



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Διπλωματική Εργασία

ΜΗ ΠΑΡΑΜΕΤΡΙΚΕΣ ΜΕΘΟΔΟΙ ΚΑΙ ΕΦΑΡΜΟΓΕΣ ΜΕ ΧΡΗΣΗ ΤΟΥ
ΣΤΑΤΙΣΤΙΚΟΥ ΠΑΚΕΤΟΥ ΤΗΣ R

Κωνσταντίνου Αθηνά

Τριμελής Επιτροπή: Φουσκάκης Δημήτρης (Επιβλέπων Καθηγητής)
Κολέτσος Ιωάννης
Λουλάκης Μιχάλης

Αθήνα, Σεπτέμβρης 2018

Ευχαριστίες

Αρχικά νιώθω την ανάγκη να ευχαριστήσω τον κύριο Δημήτρη Φουσκάκη τόσο για την ανάθεση της διπλωματικής εργασίας όσο και για την συνεισφορά του στο να μπορέσω να πραγματοποιήσω το επόμενο βήμα στις σπουδές μου. Επιπλέον, θα ήθελα να πω ένα μεγάλο ευχαριστώ στην οικογένεια μου καθώς τίποτα από όλα αυτά δεν θα μπορούσε να επιτευχθεί αν δεν είχα την άφθονη αγάπη και υποστήριξη της. Τέλος, ευχαριστώ τους φίλους μου που έκαναν υπομονή και ήταν εκεί σε κάθε δύσκολη στιγμή, δε θα το ξεχάσω ποτέ. Σας αγαπώ όλους πολύ.

Περίληψη

Στην παρούσα διπλωματική εργασία παρουσιάζονται πέντε από τους δημοφιλέστερους ελέγχους υποθέσεων που βασίζονται στους βαθμούς και όχι στις τιμές των παρατηρήσεων. Σε κάθε κεφάλαιο υπάρχει το θεωρητικό υπόβαθρο της μεθόδου ενώ στο τέλος αναφέρεται η εντολή στο στατιστικό πακέτο της R που πραγματοποιεί τον έλεγχο για εμάς γρήγορα και βοηθάει να γλυτώσουμε πολύτιμο χρόνο. Αναλυτικότερα, το Κεφάλαιο 1 είναι μια εισαγωγή στους μη παραμετρικούς ελέγχους. Στο Κεφάλαιο 2 παρουσιάζεται ο έλεγχος των Wilcoxon-Mann-Whitney που αφορά στη σύγκριση δύο θεραπειών που εφαρμόζονται σε ανεξάρτητα δείγματα. Στην συνέχεια, στο Κεφάλαιο 3, αναφέρεται ο έλεγχος προσημασμένων τάξεων μεγέθους του Wilcoxon ο οποίος συγκρίνει επίσης δύο θεραπείες, μόνο που αυτή τη φορά το δείγμα αποτελείται από ζεύγη παρατηρήσεων. Στο Κεφάλαιο 4 έχουμε τον έλεγχο των Kruskal – Wallis για την σύγκριση περισσότερων από δύο θεραπειών που είναι στην ουσία μια γενίκευση του ελέγχου που είδαμε στο Κεφάλαιο 2. Στο τελευταίο κεφάλαιο της διπλωματικής, το Κεφάλαιο 5, έχουμε τον έλεγχο του Friedman για την σύγκριση περισσότερων από δύο θεραπειών αλλά όταν έχουμε ομαδοποιημένες παρατηρήσεις. Παρουσιάζεται επίσης μια ειδική περίπτωση του ελέγχου που είναι ο έλεγχος του McNemar.

Abstract

This dissertation showcases five of the most popular hypothesis testing methods that are based on ranks. Each chapter analyses the theory behind each method and concludes with commands that are needed to be applied in R to produce these tests saving us valuable time from executing them ourselves.

In more detail, chapter 1 introduces nonparametric statistics. Chapter two focuses on the Wilcoxon rank sum test which is used to compare two independent samples randomly selected from a population. Following, chapter three where the Wilcoxon signed-rank test is presented to compare two related samples.

In chapter four we analyze the Kruskal – Wallis test which is used to compare more than two treatments and pinpoint differences among them. Last but not least, chapter five presents the tests of Friedman and McNemar.

Περιεχόμενα

| | |
|--|-----------|
| Περίληψη | 4 |
| Abstract | 5 |
| | |
| Κεφάλαιο 1 | |
| Εισαγωγή | |
| 1.1 Μη παραμετρική στατιστική. | 8 |
| | |
| Κεφάλαιο 2 | |
| Ο έλεγχος του Wilcoxon και των Mann - Whitney | |
| 2.1 Η κατανομή των στατιστικών του Wilcoxon και των Mann- Whitney υπό την μηδενική υπόθεση. | 10 |
| 2.2 Η κατανομή των στατιστικών του Wilcoxon και των Mann- Whitney υπό την μηδενική υπόθεση για δείγματα που έχουν επιλεγθεί κατά τυχαίο τρόπο από έναν πληθυσμό. | 17 |
| 2.3 Ασυμπτωτική κατανομή των στατιστικών του Wilcoxon και των Mann- Whitney. | 19 |
| 2.4 Όταν δύο ή περισσότερες παρατηρήσεις έχουν ίσες τιμές. | 20 |
| 2.5 Η εντολή του έλεγχου στο στατιστικό πακέτο της R. | 24 |
| | |
| Κεφάλαιο 3 | |
| Ο έλεγχος προσημασμένων τάξεων μεγέθους του Wilcoxon | |
| 3.1 Η κατανομή του στατιστικού υπό την μηδενική υπόθεση. | 29 |
| 3.2 Ασυμπτωτική κατανομή του στατιστικού ελέγχου. | 31 |
| 3.3 Μια διαφορετική παρουσίαση του στατιστικού ελέγχου. | 32 |
| 3.4 Όταν έχουμε παρατηρήσεις με ίσες τιμές. | 33 |
| 3.5 Η εντολή του ελέγχου στο στατιστικό πακέτο της R. | 35 |
| | |
| Κεφάλαιο 4 | |
| Ο έλεγχος των Kruskal - Wallis | |
| 4.1 Η κατανομή υπό την μηδενική υπόθεση για τους βαθμούς των παρατηρήσεων από s το πλήθος θεραπείες. | 39 |
| 4.2 Περιγραφή του ελέγχου των Kruskal – Wallis. | 41 |
| 4.3 Όταν έχουμε ίσες παρατηρήσεις. | 42 |
| 4.4 Η εντολή του ελέγχου στο στατιστικό πακέτο της R. | 43 |
| | |
| Κεφάλαιο 5 | |
| Ο έλεγχος του Friedman | |
| 5.1 Οι βαθμοί σε τυχαιοποιημένα πλήρη μπλοκ σχεδιασμών. | 49 |
| 5.2 Περιγραφή του έλεγχου Friedman και παρουσίαση ειδικής περίπτωσης αυτού. | |
| 5.3 Όταν έχουμε παρατηρήσεις με ίσες τιμές. | 50 |
| 5.4 Όταν θέλουμε να συγκρίνουμε δύο θεραπείες για ομαδοποιημένα δεδομένα. | 52 |
| 5.5 Μια ειδική μορφή του έλεγχου Friedman, ο έλεγχος McNemar. | 53 |
| 5.6 Η εντολή του έλεγχου του Friedman και του McNemar στο στατιστικό πακέτο της R. | 54 |
| | |
| Βιβλιογραφία | 58 |

1. Εισαγωγή

1.1 Μη παραμετρική στατιστική – Μη παραμετρικοί έλεγχοι

Είναι σημαντικό να αναφερθούμε αρχικά με συντομία στους παραμετρικούς ελέγχους υποθέσεων. Οι παραμετρικοί έλεγχοι είναι έλεγχοι υποθέσεων που εφαρμόζονται προκειμένου να εκτιμηθεί μια άγνωστη παράμετρος θ , όπως η μέση τιμή ή η διασπορά, (ή μια συνάρτηση του θ) της κατανομής μιας τυχαίας μεταβλητής X και βασίζονται σε κάποιες προϋποθέσεις προκειμένου να μπορούν να εφαρμοστούν. Για παράδειγμα κάποιοι έλεγχοι απαιτούν δεδομένα που προέρχονται από κανονικό πληθυσμό. Σε περιπτώσεις που αυτό δεν ισχύει, μπορούμε να με τη βοήθεια του Κεντρικού Οριακού Θεωρήματος (Κ.Ο.Θ) (Central Limit Theorem) να προσεγγίσουμε την κατανομή των δεδομένων από την κανονική κατανομή. Πολλές φορές ωστόσο συμβαίνει μία ή και περισσότερες προϋποθέσεις των παραμετρικών ελέγχων να παραβιάζονται, πχ. οι παρατηρήσεις δεν προέρχονται από κανονικό πληθυσμό και το πλήθος τους είναι μικρό για να μπορέσουμε να εφαρμόσουμε το Κ.Ο.Θ. Σε περιπτώσεις όπως αυτή υπάρχουν εναλλακτικές μέθοδοι για να βγάλουμε τα απαραίτητα για την έρευνα συμπεράσματα, που καλούνται μη παραμετρικές μέθοδοι (non-parametric techniques). Το βασικό προτέρημα των μεθόδων αυτών είναι ότι μπορούν να εφαρμοστούν σε δεδομένα που δεν ακολουθούν κάποια γνωστή κατανομή καθώς τις περισσότερες φορές δεν κάνουν χρήση των τιμών των παρατηρήσεων αλλά των βαθμών τους (σειρά κατάταξης rank) και για αυτό εφαρμόζονται σε δεδομένα που παίρνουν τιμές ονομαστικές ή διάταξης και όχι μόνο αριθμητικές. Για παράδειγμα μπορεί σαν δεδομένα να έχουμε την γνώμη κάποιων ατόμων σχετικά με ένα πολιτικό ζήτημα. Η κάθε παρατήρηση μπορεί επομένως να ανήκει σε μια από τις κατηγορίες «συμφωνώ απόλυτα», «συμφωνώ», «μου είναι αδιάφορο», «διαφωνώ», «διαφωνώ απόλυτα». Στην πλειοψηφία τους οι μη παραμετρικοί έλεγχοι κάνουν χρήση απλών στατιστικών ελέγχου που δεν απαιτούν πολλές πράξεις για τον υπολογισμό τους. Τα στατιστικά ελέγχων συνήθων υπολογίζονται από τους βαθμούς διάταξης και για τον λόγο αυτό οι τιμές τους δεν επηρεάζονται από ακραίες παρατηρήσεις όπως συμβαίνει στους παραμετρικούς ελέγχους. Τέλος, ακόμα και αν το μέγεθος του υπό μελέτη δείγματος είναι μικρό, οι μη παραμετρικοί έλεγχοι μπορούν να εφαρμοστούν και μάλιστα δίνουν εξαιρετικά αξιόπιστα αποτελέσματα. Ωστόσο, σε περιπτώσεις που μπορεί να εφαρμοστεί κάποιος από τους γνωστούς παραμετρικούς ελέγχους αποφεύγουμε να χρησιμοποιήσουμε κάποιον μη παραμετρικό καθώς η ισχύς του θα είναι συγκριτικά πολύ μικρότερη και είναι πιθανό να οδηγηθούμε σε τελείως λανθασμένα συμπεράσματα. Οι μη παραμετρικοί έλεγχοι βρίσκουν εφαρμογή σε ένα μεγάλο εύρος προβλημάτων των κλάδων της ιατρικής, της οικονομίας, του marketing και πολλών ακόμα.

2. Ο έλεγχος του Wilcoxon και των Mann-Whitney

2.1 Η κατανομή των στατιστικών του Wilcoxon και των Mann-Whitney υπό την μηδενική υπόθεση για τυχαία δείγματα.

Υποθέτουμε ότι έχουμε μετρήσεις της ίδιας ποσοτικής μεταβλητής σε δύο ομάδες διαφορετικών μονάδων του πληθυσμού (πρόκειται στην ουσία για δύο διαφορετικούς υποπληθυσμούς) και θέλουμε να ελέγξουμε αν υπάρχει κάποιος είδους σχέση ανάμεσα τους. Θα συμβολίζουμε με X το χαρακτηριστικό από τον πρώτο υποπληθυσμό και με Y το ίδιο χαρακτηριστικό αλλά από τον δεύτερο υποπληθυσμό και υποθέτουμε ότι οι τυχαίες μεταβλητές X, Y είναι ανεξάρτητες και επομένως η τιμή του χαρακτηριστικού X δεν επηρεάζεται από την τιμή του χαρακτηριστικού Y . Έστω X_1, \dots, X_n ένα τυχαίο δείγμα από τον πρώτο υποπληθυσμό και Y_1, \dots, Y_m ένα τυχαίο δείγμα από τον δεύτερο υποπληθυσμό. Προκειμένου να ελέγξουμε αν υπάρχει κάποια διαφορά του υπό μελέτη χαρακτηριστικού στους δύο υποπληθυσμούς θα ελέγξουμε την μηδενική υπόθεση H_0 : Δεν διαφέρουν, σε κάποιο επίπεδο σημαντικότητας α . Στην πραγματικότητα θα ελέγξουμε αν οι κατανομές των δύο υποπληθυσμών είναι ίσες, δηλαδή αν τα δύο δείγματα προέρχονται από πληθυσμούς με την ίδια συνάρτηση κατανομής και υπό την προϋπόθεση ότι οι δύο υποπληθυσμοί έχουν το ίδιο σχήμα και περίπου ίδια διασπορά, θα πάρουμε μικρή P-τιμή όταν τα μέτρα θέσης των κατανομών (για παράδειγμα η διάμεσος) θα διαφέρουν στους δύο υποπληθυσμούς. Αξίζει να σημειωθεί πως δεν είναι απαραίτητο τα δύο δείγματα να είναι ίδιου μεγέθους καθώς πολλές φορές είναι ευκολότερο να συλλέξουμε δείγματα για έναν πληθυσμό παρά για κάποιον άλλο.

Συγκεκριμένα, έστω ότι επιθυμούμε να πραγματοποιήσουμε μια έρευνα για την αποτελεσματικότητα ενός νέου φαρμάκου για την χοληστερίνη. Για τον λόγο αυτό N το πλήθος ασθενείς είναι διαθέσιμοι να συμμετάσχουν στην έρευνα και θα συμβολίζουμε με n το πλήθος αυτών που θα λάβουν την νέα αγωγή ενώ θα συμβολίζουμε με $m = N - n$ το πλήθος όσων θα λάβουν ένα εικονικό φάρμακο που δεν περιέχει ενεργές ουσίες. Συνολικά μπορούμε να πραγματοποιήσουμε $\binom{N}{n}$ τυχαίους συνδυασμούς των n παρατηρήσεων που θα λάβουν την νέα αγωγή. Ο αριθμός αυτός καλείται διωνυμικός συντελεστής και είναι όλοι οι πιθανοί συνδυασμοί που μπορούν να προκύψουν αν από N στοιχεία διαλέξω n και δίνεται από την σχέση:

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

Το δείγμα μεγέθους n επιλέγεται κατά τυχαίο τρόπο και επομένως κάθε $\binom{N}{n}$ συνδυασμός είναι εξίσου πιθανός να προκύψει με πιθανότητα $1/\binom{N}{n}$ ο καθένας.

Στο τέλος της έρευνας οι N το πλήθος ασθενείς βαθμολογούνται ως ένα ενιαίο δείγμα ανάλογα με τα επίπεδα χοληστερίνης τους. Η μηδενική υπόθεση του ελέγχου είναι ότι το νέο φάρμακο δεν έχει καμία απολύτως επίδραση στα επίπεδα χοληστερίνης του ασθενή που το λαμβάνει. Υπό την μηδενική υπόθεση επομένως, ο βαθμός που θα λάβει ο κάθε ασθενής δεν εξαρτάται από το αν πήρε το νέο φάρμακο ή όχι παρά μόνο από το επίπεδο της χοληστερίνης του μετά το τέλος της έρευνας που θα είναι αμετάβλητο (αγνοούμε τον ρόλο που παίζει η διατροφή στα επίπεδα της χοληστερίνης) αφού το νέο φάρμακο δεν έχει καμία απολύτως επίδραση και επομένως μπορούμε να σκεφτούμε πως οι βαθμοί θα μπορούσαν να έχουν δοθεί στους ασθενείς πριν την έρευνα και πως όταν διαλέξαμε τυχαία το δείγμα μεγέθους n για να λάβει την νέα θεραπεία, διαλέξαμε τυχαία και τους βαθμούς του δείγματος αυτού. Επομένως, υπό την

μηδενική υπόθεση όλοι οι $\binom{N}{n}$ δυνατοί συνδυασμοί των $1, \dots, N$ βαθμών για τους ασθενείς που έλαβαν την θεραπεία έχουν πιθανότητα $1/\binom{N}{n}$ ο καθένας να προκύψει.

Έστω ότι το σύνολο S_1, \dots, S_n είναι οι βαθμοί των ασθενών που έλαβαν την νέα θεραπεία δοσμένοι κατά αύξουσα σειρά, ενώ οι R_1, \dots, R_m είναι οι βαθμοί εκείνων των ασθενών που έλαβαν το εικονικό φάρμακο δοσμένοι επίσης κατά αύξουσα σειρά. Αυτοί οι $n + m$ το πλήθος βαθμοί είναι στην ουσία μια διάταξη των ακεραίων $1, \dots, N$ και για να τους προσδιορίσουμε αρκεί να γνωρίζουμε το ένα από τα δύο σύνολα και έστω πως γνωρίζουμε τους βαθμούς των ασθενών που έλαβαν την θεραπεία, δηλαδή πως γνωρίζουμε στο σύνολο S_1, \dots, S_n . Όπως αναφέραμε υπάρχουν $\binom{N}{n}$ δυνατοί τρόποι να παρατηρήσουμε την n -άδα των S_1, \dots, S_n βαθμών και επομένως υπό την μηδενική υπόθεση η πιθανότητα να παρατηρήσουμε έναν από τους $\binom{N}{n}$ συνδυασμούς είναι:

$$P_{H_0}(S_1 = s_1, \dots, S_n = s_n) = 1/\binom{N}{n}.$$

Υπό την υπόθεση ότι το νέο φάρμακο δεν έχει καμία επίδραση περιμένουμε οι βαθμοί S_1, \dots, S_n των ασθενών που έλαβαν την θεραπεία να παίρνουν τελείως τυχαίες τιμές από το $(1, \dots, N)$ σύνολο των δυνατών τιμών που μπορούν να πάρουν. Απορρίπτουμε την μηδενική υπόθεση επομένως και αναγνωρίζουμε την αποτελεσματικότητα του νέου φαρμάκου για τιμές του συνόλου των βαθμών S_1, \dots, S_n που είναι αρκετά μεγάλες (αν υποθέσουμε πως ο ασθενής με το χαμηλότερο επίπεδο χοληστερίνης παίρνει τον μεγαλύτερο βαθμό N , αυτός με το αμέσως χαμηλότερο επίπεδο τον βαθμό $N - 1$ κ.ο.κ μέχρι αυτός που έχει την περισσότερη χοληστερίνη να πάρει τον βαθμό 1).

Πως όμως μπορούμε να αποφασίσουμε ότι οι βαθμοί των ασθενών που έλαβαν την θεραπεία είναι αρκετά υψηλοί; Για να απαντήσουμε στην ερώτηση αυτή αρκεί να βρούμε ένα κατάλληλο στατιστικό ελέγχου. Αν οι βαθμοί των ασθενών είναι υψηλοί τότε και το άθροισμα τους θα αντιστοιχεί σε έναν μεγάλο αριθμό. Έτσι ορίζουμε ως στατιστικό ελέγχου το άθροισμα των βαθμών εκείνων των ασθενών που έλαβαν την θεραπεία. Το στατιστικό ελέγχου που περιγράψαμε μπορεί να υπολογιστεί από την σχέση:

$$W_s = S_1 + \dots + S_n$$

και η μηδενική υπόθεση θα απορρίπτεται για μια τιμή του στατιστικού W_s πολύ μεγάλη παραδείγματος χάριν

$$W_s \geq c.$$

Η τιμή της σταθεράς c καθορίζεται από την σχέση:

$$P_{H_0}(W_s \geq c) = \alpha$$

όπου η παραπάνω πιθανότητα υπολογίζεται υπό την μηδενική υπόθεση.

Υπό την μηδενική υπόθεση, ότι δηλαδή η θεραπεία δεν έχει αποτελέσματα το στατιστικό W_s δεν θα παίρνει μεγάλες τιμές και η μόνη περίπτωση να πάρει μεγάλη τιμή θα είναι κατά τύχη με πιθανότητα ίση με α , που ορίζεται εξαρχής ως ένας πολύ μικρός αριθμός (0.05 ή 0.01 συνήθως). Το α ονομάζεται επίπεδο σημαντικότητας του ελέγχου.

Για να υπολογίσουμε την κατανομή του W_s υπό την μηδενική υπόθεση χρησιμοποιούμε τη σχέση:

$$P_{H_0}(W_s = w) = \frac{\#(w; n, m)}{\binom{N}{n}},$$

όπου w είναι η παρατηρούμενη τιμή του στατιστικού W_s και ο όρος $\#(w; n, m)$ στον αριθμητή του κλάσματος είναι το πλήθος των συνδυασμών των βαθμών των ασθενών που θα λάβουν την αγωγή και θα δώσουν παρατηρούμενη τιμή για το στατιστικό ίση με w . Μπορούμε με τη βοήθεια του Παραδείγματος 2.1.1 να κατανοήσουμε καλύτερα τον τρόπο υπολογισμού της κατανομής του στατιστικού W_s υπό την μηδενική υπόθεση.

