Ενότητα 3: Έλεγχοι καλής προσαρμογής (Goodness of fit tests)

Ένα σημαντικό πρόβλημα στην στατιστική είναι η εξεύρεση πληροφορίας σχετικά με την μορφή της κατανομής από την οποία προέρχεται ένα τυχαίο δείγμα. Είναι π.χ. γνωστό ότι οι περισσότεροι έλεγχοι γίνονται με την προϋπόθεση ότι (υπό την H_0) το τυχαίο δείγμα προέρχεται από μια συγκεκριμένη κατανομή. Μια τέτοια περίπτωση είναι τα t-tests που εξετάσαμε σε προηγούμενη ενότητα και τα οποία, ιδιαίτερα για μικρά δείγματα, προϋποθέτουν ότι το δείγμα προέρχεται από κανονικό πληθυσμό (υπό την H_0). Εάν το τυχαίο δείγμα δεν προέρχεται από την κατανομή κάτω από την οποία έχει κατασκευασθεί κάποιος έλεγχος τότε προφανώς το αντίστοιχο p-value που λαμβάνεται δεν είναι ακριβές (και επομένως η πιθανότητα σφάλματος τύπου I δεν είναι ακριβώς a). Συνεπώς είναι αρκετά χρήσιμη η δυνατότητα να ελέγχουμε αν κάποια δεδομένα προέρχονται από μια συγκεκριμένη κατανομή ή όχι.

Έλεγχοι αυτής της μορφής καλούνται «έλεγχοι καλής προσαρμογής» των δεδομένων σε μια συγκεκριμένη κατανομή και έχουν προταθεί αρκετοί. Σε αυτήν την ενότητα αρχικά θα εξετάσουμε κάποιους «εμπειρικούς» ελέγχους οι οποίοι γίνονται μέσω κάποιων γραφημάτων (P-P και Q-Q plots) ώστε να πάρουμε μια πρώτη εποπτική εικόνα για τα δεδομένα (τα γραφήματα αυτά δεν οδηγούν με σχετική «ασφάλεια» σε κάποια απόφαση) ενώ στη συνέχεια θα περάσουμε στους πιο σημαντικούς ελέγχους καλής προσαρμογής: το χι-τετράγωνο τεστ καλής προσαρμογής και το Kolmogorv-Smirnov τεστ.

Τέλος, ένα ενδιαφέρον παρεμφερές πρόβλημα αφορά δύο δείγματα και τον έλεγχο της υπόθεσης ότι τα δείγματα αυτά προέρχονται από τον ίδιο πληθυσμό (δηλαδή από την ίδια κατανομή). Για τον έλεγχο αυτό παρουσιάζονται εν συντομία τρία απαραμετρικά τεστ, το Kolmogorov-Smirnov για δυο δείγματα, το το Wald-Wolfowitz τέστ των ροών και το Mann-Whitney U τέστ.

3.1. P-P Plot και Q-Q Plot

Τα P-P Plot και Q-Q plot (probability-probability plot και Quantile-Quantile plot) είναι δύο γραφήματα τα οποία μας βοηθούν να ελέγξουμε αν κάποια δεδομένα προέρχονται από κάποια συγκεκριμένη κατανομή (π.χ. κανονική). Τα γραφήματα αυτά βασίζονται στην ακόλουθη παρατήρηση:

Αν $X_1,X_2,\ldots,X_n$ είναι ένα τυχαίο δείγμα (ανεξ. τ.μ.) από μια (συνεχή) κατανομή με σ.κ. F τότε οι νέες τ.μ. $Y_1 = F(X_1)$, $Y_2 = F(X_2)$, …, $Y_n = F(X_n)$ είναι και αυτές ανεξάρτητες και ακολουθούν την ομοιόμορφη U(0,1) κατανομή διότι

$$P(F(X) \le x) = P(X \le F^{-1}(x)) = F(F^{-1}(x)) = x,\ x \in [0,1].$$

Είναι εύκολο να αποδειχθεί ότι αν $Y_1,Y_2,\ldots,Y_n \sim$ U(0,1) τότε κάθε μια από τις διατεταγμένες τ.μ. $Y_{(1)},Y_{(2)},\ldots,Y_{(n)}$ ακολουθεί την κατανομή βήτα και συγκεκριμένα $Y_{(i)} \sim$ Beta(i, n–i+1) με $E(Y_{(i)}) = i/(n+1)$. Επομένως, για μεγάλο n θα ισχύει προσεγγιστικά ότι, για $i = 1,2,\ldots,n$,

$$Y_{(i)} = F(X_{(i)}) \approx \frac{i}{n+1} \quad \text{ή ισοδύναμα} \quad X_{(i)} \approx F^{-1}\left(\frac{i}{n+1}\right) \quad (\text{διότι και } V(Y_{(i)}) \to_{n\to\infty} 0)$$

Με άλλα λόγια, αν $X_i \sim F_0$ περιμένουμε ότι τα n σημεία του επιπέδου

$$\left(F_0(X_{(i)}), \tfrac{i}{n+1}\right),\ i = 1, 2, \ldots, n$$

ή ισοδύναμα τα n σημεία του επιπέδου

$$\left(X_{(i)}, F_0^{-1}\left(\tfrac{i}{n+1}\right)\right),\ i = 1, 2, \ldots, n$$

θα βρίσκονται «κοντά» στην διαγώνιο ($x = y$) που περνά από την αρχή των αξόνων. Το P-P plot ακριβώς είναι το γράφημα των πρώτων n σημείων (μαζί με τη διαγώνιο) ενώ το Q-Q Plot είναι το γράφημα των δεύτερων n σημείων (μαζί με τη διαγώνιο). Και στα δύο γραφήματα, αν τα σημεία βρίσκονται «κοντά» στη διαγώνιο (και «τυχαία» γύρω από αυτήν) τότε μπορεί να θεωρηθεί ότι τα

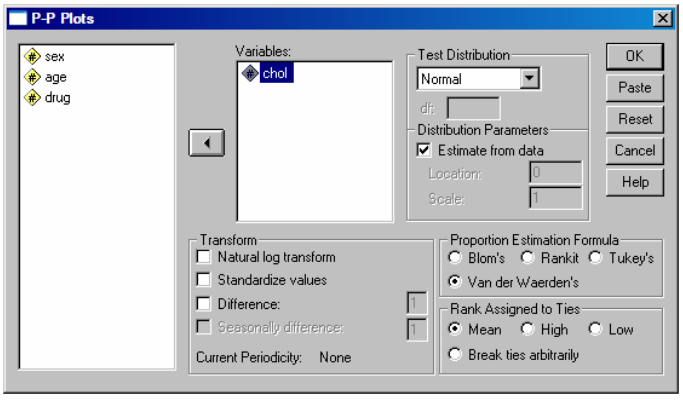
δεδομένα προέρχονται από την F_0. Ενδέχεται να μην είναι γνωστές όλες οι παράμετροι της κατανομής F_0 (π.χ. μπορεί να είναι κανονική με άγνωστο μ, σ^2). Σε αυτή την περίπτωση οι άγνωστοι παράμετροι εκτιμώνται από τα δεδομένα.

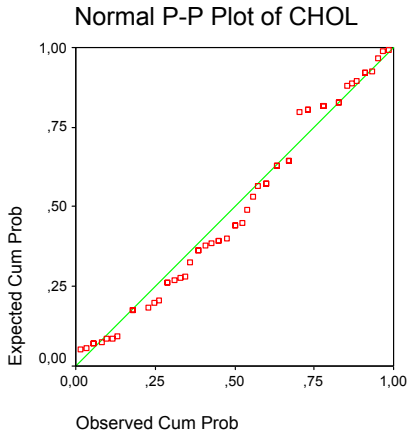
Όπως υπονοήθηκε και στην εισαγωγή της ενότητας αυτής, ο έλεγχος μέσω των παραπάνω γραφημάτων δεν μπορεί να είναι αξιόπιστος διότι δεν βασίζεται σε κάποιο στατιστικό κριτήριο που μας οδηγεί σε σωστή απόφαση π.χ. στο $1-a$ % των περιπτώσεων. Συνήθως γίνεται για να πάρουμε μια πρώτη εποπτική εικόνα και για να δούμε αν υπάρχουν κάποιες έκτροπες, σε σχέση με τις αναμενόμενες υπό την F_0, παρατηρήσεις.

Είναι προφανές ότι, εκτός της κανονικής, μπορούμε γραφικά να ελέγξουμε την καλή προσαρμογή των δεδομένων και σε άλλες κατανομές (αλλάζουμε την Test distribution).

Εφαρμογή 1. Να ελεγχθεί γραφικά (μέσω P-P plot ή Q-Q plot) αν οι παρατηρήσεις του δείκτη χοληστερίνης (μεταβλητή chol εφαρμογής Ενότητας 1) προέρχονται από την κανονική κατανομή.

