

ΚΕΦΑΛΑΙΟ 5

ΕΛΕΓΧΟΙ ΥΠΟΘΕΣΕΩΝ ΓΙΑ ΙΣΟΤΗΤΑ ΔΥΟ ΚΑΤΑΝΟΜΩΝ

Στο προηγούμενο κεφάλαιο εξετάσαμε διάφορες μορφές ελέγχου της υπόθεσης ότι ένα δείγμα παρατηρήσεων προέρχεται από κάποια συγκεκριμένη κατανομή. Στην περίπτωση που έχουμε δύο ανεξάρτητα δείγματα παρατηρήσεων, ενδέχεται να ενδιαφερόμαστε να ξέρουμε εάν κάθε ένα από αυτά μπορεί εύλογα να υποθεθεί ότι προέρχεται από την ίδια άγνωστη κατανομή. Παρόμοια είναι και η περίπτωση δύο δειγμάτων τα οποία προέρχονται από διαφορετικούς πληθυσμούς. Στην περίπτωση αυτή, ο ερευνητής ενδέχεται να ενδιαφέρεται να προσδιορίσει αν οι δύο συναρτήσεις κατανομής που σχετίζονται με τους δύο πληθυσμούς ταυτίζονται ή όχι. Ελεγχος, όπως ο έλεγχος της διαμέσου, ο έλεγχος Mann-Whitney ή ο παραμετρικός έλεγχος t ενδέχεται να είναι κατάλληλοι στις περιπτώσεις αυτές. Ομως, οι έλεγχοι αυτοί είναι ευαίσθητοι μόνο σε διαφορές μεταξύ των πληθυσμών οι οποίες αναφέρονται στις μέσες τιμές τους ή στις διαμέσους τους, αλλά δεν έχουν την δυνατότητα να ανιχνεύσουν διαφορές άλλου τύπου, όπως διαφορές στις διασπορές των πληθυσμών. Ένα από τα πλεονεκτήματα των ελέγχων οι οποίοι θα εξετασθούν στο κεφάλαιο αυτό είναι ότι είναι συνεπείς έναντι όλων των τύπων των διαφορών, οι οποίες ενδέχεται να υπάρχουν μεταξύ δύο συναρτήσεων κατανομών: Οι κατανομές μπορεί να έχουν την ίδια μέση τιμή, αλλά διαφορετικές διασπορές, η μία μπορεί να είναι ασύμμετρη και η άλλη συμμετρική ή μπορεί να έχουν ίσες μέσες τιμές και ίσες διασπορές, αλλά η μία να είναι συμμετρική και η άλλη ασύμμετρη κ.ο.κ.

Ο πρώτος έλεγχος ο οποίος εξετάζεται είναι γνωστός ως έλεγχος Smirnov, που εξετάστηκε από τον Smirnov το 1939. Πρόκειται για μία μορφή ελέγχου Kolmogorov προσαρμοσμένη για την περίπτωση δύο ανεξαρτήτων δειγμάτων και ονομάζεται μερικές φορές έλεγχος *Kolmogorov - Smirnov για δύο δείγματα*. Αντίστοιχα, ο έλεγχος Kolmogorov ονομάζεται συχνά *έλεγχος Kolmogorov-Smirnov βασισμένος σε ένα δείγμα*. Ο έλεγχος Smirnov διακρίνεται σε μονόπλευρο και αμφίπλευρο έλεγχο. Για την περίπτωση αμφίπλευρης εναλλακτικής, εξετάζεται και μια άλλη μορφή ελέγχου, ο έλεγχος Cramér-Von Mises για δύο δείγματα. Αν και η ελεγχοσυνάρτηση του ελέγχου αυτού είναι περισσότερο δύσκολο να υπολογισθεί σε σύγκριση με την ελεγχοσυνάρτηση του ελέγχου Smirnov, ο έλεγχος Cramér-Von Mises προτιμάται αρκετά συχνά, κυρίως γιατί φαίνεται ότι κάνει περισσότερο αποτελεσματική χρήση των δεδομένων. Στην πραγματικότητα, όμως, υπάρχει πολύ μικρή διαφορά στην ισχύ των δύο ελέγχων.

5.1 Ο ΕΛΕΓΧΟΣ SMIRNOV

Έστω δύο ανεξάρτητα τυχαία δείγματα X_1, X_2, \dots, X_n και Y_1, Y_2, \dots, Y_m n και m παρατηρήσεων πάνω στις τυχαίες μεταβλητές X και Y , αντίστοιχα. Έστω, επίσης, ότι $F_X(x)$, $x \in (-\infty, +\infty)$ και $F_Y(y)$, $y \in (-\infty, +\infty)$ είναι οι αντίστοιχες άγνωστες συναρτήσεις κατανομής των τυχαίων μεταβλητών X και Y . Οι υποθέσεις που ενδιαφερόμαστε να ελέγξουμε έχουν μία από τις εξής τρεις μορφές:

A. (Αμφίπλευρη εναλλακτική υπόθεση)

$H_0: F_X(x) = F_Y(x)$, για κάθε $x \in (-\infty, +\infty)$

$H_1: F_X(x) \neq F_Y(x)$, για τουλάχιστον ένα $x \in (-\infty, +\infty)$.

Η εναλλακτική υπόθεση της περίπτωσης αυτής διατυπώνεται συχνά ως εξής: "τα X τείνουν να είναι μικρότερα από τα Y ". Είναι, δηλαδή, η μορφή αυτής της εναλλακτικής υπόθεσης γενικότερη από κάποια εναλλακτική η οποία θα έλεγε ότι τα X και Y διαφέρουν μόνο κατά μία παράμετρο θέσης (μέση τιμή ή διάμεσο).

B. (Μονόπλευρη εναλλακτική υπόθεση)

$H_0: F_X(x) = F_Y(x)$, για κάθε $x \in (-\infty, +\infty)$

$H_1: F_X(x) > F_Y(x)$, για τουλάχιστον ένα $x \in (-\infty, +\infty)$.

Γ. (Μονόπλευρη εναλλακτική υπόθεση)

$H_0: F_X(x) = F_Y(x)$, για κάθε $x \in (-\infty, +\infty)$

$H_1: F_X(x) < F_Y(x)$, για τουλάχιστον ένα $x \in (-\infty, +\infty)$.

Η περίπτωση B είναι κατάλληλη για προβλήματα όπου έχει έννοια να ελέγξουμε αν τα X είναι μετατοπισμένα προς τα αριστερά των Y (αν τα X τείνουν να είναι μικρότερα από τα Y). Αντίστοιχα, η περίπτωση Γ είναι κατάλληλη για τα προβλήματα στα οποία επιθυμούμε να ελέγξουμε αν τα X είναι μετατοπισμένα προς τα δεξιά των Y (δηλαδή αν τα X τείνουν να είναι μεγαλύτερα από τα Y).

Όπως και στην περίπτωση του ελέγχου Kolmogorov, είναι φυσικό η στατιστική συνάρτηση για τον έλεγχο των παραπάνω υποθέσεων να βασίζεται στις εμπειρικές κατανομές των δύο δειγμάτων παρατηρήσεων.