Παράδειγμα 2.1.1

Έστω ότι έχουμε ένα δείγμα $N=8$ ατόμων και ότι 4 από τα άτομα αυτά θα λάβουν το νέο φάρμακο για την χοληστερίνη ενώ τα υπόλοιπα 4 θα λάβουν το εικονικό φάρμακο που δεν περιέχει ενεργά συστατικά. Στο τέλος της έρευνας τα 8 άτομα θα βαθμολογηθούν ως ένα ενιαίο δείγμα με βάση τα επίπεδα χοληστερίνης στον οργανισμό τους. Οι βαθμοί για το δείγμα είναι στο σύνολο $1, \dots, 8$. Μπορούν να προκύψουν συνολικά $\binom{8}{4}$ συνδυασμοί των βαθμών των 4 ατόμων που θα λάβουν την θεραπεία. Από την R με χρήση της εντολής

```
> choose(8,4)
[1] 70
```

βλέπουμε πως υπάρχουν 70 δυνατοί συνδυασμοί των τεσσάρων βαθμών για τα άτομα που έλαβαν το νέο φάρμακο. Ο κάθε ένας συνδυασμός των τεσσάρων βαθμών έχει πιθανότητα ίση με $1/70$ να προκύψει. Μπορούμε να δούμε αυτούς τους 70 συνδυασμούς στην R με την εντολή

```
> combn(8,4)
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
[1,]  1  1  1  1  1  1  1  1  1  1  1  1  1
[2,]  2  2  2  2  2  2  2  2  2  2  2  2  2
[3,]  3  3  3  3  3  4  4  4  4  5  5  5  6
[4,]  4  5  6  7  8  5  6  7  8  6  7  8  7
     [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24] [,25]
[1,]  1  1  1  1  1  1  1  1  1  1  1  1  1
[2,]  2  2  3  3  3  3  3  3  3  3  3  3  3
[3,]  6  7  4  4  4  4  4  5  5  5  6  6  7
[4,]  8  8  5  6  7  8  8  6  7  8  7  8  8
     [,26] [,27] [,28] [,29] [,30] [,31] [,32] [,33] [,34] [,35] [,36] [,37]
[1,]  1  1  1  1  1  1  1  1  1  1  2  2
[2,]  4  4  4  4  4  4  5  5  5  6  3  3
[3,]  5  5  5  6  6  7  6  6  7  7  4  4
[4,]  6  7  8  7  8  8  7  8  8  8  5  6
     [,38] [,39] [,40] [,41] [,42] [,43] [,44] [,45] [,46] [,47] [,48] [,49]
[1,]  2  2  2  2  2  2  2  2  2  2  2  2
[2,]  3  3  3  3  3  3  3  3  4  4  4  4
[3,]  4  4  5  5  5  6  6  7  5  5  5  6
[4,]  7  8  6  7  8  7  8  8  6  7  8  7
     [,50] [,51] [,52] [,53] [,54] [,55] [,56] [,57] [,58] [,59] [,60] [,61]
[1,]  2  2  2  2  2  2  3  3  3  3  3  3
[2,]  4  4  5  5  5  6  4  4  4  4  4  4
[3,]  6  7  6  6  7  7  5  5  5  6  6  7
[4,]  8  8  7  8  8  8  6  7  8  7  8  8
     [,62] [,63] [,64] [,65] [,66] [,67] [,68] [,69] [,70]
[1,]  3  3  3  3  4  4  4  4  5
[2,]  5  5  5  6  5  5  5  6  6
[3,]  6  6  7  7  6  6  7  7  7
[4,]  7  8  8  8  7  8  8  8  8
```

που είναι ένας πίνακας 4×70 και σε κάθε στήλη έχουμε 4 υποψήφιους συνδυασμούς βαθμών.

Όπως αναφέραμε πριν η τιμή του στατιστικού ελέγχου δίνεται ως το άθροισμα των βαθμών των ασθενών που έλαβαν τη νέα θεραπεία και η κατανομή του υπό την μηδενική υπόθεση δίνεται από τον τύπο

$$P_{H_0}(W_s = w) = \frac{\#(w; n, m)}{\binom{N}{n}}$$

Από τους παραπάνω συνδυασμούς, θα υπάρχουν κάποιοι που θα δίνουν ίδια παρατηρούμενη τιμή w για το στατιστικό W_s . Για να υπολογίσουμε την παρατηρούμενη τιμή w στο παραπάνω παράδειγμα επειδή έχουμε πολλές τιμές θα κάνουμε χρήση της R και της εντολής

> `combn(8,4, FUN=sum)`

```
[1] 10 11 12 13 14 12 13 14 15 14 15 16 16 17 18 13 14 15 16 15 16 17 17 18
[25] 19 16 17 18 18 19 20 19 20 21 22 14 15 16 17 16 17 18 18 19 20 17 18 19
[49] 19 20 21 20 21 22 23 18 19 20 20 21 22 21 22 23 24 22 23 24 25 26
```

που μας επιτρέπει να δούμε όλες τις παρατηρούμενες τιμές του W_s που μπορούν να προκύψουν για δείγμα $N = 8$ και $n = 4$ ασθενείς που θα λάβουν το νέο φάρμακο. Από τις τιμές αυτές μπορούμε να υπολογίσουμε την κατανομή για το στατιστικό W_s υπό την μηδενική υπόθεση από την τελευταία σχέση. Στον Πίνακα 2.1.1 φαίνονται τα αποτελέσματα για το Παράδειγμα 2.1.1.

| | | | | | | |
|--------------------|-----------|-----------|-----------|-----------|-----------|-----------|
| w | 10 | 11 | 12 | 13 | 14 | 15 |
| $P_{H_0}(W_s = w)$ | 1/70 | 1/70 | 2/70 | 3/70 | 5/70 | 5/70 |
| w | 16 | 17 | 18 | 19 | 20 | 21 |
| $P_{H_0}(W_s = w)$ | 7/70 | 7/70 | 8/70 | 7/70 | 7/70 | 5/70 |
| w | 22 | 23 | 24 | 25 | 26 | |
| $P_{H_0}(W_s = w)$ | 5/70 | 3/70 | 2/70 | 1/70 | 1/70 | |

Πίνακας 2.1.1: Η κατανομή του στατιστικού W_s υπό την μηδενική υπόθεση για το Παράδειγμα 2.1.1.

Με την εντολή

> `table(sort(combn(8,4, FUN=sum)))`

```
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
 1  1  2  3  5  5  7  7  8  7  7  5  5  3  2  1  1
```

μπορούμε να υπολογίσουμε στην R πόσες φορές εμφανίζεται μια συγκεκριμένη τιμή w για τους 70 συνδυασμούς και κάνοντας χρήση του τύπου

$$P_{H_0}(W_s = w) = \frac{\#(w; n, m)}{\binom{N}{n}}$$

υπολογίζουμε τις πιθανότητες του Πίνακα 2.1.1 που φυσικά αθροίζουν στην μονάδα αφού πρόκειται για κατανομή.

Έτσι για να βρούμε την πιθανότητα το στατιστικό W_s να πάρει μια τιμή τόσο ακραία ή και περισσότερο ακραία από το 19 αρκεί να υπολογίσουμε την πιθανότητα

$$\begin{aligned} P_{H_0}(W_s \geq 19) &= P_{H_0}(W_s = 19) + P_{H_0}(W_s = 20) + P_{H_0}(W_s = 21) + P_{H_0}(W_s = 22) \\ &\quad + P_{H_0}(W_s = 23) + P_{H_0}(W_s = 24) + P_{H_0}(W_s = 25) + P_{H_0}(W_s = 26) \\ &= \frac{7}{70} + \frac{7}{70} + \frac{5}{70} + \frac{5}{70} + \frac{3}{70} + \frac{2}{70} + \frac{1}{70} + \frac{1}{70} = \frac{31}{70} = 0.4428571. \end{aligned}$$

Μπορούμε να δούμε από τον Πίνακα 2.1.2 ότι η πιθανότητα που υπολογίζουμε μπορεί να πάρει συγκεκριμένες τιμές μόνο. Όλες οι δυνατές τιμές δίνονται στον Πίνακα 2.1.2

| | | | | |
|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|
| $P_{H_0}(W_s \geq 26)$ = 0.014 | $P_{H_0}(W_s \geq 25)$ = 0.029 | $P_{H_0}(W_s \geq 24)$ = 0.057 | $P_{H_0}(W_s \geq 23)$ = 0.1 | $P_{H_0}(W_s \geq 22)$ = 0.171 |
| $P_{H_0}(W_s \geq 21)$ = 0.243 | $P_{H_0}(W_s \geq 20)$ = 0.343 | $P_{H_0}(W_s \geq 19)$ = 0.443 | $P_{H_0}(W_s \geq 18)$ = 0.557 | $P_{H_0}(W_s \geq 17)$ = 0.657 |
| $P_{H_0}(W_s \geq 16)$ = 0.757 | $P_{H_0}(W_s \geq 15)$ = 0.829 | $P_{H_0}(W_s \geq 14)$ = 0.9 | $P_{H_0}(W_s \geq 13)$ = 0.943 | $P_{H_0}(W_s \geq 12)$ = 0.971 |
| $P_{H_0}(W_s \geq 11)$ = 0.986 | $P_{H_0}(W_s \geq 10)$ = 1 | | | |

Πίνακας 2.1.2: Τιμές του α για το Παράδειγμα 2.1.1.

Συνηθίζεται για να υπολογίσουμε την κρίσιμη τιμή c του ελέγχου να χρησιμοποιούμε ως α την πλησιέστερη τιμή σε αυτές τις τιμές που έχουμε στον Πίνακα 2.1.2.

Υπάρχουν έτοιμοι πίνακες με την κατανομή του W_s καθώς όπως μπορούμε να αντιληφθούμε ο αριθμός $\binom{N}{n}$ μεγαλώνει πάρα πολύ γρήγορα και ο υπολογισμός των πιθανοτήτων με τον παραπάνω τρόπο είναι αρκετά χρονοβόρος. Ωστόσο, στους πίνακες έχει παρατηρηθεί πως η κατανομή του στατιστικού W_s δεν είναι η ίδια για τις τιμές $m = 2, n = 4$ με αυτή που δίνεται για τιμές $m = 4, n = 2$. Για τον λόγο αυτό χρησιμοποιήθηκε ένα άλλο στατιστικό που ορίζεται από τον τύπο

$$W_{XY} = W_s - \frac{1}{2}n(n+1)$$

και του οποίου η κατανομή είναι ίδια για κάθε ακέραιο $k_1 \leq k_2$ είτε για $m = k_1, n = k_2$ είτε για $m = k_2, n = k_1$.

Υπάρχουν έτοιμοι πίνακες που δίνουν την πιθανότητα

$$P_{H_0}(W_{XY} \leq c)$$

την οποία όμως μπορούμε εύκολα να υπολογίσουμε στην R με την εντολή `pwilcox(c,m,n)` για γνωστές τιμές του αριθμού c και γνωστά μεγέθη m, n δειγμάτων.

Ένας συνηθισμένος τρόπος να απορρίπτουμε την μηδενική υπόθεση ότι τα υπο μελέτη χαρακτηριστικά δεν διαφέρουν στους δύο υποπληθυσμούς είναι σύμφωνα με την σχέση

$$P_{H_0}(W_s \geq c) = \alpha$$

και για να αποφανθούμε για την μηδενική υπόθεση κάνοντας χρήση του στατιστικού W_{XY} θα πρέπει να αναγάγουμε την σχέση $P_{H_0}(W_s \geq c)$ στην $P_{H_0}(W_{XY} \leq c)$ κάνοντας χρήση της ιδιότητας της συμμετρίας του στατιστικού W_s γύρω από το σημείο $n(N+1)/2$. Αφού η κατανομή του στατιστικού W_s είναι συμμετρική στο $n(N+1)/2$, θα ισχύει για κάθε k ότι

$$P_{H_0}\left[W_s = \frac{1}{2}n(N+1) - k\right] = P_{H_0}\left[W_s = \frac{1}{2}n(N+1) + k\right]$$

καθώς επίσης και ότι

$$P_{H_0} \left[W_s \leq \frac{1}{2}n(N+1) - k \right] = P_{H_0} \left[W_s \geq \frac{1}{2}n(N+1) + k \right].$$

Παράδειγμα 2.1.2

Έστω πως έχουμε $m = 3$ και $n = 6$ το πλήθος δείγματα και επιθυμούμε να υπολογίσουμε την πιθανότητα $P_{H_0}(W_s \geq 38)$. Αφού υπολογίσουμε τον αριθμό $\frac{1}{2}n(N+1) = \frac{1}{2}6(9+1) = 30$ μπορούμε να υπολογίσουμε την πιθανότητα με τον ακόλουθο τρόπο:

$$\begin{aligned} P_{H_0}(W_s \geq 38) &= P_{H_0}(W_s \geq 30 + 8) = P_{H_0}(W_s \leq 30 - 8) = P_{H_0}(W_s \leq 22) \\ &= P_{H_0}(W_{XY} \leq 1) = 0.02380952. \end{aligned}$$

Η τελική πιθανότητα μπορεί να βρεθεί από τους πίνακες της κατανομής του στατιστικού όμως εδώ την υπολογίσαμε με τη βοήθεια του στατιστικού πακέτου της R με την εντολή

```
> pwilcox(1,6,3)
[1] 0.02380952.
```

Για δεδομένο επίπεδο σημαντικότητας α μπορούμε να υπολογίσουμε μια κρίσιμη τιμή c για την πιθανότητα $P_{H_0}(W_{XY} \leq c) = \alpha$ στην R με την εντολή `qwilcox(α , m , n)` αντί να ψάχνουμε να βρούμε στους πίνακες της κατανομής την τιμή c .

Εναλλακτικά έχουμε την δυνατότητα αντί για το στατιστικό W_s να χρησιμοποιήσουμε το στατιστικό W_r που ορίζεται ως εξής:

$$W_r = R_1 + \dots + R_m$$

Όπου R_1, \dots, R_m είναι το σύνολο των βαθμών των ασθενών που έλαβαν το εικονικό φάρμακο.

Το άθροισμα των συνολικών $1, \dots, N$ βαθμών του ενιαίου δείγματος είναι σταθερό και ίσο με $\frac{1}{2}N(N+1)$ οπότε θα ισχύει:

$$W_r + W_s = \frac{1}{2}N(N+1).$$

Από την παραπάνω σχέση μπορούμε να δούμε ότι μεγάλες τιμές του στατιστικού W_r αντιστοιχούν σε μικρές τιμές του στατιστικού W_s και αντίστροφα. Επομένως το να απορρίψουμε την μηδενική υπόθεση για μεγάλες τιμές του W_s είναι ανάλογο του να απορρίψουμε την μηδενική υπόθεση για μικρές τιμές του στατιστικού W_r , παραδείγματος χάριν για τιμές

$$W_r \leq c'.$$

Η τιμή της σταθεράς c' καθορίζεται σύμφωνα με την πιθανότητα $P_{H_0}(W_r \leq c') = \alpha$.

Όμοια με πριν είναι προτιμότερο να κάνουμε χρήση του στατιστικού

$$W_{YX} = W_r - \frac{1}{2}m(m+1).$$

Τα στατιστικά W_{YX} , W_{XY} έχουν την ίδια κατανομή υπό την μηδενική υπόθεση, επομένως θα ισχύει

$$P_{H_0}(W_{XY} \leq c) = P_{H_0}(W_{YX} \leq c) \quad , \quad \text{για όλες τις τιμές του } c.$$

Σε έναν έλεγχο υποθέσεων, αντί να αναφέρουμε απλά αν απορρίπτεται ή όχι η μηδενική υπόθεση για ένα καθορισμένο επίπεδο σημαντικότητας α θα πρέπει να υπολογίζουμε επιπρόσθετα και την p -τιμή του ελέγχου, την πιθανότητα δηλαδή να παρατηρήσουμε μια τιμή τόσο ακραία ή περισσότερο ακραία από την παρατηρούμενη τιμή w του στατιστικού.

Η πιθανότητα αυτή ορίζεται ως $P_{H_0}(W_s \geq w)$ και η μηδενική υπόθεση απορρίπτεται όταν είναι $P_{H_0}(W_s \geq w) < \alpha$ όπου α είναι το επίπεδο σημαντικότητας του ελέγχου. Την p -τιμή μας την δίνει το πρόγραμμα της R όταν τρέχουμε τον έλεγχο και δεν χρειάζεται να κάνουμε όλους του υπολογισμούς όπως στο Παράδειγμα 2.1.1.

Έστω οι παρατηρήσεις X_1, \dots, X_m για τα άτομα που έλαβαν το εικονικό φάρμακο ενώ Y_1, \dots, Y_n οι παρατηρήσεις όσων έλαβαν τη νέα θεραπεία. Μπορούν να προκύψουν συνολικά mn ζεύγη παρατηρήσεων (X_i, Y_j) για κάποια εκ των οποίων θα ισχύει

$X_i < Y_j$ ενώ για άλλα θα είναι $X_i > Y_j$. Τότε μπορούμε εναλλακτικά να ορίσουμε τα ακόλουθα στατιστικά για να ελέγξουμε την μηδενική υπόθεση

$$W_{XY} = \text{το πλήθος των ζευγών για τα οποία ισχύει } X_i < Y_j$$

και

$$W_{YX} = \text{το πλήθος των ζευγών για τα οποία ισχύει } Y_j < X_i$$

που είναι γνωστά ως στατιστικά Mann-Whitney.

Πρόταση 2.1.1

$$W_{XY} = W_s - \frac{1}{2}n(n+1)$$

Απόδειξη 2.1.1:

Γνωρίζουμε ότι το στατιστικό W_{XY} ορίζεται ως

$$W_{XY} = \text{το πλήθος των ζευγών για τα οποία ισχύει } X_i < Y_j.$$

Έστω $Y_{(1)} < \dots < Y_{(n)}$ οι διατεταγμένες παρατηρήσεις που αφορούν στα άτομα που έλαβαν την νέα θεραπεία. Τότε για κάθε j θα βρούμε το πλήθος των παρατηρήσεων της τυχαίας μεταβλητής X που είναι μικρότερες από την Y_j παρατήρηση.

Αν ο βαθμός της παρατήρησης $Y_{(1)}$ είναι S_1 τότε υπάρχουν $S_1 - 1$ το πλήθος μικρότερες παρατηρήσεις που φυσικά μπορούν να είναι όλες μόνο παρατηρήσεις της τυχαίας μεταβλητής X αφού η $Y_{(1)}$ είναι η μικρότερη παρατήρηση από τις Y . Όμοια υπάρχουν $S_2 - 1$ παρατηρήσεις μικρότερες από την $Y_{(2)}$ παρατήρηση εκ των οποίων η μια είναι η $Y_{(1)}$ ενώ οι υπόλοιπες $(S_2 - 1) - 1 = S_2 - 2$ είναι παρατηρήσεις της τυχαίας μεταβλητής X . Συνεχίζοντας με αυτή τη λογική, αφού το πλήθος των παρατηρήσεων Y που είναι μικρότερες από την παρατήρηση $Y_{(j)}$ είναι $j - 1$, το πλήθος των X παρατηρήσεων που είναι μικρότερες από την $Y_{(j)}$ είναι $(S_j - 1) - (j - 1) = S_j - j$.

Το συνολικό πλήθος των ζευγών (X_i, Y_j) για τα οποία ισχύει $X_i < Y_j$ είναι επομένως το άθροισμα

$$(S_1 - 1) + (S_2 - 2) + \dots + (S_n - n) = W_s - (1 + 2 + \dots + n) = W_s - 1/2 n(n + 1) = W_{XY}.$$

Πρόταση 2.1.2

Η κατανομή του στατιστικού W_s είναι συμμετρική γύρω από το $1/2 n(N + 1)$

Απόδειξη 2.1.2

Αν τα $N = n + m$ το πλήθος υποκείμενα που λαμβάνουν μέρος στην έρευνα βαθμολογηθούν με αντίθετο τρόπο, δηλαδή η παρατήρηση που πριν είχε τον βαθμό 1 θα έχει τώρα τον βαθμό N , αυτή με τον βαθμό 2 θα έχει τον βαθμό $N - 2$ και γενικά η παρατήρηση με τον βαθμό S θα έχει τώρα βαθμό ίσο με $N - S + 1$, τότε ο νέος βαθμός της παρατήρησης Y που πριν είχε βαθμό S_i είναι τώρα S'_i . Ακολουθώντας την ίδια επιχειρηματολογία με αυτή στην αρχή του κεφαλαίου τότε μπορούμε να βρούμε πως η πιθανότητα να παρατηρήσουμε την n -αδα (S'_1, \dots, S'_n) δίνεται πάλι από την σχέση

$$P_{H_0}(S'_1 = s_1, \dots, S'_n = s_n) = 1 / \binom{N}{n}$$

και επομένως το στατιστικό $W'_s = S'_1 + \dots + S'_n$ έχει την ίδια κατανομή με το W_s υπό την μηδενική υπόθεση. Αφού ισχύει ότι $S'_i = N - S_i + 1$ η τελευταία σχέση γίνεται

$$W'_s = S'_1 + \dots + S'_n = (N - S_1 + 1) + \dots + (N - S_n + 1) = n(N + 1) - (S_1 + \dots + S_n) = n(N + 1) - W_s$$

και προκύπτει ότι

$$W'_s = n(N + 1) - W_s \leftrightarrow W'_s - \frac{1}{2}n(N + 1) = \frac{1}{2}n(N + 1) - W_s.$$

Επομένως

$$P_{H_0}\left(W'_s - \frac{1}{2}n(N + 1) = k\right) = P_{H_0}\left(\frac{1}{2}n(N + 1) - W_s = k\right)$$

άρα τα στατιστικά έχουν την ίδια κατανομή που πράγματι είναι συμμετρική.