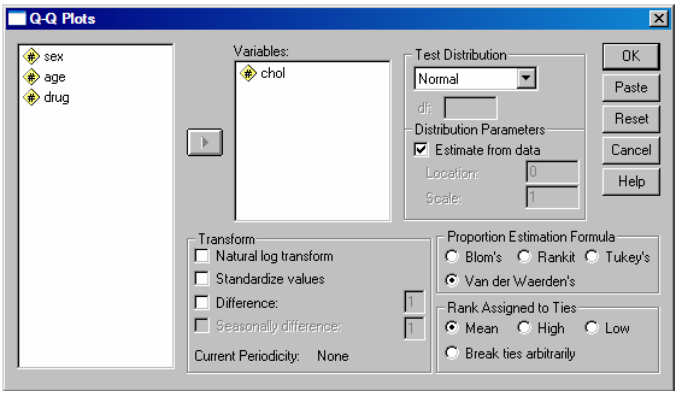
Ανοίγουμε την ανάλυση Graphs/P-P και επιλέγοντας την μεταβλητή chol (60 παρατηρήσεις) λαμβάνουμε το γράφημα στα δεξιά (ως proportion estimation formula επιλέγουμε την Van der Waerden's (=$r/(n+1)$) η οποία συμφωνεί με την προηγηθείσα ανάλυση θεωρώντας ότι θα πρέπει $F_0(X_{(r)}) \approx \frac{r}{n+1}$ (κάποιοι άλλοι ερευνητές παραθέτοντας κάποια δικαιολόγηση έχουν προτείνει την απεικόνιση των σημείων $(F_0(X_{(r)}), \frac{r-3/8}{n+1/4})$ (Blom's) ή των $(F_0(X_{(r)}), \frac{r-1/2}{n})$ (Rankit) κ.ο.κ.). Για μέτρια ή μεγάλα δείγματα δεν υπάρχει ουσιαστική διαφορά οποιαδήποτε proportion estimation formula και αν επιλέξουμε.

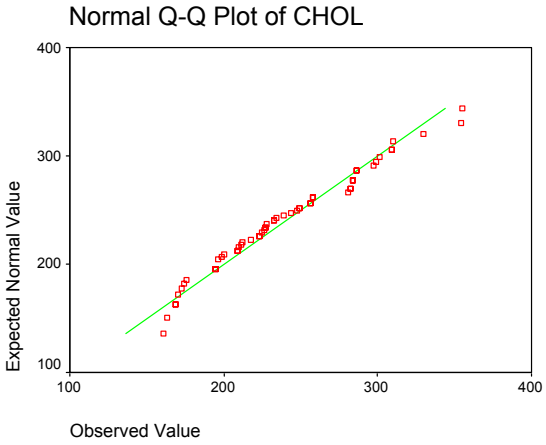




Φαίνεται ότι τα 60 σημεία του επιπέδου δεν «απέχουν» πολύ από την διαγώνιο, ούτε φαίνονται κάποιες «έκτροπες» παρατηρήσεις. Επομένως, τουλάχιστον γραφικά, δεν φαίνεται να υπάρχει επαρκής λόγος ώστε να μην θεωρήσουμε τα δεδομένα ως κανονικά (για να είμαστε πιο ακριβείς θα πρέπει να προχωρήσουμε και σε έλεγχο με δεδομένο ε.σ. a, π.χ. χ^2 ή K-S που θα εξετάσουμε παρακάτω). Αξίζει να παρατηρήσουμε ότι αν υπήρχαν κάποιες έκτροπες παρατηρήσεις (παρατηρήσεις αρκετά «απομακρυσμένες» από την διαγώνιο) τότε θα έπρεπε αυτές να επανεξεταστούν λεπτομερέστερα ώστε να βεβαιωθούμε ότι δεν ισχύουν κάποιες ειδικές συνθήκες για αυτές ή ότι δεν έχουν περαστεί λάθος στο SPSS.

Το Q-Q plot λαμβάνεται με παρόμοιο τρόπο: Ανοίγουμε την ανάλυση Graphs/Q-Q και επιλέγοντας την μεταβλητή chol λαμβάνουμε το γράφημα στα δεξιά:





Το εμπειρικό αυτό τεστ είναι ισοδύναμο με το προηγούμενο και επομένως δεν περιμένουμε να δούμε κάτι διαφορετικό. Και εδώ φαίνεται ότι τα 60 σημεία του επιπέδου δεν «απέχουν» πολύ από την διαγώνιο και επομένως δεν υπάρχει επαρκής λόγος ώστε απορρίψουμε ότι τα δεδομένα είναι κανονικά.

3.2. Ο έλεγχος χ^2 (χι-τετράγωνο) καλής προσαρμογής

Επιθυμούμε και πάλι να ελέγξουμε αν κάποιες παρατηρήσεις ενός τ.δ. $X_1, X_2, \ldots, X_n$ προέρχονται από μια συγκεκριμένη κατανομή με σ.κ. F_0. Ο Pearson, ήδη από τις αρχές του προηγούμενου αιώνα (1900), πρότεινε για το σκοπό αυτό τη χρήση μιας στατιστικής συνάρτησης η οποία, υπό την H_0: $X_i \sim F_0$, ακολουθεί (προσεγγιστικά) κατανομή χ^2 (με κάποιους β.ε.) ενώ όταν δεν ισχύει η H_0 λαμβάνει «μεγάλες» τιμές. Πριν δούμε ποια είναι η μορφή αυτής της στατιστικής συνάρτησης στο συγκεκριμένο πρόβλημα, αξίζει να θυμηθούμε ένα σημαντικό θεωρητικό αποτέλεσμα το οποίο αφορά την πολυωνυμική κατανομή και αποτελεί την βάση του χ^2 ελέγχου καλής προσαρμογής. Επίσης αποτελεί την βάση και για άλλους ελέγχους που θα εξετάσουμε σε επόμενες ενότητες (π.χ. χ^2 έλεγχοι για πίνακες συνάφειας).

Πρόταση. *Αν το τυχαίο διάνυσμα* $\mathbf{N} = (N_1,N_2,...,N_k)$ *ακολουθεί πολυωνυμική*[1] *κατανομή με παραμέτρους* n *και* $p_1,p_2,...,p_k$ (με $\sum_{i=1}^{k} p_i = 1$) *τότε η στατιστική συνάρτηση*

$$T = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i},$$

ακολουθεί ασυμπτωτικά ($n \to \infty$) *κατανομή* χ^2_{k-1} (*χι-τετράγωνο με* $k-1$ *βαθμούς ελευθερίας*).

Έστω τώρα $X_1,X_2,\ldots,X_n$ ένα τυχαίο δείγμα και έστω ότι επιθυμούμε να ελέγξουμε την H_0: $X_i \sim F_0$. Προκειμένου να χρησιμοποιήσουμε το αποτέλεσμα της παραπάνω πρότασης εργαζόμαστε ως εξής: διαμερίζουμε το πεδίο τιμών των X_i (υπό την H_0) σε k σύνολα $A_1, A_2, \ldots, A_k$ (συνήθως έτσι ώστε στο κάθε σύνολο να αναμένονται τουλάχιστον 5 παρατηρήσεις). Στη συνέχεια θεωρούμε τις τ.μ.

$$N_i = \text{πλήθος των } X_1,X_2,\ldots,X_n \text{ που ανήκουν στο σύνολο } A_i,\ i = 1,2,\ldots,n.$$

Είναι προφανές ότι όταν ισχύει η υπόθεση H_0: $X_i \sim F_0$ τότε το τυχαίο διάνυσμα $(N_1,N_2,\ldots,N_k)$ ακολουθεί πολυωνυμική κατανομή με παραμέτρους n και $p_1,p_2,...,p_k$ όπου

$$p_i = P(X_1 \in A_i / H_0 : X_i \sim F_0),\ i = 1,2,...,n.$$

Επομένως, υπό την H_0, η στατιστική συνάρτηση

[1] Η πολυωνυμική κατανομή είναι η από κοινού κατανομή του πλήθους των επιτυχιών $1^{ου}$-είδους, $2^{ου}$-είδους,...,k-είδους σε μία ακολουθία n ανεξάρτητων και ισόνομων δοκιμών με k δυνατά είδη επιτυχιών (πιθ. επιτ. i-είδους $= p_i$)

$$T(\mathbf{X}) = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$$

ακολουθεί προσεγγιστικά κατανομή χ^2 με $k-1$ β.ε. ενώ υπό την H_1: $X_i \sim G \neq F_0$ θα λαμβάνει «μεγάλες» τιμές. Το τελευταίο συμβαίνει διότι, το np_i είναι το αναμενόμενο πλήθος παρατηρήσεων στο A_i υπό την H_0 ($E(N_i) = np_i = nP(X_1 \in A_i \ / \ H_0)$) και επομένως όταν δεν ισχύει η H_0 κάθε N_i (παρατηρούμενη συχνότητα) θα διαφέρει αρκετά από το np_i (αναμενόμενη συχνότητα υπό την H_0).

