριθμοί αυτοί θα μπορούσαν να αποτελούν παρατηρήσεις πάνω σε μια κανονική τυχαία μεταβλητή.

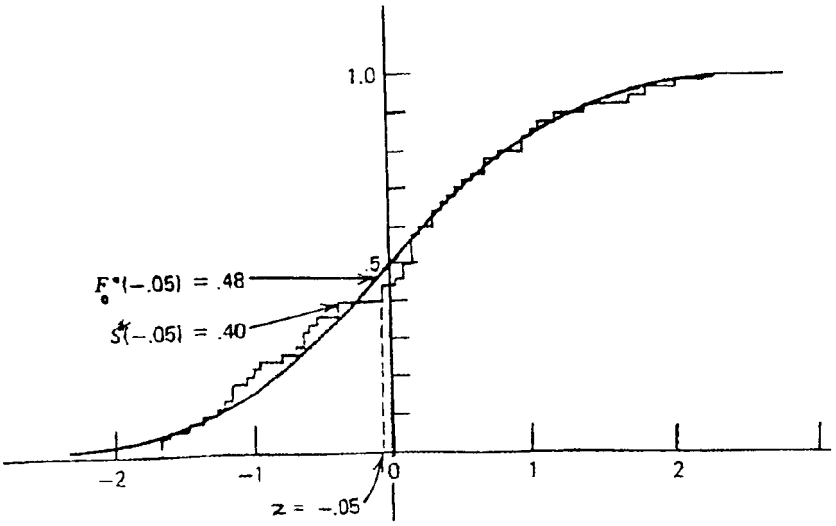
Λύση: Παρά το γεγονός ότι οι παρατηρήσεις προέρχονται από ένα σαφώς διακριτό δειγματοληπτικό πλαίσιο, έχει έννοια να ελεγχθεί η υπόθεση της κανονικότητας. Αυτό, γιατί η μη απόρριψη της μηδενικής υπόθεσης δεν συνεπάγεται ότι ο πληθυμός είναι κανονικός και, επομένως, συνεχής, αλλά ότι η διαφορά μεταξύ της κανονικής και της πραγματικής συνάρτησης κατανομής είναι αρκετά μικρή (ασήμαντη), ώστε να μην είναι δυνατή η ανίχνευσή της.

Τα στοιχεία X_i, i = 1, 2, ..., 50 του παραπάνω πίνακα τυποποιούνται αφαιρώντας από κάθε ένα από αυτά τον μέσο τους $\overline{x}$=55.04 και διαιρώντας το αποτέλεσμα με s^* =19.00.

X_i	Z_i	X_i	Z_i	X_i	Z_i	X_i	Z_i	X_i	Z_i
23	–1.69	36	–1.00	54	–0.05	61	0.31	73	0.95
23	–1.69	37	–0.95	54	–0.05	61	0.31	73	0.95
24	–1.63	40	–0.79	56	0.05	62	0.37	74	1.00
27	–1.48	42	–0.69	57	0.10	63	0.42	75	1.05
29	–1.37	43	–0.63	57	0.10	64	0.47	77	1.16
31	–1.27	43	–0.63	58	0.16	65	0.52	81	1.37
32	–1.21	44	–0.58	58	0.16	66	0.58	87	1.68
33	–1.16	45	–0.53	58	0.16	68	0.68	89	1.79
33	–1.16	48	–0.37	58	0.16	68	0.68	93	2.00
35	–1.05	48	–0.37	59	0.21	70	0.79	97	2.21

Τότε, η τιμή της ελεγχοσυνάρτησης του Lilliefors $T_1 = \sup_z \left|F_0^*(z) - S^*(z)\right|$.μπορεί να προσδιορισθεί από το σχήμα 4.3.2. Από το σχήμα αυτό, φαίνεται ότι η μέγιστη απόσταση μεταξύ $F_0^*(z)$ και $S^*(z)$ επιτυγχάνεται στα αριστερά του σημείου z = –0.05, οπότε $S^*(-0.05)$ = 0.40 και $F_0^*(-0.05)$=0.48. Τότε, η τιμή της T_1 είναι τ_1=0.08.

Σύμφωνα με τον έλεγχο Lilliefors, η μηδενική υπόθεση θα απορριφθεί σε επίπεδο σημαντικότητας α=0.05 αν η τιμή της T_1 υπερβαίνει το 0.95-ποσοστιαίο σημείο της κατανομής της.



Σχήμα 4.3.2

Γραφήματα των συνάρτησεων $F_0^*(z)$ και $S^*(z)$ που δείχνουν την μεγιστή απόκλισή τους

Από τον πίνακα 15 του παραρτήματος όμως, το σημείο αυτό είναι ίσο με

$$w_{0.95} = \frac{0.886}{\sqrt{50}} = 0.125.$$

Επομένως, η μηδενική υπόθεση δεν απορρίπτεται σε επίπεδο σημαντικότητας 0.05. (Στην πραγματικότητα, $\hat{\alpha} > 0.20$).

Για λόγους σύγκρισης, αξίζει να ελεγχθεί η ίδια υπόθεση με τον έλεγχο χ^2 : Ετσι, για παράδειγμα, θεωρώντας την ομαδοποίηση των δεδομένων στις κατηγορίες
"x<20", "20≤x<40", "40≤x<60", "60≤x<80", "80≤x<100" και "x≥100" προκύπτει ο εξής πίνακας:

Κατηγορία i	<20	20≤x<40	40≤x<60	60≤x<80	80≤x<100	x≥100
Παρατηρούμενη Συχνότητα O_i	0	12	18	15	5	0
Αναμενόμενη Συχνότητα E_i	1.5	9.0	19.5	15.5	4	0.5

Εδώ, $E_i = 50$ P (η μεταβλητή $\frac{X - \overline{x}}{S^*}$ ανήκει στην κατηγορία i), όπου, από τα δεδομένα, $\overline{x}$ =55.2 και s^* =18.7.

Θεωρώντας την ομαδοποίηση που φαίνεται στον παραπάνω πίνακα, προκύπτει ότι η τιμή της στατιστικής συνάρτησης

$$T = \sum_{i=4}^{4} \frac{(O_i - E_I)^2}{E_i}$$

είναι ίση με 0.395. Η τιμή αυτή δεν υπερβαίνει το 0.95-ποσοστιαίο σημείο της κατανομής χ^2 με 4–2–1=1 βαθμό ελευθερίας που είναι ίσο με $\chi^2_{1,0.95}$ =3.841, όπως προκύπτει από τον πίνακα 4 του παραρτήματος. Επομένως, η μηδενική υπόθεση δεν απορρίπτεται και με αυτόν τον έλεγχο σε επίπεδο σημαντικότητας 0.05. Το παρατηρούμενο επίπεδο σημαντικότητας â προκύπτει ότι είναι μεγαλύτερο του 0.25.

Λύση με το MINITAB: Οπως αναφέρθηκε στο παράδειγμα 4.3.1, το MINITAB δεν προσφέρεται για τον έλεγχο Lilliefors όταν έχουμε παρατηρήσεις με τουλάχιστον δύο εμφανίσεις στο δείγμα.

Στην συνέχεια, δίνεται η λύση του παραδείγματος με τον έλεγχο χ^2. Έχουμε ήδη δει ότι ο έλεγχος αυτός δεν διατίθεται αυτοματοποιημένος στο MINITAB και πρέπει οι διαδικασίες του να εκτελεσθούν βήμα – βήμα. Υπολογίζουμε, κατά τα ήδη γνωστά, εκτιμήσεις της μέσης τιμής και της διακύμανσης της κατανομής των δεδομένων, ορίζουμε τα όρια των κλάσεων, καταχωρίζουμε σε μία

στήλη, έστω με όνομα **o**, τις παρατηρούμενες συχνότητες, υπολογίζουμε τις αναμενόμενες συχνότητες κάτω από μια κανονική κατανομή με μέση τιμή και διακύμανση ίσες με τις εκτιμήσεις από το δείγμα και καταχωρίζουμε τις συχνότητες αυτές σε μία άλλη στήλη, έστω με όνομα **e**. Μετά το πέρας των διαδικασιών αυτών, οι δύο στήλες με τις συχνότητες δείχνουν ως εξής:

C2	C3
o	e
0.0	1.5
12.0	9.0
18.0	19.5
15.0	15.5
5.0	4.0
0.0	0.5

Από το κυρίως μενού, επιλέγουμε **Calc, Calculator** και πληκτρολογούμε **sum((o–e)**2/e)** στο πεδίο **Numeric Expression**. Προκύπτει ότι η τιμή της ελεγχοσυνάρτησης T είναι τ=3.3815. Αν θεωρήσουμε την ομαδοποίηση των κλάσεων που ακολουθήθηκε στην αναλυτική λύση, η διαδικασία επίλυσης είναι η ίδια, αλλά αναφέρεται στις 4 προκύπτουσες κλάσεις. Στην περίπτωση αυτή, οι δύο στήλες με τις συχνότητες θα είναι οι εξής:

C2	C3
o	**e**
12	10.5
18	19.5
15	15.5
5	4.5

Η προκύπτουσα τιμή της T είναι τ=0.401.

Λύση με το SPSS: Η διεξαγωγή του ελέγχου κανονικότητας του Lilliefors με το SPSS κατά τα ήδη γνωστά, δίνει τα εξής αποτελέσματα:

One-Sample Kolmogorov-Smirnov Test

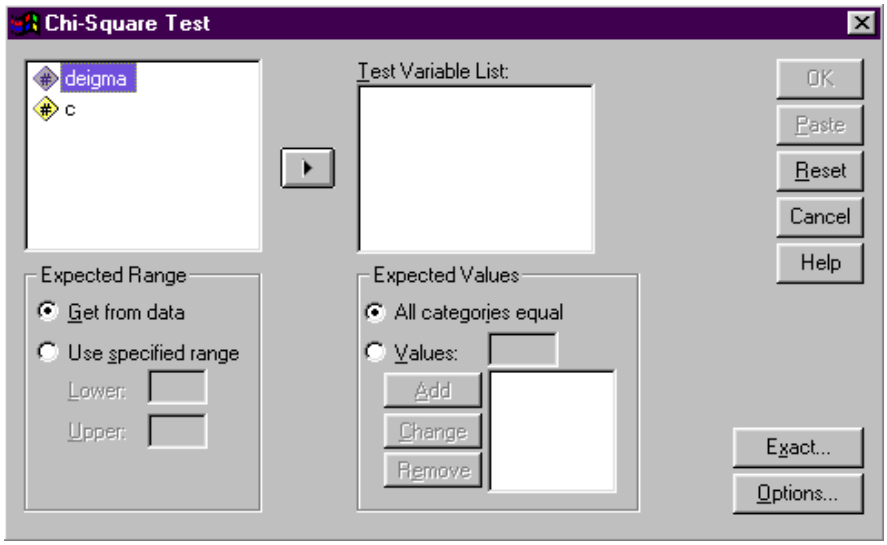
		X
N		50
Normal Parameters	Mean	55.04
	Std. Deviation	19.00
Most Extreme Differences[a,b]	Absolute	.081
	Positive	.069
	Negative	-.081
Kolmogorov-Smirnov Z		.573
Asymp. Sig. (2-tailed)		.898
Exact Sig. (2-tailed)		.871
Point Probability		.000

a Test distribution is Normal.
b Calculated from data.

Η τιμή της ελεγχοσυνάρτησης T_1 είναι τ_1=0.081. Οπως παρατηρήθηκε στην αναλυτική λύση του παραδείγματος η τιμή του κρίσιμου επιπέδου υπερβαίνει το 0.20. Συνεπώς, τα δεδομένα μπορούν εύλογα να θεωρηθούν ως προερχόμενα από την κανονική κατανομή.

Για να επιλύσουμε το παράδειγμα και με τον έλεγχο χ^2, ακολουθούμε την ήδη γνωστή διαδικασία. Από το δείγμα, υπολογίζουμε εκτιμήσεις των παραμέτρων της κατανομής, από την οποία προέρχονται τα δεδομένα. Στην συνέχεια, σε μία μεταβλητή, έστω με όνομα **c**, καταχωρίζουμε τα δεδομένα σε κωδικοποιημένη μορφή. Στην πρώτη κλάση αντιστοιχίζουμε τον κωδικό 1, στην δεύτερη τον κωδικό 2, κ.ο.κ. Στην συνέχεια, καταχωρίζουμε, στην μεταβλητή **c**, 12 φορές την τιμή 2, 18 φορές την τιμή 3, κ.ο.κ., δηλαδή, αντικαθιστούμε κάθε παρατήρηση του δείγματος με τον κωδικό της κλάσης της. Αφού υπολογίσουμε τις αναμενόμενες συχνότητες των κλάσεων κάτω από κανονική κατανομή με μέση τιμή και διακύμανση

ίσες με τις εκτιμήσεις τους από το δείγμα, επιλέγουμε από το κυρίως μενού **Analyze, Nonparametric Statistics, Chi-Square** οδηγούμενοι στο γνωστό πλαίσιο διαλόγου



Η μόνη διαφορά από τις προηγούμενες φορές που έχουμε χρησιμοποιήσει το πλαίσιο αυτό είναι ότι, τώρα, οι δύο ακραίες κλάσεις έχουν παρατηρούμενη συχνότητα 0. Για τον λόγο αυτό, χρειάζεται να δηλωθούν στο πρόγραμμα οι κωδικές τιμές όλων των κλάσεων. Τότε, το πρόγραμμα θα διαπιστώσει ότι οι κωδικοί των δύο ακραίων κλάσεων λείπουν από το δείγμα και θα τους αποδώσει παρατηρούμενη συχνότητα 0. Αυτό επιτυγχάνεται επιλέγοντας **Use specified range** στο πεδίο **Get from data** και δηλώνοντας 1 (τον κωδικό της πρώτης κλάσης) στο πεδίο **Lower** και 6 (τον κωδικό της τελευταίας κλάσης) στο πεδίο **Upper**. Αφού δώσουμε και τα υπόλοιπα κατάλληλα στοιχεία στα πεδία του παραθύρου, πιέζουμε **OK** και οδηγούμεθα στα εξής αποτελέσματα:

Test Statistics

	C
Chi-Square[a]	3.382
df	5
Asymp. Sig.	.641

a 3 cells (50.0%) have expected frequencies less than 5. The minimum expected cell frequency is .5.

Παρατηρούμε ότι το SPSS χρησιμοποιεί τους βαθμούς ελευθερίας που προκύπτουν όταν η κατανομή που καθορίζεται κάτω από την μηδενική υπόθεση είναι πλήρως καθορισμένη. Για τον λόγο αυτό, χρησιμοποιούμε μόνο την τιμή της ελεγχοσυνάρτησης T (πεδίο **Chi-Square**) που είναι ίση με 3.382. Την τιμή αυτή πρέπει να συγκρίνουμε με τα ποσοστιαία σημεία της κατανομής της T, για να μπορέσουμε να διαμορφώσουμε άποψη γαι το εύλογο της μηδενικής υπόθεσης. Οπως αναφέρθηκε στην αναλυτική λύση του παραδείγματος, το κρίσιμο επίπεδο του ελέγχου είναι >0.20. Επομένως, η υπόθεση ότι τα δεδομένα προέρχονται από κανονική κατανομή είναι εύλογη.

