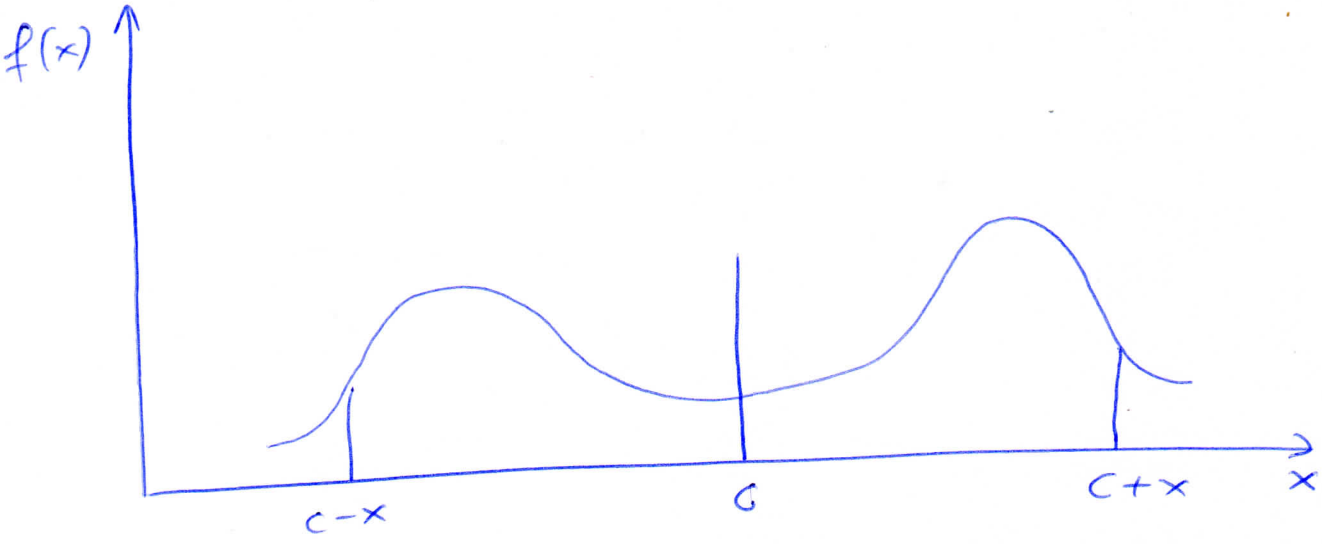
Μη-Παραμετρική Στατιστική

Φροντιστήριο #2

Ο έλεγχος των Προσημασμένων Τάξεων Μεγέθους του Wilcoxon για τη διάμεσο ενός πληθυσμού

Εισαγωγή Ο προσημικός έλεγχος για τη διάμεσο χρησιμοποιεί μόνο τα πρόσημα των διαφορών κάθε παρατήρησης και της υποτιθέμενης τιμής της διαμέσου M_0 κάτω από τη μηδενική υπόθεση H_0. Σε μία τέτοια περίπτωση τα "μεγέθη" αυτών των παρατηρήσεων σχετικά με την τιμή M_0 δεν λαμβάνονται υπόψη. Υποθέτοντας ότι αυτή η πληροφορία είναι διαθέσιμη μία κατάλληλη ελεγχοσυνάρτηση μπορεί να κατασκευαστεί που να λαμβάνει υπόψη το σχετικό μέγεθος κάθε παρατήρησης σχετικά με την τιμή M_0.

Αν μπορούμε να υποθέσουμε ότι ο πληθυσμός των διαφορών $X_i - M_0$ είναι συμμετρικός τότε ο έλεγχος Wilcoxon (Wilcoxon signed-rank test) παρέχει έναν εναλλακτικό τρόπο ελέγχου της διαμέσου λαμβάνοντας υπόψη και τα μεγέθη αλλά και τα πρόσημα των διαφορών $X_i - M_0$. Τι εννοούμε με τον όρο συμμετρική κατανομή; Η κατανομή μίας συνεχούς τ.μ. X καλείται συμμετρική γύρω από την ευθεία $x=c$, για κάποια σταθερά c αν $P(X \leq c-x) = P(X \geq c+x)$ για κάθε δυνατή τιμή x. Μία εικόνα συμμετρίας έχουμε στο ακόλουθο σχήμα.