Εστω $S_X(x)$ η εμπειρική συνάρτηση κατανομής βασισμένη στο τυχαίο δείγμα X_1, X_2, \dots, X_n και έστω $S_Y(y)$ η εμπειρική συνάρτηση κατανομής βασισμένη στο δεύτερο τυχαίο δείγμα $Y_1,$

Y_2, \dots, Y_m . Η στατιστική συνάρτηση ελέγχου ορίζεται διαφορετικά για κάθε ένα από τα τρία διαφορετικά σύνολα υποθέσεων:

A. (Αμφίπλευρος έλεγχος): Η φυσική επιλογή στατιστικής συνάρτησης είναι η μέγιστη κατακόρυφη απόσταση μεταξύ των δύο εμπειρικών συναρτήσεων κατανομής:

$$T_1 = \sup_x |S_X(x) - S_Y(x)|.$$

B. (Μονόπλευρος έλεγχος): Ως ελεγχοσυνάρτηση για τον έλεγχο των υποθέσεων της μορφής B, ορίζεται η μέγιστη κατακόρυφη απόσταση που επιτυγχάνεται από την $S_X(x)$ υπεράνω της $S_Y(x)$. Δηλαδή,

$$T_1^+ = \sup_x [S_X(x) - S_Y(x)].$$

Γ. (Μονόπλευρος έλεγχος): Για τον μονόπλευρο έλεγχο της περίπτωσης Γ, η ελεγχοσυνάρτηση ορίζεται ως η μέγιστη κατακόρυφη απόσταση που επιτυγχάνεται από την $S_Y(x)$ υπεράνω της $S_X(x)$. Δηλαδή,

$$T_1^- = \sup_x [S_Y(x) - S_X(x)].$$

Είναι προφανές ότι και στις τρεις περιπτώσεις εναλλακτικών υποθέσεων, μεγάλες τιμές της ελεγχοσυνάρτησης αποτελούν ένδειξη εναντίον της μηδενικής υπόθεσης. Επομένως, ο κανόνας απόφασης έχει την εξής μορφή:

Η μηδενική υπόθεση H_0 απορρίπτεται σε επίπεδο σημαντικότητας α , αν η τιμή της κατάλληλης στατιστικής συνάρτησης (T_1, T_1^+ ή T_1^-) υπερβαίνει το $(1-\alpha)$ -ποσοστιαίο σημείο της κατανομής της, όπως αυτό δίνεται από τον πίνακα 20 του παραρτήματος, για $n=m$ ή από τον πίνακα 21, για $n \neq m$. Προσεγγίσεις των ποσοστιαίων σημείων της κατανομής της στατιστικής συνάρτησης Smirnov δίνονται

στο τέλος του πίνακα για τιμές του μεγέθους του δείγματος που δεν καλύπτονται από αυτόν.

Παράδειγμα 5.1.1: Ένα τυχαίο δείγμα μεγέθους 9, X_1, X_2, \dots, X_9 επιλέγεται από ένα πληθυσμό. Ένα άλλο τυχαίο δείγμα μεγέθους 15, Y_1, Y_2, \dots, Y_{15} επιλέγεται ανεξάρτητα από το πρώτο δείγμα από έναν δεύτερο πληθυσμό. Η μηδενική υπόθεση είναι ότι οι δύο πληθυσμοί έχουν την ίδια συνάρτηση κατανομής. Δηλαδή, αν οι αντίστοιχες συναρτήσεις κατανομής συμβολίζονται με $F_X(x)$, $x \in (-\infty, +\infty)$ και $F_Y(y)$, $y \in (-\infty, +\infty)$, τότε η μηδενική υπόθεση μπορεί να γραφεί με την μορφή

$$H_0: F_X(x) = F_Y(x), \text{ για κάθε } x \in (-\infty, +\infty).$$

Η εναλλακτική υπόθεση μπορεί να διατυπωθεί με την μορφή

$$H_1: F_X(x) \neq F_Y(x), \text{ για τουλάχιστον μία τιμή του } x.$$

Για την ευχερέστερη διεξαγωγή του ελέγχου, τα δύο δείγματα διατάσσονται κατά αύξουσα σειρά μεγέθους, ως εάν αποτελούσαν ένα ενιαίο δείγμα. Ο πίνακας που ακολουθεί περιέχει τις τιμές των δύο αυτών δειγμάτων διατεταγμένες κατά αύξουσα σειρά μεγέθους μαζί με τις τιμές της διαφοράς $S_X(x) - S_Y(x)$ των εμπειρικών συναρτήσεων κατανομής.

X_i	Y_i	$S_X(x) - S_Y(x)$	X_i	Y_i	$S_X(x) - S_Y(x)$
	5.2	$0 - 1/15 = -1/15$		9.8	$5/9 - 8/15 = 1/45$
	5.7	$0 - 2/15 = -2/15$	9.9		$6/9 - 8/15 = 2/15$
	5.9	$0 - 3/15 = -1/5$	10.1		$7/9 - 8/15 = 11/45$
	6.5	$0 - 4/15 = -4/15$	10.6		$8/9 - 8/15 = 16/45$
	6.8	$0 - 5/15 = -1/3$		10.8	$8/9 - 9/15 = 13/45$
7.6		$1/9 - 5/15 = -2/9$	11.2		$1 - 9/15 = 2/5$
	8.2	$1/9 - 6/15 = -13/45$		11.3	$1 - 10/15 = 1/3$
8.4		$2/9 - 6/15 = -8/45$		11.5	$1 - 11/15 = 4/15$
8.6		$3/9 - 6/15 = -1/15$		12.3	$1 - 12/15 = 1/5$
8.7		$4/9 - 6/15 = 2/45$		12.5	$1 - 13/15 = 2/15$
	9.1	$4/9 - 7/15 = -1/45$		13.4	$1 - 14/15 = 1/15$
9.3		$5/9 - 7/15 = 4/45$		14.6	$1 - 1 = 0$

Η κατάλληλη ελεγχοσυνάρτηση για τον έλεγχο των παραπάνω υποθέσεων είναι η

$$T_1 = \sup_x |S_X(x) - S_Y(x)|.$$

Από τον παραπάνω πίνακα, προκύπτει ότι η μεγαλύτερη τιμή της στατιστικής συνάρτησης T_1 επιτυγχάνεται στην τιμή 11.2. Είναι δηλαδή

$$\begin{aligned} T_1 &= \sup_x |S_X(x) - S_Y(x)| \\ &= |S_X(11.2) - S_Y(11.2)| \\ &= 2/5 \\ &= 0.4. \end{aligned}$$

Η τιμή 0.4 της στατιστικής συνάρτησης T μπορούσε επίσης να προσδιορισθεί γραφικά με την κατασκευή της γραφικής παράστασης των εμπειρικών συναρτήσεων κατανομής των δύο δειγμάτων. Από τα γραφήματα των συναρτήσεων αυτών, εύκολα μπορεί να δει κανείς ότι η διαφορά $S_X(x) - S_Y(x)$ αλλάζει τιμή μόνο στις παρατηρούμενες τιμές των $X_i, i = 1, 2, \dots, 9$ ή των $Y_i, i = 1, 2, \dots, 15$. Αυτός είναι και ο λόγος για τον οποίο αρκεί να υπολογίσουμε την διαφορά $S_X(x) - S_Y(x)$ μόνο στις παρατηρηθείσες τιμές του δείγματος, όπως κάναμε στον παραπάνω πίνακα.

Από τον πίνακα 21 του παραρτήματος βλέπουμε ότι το 0.95-ποσοστιαίο σημείο της κατανομής της στατιστικής συνάρτησης T_1 , για τον αμφίπλευρο έλεγχο και για $n=9=N_1$ και $m=15=N_2$, είναι

$$w_{0.95} = 8/15.$$

Για τα δεδομένα του προβλήματός μας, η τιμή της στατιστικής συνάρτησης T_1 είναι

$$\tau_1 = 2/5 < w_{0.95}.$$

Επομένως, η μηδενική υπόθεση H_0 δεν απορρίπτεται σε επίπεδο σημαντικότητας $\alpha=0.05$. Από τον πίνακα, μπορεί εύκολα να διαπιστωθεί ότι το κρίσιμο επίπεδο $\hat{\alpha}$ είναι ελαφρώς μεγαλύτερο από 0.20.