Άρα, με βάση την παραπάνω στατιστική συνάρτηση μπορούμε να κατασκευάσουμε έναν έλεγχο για την υπόθεση H_0: $X_i \sim F_0$. Συγκεκριμένα θα απορρίπτουμε την H_0 (σε ε.σ. a περίπου) όταν, με βάση τις παρατηρήσεις $x_1, x_2, \ldots, x_n$,

$$T(\mathbf{x}) > c = \chi^2_{k-1}(a) \ : \textit{άνω a-σημείο της} \ \chi^2_{k-1}$$

με αντίστοιχο (προσεγγιστικό) p-value

$$p - value = P(T(\mathbf{X}) > T(\mathbf{x})) = 1 - F_{\chi^2_{k-1}}(T(\mathbf{x})).$$

Παρατήρηση 1. (*έλεγχος χ^2 όταν υπάρχουν άγνωστες παράμετροι*). Παραπάνω προφανώς θεωρήσαμε ότι τα p_i είναι γνωστά (καθορίζονται πλήρως από την κατανομή F_0). Υπάρχουν όμως περιπτώσεις όπου τα p_i δεν είναι απολύτως γνωστά, αλλά εξαρτώνται από κάποιες άγνωστες παραμέτρους, δηλαδή $p_i = p_i(\boldsymbol{\theta})$ με $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_r)$ άγνωστο. Η περίπτωση αυτή εμφανίζεται π.χ. κατά τον έλεγχο καλής προσαρμογής δεδομένων σε μία γνωστή κατανομή (π.χ. κανονική) με άγνωστες όμως παραμέτρους (π.χ. μ, σ, δηλ. $p_i = p_i(\mu, \sigma)$) ή π.χ. κατά τον έλεγχο ανεξαρτησίας σε πίνακες συνάφειας (χρησιμοποιώντας το χι-τετράγωνο τεστ). Στην περίπτωση αυτή χρησιμοποιούμε την τροποποιημένη στατιστική συνάρτηση

$$T'(\mathbf{X}) = \sum_{i=1}^{k} \frac{(N_i - np_i(\hat{\boldsymbol{\theta}}))^2}{np_i(\hat{\boldsymbol{\theta}})},$$

όπου $\hat{\boldsymbol{\theta}}$ είναι η εκτίμηση του $\boldsymbol{\theta}$ από τα δεδομένα. Τώρα, υπό την H_0, αποδεικνύεται ότι η T' ακολουθεί ασυμπτωτικά χι-τετράγωνο κατανομή με $k - r - 1$ βαθμούς ελευθερίας, όπου r είναι το πλήθος των παραμέτρων που χρειάστηκε να εκτιμηθούν από τα δεδομένα (αρκεί να χρησιμοποιηθούν οι εκτιμήτριες μέγιστης πιθανοφάνειας των παραμέτρων από τα ομαδοποιημένα στις k κλάσεις δεδομένα). Επομένως τώρα, απορρίπτουμε την H_0 σε ε.σ. a (περίπου) όταν $T'(\mathbf{x}) > \chi^2_{k-1-r}(a)$ με αντίστοιχο (προσεγγιστικό) p-value

$$p - value \approx P(T'(\mathbf{X}) \geq T'(\mathbf{x}) \mid H_0) = 1 - F_{\chi^2_{k-r-1}}(T'(\mathbf{x})).$$

Παρατήρηση 2. Ο έλεγχος χ^2 τις περισσότερες φορές δεν είναι ο καλύτερος έλεγχος καλής προσαρμογής για συνεχή δεδομένα διότι προϋποθέτει ομαδοποίηση των δεδομένων (διαμερίζουμε το πεδίο τιμών των παρατηρήσεων σε k σύνολα $A_1, A_2, \ldots, A_k$) με συνέπεια την απώλεια πληροφορίας (επίσης η διαμέριση είναι τις περισσότερες φορές αυθαίρετη). Σε αυτήν την περίπτωση (δεδομένα από συνεχή κατανομή) συνήθως προτιμάται ο έλεγχος Kolmogorov-Smirnov (K-S) ο οποίος βασίζεται στην εμπειρική συνάρτηση κατανομής του δείγματος και δεν προϋποθέτει κάποια ομαδοποίηση των δεδομένων. Ο έλεγχος χ^2 προτιμάται όταν έχουμε κατηγορικά δεδομένα που παίρνουν τιμές σε ένα πεπερασμένο σύνολο (βλ. εφαρμογή 2 παρακάτω).

Για τους παραπάνω λόγους το SPSS δεν δίνει μεγάλο βάρος στους ελέγχους καλής προσαρμογής μέσω του χ^2 τεστ. Συγκεκριμένα, με το SPSS είναι δυνατός μόνο ο έλεγχος προσαρμογής κάποιων παρατηρήσεων σε μια διακριτή κατανομή η οποία λαμβάνει k τιμές, ενώ οι πιθανότητες p_i, $i = 1, 2, \ldots, k$ θα πρέπει να δοθούν από τον χρήστη του προγράμματος. Παρόλα αυτά μπορούμε με έμμεσο τρόπο να πραγματοποιήσουμε χ^2 έλεγχο καλής προσαρμογής για οποιαδήποτε διακριτή (βλ.

εφαρμ. 4) ή συνεχή κατανομή (βλ. εφαρμογές 3, 5) αφού κάνουμε μόνοι μας ομαδοποίηση και υπολογισμό των p_i (π.χ. χρησιμοποιώντας τις εντολές compute ή recode του SPSS).

Εφαρμογή 2. Ρίχνοντας ένα ζάρι 120 φορές καταγράφουμε τα εξής αποτελέσματα:

το 1 εμφανίστηκε 18 φορές, το 2 εμφανίστηκε 22 φορές, το 3 εμφανίστηκε 30 φορές
το 4 εμφανίστηκε 21 φορές, το 5 εμφανίστηκε 17 φορές, το 6 εμφανίστηκε 12 φορές

Να ελέγξετε (ε.σ. 5%) αν το ζάρι αυτό είναι αμερόληπτο.

Αρχικά εισάγουμε τα δεδομένα στο SPSS. Κανονικά θα πρέπει να εισάγουμε 120 αποτελέσματα, τα 18 από τα οποία θα είναι 1, τα 22 επόμενα να είναι 2 κ.ο.κ. (δηλ 120 cases – γραμμές με μία μεταβλητή – στήλη). Έχουμε δει όμως ότι σε τέτοιες περιπτώσεις (όπου έχουμε επαναλήψεις γραμμών) είναι ισοδύναμο αλλά αρκετά βολικότερο να χρησιμοποιούμε βάρη. Εισάγουμε επομένως μία μεταβλητή apot με τα αποτελέσματα 1, 2, 3, 4, 5, 6 και μία άλλη μεταβλητή w (βάρη) τις αντίστοιχες εμφανίσεις 18, 22, 30, 21, 17, 12 και επιλέγουμε Data/weight cases/weight cases by w. Για το χ^2 τεστ θα χρησιμοποιήσουμε 6 κλάσεις, τις προφανείς: $A_1 = \{1\}$, $A_2 = \{2\}$, …, $A_6 = \{6\}$.

Στη συνέχεια επιλέγουμε Analyze/Non parametric tests/chi-square/test variable list: apot. Επίσης θα πρέπει να εισάγουμε τις αναμενόμενες πιθανότητες p_1, p_2, …, p_k (εδώ k=6) υπό την H_0 στο πεδίο expected values. Επειδή όπως εδώ $p_1 = p_2 = \ldots = p_6 = 1/6$ (H_0: αμερόληπτο ζάρι) μπορούμε πολύ απλά να επιλέξουμε να είναι all categories equal (είναι η default επιλογή). Από την ανάλυση αυτή λαμβάνουμε τους πίνακες

APOT

	Observed N	Expected N	Residual
1	18	20,0	-2,0
2	22	20,0	2,0
3	30	20,0	10,0
4	21	20,0	1,0
5	17	20,0	-3,0
6	12	20,0	-8,0
Total	120		

Test Statistics

	RESULTS
Chi-Square	9,100
df	5
Asymp. Sig.	,105

0 cells (,0%) have expected frequencies less than 5.
The minimum expected cell frequency is 20,0.

Ο πρώτος πίνακας απεικονίζει τις παρατηρούμενες και τις αναμενόμενες συχνότητες σε κάθε ένα από τα 6 σύνολα $A_1=\{1\},\ldots,A_6=\{6\}$ (κατηγορίες ή κελιά) ενώ ο δεύτερος πίνακας δίνει την τιμή της στατιστικής συνάρτησης δείγμα, $T(\mathbf{x}) = 9.1$ (β.ε. = k–1 =5) και το αντίστοιχο p-value = 0.105. Το p-value δεν είναι μικρότερο του a = 5% οπότε, με βάση τις 120 αυτές παρατηρήσεις, δεν μπορούμε να απορρίψουμε ότι το ζάρι είναι αμερόληπτο.

Εφαρμογή 3. Να ελέγξετε (ε.σ. 5%) αν οι παρακάτω 45 παρατηρήσεις προέρχονται από την ομοιόμορφη κατανομή στο (0,1).

0.00	0.06	0.38	0.38	0.56	0.44	0.44	0.60	0.77
0.17	0.18	0.38	0.34	0.46	0.44	0.53	0.75	0.72
0.14	0.04	0.28	0.32	0.43	0.42	0.56	0.61	0.98
0.04	0.20	0.36	0.54	0.48	0.40	0.79	0.66	0.85
0.16	0.26	0.22	0.51	0.52	0.43	0.71	0.78	0.86

Εισάγουμε τα δεδομένα στο SPSS: 45 cases (γραμμές) με μία μεταβλητή (στήλη) με όνομα data (στο SPSS μερικές φορές, π.χ. σε ελληνικά windows, ως υποδιαστολή θεωρείται το «,» αντί του «.»). Τώρα τα δεδομένα δεν μπορούν να θεωρηθούν κατηγορικά (όπως στην προηγ. εφαρμογή) ώστε να εφαρμόσουμε απευθείας το χ^2 τεστ, αλλά θα πρέπει να τα «ομαδοποιήσουμε». Για να έχουμε τουλάχιστον 5 αναμενόμενες παρατηρήσεις σε κάθε κελί θα χρησιμοποιήσουμε 8 κλάσεις, τις

$$A_1 = [0,1/8),\ A_2 = [1/8,2/8),\ \ldots,\ A_8 = [7/8,1).$$

Κατασκευάζουμε μία νέα μεταβλητή η οποία δείχνει τις κλάσεις με Transform/compute: categ = Trunc(data*8) και εφαρμόζουμε το χ^2 τεστ σε αυτή την μεταβλητή: επιλέγουμε Analyze/Non parametric tests/chi-square/test variable list: categ. Όπως και στην προηγούμενη εφαρμογή, οι αναμενόμενες πιθανότητες $p_1, p_2, \ldots, p_k$ (εδώ $k = 8$) υπό την H_0: $X_i \sim$ *ομοιόμορφη κατανομή* (0,1) είναι όλες ίσες (με 1/8 διότι και τα A_i έχουν πλάτος 1/8). Και έτσι μπορούμε πολύ απλά να επιλέξουμε all categories equal. Λαμβάνουμε τους πίνακες

CATEG

	Observed N	Expected N	Residual
0	4	5,625	-1,625
1	6	5,625	,375
2	5	5,625	-,625
3	12	5,625	6,375
4	8	5,625	2,375
5	3	5,625	-2,625
6	6	5,625	,375
7	1	5,625	-4,625
Total	45		

Test Statistics

	CAT
Chi-Square	13,844
df	7
Asymp. Sig.	,054

0 cells (,0%) have expected frequencies less than 5.
The minimum expected cell frequency is 5,6.