Ο έλεγχος Wilcoxon έχει ως παραμετρικό ανάλογο τον έλεγχο t για ένα δείγμα παρατηρήσεων. Έχουμε ένα τ.δ. N παρατηρήσεων $X_1, X_2, \ldots, X_N$ από μία συνεχή σ.κ. F με διάμεσο M για την οποία υποθέτουμε ότι F είναι συμμετρική γύρω από την τιμή M.

Η μηδενική προς έλεγχο υπόθεση H_0 είναι:

$$H_0: M = M_0$$

και θεωρούμε ότι οι διαφορές $D_i = X_i - M_0$ κατανέμονται συμμετρικά γύρω από την τιμή 0.

Αφού έχουμε υποθέσει συνεχή πληθυσμό, δεν ενδιαφερόμαστε ιδιαίτερα για μηδενικές διαφορές ή για διαφορές με την ίδια απόλυτη διαφορά $|D_i|$. Αυτές μπορούμε να τις αγνοήσουμε. Διατάσσουμε αυτές τις απόλυτες διαφορές $|D_1|, |D_2|, \ldots, |D_N|$ από τη μικρότερη στη μεγαλύτερη και αποδίδουμε σε αυτές τους βαθμούς $1, 2, \ldots, N$, "κρατώντας" τα αρχικά πρόσημα για τις διαφορές D_i.

Αν M_0 είναι η πραγματική τιμή της διαμέσου, τότε περιμένουμε ότι το άθροισμα των βαθμών των θετικών διαφορών T^+ είναι ίσο με το άθροισμα των βαθμών των αρνητικών διαφορών T^-.

Όμως $T^+ + T^- = \sum_{i=1}^{N} i = \frac{N(N+1)}{2}$ και συνεπώς ελεγχοσυναρτήσεις που βασίζονται μόνο στην T^+ ή μόνο στην T^- ή στη διαφορά $T^+ - T^-$ είναι ισοδύναμες μεταξύ τους. Σε σύγκριση με το sign test, η τιμή της T^+ επηρεάζεται όχι μόνο από τον αριθμό των θετικών διαφορών αλλά και από το σχετικό μέγεθός τους σε σχέση με την τιμή M_0. Συνεπώς, σε συνδυασμό με τη συμμετρική υπόθεση, η T^+ μπορεί να προσφέρει ένα αποτελεσματικότερο test, για κάποιες κατανομές.

Η τάξη μεγέθους i σχετίζεται με ένα "+" ή "−" πρόσημο της διαφοράς $D_j = X_j - M_0$, όπου D_j είναι η i-οστή μεγαλύτερη διαφορά $|D_j|$. Αν με $r(X)$ συμβολίσουμε το βαθμό (την τάξη) μίας τ.μ. X τότε το Wilcoxon signed-rank test μπορεί να γραφεί συμβολικά ως εξής:

$$T^+ = \sum_{i=1}^{N} Z_i\, r(|D_i|), \quad T^- = \sum_{i=1}^{N} (1 - Z_i)\, r(|D_i|)$$

όπου,

$$Z_i = \begin{cases} 1, & \text{αν } D_i > 0 \\ 0, & \text{αν } D_i < 0 \end{cases}$$