2.2 Η κατανομή των στατιστικών του Wilcoxon και των Mann-Whitney υπό την μηδενική υπόθεση για δείγματα που έχουν επιλεγθεί κατά τυχαίο τρόπο από έναν πληθυσμό.

Στην έρευνα που περιγράψαμε στην προηγούμενη ενότητα έχει γίνει μια σημαντική παράλειψη. Αφού δεν έχουμε κάνει καμία υπόθεση που να αφορά τα N το πλήθος άτομα που συμμετείχαν στην έρευνα αυτή, τα αποτελέσματα της έρευνας θα αφορούν αποκλειστικά και μόνο τα συγκεκριμένα άτομα που συμμετείχαν και δε μπορούμε σε καμία περίπτωση να τα γενικεύσουμε για ολόκληρο τον πληθυσμό. Δηλαδή, αν το φάρμακο για την μείωση της χοληστερίνης αποδειχτεί πως είχε θετικά αποτελέσματα για το δείγμα που το έλαβε δεν

μπορούμε να είμαστε απολύτως σίγουροι ότι τα αποτελέσματα θα είναι αντίστοιχα θετικά και σε όλους τους πιθανούς χρήστες του. Για να συμβεί αυτό θα πρέπει το δείγμα να έχει επιλεγεί με τέτοιο τρόπο ώστε να είναι αντιπροσωπευτικό ολόκληρου του πληθυσμού. Ο τρόπος δειγματοληψίας για να διαθέτει το δείγμα το χαρακτηριστικό αυτό είναι η απλή τυχαία δειγματοληψία, όπου κάθε μονάδα του πληθυσμού έχει την ίδια πιθανότητα να επιλεγεί στο τυχαίο δείγμα.

Έστω ότι $N = n + m$ το πλήθος άτομα επιλέγονται από τον πληθυσμό με απλή τυχαία δειγματοληψία και ότι n το πλήθος από αυτά λαμβάνουν μια νέα θεραπεία ενώ m το πλήθος από αυτά λαμβάνουν την υπάρχουσα θεραπεία. Έστω X, Y οι τυχαίες μεταβλητές που περιέχουν τις παρατηρήσεις όσων έλαβαν την παλιά και την νέα θεραπεία αντίστοιχα.

Οι αθροιστικές συναρτήσεις των τυχαίων μεταβλητών X, Y είναι οι:

$$P(X \leq x) = F(x)$$

$$P(Y \leq y) = G(y)$$

και η μηδενική υπόθεση, η υπόθεση δηλαδή ότι η νέα θεραπεία δεν έχει κάποιο αποτέλεσμα είναι στην ουσία η υπόθεση ότι οι μεταβλητές X, Y έχουν την ίδια συνάρτηση κατανομής. Δηλαδή είναι

$$H_0: F = G.$$

Δύο επιπρόσθετες υποθέσεις που πρέπει να κάνουμε είναι ότι ο πληθυσμός από τον οποίο επιλέγουμε το δείγμα είναι αρκετά μεγάλος ώστε οι μεταβλητές X, Y να μπορούν να θεωρηθούν ανεξάρτητες και ότι δεν υπάρχουν παρατηρήσεις με την ίδια τιμή. Αναγκαία και ικανή συνθήκη για να μην υπάρχουν παρατηρήσεις με ίδιες τιμές είναι ότι οι κατανομές των μεταβλητών είναι συνεχείς. Υπό την μηδενική υπόθεση, η από κοινού κατανομή των βαθμών δίνεται από την πρόταση 2.2.1.

Πρόταση 2.2.1

Έστω ότι οι παρατηρήσεις X_1, \dots, X_m και Y_1, \dots, Y_n είναι ανεξάρτητα κατανεμημένες σύμφωνα με μια κοινή συνεχή κατανομή F και έστω S_1, \dots, S_n είναι οι βαθμοί της Y μεταβλητής στο ενιαίο δείγμα των βαθμών για τις $N = n + m$ παρατηρήσεις, ώστε να ισχύει $S_1 < \dots < S_n$. Τότε κάθε δυνατή n -άδα έστω s_1, \dots, s_n βαθμών που μπορεί να παρατηρηθεί έχει πιθανότητα που δίνεται από την παρακάτω σχέση:

$$P_{H_0}(S_1 = s_1, \dots, S_n = s_n) = \frac{1}{\binom{N}{n}}.$$

Απόδειξη 2.2.1:

Θεωρούμε το ενιαίο δείγμα με τις X_1, \dots, X_m και τις Y_1, \dots, Y_n και το συμβολίζουμε με $Z_1, \dots, Z_{N=m+n}$. Τότε η πιθανότητα που θέλουμε να υπολογίσουμε είναι η πιθανότητα ότι οι βαθμοί των παρατηρήσεων Z_{m+1}, \dots, Z_{m+n} είναι οι s_1, \dots, s_n όχι απαραίτητα με αυτή τη σειρά. Αφού οι s_1, \dots, s_n είναι οι βαθμοί από n παρατηρήσεις της μεταβλητής Z και η από κοινού κατανομή των N παρατηρήσεων Z είναι συμμετρική, κάθε υποσύνολο n το πλήθος των Z είναι εξίσου πιθανό να έχει ως βαθμούς το σύνολο s_1, \dots, s_n . Υπάρχουν $\binom{N}{n}$ τέτοια υποσύνολα των Z οι βαθμοί των οποίων έχουν πιθανότητα $\frac{1}{\binom{N}{n}}$ να είναι ίσοι με s_1, \dots, s_n . Έτσι αυτή είναι η πιθανότητα το σύνολο s_1, \dots, s_n να είναι οι βαθμοί των παρατηρήσεων Z_{m+1}, \dots, Z_{m+n} , δηλαδή των παρατηρήσεων Y_1, \dots, Y_n .

Επομένως υπό την μηδενική υπόθεση η κατανομή του στατιστικού δεν εξαρτάται από την συνάρτηση κατανομής F .

2.3 Ασυμπτωτική κατανομή των στατιστικών του Wilcoxon και των Mann-Whitney

Για μεγάλες τιμές των μεγεθών των δειγμάτων, δηλαδή για μεγάλα m , n μπορούμε να εφαρμόσουμε το Κεντρικό Οριακό Θεώρημα (central limit theorem) για να προσεγγίσουμε τις κατανομές των στατιστικών που αναφέραμε στην προηγούμενη ενότητα. Σύμφωνα με το Κ.Ο.Θ το άθροισμα ενός μεγάλου αριθμού ανεξάρτητων τυχαίων μεταβλητών ακολουθεί προσεγγιστικά την κανονική κατανομή.

Έστω $\{X_1, \dots, X_n\}$ τυχαίο δείγμα μεγέθους n ($n \geq 30$) ανεξάρτητων τυχαίων μεταβλητών που προέρχονται από κατανομή με πεπερασμένη μέση τιμή μ και διασπορά σ^2 τότε:

$$S_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow{D} Z \sim N(0,1)$$

με την παραπάνω σύγκλιση να είναι κατά νόμο καθώς το $n \rightarrow \infty$.

Αν το πλήθος του δείγματος, δηλαδή ο αριθμός n είναι μεγάλος αριθμός, τότε και το άθροισμα των βαθμών των παρατηρήσεων που έλαβαν την νέα θεραπεία θα είναι μεγάλος αριθμός. Ωστόσο, αυτές οι τυχαίες μεταβλητές δεν είναι ανεξάρτητες μεταξύ τους. Αν όμως και το μέγεθος από το δεύτερο δείγμα, δηλαδή ο αριθμός m είναι μεγάλος αριθμός, τότε η εξάρτηση εξασθενεί και μπορούμε να εφαρμόσουμε το Κ.Ο.Θ για να προσεγγίσουμε την κατανομή του Wilcoxon από την κανονική κατανομή.

Χρειαζόμαστε την μέση τιμή και την διασπορά για κάθε στατιστικό. Αυτές οι τιμές δίνονται από τις ακόλουθες σχέσεις:

για το στατιστικό W_s η μέση τιμή $E(W_s)$ και η διασπορά $Var(W_s)$ δίνονται από τις σχέσεις:

$$E(W_s) = \frac{1}{2}n(N+1)$$

$$Var(W_s) = \frac{1}{12}mn(N+1)$$

ενώ οι αντίστοιχες σχέσεις για τα στατιστικά W_r και W_{XY} είναι οι

$$E(W_r) = \frac{1}{2}m(N+1)$$

$$E(W_{XY}) = \frac{1}{2}mn$$

και για την διασπορά

$$Var(W_r) = Var(W_{XY}) = \frac{1}{12}mn(N+1).$$

Επομένως η προσέγγιση από την κανονική κατανομή θα γίνεται με την χρήση της σχέσης:

$$P(W_s \leq c) = P\left(\frac{W_s - E(W_s)}{\sqrt{Var(W_s)}} \leq \frac{c - E(W_s)}{\sqrt{Var(W_s)}}\right) \approx \Phi\left(\frac{c - E(W_s)}{\sqrt{Var(W_s)}}\right).$$

Παράδειγμα 2.3.1

Έστω ότι θέλουμε για μεγέθη δειγμάτων ίσα με $n = m = 10$ να υπολογίσουμε την πιθανότητα $P(W_s \leq 79)$. Τότε υπολογίζουμε την μέση τιμή και την διασπορά από τις σχέσεις και παίρνουμε $E(W_s) = 105$ και $Var(W_s) = 175$, επομένως η πιθανότητα που επιθυμούμε να υπολογίσουμε θα είναι στην ουσία η πιθανότητα

$$P(W_s \leq 79) = P\left(\frac{W_s - 105}{\sqrt{175}} \leq \frac{79 - 105}{\sqrt{175}}\right) \approx \Phi(-1.965)$$

και από την R βρίσκουμε

```
> pnorm(-1.965)
[1] 0.02470712.
```

Μια καλύτερη προσέγγιση της κατανομής μπορεί να δοθεί από την σχέση

$$P(W_s \leq c) \approx \Phi\left(\frac{c - E(W_s) + 1/2}{\sqrt{Var(W_s)}}\right)$$

που καλείται διόρθωση συνέχειας.

2.4 Όταν δύο ή περισσότερες παρατηρήσεις έχουν ίσες τιμές

Μέχρι τώρα έχουμε υποθέσει ότι στις παρατηρήσεις μας δεν υπήρχαν ίσες τιμές. Βέβαια αυτή η συνθήκη δεν ισχύει καθώς σε πραγματικές έρευνες συναντάμε πολλές φορές ίσες τιμές παρατηρήσεων και ειδικότερα όταν οι παρατηρήσεις είναι κατηγορικές παρατηρήσεις.

Παράδειγμα 2.4.1

Έστω ότι έχουμε μια έρευνα με πλήθος $N = n + m = 4$ παρατηρήσεις (όπου $m = n = 2$) και οι τιμές των παρατηρήσεων είναι οι 12.4, 11.9, 12.4, 13.2, τότε εύκολα διαπιστώνουμε ότι η παρατήρηση 11.9 θα πάρει τον βαθμό 1 αφού είναι η παρατήρηση με την μικρότερη τιμή, ενώ η 13.2 που είναι η παρατήρηση με την μεγαλύτερη τιμή θα πάρει τον βαθμό 4. Τι συμβαίνει όμως με τις άλλες δύο τιμές;

Φαίνεται λογικό ότι αφού οι παρατηρήσεις έχουν την ίδια τιμή θα πρέπει να έχουν και τον ίδιο βαθμό και αυτό μπορεί να γίνει αν αποδώσουμε στην κάθε παρατήρηση τον μέσο όρο του αθροίσματος των βαθμών που θα είχαν αν οι τιμές τους ήταν διακεκριμένες. Επομένως στο συγκεκριμένο παράδειγμα οι βαθμοί των τιμών που είναι ίσες με 12.4 θα είναι

$$\frac{(2 + 3)}{2} = 2.5$$

και η μέση βαθμολογία (midranks) για όλες τις παρατηρήσεις για το Παράδειγμα 2.4.1 θα είναι

$$1, 2.5, 2.5, 4.$$

Συμβολίζουμε τους μέσους βαθμούς για όσους ασθενείς έλαβαν την νέα θεραπεία με S_1^*, S_2^* ενώ τους μέσους βαθμούς για όσους ασθενείς έλαβαν την παλιά θεραπεία τους συμβολίζουμε με R_1^*, R_2^* .

Όταν υπάρχουν δεσμοί (ίσες τιμές) επομένως, το νέο στατιστικό ελέγχου θα δίνεται από την σχέση

$$W_s^* = S_1^* + S_2^*.$$

Η κατανομή του στατιστικού υπό την μηδενική υπόθεση που είδαμε στις προηγούμενες ενότητες δεν ισχύει καθώς οι τιμές των S_1^*, S_2^* δεν είναι ίδιες με αυτές των S_1, S_2 , ωστόσο, ισχύει η ίδια επιχειρηματολογία. Υπό την μηδενική υπόθεση ότι η θεραπεία δεν έχει κανένα αποτέλεσμα, οι τέσσερις ασθενείς που συμμετέχουν στην έρευνα θα πάρουν τους ίδιους μέσους βαθμούς που θα έπαιρναν και πριν την έρευνα, ανεξάρτητα από το αν έλαβαν την νέα ή την παλιά αγωγή. Επομένως η τυχαία επιλογή των δύο ατόμων που θα λάβουν την νέα θεραπεία έχει ως αποτέλεσμα την τυχαία επιλογή και των μέσων βαθμών τους, επομένως οι $\binom{4}{2}$ δυνατοί συνδυασμοί των μέσων βαθμών S_1^*, S_2^* είναι εξίσου πιθανοί.

Υπάρχουν συνολικά 6 δυνατοί συνδυασμοί των S_1^*, S_2^* αφού από την R προκύπτει ότι είναι

```
> choose(4,2)
[1] 6
```

και παρουσιάζονται όλοι στον Πίνακα 2.4.1.

| | |
|-----------------------------------|-----------------------------------|
| $P(S_1^* = 1, S_2^* = 2.5) = 2/6$ | $P(S_1^* = 2.5, S_2^* = 4) = 2/6$ |
| $P(S_1^* = S_2^* = 2.5) = 1/6$ | $P(S_1^* = 1, S_2^* = 4) = 1/6$ |

Πίνακας 2.4.1: Όλοι οι δυνατοί συνδυασμοί για τις τιμές των S_1^*, S_2^* και η πιθανότητα με την οποία προκύπτει ο κάθε ένας.

Στον Πίνακα 2.4.2 φαίνεται η κατανομή του στατιστικού $W_s^* = S_1^* + S_2^*$:

| | | | |
|----------------------|-----|-----|-----|
| w | 3.5 | 5 | 7.5 |
| $P_{H_0}(W_s^* = w)$ | 2/6 | 2/6 | 2/6 |

Πίνακας 2.4.2 Η κατανομή του στατιστικού W_s^* .

Όπως καταλαβαίνουμε η κατανομή του στατιστικού W_s^* δεν εξαρτάται μόνο από τα μεγέθη (m και n) των δειγμάτων αλλά και από το πλήθος των παρατηρήσεων που έχουν ίδια τιμή (το πλήθος των δεσμών) και για το λόγο αυτό είναι πιο πρακτικό να προσεγγίσουμε την κατανομή του W_s^* από την κανονική κατανομή. Αυτό μπορεί να συμβεί για μεγέθη (m και n) δειγμάτων που δεν είναι πολύ μικρά και το μέγιστο πλήθος των παρατηρήσεων που έχουν ίδια τιμή δεν είναι κοντά στην μονάδα.

Η μέση τιμή και η διασπορά του στατιστικού W_s^* δίνονται από τον τύπο

$$E(W_s^*) = \frac{1}{2} n (N + 1)$$

για την μέση τιμή και τον τύπο

$$Var(W_s^*) = \frac{mn(N+1)}{12} - \frac{mn \sum_{i=1}^e (d_i^3 - d_i)}{12N(N-1)}$$

για την διασπορά, όπου για την διασπορά το e είναι οι διακεκριμένες τιμές που μπορούν να πάρουν οι N το πλήθος παρατηρήσεις και το d_1 συμβολίζει το πλήθος των N παρατηρήσεων που έχουν την μικρότερη τιμή από τις e διακεκριμένες τιμές, το d_2 το πλήθος των N παρατηρήσεων που έχουν της αμέσως μικρότερη τιμή από τις e διακεκριμένες τιμές, ..., d_e το πλήθος των N παρατηρήσεων που έχουν την μεγαλύτερη τιμή από τις e διακεκριμένες τιμές.

Παράδειγμα 2.4.2

Αν οι παρατηρήσεις σε μια έρευνα είναι οι 2, 4, 5, 2, 4, 5, 2, 1 τότε το $e = 4$ και $d_1 = 1, d_2 = 3, d_3 = 2, d_4 = 2$.

Ο δεύτερος όρος στον τύπο της διασποράς αφορά της διόρθωση για τις ισοβαθμίες που προκύπτουν.

Επομένως ισχύει

$$\frac{W_s^* - E(W_s^*)}{\sqrt{Var(W_s^*)}} \rightarrow \Phi$$

για

$$\max_{i=1, \dots, e} \left(\frac{d_i}{N} \right) \text{ που απομακρύνεται από το } 1 \text{ καθώς } N \rightarrow \infty.$$

Είχαμε ορίσει το στατιστικό W_{XY} ως εξής

$$W_{XY} = \text{το πλήθος των ζευγών για τα οποία ισχύει } X_i < Y_j.$$

Κάθε ζεύγος παρατηρήσεων (X_i, Y_j) θα έχει την τιμή 1 αν είναι $X_i < Y_j$ ή την τιμή 0 διαφορετικά και το W_{XY} είναι το άθροισμα των τιμών αυτών. Όπως είδαμε σε αυτή την ενότητα, μπορούν να υπάρχουν σε μια έρευνα παρατηρήσεις για τις οποίες ισχύει $X_i = Y_j$ και επομένως σύμφωνα με όσα αναφέραμε είναι λογικό να δώσουμε την τιμή $\frac{1}{2}$ στα ζεύγη αυτά, επομένως το στατιστικό των Mann-Whitney ορίζεται τώρα ως

$$W_{XY}^* = \text{το πλήθος των ζευγών για τα οποία ισχύει } X_i < Y_j + \frac{1}{2} (\text{το πλήθος των ζευγών για τα οποία ισχύει } X_i = Y_j).$$

Αντίστοιχα, για την προσέγγιση της κατανομής του W_{XY}^* από την κανονική κατανομή η μέση τιμή και η διασπορά του δίνονται από τους τύπους

$$E(W_{XY}^*) = \frac{1}{2} mn$$

και

$$Var(W_{XY}^*) = Var(W_s^*).$$

Πρόταση 2.4.1

Ισχύει $W_{XY}^* = W_s^* - \frac{1}{2}n(n+1)$.

Απόδειξη 2.4.1

Αν d_1 το πλήθος παρατηρήσεις έχουν την μικρότερη από τις e διακεκριμένες τιμές των παρατηρήσεων τότε ο βαθμός των d_1 αυτών παρατηρήσεων είναι ο μέσος όρος των βαθμών $1, \dots, d_1$, δηλαδή

$$\frac{1 + \dots + d_1}{d_1} = \frac{1}{2}(d_1 + 1).$$

Αντίστοιχα οι παρατηρήσεις που έχουν την αμέσως μικρότερη τιμή καταλαμβάνουν τις θέσεις $d_1 + 1, d_1 + 2, \dots, d_1 + d_2$ και επομένως ο μέσος όρος των βαθμών τους θα είναι

$$\frac{(d_1 + 1) + (d_1 + 2) + \dots + (d_1 + d_2)}{d_2} = d_1 + \frac{1}{2}(d_2 + 1).$$

Έτσι με την ίδια λογική, ο μέσος όρος των βαθμών των παρατηρήσεων που έχουν την i -οστή τιμή θα είναι

$$\frac{(d_1 + \dots + d_{i-1} + 1) + \dots + (d_1 + \dots + d_{i-1} + d_i)}{d_i} = d_1 + \dots + d_{i-1} + \frac{1}{2}(d_i + 1).$$

Έστω τώρα πως από τις d_1 παρατηρήσεις που έχουν την μικρότερη τιμή, οι A_1 είναι παρατηρήσεις της τυχαίας μεταβλητής X ενώ οι B_1 είναι παρατηρήσεις της τυχαίας μεταβλητής Y , αντίστοιχα από τις d_i παρατηρήσεις, οι A_i είναι της μεταβλητής X ενώ οι B_i της μεταβλητής Y κ.ο.κ.