Η τιμή της στατιστικής συνάρτησης δείγμα είναι 18.844 με αντίστοιχο p-value = 0.0504. Το p-value δεν είναι μικρότερο του a = 5% οπότε δεν μπορούμε να απορρίψουμε ότι τα δεδομένα προέρχονται από την ομοιόμορφη. Όπως έχει αναφερθεί και παραπάνω, σε αυτή την περίπτωση (συνεχής κατανομή) είναι ίσως προτιμότερο να κάνουμε τεστ καλής προσαρμογής χρησιμοποιώντας το Kolmogorov –Smirnov τεστ που θα εξετάσουμε σε επόμενη παράγραφο.

Εφαρμογή 4. Να ελέγξετε αν τα παρακάτω 30 δεδομένα προέρχονται από την κατανομή Poisson με λ = 3 (ε.σ. 5%).

2	1	2	1	5	3	4	1	4	4
1	0	2	2	2	3	2	2	3	3
3	4	3	1	5	2	0	3	7	1

Πως θα ελέγχαμε αν τα δεδομένα προέρχονται από την Poisson (με άγνωστο λ);

Εισάγουμε τα δεδομένα στο SPSS: 30 cases (γραμμές) με μία μεταβλητή (στήλη) με όνομα data. Σε αυτή την περίπτωση τα δεδομένα μπορούν μεν να θεωρηθούν κατηγορικά αλλά το πλήθος των κατηγοριών δεν είναι πεπερασμένο (μια τ.μ. που ακολουθεί την Poisson μπορεί να πάρει τιμές στο {0,1,2,…}). Για το λόγο αυτό θα πρέπει να «ενώσουμε» κάποια δυνατά αποτελέσματα σε κλάσεις (έτσι ώστε και οι αναμενόμενες συχνότητες σε αυτές τις κλάσεις να είναι τουλάχιστον 5). Παρατηρούμε ότι αν $X \sim$ Poisson(λ = 3) θα είναι

$$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!},\ x = 0,1,2,\ldots$$

και επομένως,

x	$\approx P(X \le x)$	$\approx P(X = x)$	$\approx n\,P(X = x)$
0	0.0498	0.0498	1.49
1	0.1991	0.1494	4.48
2	0.4232	0.2240	6.72
3	0.6472	0.2240	6.72
4	0.8153	0.1680	5.04
5	0.9161	0.1008	3.02
6	0.9665	0.0504	1.51

Ο παραπάνω πίνακας μπορεί πολύ εύκολα να κατασκευασθεί χρησιμοποιώντας το SPSS. Αρχικά φτιάχνουμε μόνοι μας την μεταβλητή x (με τιμές 0,1,2,3,4,5,6). Στη συνέχεια κατασκευάζουμε την αθροιστική συνάρτηση κατανομής $P(X \le x)$ της Poisson(λ=3) μέσω της Transform/compute: cdfP = CDF.POISSON(x,3). Η επόμενη στήλη μπορεί τώρα να ληφθεί από την cdfP μέσω της εντολής Transform/create time series εισάγοντας στο πεδίο new variable την cdfP με Function:difference (order 1). Η διαδικασία αυτή κατασκευάζει μια νέα μεταβλητή, την cdf_1 η οποία απαρτίζεται από τις διαφορες των διαδοχικών τιμών της cdfP δηλ. τις $P(X = x)$. Τέλος, η στήλη με τις τιμές $nP(X = x)$ μπορεί να ληφθεί από την Transform/compute, π.χ. expval = 30*cdf_1.

Από τον παραπάνω πίνακα παρατηρούμε ότι μπορούμε να χρησιμοποιήσουμε τα σύνολα (κλάσεις ή κατηγορίες) $A_1=\{0,1\}$, $A_2=\{2\}$, $A_3=\{3\}$, $A_4=\{4\}$, $A_5=\{5,6,\ldots\}$ (για να έχουμε αναμενόμενες συχνότητες σε όλες τις κλάσεις τουλάχιστον 5) με αντίστοιχες αναμενόμενες πιθανότητες p_i:

$$p_1 = P(X \le 1) \approx 0.1991,\ p_2 = P(X = 2) \approx 0.2240,\ p_3 = P(X = 3) \approx 0.2240,$$
$$p_4 = P(X = 4) \approx 0.1680,\ p_5 = P(X \ge 5) \approx 1 - 0.8153 = 0.1847$$

Στη συνέχεια θα πρέπει να κατασκευάζουμε μία νέα μεταβλητή categ η οποία δείχνει τις παραπάνω κλάσεις (μπορεί να γίνει με τον γνωστό τρόπο χρησιμοποιώντας Transform/recode) και εφαρμόζουμε το χ^2 τεστ σε αυτή την μεταβλητή. Επιλέγουμε Analyze/Non parametric tests/chi-square/test variable list: categ. Σε αυτή όμως την περίπτωση δεν επιλέγουμε all categories equal αλλά περνάμε τις παραπάνω αναμενόμενες πιθανότητες 0.199, 1.224, 0.224, 0.168, 1.185 (τις εισάγουμε με add με αυτή την σειρά). Λαμβάνουμε τους πίνακες

CATEG

	Observed N	Expected N	Residual
1	8	6,0	2,0
2	8	6,7	1,3
3	7	6,7	,3
4	4	5,0	-1,0
5	3	5,5	-2,5
Total	30		

Test Statistics

	CATEG
Chi-Square	2,332
df	4
Asymp. Sig.	,675

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

Η τιμή της στατιστικής συνάρτησης στο δείγμα είναι 2.332 (4 β.ε.) με αντίστοιχο p-value = 0.675. Άρα δεν μπορούμε να απορρίψουμε ότι τα δεδομένα προέρχονται από την Poisson (λ = 3).

Τέλος, εάν έπρεπε να ελέγξουμε αν τα δεδομένα προέρχονται από κάποια Poisson (δηλ, Poisson με άγνωστο λ) τότε σύμφωνα με την Παρατήρηση 1 της Παραγράφου 3.2 θα έπρεπε να κάνουμε όλα τα παραπάνω αυτή τη φορά χρησιμοποιώντας την εκτίμηση του λ από το δείγμα και όχι το λ = 3. Σε αυτή την περίπτωση χάνουμε έναν β.ε. (διότι κάναμε εκτίμηση μιας παραμέτρου) και θα πρέπει να βρούμε μόνοι μας το p-value (το SPSS θα μας δώσει την τιμή της στατιστικής συνάρτησης *chi-square* αλλά το p-value που θα λάβουμε θα αντιστοιχεί σε 4 β.ε. και επομένως θα είναι μεγαλύτερο του p-value που αντιστοιχεί σε 3 β.ε.). Το p-value μπορεί να υπολογιστεί π.χ. από το compute : pvalue = 1 – CDF.CHISQ(*chi-square*,3).

Εφαρμογή 5. Να ελέγξετε αν οι παρακάτω παρατηρήσεις προέρχονται από την κανονική κατανομή με μέση τιμή 100 και διασπορά 65 (ε.σ. 5%).

98.47	95.53	106.67	106.01	94.84	116.79	103.59	99.57
94.56	111.66	104.08	92.50	103.95	102.77	107.09	115.08
104.14	76.89	98.34	98.20	95.70	99.28	104.59	101.44
87.50	97.59	117.26	98.91	96.79	91.80	107.68	107.31
98.82	100.99	105.91	97.69	109.03	104.37	91.92	107.63