Αν επαναπροσδιορίσουμε τις κλάσεις, ακολουθώντας την ομαδοποίηση που θεωρήθηκε στην αναλυτική λύση, η διαδικασία επίλυσης η ίδια, αλλά αναφέρεται στις τέσσερις νέες κλάσεις δηλαδή, στην μεταβλητή **c**, καταχωρίζουμε 12 φορές την τιμή 1, 18 φορές την τιμή 2, 15 φορές την τιμή 3 και 5 φορές την τιμή 4. Στην περίπτωση αυτή, προκύπτουν τα εξής αποτελέσματα:

Test Statistics

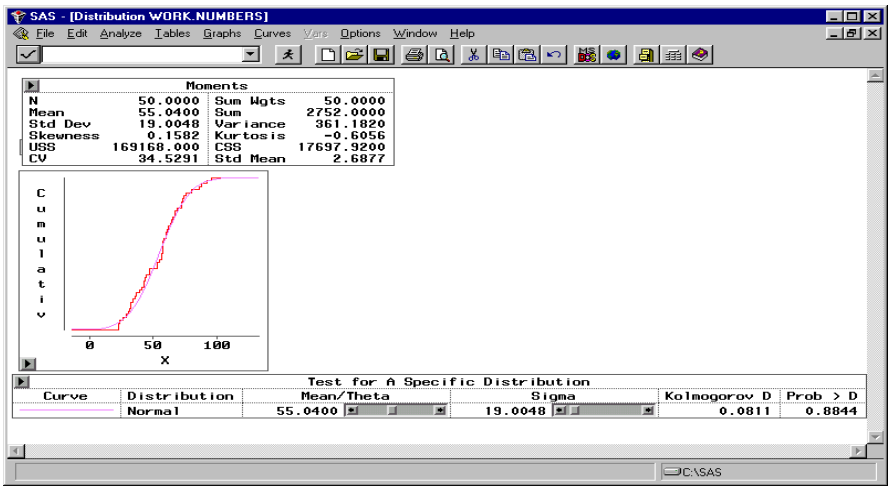
	C
Chi-Square[a]	.401
df	3
Asymp. Sig.	.940

a 1 cells (25.0%) have expected frequencies less than 5. The minimum expected cell frequency is 4.5.

Λύση με το SAS: Οπως και στο προηγούμενο παράδειγμα, η διεξαγωγή του ελέγχου μπορεί να γίνει με την χρήση του μενού του πακέτου. Ετσι, εισάγοντας τις εντολές σε ένα αρχείο με τις εντολές

```
data numbers;
input x @@;
cards;
23 23 24 27 29 31 32 33 33 35
36 37 40 42 43 43 44 45 48 48
54 54 56 57 57 58 58 58 58 59
61 61 62 63 64 65 66 68 68 70
73 73 74 75 77 81 87 89 93 97
;
run;
```

και επιλέγοντας το κατάλληλο μενού, οδηγούμεθα στα εξής αποτελέσματα:



Παρατηρούμε ότι η τιμή της ελεγχοσυνάρτησης υπολογίσθηκε ίση με 0.0811. Οπως αναφέρθηκε στην αναλυτική λύση του παραδείγματος, το κρίσιμο επίπεδο του ελέγχου είναι >0.20. Επομένως, η υπόθεση ότι τα δεδομένα προέρχονται από κανονική κατανομή είναι εύλογη. Το συμπέρασμα αυτό υποστηρίζεται και από το σχήμα που περιέχεται στον πίνακα αποτελεσμάτων όπου το απεικονιζόμενο γράφημα της εμπειρικής αθροιστικής συνάρτησης κατανομής είναι πολύ κοντά (σχεδόν ταυτίζεται) με την θεωρητική αθροιστική συνάρτηση κατανομής της κανονικής κατανομής.

4.3.2 Ο Ελεγχος Lilliefors για την Εκθετική Κατανομή

Μια δεύτερη παραλλαγή του ελέγχου Kolmogorov εξετάσθηκε από τον Lilliefors το 1969. Ο έλεγχος αυτός χρησιμοποιείται για τον έλεγχο της υπόθεσης ότι ο γεννήτορας πληθυσμός είναι εκθετικός με συνάρτηση κατανομής

$$F(x) = 1 - e^{-x/\mu}, \quad x > 0,$$

όπου μ είναι μία άγνωστη παράμετρος, η οποία πρέπει να εκτιμηθεί με βάση τα δεδομένα.

Οπως είναι γνωστό, η εκθετική κατανομή χρησιμοποιείται για την περιγραφή της κατανομής του χρονικού διαστήματος μεταξύ δύο διαδοχικών γεγονότων, όταν αυτά συμβαίνουν τυχαία μέσα στον χρόνο. Επομένως, ένας έλεγχος για την εκθετική κατανομή, όπως αυτός που θα περιγράψουμε, χρησιμοποιείται στην πραγματικότητα κυρίως ως έλεγχος τυχαιότητας.

Εστω $X_1, X_2, ..., X_n$ ένα τυχαίο δείγμα n παρατηρήσεων πάνω στην τυχαία μεταβλητή X της οποίας η συνάρτηση κατανομής είναι

$$F_x(x), \; x \in R.$$

Οι υποθέσεις που ενδιαφερόμαστε να ελέγξουμε είναι

$$H_0: F_x(x) = \begin{cases} 1 - e^{-x/\mu} & x > 0, \ \mu \text{ άγνωστη παράμετρος} \\ 0, & \text{διαφορετικά} \end{cases}$$

H_1: η κατανομή της X δεν είναι εκθετική.

Από τη μορφή της *μηδενικής* κατανομής, είναι προφανές ότι ο κατάλληλος μετασχηματισμός των δεδομένων, ο οποίος θα οδηγήσει στην παραλλαγή του ελέγχου Kolmogorov που απαιτείται στην προκειμένη περίπτωση, είναι ο

$$Z_i = X_i / \overline{X} \ , \quad i = 1, 2, ..., n,$$

όπου $\overline{X} = \sum_{i=1}^{n} X_i / n$. Έτσι, η στατιστική συνάρτηση που θα χρησιμοποιηθεί, αντί να εκφράζει ένα μέτρο της απόστασης μεταξύ της εμπειρικής συνάρτησης κατανομής S(x) από την συνάρτηση κατανομής κάτω από την μηδενική υπόθεση $F_0(x) = 1 - e^{-x/\mu}$, x>0, θα εκφράζει την απόσταση μεταξύ της εμπειρικής συνάρτησης κατανομής $S^*(z)$ των μετασχηματισμένων δεδομένων Z_1, Z_2, ..., Z_n από την συνάρτηση κατανομής

$$F^*(z) = 1 - e^{-z} \ , \ z>0.$$

Δηλαδή, ως στατιστική συνάρτηση, στην προκειμένη περίπτωση, χρησιμοποιείται η μέγιστη κατακόρυφη απόσταση μεταξύ των συναρτήσεων $S^*(z)$ και $F^*(z)$:

$$T_2 = \sup_z \left| F^*(z) - S^*(z) \right|$$

Είναι προφανές ότι μεγάλες τιμές της στατιστικής συνάρτησης T_2 αποτελούν ένδειξη ότι η μηδενική υπόθεση δεν αληθεύει. Επομένως, ο κανόνας απόφασης έχει την μορφή:

Απορρίπτουμε την μηδενική υπόθεση H_0 σε επίπεδο σημαντικότητας α αν η τιμή της στατιστικής συνάρτησης T_2 υπερβαίνει το (1–α)-ποσοστιαίο σημείο της κατανομής της, όπως αυτό δίνεται στον πίνακα 16 του παραρτήματος.

(Ο Lilliefors μελέτησε την κατανομή της στατιστικής συνάρτησης T_2 και προσδιόρισε κατά προσέγγιση τα ποσοστιαία σημεία της, αλλά η πραγματική κατανομή της συνάρτησης αυτής μελετήθηκε αργότερα από τον Durbin το 1975. Ο πίνακας 16 του παραρτήματος αναφέρεται στην πραγματική κατανομή της στατιστικής συνάρτησης T_2).

Παράδειγμα 4.3.3: Πιστεύεται ότι ο αριθμός των υπεραστικών τηλεφωνημάτων μέσω κάποιου τηλεφωνικού κέντρου είναι μία τυχαία διαδικασία με χρόνους μεταξύ των διαδοχικών τηλεφωνημάτων οι οποίοι ακολουθούν την εκθετική κατανομή. Ας υποθέσουμε ότι τα 10 πρώτα τηλεφωνήματα μετά την 1:00 το μεσημέρι κάποιας Δευτέρας έγιναν κατά τις εξής ώρες:

1:06 1:08 1:16 1:22 1:23 1:34 1:44 1:47 1:51 1:57.

Να ελεγχθεί η υπόθεση ότι ο χρόνος μεταξύ διαδοχικών τηλεφωνημάτων ακολουθεί την εκθετική κατανομή έναντι της εναλλακτικής ότι η κατανομή του χρόνου αυτού δεν είναι η εκθετική.

Λύση: Οι διαδοχικοί χρόνοι μεταξύ τηλεφωνημάτων, μετρώντας ως πρώτο διάστημα το διάστημα μεταξύ 1:00 και 1:06, είναι (σε πρώτα λεπτά) 6, 2, 8, 6, 1, 11, 10, 3, 4, 6, με μέσο $\overline{x} = 5.7$. Οι προκύπτουσες τιμές των Z_i, $S^*(z_i)$ και $F^*(z_i) = 1 - e^{-z_i}$, $= 1 - e^{-z_i}$, $i = 1, 2, ..., n$, καθώς και των διαφορών μεταξύ $S^*(z)$ και $F^*(z)$ και στις δύο πλευρές καθενός από τα άλματα της $S^*(z)$ δίνονται στον πίνακα που ακολουθεί. (Σημειώνεται ότι οι τιμές των X_i, $i = 1, 2, ..., n$ έχουν διαταχθεί κατά αύξουσα σειρά μεγέθους).

i	x_i	$z_i = x_i/\overline{x}$	$F^*(z_i) = 1 - e^{-z_i}$	$S^*(z_i)$	$F^*(z_i) - S^*(z_i)$	$F^*(z_i) - S^*(z_{i-1})$
1	1	0.1754	0.1609	0.1	0.0609	0.1609
2	2	0.3508	0.2959	0.2	0.0959	0.1959
3	3	0.5263	0.4092	0.3	0.2092	0.2092
4	4	0.7018	0.5043	0.4	0.1043	0.2043
5	6	1.0526	0.6510	0.7	–0.0490	0.2510
6	8	1.4035	0.7543	0.8	–0.0457	0.0543
7	10	1.7544	0.8270	0.9	–0.0730	0.0270
8	11	1.9298	0.8548	1.0	–0.1452	–0.0452

Από τις δύο τελευταίες στήλες του πίνακα, είναι προφανές ότι η μέγιστη απόλυτη απόκλιση μεταξύ $S^*(z)$ και $F^*(z)$ είναι ίση με 0.2510. Η μηδενική υπόθεση ότι η κατανομή της τυχαίας μεταβλητής X είναι εκθετική πρέπει να απορριφθεί σε επίπεδο σημαντικότητας α=0.05 μόνο εάν η τιμή της στατιστικής συνάρτησης T_2 υπερβαίνει την τιμή 0.3244 (η οποία προκύπτει από τον πίνακα 16 του παραρτήματος για n=10 και 1–α=0.95). Επειδή, όμως, $T_2 = 0.2510$, η μηδενική υπόθεση δεν απορρίπτεται σε επίπεδο σημαντικότητας 5%. Το κρίσιμο επίπεδο προκύπτει από τον σχετικό πίνακα, με την χρήση γραμμικής παρεμβολής, ότι είναι ίσο με $\hat{\alpha} = 0.25$. Επομένως, η υπόθεση ότι ο χρόνος μεταξύ διδαδοχικών τηλεφωνημάτων ακολουθεί την εκθετική κατανομή είναι μία *εύλογη* υπόθεση.

Ο έλεγχος αυτός, στην πράξη, γίνεται κυρίως με γραφικό τρόπο, χρησιμοποιώντας την γραφική παράσταση της εμπειρικής συνάρτησης κατανομής $S^*(z)$ και της συνάρτησης $F^*(z)$. Οι γραφικές αυτές παραστάσεις γίνονται με βάση τιμές μόνο στα n σημεία των μετασχηματισμένων δεδομένων $Z_1, Z_2, ..., Z_n$.

Λύση με το MINITAB: Ο έλεγχος Lilliefors για την εκθετική κατανομή δεν είναι διαθέσιμος στο MINITAB. Μπορούμε μόνο να χρησιμοποιήσουμε το πακέτο για να υπολογίσουμε την εμπειρική συνάρτηση κατανομής, να εκτιμήσουμε την μέση τιμή της εκθετικής κατανομής από την οποία υποτίθεται ότι προέρχονται τα δεδομένα, να τυποποιήσουμε τα δεδομένα με βάση την εκτίμηση της μέσης τιμής και να ελέγξουμε αν τα τυποποιημένα δεδομένα προέρχονται από την εκθετική κατανομή με παράμετρο 1.

Η εμπειρική συνάρτηση κατανομής υπολογίζεται με τον τρόπο που έχουμε δει και σε προηγούμενα παραδείγματα (υπολογισμός τάξεων μεγέθους και διαίρεση τους με το μέγεθος του δείγματος). Στη στήλη **x** καταχωρίζουμε το δείγμα και ταξινομούμε τις παρατηρήσεις του κατ' αύξουσα τάξη μεγέθους. Αποθηκεύουμε στην στήλη **rx** τις τάξεις μεγέθους των παρατηρήσεων και, στην στήλη **s**, την εμπειρική συνάρτηση κατανομής.

Με την επιλογή **Stat, Basic Statistics, Display Descriptive Statistics**, προκύπτει ότι η μέση τιμή του δείγματος είναι 5.7. Οι παρατηρήσεις του δείγματος *τυποποιούνται* με διαίρεση τους με 5.7 και καταχωρίζονται στην στήλη **zx**. Με την επιλογή **Calc**, **Probability Distributions**, **Exponential**, υπολογίζουμε τις τιμές της αθροιστικής συνάρτησης κατανομής της εκθετικής κατανομής με παράμετρο 1. (**Exp(1)**) στα σημεία που αντιστοιχούν στις παρατηρήσεις της **zx** και τις καταχωρίζουμε στην στήλη **f**. Τέλος, δημιουργούμε μία στήλη **s1**, στην οποία καταχωρίζουμε, ως πρώτη παρατήρηση, το 0 και, κατόπιν, τις 9 πρώτες παρατηρήσεις της **s**. Η μέγιστη απόλυτη διαφορά μεταξύ των τιμών των μεταβλητών **f** και **s** και μεταξύ των τιμών των μεταβλητών **f** και **s1** είναι η τιμή της ελεγχοσυνάρτησης T_2. Οι απόλυτες αυτές διαφορές καταχωρίζονται στις στήλες **d1** και **d2**, αντίστοιχα και το φύλλο δεδομένων δείχνει ως εξής:

x	rx	s	s1	zx	f	d1	d2
1	1	0.1	0.0	0.17544	0.160911	0.060911	0.160911
2	2	0.2	0.1	0.35088	0.295930	0.095930	0.195930
3	3	0.3	0.2	0.52632	0.409222	0.109222	0.209222
4	4	0.4	0.3	0.70175	0.504285	0.104285	0.204285
6	7	0.7	0.4	1.05263	0.650982	0.049018	0.250982
6	7	0.7	0.7	1.05263	0.650982	0.049018	0.049018
6	7	0.7	0.7	1.05263	0.650982	0.049018	0.049018
8	8	0.8	0.7	1.40351	0.754267	0.045733	0.054267
10	9	0.9	0.8	1.75439	0.826987	0.073013	0.026987
11	10	1.0	0.9	1.92982	0.854826	0.145174	0.045174

Η τιμή που προκύπτει για την ελεγχοσυνάρτηση T_2 είναι 0.251. Όπως έχει ήδη παρατηρηθεί στην αναλυτική λύση του παραδείγματος, με βάση τον πίνακα 16 του παραρτήματος, η τιμή του κρίσιμου επιπέδου είναι $\hat{a}=0.25$. Επομένως, η υπόθεση εκθετικότητας μπορεί να θεωρηθεί εύλογη.