Τότε το άθροισμα των μέσων βαθμών των Y παρατηρήσεων μπορεί να γραφτεί ως

$$W_s^* = B_1 \left(\frac{1}{2}(d_1 + 1) \right) + B_2 \left(d_1 + \frac{1}{2}(d_2 + 1) \right) + B_3 \left(d_1 + d_2 + \frac{1}{2}(d_3 + 1) \right) + \dots$$

και για να βρούμε τον αντίστοιχο τύπο για το στατιστικό W_{XY}^* αρκεί να σκεφτούμε πως για την μικρότερη τιμή του Y , δεν υπάρχουν X μικρότερα από το Y και το πλήθος των X που είναι ίσα με Y είναι ίσο με A_1 , επομένως η συμβολή των B_1 τέτοιων Y για τον υπολογισμό του W_{XY}^* θα είναι $\frac{1}{2}A_1$. Για την αμέσως επόμενη τιμή, θα υπάρχουν A_1 X που είναι μικρότερα από το Y και A_2 που θα έχουν ίση τιμή με το Y , επομένως η συμβολή των B_2 θα είναι ίση με $A_1 + \frac{1}{2}A_2$ και αν συνεχίσουμε με αυτή τη λογική βρίσκουμε ότι θα είναι

$$W_{XY}^* = B_1 \left(\frac{1}{2}A_1 \right) + B_2 \left(A_1 + \frac{1}{2}A_2 \right) + B_3 \left(A_1 + A_2 + \frac{1}{2}A_3 \right) + \dots$$

Αφού ισχύει ότι $A_i + B_i = d_i$ η παραπάνω σχέση γίνεται

$$W_s^* - W_{XY}^* = B_1 \left[\frac{1}{2}(B_1 + 1) \right] + B_2 \left[B_1 + \frac{1}{2}(B_2 + 1) \right] + B_3 \left[B_1 + B_2 + \frac{1}{2}(B_3 + 1) \right] + \dots$$

Το άθροισμα των όρων που δεν έχουν μπροστά το $\frac{1}{2}$ δίνεται από την σχέση

$$B_2B_1 + B_3(B_1 + B_2) + \dots$$

και ισχύει

$$\sum_{i>j} B_iB_j = \frac{1}{2} \sum_{i \neq j} B_iB_j,$$

επομένως η σχέση που υπολογίσαμε μπορεί να γραφτεί ως

$$W_s^* - W_{XY}^* = \frac{1}{2} \left[\sum B_i + \sum B_i^2 + \sum_{i \neq j} B_iB_j \right] = \frac{1}{2} \left[\sum B_i + \left(\sum B_i \right)^2 \right]$$

και επειδή ισχύει

$$\sum B_i = n$$

η σχέση αποδείχθηκε.

2.5 Η εντολή του ελέγχου στο στατιστικό πακέτο της R

Η εντολή στην R για να πραγματοποιήσουμε έναν έλεγχο Mann-Whitney για ανεξάρτητες παρατηρήσεις είναι η `wilcox.test(x, y)`.

Ορίσματα εντολής

Η εντολή παίρνει ως ορίσματα δύο διανύσματα **x**, **y** που περιέχουν τις παρατηρήσεις από το 1^ο και το 2^ο δείγμα αντίστοιχα. Επιπλέον, το όρισμα **alternative** είναι τύπου string και καθορίζει ανάλογα την τιμή του αν ο έλεγχος θα είναι αμφίπλευρος ή μονόπλευρος. Έχει επομένως ως δυνατές τιμές τις **“two.sided”**, η οποία είναι και η προκαθορισμένη από την R τιμή, **“greater”** ή **“less”**. Μπορούμε να το γράψουμε είτε ολογράφως είτε δίνοντας μόνο το αρχικό γράμμα. Π.χ.

```
wilcox.test(x, y, alternative = "l")
```

Τα ορίσματα **exact** και **correct** παίρνουν λογικές τιμές TRUE/FALSE και το πρώτο είναι για τον υπολογισμό της ακριβής τιμής της P-τιμής ενώ το δεύτερο για το αν πρέπει να εφαρμοστεί προσέγγιση από την κανονική τιμή ή όχι.

Αφού περάσουμε τα δεδομένα και τρέξουμε την εντολή του ελέγχου, η R θα μας δώσει ως απάντηση τα αποτελέσματα του ελέγχου που θα είναι της μορφής:

```
> wilcox.test(x, y, correct=FALSE)
```

```
Wilcoxon rank sum test
```

```
data: x and y
```

```
W = ..., p-value = ...
```

```
alternative hypothesis: true location shift is not equal to 0
```

όπου,

x, **y** με το πορτοκαλί θα είναι τα δύο διανύσματα με τις παρατηρήσεις.

W θα είναι η τιμή του στατιστικού ελέγχου.

p-value θα είναι η p-τιμή η οποία αν είναι < του επίπεδου σημαντικότητας α θα σημαίνει ότι έχουμε σημαντικές ενδείξεις για απόρριψη της μηδενικής υπόθεσης.

Alternative hypothesis δείχνει ποια είναι η εναλλακτική υπόθεση H_1 , δηλαδή τη μορφή του ελέγχου, αμφίπλευρος ή μονόπλευρος.

Ο εν λόγω έλεγχος είναι ο μη παραμετρικός έλεγχος που είναι ανάλογος του παραμετρικού **two sample t-test**.

2.5.1 Εφαρμογή

Μια φαρμακευτική δημιούργησε ένα νέο φάρμακο για την υπνοβασία και προκειμένου να ελέγξει την αποτελεσματικότητά του εξέτασε 6 ανθρώπους ένα μήνα αφού λάμβαναν την νέα αγωγή. Μια άλλη ομάδα 6 ατόμων λάμβανε το παλιό χάπι για το ίδιο χρονικό διάστημα. Για την μελέτη, ζητήθηκε από τους 6 αυτούς ανθρώπους να καταγράψουν τις φορές που υπνοβάτησαν μέσα σε αυτό τον ένα μήνα. Τα αποτελέσματα της μελέτης παρουσιάζονται στον Πίνακα 2.5.1

| Πόσες φορές υπνοβάτησε ο ασθενής στη διάρκεια του ενός μήνα | |
|---|----------------------------------|
| Ο ασθενής λαμβάνει την παλιά αγωγή | Ο ασθενής λαμβάνει την νέα αγωγή |
| 4 | 3 |
| 7 | 1 |
| 8 | 2 |
| 5 | 4 |
| 9 | 1 |
| 6 | 1 |

Πίνακας 2.5.1: Πόσες φορές υπνοβάτησε κάθε ασθενής σε ένα μήνα.

Αυτό που μας ενδιαφέρει να ελέγξουμε είναι αν οι δύο ομάδες ασθενών κατέγραψαν τα ίδια επεισόδια υπνοβασίας. Ο έλεγχος δηλαδή είναι ο παρακάτω:

H_0 : οι δύο ομάδες ανέφεραν ίσα σε αριθμό επεισόδια υπνοβασίας

H_1 : οι δύο ομάδες ανέφεραν διαφορετικούς αριθμούς επεισοδίων υπνοβασίας.

Θα πρέπει τώρα να διατάξουμε τις παρατηρήσεις μας κατά αύξουσα σειρά και να δώσουμε τους αντίστοιχους βαθμούς (ranks). Αυτό φαίνεται στον Πίνακα 2.5.2.

| | N.A | N.A | N.A | N.A | N.A | N.A | Π.A | Π.A | Π.A | Π.A | Π.A | Π.A |
|---------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Δείγμα | 1 | 1 | 1 | 2 | 3 | 4 | 4 | 5 | 6 | 7 | 8 | 9 |

Πίνακας 2.5.2: Ενιαίο δείγμα παρατηρήσεων.

Βλέπουμε από το συνολικό δείγμα ότι υπάρχουν δεσμοί, πρέπει επομένως να προσέξουμε να δώσουμε στις παρατηρήσεις που είναι ίσες, ίσους μέσους βαθμούς.

| | N.A | N.A | N.A | N.A | N.A | N.A | Π.A | Π.A | Π.A | Π.A | Π.A | Π.A |
|---------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Δείγμα | 1 | 1 | 1 | 2 | 3 | 4 | 4 | 5 | 6 | 7 | 8 | 9 |
| Βαθμός | 2 | 2 | 2 | 4 | 5 | 6.5 | 6.5 | 8 | 9 | 10 | 11 | 12 |

Πίνακας 2.5.3: Βαθμοί του ενιαίου δείγματος παρατηρήσεων.

Αρχικά πρέπει να περάσουμε τις παρατηρήσεις μας στο πρόγραμμα. Αυτό θα γίνει με την δημιουργία δύο διανυσμάτων.

```
> old <- c(4,7,8,5,9,6)
```

```
> old
```

```
[1] 4 7 8 5 9 6
```

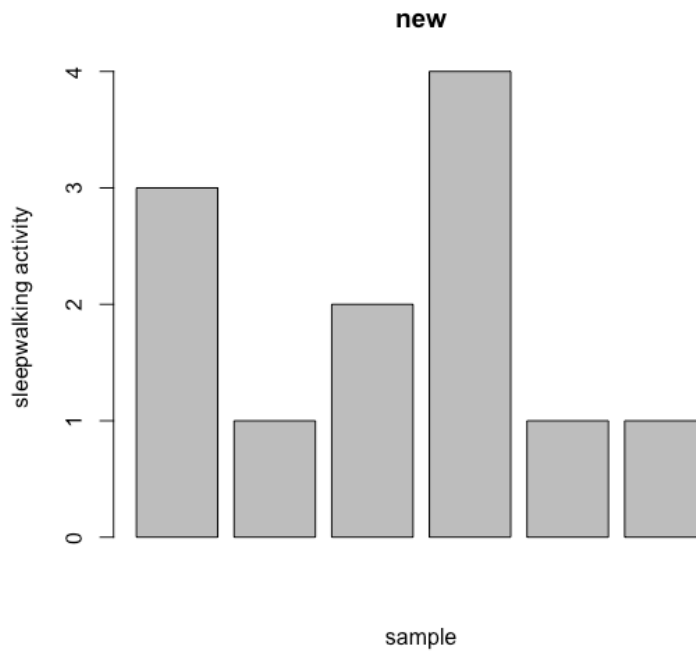
```
> new <- c(3,1,2,4,1,1)
```

```
> new
```

```
[1] 3 1 2 4 1 1
```

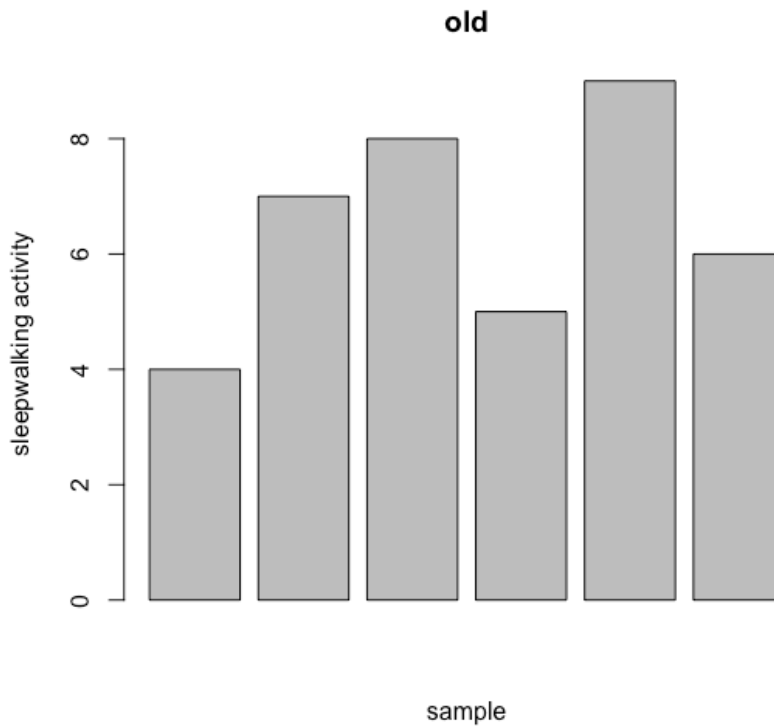
Μπορούμε να κάνουμε τα ραβδογράμματα για το κάθε δείγμα για να έχουμε μια εικόνα αναφορικά με τις συχνότητες υπνοβασίας του κάθε δείγματος.

```
> barplot(new,xlab = "sample" ,ylab="sleepwalking activity",main = "new")
```



Διάγραμμα 2.5.1: Για τα δεδομένα του Πίνακα 2.5.1 που αφορά όσους έλαβαν τη νέα θεραπεία.

```
> barplot(old,xlab = "sample" ,ylab="sleepwalking activity",main = "old")
```



Διάγραμμα 2.5.2: Για τα δεδομένα του Πίνακα 2.5.1 που αφορά όσους έλαβαν την παλιά θεραπεία.

Με μια πρώτη ματιά μπορούμε να δούμε αμέσως ότι τα περιστατικά υπνοβασίας ήταν περισσότερα στα άτομα που έλαβαν την παλιά αγωγή.

Για να βγάλουμε καλύτερα συμπεράσματα ωστόσο θα τρέξουμε τον έλεγχο:

```
> wilcox.test(old, new, correct = F)
```

Wilcoxon rank sum test

data: old and new

W = 0.5, p-value = 0.004698

alternative hypothesis: true location shift is not equal to 0

Η p-τιμή του ελέγχου είναι μικρότερη από το επίπεδο σημαντικότητας α , το οποίο έχουμε ορίσει ίσο με 0.05 και επομένως έχουμε σοβαρές ενδείξεις κατά της μηδενικής υπόθεσης.

3. Ο έλεγχος των προσημασμένων τάξεων μεγέθους του Wilcoxon (Wilcoxon Signed Rank Test)

3.1 Η κατανομή του στατιστικού υπό την μηδενική υπόθεση

Σε πολλές περιπτώσεις στατιστικών μελετών τα δείγματα μας είναι εξαρτημένα. Αυτό μπορεί να συμβεί όταν για παράδειγμα έχουμε μετρήσεις της ίδιας ποσοτικής μεταβλητής στην ίδια μονάδα μελέτης αλλά σε διαφορετικές χρονικές στιγμές (η κατάσταση της υγείας ενός ασθενή πριν και μετά από την εφαρμογή μιας θεραπείας), μετρήσεις για τα ίδια άτομα σε δύο διαφορετικές μεταβλητές που όμως συνδέονται μεταξύ τους (η επίδοση του ίδιου δείγματος ανθρώπων σε δύο διαφορετικά αθλήματα) κλπ. Στα παραπάνω παραδείγματα δεν μπορούμε σαφώς να χρησιμοποιήσουμε την μέθοδο που αναφέραμε στο προηγούμενο Κεφάλαιο, καθώς έτσι θα αγνοούσαμε πλήρως την συσχέτιση που υπάρχει ανάμεσα στα δύο δείγματα. Θα θεωρήσουμε ότι έχουμε μετρήσεις της ίδιας ποσοτικής μεταβλητής σε διαφορετικές μονάδες μελέτης που όμως σχετίζονται (για παράδειγμα δίδυμα αδέρφια ή συγγενείς). Έστω ότι καλούμε X την μεταβλητή για την πρώτη μονάδα (το μεγαλύτερο, κατά μερικά λεπτά, δίδυμο) και υποθέτουμε ότι αυτή ακολουθεί μια γνωστή κατανομή και ότι καλούμε Y την μεταβλητή για την δεύτερη μονάδα (το μικρότερο, κατά μερικά λεπτά, δίδυμο) που ακολουθεί επίσης μια γνωστή κατανομή. Υποθέτουμε ότι X_1, \dots, X_n είναι ένα τυχαίο δείγμα από παρατηρήσεις για την πρώτη μονάδα και ότι Y_1, \dots, Y_n είναι ένα τυχαίο δείγμα από παρατηρήσεις για την δεύτερη μονάδα. Προκειμένου να ελέγξουμε αν υπάρχει κάποια διαφορά του υπό μελέτη χαρακτηριστικού στις δύο μονάδες θα ελέγξουμε την μηδενική υπόθεση H_0 : Δεν υπάρχει διαφορά, σε κάποιο επίπεδο σημαντικότητας α . Στην πραγματικότητα θα ελέγξουμε αν οι κατανομές των X, Y είναι ίσες, δηλαδή αν τα δύο δείγματα προέρχονται από πληθυσμούς με την ίδια συνάρτηση κατανομής και υπό την προϋπόθεση ότι οι δύο υποπληθυσμοί έχουν το ίδιο σχήμα και περίπου ίδια διασπορά, θα πάρουμε μικρή P-τιμή όταν τα μέτρα θέσης των κατανομών (για παράδειγμα η διάμεσος) θα διαφέρουν στους δύο υποπληθυσμούς. Έστω πως έχουμε ένα τυχαίο δείγμα $(X_1, Y_1), \dots, (X_n, Y_n)$ το οποίο αποτελείται από ζεύγη των δύο συσχετισμένων τυχαίων μεταβλητών και θεωρούμε τις διαφορές $(D_1 = X_1 - Y_1), \dots, (D_n = X_n - Y_n)$ για το κάθε ζεύγος. Τότε θα έχουμε συνολικά N ζεύγη παρατηρήσεων όπου σε κάθε ζεύγος επιλέγουμε με τυχαίο τρόπο την μια παρατήρηση να λάβει την νέα θεραπεία ενώ η άλλη θα λάβει την υπάρχουσα θεραπεία. Συμβολίζουμε με $N_+ = n$ τις θετικές διαφορές των τιμών των παρατηρήσεων που έλαβαν την νέα θεραπεία με αυτών που έλαβαν την παλιά και με $N_- = m = N - n$ τις αρνητικές διαφορές. Αυτό που πρέπει να κάνουμε είναι να διατάξουμε κατά αύξουσα σειρά τις απόλυτες τιμές των διαφορών και να αποδώσουμε σε κάθε μια τον βαθμό της. Στην συνέχεια ανάλογα αν η διαφορά είναι θετική ή αρνητική δίνουμε στον βαθμό τα πρόσημα με $+$ και με $-$ αντίστοιχα. Συμβολίζουμε με $R_1 < \dots < R_m$ τους βαθμούς που έχουν αρνητικό πρόσημο και με $S_1 < \dots < S_n$ τους βαθμούς που έχουν θετικό πρόσημο και παίρνουν τιμές στο σύνολο των ακεραίων $1, \dots, N$. Συνολικά υπάρχουν 2^N δυνατοί συνδυασμοί των πρόσημων $\pm 1, \pm 2, \dots, \pm N$ και κάθε συνδυασμός μπορεί να χαρακτηριστεί μόνο από το υποσύνολο (S_1, \dots, S_n) των βαθμών με θετικό πρόσημο. Μια ειδική περίπτωση είναι όλες οι διαφορές που θα προκύψουν να είναι αρνητικές και τότε το πλήθος των θετικών διαφορών είναι $n = 0$ και το σύνολο (S_1, \dots, S_n) των βαθμών με θετικό πρόσημο είναι κενό. Για κάθε τιμή του αριθμού n υπάρχουν $\binom{N}{n}$ δυνατές επιλογές του συνόλου (S_1, \dots, S_n) των βαθμών με θετικό πρόσημο.

Αφού το n μπορεί να πάρει τιμές στο σύνολο $1, \dots, N$ το συνολικό πλήθος των δυνατών επιλογών για τα θετικά πρόσημα S είναι

$$\binom{N}{0} + \binom{N}{1} + \dots + \binom{N}{N} = (1 + 1)^N = 2^N$$

που ισούται με το συνολικό πλήθος των συνδυασμών των πρόσημων και για τα παραπάνω έχουμε υποθέσει πως δεν προκύπτουν διαφορές ίσες με μηδέν.

Υπό την μηδενική υπόθεση ότι η θεραπεία δεν έχει κάποιο αποτέλεσμα και αφού το ποια παρατήρηση μέσα στο ζεύγος θα λάβει την νέα θεραπεία και ποια όχι αποφασίζεται με τυχαίο τρόπο, προκύπτει ότι κάθε διαφορά έχει πιθανότητα ίση με $1/2$ να είναι θετική ή αρνητική και επομένως τα N το πλήθος πρόσημα που μπορούν να προκύψουν είναι ανεξάρτητα και κάθε ένας από τους συνολικά 2^N δυνατούς συνδυασμούς μπορεί να προκύψει με πιθανότητα $1/2^N$. Αφού κάθε συνδυασμός αντιστοιχεί σε μια τιμή n πλήθους θετικών διαφορών και ένα σύνολο (S_1, \dots, S_n) , η από κοινού κατανομή των μεταβλητών είναι

$$P_{H_0}(N_+ = n; S_1 = s_1, \dots, S_n = s_n) = 1/2^N,$$

οπου το δεξί μέλος της σχέσης αντιστοιχεί στην πιθανότητα το πλήθος των θετικών διαφορών να είναι ίσο με n και ότι οι βαθμοί τους είναι οι τιμές (s_1, \dots, s_n) από το σύνολο όλων των βαθμών των απόλυτων διαφορών.