Όπως και στην εφαρμογή 3, θα ήταν ίσως προτιμότερο να κάνουμε το τεστ καλής προσαρμογής χρησιμοποιώντας το Kolmogorov –Smirnov τεστ που θα εξετάσουμε σε επόμενη παράγραφο. Είναι ενδιαφέρον όμως να δούμε πως μπορούμε να χρησιμοποιήσουμε το χ^2 τεστ για έλεγχο καλής προσαρμογής. Προφανώς θα πρέπει και πάλι να ομαδοποιήσουμε τα δεδομένα για να χρησιμοποιήσουμε το χ^2 τεστ. Η ομαδοποίηση θα πρέπει να γίνει έτσι ώστε σε κάθε κλάση η αναμενόμενη συχνότητα να είναι ≥ 5. Μπορούμε να ορίσουμε μόνοι μας τις κλάσεις, να δημιουργήσουμε μια νέα μεταβλητή που να δείχνει τις κλάσεις και στην συνέχεια να υπολογίσουμε τις αναμενόμενες πιθανότητες p_i και να τις εισάγουμε στην ανάλυση του SPSS (όπως στην Εφαρμογή 4). Για να γλιτώσουμε όμως τον υπολογισμό των p_i (κάτι όχι τόσο εύκολο) μπορούμε να κάνουμε κάτι απλούστερο. Αντί να ελέγξουμε αν οι παραπάνω παρατηρήσεις $X_1,X_2,\ldots,X_n$ προέρχονται από την N(100,65), μπορούμε ισοδύναμα να ελέγξουμε αν οι μετασχηματισμένες παρατηρήσεις $Y_1 = F(X_1)$, $Y_2 = F(X_2)$, … $Y_n = F(X_n)$ προέρχονται από την ομοιόμορφη στο (0,1), όπου F είναι η σ.κ. της N(100,65) (στην παρ. 3.1. είδαμε ότι αν $X \sim F$:συνεχής τότε η τ.μ. $Y = F(X) \sim$ Ομοιόμορφη στο (0,1)). Τις νέες παρατηρήσεις Y_1, Y_2, …, Y_n μπορούμε να τις κατασκευάσουμε χρησιμοποιώντας την εντολή Transform / compute Y = CDF.NORMAL(X,100,SQRT(65)). Ο έλεγχος αν οι τ.μ. Y_1, Y_2, …, Y_n προέρχονται από την ομοιόμορφη στο (0,1) γίνεται όμοια με την Εφαρμογή 3. Αν π.χ. χρησιμοποιήσουμε 8 κλάσεις βρίσκουμε p-value = 0.189 και επομένως δεν απορρίπτουμε ότι τα δεδομένα προέρχονται από την N(100,65).

3.3. Το κριτήριο Kolmogorov-Smirnov (K-S) για ένα δείγμα

Το κριτήριο K-S χρησιμοποιείται και αυτό για το έλεγχο καλής προσαρμογής ενός τυχαίου δείγματος σε μία δεδομένη *συνεχή* κατανομή (H_0: $X_i \sim F_0$). Το κριτήριο K-S βασίζεται στην διαφορά της εμπειρικής συνάρτηση κατανομής (που προέρχεται από το δείγμα) και της αναμενόμενης F_0 (υπό την H_0). Πιο συγκεκριμένα, αν $X_1,X_2,...,X_n$ είναι ένα τ.δ., η εμπειρική συνάρτηση κατανομής (ΕΣΚ) του δείγματος αυτού είναι

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x) = \frac{\#\{X_i \le x\}}{n},$$

(όπου $I(X_i \le x) = 1$ ή 0 ανάλογα με το αν $X_i \le x$ ή όχι) η οποία ως γνωστό αποτελεί εκτίμηση της συνάρτησης κατανομής των X_i διότι (από το νόμο των μεγάλων αριθμών, θέτοντας $Y_i = I(X_i \le x)$)

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x) = \frac{1}{n}\sum_{i=1}^{n} Y_i \underset{n\to\infty}{\to} E(Y_1) = 0P(Y_1 = 0) + 1P(Y_1 = 1) = P(Y_1 = 1) = P(X_1 \le x) = F(x)$$

για κάθε x. Επομένως, υπό την H_0, η ΕΣΚ θα πρέπει να είναι «κοντά» στην F_0. Αντίθετα, αν δεν ισχύει η H_0 αναμένουμε σημαντική απόκλιση της ΕΣΚ από την F_0. Για να κατασκευάσουμε έναν έλεγχο με βάση αυτόν τον συλλογισμό, θα πρέπει να ορίσουμε μία «απόσταση» μεταξύ των δύο κατανομών (της ΕΣΚ και της F_0) και να απορρίπτουμε την H_0 όταν αυτή η απόσταση γίνεται «μεγάλη». Σχετικά έχουμε τον επόμενο ορισμό.

Ορισμός. *Αν F, G είναι δύο συναρτήσεις κατανομής στον R, τότε η ποσότητα*

$$d_K(F,G) = \sup_{x\in R}\{|F(x) - G(x)|\}$$

καλείται απόσταση Kolmogorov μεταξύ της F και της G.

Σύμφωνα με τα παραπάνω, θα απορρίπτουμε την H_0: $X_i \sim F_0$ όταν η στατιστική συνάρτηση

$$D_n = d_K(\hat{F}_n, F_0) = \sup_{x\in R}\{|\hat{F}_n(x) - F_0(x)|\},$$

λαμβάνει «ασυνήθιστα» μεγάλες τιμές, δηλαδή όταν $D_n > c$. Το κριτήριο αυτό είναι γνωστό ως κριτήριο Kolmogorov – Smirnov (και η στατιστική συνάρτηση D_n καλείται ελεγχοσυνάρτηση Kolmogorov – Smirnov). Προκειμένου να χρησιμοποιήσουμε το συγκεκριμένο κριτήριο θα πρέπει

να προσδιορίσουμε την κατανομή της τ.μ. D_n κάτω από την H_0 έτσι ώστε να υπολογίσουμε το c (για δεδομένο επίπεδο σημαντικότητας a) και το *p-value* ενός δείγματος.

Σε αυτό το σημείο ίσως κάποιος αναλογιστεί ότι το κριτήριο αυτό έχει ένα σοβαρό μειονέκτημα: η κατανομή της D_n θα πρέπει να εξαρτάται από την F_0 (την κατανομή από την οποία προέρχεται το δείγμα, υπό την H_0) και επομένως θα πρέπει να βρούμε την κατανομή της D_n για κάθε διαφορετική κατανομή F_0. Ευτυχώς, αντίθετα με αυτό που θα περίμενε κανείς, αποδεικνύεται ότι η κατανομή της στατιστικής συνάρτησης D_n *δεν* εξαρτάται από την F_0! Το γεγονός αυτό μας δίνει την δυνατότητα να χρησιμοποιήσουμε το κριτήριο αυτό οποιαδήποτε και αν είναι η κατανομή από την οποία προέρχεται το δείγμα (υπό την H_0). Τέτοιοι έλεγχοι καλούνται απαραμετρικοί έλεγχοι (η κατανομή της στατιστικής συνάρτησης που χρησιμοποιούμε και επομένως η κρίσιμη περιοχή και το p-value δεν εξαρτώνται από την κατανομή του δείγματος υπό την H_0). Το χ^2 τεστ, το Kolmogorov – Smirnov τεστ καθώς και τα τεστ που θα εξετάσουμε στη συνέχεια στην ενότητα αυτή είναι απαραμετρικά.

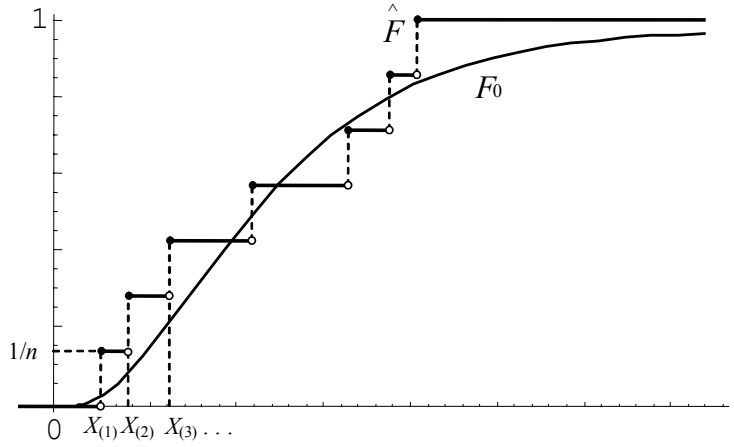
Πριν προχωρήσουμε, έχει ενδιαφέρον να δούμε γιατί η κατανομή του D_n δεν εξαρτάται από την F_0. Ξεκινάμε αναζητώντας μία απλούστερη έκφραση της τ.μ. D_n ώστε να υπολογίζεται εύκολα από το τ.δ. $X_1, X_2, ..., X_n$ αλλά και να φαίνεται αμεσότερα η εξάρτησή της από τα X_i. Έστω $X_{(1)}, X_{(2)}, ...,X_{(n)}$ οι διατεταγμένες τιμές των $X_1, X_2, ...,X_n$ ($X_{(1)} < X_{(2)} < \ ... < X_{(n)}$). Παρατηρούμε ότι η εμπειρική συνάρτηση κατανομής γράφεται ως εξής:

$$\hat{F}_n(x) = \begin{cases} 0, & x < X_{(1)} \\ 1/n, & X_{(1)} \le x < X_{(2)} \\ 2/n, & X_{(2)} \le x < X_{(3)} \\ \vdots & \\ n/n, & X_{(n)} \le x \end{cases}$$

δηλαδή είναι σταθερή στα διαστήματα $[X_{(i-1)}, X_{(i)})$ ενώ παρουσιάζει άλματα ύψους $1/n$ στα σημεία $X_{(1)}, ...,X_{(n)}$. Εφόσον τώρα η F_0 είναι αύξουσα συνάρτηση, η μέγιστη τιμή της $\hat{F}_n(x) - F_0(x)$ θα λαμβάνεται πάνω σε κάποιο από τα σημεία $X_{(1)}, ..., X_{(n)}$. Δηλαδή,

$$D_n^+ = \sup_{x \in R} \{\hat{F}_n(x) - F_0(x)\} = \max_{i=1,2,...,n} \{\hat{F}_n(X_{(i)}) - F_0(X_{(i)})\} = \max_{i=1,2,...,n} \{\frac{i}{n} - F_0(X_{(i)})\} \ge 0 .$$

Αυτό μπορεί να φανεί και από το παρακάτω σχήμα ($n = 7$):