Λύση με το SPSS: Η διεξαγωγή του ελέγχου Lilliefors για την εκθετική κατανομή με το SPSS, μπορεί να γίνει μέσω του ελέγχου Kolmogorov-Smirnov με **Test Distribution: Exponential** με παράμετρο ίση με την τιμή της εκτιμήτριάς της με βάση το δείγμα. Τα αποτελέσματα του ελέγχου είναι τα εξής:

One-Sample Kolmogorov-Smirnov Test

		X
N		10
Exponential parameter[a,b].	Mean	5.70
Most Extreme Differences	Absolute	.251
	Positive	.145
	Negative	-.251
Kolmogorov-Smirnov Z		.794
Asymp. Sig. (2-tailed)		.554
Exact Sig. (2-tailed)		.867
Point Probability		.005

a Test Distribution is Exponential.
b Calculated from data.

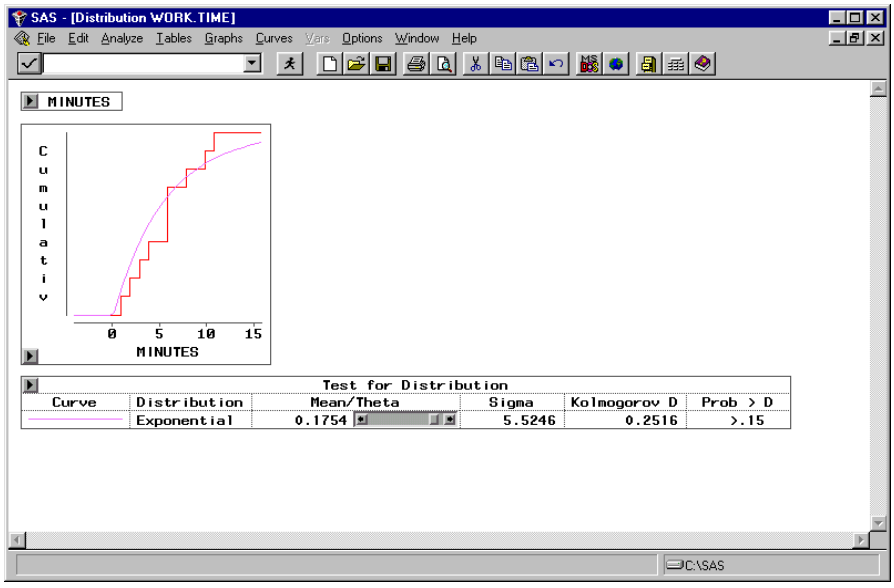
Η τιμή της ελεγχοσυνάρτησης T_2 είναι 0.251 και το κρίσιμο επίπεδο, όπως και στην περίπτωση του ελέγχου κανονικότητας, πρέπει να υπολογισθεί με βάση τον σχετικό πίνακα του παραρτήματος, αφού αυτό που δίνει το SPSS αντιστοιχεί σε έλεγχο Kolmogorov-Smirnov.

Όπως έχει ήδη παρατηρηθεί, με βάση τον πίνακα 16 του παραρτήματος, το κρίσιμο επίπεδο του ελέγχου προκύπτει ίσο με $\hat{\alpha} = 0.25$. Άρα, η υπόθεση ότι τα δεδομένα προέρχονται από την εκθετική κατανομή μπορεί να θεωρηθεί μία εύλογη υπόθεση.

Λύση με το SAS: Αρχικά εισάγουμε τα δεδομένα σε αρχείο με τις εντολές

```
data time;
input minutes @@;
cards;
6 2 8 6 1 11 10 3 4 6
;
run;
```

Στην συνέχεια, από το κεντρικό μενού επιλέγουμε **Globals**, **Analyze** και **Interactive Data Analysis**. Αφού επιλέξουμε το αρχείο και την υπό εξέταση μεταβλητή, από το κεντρικό μενού επιλέγουμε **Analyze** και **Distribution (Y)**. Όπως και προηγουμένως, αφού το πρόγραμμα εμφανίσει τα διαθέσιμα περιγραφικά μέτρα της μεταβλητής, επιλέγουμε, από το κεντρικό μενού, **Curves** και **Test for distribution**. Από την λίστα με τις διαθέσιμες κατανομές, επιλέγουμε την εκθετική κατανομή και, στο πεδίο **Parameter theta** δηλώνουμε την τιμή 0.17544, που είναι η αντίστροφη της τιμής 5.7 του δειγματικού μέσου. Τα αποτελέσματα της διαδικασίας περιέχονται στο εξής πλαίσιο:



Η τιμή της ελεγχοσυνάρτησης υπολογίζεται ίση με 0.2516. Και πάλι, το κρίσιμο επίπεδο σημαντικότητας που δίνεται από το πακέτο, αντιστοιχεί στον έλεγχο Kolmogorov. Επομένως, η ευλογοφάνεια της μηδενικής υπόθεσης μπορεί να εξετασθεί με βάση την τιμή του κρίσιμου επιπέδου όπως αυτή προσδιορίζεται μέσω του πίνακα ποσοστιαίων σημείων της κατανομής του παραρτήματος. Όπως έχει ήδη παρατηρηθεί το κρίσιμο επίπεδο αυτού του ελέγχου είναι 0.25 και, κατά συνέπεια, η H_0 μπορεί να θεωρηθεί εύλογη. Είναι ενδιαφέρον να παρατηρηθεί ότι και με βάση τον έλεγχο Kolmogorov που παρέχει το πακέτο, η μηδενική υπόθεση μπορεί να θεωρηθεί εύλογη σε όλα τα συνήθη επίπεδα σημαντικότητας. Το παρατηρούμενο επίπεδο στατιστικής σημαντικότητας του ελέγχου αυτού που δίνεται και αφορά στην εκθετική κατανομή, υπολογίζεται με τροποποίηση της ελεγχοσυνάρτησης Kolmogorov D: $(\sqrt{n} + 0.26 + \frac{0.5}{\sqrt{n}})(D - \frac{0.2}{n})$.

Σημείωση: Ο έλεγχος Kolmogorov έχει επίσης επεκταθεί από τον Lilliefors το 1973 για τον έλεγχο της υπόθεσης ότι ο γεννήτορας

πληθυσμός ακολουθεί την κατανομή γάμμα στην περίπτωση που υπάρχουν άγνωστες παράμετροι.

4.3.3 Ο Ελεγχος των Shapiro-Wilk για την Κανονική Κατανομή

Ένας άλλος πολύ γνωστός έλεγχος καλής προσαρμογής για την κανονική κατανομή, ο οποίος μπορεί να χρησιμοποιηθεί στην θέση του ελέγχου Lilliefors, είναι ο έλεγχος κανονικότητας των Shapiro και Wilk. Εμπειρικές μελέτες έχουν δείξει ότι αυτός ο έλεγχος έχει υψηλή ισχύ σε πολλές περιπτώσεις σε σύγκριση με πολλούς άλλους ελέγχους της σύνθετης υπόθεσης της κανονικότητας, περιλαμβανομένου και του ελέγχου του Lilliefors και του ελέγχου χ^2. Θα πρέπει να τονισθεί, βέβαια, ότι ο έλεγχος αυτός δεν είναι τύπου Kolmogorov. Παρ' όλα αυτά, περιλαμβάνεται στο κεφάλαιο αυτό λόγω της μεγάλης του χρησιμότητας.

Έστω $X_1, X_2, ..., X_n$ δείγμα n παρατηρήσεων πάνω στην τυχαία μεταβλητή X, της οποίας η άγνωστη συνάρτηση κατανομής είναι

$$F_X(x),\ x \in R.$$

Οι προς έλεγχο υποθέσεις είναι οι εξής:

H_0: η $F_X(x)$ είναι η συνάρτηση κατανομής της κανονικής κατανομής με άγνωστη μέση τιμή και άγνωστη διασπορά

H_1: η $F_X(x)$ είναι η συνάρτηση κατανομής μίας μη κανονικής κατανομής.

Η στατιστική συνάρτηση για τον έλεγχο των υποθέσεων αυτών είναι η

$$T_3 = \frac{\left[\sum_{i=1}^{k} \alpha_i \left(X^{(n-i+1)} - X^{(i)}\right)\right]^2}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2},$$

όπου $X^{(i)}$ είναι η i παρατήρηση του διατεταγμένου κατά αύξουσα τάξη μεγέθους δείγματος, k είναι ένας ακέραιος αριθμός περίπου ίσος με n/2 και α_i , i = 1, 2, ..., k είναι σταθεροί συντελεστές. Εξαρτάται, δηλαδή, η T_3 τόσο από τις τετραγωνικές αποκλίσεις των παρατηρήσεων $\overline{X}^{(i)}$ από τον μέσο τους $\overline{X}$, όσο και από τις αποκλίσεις που έχουν στο διατεταγμένο δείγμα η πρώτη (ελάχιστη) παρατήρηση από την τελευταία (μέγιστη) παρατήρηση, η δεύτερη από την προτελευταία κ.ο.κ. (Στην πράξη, για τον καθορισμό της τιμής της στατιστικής συνάρτησης T_3, υπολογίζουμε πρώτα τον παρονομαστή

$$D = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

όπου $\overline{X}$ είναι ο μέσος των παρατηρήσεων. Στην συνέχεια, διατάσσουμε κατά αύξουσα σειρά μεγέθους το δείγμα των παρατηρήσεων $X^{(1)} \le X^{(2)} \le .... \le X^{(n)}$. Από τον πίνακα 17 του παραρτήματος προσδιορίζουμε τους συντελεστές α_i, i = 1, 2, ..., k, για το δοθέν μέγεθος δείγματος n, και για k = n/2. Η στατιστική συνάρτηση T_3 υπολογίζεται, τότε, από τον τύπο

$$T_3 = \left[\sum_{i=1}^{k} \alpha_i \left(X^{(n-i+1)} - X^{(i)}\right)\right]^2 \Big/ D.$$

Η στατιστική συνάρτηση T_3 συχνά συμβολίζεται με W και ο έλεγχος συχνά ονομάζεται έλεγχος W.

Παρατηρούμε ότι οι μικρές τιμές της στατιστικής συνάρτησης T_3 είναι εκείνες οι οποίες αποτελούν ένδειξη ότι η μηδενική υπόθεση δεν είναι αληθής. Επομένως, ο κανόνας απόφασης είναι ο εξής:

Η μηδενική υπόθεση H_0 απορρίπτεται σε επίπεδο σημαντικότητας α εάν η τιμή της στατιστικής συνάρτησης T_3 είναι μικρότερη από το α-ποσοστιαίο σημείο της κατανομής της, όπως αυτό δίνεται στον πίνακα 18 του παραρτήματος.

Σημείωση: Οπως φαίνεται, ο σχετικός πίνακας επιτρέπει την χρήση του ελέγχου W μόνο στην περίπτωση που $n \leq 50$. Για την περίπτωση $n > 50$, έχουν μελετηθεί εναλλακτικοί και παρόμοιας φύσης έλεγχοι από τους D' Agostino (1971) και από τους Shapiro και Francia (1972).

Παράδειγμα 4.3.4: Ας θεωρήσουμε τους 50 διψήφιους αριθμούς του παραδείγματος 4.3.2 που αναφέρεται στον έλεγχο κανονικότητας του Lilliefors. Οπως είδαμε εκεί, ο έλεγχος του Lilliefors οδήγησε στην μη απόρριψη της μηδενικής υπόθεσης με κρίσιμο επίπεδο $\hat{\alpha} > 0.20$. Είδαμε, επίσης εκεί, ότι ο έλεγχος χ^2 οδήγησε και αυτός στην μη απόρριψη της μηδενικής υπόθεσης με κρίσιμο επίπεδο $\hat{\alpha}$ αρκετά μεγαλύτερο από το 0.25. Η υπόθεση της κανονικότητας θα ελεγχθεί τώρα με τον έλεγχο W.

Λύση: Από τον σχετικό πίνακα του παραρτήματος, υπολογίζονται οι σταθεροί συντελεστές α_i, i = 1, 2, ..., 25. Οι τιμές αυτές μαζί με τις τιμές των στατιστικών συναρτήσεων $X^{(50-i+1)} - X^{(i)}$, i = 1, 2, ..., 25 δίνονται στον πίνακα που ακολουθεί.

i	α_i	$X^{(50-i+1)} - X^{(i)}$	i	α_i	$X^{(50-i+1)} - X^{(i)}$
1	0.3751	97–23	14	0.0846	66–42
2	0.2574	93–23	15	0.0764	65–43
3	0.2260	98–24	16	0.0685	64–43
4	0.2032	87–27	17	0.0608	63–44
5	0.1847	81–29	18	0.0532	62–45
6	0.1691	77–31	19	0.0459	61–48
7	0.1554	75–32	20	0.0386	61–48
8	0.1430	74–33	21	0.0314	59–54
9	0.1317	73–33	22	0.0244	58–54
10	0.1212	73–35	23	0.0174	58–56
11	0.1113	70–36	24	0.0104	58–57
12	0.1020	68–37	25	0.0035	58–57
13	0.0932	68–40			

Από τα στοιχεία του πίνακα, προκύπτει ότι ο αριθμητής της στατιστικής συνάρτησης T_3 είναι

$$\left[\sum_{i=1}^{25} \alpha_i \left(X^{(50-i+1)} - X^{(i)}\right)\right]^2 = \left[(0.3751)(97-23) + \ldots + (0.0035)(58-57)\right]^2$$

$$= \left[130.63\right]^2 = 17064,$$

ενώ ο παρονομαστής έχει την τιμή

$$D = \sum_{i=1}^{50} \left(X_i - \overline{X}\right)^2 = 17698 \ .$$

Επομένως, η τιμή τ_3 της στατιστικής συνάρτησης T_3 είναι

$$\tau_3 = \frac{17064}{17698} = 0.9642.$$

Παρατηρούμε ότι η τιμή αυτή βρίσκεται μεταξύ των 0.10 και 0.50 ποσοστιαίων σημείων της κατανομής της στατιστικής συνάρτησης T_3. Με την μέθοδο της παρεμβολής, βρίσκουμε ότι $\hat{\alpha} = 0.29$, κατά προσέγγιση.

Παρατήρηση 1: Συχνά, η στατιστική συνάρτηση T_3 μετασχηματίζεται σε μία κατά προσέγγιση κανονική μεταβλητή, της οποίας η τιμή συγκρίνεται, στην συνέχεια, με τα ποσοστιαία σημεία της τυποποιημένης κανονικής κατανομής οδηγώντας, έτσι, στην τιμή του κρίσιμου επιπέδου $\hat{\alpha}$. Η εκτίμηση του κρισίμου επιπέδου με την μέθοδο αυτή είναι, εν γένει, περισσότερο ακριβής. Ο μετασχηματισμός της στατιστικής συνάρτησης T_3 σε μία κατά προσέγγιση κανονική μεταβλητή γίνεται με την βοήθεια του πίνακα 19 του παραρτήματος. Στο πλαίσιο του παραδείγματός μας, έχουμε από τον πίνακα αυτό για n=50, $b_{50} = -7.677$, $c_{50} = 2.212$ και $d_{50} = 0.1436$. Η παρατηρηθείσα τιμή της στατιστικής συνάρτησης T_3 αντικαθίσταται, τότε, στον τύπο

$$G = b_{50} + c_{50}\,\ell n\left[\frac{T_3 - d_{50}}{1 - T_3}\right],$$

οδηγώντας, έτσι, στην τιμή

$$\begin{aligned} G &= -7.677 + (2.212)\,\ell n\left[\frac{0.9642 - 0.1436}{1 - 0.9642}\right] \\ &= -0.7488. \end{aligned}$$

Η τιμή αυτή είναι τιμή μίας κατά προσέγγιση τυποποιημένης κανονικής κατανομής και οδηγεί στο κρίσιμο επίπεδο $\hat{\alpha}$=0.227 με βάση τον πίνακα της τυποποιημένης κανονικής κατανομής.