Αν, για τα δεδομένα του προβλήματός μας ήταν περισσότερο κατάλληλος ένας μονόπλευρος έλεγχος αντί του αμφίπλευρου που εξετάσαμε, η στατιστική συνάρτηση για την περίπτωση B των υποθέσεων έχει τιμή

$$T_1^+ = \sup_x [S_X(x) - S_Y(x)] = 2/5 = 0.4$$

Αντίστοιχα, για την περίπτωση Γ των υποθέσεων, η κατάλληλη στατιστική συνάρτηση έχει τιμή

$$T_1^- = \sup_x [S_Y(x) - S_X(x)] = 1/3 = 0.333.$$

Δηλαδή, στην περίπτωση B, η τιμή της στατιστικής συνάρτησης T_1^+ είναι η μέγιστη κατακόρυφη απόσταση της $S_X(x)$ υπεράνω της $S_Y(y)$ και την αναζητούμε μεταξύ των θετικών διαφορών, ενώ στην περίπτωση Γ η τιμή της στατιστικής συνάρτησης T_1^- είναι η μέγιστη κατακόρυφη απόσταση της $S_X(x)$ κάτω από την $S_Y(y)$ και την αναζητούμε μεταξύ των αρνητικών διαφορών. Τα κρίσιμα επίπεδα και για τους δύο μονόπλευρους ελέγχους εύκολα διαπιστώνονται ότι είναι μεγαλύτερα από 0.10.

Λύση με το MINITAB: Το MINITAB δεν δίνει την δυνατότητα διεξαγωγής του ελέγχου Kolmogorov-Smirnov για δύο ανεξάρτητα δείγματα. Είναι όμως δυνατή η διεξαγωγή του με έμμεσο τρόπο. Αυτό γίνεται ως εξής: Σε μία στήλη (έστω **a**), καταχωρίζουμε και τα δύο δείγματα, ενώ, σε μία άλλη στήλη (έστω **d**), καταχωρίζουμε μία κωδική μεταβλητή, της οποίας οι τιμές δείχνουν από ποιο δείγμα προέρχεται κάθε παρατήρηση (π.χ. 0 αν η παρατήρηση προέρχεται από το δείγμα τιμών της X και 1 αν η παρατήρηση προέρχεται από αυτό της Y). Στην συνέχεια, διατάσσουμε τις τιμές της στήλης που περιέχει το συνενωμένο δείγμα κατά αύξουσα σειρά μεγέθους. Υπολογίζουμε, για κάθε δείγμα χωριστά, τις τιμές της εμπειρικής συνάρτησης κατανομής και τις αποθηκεύουμε σε μία άλλη μεταβλητή, έστω **ra**. Στην συνέχεια, ορίζουμε δύο ακόμη μεταβλητές (έστω **sx** και **sy**). Σε κάθε κελλί της **sx**, καταχωρίζουμε την τιμή του αντίστοιχου κελλιού της **ra** αν αυτό περιέχει παρατήρηση του δείγματος X (μας το δείχνει η **d**) ή την μεγαλύτερη από τις τιμές της **ra** που είναι πάνω από το συγκεκριμένο κελλί και αντιστοιχούν σε παρατηρήσεις του δείγματος X αν το συγκεκριμένο κελλί περιέχει παρατήρηση του δείγματος Y. Αν δεν

υπάρχουν παρατηρήσεις του δείγματος X πάνω από το εν λόγω κελλί, καταχωρίζουμε την τιμή 0. Με όμοιο τρόπο, εργαζόμαστε και στην περίπτωση της μεταβλητής sy . Μόλις αυτή η διαδικασία ολοκληρωθεί, η sx θα δίνει τις τιμές της $S_x(\cdot)$ και η sy τις τιμές της $S_y(\cdot)$ στις τιμές που αντιστοιχούν στις παρατηρήσεις των δειγμάτων. Η κορυφή του παραθύρου δεδομένων δείχνει ως εξής:

a	d	ra	sx	sy
5.2	1	0.0667	0.0000	0.0667
5.7	1	0.1333	0.0000	0.1333
5.9	1	0.2000	0.0000	0.2000
6.5	1	0.2667	0.0000	0.2667
6.8	1	0.3333	0.0000	0.3333
7.6	0	0.1111	0.1111	0.3333
8.2	1	0.4000	0.1111	0.4000
8.4	0	0.2222	0.2222	0.4000
8.6	0	0.3333	0.3333	0.4000
8.7	0	0.4444	0.4444	0.4000
9.1	1	0.4667	0.4444	0.4667
9.3	0	0.5555	0.5555	0.4667

Τέλος, σε μία άλλη στήλη (έστω ds) καταχωρίζουμε τις τιμές των απόλυτων διαφορών μεταξύ sx και sy . Η μέγιστη τιμή της στήλης αυτής είναι η τιμή της στατιστικής συνάρτησης T_1 . Στο παράδειγμα μας, η τιμή της T_1 προκύπτει ίση με 0.4, και δεν υπερβαίνει το 0.95 ποσοστιαίο σημείο της κατανομής της κάτω από την H_0 , η τιμή του οποίου είναι 8/15. Κατά συνέπεια, τα δύο δείγματα μπορούν να θεωρηθούν ως προερχόμενα από πληθυσμούς με την ίδια κατανομή σε επίπεδο σημαντικότητας 5%. Για να υπολογίσουμε τις τιμές των στατιστικών συναρτήσεων T_1^+ και T_1^- , αρκεί να αναζητήσουμε την ελάχιστη και την μέγιστη των τιμών της ds , αντίστοιχα.

Λύση με το SPSS: Για την διεξαγωγή του ελέγχου Kolmogorov-Smirnov για δύο ανεξάρτητα δείγματα με το SPSS, τα δεδομένα εισάγονται με την εξής μορφή: σε μία μεταβλητή (έστω x), καταχωρίζουμε και τα δυο δείγματα και, σε μία άλλη (έστω y), καταχωρίζουμε κωδικές τιμές που δείχνουν σε ποιο δείγμα ανήκει η κάθε παρατήρηση (π.χ. 0 για τις τιμές του πρώτου δείγματος και 1 για τις τιμές του δεύτερου).

Στην συνέχεια, επιλέγουμε **Analyze, Nonparametric Tests, 2 Independent Samples** και οδηγούμεθα στο ίδιο πλαίσιο διαλόγου που έχουμε χρησιμοποιήσει και για τον έλεγχο Wilcoxon-Mann-Whitney. Η μόνη διαφορά τώρα είναι ότι, στο πεδίο **Test Type**, επιλέγουμε **Kolmogorov-Smirnov Z**. Επειδή τα δείγματα είναι μικρά, επιλέγουμε **Exact** από το πλαίσιο διαλόγου **Exact Tests**.