Όμοια, το supremum της $F_0(x) - \hat{F}_n(x)$ θα είναι

$$D_n^- = \sup_{x \in R} \{F_0(x) - \hat{F}_n(x)\} = \max_{i=1,2,...,n} \{F_0(X_{(i)}) - \hat{F}_n(X_{(i)}^-)\} = \max_{i=1,2,...,n} \{F_0(X_{(i)}) - \frac{i-1}{n}\} \ge 0 .$$

(όπου $F(x^-) = \lim_{t\uparrow x} F(t)$) και τελικά,

$$D_n = \sup_{x\in R}\{|\hat{F}_n(x) - F_0(x)|\} = \max\left\{D_n^+, D_n^-\right\} = \max\left\{\frac{i}{n} - F_0(X_{(i)}),\ F_0(X_{(i)}) - \frac{i-1}{n},\ i = 1,2,...,n\right\}.$$

Παρατηρούμε τώρα ότι οι τ.μ. $U_i = F_0(X_i)$, i=1,2,…,n είναι ανεξάρτητες και ακολουθούν την ομοιόμορφη στο (0,1) κατανομή (βλ. και παρ. 3.1) και επομένως οι τ.μ. $U_{(i)} = F_0(X_{(i)})$ μπορεί να θεωρηθεί ότι αποτελούν ένα διατεταγμένο δείγμα από την ομοιόμορφη στο (0,1) κατανομή. Συνεπώς, οποιαδήποτε και αν είναι η F_0, η D_n έχει ίδια κατανομή (υπό την H_0) με την τ.μ.

$$\max\left\{\frac{i}{n} - U_{(i)},\ U_{(i)} - \frac{i-1}{n},\ i = 1,2,...,n\right\},$$

όπου $U_1,U_2,\ldots,U_n$ είναι ανεξάρτητες τ.μ. από την U(0,1) η οποία προφανώς δεν εξαρτάται από την F_0. Επομένως, θα απορρίπτουμε την H_0 όταν

$$D_n > c = D_n(a),$$

όπου $D_n(a)$ είναι το άνω a-σημείο της κατανομής της τ.μ. D_n (το οποίο δεν εξαρτάται από την F_0). Η ακριβής κατανομή της τ.μ. D_n είναι δύσκολο να υπολογιστεί και για αυτό έχουν κατασκευαστεί πίνακες με τα άνω a-σημεία της. Αποδεικνύεται όμως ότι η κατανομή της τ.μ. $Z_n = \sqrt{n}\cdot D_n$ (καλείται και Kolmogorov-Smirnov Z) έχει ασυμπτωτικά (υπό την H_0 και για συνεχή σ.κ. F_0) τη συνάρτηση κατανομής,

$$P(Z_n \le z) = P(\sqrt{n}\cdot D_n \le z) \to_{n\to\infty} 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1} e^{-2i^2z^2} \quad \text{για κάθε } z \ge 0,$$

και επομένως το *p-value* ενός δείγματος που έδωσε $D_n = d$ θα είναι (ασυμπτωτικά)

$$p-value = P(D_n > d / H_0) = 1 - P(\sqrt{n}\cdot D_n < \sqrt{n}\cdot d) \approx 2\sum_{i=1}^{\infty}(-1)^{i-1} e^{-2i^2nd^2}.$$

Παρατήρηση. Παραπάνω εξετάσαμε τον έλεγχο της υπόθεσης H_0: $X_i \sim F_0$ όπου η F_0 ήταν πλήρως καθορισμένη. Συνηθέστερη όμως περίπτωση είναι να γνωρίζουμε την οικογένεια στην οποία ανήκει η F_0 με άγνωστες όμως παραμέτρους $\boldsymbol{\theta}$ (π.χ. κανονική με άγνωστα μ, σ). Στην περίπτωση αυτή συνήθως εκτιμούμε τις παραμέτρους $\boldsymbol{\theta}$ από τα δεδομένα και χρησιμοποιούμε την ίδια στατιστική συνάρτηση

$$D_n(\hat{\boldsymbol{\theta}}) = \sup_{x\in R}\{|\hat{F}_n(x) - F_0(x;\hat{\boldsymbol{\theta}})|\},$$

όπου $F_0(x;\hat{\boldsymbol{\theta}})$ είναι η σ.κ. που προκύπτει αν θεωρήσουμε ότι οι άγνωστες παράμετροι της F_0 έχουν εκτιμηθεί από τα δεδομένα[2]. Το αντίστοιχο *p-value* είναι περίπου ίσο (για μεγάλα δείγματα) με αυτό που θα προέκυπτε αγνοώντας το γεγονός της εκτίμησης του $\boldsymbol{\theta}$, δηλαδή,

$$p-value = \Pr(D_n(\hat{\boldsymbol{\theta}}) \ge d / H_0) \approx P(D_n \ge d / H_0),$$

και έτσι μπορούμε να χρησιμοποιήσουμε και πάλι την ασυμπτωτική κατανομή της D_n[3].

---

[2] Για ορισμένες κατανομές (π.χ. κανονική με άγνωστες παραμέτρους) συνήθως χρησιμοποιείται μία τροποποίηση του K-S τεστ (π.χ. Lilliefors K-S).

[3] Στην πραγματικότητα, η συγκεκριμένη προσεγγιστική τιμή του *p-value* είναι μεγαλύτερη από την αντίστοιχη ακριβή τιμή. Αυτό είναι διαισθητικά προφανές, διότι η $F_0(x;\hat{\boldsymbol{\theta}})$ θα ταιριάζει περισσότερο στα δεδομένα από την $F_0(x;\boldsymbol{\theta})$ (ακριβώς διότι χρησιμοποιούμε τις παραμέτρους που ταιριάζουν περισσότερο στα δεδομένα) με αποτέλεσμα να λαμβάνουμε μεγαλύτερο *p-value*.

Εφαρμογή 6. Να ελέγξετε αν οι παρακάτω παρατηρήσεις (είναι ίδιες με της Εφαρμογής 5) προέρχονται από την κανονική κατανομή (ε.σ. 5%).

98.47	95.53	106.67	106.01	94.84	116.79	103.59	99.57
94.56	111.66	104.08	92.50	103.95	102.77	107.09	115.08
104.14	76.89	98.34	98.20	95.70	99.28	104.59	101.44
87.50	97.59	117.26	98.91	96.79	91.80	107.68	107.31
98.82	100.99	105.91	97.69	109.03	104.37	91.92	107.63

Εισάγουμε τα δεδομένα στο SPSS (40 περιπτώσεις – cases με μια μεταβλητή: data). Στη συνέχεια επιλέγουμε Analyze/non parametric tests/1 sample K-S/test variable: data, test distribution: Normal. Λαμβάνεται ο πίνακας:

One-Sample Kolmogorov-Smirnov Test

		DATA
N		40
Normal Parameters[a,b]	Mean	101,3235
	Std. Deviation	7,8628
Most Extreme Differences	Absolute	,084
	Positive	,084
	Negative	-,070
Kolmogorov-Smirnov Z		,534
Asymp. Sig. (2-tailed)		**,938**

a Test distribution is Normal, b Calculated from data.

Από τον πίνακα αυτό βλέπουμε ότι

$$D_n^+ = 0.084,\ D_n^- = 0.070,\ D_n = \max\{D_n^+, D_n^-\} = 0.084\,,\ Z_n = \sqrt{n}\cdot D_n \approx 0.534$$

και

$$p - value \approx 2\sum_{i=1}^{\infty}(-1)^{i-1} e^{-2i^2(0.534)^2} \approx 0.938$$

από όπου δεν μπορούμε να απορρίψουμε ότι τα δεδομένα προέρχονται από την κανονική.

Είναι σε αυτό το σημείο ενδιαφέρον να παρατηρήσουμε ότι ο ίδιος έλεγχος μέσω του χ^2 τεστ έδωσε p-value=0.189 (βλ. Εφαρμογή 5). Η μεγάλη αυτή διαφορά οφείλεται στο γεγονός ότι στην Εφαρμογή 5 ελέγξαμε αν το δείγμα προέρχεται από την N(100,65) ενώ τώρα οι παράμετροι της κανονικής κατανομής δεν είχαν καθοριστεί και για αυτό εκτιμήθηκαν από το δείγμα. Είναι εύκολο να δούμε ότι $\overline{X} = 101.3235$, $S^2 = 7.8628^2$ και επομένως τώρα ουσιαστικά ελέγχθηκε η υπόθεση H_0: $X_i \sim$ N(101,3235, 7.8628^2). Εάν είχαμε ελέγξει την ίδια υπόθεση με το χ^2 τεστ (με την ίδια διαδικασία που περιγράφηκε στην Εφαρμογή 5) τότε θα βρίσκαμε ότι p-value = 0.934.