Έστω ότι η εναλλακτική υπόθεση ορίζεται ως ότι η θεραπεία έχει ως αποτέλεσμα να αυξάνει την απόκριση. Υπό την συνθήκη αυτή, η μηδενική υπόθεση θα απορρίπτεται αν για πολλά ζεύγη, η παρατήρηση που αντιστοιχεί στο άτομο που έλαβε την νέα θεραπεία έχει μεγαλύτερη τιμή από αυτή που αντιστοιχεί στο άτομο που έλαβε την παλιά θεραπεία και αν οι διαφορές αυτές είναι μεγαλύτερες από τις διαφορές που προκύπτουν όταν η τιμή που αντιστοιχεί στο άτομο που έλαβε την παλιά θεραπεία είναι μεγαλύτερη από την τιμή που αντιστοιχεί στο άτομο που έλαβε την νέα θεραπεία. Δηλαδή, η μηδενική υπόθεση θα απορρίπτεται όταν το πλήθος n είναι μεγάλο και όταν οι βαθμοί (S_1, \dots, S_n) με θετικό πρόσημο τείνουν να είναι μεγαλύτεροι από τους βαθμούς (R_1, \dots, R_m) με αρνητικό πρόσημο.

Ένα απλό στατιστικό ελέγχου που να συνδυάζει αυτές τις δύο συνθήκες είναι το

$$V_s = S_1 + \dots + S_n$$

και η μηδενική υπόθεση, ότι η νέα θεραπεία δεν είναι αποτελεσματική, απορρίπτεται για μεγάλες τιμές του V_s , παραδείγματος χάριν

$$V_s \geq c.$$

Έστω ότι $N = 2$. Τότε υπάρχουν συνολικά $2^N = 4$ πιθανοί συνδυασμοί βαθμών και πρόσημων αυτών, που παρουσιάζονται στον Πίνακα 3.1.1.

| Βαθμοί πρόσημα | με | +1, +2 | +1, -2 | -1, +2 | -1, -2 |
|-------------------|----|--------|--------|--------|--------|
| v | | 3 | 1 | 2 | 0 |

Πίνακας 3.1.1 Όλοι οι δυνατοί συνδυασμοί βαθμών για $N = 2$

Αφού κάθε συνδυασμός έχει πιθανότητα $1/4$ να προκύψει, η κατανομή του στατιστικού V_s υπό την μηδενική υπόθεση δίνεται στον Πίνακα 3.1.2.

| | | | | |
|--------------------|-----|-----|-----|-----|
| v | 0 | 1 | 2 | 3 |
| $P_{H_0}(V_s = v)$ | 1/4 | 1/4 | 1/4 | 1/4 |

Πίνακας 3.1.2 Η κατανομή του στατιστικού V_s .

Ο υπολογισμός της κατανομής του στατιστικού V_s υπό την μηδενική υπόθεση μπορεί να γενικευθεί. Είναι

$$P_{H_0}(V_s = v) = \frac{\#(v; N)}{2^N},$$

οπου με $\#(v; N)$ συμβολίζουμε το πλήθος των συνδυασμών των πρόσημων των $1, \dots, N$ βαθμών για τους οποίους το άθροισμα των βαθμών με το θετικό πρόσημο είναι ίσο με v . Για περιπτώσεις όπου το πλήθος των βαθμών με αρνητικό πρόσημο είναι μικρότερο από το πλήθος των βαθμών με θετικό πρόσημο, μπορούμε να χρησιμοποιήσουμε ως στατιστικό ελέγχου το άθροισμα των βαθμών με αρνητικό πρόσημο, δηλαδή το

$$V_r = R_1 + \dots + R_m.$$

Το άθροισμα των στατιστικών V_r και V_s είναι ίσο με το άθροισμα των ακεραίων από το 1 ως το N , δηλαδή ισχύει

$$V_r + V_s = \frac{1}{2}N(N + 1).$$

Επομένως μεγάλες τιμές του στατιστικού V_r αντιστοιχούν σε μικρές τιμές του στατιστικού V_s και αντίστροφα. Μάλιστα τα δύο στατιστικά έχουν ίδια κατανομή υπό την μηδενική υπόθεση, η οποία είναι συμμετρική γύρω από το σημείο $N(N + 1)/4$ και επομένως ισχύει:

$$P_{H_0}(V_s \leq v) = P_{H_0}(V_r \leq v)$$

που στην R η πιθανότητα αυτή υπολογίζεται με την εντολή `psignrank(v,N)` ενώ η κρίσιμη τιμή για επίπεδο σημαντικότητας α δίνεται από την εντολή `qsignrank(alpha,N)`.

3.2 Ασυμπτωτική κατανομή του στατιστικού ελέγχου

Όπως και στο προηγούμενο Κεφάλαιο έτσι και σε αυτό, για μεγάλες τιμές δειγμάτων μπορούμε να προσεγγίσουμε την κατανομή των στατιστικών V_s και V_r από την κανονική κατανομή κάνοντας χρήση του Κ.Ο.Θ.

Η μέση τιμή και η διασπορά των στατιστικών δίνονται από τους τύπους

$$E_{H_0}(V_r) = E_{H_0}(V_s) = \frac{N(N + 1)}{4}$$

και

$$Var_{H_0}(V_r) = Var_{H_0}(V_s) = \frac{N(N + 1)(2N + 1)}{24}.$$

Παράδειγμα 3.2.1

Έστω ότι θέλουμε να υπολογίσουμε την πιθανότητα $P_{H_0}(V_r \leq 10)$ όταν το συνολικό δείγμα έχει μέγεθος $N = 12$. Από τους παραπάνω τύπους βρίσκουμε

$$E_{H_0}(V_r) = 39$$

και

$$Var_{H_0}(V_r) = 162.5.$$

Επομένως θα είναι

$$P(V_r \leq 10) \approx \Phi\left(\frac{10.5 - 29}{12.75}\right) = \Phi(-1.45098)$$

οπου την πιθανότητα αυτή μπορούμε να την υπολογίσουμε στην R με την εντολή

```
> pnorm(-1.45098)
[1] 0.07339271
```

ή απευθείας με την εντολή

```
> pnorm(10.5,mean=29,sd=12.75)
[1] 0.07339266.
```

3.3 Μια διαφορετική παρουσίαση του στατιστικού ελέγχου

Υπάρχει ένας εναλλακτικός τρόπος παρουσίασης του στατιστικού ελέγχου V_s . Ας συμβολίσουμε τις N το πλήθος διαφορές των παρατηρήσεων με Z_1, \dots, Z_N και με $\frac{1}{2}(Z_i + Z_j)$ το άθροισμα όλων των μέσων διαφορών για $i \leq j$.

Συνολικά το πλήθος του αθροίσματος είναι ίσο με $\binom{N}{n} + N$.

Πρόταση 3.3.1

Μπορούμε να ορίσουμε το στατιστικό V_s ως

$$V_s = \text{το πλήθος των θετικών } \frac{1}{2}(Z_i + Z_j) \text{ με } i \leq j.$$

Απόδειξη 3.3.1

Το $\frac{1}{2}(Z_i + Z_j)$ είναι θετικό όταν για $i < j$ η μεγαλύτερη κατά απόλυτη τιμή διαφορά είναι θετική, ενώ για $i = j$ όταν η μοναδική αυτή διαφορά είναι θετική. Έστω, χωρίς βλάβης της γενικότητας, ότι οι διαφορές έχουν καταταχθεί σε αύξουσα σειρά σύμφωνα με τις απόλυτες τιμές τους έτσι ώστε ο προσημασμένος βαθμός της διαφοράς Z_j να είναι $+j$ αν η διαφορά είναι θετική ή $-j$ αν η διαφορά Z_j είναι αρνητική. Τότε για μια συγκεκριμένη τιμή του j η συνεισφορά των (Z_i, Z_j) ζευγών με $i \leq j$ στο δεξί μέλος της σχέσης που θέλουμε να αποδείξουμε θα είναι μηδενική όταν $Z_j < 0$ ενώ θα είναι ίση με j όταν $Z_j > 0$. Επομένως πράγματι το δεξί μέλος της σχέσης θα είναι το άθροισμα των θετικών βαθμών.

3.4 Όταν έχουμε παρατηρήσεις με ίσες τιμές

Ο τρόπος με τον οποίο διαχειριζόμαστε διαφορές με ίσες τιμές είναι εύκολο να εξηγηθεί μέσα από το Παράδειγμα 3.4.1.

Παράδειγμα 3.4.1

Έστω πως έχουμε τα παρακάτω ζεύγη παρατηρήσεων σε μια έρευνα

$$(1, -1), (0,1), (1,1), (2,1), (-2,0).$$

Η πρώτη τιμή σε κάθε ζεύγος αφορά στην παρατήρηση που έλαβε την παλιά θεραπεία ενώ η δεύτερη τιμή αφορά στην παρατήρηση που έλαβε την νέα θεραπεία. Οι διαφορές μεταξύ των παρατηρήσεων για το κάθε ζεύγος είναι $-2, 1, 0, -1$ και 2 και οι απόλυτες τιμές τους κατά αύξουσα σειρά είναι οι $0, 1, 1, 2, 2$. Οι βαθμοί δίνονται με τον ίδιο τρόπο που δόθηκαν και στο προηγούμενο κεφάλαιο όταν είχαμε ισοβαθμίες, επομένως θα είναι $1, 2.5, 2.5, 4.5$ και 4.5 . Για να βρούμε τον προσημασμένο βαθμό κάθε διαφοράς αρκεί να πολλαπλασιάσουμε τον κάθε βαθμό με $+1, -1$ ή 0 αν η διαφορά είναι θετική, αρνητική ή ίση με το μηδέν αντίστοιχα.

Οι προσημασμένοι βαθμοί παρουσιάζονται στον Πίνακα 3.4.1

| | | | | | |
|----------------------|------|------|---|-----|-----|
| Z_i | -2 | -1 | 0 | 1 | 2 |
| $ Z_i $ | 2 | 1 | 0 | 1 | 2 |
| Προσημασμένοι Βαθμοί | -4.5 | -2.5 | 0 | 2.5 | 4.5 |

Πίνακας 3.4.1 Προσημασμένοι βαθμοί των διαφορών των ζευγών για το Παράδειγμα 3.4.1.

Το άθροισμα των βαθμών με θετικό πρόσημο είναι ίσο με

$$V_s^* = 2.5 + 4.5 = 7.$$

Για να υπολογίσουμε την κατανομή υπό την μηδενική υπόθεση για το στατιστικό V_s^* μπορούμε να παραλείψουμε τους μηδενικούς βαθμούς, αν υπάρχουν και να ασχοληθούμε με τους υπόλοιπους βαθμούς. Η διαδικασία αυτή γίνεται αυστηρά μόνο αφού έχουν αποδοθεί οι βαθμοί των παρατηρήσεων. Υπάρχουν συνολικά 2^4 δυνατοί συνδυασμοί βαθμών με θετικά πρόσημα. Όλοι οι συνδυασμοί δίνονται στον Πίνακα 3.4.2 μαζί με τις πιθανότητες τους.

| | | | | | |
|-------------------------|---------------|---------------|--------------------|---------------|----------|
| (s_1^*, \dots, s_n^*) | 2.5 | 4.5 | 2.5, 2.5 | 2.5, 4.5 | 4.5, 4.5 |
| V_s^* | 2.5 | 4.5 | 5 | 7 | 9 |
| Πιθανότητα | 2/16 | 2/16 | 1/16 | 4/16 | 1/16 |
| (s_1^*, \dots, s_n^*) | 2.5, 2.5, 4.5 | 2.5, 4.5, 4.5 | 2.5, 2.5, 4.5, 4.5 | Κανένα θετικό | |
| V_s^* | 9.5 | 11.5 | 14 | 0 | |
| Πιθανότητα | 2/16 | 2/16 | 1/16 | 1/16 | |

Πίνακας 3.4.2: Η κατανομή του V_s^* υπό την μηδενική υπόθεση.

Στο κελί του πίνακα που έχουμε τις τιμές $2.5, 4.5, 4.5$ με τιμή στατιστικού $V_s^* = 11.5$ η πιθανότητα έχει υπολογιστεί να είναι ίση με $2/16$. Αυτό συμβαίνει διότι υπάρχουν 2 διαφορετικοί συνδυασμοί βαθμών που μπορούν να μας δώσουν την τιμή 11.5 για το στατιστικό ελέγχου. Με την ίδια λογική έχουν υπολογιστεί και οι υπόλοιπες πιθανότητες στα κελιά του Πίνακα 3.4.2. Αφού δεν υπάρχουν ίδιες τιμές για το στατιστικό V_s^* στον πίνακα, τα κελιά με

τις δυνατές τιμές του στατιστικού μαζί με τις πιθανότητες δίνουν την κατανομή του στατιστικού υπό την μηδενική υπόθεση.

Αν θέλουμε να υπολογίσουμε την πιθανότητα το στατιστικό V_s^* να πάρει μια τιμή τόσο ακραία ή και περισσότερο ακραία από την τιμή 7 αρκεί να υπολογίσουμε με την βοήθεια του Πίνακα 3.4.2 την πιθανότητα

$$P_{H_0}(V_s^* \geq 7) = P_{H_0}(V_s^* = 7) + P_{H_0}(V_s^* = 9) + P_{H_0}(V_s^* = 9.5) + P_{H_0}(V_s^* = 11.5) + P_{H_0}(V_s^* = 14) = 10/16.$$

Καθώς το μέγεθος του N μεγαλώνει η παραπάνω διαδικασία γίνεται αρκετά πολύπλοκη. Για το λόγο αυτό μπορούμε να υπολογίσουμε την κατανομή του στατιστικού V_s^* προσεγγιστικά από την κανονική κατανομή. Η μέση τιμή και η διασπορά του στατιστικού υπολογίζονται από τους ακόλουθους τύπους:

$$E_{H_0}(V_s^*) = \frac{N(N+1) - d_0(d_0+1)}{4}$$

οπου d_0 είναι το πλήθος των μηδενικών διαφορών και

$$Var_{H_0}(V_s^*) = \frac{1}{24} [N(N+1)(2N+1) - d_0(d_0+1)(2d_0+1)] - \frac{1}{48} \sum_{i=1}^e d_i(d_i-1)(d_i+1)$$

οπου d_i, \dots, d_e είναι το πλήθος των δεσμών για όλες τις τιμές που είναι διάφορες του μηδενός.

Ισχύει

$$P_{H_0}(V_s^* \geq c) \approx 1 - \Phi \left(\frac{c - E_{H_0}(V_s^*)}{\sqrt{Var_{H_0}(V_s^*)}} \right)$$

καθώς το πλήθος των μη μηδενικών διαφορών τείνει στο άπειρο.

Η παραπάνω πιθανότητα στην R υπολογίζεται με την εντολή `pnorm(c,mean=,sd=,lower.tail=F)` για όποια τιμή της σταθεράς c επιθυμούμε, αρκεί να δώσουμε την μέση τιμή και την τυπική απόκλιση του στατιστικού.

3.5 Η εντολή του ελέγχου στο στατιστικό πακέτο της R

Η εντολή στην R για να πραγματοποιήσουμε έναν έλεγχο προσημασμένων τάξεων μεγέθους του Wilcoxon για εξαρτημένα δεδομένα είναι ο `wilcox.test(x, y, paired=TRUE)`.

Ορίσματα εντολής

Η εντολή παίρνει ως ορίσματα δύο διανύσματα **x**, **y** που περιέχουν τις παρατηρήσεις από το 1^ο και το 2^ο δείγμα αντίστοιχα. Το **paired** είναι ένα λογικό όρισμα οπου στον έλεγχο αυτό παίρνει την τιμή TRUE πάντα για να δείξουμε ότι τα δεδομένα είναι σε ζεύγη (εξαρτημένα) και πρέπει να πραγματοποιηθεί ο κατάλληλος έλεγχος. Σε αντίθετη περίπτωση ο έλεγχος που θα πραγματοποιηθεί θα είναι αυτός που είδαμε στο προηγούμενο κεφάλαιο και επομένως τα αποτελέσματα που θα πάρουμε δε θα είναι σωστά. Επιπλέον, το όρισμα **alternative** είναι τύπου string και καθορίζει ανάλογα την τιμή του αν ο έλεγχος θα είναι αμφίπλευρος ή μονόπλευρος. Έχει επομένως ως δυνατές τιμές τις **“two.sided”**, η οποία είναι και η προκαθορισμένη από την

R τιμή, “greater” ή “less”. Μπορούμε να το γράψουμε είτε ολογράφως είτε δίνοντας μόνο το αρχικό γράμμα. Π.χ.

```
wilcox.test(x, y, paired = TRUE, alternative = "l")
```

Τα ορίσματα exact και correct παίρνουν λογικές τιμές TRUE/FALSE και το πρώτο είναι για τον υπολογισμό της ακριβούς τιμής της P-τιμής ενώ το δεύτερο για το αν πρέπει να εφαρμοστεί προσέγγιση από την κανονική τιμή ή όχι.

Αφού περάσουμε τα δεδομένα και τρέξουμε την εντολή του ελέγχου, η R θα μας δώσει ως απάντηση τα αποτελέσματα του ελέγχου που θα είναι της μορφής:

Wilcoxon signed rank test

data: x and y

V = ..., p-value = ...

alternative hypothesis: true location shift is ... 0

όπου,

x, y με το πορτοκαλί θα είναι τα δύο διανύσματα με τις παρατηρήσεις.

V θα είναι η τιμή του στατιστικού ελέγχου.

p-value θα είναι η p-τιμή η οποία αν είναι < του επίπεδου σημαντικότητας α θα οδηγεί στην απόρριψη της μηδενικής υπόθεσης.

Alternative hypothesis δείχνει ποια είναι η εναλλακτική υπόθεση H_1 , δηλαδή τη μορφή του ελέγχου, αμφίπλευρος ή μονόπλευρος. Ο εν λόγω έλεγχος είναι ο μη παραμετρικός έλεγχος που είναι ανάλογος του παραμετρικού **paired t-test**.

Εφαρμογή 3.5.1

Ένας δήμαρχος θέλει να ελέγξει αν τα επίπεδα της ρύπανσης μειώνονται με το κλείσιμο των δρόμων. Για το λόγο αυτό έγιναν συνολικά 15 μετρήσεις ανά μια ώρα μια μέρα που οι δρόμοι ήταν ανοιχτοί για τα αυτοκίνητα και μια που ήταν κλειστοί. Εύκολα καταλαβαίνουμε ότι πρόκειται για εξαρτημένα δείγματα καθώς οι μετρήσεις πραγματοποιούνται στην ίδια πόλη, με τις ίδιες καιρικές συνθήκες σε γενικό πλαίσιο κλπ.

Στον Πίνακα 3.5.1 δίνονται οι τιμές των μετρήσεων.

| Μετρήσεις | Με κίνηση | Χωρίς κίνηση |
|-----------|-----------|--------------|
| 1 | 214 | 159 |
| 2 | 159 | 135 |
| 3 | 169 | 141 |
| 4 | 202 | 101 |
| 5 | 103 | 102 |
| 6 | 119 | 168 |
| 7 | 200 | 62 |
| 8 | 109 | 167 |
| 9 | 132 | 174 |
| 10 | 142 | 159 |
| 11 | 194 | 66 |
| 12 | 104 | 118 |
| 13 | 219 | 181 |
| 14 | 119 | 171 |
| 15 | 234 | 112 |

Πίνακας 3.5.1: Επίπεδα ρύπανσης.

Πρώτα πρέπει να υπολογίσουμε τις διαφορές που προέκυψαν στις μετρήσεις.

| Με κίνηση | Χωρίς κίνηση | Διαφορές (Με-Χωρίς) |
|-----------|--------------|------------------------|
| 214 | 159 | 55 |
| 159 | 135 | 24 |
| 169 | 141 | 28 |
| 202 | 101 | 101 |
| 103 | 102 | 1 |
| 119 | 168 | -49 |
| 200 | 62 | 138 |
| 109 | 167 | -58 |
| 132 | 174 | -42 |
| 142 | 159 | -17 |
| 194 | 66 | 128 |
| 104 | 118 | -14 |
| 219 | 181 | 38 |
| 119 | 171 | -52 |
| 234 | 112 | 122 |

Πίνακας 3.5.2: Διαφορές επιπέδων ρύπανσης με και χωρίς κλείσιμο των δρόμων.

Στο επόμενο βήμα θα πρέπει να διατάξουμε τις διαφορές και να τους δώσουμε τον βαθμό που αντιστοιχεί σε κάθε μια. Αρχικά διατάσσουμε τις απόλυτες διαφορές των παρατηρήσεων και δίνουμε το βαθμό 1 στην μικρότερη και συνεχίζουμε ως την μεγαλύτερη παρατήρηση που θα λάβει τον βαθμό n. Προσέχουμε σε περίπτωση που προκύψουν ίσες τιμές να δώσουμε την μέση βαθμολογία. Οι βαθμοί των διαφορών δίνονται στον Πίνακα 3.5.3.

| Διαφορές (Με-Χωρίς) | Διατεταγμένες απόλυτες διαφορές | Βαθμοί |
|------------------------|------------------------------------|--------|
| 55 | 1 | 1 |
| 24 | -14 | 2 |
| 28 | -17 | 3 |
| 101 | 24 | 4 |
| 1 | 28 | 5 |
| -49 | 38 | 6 |
| 138 | -42 | 7 |
| -58 | -49 | 8 |
| -42 | -52 | 9 |
| -17 | 55 | 10 |
| 128 | -58 | 11 |
| -14 | 101 | 12 |
| 38 | 122 | 13 |
| -52 | 128 | 14 |
| 122 | 138 | 15 |

Πίνακας 3.5.3: Βαθμοί των παρατηρήσεων των διαφορών.