Παρατήρηση 2: Ενα πολύ χρήσιμο χαρακτηριστικό του ελέγχου Shapiro-Wilk είναι ότι αρκετοί ανεξάρτητοι έλεγχοι καλής προσαρμογής μπορούν να συνδυασθούν (ενοποιηθούν) σε έναν ενιαίο έλεγχο κανονικότητας. Αυτό βοηθά πολύ στην περίπτωση όπου αρκετά μικρά δείγματα από, ενδεχομένως, διαφορετικούς πληθυσμούς είναι ανεπαρκή από μόνα τους να οδηγήσουν σε απόρριψη της υπόθεσης της κανονικότητας, αλλά συνδυαζόμενα παρέχουν ενδείξεις που είναι αρκετές για την απόρριψη της υπόθεσης της κανονικότητας. Η τεχνική αυτή εφαρμόζεται στο παράδειγμα που ακολουθεί.

Λύση με το MINITAB: Ο έλεγχος των Shapiro-Wilk για την κανονική κατανομή δεν είναι διθέσιμος από το MINITAB. Η διεξαγωγή του μπορεί να γίνει μόνο έμμεσα. Από τον ορισμό της ελεγχοσυνάρτησης T_3, προκύπτει ότι μπορούμε να εργασθούμε ως εξής: Αντιστοιχίζουμε στην μικρότερη τιμή του δείγματος ($X^{(1)}$) τον αριθμό $-\alpha_1$, στην δεύτερη μικρότερη ($X^{(2)}$) τον αριθμό $-\alpha_2$ και, συνεχίζουμε με αυτόν τον τρόπο ώσπου να φθάσουμε στα μισά του δείγματος, όπου στην k κατά σειρά μεγέθους παρατήρηση ($X^{(k)}$) αντιστοιχίζουμε την τιμή $-\alpha_k$, στην k+1 κατά σειρά μεγέθους παρατήρηση($X^{(k+1)}$) αντιστοιχίζουμε την τιμή α_k και συνεχίζουμε με τον ίδιο τρόπο μέχρι την μεγαλύτερη παρατήρηση ($X^{(n)}$), στην οποία αντιστοιχίζουμε την τιμή α_1. Τότε, η τιμή της T_3 δεν είναι παρά το τετράγωνο της τιμής του συντελεστή συσχέτισης μεταξύ των διατεταγμένων παρατηρήσεων του δείγματος και των αριθμών α_i.

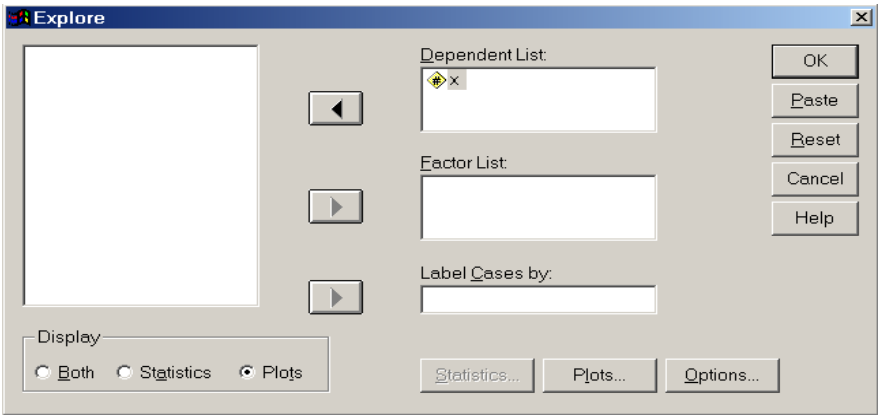
Κατά συνέπεια, ο υπολογισμός της τιμής της T_3 μπορεί να γίνει ως εξής: Σε μία στήλη (έστω **x**) καταχωρίζουμε το δείγμα και διατάσσουμε τις παρατηρήσεις του κατ' αύξουσα τάξη μεγέθους. Σε μία άλλη στήλη (έστω **a**), καταχωρίζουμε τις τιμές α_i. Ο συντελεστής

συσχέτισης μεταξύ των μεταβλητών αυτών υπολογίζεται από το πακέτο ίσος με 0.982 .

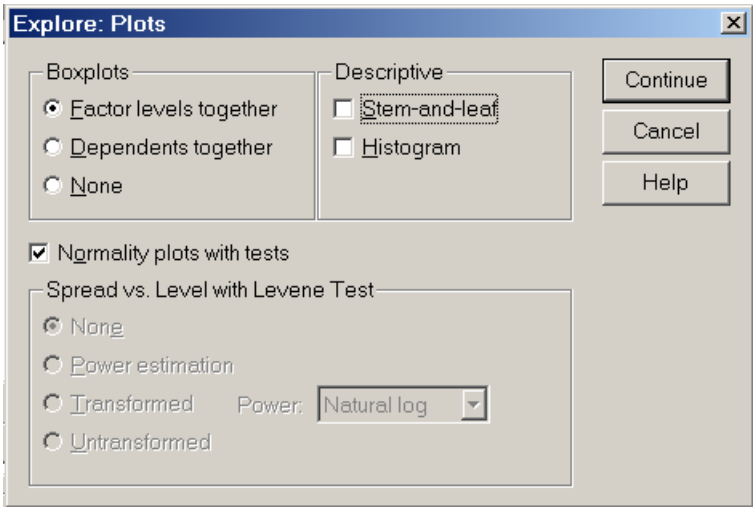
```
Correlation of x and a = 0.982
```

Επομένως, η τιμή της ελεγχοσυνάρτησης T_3 είναι 0.9643. Όπως παρατηρήθηκε στην αναλυτική λύση του παραδείγματος, η τιμή του κρίσιμου επιπέδου του ελέγχου αυτού είναι 0.29. Κατά συνέπεια, η υπόθεση της κανονικότητας μπορεί να θεωρηθεί εύλογη.

Λύση με το SPSS: Καταχωρίζουμε το δείγμα σε μία μεταβλητή (έστω **x**) και επιλέγουμε **Analyze, Descriptive Statistics, Explore** οδηγούμενοι στο εξής πλαίσιο διαλόγου:



Στο πεδίο **Dependent List**, δηλώνουμε την μεταβλητή που περιέχει το δείγμα (**x**). Στο πεδίο **Display**, επιλέγουμε **Plots**. Πιέζοντας **Plots**, εμφανίζεται το εξής πλαίσιο διαλόγου:



Επιλέγοντας **Normality plots with tests**, το πρόγραμμα δίνει κάποια γραφήματα ελέγχου κανονικότητας μαζί με τα αποτελέσματα δύο ελέγχων, ένας από τους οποίους είναι αυτός των **Shapiro-Wilk**. Τα αποτελέσματα που μας ενδιαφέρουν αποτελούν τμήμα των αποτελεσμάτων που δίνει το πρόγραμμα και είναι τα εξής:

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
X	.081	50	.200*	.964	50	.294

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

Η τιμή της ελεγχοσυνάρτησης T_3 (πεδίο **Shapiro-Wilk Statistic**) είναι 0.964 και η τιμή του κρίσιμου επιπέου είναι 0.294. Συνεπώς, η υπόθεση της κανονικότητας των δεδομένων μπορεί να θεωρηθεί εύλογη.

Σημείωση: Παρατηρούμε ότι εκτός από την τιμή της ελεγχοσυνάρτησης T_3 (W, στον παραπάνω πίνακα), το SPSS δίνει και την τιμή μιας στατιστικής συνάρτησης τύπου Kolmogorov-Smirnov συνοδευόμενη από το αντίστοιχο κρίσιμο επίπεδο. Αυτή η στατιστική συνάρτηση είναι η ελεγχοσυνάρτηση του ελέγχου Lilliefors με το

αντίστοιχο κρίσιμο επίπεδο. Επομένως, μπορούμε από το ίδιο πλαίσιο διαλόγου να διεξάγουμε τόσο τον έλεγχο κανονικότητας κατά Lilliefors όσο και τον έλεγχο κατά Shapiro-Wilk.

Λύση με το SAS: Για την διεξαγωγή του ελέγχου αυτού, χρησιμοποιούμε την εντολή **proc univariate**, η οποία παρέχει εξ ορισμού τον έλεγχο W.

Ετσι, χρησιμοποιώντας τις εντολές

```
data numbers;
input x @@;
cards;
23 23 24 27 29 31 32 33 33 35
36 37 40 42 43 43 44 45 48 48
54 54 56 57 57 58 58 58 58 59
61 61 62 63 64 65 66 68 68 70
73 73 74 75 77 81 87 89 93 97
;
run;
proc univariate normal;
var x;
run;
```

προκύπτουν τα εξής αποτελέσματα:

```
                                  Univariate Procedure

Variable=X

       Moments                                      Quantiles(Def=5)
N        50         Sum Wgts    50      100% Max     97      99%         97
Mean     55.04      Sum        2752     75% Q3       68      95%         89
Std Dev 19.00479    Variance   361.182   50% Med     57.5     90%         79
Skewness 0.158196    Kurtosis  -0.60557  25% Q1      40       10%         30
USS       169168     CSS       17697.92   0% Min     23        5%         24
CV      34.52905    Std Mean   2.687683                        1%         23
T:Mean=0 20.4786     Pr>|T|    0.0001                Range          74
Num ^= 0    50       Num > 0          50            Q3-Q1          28
M(Sign)     25       Pr>=|M|       0.0001           Mode           58
Sgn Rank   637.5     Pr>=|S|       0.0001
W:Normal   0.964217  Pr<W          0.2309
```

Η τιμή της ελεγχοσυνάρτησης υπολογίσθηκε ίση με 0.964217 (στο πεδίο **W:Normal**) και το παρατηρούμενο επίπεδο στατιστικής σημαντικότητας που αντιστοιχεί στην τιμή αυτή είναι 0.2309. Επομένως, η μηδενική υπόθεση μπορεί να θεωρηθεί εύλογη σε οποιοδήποτε από τα συνήθη επίπεδα σημαντικότητας.

Παράδειγμα 4.3.5: Οταν ζητείται η εκδήλωση ενδιαφέροντος για την εκμίσθωση θαλάσσιας περιοχής, η οποία πιστεύεται ότι περιέχει πετρέλαιο, αρκετές εταιρείες πετρελαίου συνήθως υποβάλλουν προσφορές για να εξασφαλίσουν το δικαίωμα γεώτρησης για την αναζήτηση πετρελαίου στην περιοχή αυτή. Η κατανομή των προσφορών αυτών συχνά θεωρείται ότι περιγράφεται από την λογαριθμοκανονική κατανομή. Δηλαδή, ο λογάριθμος των προσφορών θεωρείται ότι ακολουθεί την κανονική κατανομή. Αλλά, οι μέσες τιμές και οι διασπορές ενδέχεται να διαφέρουν από μίσθωση σε μίσθωση. Επίσης, ο αριθμός των προσφορών για οποιαδήποτε μίσθωση είναι συνήθως πολύ μικρός για να αποτελεί ένδειξη για το εάν η υπόθεση της κανονικότητας των λογαρίθμων των προσφορών είναι εύλογη ή όχι.

Για να ελεγχθεί η υπόθεση

H_0: Οι προσφορές κατανέμονται λογαριθμοκανονικά,

εναντίον της εναλλακτικής ότι δεν κατανέμονται λογαριθμοκανονικά, μαζεύτηκαν στοιχεία σχετικά με τις προσφορές 16 διαφορετικών μισθώσεων, όπως φαίνεται στον πίνακα που ακολουθεί. Ο έλεγχος Shapiro-Wilk εφαρμοζόμενος στους λογαρίθμους των προσφορών κάθε μιας από τις μισθώσεις ξεχωριστά, οδήγησε στην απόρριψη της μηδενικής υπόθεσης σε 4 από τις 16 περιπτώσεις σε επίπεδο σημαντικότητας $\alpha=0.05$. Όμως, μερικές από τις μισθώσεις επέδειξαν ικανοποιητική συμφωνία με την μηδενική υπόθεση κρίνοντας από τα αντίστοιχα κρίσιμα επίπεδα τα οποία ήταν μεγαλύτερα από 0.50. Για να συνδυασθούν τα αποτελέσματα των 16 επιμέρους ελέγχων, ακολουθούνται τα εξής βήματα:

1. Η τιμή κάθε μιας των στατιστικών συναρτήσεων T_3 μετασχηματίζεται σε τιμή της μεταβλητής G, όπως περιγράφεται στον σχετικό πίνακα του παραρτήματος.
2. Οι 16 διαφορετικές τιμές της μεταβλητής G προστίθενται και το άθροισμά τους διαιρείται με $\sqrt{n}$. Δηλαδή αν G_i , i = 1, 2, ..., 16 είναι οι διαφορετικές τιμές της μεταβλητής G, θεωρούμε την τυχαία μεταβλητή $Z = \sum_{i=1}^{n} G_i \Big/ \sqrt{n}$. Είναι προφανές ότι η μεταβλητή Z ακολουθεί την τυποποιημένη κανονική κατανομή κάτω από την μηδενική υπόθεση.
3. Η μηδενική υπόθεση H_0 απορρίπτεται σε επίπεδο σημαντικότητας α, αν η τιμή της στατιστικής συνάρτησης Z είναι μικρότερη από το α-ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής.

Οι παραπάνω υπολογισμοί για τις 16 θεωρηθείσες μισθώσεις συνοψίζονται στον πίνακα που ακολουθεί.

Με βάση τα στοιχεία του πίνακα αυτού, έχουμε ότι η τιμή της στατιστικής συνάρτησης Z είναι

$$z = \frac{-8.2099}{\sqrt{16}} = -2.0525.$$

Η τιμή αυτή είναι μικρότερη από το 0.05–ποσοστιαίο σημείο της τυποποιημενης κανονικής κατανομής, δηλαδή

$$z < z_{0.05} \equiv -1.645.$$

Επομένως, η υπόθεση H_0 απορρίπτεται σε επίπεδο σημαντικότητας α=0.05.

Το κρίσιμο επίπεδο του ελέγχου είναι $\hat{\alpha}$=0.020. Κατά συνέπεια, η υπόθεση της λογαριθμοκανονικότητας των προσφορών δεν φαίνεται να είναι εύλογη.

Μίσθωση	Αριθμός προσφορών	T_3	G
1	14	0.9243	–0.6550
2	14	0.9757	1.3559
3	14	0.9717	1.0939
4	14	0.8772	–1.5848
5	14	0.9537	0.2345
6	15	0.9135	–1.0093
7	15	0.8629*	–1.9321
8	15	0.8786*	–1.6806
9	15	0.8515*	–2.1011
10	15	0.9226	–0.7966
11	15	0.9581	0.3354
12	15	0.9625	0.5344
13	16	0.9178	–1.0151
14	16	0.8596*	–2.1011
15	15	0.9603	0.4323
16	16	0.9669	0.6795
		Σύνολο	–8.2099

* Τιμή στατιστικά σημαντική στο επίπεδο α=0.05

Λύση με το MINITAB: Καταχωρίζουμε τις τιμές των στατιστικών συναρτήσεων T_3 σε μία στήλη (με το όνομα **t**). Σε τρεις άλλες στήλες (με ονόματα **b**, **c**, **d**), καταχωρίζουμε τις τιμές των σταθερών b_n, c_n, d_n, που αντιστοιχούν σε κάθε μία από τις τιμές της στατιστικής συνάρτησης T_3 βάσει του αριθμού n των παρατηρήσεων του δείγματος, από το οποίο αυτή έχει υπολογισθεί. Για παράδειγμα, στην πρώτη τιμή της T_3 που προέρχεται από ένα δείγμα 14 παρατηρήσεων, θα αντιστοιχίσουμε τις τιμές των b_{14}, c_{14}, d_{14}. Με βάση όλες αυτές τις στήλες, από την επιλογή **Calc**, **Calculator** με **expression**

b+c * LOGE((t-d)/(1-t))

δημιουργούμε την στήλη **g** που περιέχει τις τιμές των στατιστικών συναρτήσεων G που αντιστοιχούν στις στατιστικές συναρτήσεις T_3. Στην συνέχεια, αθροίζουμε τις τιμές αυτής της στήλης και διαιρούμε το προκύπτον άθροισμα με την τετραγωνική ρίζα του αριθμού των στοιχείων της. Αυτό γίνεται πάλι από την επιλογή **Calc**, **Calculator** με **expression**

SUM(g)/SQRT(COUNT(g)).