3.4. Το κριτήριο Kolmogorov-Smirnov (K-S) για δύο δείγματα

Το κριτήριο K-S μπορεί να τροποποιηθεί ώστε να χρησιμοποιηθεί για να ελέγξουμε αν δύο δείγματα προέρχονται από την ίδια κατανομή. Συγκεκριμένα, έστω $X_1, X_2,\ldots,X_m$ και $Y_1,Y_2,\ldots,Y_n$ δύο τυχαία δείγματα από τις κατανομές F και G αντίστοιχα και έστω ότι επιθυμούμε να ελέγξουμε την υπόθεση

$$H_0 : F = G \quad \text{έναντι της} \quad H_1 : F \neq G$$

Όπως και στο K-S για ένα δείγμα, θα χρησιμοποιήσουμε τις εμπειρικές συναρτήσεις κατανομής των δύο δειγμάτων. Αυτή την φορά δεν θα τις συγκρίνουμε με κάποια θεωρητική κατανομή, αλλά μεταξύ τους. Θέτουμε

$$D_{m,n} = d_K(\hat{F}_m, \hat{G}_n) = \sup_{x\in R}\{|\hat{F}_m(x) - \hat{G}_n(x)|\}$$

Όπως και στην περίπτωση του ενός δείγματος, υπό την $H_0 : F = G$, η κατανομή του $D_{m,n}$ δεν εξαρτάται από την κοινή κατανομή των X_i, Y_i ενώ υπό την H_1: $F \neq G$, η $D_{m,n}$ λαμβάνει μεγάλες τιμές. Επομένως θα απορρίπτουμε την H_0 όταν

$$D_{m,n} > c = D_{m,n}(a) : \textit{άνω a-σημείο της κατανομής της τ.μ. } D_{m,n}(a)$$

Η ακριβής κατανομή της τ.μ. $D_{m,n}$ είναι δύσκολο να υπολογιστεί και για αυτό έχουν κατασκευαστεί πίνακες με τα άνω a-σημεία της. Αποδεικνύεται όμως ότι (υπό την H_0 και για συνεχείς F, G), όπως και στην περίπτωση του ενός δείγματος,

$$P(\sqrt{\frac{mn}{m+n}} \cdot D_{m,n} \le z) \to_{n\to\infty} 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1} e^{-2i^2 z^2} \quad \text{για κάθε } z \ge 0,$$

και επομένως το *p-value* ενός δείγματος που έδωσε $D_{m,n} = d$ θα είναι (ασυμπτωτικά)

$$p - value = P(D_{n,m} > d / H_0) = 1 - P(\sqrt{\frac{mn}{m+n}} \cdot D_{n,m} < \sqrt{\frac{mn}{m+n}} \cdot d) \approx 2\sum_{i=1}^{\infty}(-1)^{i-1} e^{-2i^2 \frac{mn}{m+n} d^2} .$$

3.5. Το Wald-Wolfowitz τέστ των ροών (runs test)

Έστω και πάλι ότι έχουμε $X_1, X_2,\ldots,X_m$ και $Y_1,Y_2,\ldots,Y_n$ δύο τυχαία δείγματα από τις κατανομές F και G αντίστοιχα και έστω ότι επιθυμούμε να ελέγξουμε την υπόθεση

$$H_0 : F = G \quad \text{έναντι της} \quad H_1 : F \neq G$$

Εκτός από το K-S τεστ, για τον έλεγχο αυτό έχουν προταθεί και άλλα τεστ όπως το τεστ των ροών (Wald-Wolfowitz runs τεστ) που θα περιγράψουμε εν συντομία στη συνέχεια.

Θεωρούμε τα δύο παραπάνω δείγματα ως ένα δείγμα $n+m$ παρατηρήσεων και στη συνέχεια διατάσσουμε (από την μικρότερη προς την μεγαλύτερη) τις παρατηρήσεις στο κοινό αυτό δείγμα. Για παράδειγμα αν το πρώτο δείγμα είναι το 1.5, 2.4, 1.2, 4.6, 0.8 και το δεύτερο δείγμα είναι το **2.5, 3.4, 0.2, 1.6, 1.8, 5.1** τότε λαμβάνουμε το διατεταγμένο δείγμα

0.2, 0.8, 1.2, 1.5, **1.6**, **1.8**, 2.4, **2.5**, **3.4**, 4.6, **5.1**

Αν στην παραπάνω ακολουθία συμβολίσουμε με X τις παρατηρήσεις από το 1ο δείγμα και με Y τις παρατηρήσεις από το 2ο δείγμα τότε λαμβάνουμε την ακολουθία συμβόλων

Y, X, X, X, **Y**, **Y**, X, **Y**, **Y**, X, **Y**

Κάτω από την H_0, οι $n+m$ παρατηρήσεις προέρχονται από την ίδια κατανομή και επομένως οι παρατηρήσεις από το 1ο και το 2ο δείγμα θα πρέπει να βρεθούν σε «τυχαίες» θέσεις στο διατεταγμένο από κοινού δείγμα. Αντίθετα υπό την H_1 θα πρέπει να διαφαίνονται κάποιες συγκεντρώσεις των μεν ή των δε (π.χ. X,X,X,X,Y,X,Y,Y,Y,Y,Y). Μια στατιστική συνάρτηση που κατά κάποιο τρόπο εκφράζει το πόσο τυχαία βρίσκονται οι παρατηρήσεις από το 1ο και το 2ο δείγμα στο διατεταγμένο από κοινού είναι η

R = *πλήθος από ομάδες συνεχόμενων όμοιων συμβόλων* X *ή* Y (*πλήθος «ροών»*)

(στην ακολουθία συμβόλων που δείχνει τις θέσεις των X_i, Y_i στο διατεταγμένο από κοινού δείγμα). Στο παραπάνω παράδειγμα οι ροές ομοίων συμβόλων είναι R = 7: (Y), (XXX), (YY), (X), (YY), (X), (Y) (ενώ στην ακολουθία X,X,X,X,Y,X,Y,Y,Y,Y,Y υπάρχουν μόνο R = 4 ροές). Είναι προφανές ότι όταν ισχύει η H_1 η R θα λαμβάνει μικρές τιμές. Επομένως θα απορρίπτουμε την H_0 όταν $R < c$. Για να βρεθεί το p-value που αντιστοιχεί στο συγκεκριμένο κριτήριο θα πρέπει να γνωρίζουμε την κατανομή της R υπό την H_0 (οι $n+m$ τ.μ. ακολουθούν την ίδια κατανομή). Η κατανομή αυτή μπορεί να εκφρασθεί αναλυτικά αλλά έχει σύνθετη μορφή και για αυτό συνήθως χρησιμοποιούμε το γεγονός ότι ασυμπτωτικά (πρακτικά m, $n > 10$),

$$Z = \frac{R - \left(\frac{2mn}{m+n} + 1\right)}{\sqrt{\frac{2mn(2mn-m-n)}{(m+n)^2(m+n-1)}}} \underset{H_0}{\sim} N(0,1) \quad (m,n \to \infty)$$

Με βάση το παραπάνω (αν r είναι το πλήθος ροών στο δείγμα),

$$\text{p-value} = P(R < r) \approx \Phi\left(\frac{r - \left(\frac{2mn}{m+n} + 1\right)}{\sqrt{\frac{2mn(2mn-m-n)}{(m+n)^2(m+n-1)}}} \right).$$

(Σε περίπτωση που υπάρχουν ίσες παρατηρήσεις στο από κοινού δείγμα, τις διατάσσουμε έτσι ώστε να προκύψει ο μεγαλύτερος δυνατός αριθμός ροών).

Παρατήρηση. (*Έλεγχος τυχαιότητας με βάση το πλήθος των ροών*). Αξίζει να σημειωθεί ότι το πλήθος των ροών μπορεί να χρησιμοποιηθεί και για *ελέγχους τυχαιότητας*. Συγκεκριμένα, έστω ότι έχουμε ένα δείγμα $X_1,X_2,\ldots,X_n$ και θέλουμε να ελέγξουμε αν οι X_i αποτελούν *τυχαίο δείγμα* από κάποια F (δηλ. είναι ανεξάρτητες τ.μ. από την F). Παρατηρούμε ότι αν π.χ. $P(X_i = 0) = P(X_i = 1) = 0.5$ (δηλ $F \sim$ Bernoulli(0.5)) τότε μια πραγματοποίηση της μορφής

$$1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1 \quad \textit{ή της μορφής} \quad 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1$$

θα μας γεννούσε υποψίες (για την «τυχαιότητα» με την οποία παράγονται οι ακολουθίες). Στην περίπτωση που οι $X_1,X_2,\ldots,X_n$ προέρχονται από μια F που δεν είναι δίτιμη, τότε μπορούμε να θέσουμε ίσες με 0 τις παρατηρήσεις που είναι κάτω του δειγματικού μέσου (ή κάτω της δειγματικής διάμεσου) και με 1 τις υπόλοιπες (αν κάποιες είναι ίσες με το μέσο εξαιρούνται από την ανάλυση) και να πάρουμε μια ανάλογη με το παραπάνω παράδειγμα ακολουθία από 0,1.

Το πλήθος των ροών R μπορεί και εδώ να χρησιμοποιηθεί αναλογιζόμενοι ότι ασυνήθιστα μεγάλες ή μικρές τιμές του R οδηγούν στο συμπέρασμα ότι το δείγμα δεν πρέπει να είναι τυχαίο (ανεξ. ισόνομες τ.μ.). Επομένως θα απορρίπτεται η

$$H_0\text{: } \textit{το δείγμα είναι τυχαίο}$$

όταν

$$R < c_1 \text{ ή } R > c_2.$$

Χρησιμοποιώντας το παραπάνω ασυμπτωτικό αποτέλεσμα, το αντίστοιχο p-value θα είναι (τώρα ο έλεγχος είναι αμφίπλευρος)

$$\text{p-value} \approx 2\left(1 - \Phi\left(\frac{r - \left(\frac{2mn}{m+n} + 1\right)}{\sqrt{\frac{2mn(2mn-m-n)}{(m+n)^2(m+n-1)}}} \right)\right).$$

όπου m, n είναι το πλήθος από 0, 1 αντίστοιχα στο δείγμα. Προφανώς θα μπορούσε κανείς εδώ αντί για αμφίπλευρο έλεγχο να απορρίπτει μόνο όταν $R < c$ (π.χ. έχοντας ως εναλλακτική την θετική εξάρτηση μεταξύ των παρατηρήσεων) ή όταν $R > c$ (π.χ. έχοντας ως εναλλακτική την αρνητική εξάρτηση μεταξύ των παρατηρήσεων)

Όπως ίσως μπορεί κανείς να φαντασθεί, υπάρχουν πολλά διαφορετικά (απαραμετρικά) κριτήρια που θα μπορούσαν να χρησιμοποιηθούν για έναν έλεγχο τυχαιότητας ή για τον έλεγχο ισότητας των κατανομών δύο δειγμάτων (πράγματι στην βιβλιογραφία έχουν προταθεί απαραμετρικά κριτήρια που βασίζονται π.χ. στην μεγαλύτερη ροή, σε ανοδικές ροές, στους βαθμούς (ranks) των παρατηρήσεων κ.α.). Το κάθε ένα από αυτά τα τεστ είναι «ευαίσθητο» σε διαφορετικού είδους εναλλακτική υπόθεση. Το τεστ των ροών που εξετάσαμε παραπάνω αν και δεν είναι το πιο ισχυρό, είναι το πιο απλό και το πιο γενικό (όσον αφορά την εναλλακτική υπόθεση) τεστ αυτής της μορφής.