Τα αποτελέσματα του ελέγχου είναι:

Frequencies

	Y	N
X	0	9
	1	15
	Total	24

Test Statistics^a

		X
Most Extreme Differences	Absolute	.400
	Positive	.333
	Negative	-.400
Kolmogorov-Smirnov Z		.949
Asymp. Sig. (2-tailed)		.329
Exact Sig. (2-tailed)		.265
Point Probability		.051

a Grouping Variable: Y

Στον πρώτο πίνακα, βλέπουμε πόσες παρατηρήσεις προέρχονται από κάθε δείγμα. Στον δεύτερο, μας δίνονται (στα πεδία **Most Extreme Differences**) οι τιμές των συναρτήσεων T_1 , T_1^+ και T_1^- . Το παρατηρούμενο επίπεδο σημαντικότητας δίνεται μόνο για δίπλευρο έλεγχο. Παρατηρούμε ότι οι τιμές του ακριβούς και του ασυμπτωτικού

κρίσιμου επιπέδου (**Exact Sig.** και **Asymp. Sig.** αντίστοιχα) διαφέρουν (λόγω του μικρού μεγέθους των δειγμάτων). Η σωστή επιλογή είναι η τιμή του ακριβούς κρίσιμου επιπέδου (0.265). Με βάση την τιμή η μηδενική υπόθεση μπορεί να θεωρηθεί εύλογη.

Σημειώνεται ότι, για μονόπλευρο έλεγχο, χρειάζεται να χρησιμοποιήσουμε πίνακες αφού το πακέτο δεν δίνει κρίσιμο επίπεδο. Πρέπει όμως να προσέξουμε ποια στατιστική συνάρτηση χρησιμοποιούμε. Θα είναι μία από τις **positive** και **negative** του πεδίου **Most Extreme Differences**. Το SPSS ορίζει το πρόσημό τους με βάση τις διαφορές των τιμών του μικρότερου δείγματος από τις τιμές του μεγαλύτερου δείγματος (Συμβολικά, *παρατηρήσεις δείγματος με τις περισσότερες παρατηρήσεις – παρατηρήσεις δείγματος με τις λιγότερες*). Αν λοιπόν εδώ συμβολίσουμε με Y το δεύτερο δείγμα (**POP=1**) και με X το πρώτο (**POP=0**), έχουμε **positive**=sup($S_Y(\cdot) - S_X(\cdot)$) και **negative**=sup($S_X(\cdot) - S_Y(\cdot)$). Όταν τα δείγματα έχουν το ίδιο μέγεθος, τότε στις παραπάνω διαφορές μειωτέοι είναι οι τιμές του δείγματος του οποίου τον κωδικό έχουμε δηλώσει πρώτο στο πλαίσιο **Define Groups**.

Λύση με το SAS: Η διεξαγωγή του ελέγχου Smirnov γίνεται με την εντολή **proc npar1way edf;** όπως φαίνεται στα παρακάτω.

```
data sample;
input x code @@;
cards;
7.6 1 8.4 1 8.6 1 8.7 1 9.3 1
9.9 1 10.1 1 10.6 1 11.2 1
5.2 2 5.7 2 5.9 2 6.5 2 6.8 2 8.2 2 9.1 2
9.8 2 10.8 2 11.3 2 11.5 2 12.3 2 12.5 2 13.4 2 14.6 2
;
run;
proc npar1way edf;
class code;
var x;
run;
```

Το αποτέλεσμα που δίνεται είναι:

```
The SAS System
NPAR1WAY PROCEDURE
```

Kolmogorov-Smirnov Test for Variable X
Classified by Variable CODE

CODE	N	EDF at Maximum	Deviation from Mean at Maximum
1	9	1.0000000	0.75000000
2	15	0.6000000	-.580947502
	----- 24	----- 0.7500000	

Maximum Deviation Occurred at Observation 9
 Value of X at Maximum 11.2000000
 Kolmogorov-Smirnov 2-Sample Test (Asymptotic)
 KS = 0.193649 D = 0.400000
 KSa = 0.948683 Prob > KSa = 0.3291

Το πακέτο δίνει την επισήμανση ότι η μεγαλύτερη τιμή της στατιστικής συνάρτησης T_1 επιτυγχάνεται στην παρατήρηση με αύξοντα αριθμό 9, που αντιστοιχεί στην τιμή 11.2. Η τιμή της ελεγχουσυνάρτησης δίνεται στο πεδίο **D=0.400000**. Η ασυμπτωτική τιμή του κρίσιμου επιπέδου υπολογίζεται ίση με 0.3291.

Παράδειγμα 5.1.2: Ας υποθέσουμε ότι ενδιαφερόμαστε να ελέγξουμε την υπόθεση ότι ο πληθυσμός των βιβλίων στατιστικού περιεχομένου στα ράφια της βιβλιοθήκης του Σκωτσέτζου Καθηγητή του παραδείγματος 3.3.1 διαφέρει από αυτόν των βιβλίων γενικού ενδιαφέροντος ως προς τον αριθμό σελίδων, με βάση τα ανεξάρτητα τυχαία δείγματα των 12 βιβλίων στατιστικής και των 16 βιβλίων γενικού ενδιαφέροντος που επέλεξε. Οι αριθμοί των σελίδων των βιβλίων αυτών που κατεγράφησαν δίνονται, για ευκολία, στον πίνακα που ακολουθεί κατά αύξουσα σειρά τάξης μεγέθους.

	Αριθμός σελίδων											
Βιβλία Στατιστικού Περιεχομένου	126	142	156	228	245	246	370	419	433	454	478	503
Βιβλία Γενικού Ενδιαφέροντος	29	39	60	78	82	112	125	170	192	224	263	275
	276	286	369	756								

Με βάση τα στοιχεία του πίνακα αυτού, θα μπορούσε να συμπεράνει κανείς ότι υπάρχουν ενδείξεις ότι τα δύο τυχαία δείγματα βιβλίων προήλθαν από πληθυσμούς που διαφέρουν ως προς τον αριθμό σελίδων;

Λύση: Ακολουθώντας τα ίδια βήματα που ακολουθήσαμε στο προηγούμενο παράδειγμα, κατασκευάζουμε τον πίνακα 5.1.1.

Από τον πίνακα, προκύπτει ότι η τιμή της στατιστικής συνάρτησης T_1 είναι 0.4375 και παρατηρείται στην τιμή 369 του δείγματος των αριθμών των σελίδων των βιβλίων γενικού ενδιαφέροντος, δηλαδή

$$\begin{aligned} T_1 &= \sup_x |S_X(x) - S_Y(x)| \\ &= |S_X(369) - S_Y(369)| \\ &= 0.4375 \end{aligned}$$

Από τον πίνακα 21 του παραρτήματος βλέπουμε ότι το 0.95-ποσοστιαίο σημείο της κατανομής της στατιστικής συνάρτησης T_1 , για τον αμφίπλευρο έλεγχο και για $n=12=N_1$ και $m=16=N_2$, είναι

$$w_{0.95} = 23/48 = 0.4792.$$

Επομένως, σε επίπεδο σημαντικότητας 5% δεν απορρίπτεται η μηδενική υπόθεση ταυτοτικής ισότητας των συναρτήσεων κατανομών των αριθμών των σελίδων των δύο πληθυσμών βιβλίων.