Η τιμή που προκύπτει, είναι η τιμή της ελεγχοσυνάρτησης Z που χρησιμοποιείται για τον έλεγχο.

Στο παράδειγμα μας προκύπτει Z=-2.0525. Η τιμή του κρίσιμου επιπέδου του ελέγχου έχει προσδιορισθεί (στην αναλυτική λύση του παραδείγματος) ίση με 0.02. Επομένως, η υπόθεση ότι τα δεδομένα προέρχονται από κανονική κατανομή δεν μπορεί να θεωρηθεί εύλογη.

Λύση με το SPSS: Καταχωρίζουμε τις τιμές των στατιστικών συναρτήσεων T_3 σε μία στήλη (με το όνομα **t**). Σε τρεις άλλες στήλες (με ονόματα **b**, **c**, **d**) καταχωρίζουμε τις τιμές των σταθερών b_n, c_n, d_n, που αντιστοιχούν σε κάθε μία από τις τιμές της στατιστικής συνάρτησης T_3 βάσει του αριθμού n των παρατηρήσεων του δείγματος από το οποίο αυτή έχει υπολογισθεί. Για παράδειγμα, στην πρώτη τιμή της T_3 που προέρχεται από ένα δείγμα 14 παρατηρήσεων θα αντιστοιχίσουμε τις τιμές των b_{14}, c_{14}, d_{14}. Με βάση όλες αυτές τις στήλες, από την επιλογή **Transform**, **Compute** και με **expression**

b+c * LN((t-d)/(1-t))

δημιουργούμε την στήλη **g** που περιέχει τις τιμές των στατιστικών συναρτήσεων G που αντιστοιχούν στις στατιστικές συναρτήσεις T_3. Στην συνέχεια, αθροίζουμε αυτή την στήλη και διαιρούμε το

προκύπτον άθροισμα με την τετραγωνική ρίζα του αριθμού των στοιχείων της. Αυτό γίνεται με υπολογισμό του αθροίσματος από την επιλογή **Summarize**, **Descriptive Statistics**, **Frequencies** και διαίρεσή του με την τετραγωνικήρίζα του 16. Η τιμή που προκύπτει είναι η τιμή της ελεγχοσυνάρτησης Z που χρησιμοποιούμε για τον έλεγχο.
Στο παράδειγμά μας, προκύπτει Z=-2.0525. Η τιμή του κρίσιμου επιπέδου που αντιστοιχεί σ΄αυτήν έχει υπολογισθεί ίση με 0.02 (αναλυτική λύση). Άρα τα δεδομένα δεν είναι εύλογο να θεωρηθούν ως προερχόμενα από Λογαριθμοκανονική κατανομή.

Ολοκληρώνοντας το κεφάλαιο αυτό, θα πρέπει να τονισθεί ότι, τα πραγματικά δεδομένα σχεδόν ποτέ δεν κατανέμονται ακριβώς σύμφωνα με κάποια συγκεκριμένη κατανομή. Ομως, συχνά, τα δεδομένα είναι *αρκετά κοντά* με κάποια κατανομή, ώστε να είναι δυνατόν να ληφθούν ικανοποιητικά ακριβή αποτελέσματα από την υπόθεση ότι η μηδενική υπόθεση είναι ορθή. Ενας έλεγχος καλής προσαρμογής, δηλαδή, είναι ένας τρόπος να συμπεράνουμε το κατά πόσον η *συμφωνία* των δεδομένων με την *μηδενική* κατανομή είναι πολύ *καλή*.

4.4. ΕΛΕΓΧΟΙ ΤΥΧΑΙΟΤΗΤΑΣ

Οι έλεγχοι καλής προσαρμογής, στην πραγματικότητα, αποτελούν εργαλεία ανίχνευσης του τρόπου με τον οποίο άτομα, αντικείμενα ή μετρήσεις κατανέμονται σε διάφορες κατηγορίες (κλάσεις). Με αυτή την έννοια μπορούν, επίσης, να χρησιμοποιηθούν ως *έλεγχοι τυχαιότητας*, δηλαδή ως έλεγχοι υποθέσεων της μορφής

H_0: Οι τιμές μιας ακολουθίας τιμών εμφανίζονται με τυχαία σειρά

έναντι εναλλακτικών υποθέσεων μη τυχαιότητας.

Για παράδειγμα, για να ελεγχθεί η υπόθεση ότι μία ακολουθία n αριθμών αποτελεί ακολουθία n τυχαίων αριθμών, θα μπορούσε να χρησιμοποιηθεί ο έλεγχος χ^2 για να ελεγχθεί η υπόθεση ότι η κατανομή του αριθμού των περιττών (ή των άρτιων) αριθμών που περιέχονται σε ισομήκη μη επικαλυπτόμενα τμήματα της ακολουθίας είναι η διωνυμική με παραμέτρους n και p=1/2. Αυτό είναι άμεση συνέπεια του ότι, αν η ακολουθία είναι ακολουθία τυχαίων αριθμών και χωρισθεί σε μη επικαλυπτόμενες ισομήκεις υποακολουθίες, η πιθανότητα ενός τυχόντος όρου της να είναι περιττός (ή άρτιος) είναι 5/10=1/2.

Αν η μηδενική υπόθεση δεν αληθεύει, κάποιες μορφές μη τυχαιότητας θα «*ανιχνεύονται*» μέσω μη ικανοποιητικής προσαρμογής της διωνυμικής $(n, p=\frac{1}{2})$ στα δεδομένα, ενώ, άλλες μορφές μη τυχαιότητας δεν θα ανιχνεύονται καθόλου. Στην πραγματικότητα, κανείς από τους ελέγχους τυχαιότητας δεν είναι συνεπής έναντι όλων των μορφών μη τυχαιότητας. Επιπλέον, οι έλεγχοι τυχαιότητας που αποτελούν παραλλαγές ελέγχων καλής προσαρμογής, βασίζονται στην συχνότητα εμφάνισης των όρων μιας ακολουθίας ή των αποτελεσμάτων ενός πειράματος (όπως,

εξάλλου συμβαίνει με τους ελέγχους καλής προσαρμογής), η οποία δεν παρέχει ουδεμία πληροφορία για την σειρά εμφάνισης των ενδεχομένων. Για παράδειγμα, αν σε 20 ρίψεις ενός νομίσματος, τα διαδοχικά αποτελέσματα ήταν

Κ Γ Κ Γ Κ Γ Κ Γ Κ Γ Κ Γ Κ Γ Κ Γ Κ Γ

και η τυχαιότητα εξεταζόταν με βάση την συχνότητα εμφάνισης των αποτελεσμάτων Κ και Γ, π.χ. με την χρήση του ελέγχου χ^2 ή και του διωνυμικού ελέγχου, δεν θα είχαμε λόγο να αμφισβητήσουμε την αμεροληψία του νομίσματος. Η έλλειψη τυχαιότητας, δηλαδή, αν και είναι καταφανής, δεν αποκαλύπτεται από ελέγχους που βασίζονται στην συχνότητα εμφάνισης των αποτελεσμάτων ενός πειράματος (ενδεχομένων). Η ανίχνευση αυτής της μορφής μη τυχαιότητας είναι δυνατή μόνο με ελέγχους που βασίζονται στην σειρά με την οποία έγινε η λήψη ή καταγραφή των παρατηρήσεων. Στην κατηγορία αυτών των ελέγχων περιλαμβάνονται παραλλαγές ελέγχων για ύπαρξη τάσης σε μία ακολουθία παρατηρήσεων $X_1, X_2, \ldots, X_n$, όπως ο έλεγχος των Cox και Stuart ή έλεγχοι βασισμένοι στον βαθμό συσχέτισης μεταξύ των τιμών των X_i και της σειράς εμφάνισής τους. Η χρήση του ελέγχου Cox – Stuart για τον έλεγχο υποθέσεων τυχαιότητας έχει ήδη περιγραφεί στο κεφάλαιο 2. Επίσης, η περίπτωση ελέγχων βασισμένων στον βαθμό συσχέτισης έχει εξετασθεί στο κεφάλαιο 3, στο πλαίσιο ελέγχων βασισμένων στις τάξεις μεγέθους των παρατηρήσεων.

Στην συνέχεια, παρουσιάζονται δύο έλεγχοι τυχαιότητας που βασίζονται στην χρονική σειρά εμφάνισης των τιμών των όρων μιας ακολουθίας παρατηρήσεων $X_1, X_2, \ldots, X_n$.

4.4.1 Ελεγχος Σημείων Πρώτων Διαφορών των Moore και Wallis

Ο έλεγχος τυχαιότητας της ενότητας αυτής χρησιμοποιείται συνήθως, στις περιπτώσεις που επιθυμούμε να ελέγξουμε κατά πόσο οι τιμές μιας ακολουθίας παρατηρήσεων θεωρούμενες με την χρονική σειρά εμφάνισής τους, έστω X_1, X_2, …, X_n αποτελούν μια ακολουθία τυχαίων παρατηρήσεων πάνω στην μεταβλητή Χ. (Οποτεδήποτε, δηλαδή, επιθυμούμε να ελέγξουμε κατά πόσο οι παρατηρήσεις X_1, X_2, …, X_n έχουν προκύψει κατά τύχη με αυτή τη σειρά εμφάνισης).

Η προς έλεγχο μηδενική υπόθεση διατυπώνεται ως εξής:

H_0: η ακολουθία των παρατηρήσεων, με την χρονική σειρά εμφάνισής τους, είναι τυχαία.

Ο έλεγχος της υπόθεσης αυτής στηρίζεται στην εξής θεωρία:

Ας ονομάσουμε *«+» ενδεχόμενο* το ενδεχόμενο $\{X_i<X_{i+1}\}$, i = 1, 2, …, n , *«–» ενδεχόμενο* το ενδεχόμενο $\{X_i>X_{i+1}\}$ και *«0» ενδεχόμενο* το ενδεχόμενο $\{X_i = X_{i+1}\}$. Αν η σειρά των τιμών X_1, X_2, …, X_n δεν είναι τυχαία, ο αριθμός Τ των «+» ενδεχομένων θα πρέπει να είναι πολύ μεγάλος ή πολύ μικρός. Επομένως, η κατάλληλη ελεγχοσυνάρτηση για τον έλεγχο της υπόθεσης H_0 έναντι της εναλλακτικής ότι η ακολουθία των παρατηρήσεων με την χρονική σειρά εμφάνισής τους δεν είναι τυχαία, είναι η

Τ = αριθμός των «+» ενδεχομένων.

Τα ποσοστιαία σημεία της κατανομής της Τ έχουν υπολογισθεί από τους Moore και Wallis (1943).

Μπορεί να δειχθεί ότι η ακριβής κατανομή της τυχαίας μεταβλητής T προσεγγίζεται από την κανονική κατανομή με μέση τιμή $\mu = E(T) = (n-1) / 2$ και διασπορά $\sigma^2 = V(T) = (n+1) / 12$. Δηλαδή,

$$T \sim N\left(\frac{n-1}{2}, \frac{n+1}{12}\right).$$

Επομένως, η μηδενική υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας α αν

$$\left| \frac{T - \frac{n-1}{2}}{\sqrt{\frac{n+1}{12}}} \right| > z_{1-\alpha/2},$$

όπου $z_{1-\alpha/2}$ είναι το (α-α/2)-ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής (πίνακας 2 του παραρτήματος).

Παράδειγμα 4.4.1: Εστω ότι οι τριμηνιαίες πωλήσεις ενός αναψυκτικού (σε εκατοντάδες χιλιάδες δραχμές) κατά τη διάρκεια μιας τετραετίας διαμορφώθηκαν ως εξής:

Έτος	31 Μαρτίου	30 Ιουνίου	30 Σεπτ.	31 Δεκεμ.
1976	41234	50225	54462	41295
1977	44555	59893	68958	53344
1978	54684	74238	79430	62656
1979	62691	79865	83637	65569

Ας υποθέσουμε ότι ενδιαφερόμαστε να ελέγξουμε κατά πόσο οι τιμές της ακολουθίας των παρατηρήσεων των όρων της ακολουθίας με την χρονική σειρά εμφάνισής τους, είναι μια τυχαία ακολουθία.

Για τον σκοπό αυτό, κατασκευάζουμε τον εξής πίνακα:

Ετος	Τρίμηνο i	X_i	X_{i+1}	Πρόσημο
1976	1	41234	50225	+
	2	50225	54462	+
	3	54462	41295	−
	4	41295	44555	+
1977	5	44555	58893	+
	6	58893	68958	+
	7	68958	53344	−
	8	53344	54684	+
1978	9	54684	74238	+
	10	74238	79430	+
	11	79430	62656	−
	12	62656	62691	+
1979	13	62691	79865	+
	14	79865	83637	+
	15	83637	65569	−
	16	65569		

Από τα στοιχεία του πίνακα, προκύπτει ότι η τιμή της ελεγχοσυνάρτησης T = αριθμός των «+» είναι τ = 11. Επιπλέον, επειδή n = 16, έχουμε, για την μέση τιμή και την διασπορά της T,

$$E(T) = \frac{n-1}{2} = 7.5 \text{ και } V(T) = \frac{n+1}{12} = \frac{17}{12} = 1.417 .$$

Η κρίσιμη περιοχή του ελέγχου ορίζεται, επομένως, από την ανισότητα

$$\left| \frac{T - 7.5}{1.417} \right| > z_{0.975} \equiv 1.96 .$$

Η τυποποιημένη τιμή της T είναι

$$\frac{11 - 7.5}{17/12} = 5.46 > z_{0.975} = 1.96 .$$

Επομένως, σε επίπεδο σημαντικότητας 0.05 δεν υπάρχουν ενδείξεις ότι η παραπάνω ακολουθία παρατηρήσεων αποτελεί ακολουθία τυχαίων παρατηρήσεων.

Λύση με το MINITAB: Για να εφαρμόσουμε τον έλεγχο των Moore και Wallis καταχωρούμε σε μία στήλη (έστω **x1**) όλες τις παρατηρήσεις εκτός από την τελευταία και σε μία άλλη στήλη (έστω **x2**) όλες τις παρατηρήσεις εκτός από την πρώτη. Η καταχώρηση των παρατηρήσεων γίνεται με τη σειρά που έχουν ληφθεί. Κατόπιν επιλέγουμε **Calc**, **Calculator** και δημιουργούμε μία μεταβλητή (έστω **s**) με expression x2>x1. Αυτό δρα ως λογική συνάρτηση και στη νέα μεταβλητή καταχωρείται η τιμή 1 όταν η συνθήκη ισχύει και η τιμή 0 όταν δεν ισχύει. Πράγματι, η κορυφή του φύλλου δεδομένων δείχνει ως εξής:

x1	x2	s
41234	50225	1
50225	54462	1
54462	41295	0
41295	44555	1
44555	58893	1
58893	68958	1

Η τιμή του στατιστικού T είναι το άθροισμα των παρατηρήσεων της s το οποίο προκύπτει ίσο με 11. Η μέση τιμή και διασπορά του T κάτω από την H_0 είναι 7.5 και 1.417 αντίστοιχα άρα το Z στατιστικό είναι 5.46. Αυτό είναι μέσα στην περιοχή απόρριψης (είναι μεγαλύτερο του 1.96) άρα η υπόθεση της τυχαιότητας των παρατηρήσεων απορρίπτεται.