Το τελικό βήμα είναι να αντιστοιχήσουμε στους βαθμούς το πρόσημο (+ ή -) όπως φαίνεται από κάτω.

| Διαφορές (Με-Χωρίς) | Διατεταγμένες απόλυτες διαφορές | Βαθμοί | Προσημασμένοι Βαθμοί |
|------------------------|------------------------------------|--------|-------------------------|
| 55 | 1 | 1 | 1 |
| 24 | -14 | 2 | -2 |
| 28 | -17 | 3 | -3 |
| 101 | 24 | 4 | 4 |
| 1 | 28 | 5 | 5 |
| -49 | 38 | 6 | 6 |
| 138 | -42 | 7 | -7 |
| -58 | -49 | 8 | -8 |
| -42 | -52 | 9 | -9 |
| -17 | 55 | 10 | 10 |
| 128 | -58 | 11 | -11 |
| -14 | 101 | 12 | 12 |
| 38 | 122 | 13 | 13 |
| -52 | 128 | 14 | 14 |
| 122 | 138 | 15 | 15 |

Πίνακας 3.5.4: Προσημασμένοι βαθμοί.

Αρχικά πρέπει να περάσουμε τις παρατηρήσεις μας στο πρόγραμμα. Αυτό θα γίνει με την δημιουργία δύο διανυσμάτων.

```
> a <- c(214, 159, 169, 202, 103, 119, 200, 109, 132, 142, 194, 104, 219, 119, 234)
```

```
> a
```

```
[1] 214 159 169 202 103 119 200 109 132 142
```

```
[11] 194 104 219 119 234
```

```
> b <- c(159, 135, 141, 101, 102, 168, 62, 167, 174, 159, 66, 118, 181, 171, 112)
```

```
> b
```

```
[1] 159 135 141 101 102 168 62 167 174 159
```

```
[11] 66 118 181 171 112
```

Τρέχουμε τον έλεγχο για παίρνουμε τα αποτελέσματα

```
> wilcox.test(a,b,paired = TRUE)
```

Wilcoxon signed rank test

data: a and b

V = 80, p-value = 0.2769

alternative hypothesis: true location shift is not equal to 0

Η P-τιμή του ελέγχου είναι p-value = 0.2769, δηλαδή μεγαλύτερη από το επίπεδο σημαντικότητας 0.05 και επομένως αποδεχόμαστε τη μηδενική υπόθεση αφού δεν έχουμε ισχυρές ενδείξεις εναντίον της.

4. Ο έλεγχος των Kruskal-Wallis

4.1 Η κατανομή υπό την μηδενική υπόθεση για τους βαθμούς των παρατηρήσεων από s το πλήθος θεραπειές.

Σε πολλές έρευνες δε θέλουμε να συγκρίνουμε μόνο δύο, αλλά τρεις και περισσότερες θεραπείες ή συνθήκες. Για λόγο αυτό, στο παρόν κεφάλαιο θα υποθέσουμε ότι έχουμε μια συνεχή μεταβλητή απόκρισης Y και έναν παράγοντα X , τον οποίο μπορούμε να συμβολίσουμε και ως A , ο οποίος έχει k επίπεδα τιμών (στάθμες/ομάδες) που μπορεί να πάρει. Έστω ότι η μεταβλητή απόκρισης Y εκφράζει τα επίπεδα χοληστερίνης των ασθενών που λαμβάνουν μέρος στην έρευνα, ενώ ο παράγοντας A είναι η μέθοδος-αγωγή που θα χρησιμοποιηθεί προκειμένου να ελαττωθούν τα επίπεδα χοληστερίνης και παίρνει $s = 4$ διαφορετικές τιμές. Στόχος της έρευνας είναι να αποφασίσουμε αν διαφοροποιείται η εξαρτημένη μεταβλητή ανάλογα τη θεραπεία(στάθμη/ομάδα) που εφαρμόζουμε κάθε φορά. Θα συμβολίζουμε με y_{ij} την τιμή της παρατήρησης i στην ομάδα j . Έστω πως N το πλήθος άτομα είναι διαθέσιμα για να συμμετάσχουν στην έρευνα που αφορά στην σύγκριση s το πλήθος θεραπειών και για τον λόγο αυτό επιλέγονται n_i το πλήθος άτομα τυχαία για να λάβουν την i – οστή θεραπεία έτσι ώστε να ισχύει συνολικά

$$N = n_1 + \dots + n_s.$$

Το πλήθος όλων των δυνατών επιλογών των n_i το πλήθος ατόμων για την i – οστή θεραπεία δίνεται από τον αριθμό $\binom{N}{n_1, \dots, n_s}$ που είναι στην ουσία μια γενίκευση του διωνυμικού συντελεστή και ονομάζεται πολυωνυμικός συντελεστής και υπολογίζεται από τον τύπο

$$\binom{N}{n_1, \dots, n_s} = \binom{N}{n_1} \binom{N - n_1}{n_2} \dots \binom{N - n_1 - \dots - n_{s-2}}{n_{s-1}}.$$

Στην R ο πολυωνυμικός συντελεστής υπολογίζεται εύκολα με την εντολή `multinom(x, counts = T)` αρκεί να δώσουμε ένα διάνυσμα x που θα περιέχει τα μεγέθη των δειγμάτων n_1, \dots, n_s .

Από την αρχική υπόθεση, τα N το πλήθος άτομα θα χωριστούν σε κάθε μια από τις s το πλήθος δυνατές θεραπείες κατά τυχαίο τρόπο και επομένως όλοι οι $\binom{N}{n_1, \dots, n_s}$ πιθανοί συνδυασμοί έχουν ίδια πιθανότητα να προκύψουν, που είναι ίση με $1 / \binom{N}{n_1, \dots, n_s}$.

Στο τέλος της θεραπείας τα N το πλήθος άτομα βαθμολογούνται ανάλογα με την τιμή της απόκρισης του χαρακτηριστικού που μελετάμε. Συμβολίζουμε με

$$R_{11}, \dots, R_{1n_1}; R_{21}, \dots, R_{2n_2}; \dots; R_{s1}, \dots, R_{sn_s}$$

τους κατά αύξουσα σειρά διατεταγμένους βαθμούς για κάθε ένα από τα n_1, \dots, n_s το πλήθος δείγματα.

Υπό την μηδενική υπόθεση ότι οι θεραπείες δεν διαφέρουν μεταξύ τους ο βαθμός που θα λάβει το κάθε ένα από τα N το πλήθος άτομα θα είναι ανεξάρτητος από το αν το άτομο έλαβε την θεραπεία 1 του δείγματος πλήθους n_1 ή την θεραπεία j του δείγματος πλήθους n_j και μπορούμε να θεωρήσουμε πως όταν διαλέξαμε τα n_1 άτομα για την θεραπεία 1 διαλέξαμε ουσιαστικά και τους βαθμούς τους.

Επομένως όλοι οι $\binom{N}{n_1, \dots, n_s}$ πιθανοί τρόποι να χωριστούν οι βαθμοί $1, \dots, N$ στις s θεραπείες με μεγέθη δειγμάτων n_1, \dots, n_k έχουν πιθανότητα

$$P_{H_0}(R_{11} = r_{11}, \dots, R_{1n_1} = r_{1n_1}; \dots; R_{s1} = r_{s1}, \dots, R_{sn_s} = r_{sn_s}) = 1 / \binom{N}{n_1, \dots, n_s}.$$

Η παραπάνω πιθανότητα είναι η κατανομή των βαθμών των δειγμάτων υπό την μηδενική υπόθεση ότι δεν διαφέρουν οι θεραπείες μεταξύ τους.

Παράδειγμα 4.1.1

Για να ελέγξουμε αν υπάρχει διαφορά στην δράση τριών φαρμάκων επιλέξαμε ένα δείγμα 5 ανθρώπων από τους οποίους οι 2 επιλέχθηκαν κατά τυχαίο τρόπο να λάβουν το φάρμακο 1, άλλοι 2 με τυχαίο τρόπο επιλέχθηκαν για να λάβουν το φάρμακο 2 και ένας επιλέχθηκε κατά τυχαίο τρόπο για να λάβει το φάρμακο 3. Μετά από εξετάσεις και διαγνώσεις που έγιναν σε όλους τους ασθενείς, έλαβαν βαθμολογία ανάλογα με την επίδραση που είχε το φάρμακο σε αυτούς. Ο βαθμός του κάθε ασθενή ανάλογα με την φαρμακευτική αγωγή που έλαβε παρουσιάζεται παρακάτω στις παρενθέσεις:

Φάρμακο 1 (3,5)

Φάρμακο 2 (2,4)

Φάρμακο 3 (1).

Δηλαδή οι ασθενείς που έλαβαν το φάρμακο 2 πήραν βαθμούς 2 και 4. Όσο περισσότερο θετική επίδραση είχε το φάρμακο σε έναν ασθενή, τόσο μεγαλύτερη θα είναι η βαθμολογία του συγκεκριμένου ασθενή.

Ορίζουμε ως μηδενική υπόθεση ότι δεν υπάρχει διαφορά ανάμεσα στις επιδράσεις των τριών φαρμάκων. Υπό την μηδενική υπόθεση, ο βαθμός του κάθε ασθενή θα προσδιορίζει την κατάσταση της υγείας του και δεν εξαρτάται από το ποιο από τα τρία φάρμακα έλαβε ο ασθενής επομένως μπορούμε να σκεφτούμε ότι όταν διαλέξαμε τους ασθενείς για κάθε φάρμακο τυχαία, επιλέξαμε τυχαία και τους βαθμούς αυτών. Επομένως όταν χωρίζουμε τους 5 ασθενείς σε τρεις κατηγορίες χωρίζουμε στην ουσία και τους 5 βαθμούς στις κατηγορίες αυτές και αφού αυτό γίνεται κατά τρόπο τυχαίο, όλοι οι δυνατοί τρόποι να χωρίσουμε τους βαθμούς στις τρεις αυτές κατηγορίες είναι εξίσου πιθανοί.

Για το πρώτο φάρμακο μπορούμε από το δείγμα 5 ατόμων να επιλέξουμε 2 κατά τυχαίο τρόπο. Υπάρχουν επομένως συνολικά $\binom{5}{2}$ δυνατοί συνδυασμοί των 2 αυτών βαθμών για το πρώτο φάρμακο. Με την R υπολογίζουμε το πλήθος των δυνατών συνδυασμών που είναι ίσο με

```
> choose(5,2)
```

```
[1] 10.
```

Στον Πίνακα 4.1.1 παρουσιάζονται και οι 10 δυνατοί συνδυασμοί των βαθμών για το φάρμακο 1.

| | | | | |
|-----|-----|-----|-----|-----|
| 1,2 | 1,3 | 1,4 | 1,5 | 2,3 |
| 2,4 | 2,5 | 3,4 | 3,5 | 4,5 |

Πίνακας 4.1.1: Δυνατοί συνδυασμοί βαθμών για το Φάρμακο 1.

Για κάθε μια από αυτές τις δυνατές επιλογές για το φάρμακο 1 μένουν $\binom{3}{2}$ δυνατές επιλογές για τους βαθμούς του φαρμάκου 2, όπου με την R βρίσκουμε

```
> choose(3,2)
[1] 3
```

δυνατούς συνδυασμούς βαθμών. Οι βαθμοί για το φάρμακο 2 μπορούν να καθοριστούν μόνο εφόσον έχουμε επιλέξει τους βαθμούς για το φάρμακο 1. Για παράδειγμα αν το φάρμακο 1 έχει τους βαθμούς (2,5) τότε το φάρμακο 2 μπορεί να έχει είτε τους βαθμούς (1,3) ή (1,4) ή (3,4). Επομένως όλες οι δυνατές επιλογές για κάθε βαθμό των φαρμάκων 1 και 2 μπορούν να παρουσιαστούν σε έναν πίνακα με 10 σειρές και 3 στήλες. Οι δυνατοί συνδυασμοί θα είναι συνολικά 30, όπου ο καθένας από αυτούς θα έχει υπό την μηδενική υπόθεση ίδια πιθανότητα να προκύψει, που είναι ίση με 1/30. Κάνοντας χρήση της R μπορούμε να βρούμε πως για $n_1 = 2, n_2 = 2$ και $n_3 = 1$ οι δυνατοί συνδυασμοί των 5 βαθμών είναι

```
> x<-c(2,2,1)
> multinom(x,counts = T)
[1] 30
```

πράγματι. Όταν επιλέξουμε τους βαθμούς για το φάρμακο 1 και το φάρμακο 2, οι βαθμοί που περισεύουν (ένας στο συγκεκριμένο παράδειγμα) θα πάνε απευθείας στο φάρμακο 3.

4.2 Περιγραφή του ελέγχου των Kruskal - Wallis

Μια από τις πιο συνηθισμένες αλλά και απλές ερωτήσεις που μας βοηθάει ο έλεγχος των Kruskal-Wallis να απαντήσουμε είναι αν παρατηρούμε κάποια διαφορά ανάμεσα στα s το πλήθος δείγματα- θεραπείες κλπ. Η διαφορά αυτή θα παρατηρείται από τις αποκρίσεις – τις τιμές δηλαδή – των παρατηρήσεων κάθε δείγματος που λαμβάνει μέρος στην έρευνα. Ως μηδενική υπόθεση για τον έλεγχο ορίζουμε ότι οι s το πλήθος θεραπείες δεν παρουσιάζουν καμία διαφορά με εναλλακτική υπόθεση ότι διαφέρουν. Για παράδειγμα οι παρατηρήσεις κάποιας θεραπείας τείνουν να είναι οι μικρότερες από όλες τις άλλες παρατηρήσεις, ενώ οι παρατηρήσεις μιας άλλης θεραπείας τείνουν να είναι οι αμέσως επόμενες μικρότερες κ.ο.κ. Επομένως το μέτρο θέσης των αποκρίσεων αποτελεί χαρακτηριστικό για την κάθε θεραπεία.

Ένα μέτρο θέσης που μπορούμε να χρησιμοποιήσουμε είναι ο μέσος όρος των βαθμών των παρατηρήσεων σε κάθε θεραπεία. Αν οι θεραπείες διαφέρουν σε μεγάλο βαθμό τότε θα αναμένεται και ο μέσος όρος των βαθμών για κάθε θεραπεία να διαφέρει. Αντίθετα, αν οι θεραπείες δεν διαφέρουν σημαντικά τότε θα περιμένουμε και ο μέσος όρος των βαθμών για κάθε θεραπεία να είναι κοντά σε μια τιμή και συγκεκριμένα στον μέσο όρο όλων των βαθμών από όλες τις s θεραπείες.

Ο μέσος όρος των βαθμών για κάθε i θεραπεία δίνεται από την σχέση

$$R_{i.} = \frac{R_{i1} + \dots + R_{in_1}}{n_i}$$

ενώ ο συνολικός μέσος όρος δίνεται από την σχέση

$$R_{..} = \frac{(R_{i1} + \dots + R_{in_1}) + \dots + (R_{s1} + \dots + R_{sn_s})}{N}$$

Γνωρίζουμε ότι το άθροισμα όλων των βαθμών είναι ίσο με τον αριθμό $N(N + 1)/2$ και επομένως θα ισχύει

$$R_{..} = \frac{(N + 1)}{2}$$

και για να δούμε αν ισχύει η μηδενική υπόθεση μπορούμε να ελέγξουμε πόσο κοντά είναι οι τιμές $R_{i.}$ στην τιμή $R_{..}$.

Το στατιστικό που μας βοηθάει να πραγματοποιήσουμε αυτόν τον έλεγχο είναι το ακόλουθο

$$K = \frac{12}{N(N + 1)} \sum_{i=1}^s n_i \left(R_{i.} - \frac{N + 1}{2} \right)^2.$$

Όταν όλα τα $R_{i.}$ είναι ίσα το στατιστικό K παίρνει την τιμή μηδέν ενώ η τιμή του μεγαλώνει ανάλογα με το πόσο διαφορετικά είναι όλα τα $R_{i.}$ μεταξύ τους και η μηδενική υπόθεση απορρίπτεται για μεγάλες τιμές του K , παραδείγματος χάριν

$$K \geq c.$$

Το στατιστικό K μπορεί να γραφτεί και με τον ακόλουθο τρόπο

$$K = \frac{12}{N(N + 1)} \sum_{i=1}^s \frac{R_{i.}^2}{n_i} - 3(N + 1).$$

Για να υπολογίσουμε την κρίσιμη τιμή c για οποιοδήποτε επίπεδο σημαντικότητας α αρκεί να υπολογίσουμε την κατανομή του στατιστικού υπό την μηδενική υπόθεση. Η κατανομή του είναι αυτή που δίνεται από την τελευταία σχέση της ενότητας 4.1 και οι κρίσιμες τιμές δίνονται σε πίνακες για συγκεκριμένα μεγέθη δειγμάτων και συγκεκριμένο πλήθος θεραπειών.

Ένας εναλλακτικός τρόπος για τον υπολογισμό της πιθανότητας

$$P_{H_0}(K \geq c)$$

καθώς τα μεγέθη n_1, \dots, n_s τείνουν στο άπειρο είναι να προσεγγίσουμε την κατανομή του στατιστικού K από την κατανομή χ^2 με $s - 1$ βαθμούς ελευθερίας.

Στο στατιστικό πακέτο της R η πιθανότητα αυτή υπολογίζεται με την εντολή `pchisq(c,s - 1,lower.tail = F)`.

Για παράδειγμα, για $c = 5.7800$ και $s = 3$ παίρνουμε στην R την παρακάτω τιμή

```
> pchisq(5.7800,2,lower.tail = F)
[1] 0.05557621.
```

4.3 Όταν έχουμε ίσες παρατηρήσεις

Ως τώρα έχουμε υποθέσει πως στις s το πλήθος θεραπείες δεν προκύπτουν τιμές παρατηρήσεων που να είναι ίσες μεταξύ τους. Η συνθήκη αυτή ωστόσο στις περισσότερες έρευνες παραβιάζεται καθώς είναι αρκετά συχνό δύο και περισσότερες παρατηρήσεις να έχουν ίδιες τιμές. Ο τρόπος με τον οποίο αντιμετωπίζουμε την περίπτωση αυτή είναι ανάλογος με αυτόν

που είδαμε στα προηγούμενα κεφάλαια. Χρησιμοποιούμε δηλαδή τους μέσους βαθμούς των παρατηρήσεων.

Συγκεκριμένα, ας υποθέσουμε ότι οι N το πλήθος παρατηρήσεις μπορούν να πάρουν συνολικά e διακεκριμένες τιμές και ότι d_1 από αυτές τις N παρατηρήσεις είναι ίσες με την μικρότερη τιμή, d_2 είναι ίσες με την αμέσως επόμενη τιμή κ.ο.κ

Οι μέσοι βαθμοί των n_i το πλήθος παρατηρήσεων που έλαβαν την i θεραπεία συμβολίζονται με $R_{i1}^*, \dots, R_{in_i}^*$ και το άθροισμα τους με R_i^* .

Το στατιστικό ελέγχου δίνεται από την ακόλουθη σχέση

$$K^* = \frac{[12/N(N+1)] \sum R_i^{*2}/n_i - 3(N+1)}{1 - \sum(d_i^3 - d_i)/(N^3 - N)}$$

και μεγάλες τιμές του, έστω $K^* \geq c$, όπου η c θα είναι μια μεγάλη τιμή, αποτελούν ένδειξη κατά της μηδενικής υπόθεσης.

Σε περιπτώσεις σαν αυτή ο υπολογισμός της κατανομής του στατιστικού υπό την μηδενική υπόθεση απαιτεί αρκετές πράξεις και συνηθίζουμε να προσεγγίζουμε την κατανομή του από την κατανομή X^2 με $s - 1$ βαθμούς ελευθερίας.

4.4 Η εντολή του ελέγχου για το στατιστικό πακέτο της R

Ο έλεγχος Kruskal – Wallis στην R πραγματοποιείται με την χρήση της εντολής `Kruskal.test(x, g)`.

Σαν ορίσματα δέχεται το ενοποιημένο δείγμα που δίνεται από το διάνυσμα x και περιέχει όλες τις παρατηρήσεις του προβλήματος. Χρειαζόμαστε επίσης και ένα διάνυσμα g το οποίο στην ουσία δηλώνει από ποιο από τα k δείγματα προέρχεται κάθε παρατήρηση του ενοποιημένου δείγματος.

Δεδομένων των x, g μπορούμε να τρέξουμε τον έλεγχο και να πάρουμε ως αποτελέσματα τα ακόλουθα:

```
> kruskal.test(x ,g)
```

```
Kruskal-Wallis rank sum test
```

```
data: x and g
Kruskal-Wallis chi-squared = .., df
= .., p-value = ...
```

Ο έλεγχος επιστρέφει δηλαδή την τιμή του στατιστικού ελέγχου **Kruskal-Wallis chi-squared**, τους βαθμούς ελευθερίας (**df**) και την p -τιμή του ελέγχου από την οποία μπορούμε κατά τα γνωστά να αποφασίσουμε αν θα απορρίψουμε ή θα δεχτούμε την μηδενική υπόθεση σε επίπεδο σημαντικότητας α .