Στη συνέχεια θα δούμε ακόμη ένα τεστ που βασίζεται περίπου στην ίδια ιδέα (διάταξη του από κοινού δείγματος των X_i, Y_i).

3.6. Το Mann-Whitney U τεστ

Και αυτό το τεστ χρησιμοποιείται για τον έλεγχο ισότητας των κατανομών δύο δειγμάτων. Έστω και πάλι ότι έχουμε $X_1, X_2,\ldots,X_m$ και $Y_1,Y_2,\ldots,Y_n$ δύο τυχαία δείγματα από τις κατανομές F και G αντίστοιχα και έστω ότι επιθυμούμε να ελέγξουμε την υπόθεση

$$H_0 : F = G \quad \text{έναντι της} \quad H_1 : F \neq G$$

Ενώνουμε όπως και την περίπτωση του Wald-Wolfowitz runs τεστ τα δύο παραπάνω δείγματα σε ένα δείγμα $n+m$ παρατηρήσεων και στη συνέχεια διατάσσουμε (από την μικρότερη προς την μεγαλύτερη) τις παρατηρήσεις στο κοινό αυτό δείγμα. Αυτή τη φορά όμως δεν μετράμε το πλήθος των ροών, αλλά το πλήθος από τα Y_i που είναι μικρότερα του X_1 συν το πλήθος από τα Y_i που είναι μικρότερα του X_2 κ.ο.κ. στο διατεταγμένο από κοινού δείγμα. Χρησιμοποιώντας το ίδιο παράδειγμα με την προηγούμενη παράγραφο, αν το πρώτο δείγμα είναι το 1.5, 2.4, 1.2, 4.6, 0.8 και το δεύτερο είναι το **2.5, 3.4, 0.2, 1.6, 1.8, 5.1** τότε λαμβάνουμε το διατεταγμένο δείγμα

0.2, 0.8, 1.2, 1.5, **1.6**, **1.8**, 2.4, **2.5**, **3.4**, 4.6, **5.1**

Τώρα βλέπουμε ότι το 0.8, το 1.2 και το 1.5 είναι μεγαλύτερο από ένα Y_i (το 0.2), το 2.4 είναι μεγαλύτερο από τρία Y_i (τα 0.2, 1.6, 1.8), και το 4.6 είναι μεγαλύτερο από πέντε X_i. Άρα εδώ,

$$U = 1 + 1 + 1 + 3 + 5 = 11$$

Θα απορρίπτεται η $H_0 : F = G$ έναντι της $H_1 : F \neq G$ όταν το U είναι αδικαιολόγητα μικρό ή μεγάλο ($U < c_1$ ή $U > c_2$). Η κατανομή της τ.μ. U μπορεί να παρασταθεί αναλυτικά (κάτω από την H_0) αλλά και πάλι δεν έχει απλή μορφή. Αποδεικνύεται ότι, ασυμπτωτικά,

$$Z = \frac{U - mn/2}{\sqrt{mn(m+n+1)/12}} \underset{H_0}{\sim} N(0,1) \quad (m,n \to \infty)$$

και επομένως (για μεγάλα m, n) αν z είναι η τιμή της παραπάνω στατιστικής συνάρτησης στο δείγμα,

$$\text{p-value} = P(|Z| > |z|) \approx 2(1-\Phi(|z|)) = 2\left(1-\Phi\left(\frac{|u-mn/2|}{\sqrt{mn(m+n+1)/12}}\right)\right).$$

Το τεστ αυτό είναι σε αρκετές περιπτώσεις πιο ισχυρό από το τεστ των ροών διότι χρησιμοποιεί περισσότερη πληροφορία από το δείγμα.

Εφαρμογή 7. Να ελέγξετε αν οι παρατηρήσεις 0.1, 0.0, 1.0, 1.8, 2.0, 2.8, 3.4 και 1.7, 1.9, 2.1, 5.2, 6.9, 7.0, 7.8, 9.1 προέρχονται από την ίδια κατανομή.

Εισάγουμε τα δεδομένα σε μία στήλη του SPSS με όνομα π.χ. a (μία στήλη με 7+8 = 15 παρατηρήσεις). Όπως και στα t-tests για δυο ανεξάρτητα δείγματα χρησιμοποιούμε και μια βοηθητική μεταβλητή, την g η οποία λαμβάνει τις τιμές 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2 «δείχνοντας» το group που ανήκει κάθε παρατήρηση. Στη συνέχεια εκτελούμε Analyze/Non parametric tests/2 independent samples, test variable: a, grouping variable: g, Test type: Kolmogorov-Smirnov Z, Mann-Whitney U, Wald – Wolfowitz runs test (δηλ. επιλέγουμε και τα τρία παραπάνω τεστ). Λαμβάνονται οι πίνακες:

Mann-Whitney Test

Test Statistics[b]

	A
Mann-Whitney U	9,000
Wilcoxon W	37,000
Z	-2,199
Asymp. Sig. (2-tailed)	,028
Exact Sig. [2*(1-tailed Sig.)]	,029[a]

a Not corrected for ties, b Grouping Variable: G

Two-Sample Kolmogorov-Smirnov Test

Test Statistics

		A
Most Extreme Differences	Absolute	,625
	Positive	,000
	Negative	-,625
Kolmogorov-Smirnov Z		1,208
Asymp. Sig. (2-tailed)		,108

Wald-Wolfowitz Test

Test Statistics[b,c]

		Number of Runs	Z	Exact Sig. (1-tailed)
A	Exact Number of Runs	8[a]	,000	,514

a No inter-group ties encountered, b Wald-Wolfowitz Test, c Grouping Variable: G

Από το Mann-Whitney τεστ απορρίπτουμε ότι τα δείγματα προέρχονται από τον ίδιο πληθυσμό (σε ε.σ. 5%) διότι το p-value=0.029 ενώ τα άλλα δύο τεστ δεν κατάφεραν με βάση το συγκεκριμένο δείγμα να εντοπίσουν διαφορά στις κατανομές των δυο δειγμάτων (p-values: 0.108 και 0.514).

Εφαρμογή 8. Από ένα πείραμα λαμβάνονται κατά σειρά οι επόμενες παρατηρήσεις 2, 2, 4, 9, 12, 10, 8, 4, 7, 3, 1, 0. Οι παρατηρήσεις αυτές μπορεί να αποτελούν τυχαίο δείγμα; (δηλ. μπορεί να πρόκειται για πραγματοποίηση ανεξάρτητων τ.μ. από μια κοινή κατανομή;)

Εισάγουμε τα δεδομένα σε μια μεταβλητή, έστω x. Θα χρησιμοποιήσουμε το runs τεστ για το έλεγχο τυχαιότητας του δείγματος (βλ. παρατήρηση στην Παράγραφο 3.5). Επιλέγουμε Analyze/Non parametric tests/runs, test variable:x, cut point: median (οι παρατηρήσεις πάνω από την διάμεσο θεωρούνται «1» και αυτές κάτω από την διάμεσο «0»). Λαμβάνεται ο πίνακας

Runs Test

	X
Test Value[a]	4,00
Cases < Test Value	5
Cases >= Test Value	7
Total Cases	12
Number of Runs	3
Z	-2,082
Asymp. Sig. (2-tailed)	,037

a Median

από όπου βλέπουμε ότι βρέθηκαν 5 παρατηρήσεις μικρότερες και 7 παρατηρήσεις μεγαλύτερες ή ίσες της (δειγματικής) διαμέσου (=4), ενώ ο αριθμός των ροών ήταν μόλις 3. Το αντίστοιχο p-value είναι 0.037, δηλαδή

$$\text{p-value} \approx 2\left(1-\Phi\left(\frac{r-\left(\frac{2mn}{m+n}+1\right)+0.5}{\sqrt{\frac{2mn(2mn-m-n)}{(m+n)^2(m+n-1)}}}\right)\right) \approx 0.037 \quad (n=5,\, m=7,\, r=3)$$

(χρησιμ. διόρθωση συνέχειας) και επομένως απορρίπτουμε (σε ε.σ. 5%) ότι οι παραπάνω παρατηρήσεις αποτελούν τυχαίο δείγμα από μία κατανομή (πράγματι, βλέπουμε ότι οι τιμές των παρατηρήσεων αρχικά αυξάνονται και μετά μειώνονται κάτι που, όπως φάνηκε από το p-value, σπάνια συμβαίνει «τυχαία»).