Λύση με το SPSS: Ο έλεγχος τυχαιότητας των Moore-Wallis δεν παρέχεται αυτούσιος από το SPSS. Μπορούμε όμως εύκολα να υπολογίσουμε την τιμή του στατιστικού T με τον ακόλουθο τρόπο. Σε μία μεταβλητή (έστω **x1**) καταχωρούμε όλες τις παρατηρήσεις εκτός από την τελευταία. Σε μία άλλη μεταβλητή (έστω **x2**) καταχωρούμε όλες τις παρατηρήσεις εκτός από την πρώτη. Η καταχώρηση των παρατηρήσεων γίνεται με τη σειρά που αυτές έχουν ληφθεί.

Κατόπιν επιλέγουμε **Transform**, **Compute** και ζητάμε να δημιουργηθεί μία νέα μεταβλητή (έστω **s**) με expression x2>x1. Αυτή δρα ως λογική συνάρτηση και η s παίρνει την τιμή 1 όταν η συνθήκη ικανοποιείται και την τιμή 0 όταν δεν ικανοποιείται. Η κορυφή του φύλλου δεδομένων δείχνει τώρα ως εξής:

x1	x2	s
41234	50225	1.00
50225	54462	1.00
54462	41295	.00
41295	44555	1.00
44555	58893	1.00
58893	68958	1.00
68958	53344	.00

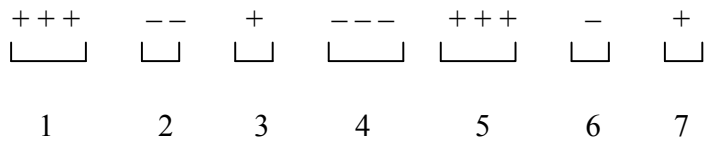
Η τιμή του T δεν είναι παρά το άθροισμα των τιμών της s. Αυτό προκύπτει ίσο με 11. Η μέση τιμή και η διακύμανση του T κάτω από την H_0 είναι 7.5 και 1.417 αντίστοιχα. Συνεπώς η τιμή του Z στατιστικού είναι 5.46, δηλαδή αυτό βρίσκεται στην περιοχή απόρριψης. Άρα οι παρατηρήσεις δεν μπορούν να θεωρηθούν τυχαίες.

4.4.2 Ο Ελεγχος των Ροών
(Runs Test)

Ο έλεγχος της ενότητας αυτής βασίζεται στον αριθμό των *ροών* που εμφανίζονται σε ένα δείγμα παρατηρήσεων θεωρούμενων με την σειρά καταγραφής τους.

Ως *ροή* ορίζεται μία διαδοχή πανόμοιων συμβόλων, της οποίας προηγούνται ή/και έπονται διαφορετικά σύμβολα.

Για παράδειγμα, η ακολουθία των συμβόλων + και – που προέκυψαν με την εξής σειρά,



εμφανίζει 7 ροές, 4 του συμβόλου + και 3 του συμβόλου –.

Ο συνολικός αριθμός των ροών σε ένα δείγμα οποιουδήποτε μεγέθους είναι ενδεικτικός της τυχαιότητας η μη του δείγματος. Ένας πολύ μικρός αριθμός ροών αποτελεί ένδειξη ύπαρξης τάσης στους όρους της ακολουθίας παρατηρήσεων, δηλαδή, εξάρτησής τους από την (χρονική) σειρά καταγραφής τους ή εξάρτησης άλλης μορφής. Ένας πολύ μεγάλος αριθμός ροών, από την άλλη μεριά, είναι ενδεικτικός κάποιας μορφής συστηματικής κυκλικής επίδρασης πάνω στις τιμές των όρων της ακολουθίας.

Ας υποθέσουμε ότι, σε μία ακολουθία n στοιχείων, κάθε ένα από τα οποία ανήκει σε ένα είδος,

n_1 = αριθμός στοιχείων τύπου 1

n_2 = αριθμός στοιχείων τύπου 2.

Οι υποθέσεις που μας ενδιαφέρει να ελέγξουμε έχουν την μορφή:

H_0: Τα δύο είδη συμβόλων εμφανίζονται με τυχαία σειρά

H_1: Η σειρά εμφάνισης των δύο συμβόλων δεν είναι τυχαία.

Σύμφωνα με όσα αναπτύχθηκαν παραπάνω, η κατάλληλη ελεγχοσυνάρτηση είναι η στατιστική συνάρτηση

T = αριθμός των ροών.

Οι πολύ μεγάλες ή οι πολύ μικρές τιμές της T αποτελούν, προφανώς, ένδειξη έλλειψης τυχαιότητας. Κατά συνέπεια, η κρίσιμη περιοχή μεγέθους α του ελέγχου των παραπάνω υποθέσεων ορίζεται από τις ανισότητες $T \le w_{\alpha/2}$ και $T \ge 1 - w_{\alpha/2}$, όπου w_p συμβολίζει το p-ποσοστιαίο σημείο της κατανομής της στατιστικής συνάρτησης T. Ποσοστιαία σημεία της κατανομής αυτής για διάφορες τιμές των n_1 και n_2 δίνονται για την περίπτωση αμφίπλευρων ελέγχων μεγέθους α=0.05 ή μονόπλευρων ελέγχων μεγέθους α=0.025 (από τον πίνακα 25 του παραρτήματος).

Αν η κατεύθυνση της απόκλισης από την τυχαιότητα προσδιορίζεται ή προκύπτει από την μορφή της εναλλακτικής υπόθεσης (μονόπλευρη εναλλακτική υπόθεση), η κρίσιμη περιοχή μεγέθους α του ελέγχου ορίζεται από την ανισότητα

$$T \le w_{\alpha}$$

ή από την ανισότητα

$$T \ge 1 - w_{1-\alpha}$$

ανάλογα με το εάν η μορφή της εναλλακτικής υπόθεσης υπαινίσσεται μικρό ή μεγάλο αριθμό ροών, αντίστοιχα.

Στην περίπτωση μεγάλων δειγμάτων (n_1>20 ή n_2>20), η κατανομή της στατιστικής συνάρτησης T προσεγγίζεται ικανοποιητικά από την κανονική κατανομή με μέση τιμή

$$\frac{2n_1n_2}{n_1 + n_2} + 1$$

και διασπορά

$$\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}.$$

Επομένως, για τον έλεγχο υποθέσεων τυχαιότητας, στην περίπτωση αυτή, μπορεί να χρησιμοποιηθεί ως ελεγχοσυνάρτηση, η τυποποιημένη μορφή της στατιστικής συνάρτησης T, δηλαδή, η στατιστική συνάρτηση

$$T_1=\frac{T-\left(\frac{2n_1n_2}{n_1+n_2}+1\right)}{\sqrt{\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}}.$$

Η κρίσιμη περιοχή μεγέθους α του ελέγχου, θα ορίζεται, επομένως, από τις ανισότητες

$$|T_1|>z_{1-\alpha/2}, \qquad T_1<-z_{1-\alpha} \qquad \text{ή} \qquad T_1>z_{1-\alpha}$$

ανάλογα με την μορφή της εναλλακτικής υπόθεσης (αμφίπλευρη, μονόπλευρη και υπαινισσόμενη μικρό αριθμό ροών ή μονόπλευρη και υπαινισσόμενη μεγάλο αριθμό ροών αντίστοιχα).

Παράδειγμα 4.4.2: Στο παράδειγμα της ακολουθίας των πωλήσεων του αναψυκτικού, η ύπαρξη τυχαιότητας στην σειρά εμφάνισης των παρατηρήσεων μπορεί, εναλλακτικά, να ελεγχθεί με τον έλεγχο των ροών.

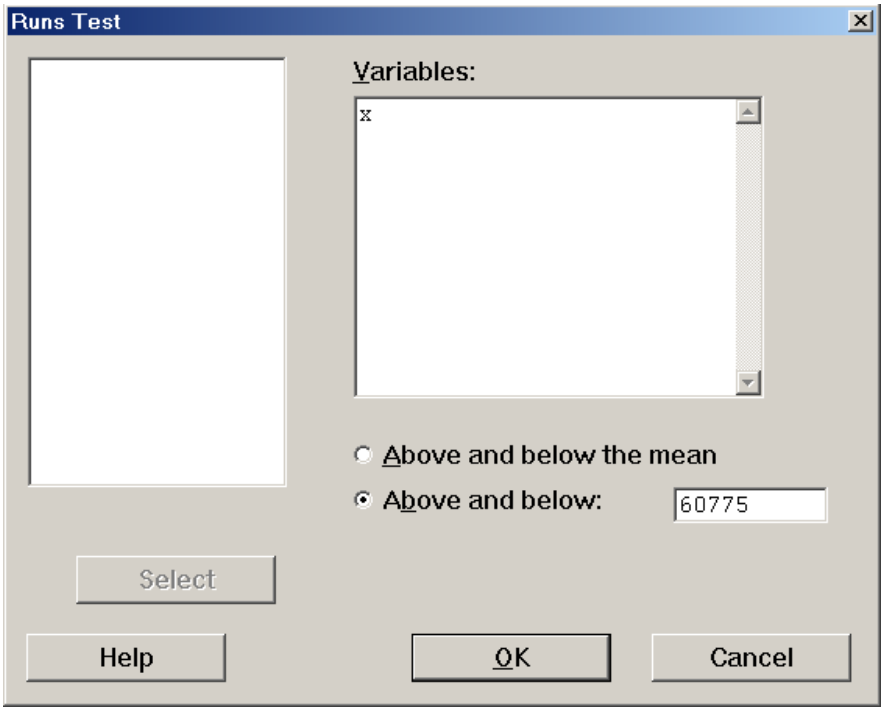
Σύμφωνα με τα παραπάνω, η μορφή του ελέγχου των ροών απαιτεί την διάκριση (ταξινόμηση) των παρατηρήσεων σε δύο διαφορετικά είδη και, εν συνεχεία, την καταγραφή των διαδοχικών εμφανίσεων κάθε ενός από αυτά τα είδη παρατηρήσεων. Ένας τρόπος με τον οποίο οι παρατηρήσεις μπορούν να διακριθούν σε δύο διαφορετικούς τύπους είναι

μέσω της σύγκρισής τους με κάποιο μέτρο θέσης του δείγματος, όπως είναι η διάμεσός του. Ετσι, σε κάθε μία από τις παρατηρήσεις, μπορούμε να αντιστοιχίσουμε το σύμβολο + ή το σύμβολο –, ανάλογα με το εάν η παρατήρηση είναι μεγαλύτερη ή μικρότερη από την διάμεσο του δείγματος. Προχωρώντας με αυτόν τον τρόπο, κατασκευάζουμε τον εξής πίνακα:

Ετος	Τρίμηνο i	X_i	Είδος	
1976	1	41234	–	
	2	50225	–	
	3	54462	–	
	4	41295	–	1
1977	5	44555	–	
	6	58893	–	
	7	68958	+	2
	8	53344	–	3
1978	9	54684	–	
	10	74238	+	
	11	79430	+	
	12	62656	+	
1979	13	62691	+	4
	14	79865	+	
	15	83637	+	
	16	65569	+	
	Διάμεσος	60972		

Από την τελευταία στήλη του πίνακα, προκύπτει ότι το δείγμα των παρατηρήσεων, με την σειρά εμφάνισής τους, παρουσιάζει 4 ροές, ενώ ο αριθμός των παρατηρήσεων που ανήκουν στο είδος + είναι $n_1 = 8$ και αυτών που ανήκουν στο είδος – είναι $n_2 = 8$. Από τον πίνακα 25 του παραρτήματος, προκύπτει ότι η κρίσιμη περιοχή μεγέθους 0.05 του ελέγχου ορίζεται από τις ανισότητες $T \le w_{0.025} = 4$ και $T \ge w_{0.975} = 14$. Η παρατηρούμενη τιμή της T είναι τ = 4. Η τιμή αυτή βρίσκεται, εντός της κρίσιμης περιοχής και, κατά συνέπεια, η υπόθεση ότι η σειρά εμφάνισης των παρατηρήσεων είναι τυχαία απορρίπτεται σε επίπεδο σημαντικότητας α = 5%.

Λύση με το MINITAB: Το MINITAB δίνει την δυνατότητα διεξαγωγής του ελέγχου των ροών, αλλά δημιουργεί τις ροές με βάση την μέση τιμή και όχι την διάμεσο του δείγματος. Για τον λόγο αυτό, καταχωρίζοντας το δείγμα σε μία στήλη (έστω **x**), υπολογίζουμε την διάμεσό του με **Stat, Basic Statistics, Display Descriptive Statistics**. (Η διάμεσος προκύπτει ίση με 60775). Κατόπιν, επιλέγουμε **Stat, Nonparametrics, Runs Test** και οδηγούμεθα στο εξής πλαίσιο διαλόγου:



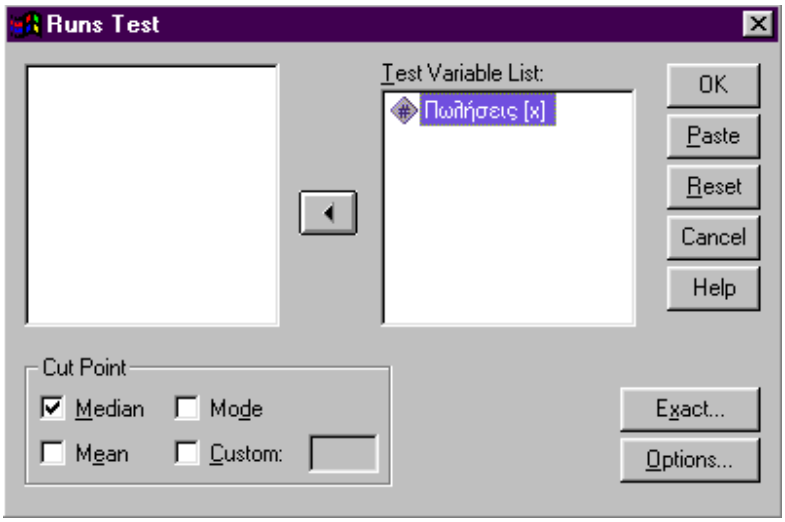
Στο πεδίο **Variables**, δηλώνουμε την μεταβλητή ή τις μεταβλητές που περιέχουν τα δείγματα των οποίων θέλουμε να ελέγξουμε την τυχαιότητα (**x**, στην περίπτωση μας). Στο πεδίο **Above and below**, δηλώνουμε την τιμή βάσει της οποίας θα υπολογισθούν οι ροές. Το αποτέλεσμα που δίνει το MINITAB είναι:

```
K = 60775.0000

  The observed number of runs =   4
  The expected number of runs =   9.0000
   8 Observations above K    8 below
* N Small -- The following approximation may be invalid
             The test is significant at  0.0097
```

Παρατηρούμε ότι έχουμε 4 ροές και η αντίστοιχη τιμή του κρίσιμου επιπέδου είναι μικρότερη από 0.01. Συνεπώς, η υπόθεση της τυχαιότητας

δεν μπορεί να θεωρηθεί εύλογη. Το πρόγραμμα δίνει επί πλέον την προειδοποίηση ότι η προσέγγιση που χρησιμοποίησε για να υπολογίσει το κρίσιμο επίπεδο ίσως δεν είναι ορθή γιατί το μέγεθος του δείγματος είναι μικρό.

Λύση με το SPSS: Σε αντίθεση με το MINITAB, το SPSS δίνει την δυνατότητα διεξαγωγής του ελέγχου των ροών με την χρησιμοποίηση της διαμέσου για τον καθορισμό τους. Καταχωρίζουμε το δείγμα σε μία μεταβλητή (έστω **x**) και επιλέγουμε **Analyze, Nonparametric Tests, Runs** οδηγούμενοι στο εξής πλαίσιο διαλόγου:



Στο πεδίο **Test Variable List**, δηλώνουμε την μεταβλητή ή τις μεταβλητές που περιέχουν τα δείγματα, των οποίων θέλουμε να ελέγξουμε την τυχαιότητα (**x**, στην περίπτωσή μας). Στο πεδίο **Cut Point**, επιλέγουμε **Median** για να χρησιμοποιηθεί η διάμεσος για τον καθορισμό των ροών. Με πίεση του πλήκτρου **Exact**, οδηγούμεθα στο γνωστό πλαίσιο διαλόγου

από το οποίο ζητάμε να χρησιμοποιηθεί η ακριβής κατανομή της στατιστικής συνάρτησης ελέγχου.