Αν p – τιμή $< \alpha$ απορρίπτουμε H_0 .

Είναι σημαντικό να γνωρίζουμε πως όταν πραγματοποιούμε έναν έλεγχο που συγκρίνουμε πολλά δείγματα αυξάνεται η πιθανότητα να προκύψει σφάλμα τύπου I, η πιθανότητα δηλαδή η πιθανότητα να απορρίψουμε την μηδενική υπόθεση ενώ αυτή είναι αληθής. Για να αποφύγουμε αυτή την πιθανότητα, μπορούμε να προσαρμόσουμε κατάλληλα το επίπεδο σημαντικότητας α , χρησιμοποιώντας τον τύπο του **Bonferroni**

$$\alpha_B = \frac{\alpha}{k}$$

όπου το α είναι το αρχικό επίπεδο σημαντικότητας και k είναι ο αριθμός των δειγμάτων. Σε περίπτωση που αποφασίσουμε ότι τα δείγματα διαφέρουν μεταξύ τους μπορούμε να προβούμε σε συγκρίσεις ανά δύο χρησιμοποιώντας τον κατάλληλο μη παραμετρικό έλεγχο για δύο δείγματα.

Εφαρμογή 4.4.1

Ο Πίνακας 4.4.1. παρουσιάζει μετρήσεις για την απαλότητα ενός συγκεκριμένου είδους χαρτιού που πραγματοποιήθηκαν σε 8 διαφορετικά εργαστήρια. Θέλουμε να ελέγξουμε ότι δεν υπάρχει διαφορά ανάμεσα στα 8 εργαστήρια.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|-----|-----|-----|-----|-----|-----|-----|-----|
| 173 | 135 | 165 | 185 | 108 | 120 | 125 | 135 |
| 185 | 165 | 175 | 137 | 112 | 135 | 138 | 155 |
| 141 | 145 | 155 | 162 | 113 | 140 | 125 | 140 |
| 133 | 160 | 135 | 125 | 137 | 135 | 155 | 155 |
| 160 | 120 | 160 | 157 | 146 | 130 | 173 | 140 |

Πίνακας 4.4.1: Απαλότητα χαρτιού.

Για αρχή πρέπει να περάσουμε τις μετρήσεις από κάθε εργαστήριο σε 8 διανύσματα.

```
> one<-c(173,185,141,133,160)
> one
[1] 173 185 141 133 160
> two<-c(135,165,145,160,120)
> two
[1] 135 165 145 160 120
> three<-c(165,175,155,135,160)
> three
[1] 165 175 155 135 160
> four<-c(185,137,162,125,157)
> four
[1] 185 137 162 125 157
> five<-c(108,112,113,137,146)
> five
[1] 108 112 113 137 146
> six<-c(120,135,140,135,130)
> six
[1] 120 135 140 135 130
> seven<-c(125,138,125,155,173)
> seven
[1] 125 138 125 155 173
> eight<-c(135,155,140,155,140)
> eight
[1] 135 155 140 155 140
```

```

> sample<-rbind(one,two,three,four,five,six,seven,eight)
> sample
      [,1] [,2] [,3] [,4] [,5]
one   173  185  141  133  160
two   135  165  145  160  120
three 165  175  155  135  160
four  185  137  162  125  157
five  108  112  113  137  146
six   120  135  140  135  130
seven 125  138  125  155  173
eight 135  155  140  155  140

> rownames(sample)<-c("one","two","three","four","five","six","seven","eight")
> sample<-as.table(sample)
> sample
      A B C D E
one   173 185 141 133 160
two   135 165 145 160 120
three 165 175 155 135 160
four  185 137 162 125 157
five  108 112 113 137 146
six   120 135 140 135 130
seven 125 138 125 155 173
eight 135 155 140 155 140

> rank(sample)
 [1] 36.5 13.0 34.5 39.5  1.0  4.5  7.0 13.0 39.5
[10] 34.5 38.0 16.5  2.0 13.0 18.0 26.5 22.0 23.0
[19] 26.5 33.0  3.0 20.0  7.0 20.0 10.0 31.0 13.0
[28]  7.0 16.5 13.0 26.5 26.5 31.0  4.5 31.0 29.0
[37] 24.0  9.0 36.5 20.0

> x<-matrix(x,ncol = 5,byrow = T)
> x
      [,1] [,2] [,3] [,4] [,5]
[1,] 36.5 13.0 34.5 39.5  1.0
[2,]  4.5  7.0 13.0 39.5 34.5
[3,] 38.0 16.5  2.0 13.0 18.0
[4,] 26.5 22.0 23.0 26.5 33.0
[5,]  3.0 20.0  7.0 20.0 10.0
[6,] 31.0 13.0  7.0 16.5 13.0
[7,] 26.5 26.5 31.0  4.5 31.0
[8,] 29.0 24.0  9.0 36.5 20.0

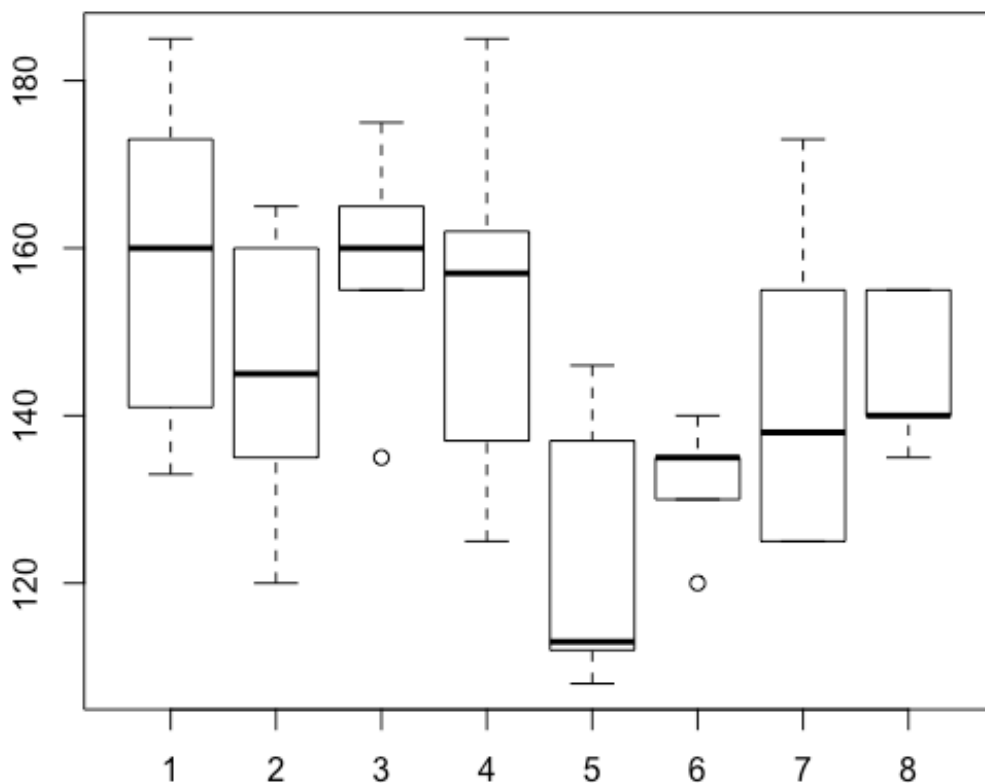
> sum(x[1,])
[1] 124.5
> sum(x[2,])
[1] 98.5
> sum(x[3,])
[1] 87.5
> sum(x[4,])
[1] 131
> sum(x[5,])
[1] 60
> sum(x[6,])
[1] 80.5
> sum(x[7,])

```

```
[1] 119.5
> sum(x[8,])
[1] 118.5
```

Μπορούμε να συγκρίνουμε τα 8 δείγματα και γραφικά με την βοήθεια των θηκογραμμάτων τους.

```
> boxplot(one,two,three,four,five,six,seven,eight)
```



Διάγραμμα 5.1.1.6.2: Θηκόγραμμα για τα οκτώ δείγματα

Στην συνέχεια δημιουργούμε το ενιαίο διάνυσμα που περιέχει τις 40 μετρήσεις που έγιναν από τα 8 εργαστήρια.

```
> laboratories<-c(one,two,three,four,five,six,seven,eight)
> laboratories
[1] 173 185 141 133 160 135 165 145 160 120
[11] 165 175 155 135 160 185 137 162 125 157
[21] 108 112 113 137 146 120 135 140 135 130
[31] 125 138 125 155 173 135 155 140 155 140
```

Για να χρησιμοποιήσουμε την εντολή του ελέγχου θα πρέπει να δημιουργήσουμε το διάνυσμα g που δηλώνει σε ποιο από τα 8 διανύσματα ανήκει η κάθε μέτρηση. Σε ποιο από τα 8 εργαστήρια δηλαδή.

```
> g<-c(rep(1,5),rep(2,5),rep(3,5),rep(4,5),rep(5,5),rep(6,5),rep(7,5),rep(8,5))
```

Τέλος, πραγματοποιούμε τον έλεγχο.

```
> kruskal.test(laboratories,g)
```

Kruskal-Wallis rank sum test

```
data: laboratories and g  
Kruskal-Wallis chi-squared = 12.56, df  
= 7, p-value = 0.0836
```

Παρατηρούμε το στατιστικό ελέγχου έχει τιμή ίση με 12.56 και ότι η p -τιμή του ελέγχου είναι > 0.05 επιπέδου σημαντικότητας και επομένως δεν έχουμε καμία ένδειξη κατά της μηδενικής υπόθεσης. Τα εργαστήρια επομένως δεν φαίνεται να διαφέρουν. Επίσης από τους πίνακες τις X^2 κατανομής και για 7 βαθμούς ελευθερίας βρίσκουμε κρίσιμη τιμή ίση με 14.067 που είναι μεγαλύτερη από την τιμή του στατιστικού H και επομένως καταλήγουμε στο να δεχτούμε την H_0 .

5. Ο έλεγχος Friedman

5.1 Οι βαθμοί σε τυχαιοποιημένα πλήρη μπλοκ σχεδιασμών

Στο προηγούμενο κεφάλαιο ασχοληθήκαμε με την σύγκριση s το πλήθος θεραπειών που εφαρμόσαμε σε N το πλήθος τυχαία επιλεγμένους ασθενείς: n_1 ασθενείς για την θεραπεία 1, n_2 ασθενείς για την θεραπεία 2 κτλ. Ο έλεγχος στις περιπτώσεις αυτές επηρεάζεται φυσικά σε έναν βαθμό από την επιλογή των μονάδων που θα λάβουν την κάθε θεραπεία και τα αποτελέσματα που θα προκύψουν μπορεί να μην είναι ακριβή. Για παράδειγμα μπορεί σε μια έρευνα για την σύγκριση s φαρμάκων κάποιο φάρμακο να παρουσιάσει ανεπιθύμητες παρενέργειες σε ένα άτομο που έχει αλλεργία σε κάποιο από τα συστατικά του. Σε αυτές τις περιπτώσεις, για να απομονώσουμε τη μεταβλητότητα που προκαλεί μία ανεπιθύμητη μεταβλητή, ακολουθούμε την τεχνική της ομαδοποίησης των πειραματικών μονάδων σε μπλοκ με παρόμοια χαρακτηριστικά και συγκρίνουμε όσες μονάδες λαμβάνουν διαφορετικές θεραπείες μέσα σε αυτά τα μπλοκ. Η διαδικασία είναι ένας τρόπος ανάλογος με αυτόν που είδαμε στο Κεφάλαιο 3 όπου οι παρατηρήσεις ήταν σε ζεύγη.

Για τη μελέτη s το πλήθος θεραπειών χρειαζόμαστε έναν μπλοκ σχεδιασμό όπου το πλήθος των επιπέδων του παράγοντα θεραπεία να ισούται με το πλήθος των πειραματικών μονάδων που αντιστοιχούν στο κάθε επίπεδο του μπλοκ παράγοντα. Οι s πειραματικές μονάδες που αντιστοιχούν σε κάθε ομάδα τοποθετούνται τυχαία στις s θεραπείες. Ενδεικτικά, σε ιατρικά πειράματα τέτοιες ομάδες μπορούν να αποτελούνται από ζώα προερχόμενα από την ίδια γέννα ή ασθενείς ίδιου φύλου, ηλικίας ή με παρόμοια νόσο. Άλλα παραδείγματα προέρχονται από περιπτώσεις στις οποίες οι s θεραπείες συγκρίνονται με την εφαρμογή τους στο ίδιο άτομο σε διαφορετικές χρονικές περιόδους, όπως για παράδειγμα, η σύγκριση διαφορετικών θεραπειών για το πονοκέφαλο.

Παράδειγμα 5.1.1

Έστω ότι θέλουμε να συγκρίνουμε 3 διαφορετικές θεραπείες (A, B, Γ) σε 12 ασθενείς. Στο προηγούμενο κεφάλαιο θεωρήσαμε ότι μπορούσαμε να ταξινομήσουμε κατά τυχαίο τρόπο τους διαθέσιμους ασθενείς στις διαθέσιμες θεραπείες. Μια τέτοια τυχαία ανάθεση δεν είναι όμως πάντα ιδανική. Θα μπορούσε για παράδειγμα κάποιος οργανισμός να έχει περισσότερες αντιστάσεις ή αντίθετα να είναι περισσότερο επιρρεπής σε κάποια θεραπεία και η τυχαία αυτή ανάθεση θα είχε ως αποτέλεσμα οι αποκρίσεις (τα αποτελέσματα) της έρευνας να διαφέρουν αρκετά μεταξύ τους και να οδηγηθούμε σε λάθος συμπεράσματα. Προκειμένου να αποφύγουμε τέτοιες καταστάσεις η πιο κατάλληλη μέθοδος θα ήταν ο κάθε ένας από τους 12 ασθενείς να δοκιμάσει και τις τρεις θεραπείες με τυχαία σειρά την κάθε μια. Αυτή η πρακτική ωστόσο απαιτεί χρόνο που δεν έχουμε πάντα διαθέσιμο. Επιπλέον πάλι μπορεί να δημιουργηθεί σύγχυση λόγω του μεγάλου χρονικού διαστήματος ανάμεσα στην λήψη της κάθε θεραπείας. Ο πιο κατάλληλος τρόπος να ομαδοποιήσουμε τις παρατηρήσεις μας είναι να χωρίσουμε τους 12 ασθενείς σε τέσσερις ομάδες των τριών έτσι ώστε να υπάρχει ομογένεια στα χαρακτηριστικά

των ασθενών που ανήκουν στην ίδια ομάδα. Θα μπορούσαν για παράδειγμα να έχουν σχετικά ίδια ηλικία. Οι τρεις ασθενείς σε κάθε ομάδα επομένως μπορούν να λάβουν κατά τυχαίο τρόπο μια από τις τρεις διαθέσιμες θεραπείες. Στο τέλος της θεραπείας βάσει εξετάσεων και μετρήσεων ταξινομούνται για κάθε μια από τις τέσσερις ομάδες οι τρεις θεραπείες. Τα αποτελέσματα δίνονται στον Πίνακα 5.1.1.

| Θεραπεία | Ομάδα 1 | Ομάδα 2 | Ομάδα 3 | Ομάδα 4 |
|----------|---------|---------|---------|---------|
| A | 2 | 2 | 3 | 2 |
| B | 3 | 1 | 2 | 1 |
| Γ | 1 | 3 | 1 | 3 |

Πίνακας 5.1.1: Τα αποτελέσματα για την έρευνα του Παραδείγματος 5.1.1

Ορίζουμε την μηδενική υπόθεση ως ότι δεν παρατηρείται καμία διαφορά στις επιδράσεις των τριών φαρμάκων. Υπάρχουν $3! = 6$ δυνατές διατάξεις των βαθμών 1, 2, 3 και η κάθε μία από αυτές είναι εξίσου πιθανή να παρατηρηθεί υπό την μηδενική υπόθεση και συγκεκριμένα αυτή η πιθανότητα είναι ίση με $1/6$ για την κάθε διάταξη. Οι αποθέσεις των ασθενών σε κάθε ένα από τα τέσσερα γκρουπ είναι ανεξάρτητες μεταξύ τους και για τον λόγο αυτό η πιθανότητα καθενός από τους 6^4 δυνατούς συνδυασμούς για τους βαθμούς των τεσσάρων γκρουπ είναι $(1/6)^4$.

Μπορούμε να γενικεύσουμε το παραπάνω παράδειγμα. Για τη σύγκριση s θεραπειών, αριθμούμε τις πειραματικές μονάδες που αντιστοιχούν σε κάθε ομάδα από το 1 έως το s . Επομένως, αν R_{ij} δηλώνει το βαθμό της i -οστής θεραπείας στο j -οστό μπλοκ, τότε τα (R_{1j}, \dots, R_{sj}) για κάθε j είναι μία παράθεση των ακεραίων $(1, \dots, s)$. Υπο την μηδενική υπόθεση ότι δεν υπάρχει διαφορά μεταξύ των θεραπειών, η τυχαία αντιστοίχιση της κάθε μιας από τις s το πλήθος θεραπείες σε κάθε ομάδα ατόμων συνεπάγεται ότι όλες οι

$$s! = 1 * \dots * s$$

δυνατές διατάξεις των (R_{1j}, \dots, R_{sj}) έχουν ίση πιθανότητα να παρατηρηθούν, και μάλιστα αυτή είναι ίση με $\frac{1}{s!}$.

Δεδομένου ότι οι αντιστοιχίσεις, και άρα οι βαθμοί, διαφορετικών μπλοκ είναι ανεξάρτητες, η συνολική κατανομή των βαθμών υπό την μηδενική υπόθεση δίνεται από τον ακόλουθο τύπο

$$P_{H_0}(R_{11} = r_{11}, \dots, R_{s1} = r_{s1}; \dots; R_{1N} = r_{1N}, \dots, R_{sN} = r_{sN}) = \left(\frac{1}{s!}\right)^N$$

όπου N είναι ο αριθμός των μπλοκ.

5.2 Περιγραφή του έλεγχου του Friedman και παρουσίαση μιας ειδικής περίπτωσης αυτού.

Πολλές διαφορετικές εναλλακτικές υποθέσεις μπορούν να ελεγχθούν έναντι της υπόθεσης μηδενικής υπόθεσης όσον αφορά τη μη υπαρξη διαφορών μεταξύ των θεραπειών. Θα έχουμε ως εναλλακτική υπόθεση ότι οι θεραπείες επηρεάζουν το βαθμό απόκρισης για κάποιες παρατηρήσεις. Ως εκ τούτου, θα προκύπτει μια διάταξη των θεραπειών που θα βασίζεται στις αποκρίσεις των παρατηρήσεων.

Η συνολικός βαθμός της i -οστής θεραπείας θα δίνεται από τον μέσο όρο των βαθμών

$$R_i = \frac{R_{i1} + R_{i2} + \dots + R_{iN}}{N}$$

που προκύπτει για τα N το πλήθος μπλοκ.

Στην περίπτωση που οι θεραπείες διαφέρουν σημαντικά μεταξύ τους, αυτό θα αντανακλάται στις τιμές των R_i . Αντιθέτως, όταν η μηδενική υπόθεση είναι αληθής, δηλαδή οι θεραπείες δε θα διαφέρουν μεταξύ τους ως προς την αποτελεσματικότητα, τότε τα R_i θα λαμβάνουν παραπλήσιες τιμές και επίσης θα βρίσκονται πολύ κοντά και στο γενικό μέσο όρο που δίνεται από την ακόλουθη σχέση

$$R_{..} = \frac{(R_{11} + \dots + R_{s1}) + \dots + (R_{1N} + \dots + R_{sN})}{sN}$$

Δεδομένου ότι ο αριθμητής, που είναι το συνολικό άθροισμα των βαθμών, είναι ίσος με $N \frac{1}{2} s(s+1)$, η προηγούμενη σχέση ανάγεται στην

$$R_{..} = \frac{1}{2} (s+1).$$

Ένα κατάλληλο μέτρο σύγκρισης για το πόσο κοντά βρίσκονται τα R_i στο $\frac{1}{2} (s+1)$ είναι το άθροισμα των διαφορών τετραγώνων $[R_i - \frac{1}{2} s(s+1)]^2$ καθώς επίσης και το στατιστικό του έλεγχου του Friedman που δίνεται από τον τύπο

$$Q = \frac{12N}{s(s+1)} \sum_{i=1}^s [R_i - \frac{1}{2} (s+1)]^2$$

όπου ο παράγοντας $\frac{12N}{s(s+1)}$ επιλέγεται ώστε να προσεγγίζει την κατανομή του στατιστικού υπό τη μηδενική υπόθεση, όταν το πλήθος N των μπλοκ είναι μεγάλο.

Το στατιστικό Q είναι ίσο με το μηδέν όταν τα R_i είναι όλα ίσα και λαμβάνει μεγάλες τιμές όταν υπάρχουν σημαντικές διαφορές μεταξύ των R_i . Επομένως, η μηδενική υπόθεση θα απορρίπτεται για μεγάλες τιμές του Q , παραδείγματος χάριν για

$$Q \geq c$$

όπου c είναι η τιμή μιας μεγάλης σταθεράς. Μια εναλλακτική έκφραση για το Q προκύπτει εάν υψώσουμε στο τετράγωνο το δεξί τμήμα της εξίσωσης και αντικαταστήσουμε το R_i με R_i/N , όπου $R_i = R_{i1} + \dots + R_{iN}$ είναι η θέση της i -οστης θεραπείας.