Πίνακας 5.1.1

Αριθμοί σελίδων βιβλίων γενικού ενδιαφέροντος και στατιστικού περιεχομένου

Αριθμός βιβλίων w_i				
Γενικού ενδιαφέροντος	Στατιστικής	$S_X(w_i)$	$S_Y(w_i)$	$S_X(w_i) - S_Y(w_i)$
29		0.0625	0	0.0625
39		0.1250	0	0.1250
60		0.1875	0	0.1875

78		0.2500	0	0.2500
82		0.3125	0	0.3125
112		0.3750	0	0.3750
125		0.4375	0	0.4375
	126	0.4375	0.0833	0.3542
	142	0.4375	0.1667	0.2708
	156	0.4375	0.2500	0.1875
170		0.5000	0.2500	0.2500
192		0.5625	0.2500	0.3125
224		0.6250	0.2500	0.3750
	228	0.6250	0.3333	0.2917
	245	0.6250	0.4167	0.2083
	246	0.6250	0.5000	0.1250
263		0.6875	0.5000	0.1875
275		0.7500	0.5000	0.2500
276		0.8125	0.5000	0.3125
286		0.8750	0.5000	0.3750
369		0.9375	0.5000	0.4375
	370	0.9375	0.5833	0.3542
	419	0.9375	0.6667	0.2708
	433	0.9375	0.7500	0.1875
	454	0.9375	0.8333	0.1042
	478	0.9375	0.9167	0.0208
	503	0.9375	1.0000	0.0625
756		1.0000	1.0000	0.0000

Λύση με το MINITAB: Εργαζόμενοι όπως στο παράδειγμα για την διεξαγωγή του ελέγχου Kolmogorov-Smirnov για δύο δείγματα, προκύπτει ότι η τιμή της ελεγχουσυνάρτησης T_1 είναι ίση με 0.4375 και δεν υπερβαίνει την τιμή του 0.95 ποσοστιαίου σημείου της κατανομής T_1 , το οποίο είναι ίσο με 0.4792.

Λύση με το SPSS: Εργαζόμενοι όπως στο παράδειγμα, για την διεξαγωγή του ελέγχου Kolmogorov-Smirnov για δύο ανεξάρτητα δείγματα, οδηγούμεθα στα εξής αποτελέσματα:

Frequencies

	Y	N
X	0	12
	1	16
	Total	28

Test Statistics^a

		X
Most Extreme Differences	Absolute	.438
	Positive	.438
	Negative	-.063
Kolmogorov-Smirnov Z		1.146
Asymp. Sig. (2-tailed)		.145
Exact Sig. (2-tailed)		.113
Point Probability		.029

a Grouping Variable: Y

Και εδώ, χρησιμοποιούμε την τιμή του ακριβούς επιπέδου σημαντικότητας (0.113) και συμπεραίνουμε ότι η H_0 μπορεί να θεωρηθεί εύλογη στα συνήθη επίπεδα σημαντικότητας.

Λύση με το SAS: Εργαζόμενοι όπως στο προηγούμενο παράδειγμα, πληκτρολογούμε τις εντολές:

```
data books;
input x code @@;
cards;
126 1 142 1 156 1 228 1 245 1 246 1
370 1 419 1 433 1 454 1 478 1 503 1
 29 2  39 2  60 2  78 2  82 2 112 2
125 2 170 2 192 2 224 2 263 2 275 2
276 2 286 2 369 2 756 2
;
run;
proc npar1way edf;
class code;
var x;
run;
```

Τα αποτελέσματα περιέχονται στον εξής πίνακα:

```

The SAS System
N P A R 1 W A Y P R O C E D U R E
Kolmogorov-Smirnov Test for Variable X
```

Classified by Variable CODE			
CODE	N	EDF at Maximum	Deviation from Mean at Maximum
1	12	0.00000000	-.866025404
2	16	0.43750000	0.75000000
	----- 28	0.25000000	
Maximum Deviation Occurred at Observation			19
Value of X at Maximum			125.000000
Kolmogorov-Smirnov 2-Sample Test (Asymptotic)			
KS = 0.216506		D = 0.437500	
KSa = 1.14564		Prob > KSa = 0.1448	

Η μεγαλύτερη τιμή της στατιστικής συνάρτησης T_1 επιτυγχάνεται στην παρατήρηση με αύξοντα αριθμό 19, που αντιστοιχεί στην τιμή 125. (Η ίδια τιμή της στατιστικής συνάρτησης επιτυγχάνεται και στην τιμή 369 που δόθηκε στην αναλυτική λύση). Η τιμή της ελεγχοσυνάρτησης δίνεται στο πεδίο **D=0.437500**. Η ασυμπτωτική τιμή του κρίσιμου επιπέδου υπολογίζεται ίση με 0.1448.

Σημείωση: Ας σημειωθεί ότι, για την εφαρμογή του ελέγχου αυτού, τα δεδομένα θα πρέπει να βρίσκονται σε διατεταγμένη κλίμακα τουλάχιστον. Ο έλεγχος Smirnov είναι ακριβής στην περίπτωση που οι τυχαίες μεταβλητές είναι συνεχείς. Αν οι τυχαίες μεταβλητές είναι διακριτές, τότε ο έλεγχος εξακολουθεί να ισχύει, αλλά είναι περισσότερο συντηρητικός.

Παρατήρηση: Αν και δεν είναι προφανές, οι στατιστικές συναρτήσεις T_1 , T_1^+ και T_1^- , εξαρτώνται μόνο από την τάξη των μεταβλητών X και Y στο διατεταγμένο ενιαίο δείγμα των X και Y και, για τον υπολογισμό τους, δεν απαιτείται γνώση των αριθμητικών τιμών των παρατηρήσεων. Για την καλύτερη κατανόηση του γεγονότος αυτού, ας υποθέσουμε ότι υπάρχουν τρεις τιμές X και δύο τιμές Y. Υπάρχουν τότε $\binom{5}{2} = 10$ δυνατές τοποθετήσεις των στοιχείων του ενιαίου

δείγματος. Ο πίνακας που ακολουθεί συνοψίζει τις δυνατές αυτές τοποθετήσεις μαζί με τις τιμές των στατιστικών συναρτήσεων T_1 , T_1^+ και T_1^- που αντιστοιχούν σ' αυτές.

Διάταξη	T_1	T_1^+	T_1^-	Διάταξη	T_1	T_1^+	T_1^-
X<X<X<Y<Y	1	1	0	X<Y<X<Y<X	1/3	1/3	1/3
X<X<Y<X<Y	2/3	2/3	0	Y<X<X<Y<X	1/2	1/6	1/2
X<Y<X<X<Y	1/2	1/2	1/6	X<Y<Y<X<X	2/3	1/3	2/3
Y<X<X<X<Y	1/2	1/2	1/2	Y<X<Y<X<X	2/3	0	2/3
X<X<Y<Y<X	2/3	2/3	1/3	Y<Y<X<X<X	1	0	1

Αν η μηδενική υπόθεση στον αμφίπλευρο έλεγχο είναι αληθής, οι δύο συναρτήσεις κατανομής ταυτίζονται και οι δυνατές τοποθετήσεις είναι ισοπίθανες κάτω από την υπόθεση ότι οι τυχαίες μεταβλητές είναι συνεχείς. Επομένως, στον αμφίπλευρο έλεγχο, η πιθανότητα κάθε μιας τοποθέτησης είναι:

$$\frac{1}{\binom{m+n}{n}} = \frac{1}{\binom{5}{3}} = \frac{1}{10}.$$

Επομένως, συνάγονται οι εξής κατανομές πιθανότητας για τις στατιστικές συναρτήσεις T_1 , T_1^+ και T_1^- :

$$\begin{aligned} P(T_1 = 1/3) &= 1/10 & P(T_1^+ = 0) &= 1/5 & P(T_1^- = 0) &= 1/5 \\ P(T_1 = 1/2) &= 3/10 & P(T_1^+ = 1/5) &= 1/10 & P(T_1^- = 1/5) &= 1/10 \\ P(T_1 = 2/3) &= 2/5 & P(T_1^+ = 1/3) &= 1/5 & P(T_1^- = 1/3) &= 1/5 \\ P(T_1 = 1) &= 1/5 & P(T_1^+ = 1/2) &= 1/5 & P(T_1^- = 1/2) &= 1/5 \\ & & P(T_1^+ = 2/3) &= 1/5 & P(T_1^- = 2/3) &= 1/5 \\ & & P(T_1^+ = 1) &= 1/10 & P(T_1^- = 1) &= 1/10. \end{aligned}$$