Άρα, $$T^+ - T^- = 2 \sum_{i=1}^{N} r(|D_i|) - \frac{N(N+1)}{2}$$

Κάτω από την H_0, οι Z_i είναι ανεξάρτητες και ταυτοτικά κατανεμημένες τ.μ. που ακολουθούν την κατανομή Bernoulli με $P(Z_i=1) = P(Z_i=0) = 1/2$ έτσι ώστε $E(Z_i) = 1/2$, $var(Z_i) = 1/4$. Επειδή T^+ είναι ένας γραμμικός συνδυασμός των Z_i έχουμε:

$$E(T^+ | H_0) = \sum_{i=1}^{N} \frac{r(|D_i|)}{2} = \frac{N(N+1)}{4}$$

και

$$var(T^+ | H_0) = \sum_{i=1}^{N} \frac{[r(|D_i|)]^2}{4} = \frac{N(N+1)(2N+1)}{24}$$

Κατάλληλοι πίνακες κρίσιμων τιμών και επιπέδων σημαντικότητας για το Wilcoxon signed-rank test δίνονται για $N \leq 50$ (βλέπε π.χ. βιβλίο Ξεκαλάκη, Πίνακας 8). Πέραν από την ακριβή κατανομή της ελεγχοσυνάρτησης T^+ για $N \leq 50$ (exact distribution) χρησιμοποιώντας το ΚΟΘ η ασυμπτωτική κατανομή της τ.μ. T^+ είναι η κανονική κατανομή. Κάτω από την H_0, η κατανομή της τ.μ.
$$Z = \frac{4T^+ - N(N+1)}{\sqrt{2N(N+1)(2N+1)/3}}$$
προσεγγίζει την τυπική κανονική κατανομή για $N \to \infty$.

Για παράδειγμα, για την $H_1: M > M_0$ για μεγάλο N η H_0 απορρίπτεται για $Z \geq z_\alpha$. Η προσέγγιση είναι γενικά "καλή" για N τουλάχιστον ίσο με 15. Η διόρθωση συνέχειας (με 0.5) γενικά βελτιώνει τη προσέγγιση.

Το πρόβλημα των μηδενικών και των ίσων διαφορών

Στη πράξη γενικά όποιες μηδενικές διαφορές (δηλ. παρατηρησεις ίσες με M_0) αγνοούνται και το N μειώνεται αναλόγως. Για τις ίσες διαφορές $|d_i|=|d_j|$ για ένα τουλάχιστόν $i \neq j$ όπου οι παρατηρήσεις πρέπει να πάρουν τον ίδιο βαθμό χρησιμοποιείται η μέθοδος των μέσων βαθμών. Για μεγάλα δείγματα όπου ο έλεγχος βασίζεται στην τυπική κανονική κατανομή, γίνεται μία διόρθωση για τις ισοβαθμείσες παρατηρήσεις. Αν έχουμε t ισοβαθμείσες παρατηρήσεις για ένα συγκεκριμένο βαθμό τότε κάνουμε μία "διόρθωση" στη διασπορά της T^+ για να συμπεριλάβουμε αυτές τις ισοβαθμίες ως εξής:

$$var(T^+ | H_0) = \frac{N(N+1)(2N+1)}{24} - \frac{\sum t(t^2-1)}{48},$$

όπου το άθροισμα αφορά όλες τις t ισοβαθμίες.

Σχόλια: ① Το Wilcoxon είναι ένα χρήσιμο test.

② Ο πληθυσμός πρέπει να προέρχεται από συμμετρική κατανομή.

③ Εύκολο στη χρήση.

④ Διαθέσιμη και η ασυμπτωτική προσέγγιση με κανονική κατανομή και για μέτριες τιμές του N.

Επίσης, η προσέγγιση αυτή δεν επηρεάζεται από μηδενικές διαφορές ή ισοβαθμίες

(5) Αν ο πληθυσμός είναι συμμετρικός και το Sign test και το Wilcoxon μπορούν να θεωρηθούν άμεσα μη-παραμετρικά ανάλογα του t-test για έλεγχο μέσων τιμών.