Προκύπτει, λοιπόν,

$$Q = \frac{12}{Ns(s+1)} \sum_{i=1}^s R_i^2 - 3N(s+1).$$

Η κρίσιμη τιμή c της τελευταίας ανίσωσης μπορεί να υπολογιστεί από την κατανομή του στατιστικού Q υπολογισμένη υπό την μηδενική υπόθεση. Υπάρχουν έτοιμοι πίνακες που δίνουν τις κρίσιμες τιμές για το στατιστικό για συγκεκριμένες τιμές των s και N . Για τις τιμές των s και N που βρίσκονται εκτός του εύρους των πινάκων, μπορούμε να χρησιμοποιήσουμε κατάλληλη προσέγγιση. Όπως και στην περίπτωση του στατιστικού Kruskal-Wallis, μπορούμε να προσεγγίσουμε την πιθανότητα $P_{H_0}(Q \geq c)$ από την επιφάνεια που βρίσκεται δεξιά του c κάτω από την χ^2 -καμπύλη με $s-1$ βαθμούς ελευθερίας, έτσι ώστε

$$P_H(Q \geq c) \approx \Psi_{s-1}(c)$$

όπου η ανωτέρω προσέγγιση βασίζεται στο οριακό θεώρημα που αναφέρει ότι η αριστερή πλευρά τείνει στη δεξιά όταν $N \rightarrow \infty$.

Η ακρίβεια της χ^2 -προσέγγισης βελτιώνεται όσο τα s και N αυξάνουν.

Στην R μπορούμε να υπολογίσουμε την πιθανότητα $P_H(Q \geq 6.45)$ με 3 βαθμούς ελευθερίας με την προσέγγιση κάνοντας χρήση της εντολής

```
> pchisq(6.45,3,lower.tail = F)
[1] 0.09165537
```

και καταλήγουμε στο συμπέρασμα ότι αφού η p-τιμή δεν είναι στατιστικά σημαντική δεν έχουμε σοβαρές ενδείξεις κατά της μηδενικής υπόθεσης που υποστηρίζει ότι δεν υπάρχει διαφορά ανάμεσα στις τέσσερις θεραπείες.

5.3 Όταν έχουμε παρατηρήσεις με ίσες τιμές

Έως τώρα έχουμε υποθέσει ότι σε ένα μπλοκ δεν υπάρχουν παρατηρήσεις που να λαμβάνουν τις ίδιες τιμές. Στην περίπτωση που αυτό συμβεί, μπορούμε να χρησιμοποιήσουμε τη μέθοδο που περιγράψαμε για πρώτη φορά στο δεύτερο κεφάλαιο και να αποδώσουμε σε κάθε παρατήρηση τις οποίες η τιμή δεν είναι διακεκριμένη, τον μέσο βαθμό που της αναλογεί. Για τα τυχαιοποιημένα μπλοκ, ως μέσο βαθμό μιας παρατήρησης ορίζουμε τη μέση τιμή των βαθμών των παρατηρήσεων που βρίσκονται στο ίδιο μπλοκ με αυτή. Υποθέτουμε ότι οι s το πλήθος παρατηρήσεις στο πρώτο μπλοκ παίρνουν e_1 διακριτές τιμές, όπου d_{11} παρατηρήσεις ισούνται με τη μικρότερη τιμή, d_{22} με την αμέσως επόμενη μικρότερη κτλ. Για παράδειγμα, έστω ότι $s = 5$ και ότι οι παρατηρήσεις στο ένα από τα δύο μπλοκ είναι 1.3, 1.1, 1.1, 1.6, 1.1 και 1.9, 1.7, 1.9, 1.9, 1.7. Οπότε $e_1 = 3$ με $d_{11} = 3$, $d_{21} = 1$, $d_{31} = 1$ και $e_2 = 2$ με $d_{12} = 2$, $d_{22} = 3$.

Για μία παρατήρηση που λαμβάνει την i -οστή θεραπεία και βρίσκεται στο j -οστό μπλοκ, συμβολίζουμε το μέσο βαθμό της με R_{ij}^* και υπολογίζουμε το άθροισμα των μέσων βαθμών της από τον τύπο

$$R_i^* = R_{i1}^* + \dots + R_{iN}^*.$$

Η τιμή του στατιστικού Q^* στην περίπτωση αυτή υπολογίζεται από τον τύπο

$$Q^* = \frac{\frac{12}{Ns(s+1)} \sum R_i^{*2} - 3N(s+1)}{1 - \sum_{j=1}^N \sum_{i=1}^{2j} (d_{ij}^3 - d_{ij}) / Ns(s^2 - 1)}$$

και η μηδενική υπόθεση για τη μη ύπαρξη διαφοράς μεταξύ των θεραπειών απορρίπτεται ξανά για μεγάλες τιμές του στατιστικού, παραδείγματος χάριν

$$Q^* \geq c.$$

Η κατανομή υπό τη μηδενική υπόθεση θα προσεγγιστεί από την χ^2 -κατανομή με $s - 1$ βαθμούς ελευθερίας, δίνοντας ικανοποιητικές προσεγγίσεις για $sN \geq 30$.

5.4 Όταν θέλουμε να συγκρίνουμε δύο θεραπείες για ομαδοποιημένα δεδομένα

Όταν έχουμε 2 θεραπείες ο τυχαιοποιημένος πλήρης μπλοκ σχεδιασμός απλουστεύεται σε συγκρίσεις ανά ζεύγη. Σε αυτήν την περίπτωση, έχουμε τους βαθμούς 1 και 2 για κάθε μπλοκ. Υποθέτουμε ότι σε A το πλήθος μπλοκ η θεραπεία 1 (αντιστοίχως 2) έχει τον βαθμό 1 (αντιστοίχως 2), ενώ στα υπόλοιπα N-A μπλοκ ισχύει το αντίθετο. Τότε το άθροισμα των βαθμών δίνεται από τις σχέσεις:

$$R_1 = A + 2(N - A) = 2N - A; R_2 = N - A + 2A = N + A.$$

Για s=2 και χρησιμοποιώντας τη σχέση για το στατιστικό Q που είδαμε στην προηγούμενη ενότητα, προκύπτει το νέο στατιστικό:

$$Q = 4N \left(\frac{A}{N} - \frac{1}{2} \right)^2$$

Η μηδενική υπόθεση απορρίπτεται όταν το Q είναι πολύ μεγάλο.

5.5 Μια ειδική μορφή του ελέγχου Friedman, ο έλεγχος McNemar.

Το στατιστικό Q^* που ορίσαμε στην ενότητα 6.3 μπορεί να πάρει μια πιο απλοποιημένη μορφή στην περίπτωση όπου οι αποκρίσεις των παρατηρήσεων είναι δίτιμες μεταβλητές που μπορούν να έχουν ως τιμές: ναι-όχι, επιτυχία-αποτυχία κλπ. Έστω ότι η επιτυχία συμβολίζεται με την τιμή 1 ενώ η αποτυχία συμβολίζεται με την τιμή 0, τότε σε κάθε μπλοκ θα έχουμε s το πλήθος ζεύγη από 0 και 1. Έστω ότι συμβολίζουμε με L_j το πλήθος των επιτυχιών στο j μπλοκ και με B_i το πλήθος των επιτυχιών για την i θεραπεία. Υπό αυτές τις συνθήκες το στατιστικό ελέγχου Q^* παίρνει την παρακάτω μορφή:

$$Q^* = \frac{s(s-1) \sum_{i=1}^s (B_i - \bar{B})^2}{s \sum_{j=1}^N L_j - \sum_{j=1}^N L_j^2},$$

όπου το \bar{B} δίνεται από την σχέση

$$\bar{B} = \frac{(B_1 + \dots + B_s)}{s}.$$

Ο υπολογισμός του στατιστικού Q^* μπορεί να απλοποιηθεί αν στη θέση του αριθμητή αντικαταστήσουμε την σχέση:

$$(s-1) \left[s \sum B_i^2 - \left(\sum B_i \right)^2 \right].$$

Μια ειδική περίπτωση του παραπάνω στατιστικού είναι όταν οι θεραπείες που θέλουμε να συγκρίνουμε είναι s = 2 και έχουμε ζεύγη παρατηρήσεων που μπορούν να πάρουν τις τιμές 0 και 1. Τότε υπάρχουν τέσσερις δυνατές αποκρίσεις για κάθε μπλοκ. Αυτές είναι οι (0,0), (0,1), (1,0) και (1,1). Συνήθως οι παρατηρήσεις στην περίπτωση αυτή παρουσιάζονται με την μορφή ενός 2 x 2 πίνακα συνάφειας. Ο Πίνακας 5.5.1 δίνει τη μορφή ενός τέτοιου πίνακα.

| | | |
|--|------------|----------|
| | Θεραπεία 2 | |
| | Επιτυχία | Αποτυχία |

| | | | |
|------------|----------|---|---|
| Θεραπεία 1 | Επιτυχία | A | B |
| | Αποτυχία | C | D |

Πίνακας 5.5.1 Μορφή ενός 2 x 2 πίνακα συνάφειας

Όπου η επιτυχία είναι το 0 και η αποτυχία το 1. Επομένως αν αντικαταστήσουμε στην τελευταία σχέση για την τιμή του στατιστικού Q^* τις τιμές από τον πίνακα παίρνουμε τον νέο τύπο για τον υπολογισμό του στατιστικού που είναι:

$$Q^* = \frac{(B - C)^2}{B + C}.$$

Το στατιστικό αυτό είναι γνωστό και ως στατιστικό ελέγχου του McNemar και η κατανομή του προσεγγίζεται από την χ^2 με 1 βαθμό ελευθερίας.

5.6 Η εντολή για τον έλεγχο του Friedman και του McNemar στο στατιστικό πακέτο της R.

5.6.1 Ο έλεγχος του Friedman

Στην R ο έλεγχος πραγματοποιείται με την εντολή `Friedman.test(x)`

Το μόνο που χρειάζεται να κάνει ο χρήστης είναι να περάσει τα δεδομένα σε έναν πίνακα x και μετά η R χωρίς καμία επιπλέον πληροφορία πραγματοποιεί τον έλεγχο. Τα αποτελέσματα της εντολής παρουσιάζονται παρακάτω

```
> friedman.test(data)
```

```
Friedman rank sum test
```

```
data: x
```

```
Friedman chi-squared = ..., df = ..., p-value =...
```

Η εντολή επιστρέφει δηλαδή την τιμή του στατιστικού F_r , τους βαθμούς ελευθερίας df και την **p-τιμή** του ελέγχου που μας βοηθάει να αποφασίσουμε αν θα αποδεχτούμε ή θα απορρίψουμε την μηδενική υπόθεση.

Υπενθυμίζουμε ότι αν **p-τιμή** < α (επίπεδο σημαντικότητας) τότε **απορρίπτουμε** την μηδενική υπόθεση του ελέγχου.

Είναι σημαντικό να γνωρίζουμε πως όταν πραγματοποιούμε έναν έλεγχο που συγκρίνουμε πολλά δείγματα αυξάνεται η πιθανότητα να προκύψει σφάλμα τύπου I, η πιθανότητα δηλαδή η πιθανότητα να απορρίψουμε την μηδενική υπόθεση ενώ αυτή είναι αληθής. Για να αποφύγουμε αυτή την πιθανότητα, μπορούμε να προσαρμόσουμε κατάλληλα το επίπεδο σημαντικότητας α , χρησιμοποιώντας τον τύπο του **Bonferroni**

$$\alpha_B = \frac{\alpha}{k},$$

όπου το α είναι το αρχικό επίπεδο σημαντικότητας και k είναι ο αριθμός των δειγμάτων.

Σε περίπτωση που αποφασίσουμε ότι τα δείγματα διαφέρουν μεταξύ τους μπορούμε να προβούμε σε συγκρίσεις ανά δύο χρησιμοποιώντας τον κατάλληλο μη παραμετρικό έλεγχο για δύο δείγματα.

Εφαρμογή 5.6.1

Μια εταιρία σε μια προσπάθεια να βελτιώσει τους χρόνους καθυστέρησης προσέλευσης επτά υπαλλήλων της καθημερινά στον χώρο εργασίας τους αποφάσισε να εφαρμόσει την εξής στρατηγική. Για τον πρώτο μήνα, κάθε υπάλληλος που φτάνει αργοπορημένα στην δουλειά θα χρεώνεται 5 €. Τον επόμενο μήνα η «ποινή» για κάθε φορά που θα φτάνει αργά θα ανέρχεται στα 10 €.

Στον Πίνακα 6. δίνονται οι φορές που κάθε ένας από τους επτά υπαλλήλους καθυστέρησε να πάει στην εταιρία. Αρχικά στην πρώτη στήλη δίνονται οι φορές που κάθε υπάλληλος αργούσε πριν εφαρμοστεί η νέα στρατηγική. Στις επόμενες 2 στήλες βλέπουμε τα αποτελέσματα για τον πρώτο και τον δεύτερο μήνα εφαρμογής της «ποινής».

Θέλουμε να ελέγξουμε αν η νέα στρατηγική είχε το επιθυμητό αποτέλεσμα. Να μειώσει δηλαδή τις φορές που οι υπάλληλοι φτάνουν στην εταιρία αργότερα από την προβλεπόμενη ώρα. Τα αποτελέσματα της έρευνας παρουσιάζονται στον Πίνακα 5.6.1.

| Υπάλληλος | Πόσες φορές καθυστέρησε | | |
|-----------|-------------------------|----------------------|----------------------|
| | πριν | 1 ^ο μήνας | 2 ^ο μήνας |
| 1 | 16 | 13 | 12 |
| 2 | 10 | 5 | 2 |
| 3 | 7 | 8 | 9 |
| 4 | 13 | 11 | 5 |
| 5 | 17 | 2 | 6 |
| 6 | 10 | 7 | 9 |
| 7 | 11 | 6 | 7 |

Πίνακας 5.6.1: Χρόνοι καθυστέρησης των υπαλλήλων, πριν, 1 μήνα και δύο μήνες μετά την εφαρμογή της νέας στρατηγικής εφαρμογής ποινών για τις καθυστερήσεις.

Η μηδενική υπόθεση είναι ότι καμία από τις 2 στρατηγικές δεν είχε αποτέλεσμα ενώ η εναλλακτική ότι με τις νέες στρατηγικές θα μειωθούν οι φορές που κάθε υπάλληλος πήγαινε αργοπορημένα στην δουλειά.

$$H_0: m_B = m_1 = m_2$$

$$H_1: \text{τουλάχιστον μια θα έχει μείωση.}$$

Στο Πίνακα 5.6.1.2 μπορούμε να δούμε τον βαθμό- κατάταξη κάθε χρόνου καθυστέρησης σε κάθε υπάλληλο.

| Υπάλληλος | Πόσες φορές έφτασε αργοπορημένα στην δουλειά | | | | | |
|-----------|--|--------|----------------------|--------|----------------------|--------|
| | Πριν | Βαθμός | 1 ^ο Μήνας | Βαθμός | 2 ^ο Μήνας | Βαθμός |
| 1 | 16 | 3 | 13 | 2 | 12 | 1 |
| 2 | 10 | 3 | 5 | 2 | 2 | 1 |
| 3 | 7 | 1 | 8 | 2 | 9 | 3 |
| 4 | 13 | 3 | 11 | 2 | 5 | 1 |
| 5 | 17 | 3 | 2 | 1 | 6 | 2 |
| 6 | 10 | 3 | 7 | 1 | 9 | 2 |
| 7 | 11 | 3 | 6 | 1 | 7 | 2 |

Πίνακας 5.6.1.2: Βαθμοί των παρατηρήσεων του Πίνακα 66.1.

Σε έναν πίνακα περνάμε τις τιμές του προβλήματος με τις εντολές

```
> data<-rbind(c(16,13,12),c(10,5,3),c(7,8,9),c(13,11,5),c(17,2,6),c(10,7,9),c(11,6,7))
> data
  [,1] [,2] [,3]
[1,] 16 13 12
[2,] 10  5  3
[3,]  7  8  9
```

```
[4,] 13 11 5
[5,] 17 2 6
[6,] 10 7 9
[7,] 11 6 7
```

Τρέχουμε τον έλεγχο του Friedman με την εντολή

```
> friedman.test(data)
```

```
Friedman rank sum test
```

```
data: data
Friedman chi-squared = 5.4286, df = 2, p-value =
0.06625
```

Βρήκαμε την τιμή του στατιστικού ίση με 5.4286 όπως πριν. Η p-τιμή του ελέγχου είναι > 0.05 επομένως δεν έχουμε σοβαρές ενδείξεις κατά της μηδενικής υπόθεσης.

5.6.2 Ο έλεγχος του McNemar.

Ο έλεγχος McNemar στην R γίνεται με την εντολή `mcnemar.test(x,correct = T)`

Η εντολή είναι απλή καθώς το μόνο που χρειάζεται ως όρισμα είναι τον πίνακα x που περιέχει τις παρατηρούμενες συχνότητες πριν και μετά την εφαρμογή κάποιας μεθόδου. Αν πραγματοποιήσουμε τον έλεγχο τα αποτελέσματα που θα πάρουμε είναι της μορφής

```
> mcnemar.test(x,correct = T)
```

```
McNemar's Chi-squared test with continuity
correction
```

```
data: x
McNemar's chi-squared = ..., df = 1, p-value =...
```

Όπου **McNemar's chi-squared** είναι η τιμή του στατιστικού που υπολογίζει ο έλεγχος, **df** είναι οι βαθμοί ελευθερίας και **p-value** είναι η p-τιμή του ελέγχου. Αν είναι

$$p - \text{τιμή} < \alpha \text{ απορρίπτουμε την μηδενική υπόθεση}$$

Εφαρμογή 5.6.2

28 υπάλληλοι πωλήσεων βαθμολογήθηκαν σχετικά με την παρουσίαση ενός προϊόντος στον υποψήφιο πελάτη. Ύστερα, αφού είδαν ένα ενημερωτικό βίντεο σχετικά με τον τρόπο παρουσίασης του κάθε προϊόντος και πως θα πρέπει να προσεγγιστεί ένας πελάτης, βαθμολογήθηκαν ξανά. Θέλουμε να ελέγξουμε αν το ενημερωτικό βίντεο βοηθάει ή όχι στην βελτίωση ενός υπαλλήλου. Τα αποτελέσματα της έρευνας παρουσιάζονται στον Πίνακα 5.6.2.

| Πριν | Μετά | | |
|--------------------|----------------|--------------------|--------|
| | Ικανοποιητικός | Όχι Ικανοποιητικός | Σύνολο |
| Ικανοποιητικός | 5 | 4 | 10 |
| Όχι Ικανοποιητικός | 13 | 6 | 19 |
| Σύνολο | 18 | 10 | 28 |

Πίνακας 5.6.2: Βελτίωση ή όχι των υπαλλήλων μετά την προβολή του επιμορφωτικού βίντεο.

Αρχικά πρέπει να περάσουμε τις συχνότητες σε έναν πίνακα.

```
> x<-rbind(c(5,4),c(13,6))
```

```
> x
```

```
  [,1] [,2]
[1,]  5   4
[2,] 13   6
```

Ύστερα μπορούμε να εκτελέσουμε τον έλεγχο με την εντολή

```
> mcnemar.test(x,correct = T)
```

```
McNemar's Chi-squared test with continuity
correction
```

```
data: x
```

```
McNemar's chi-squared = 3.7647, df = 1, p-value =
0.05235.
```

Η τιμή του στατιστικού βλέπουμε πως είναι η ίδια με αυτή που υπολογίσαμε και εμείς. Η p-τιμή του ελέγχου είναι μεγαλύτερη από το επίπεδο σημαντικότητας α και επομένως θα αποδεχτούμε την μηδενική υπόθεση όπως και πριν.

Βιβλιογραφία

A) Διεθνής Βιβλιογραφία

E.L.Lehmann and H.J.M. D'Abbrera (1974). *Nonparametrics. Statistical Methods Based on Ranks*. Holden-Day, INC. San Francisco

W.Corder and Dale.I.Foreman(2009).*Nonparametric Statistics For Non-Statisticians, A Step-by-Step Approach*. A John Wiley & Sons, INC.,Publication

Harald Cramer (1946).*Mathematical Methods of Statistics*. Princeton University Press.

Jean Dickinson Gibbons, Subhabrata Chakraborti (2003). *Nonparametric Statistical Inference*. Marcel Dekker, INC.

Legendre P. (2010). Coefficient of concordance. Pp. 164-169 in: *Encyclopedia of Research Design, Vol. 1*. N.J.Salkind, ed. SAGE Publications, Inc., Los Angeles.

B) Ελληνική Βιβλιογραφία

Δ.Φουσκάκης (2013). *Ανάλυση Δεδομένων με χρήση της R*. Εκδόσεις Τσότρας. Αθήνα.

Γ.Κοκολάκης και Δ.Φουσκάκης (2009). *Στατιστική Θεωρία & Εφαρμογές*. Εκδόσεις Συμεών. Αθήνα.

Ξεκαλάκη Ευδοκία (2001). *Μη παραμετρική στατιστική*. Αθήνα.

Γ) Ιστοσελίδες

<http://cran.r-project.org/web/packages>

<http://www.r-bloggers.com>