Το γεγονός ότι οι κατανομές των στατιστικών συναρτήσεων T_1^+ και T_1^- ταυτίζονται για $n=3$ και $m=2$, δεν είναι συμπτωματικό. Οι κατανομές αυτές ταυτίζονται για όλες τις τιμές των n και m . Όμως, η τεχνική εξοικονόμησης χώρου η οποία χρησιμοποιήθηκε στην παρουσίαση των ποσοστιαίων σημείων των πινάκων 20 και 21 του παραρτήματος, σύμφωνα με την οποία το $(1-\alpha)$ -ποσοστιαίο σημείο της στατιστικής συνάρτησης T_1 στον αμφίπλευρο έλεγχο ταυτίζεται με το $(1-\alpha/2)$ -ποσοστιαίο σημείο της στατιστικής συνάρτησης T_1^+ στον μονόπλευρο έλεγχο, ισχύει μόνο όταν το επίπεδο σημαντικότητας α είναι μικρό. Ας σημειωθεί, για παράδειγμα, ότι στην περίπτωση αυτή των πέντε τιμών,

$$P(T_1 \geq 1) = 2 P(T_1^+ \geq 1) \text{ και } P(T_1 \geq 2/3) = 2 P(T_1^+ \geq 2/3),$$

αλλά

$$P(T_1 \geq 1/2) \neq 2 P(T_1^+ \geq 1/2).$$

Η μηδενική κατανομή (δηλαδή, η κατανομή όταν η υπόθεση H_0 αληθεύει) στους μονόπλευρους ελέγχους προσδιορίζεται με τον ίδιο τρόπο, γιατί κάτω από τη μονόπλευρη μηδενική υπόθεση, το μέγεθος της κρίσιμης περιοχής είναι μέγιστο όταν $F_X(x) = F_Y(x)$, για κάθε x .

5.2 Ο ΕΛΕΓΧΟΣ ΤΩΝ CRAMÉR-VON MISES

Ο έλεγχος που εξετάζεται στην ενότητα αυτή αποτελεί μια εναλλακτική μορφή ελέγχου της υπόθεσης ισότητας δύο κατανομών έναντι αμφίπλευρης εναλλακτικής υπόθεσης μόνο. Ο υπολογισμός της τιμής της ελεγχοσυνάρτησης, η οποία χρησιμοποιείται στο πλαίσιο του ελέγχου αυτού είναι πολυπλοκότερος από τον υπολογισμό της ελεγχοσυνάρτησης του ελέγχου Smirnov.

Εστω X_1, X_2, \dots, X_n ένα τυχαίο δείγμα n παρατηρήσεων πάνω στην τυχαία μεταβλητή X . Έστω Y_1, Y_2, \dots, Y_m ένα τυχαίο δείγμα m παρατηρήσεων πάνω στην τυχαία μεταβλητή Y , ανεξάρτητο του πρώτου δείγματος. Έστω, επιπλέον, ότι $F_X(x)$, $x \in (-\infty, +\infty)$ και $F_Y(y)$, $y \in (-\infty, +\infty)$ συμβολίζουν τις συναρτήσεις κατανομής των τυχαίων μεταβλητών X και Y , αντίστοιχα. Οι υποθέσεις τις οποίες ενδιαφερόμαστε να ελέγξουμε είναι οι εξής:

$$H_0: F_X(x) = F_Y(x), \text{ για κάθε } x \in (-\infty, +\infty)$$

$$H_1: F_X(x) \neq F_Y(x), \text{ για τουλάχιστον ένα } x \in (-\infty, +\infty).$$

Όπως και στην περίπτωση του ελέγχου Smirnov, η στατιστική συνάρτηση η οποία προτείνεται ως ελεγχοσυνάρτηση βασίζεται στις εμπειρικές συναρτήσεις κατανομής των δύο δειγμάτων $S_X(x)$ και $S_Y(y)$, αντίστοιχα και εκφράζει ένα μέτρο της μέσης τετραγωνικής απόκλισής τους. Συγκεκριμένα,

$$T_2 = \frac{mn}{(m+n)^2} \sum_{\substack{x_1, \dots, x_n \\ y_1, \dots, y_m}} [S_X(x) - S_Y(x)]^2,$$

όπου οι τετραγωνικές αποκλίσεις στο άθροισμα υπολογίζονται σε κάθε τιμή του δείγματος X_1, X_2, \dots, X_n και κάθε τιμή του δείγματος $Y_1,$

Y_2, \dots, Y_m . Μια σαφέστερη μορφή της παραπάνω ελεγχουσυνάρτησης είναι η εξής:

$$T_2 = \frac{mn}{(m+n)^2} \left\{ \sum_{i=1}^n [S_X(x_i) - S_Y(x_i)]^2 + \sum_{j=1}^m [S_X(y_j) - S_Y(y_j)]^2 \right\}$$

Είναι προφανές ότι μεγάλες τιμές της στατιστικής συνάρτησης T_2 αποτελούν ένδειξη εναντίον της μηδενικής υπόθεσης. Επομένως, ο κανόνας απόφασης έχει την εξής μορφή:

Η μηδενική υπόθεση H_0 απορρίπτεται σε επίπεδο σημαντικότητας α , αν η τιμή της στατιστικής συνάρτησης T_2 υπερβαίνει το $(1-\alpha)$ -ποσοστιαίο σημείο της κατανομής της. Δηλαδή, η μηδενική υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας α , αν $T_2 > w_{1-\alpha}$.

Η ακριβής μορφή της κατανομής της στατιστικής συνάρτησης των Cramér - von Mises μπορεί να προσδιορισθεί με τον ίδιο τρόπο που προσδιορίζεται η κατανομή της ελεγχουσυνάρτησης Smirnov. Τα ακριβή ποσοστιαία σημεία της κατανομής της τυχαίας μεταβλητής της στατιστικής συνάρτησης T_2 προσδιορίστηκαν από τον Anderson το 1962 για μικρά δείγματα. Η ασυμπτωτική κατανομή της στατιστικής συνάρτησης T_2 είχε νωρίτερα προσδιορισθεί από τους Anderson και Darling (1952) σε μία εργασία τους, η οποία αναφερόταν στην ελεγχουσυνάρτηση που είχε προταθεί από τους Cramér και von Mises στο πλαίσιο προβλημάτων ελέγχου υποθέσεων καλής προσαρμογής. Αυτός είναι και ο λόγος για τον οποίο η στατιστική συνάρτηση T_2 ονομάζεται *ελεγχουσυνάρτηση των Cramér και von Mises*, παρά το γεγονός ότι ούτε ο Cramér ούτε ο von Mises την είχαν επινοήσει. Στην συνέχεια, παραθέτουμε ορισμένα ποσοστιαία σημεία της ασυμπτωτικής κατανομής της στατιστικής συνάρτησης

T_2 , όπως αυτά υπολογίσθηκαν από τους Anderson και Darling το 1952. (Τα ακριβή ποσοστιαία σημεία της κατανομής της στατιστικής συνάρτησης T_2 για $n+m \leq 17$ δόθηκαν από τον Burr (1963, 1964):

$w_{0.10} = 0.046$	$w_{0.50} = 0.119$	$w_{0.10} = 0.347$
$w_{0.20} = 0.062$	$w_{0.60} = 0.147$	$w_{0.10} = 0.461$
$w_{0.30} = 0.079$	$w_{0.70} = 0.184$	$w_{0.10} = 0.743$
$w_{0.40} = 0.097$	$w_{0.80} = 0.241$	$w_{0.10} = 1.168$.