Τα αποτελέσματα του ελέγχου είναι:

Runs Test

	Πωλήσεις
Test Value[a]	60774.50[b]
Cases < Test Value	8
Cases >= Test Value	8
Total Cases	16
Number of Runs	4
Z	-2.329
Asymp. Sig. (2-tailed)	.020
Exact Sig. (2-tailed)	.018
Point Probability	.008

a Median

b There are multiple modes. The mode with the largest data value is used.

Παρατηρούμε ότι έχουμε 4 ροές (πεδίο **Number of Runs**) και ότι οι τιμές του αντίστοιχου ασυμπτωτικού και του ακριβούς κρίσιμου επιπέδου δεν διαφέρουν πολύ.

ΑΣΚΗΣΕΙΣ

1. Ο υπεύθυνος ενός κατασκευαστικού προγράμματος θέλει να ελέγξει την υπόθεση ότι τα αυτοκίνητα που διέρχονται πάνω από μια γέφυρα διέρχονται με τυχαίο τρόπο. Για τον σκοπό αυτό, κατέγραψε τους χρόνους που μεσολαβούν μεταξύ διαδοχικών διελεύσεων αυτοκινήτων κατά την διάρκεια ενός πρωϊνού. Οι χρόνοι αυτοί σε πρώτα λεπτά δίνονται στον πίνακα που ακολουθεί. Να ελεγχθεί η μηδενική υπόθεση ότι οι χρόνοι που μεσολαβούν μεταξύ διαδοχικών διελεύσεων αυτοκινήτων ακολουθούν την εκθετική κατανομή.

Χρόνος διαδοχικών διελεύσεων αυτοκινήτων κατά τη διάρκεια ενός πρωινού		
3.6	6.2	12.7
14.2	38.0	3.8
10.8	6.1	10.1
22.1	4.2	4.6
1.4	3.3	8.2

(Οικ. Παν/μιο Αθηνών – Εξετ. Φεβρ. 2000)

2. Ας υποθέσουμε ότι, κατά την διάρκεια ενός συγκεκριμένου μήνα, συνέβησαν 20 τροχαία ατυχήματα κατά μήκος ενός συγκεκριμένου τμήματος κάποιου αυτοκινητόδρομου. Οι 19 αποστάσεις μεταξύ των σημείων στα οποία συνέβησαν τα ατυχήματα (σε μίλια) περιέχονται στον παρακάτω πίνακα.

0.3	6.1	4.3	3.3	1.9
4.8	0.3	1.2	0.8	10.3
1.2	0.1	10.0	1.6	27.6
12.0	14.2	19.7	15.5	

Αποτελούν τα παραπάνω στοιχεία ένδειξη ότι τα ατυχήματα κατανέμονται τυχαία κατά μήκος του αυτοκινητόδρομου; Ποια είναι το κρίσιμο επίπεδο, του ελέγχου αυτού, κατά προσέγγιση;

(Οικ. Παν/μιο Αθηνών – Εξετ. Φεβρ. 1998)

3. Ο ιδιοκτήτης ενός Video Club έχει παρατηρήσει ότι οι ενοικιάσεις ταινιών στο κατάστημά του είναι ιδιαίτερα αυξημένες κατά την Παρασκευή και το Σάββατο. Συγκεκριμένα, πιστεύει ότι αυτές τις δύο ημέρες νοικιάζει το 60% των ταινιών της εβδομάδας, τις Κυριακές νοικιάζει το 25%, ενώ, από την Δευτέρα μέχρι και την Πέμπτη το υπόλοιπο 15%. Προκειμένου να ελέγξει τον ισχυρισμό του, συνέλεξε ένα

τυχαίο δείγμα από 250 ενοικιάσεις ταινιών που πραγματοποιήθηκαν κατά το τελευταίο έτος και βρήκε ότι 160 από αυτές έγιναν Παρασκευή ή Σάββατο, 70 Κυριακή και μόλις 20 τις υπόλοιπες ημέρες της εβδομάδας. Να ελεγχθεί σε επίπεδο σημαντικότητας α=0.01 αν ο ισχυρισμός του ιδιοκτήτη είναι βάσιμος.

(Οικ. Παν/μιο Αθηνών – Εξετ. Σεπτ. 2000)

4. Τα στοιχεία του πίνακα που ακολουθεί αναφέρονται σε ένα τυχαίο δείγμα 30 υποψηφίων από το σύνολο των υποψηφίων στις πρόσφατες εξετάσεις του GRE. Κάθε υποψήφιος είχε δικαίωμα να δηλώσει ένα μάθημα από ένα σύνολο πέντε μαθημάτων για να εξεταστεί,. Με βάση τα στοιχεία αυτά, θα ήταν βάσιμη η υπόθεση ότι υπήρχαν ίσοι αριθμοί υποψηφίων στα πέντε μαθήματα;

Μάθημα	Αριθμός Υποψηφίων
1^o	5
2^o	7
3^o	8
4^o	5
5^o	5

(Οικ. Παν/μιο Αθηνών - Εξετ. Σεπτ. 1999)

5. Προκειμένου να ελεγχθεί αν ένα ζάρι είναι αμερόληπτο, κατεγράφη ο αριθμός των εμφανίσεων κάθε δυνατού αποτελέσματος σε 300 ρίψεις.

Αποτέλεσμα	Αριθμός εμφανίσεων	Αναμενόμενος αριθμός εμφανίσεων
1	43	*
2	*	*
3	54	*
4	49	*
5	55	*
6	48	*

Να συμπληρωθούν οι τιμές που λείπουν στον παραπάνω πίνακα. Ποιο συμπέρασμα θα ήταν εύλογο, σε επίπεδο σημαντικότητας α=0.05, αν η τιμή της ελεγχοσυνάρτησης T του ελέγχου χ^2 είναι 1.92;

α) $\chi^2_{5,0.05}$=1.1455. Άρα, το ζάρι δεν μπορεί να θεωρηθεί αμερόληπτο.

β) $\chi^2_{5,0.05}$=1.1455. Άρα, το ζάρι μπορεί να θεωρηθεί αμερόληπτο.

γ) $\chi^2_{5,0.95}$=11.0705. Άρα, το ζάρι δεν μπορεί να θεωρηθεί αμερόληπτο.

δ) $\chi^2_{5,0.95}$=11.0705. Άρα, το ζάρι μπορεί να θεωρηθεί αμερόληπτο.

(Οικ. Παν/μιο Αθηνών – Εξετ. Προόδου Δεκ. 2000)

6. Για να ελεγχθεί αν ένα τυχαίο δείγμα 50 διψήφιων αριθμών προέρχεται από μία κανονική κατανομή, εκτιμώνται, με βάση τα δεδομένα, η μέση τιμή και η τυπική της απόκλιση ως $\hat{\mu}$=55.2 και s=18.7, αντίστοιχα. Στην συνέχεια, τα δεδομένα ταξινομούνται σε κλάσεις, με αναμενόμενες συχνότητες όπως φαίνεται στον πίνακα που ακολουθεί.

Κλάση	1	2	3	4
O_i	12	18	15	5
E_i	10.5	19.5	15.5	4.5

Εφαρμόζοντας τον έλεγχο χ^2, προκύπτει ότι η τιμή της ελεγχοσυνάρτησης είναι 0.401. Τι θα ήταν εύλογο να συμπεράνουμε σε επίπεδο σημαντικότητας α=0.05;

α) $\chi^2_{1,0.95}$=3.8415. Άρα, τα δεδομένα μπορούν να θεωρηθούν ότι προέρχονται από κανονική κατανομή.

β) $\chi^2_{3,0.95}$=7.8147. Άρα, τα δεδομένα μπορούν να θεωρηθούν ότι προέρχονται από κανονική κατανομή.

γ) $\chi^2_{2,0.95}$=5.9915. Άρα, τα δεδομένα μπορούν να θεωρηθούν ότι προέρχονται από κανονική κατανομή.

δ) $\chi^2_{1,0.95}$=3.8415. Άρα, τα δεδομένα δεν μπορούν να θεωρηθούν ότι προέρχονται από κανονική κατανομή.

(Οικ. Παν/μιο Αθηνών – Εξετ. Προόδου Δεκ. 2000)

7. 400 παρατηρήσεις (O_i) λαμβάνονται με σκοπό να ελεγχθεί αν προέρχονται από μια κατανομή Poisson. Η παράμετρος λ εκτιμάται από τα δεδομένα και προκύπτει ίση με $\hat{\lambda}$=2.1. Στην συνέχεια, τα δεδομένα ταξινομούνται σε 7 κλάσεις, για τις οποίες υπολογίζονται οι πιθανότητες να ανήκει κάποια παρατήρηση σε αυτές καθώς και οι αναμενόμενες συχνότητες (E_i) (δοθέντος ότι ισχύει η κατανομή Poisson). Τα αποτελέσματα συνοψίζονται στον πίνακα που ακολουθεί.

Κλάση i	1	2	3	4	5	6	7
O_i	89	143	94	42	20	8	4
$P(X \in i)$	0.223	0.335	0.251	*	0.047	*	0.004
E_i	89.2	*	100.4	50	18.8	*	1.6

Συμπληρώστε τις πιθανότητες και τις αναμενόμενες συχνότητες που λείπουν. Εφαρμόζοντας τον έλεγχο χ^2, προκύπτει ότι η τιμή της ελεγχοσυνάρτησης είναι τ=15.1362. Τι θα ήταν εύλογο να συμπεράνουμε σε επίπεδο σημαντικότητας α=0.05;

α) $\chi^2_{6,0.95}$=12.5916. Άρα, τα δεδομένα προέρχονται από κατανομή Poisson.

β) $\chi^2_{6,0.95}$=12.5916. Άρα, τα δεδομένα δεν προέρχονται από κατανομή Poisson.

γ) $\chi^2_{5,0.95}$=11.0705. Άρα, τα δεδομένα προέρχονται από κατανομή Poisson.

δ) $\chi^2_{5,0.95}$=11.0705. Άρα, τα δεδομένα δεν προέρχονται από κατανομή Poisson.

(Οικ. Παν/μιο Αθηνών – Εξετ. Προόδου Δεκ. 2000)

8. Λαμβάνουμε 20 μετρήσεις (O_i) πάνω σε μια μεταβλητή με σκοπό να ελέγξουμε αν η κατανομή της είναι η εκθετική. Με βάση τα αποτελέσματα, εκτιμάται η παράμετρος μ της κατανομής ως $\hat{\mu}$=4 και τα δεδομένα ταξινομούνται σε τέσσερις ισοπίθανες κλάσεις, των οποίων υπολογίζουμε τις αναμενόμενες συχνότητες (E_i).

Κλάση i	1	2	3	4
Εύρος τιμών	$\leq \alpha$	α-β	β-γ	$\geq \gamma$
Ei	5	5	5	5
Oi	3	8	5	4

Εφαρμόζοντας τον έλεγχο χ^2, προκύπτει ότι η τιμή της ελεγχοσυνάρτησης είναι τ=2.8. Τι θα ήταν εύλογο να συμπεράνουμε σε επίπεδο σημαντικότητας α=0.05;

α) $\chi^2_{2,0.95}$=5.9915. Άρα, τα δεδομένα προέρχονται από εκθετική κατανομή.

β) $\chi^2_{2,0.95}$=5.9915. Άρα, τα δεδομένα δεν προέρχονται από εκθετική κατανομή.

γ) $\chi^2_{1,0.95}$=3.8415. Άρα, τα δεδομένα προέρχονται από εκθετική κατανομή.

δ) $\chi^2_{1,0.95}$=3.8415. Άρα, τα δεδομένα δεν προέρχονται από εκθετική κατανομή.

(Οικ. Παν/μιο Αθηνών – Εξετ. Προόδου Δεκ. 2000)

9. Προκειμένου να ελεγχθεί αν μία γεννήτρια τυχαίων αριθμών παράγει, όντως, τυχαίους αριθμούς κατεγράφη ο αριθμός των εμφανίσεων κάθε ψηφίου σε 100 μονοψήφιους αριθμούς που παρήγαγε η γεννήτρια αυτή.

Ψηφίο	Αριθμός εμφανίσεων	Αναμενόμενος αριθμός εμφανίσεων
0	6	*
1	16	*
2	6	*
3	13	*
4	16	*
5	6	*
6	4	*
7	11	*
8	13	*
9	9	*

Να συμπληρωθεί, στον παραπάνω πίνακα, ο αναμενόμενος αριθμός εμφανίσεων του κάθε ψηφίου, δοθέντος ότι η γεννήτρια, όντως, παράγει τυχαίους αριθμούς. Ποιο συμπέρασμα θα ήταν εύλογο για την γεννήτρια σε επίπεδο σημαντικότητας α=0.05, αν η τιμή της ελεγχοσυνάρτησης του ελέγχου χ^2 είναι 17.6;

α) Μία παρατηρούμενη συχνότητα είναι μικρότερη του 5 και, επομένως, πρέπει να συνενώσουμε την αντίστοιχη κλάση με μία γειτονική κλάση και να επαναλάβουμε τον έλεγχο.

β) Μόνο 10% των παρατηρούμενων συχνοτήτων είναι μικρότερες του 5 και, έτσι, μπορούμε να εφαρμόσουμε με τον έλεγχο.

γ) Πρέπει τουλάχιστον 3 κλάσεις στο συγκεκριμένο πρόβλημα να έχουν παρατηρούμενη συχνότητα μικρότερη του 5 για να προβούμε σε συνενώσεις κλάσεων.

δ) Στα επίπεδα σημαντικότητας α=0.01 και α=0.05, η γεννήτρια μπορεί να θεωρηθεί ότι παράγει τυχαίους αριθμούς.

ε) Μόνο σε επίπεδο σημαντικότητας α=0.01 η γεννήτρια μπορεί να θεωρηθεί ότι παράγει τυχαίους αριθμούς.

(Οικ. Παν/μιο Αθηνών – Εξετ. Δεκ. 2000)

10. Ο παρακάτω πίνακας περιέχει 37 παρατηρήσεις (O_i) ταξινομημένες σε 6 κλάσεις και τις αναμενόμενες συχνότητες (E_i) των κλάσεων κάτω από κάποια υποτεθείσα κατανομή. Οι αναμενόμενες συχνότητες υπολογίσθηκαν από την υποτεθείσα κατανομή, αφού πρώτα μία άγνωστη παράμετρος της τελευταίας εκτιμήθηκε από τα δεδομένα.

Κλάση	1	2	3	4	5	6
O_i	4	5	*	8	8	7
E_i	1.7	6	12.3	*	5.3	5.2

Συμπληρώστε τις συχνότητες που λείπουν. Αν διεξαγάγουμε έλεγχο χ^2 και προκύψει ότι η τιμή της ελεγχοσυνάρτησης είναι τ=13.836, τι θα ήταν εύλογο να συμπεράνουμε σε επίπεδο σημαντικότητας α=0.05;

α) Μία κλάση έχει αναμενόμενη συχνότητα $E_i<5$. Άρα, πρέπει να συνενωθεί με μία γειτονική κλάση πριν από την διεξαγωγή του ελέγχου.

β) $\chi^2_{5,0.95}=11.0705$. Άρα, τα δεδομένα μπορούν να θεωρηθούν ότι προέρχονται από την υποτεθείσα κατανομή.

γ) $\chi^2_{6,0.95}=12.5916$. Άρα, τα δεδομένα μπορούν να θεωρηθούν ότι προέρχονται από την υποτεθείσα κατανομή.

δ) $\chi^2_{4,0.95}$=9.4877. Άρα, τα δεδομένα δεν μπορούν να θεωρηθούν ότι προέρχονται από την υποτεθείσα κατανομή.