Παράδειγμα 5.2.1: Χρησιμοποιώντας τα δεδομένα του παραδείγματος 5.1.1 της προηγούμενης ενότητας, έχουμε ότι

$$\sum_{i=1}^9 [S_X(x_i) - S_Y(x_i)]^2 = 0.459$$

και

$$\sum_{j=1}^{15} [S_X(y_j) - S_Y(y_j)]^2 = 0.657$$

Επομένως,

$$\begin{aligned} T_2 &= \frac{(15)(9)}{(24)^2} (0.459 + 0.657) \\ &= 0.262. \end{aligned}$$

Η μηδενική υπόθεση ότι οι δύο συναρτήσεις κατανομής ταυτίζονται δεν απορρίπτεται σε επίπεδο σημαντικότητας $\alpha=0.05$, γιατί η παρατηρούμενη τιμή της T_2 είναι εκτός της κρίσιμης περιοχής. Πράγματι,

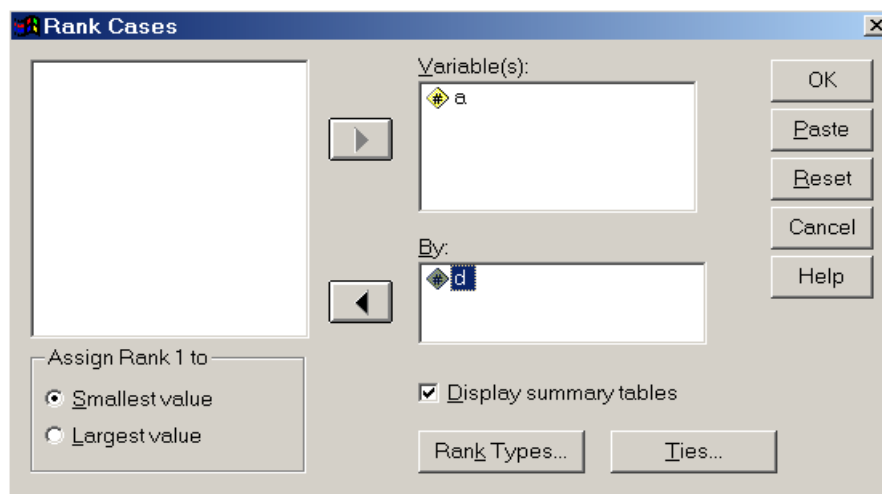
$$\tau_2 = 0.262 < w_{0.95} = 0.461.$$

Το κρίσιμο επίπεδο, μπορεί εύκολα να διαπιστωθεί ότι είναι περίπου $\hat{\alpha} = 0.18$. Όπως παρατηρούμε, το επίπεδο αυτό είναι

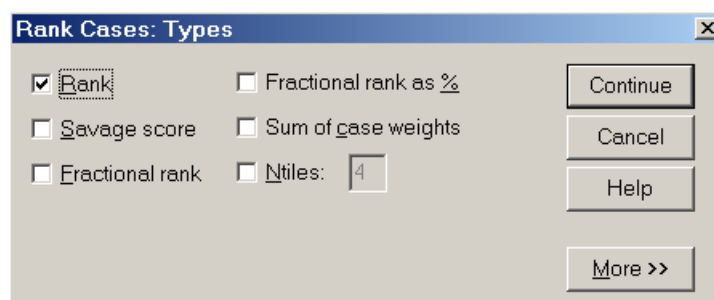
ελαφρώς χαμηλότερο από το αντίστοιχο κρίσιμο επίπεδο του ελέγχου Smirnov για τα ίδια δεδομένα.

Λύση με το MINITAB: Το MINITAB δεν διαθέτει τον έλεγχο Cramér – von Mises. Μπορεί όμως να διεξαχθεί με έμμεσο τρόπο, υπολογίζοντας την τιμή της ελεγχοσυνάρτησης T_2 στην οποία βασίζεται και, στην συνέχεια, συγκρίνοντάς την με το κατάλληλο ποσοστιαίο σημείο. Ο υπολογισμός της ελεγχοσυνάρτησης T_2 γίνεται ως εξής: Υπολογίζουμε τις τιμές των δύο εμπειρικών συναρτήσεων κατανομής και καταχωρίζουμε σε δύο στήλες τις τιμές που παίρνουν σε όλα τα σημεία που αντιστοιχούν στις παρατηρήσεις των δειγμάτων. Σε μία άλλη στήλη, καταχωρίζουμε τα τετράγωνα των διαφορών των τιμών των δύο στηλών τους. Το άθροισμα των στοιχείων της στήλης αυτής πολλαπλασιάζεται με τον κατάλληλο συντελεστή $(mn/(m+n)^2)$ δίνει την τιμή της ελεγχοσυνάρτησης T_2 .

Λύση με Το SPSS: Ο έλεγχος Cramér–von Mises δεν είναι διαθέσιμος στο SPSS. Η διεξαγωγή του όμως είναι δυνατή με έμμεσο τρόπο. Αυτό γίνεται ως εξής: Σε μία στήλη (έστω **a**) καταχωρίζουμε και τα δύο δείγματα. Σε μία άλλη στήλη (έστω **d**), καταχωρίζουμε μία κωδική μεταβλητή της οποίας οι τιμές δείχνουν από ποιο δείγμα προέρχεται κάθε παρατήρηση (π.χ. 0 για τις παρατηρήσεις του δείγματος της X και 1 για τις παρατηρήσεις του δείγματος της Y). Στην συνέχεια, για να υπολογίσουμε τις $S_X(\cdot)$ και $S_Y(\cdot)$, επιλέγουμε **Transform, Rank Cases** και οδηγούμεθα στο εξής πλαίσιο διαλόγου:



Στο πεδίο **Variable(s)**, δηλώνουμε την στήλη που περιέχει τα δείγματα (**a**), ενώ, στο πεδίο **By**, δηλώνουμε την κωδική μεταβλητή (**d**). Με αυτόν τον τρόπο, υπολογίζονται χωριστά οι τάξεις μεγέθους κάθε δείγματος. Από το πλαίσιο διαλόγου στο οποίο οδηγεί το πλήκτρο **Ties**, επιλέγουμε **High** (όπως και σε προηγούμενα παραδείγματα) και, στο επανεμφανιζόμενο αρχικό πλαίσιο διαλόγου, πιέζουμε το πλήκτρο **Rank Types**. Αυτό οδηγεί στο εξής πλαίσιο:



Ακυρώνουμε την επιλογή του πεδίου **Rank** και επιλέγουμε **Fractional rank**. Έτσι, προκύπτει απ' ευθείας η εμπειρική συνάρτηση κατανομής κάθε δείγματος χωρίς να χρειασθεί να διαιρέσουμε εμείς με το μέγεθος του κάθε δείγματος.

Αφού ολοκληρώσουμε την διαδικασία αυτή, θα έχουμε, σε μία στήλη με όνομα **ra**, τις δύο εμπειρικές συναρτήσεις κατανομής. Πρέπει όμως να τις ξεχωρίσουμε σε δύο στήλες. Αυτό γίνεται εύκολα για μικρά δείγματα όπως αυτά του παραδείγματός μας: Ταξινομούμε σε αύξουσα τάξη μεγέθους την μεταβλητή **a**. Ταυτόχρονα αναδιατάσσονται οι τιμές των μεταβλητών **d** και **ra**. Ορίζουμε δύο νέες μεταβλητές (έστω **sx** και **sy**). Σε κάθε κελλί της **sx**, καταχωρίζουμε την αντίστοιχη τιμή της **ra** αν αυτό περιέχει παρατήρηση του δείγματος X (μας το δείχνει η **d**) ή την μεγαλύτερη από τις τιμές της μεταβλητής **ra** που είναι πάνω από το εν λόγω κελλί και αντιστοιχούν σε παρατηρήσεις του δείγματος X, αν αυτό περιέχει παρατήρηση του δείγματος Y. Αν δεν υπάρχουν παρατηρήσεις του δείγματος X πάνω από το συγκεκριμένο κελλί, καταχωρίζουμε 0. Αντίστοιχη διαδικασία ακολουθείται και για την **sy**. Για να γίνουν τα παραπάνω περισσότερο κατανοητά δείχνουμε τις πρώτες καταχωρίσεις του φύλλου δεδομένων μετά από όλες τις παραπάνω διαδικασίες:

	a	d	ra	sx	sy
1	5.20	1.00	.0667	.0000	.0667
2	5.70	1.00	.1333	.0000	.1333
3	5.90	1.00	.2000	.0000	.2000
4	6.50	1.00	.2667	.0000	.2667
5	6.80	1.00	.3333	.0000	.3333
6	7.60	.00	.1111	.1111	.3333
7	8.20	1.00	.4000	.1111	.4000
8	8.40	.00	.2222	.2222	.4000
9	8.60	.00	.3333	.3333	.4000
10	8.70	.00	.4444	.4444	.4000
11	9.10	1.00	.4667	.4444	.4667

Σημειώνεται ότι η διαδικασία δημιουργίας των **sx** και **sy** από την **ra** με τον τρόπο που περιγράφεται παραπάνω, είναι εύκολη μόνο

για μικρά δείγματα. Για μεγάλα δείγματα είναι χρονοβόρα και μπορεί να οδηγήσει σε λάθη.

Αυτό που απομένει είναι να δημιουργήσουμε μία μεταβλητή που θα περιέχει τα τετράγωνα των διαφορών των καταχωρίσεων των s_x και s_y , να υπολογίσουμε το άθροισμα των στοιχείων της και να το πολλαπλασιάσουμε με τον κατάλληλο συντελεστή $(mn/(m+n)^2)$ για να υπολογίσουμε την τιμή της ελεγχοσυνάρτησης T_2 .

Λύση με το SAS: Όπως παρατηρήθηκε σε προηγούμενο παράδειγμα, οι έλεγχοι Smirnov και Cramér-von Mises παρέχονται με την εντολή **proc npar1way edf;** όπως φαίνεται στον παρακάτω πίνακα.

```
data sample;
input x code @@;
cards;
7.6 1 8.4 1 8.6 1 8.7 1 9.3 1
9.9 1 10.1 1 10.6 1 11.2 1
5.2 2 5.7 2 5.9 2 6.5 2 6.8 2 8.2 2 9.1 2
9.8 2 10.8 2 11.3 2 11.5 2 12.3 2 12.5 2 13.4 2 14.6 2
;
run;
proc npar1way edf;
class code;
var x;
run;
```

Το αποτέλεσμα που προκύπτει περιέχεται στον πίνακα που ακολουθεί.

```

The SAS System

N P A R 1 W A Y P R O C E D U R E
Cramer-von Mises Test for Variable X
Cl assi fi ed by Vari able CODE

CODE          N          Summed
                Deviat ion
                from Mean
1                9          0.163483796
2               15          0.098090278

Cramer-von Mises Statistic (Asymptotic)
CM = 0.010899          CMa = 0.261574
```

Η παρατηρούμενη τιμή της T_2 δίνεται στο πεδίο **CMa = 0.261574**. Σύγκριση της τιμής αυτής με την $w_{0.95} = 0.461$ οδηγεί σε μη απόρριψη

της μηδενικής υπόθεσης ότι οι δύο συναρτήσεις κατανομής ταυτίζονται, σε επίπεδο σημαντικότητας $\alpha=0.05$

Παρατήρηση: Μία ισοδύναμη μορφή της στατιστικής συνάρτησης T_2 λαμβάνεται αν συμβολίσουμε με $R(X^{(i)})$ και $R(Y^{(j)})$ τους βαθμούς (τις τάξεις), στο ενιαίο διατεταγμένο δείγμα, της i κατά αύξουσα σειρά μεγέθους παρατήρησης του δείγματος των X ($X^{(i)}$) και της j κατά αύξουσα σειρά μεγέθους παρατήρησης του δείγματος των Y ($Y^{(j)}$), αντίστοιχα. Τότε, αν δεν υπάρχουν περιπτώσεις ταύτισης τιμών, η T_2 γράφεται με την μορφή

$$T_2 = \frac{mn}{(m+n)^2} \left\{ \sum_{i=1}^n \left[\frac{R(X^{(i)})}{m} - i \frac{n+m}{nm} \right]^2 + \sum_{j=1}^m \left[\frac{R(Y^{(j)})}{n} - j \frac{n+m}{nm} \right]^2 \right\}.$$

Αν, επιπλέον, $n = m$, η T_2 γράφεται με την μορφή

$$T_2 = \frac{1}{4n^2} \left\{ \sum_{i=1}^n [R(X^{(i)}) - 2i]^2 + \sum_{j=1}^m [R(Y^{(j)}) - 2j]^2 \right\}.$$

Παρατήρηση: Για την εφαρμογή του ελέγχου αυτού, τα δεδομένα θα πρέπει να βρίσκονται σε κλίμακα διάταξης τουλάχιστον. Ο έλεγχος των Cramér-von Mises είναι ακριβής, εάν οι τυχαίες μεταβλητές είναι συνεχείς. Εάν οι τυχαίες μεταβλητές είναι διακριτές, ο έλεγχος ενδέχεται να είναι συντηρητικός.

ΑΣΚΗΣΕΙΣ

1. Να ελεγχθεί η μηδενική υπόθεση ότι $F(x) \leq G(x)$, αν οι παρατηρήσεις που προέρχονται από τον πληθυσμό που περιγράφεται από την συνάρτηση κατανομής $F(x)$ είναι οι 0.6, 0.8, 1.2 και 1.4 και οι παρατηρήσεις που προέρχονται από τον πληθυσμό που περιγράφεται από την συνάρτηση κατανομής $G(x)$ είναι οι 1.3, 1.3, 1.8, 2.4 και 2.9.
2. Σε ένα τυχαίο δείγμα πέντε μαθητών της έκτης τάξης ενός Δημοτικού σχολείου που ανήκε σε μία περιοχή κάποιας πόλης, δόθηκε ένα τεστ γνώσεων με τα εξής αποτελέσματα: 82, 74, 87, 86, 75. Σε ένα άλλο τυχαίο δείγμα οκτώ μαθητών της έκτης τάξης ενός άλλου Δημοτικού σχολείου που ανήκε σε μία διαφορετική περιοχή της πόλης, δόθηκε το ίδιο τεστ γνώσεων με αποτελέσματα: 88, 7, 91, 88, 94, 93, 83, 94. Θα μπορούσε να συμπεράνει κανείς ότι υπάρχει διαφορά στις γνώσεις, όπως αυτές μετριοούνται με αυτό το τεστ, των δύο πληθυσμών μαθητών της έκτης τάξης; (Να χρησιμοποιηθεί ο έλεγχος Smirnov).
3. Να χρησιμοποιηθεί ο έλεγχος των Cramér-von Mises στα δεδομένα της προηγούμενης άσκησης και να συγκριθούν τα αποτελέσματα με αυτά του ελέγχου Smirnov.