(Οικ. Παν/μιο Αθηνών – Εξετ. Προόδου Δεκ. 2000)

11. Ο Διευθυντής ενός υποκαταστήματος μιας Τράπεζας θέλει να ελέγξει την υπόθεση ότι οι πελάτες της Τράπεζας φθάνουν στο υποκατάστημα με τυχαίο τρόπο. Για το σκοπό αυτό κατέγραψε τους χρόνους που μεσολαβούν μεταξύ διαδοχικών αφίξεων πελατών κατά την διάρκεια ενός πρωϊνού. Οι χρόνοι αυτοί (σε πρώτα λεπτά της ώρας) δίνονται στον πίνακα που ακολουθεί:

3.6	6.2	12.7	14.2	38.0	3.8	10.8	6.1
10.1	22.1	4.2	4.6	1.4	3.3	8.2	

Οπως είναι γνωστό, μια σειρά γεγονότων είναι τυχαία αν οι χρόνοι που μεσολαβούν μεταξύ διαδοχικών γεγονότων ακολουθούν την εκθετική κατανομή με συνάρτηση πυκνότητας πιθανότητας

$$f(x)=\lambda e^{-\lambda x}, \quad \lambda>0 , x>0$$

Τί συμπερασματολογία θα μπορούσατε να κάνετε με βάση τα στοιχεία του παραπάνω πίνακα;

(Οικ. Παν/μιο Αθηνών – Εξετ. Φεβρ. 1998)

12. Ας υποτεθεί ότι κατά τη διάρκεια ενός Σαββατοκύριακου συνέβησαν 20 παραβάσεις του κώδικα οδικής κυκλοφορίας, κατά μήκος ενός συγκεκριμένου τμήματος του αυτοκινητόδρομου Αθηνών Πατρών. Οι 19 αποστάσεις μεταξύ των σημείων στα οποία σημειώθηκαν οι παραβάσεις

δίνονται στον πίνακα που ακολουθεί. Αποτελούν τα παρακάτω στοιχεία ένδειξη ότι οι παραβάσεις κατανέμονται τυχαία κατά μήκος του αυτοκινητόδρομου Αθηνών Πατρών;

Αποστάσεις μεταξύ των σημείων που σημειώθηκαν οι παραβάσεις (σε χιλιόμετρα)			
0.3	0.3	10.0	15.5
4.8	0.1	19.7	1.9
1.2	14.2	3.3	10.3
12.0	4.3	0.8	27.6
6.1	1.2	1.6	

(Οικ. Παν/μιο Αθηνών – Εξετ. Φεβρ. 2000)

13. Να εξηγήσετε τον λόγο για τον οποίο κατά τον υπολογισμό της τιμής της ελεγχοσυνάρτησης του ελέγχου Kolmogorov ($T=\sup_x |F(x)-S(x)|$) αρκεί να υπολογίσουμε τις διαφορές στα σημεία $F(x_i)-S(x_i)$ και $F(x_i)-S(x_{i-1})$ (όπου F είναι η θεωρητική συνάρτηση κατανομής και S είναι η εμπειρική συνάρτηση κατανομής).

(Οικ. Παν/μιο Αθηνών – Εξετ. Σεπτ. 2000)

14. Η ελεγχοσυνάρτηση, του ελέγχου Kolmogorov κατανέμεται κανονικά όταν οι παρατηρήσεις ακολουθούν την κανονική κατανομή.

α. Σωστό β. Λάθος

(Οικ. Παν/μιο Αθηνών – Εξετ. Σεπτ. 2000)

15. Ενα ζάρι ρίπτεται 300 φορές με αποτελέσματα που δίνονται στον πίνακα που ακολουθεί. Μπορεί να ισχυρισθεί κανείς ότι το ζάρι είναι αμερόληπτο;

Αποτελέσματα	*Συχνότητες*
1	43
2	*48*
3	*54*
4	*45*
5	*61*
6	*49*

(Οικ. Παν/μιο Αθηνών – Εξετ. Φεβρ. 2000)

16. Εστω ότι έχουν ληφθεί 50 παρατηρήσεις από μία κατανομή και επιθυμείται μέσω του ελέγχου χ^2 να ελεγχθεί αν αυτές προέρχονται από πληθυσμό που ακολουθεί την κατανομή Poisson. Αν οι παρατηρήσεις έχουν ταξινομηθεί σε 10 κατηγορίες, η κατανομή της ελεγχοσυνάρτησης θα είναι η:

α) χ^2 με 48 βαθμούς ελευθερίας

β) χ^2 με 49 βαθμούς ελευθερίας

γ) χ^2 με 8 βαθμούς ελευθερίας

δ) κανονική με μέσο λ διακύμανση λ

ε) Poisson με μέσο nλ

(Οικ. Παν/μιο Αθηνών – Εξετ. Σεπτ. 2000)

17. α) Προκειμένου να ελεγχθεί η καλή προσαρμογή μιας κατανομής σε κάποια δεδομένα, τα στοιχεία του δείγματος ταξινομούνται σε 6 κλάσεις. Για να υπολογισθούν οι αναμενόμενες συχνότητες των κλάσεων χρειάζεται να εκτιμηθεί η μία από τις δύο παραμέτρους της κατανομής. Έστω ότι η τιμή της ελεγχοσυνάρτησης Τ είναι ίση με 10.66. Ο πίνακας των συχνοτήτων έχει ως εξής:

Κλάση	O_i	E_i
1	3	6
2	3	6
3	4	6
4	5	6
5	11	6
6	10	6

Τότε, επειδή, $\chi^2_{\nu,p} = 11.0705$, ο ερευνητής θεωρεί ότι η μηδενική υπόθεση H_0 ότι τα δεδομένα προέρχονται από την υποτιθέμενη κατανομή δεν είναι εύλογη σε επίπεδο σημαντικότητας 0.05. Προσδιορίστε τις τιμές των παραμέτρων ν και p.

β) Στον πίνακα του ερωτήματος Ι, τρεις από τις παρατηρούμενες συχνότητες είναι μικρότερες του 5. Αυτό δημιουργεί πρόβλημα; Εξηγήστε.

i) Ναι ii) Όχι

(Οικ. Παν/μιο Αθηνών – Εξετ. Προόδου Δεκ. 2000)

18. Να εξετασθεί αν το παρακάτω τυχαίο δείγμα 20 παρατηρήσεων (διατεταγμένων κατά αύξουσα σειρά) μπορεί να θεωρηθεί ότι αποτελεί δείγμα παρατηρήσεων πάνω σε μία τυχαία μεταβλητή Χ, της οποίας η κατανομή είναι κανονική με μέση τιμή 30 και διασπορά 100.

16.7 17.4 18.1 18.2 18.8 19.3 22.4 22.5 24.0 24.7
25.9 27.0 35.1 35.8 36.5 37.6 39.8 42.1 43.2 46.2

Υπόδειξη: Να κατασκευασθούν 4 ισοπίθανες κλάσεις.

(Παν/μιο Κρήτης, Ηράκλειο – Εξετ. Προόδου Δεκ. 1984)

19. Να ελεγχθεί αν το τυχαίο δείγμα παρατηρήσεων που δίνεται παρακάτω προέρχεται από μία κατανομή Poisson με συνάρτηση πιθανότητας $P(X=x) = e^{-\lambda} \lambda^x/x!$, $\lambda>0$.

x	≤ 1	2	3	4	5	6	7	≥ 8
συχνότητα	4	9	11	7	8	9	1	3

(Παν/μιο Κρήτης, Ηράκλειο – Εξετ. Σεπτ. 1985)

20. Ενα απλό τυχαίο δείγμα εννέα φοιτητών έδειξε τις εξής αποδοχές από την μερική απασχόλησή τους κατά την διάρκεια του καλοκαιριού (σε δεκάδες χιλιάδες δραχμές)

30 12 21 42 60 24 39 9 27

Μπορεί να θεωρηθεί εύλογη η υπόθεση ότι ο πληθυσμός των «θερινών» εισοδημάτων είναι κανονικός;

21. Ενα απλό τυχαίο δείγμα επί του ύψους 8 γυναικών και των ανδρών τους έδειξε τα εξής αποτελέσματα (σε μέτρα).

Γυναίκα: 1.55 1.58 1.60 1.63 1.65 1.70 1.73 1.78

Άνδρας: 1.63 1.68 1.73 1.75 1.78 1.83 1.90 1.93

Με βάση τα παραπάνω στοιχεία θα μπορούσε να συμπεράνει ότι κοντές γυναίκες τείνουν να παντρεύονται κοντούς άνδρες;

(Οικ. Παν/μιο Αθηνών – Εξετ. Φεβρ. 1992)

22. Η απόδοση μιας δωδεκάμηνης επένδυσης σε 20 τυχαία επιλεγμένα μετοχικά κεφάλαια συνοψίζεται στον πίνακα που ακολουθεί:

9.1	5.0	7.3	7.4	5.5
8.6	7.0	4.3	4.7	8.0
4.0	8.5	6.4	6.1	5.8
9.5	5.2	6.7	8.3	9.2

α) Να ελεγχθεί η σύνθετη μηδενική υπόθεση ότι η απόδοση δωδεκάμηνων επενδύσεων σε μετοχικά κεφάλαια ακολουθεί την κανονική κατανομή, χρησιμοποιώντας τον έλεγχο του Lilliefors.
β) Να ελεγχθεί η υπόθεση αυτή χρησιμοποιώντας τον έλεγχο Shapiro-Wilk.

23. Ας υποθέσουμε ότι η βαθμολογία ενός τυχαίου δείγματος επιτυχόντων στις Γενικές Εξετάσεις κάποιας χρονιάς είναι όπως φαίνεται στον πίνακα που ακολουθεί.

481	620	642	515	740
562	395	615	596	618
525	584	540	580	598

α) Να ελεγχθεί η σύνθετη μηδενική υπόθεση ότι η βαθμολογία ακολουθεί την κανονική κατανομή χρησιμοποιώντας τον έλεγχο του Lilliefors.
β) Να ελεγχθεί η υπόθεση αυτή χρησιμοποιώντας τον έλεγχο Shapiro-Wilk.

24. Ο διευθυντής ενός πολυκαταστήματος θέλει να ελέγξει την υπόθεση ότι οι πελάτες φθάνουν στο πολυκατάστημα με τυχαίο τρόπο.

Για τον σκοπό αυτό, κατέγραψε τους χρόνους που μεσολαβούν μεταξύ διαδοχικών αφίξεων πελατών κατά την διάρκεια ενός πρωϊνού. Οι χρόνοι αυτοί (σε πρώτα λεπτά) δίνονται στον πίνακα που ακολουθεί.

3.6	6.2	12.7
14.2	38.0	3.8
10.8	6.1	10.1
22.1	4.2	4.6
1.4	3.3	8.2

Να ελεγχθεί η μηδενική υπόθεση ότι οι χρόνοι που μεσολαβούν μεταξύ διαδοχικών αφίξεων πελατών ακολουθούν την εκθετική κατανομή.

25. Ας υποθέσουμε ότι, κατά την διάρκεια ενός συγκεκριμένου μήνα, συνέβησαν 20 τροχαία ατυχήματα κατά μήκος ενός συγκεκριμένου τμήματος κάποιου αυτοκινητόδρομου. Οι 19 αποστάσεις μεταξύ των σημείων στα οποία συνέβησαν τα ατυχήματα (σε χιλιόμετρα) δίνονται στον πίνακα που ακολουθεί.

0.3	6.1	4.3	3.3	1.9
4.8	0.3	1.2	0.8	10.3
1.2	0.1	10.0	1.6	27.6
12.0	14.2	19.7	15.5	

Αποτελούν τα παραπάνω στοιχεία ένδειξη ότι τα ατυχήματα κατανέμονται τυχαία κατά μήκος του αυτοκινητόδρομου;

26. Δεδομένα που αναφέρονται στην ροή νερού (στην ποσότητα του νερού που περνά από ένα συγκεκριμένο σημείο ενός ποταμού) θεωρούνται πολύ συχνά ότι ακολουθούν τη λογαριθμοκανονική κατανομή. Για να ελεγχθεί η υπόθεση αυτή, μαζεύτηκαν στοιχεία από 8 ρυάκια και ποτάμια διαφόρων μεγεθών. Τα δεδομένα παριστάνουν μετρήσεις πάνω στον όγκο του νερού ανά μονάδα χρόνου (κυβικά πόδια ανά δευτερόλεπτο) που ελήφθησαν μία φορά ανά εβδομάδα για διαφορετικό αριθμό εβδομάδων. Οι λογάριθμοι των στοιχείων ελέγχθηκαν για κανονικότητα με τον έλεγχο των Shapiro-Wilk με τα εξής αποτελέσματα:

Ποτάμι	Αριθμός Εβδομάδων	T_3
1	8	0.972
2	10	0.858
3	6	0.875
4	14	0.840
5	9	0.966
6	10	0.924
7	14	0.881
8	12	0.868

Να εξετασθεί εάν ο συνδυασμός των παραπάνω αποτελεσμάτων παρέχει ενδείξεις ότι τα δεδομένα ροής του νερού τείνουν να ακολουθούν μία λογαριθμοκανονική κατανομή.

27. Η συνολική ετήσια βροχόπτωση μιας χώρας θεωρείται ότι ακολουθεί κανονική κατανομή. Για να ελεγχθεί η υπόθεση αυτή, επελέγησαν τυχαία 10 πόλεις από την επικράτεια της χώρας αυτής και κατεγράφη η μέση ετήσια βροχόπτωση για τα έτη για τα οποία υπήρχαν διαθέσιμα στοιχεία. Τα στοιχεία για την ετήσια βροχόπτωση αναλύθηκαν χρησιμοποιώντας τον έλεγχο Shapiro-Wilk με τα εξής αποτελέσματα:

Πόλη	Αριθμός Ετών	T_3
1	18	0.875
2	34	輶緑ㄣ洌
3	26	0.948
4	43	0.980
5	40	0.937
6	29	0.915
7	35	0.915
8	38	0.890
9	42	0.963
10	47	0.941

Θα μπορούσατε να πείτε αν ο συνδυασμός των αποτελεσμάτων αυτών αποτελεί ένδειξη ότι η ετήσια βροχόπτωση ακολουθεί στην χώρα αυτή κανονική κατανομή;

28. Ενας μετεωρολόγος διεξάγει ένα πείραμα για να προσδιορίσει αν τα επίπεδα υγρασίας, που κατεγράφησαν στις 12:00 το μεσημέρι για 20 διαδοχικές ημέρες τον Ιούλιο του 2000, κατανέμονται με τυχαίο τρόπο ως προς το εάν αυτά βρίσκονται πάνω ή κάτω από το μέσο επίπεδο υγρασίας

που κατεγράφη κατά την διάρκεια του μηνός Ιουλίου των ετών 1995-1999. Συμβολίζοντας με «+» τα επίπεδα υγρασίας που ήταν υπεράνω του μέσου επιπέδου υγρασίας του Ιουλίου και με «–» αυτά που ήταν κάτω από το μέσο επίπεδο υγρασίας του Ιουλίου, οι 20 συγκρίσεις έδωσαν τα εξής αποτελέσματα:

+ + + – – – + + – – + – + – + – – – + +

Αποτελούν τα δεδομένα ένδειξη ότι η χρονοσειρά των επιπέδων υγρασίας είναι τυχαία;

29. Τα γένη των 20 διαδοχικών ασθενών που εισάγονται στην αίθουσα πρώτων βοηθειών ενός τοπικού νοσοκομείου (Α = άνδρας, Γ = γυναίκα) είχαν ως εξής:

Γ Γ Γ Α Α Α Γ Γ Α Α Γ Α Γ Α Γ Α Α Α Γ Γ

Παρέχουν τα δεδομένα ένδειξη ότι η κατανομή του φύλου των εισαγομένων ασθενών στο εν λόγω νοσοκομείο δεν είναι τυχαία;