Παρατήρηση: Συχνά στον έλεγχο αυτό ως σ.σ. ελέγχου μπορεί να χρησιμοποιηθεί και η συνάρτηση:

$$T = \frac{\sum_{i=1}^{n} R_i}{\sqrt{\sum_{i=1}^{n} R_i^2}} \quad \text{όπου}$$

$$R_i := \begin{cases} +R(|D_i|), & \text{αν } D_i = m - X_i > 0, \text{ m διάμεσος} \\ -R(|D_i|), & \text{αν } D_i = m - X_i < 0, \; i=1,\dots,n \end{cases}$$

Η μεταβλητή $R_i, i=1,\dots,n$ καλείται προσημασμένος βαθμός ή προσημασμένη τάξη μεγέθους (signed rank).

$S = \sum_{i=1}^{n} R_i$: άθροισμα των προσημασμένων τάξεων μεγέθους των διαφορών $|D_i| = |m - X_i|$ $i=1,\dots,n$

Οι όροι $R_1, R_2, \dots, R_n$ συνιστούν ένα τ.δ. από συμμετρικό πληθυσμό με διάμεσο μηδέν άρα η S έχει μέση τιμή $E(S)=0$, $var(S) = \sum_{i=1}^{n} R_i^2$ και η T είναι η τυποποιημένη μορφή του αθροίσματος των προσημασμένων τάξεων μεγέθους των διαφορών $|D_i|, i=1,\dots,n$

Αν δεν υπάρχουν περιπτώσεις ταύτισης τιμών στα $R_1, R_2, \ldots, R_n$ τότε για τον παρονομαστή της T, έχουμε:

$$Var(S) = \sum_{i=1}^{n} R_i^2 = \sum_{i=1}^{n} n^2 = \frac{n(n+1)(2n+1)}{6}$$

Άρα,
$$T = \frac{\sum_{i=1}^{n} R_i}{\sqrt{\frac{n(n+1)(2n+1)}{6}}}$$

Η κατανομή της T προσεγγίζεται ικανοποιητικά από την τυποποιημένη κανονική κατανομή.

Ας δούμε ένα παράδειγμα.

① Ο μέσος βαθμός ενός μαθήματος Ιστορίας σε ένα Τμήμα είναι 6.5 μονάδες. Την τελευταία χρονιά άλλαξε το σύστημα εισαγωγής των φοιτητών και ο υπεύθυνος καθηγητής επιθυμεί να ελέγξει αν διαφοροποιήθηκαν σημαντικά οι βαθμολογίες των φοιτητών σε σχέση με εκείνες των φοιτητών των προηγούμενων 15 ετών. Για το λόγο αυτό συνέλεξε ένα τ.δ. $n=15$ φοιτητών των οποίων οι βαθμοί είναι:

4 4.5 1 9.5 2.5 6 7 9 9.5 8 6.5 10 9 6 7

Τα παραπάνω δεδομένα παρέχουν σημαντικές ενδείξεις ότι η βαθμολογία των νέων φοιτητών διαφέρει σημαντικά από τη βαθμολογία των φοιτητών των προηγούμενων ετών; Υποθέτουμε ότι η κατανομή της βαθμολογίας είναι συμμετρική.

<u>Λύση</u> Για τα δεδομένα μας, σχηματίζουμε τον ακόλουθο πίνακα υπολογισμών:

x_i	$D_i = x_i - m_x$	$\lvert D_i \rvert$	$R(\lvert D_i \rvert)$	R_i	R_i^2
4	-2,5	2,5	8	-8	64
4.5	-2	2	6	-6	36
1	-5,5	5,5	14	-14	196
9.5	3	3	10,5	10,5	110,25
2.5	-4	4	13	-13	169
6	-0,5	0,5	2,5	-2,5	6,25
7	0,5	0,5	2,5	2,5	6,25
9	2,5	2,5	8	8	64
9.5	3	3	10.5	10,5	110,25
8	1,5	1,5	5	5	25
6.5	0	0	—	—	—
10	3,5	3,5	12	12	144
9	2,5	2,5	8	8	64
6	-0,5	0,5	2,5	-2,5	6,25
7	0,5	0,5	2,5	2,5	6,25
				13	1007.5

$H_0: m_x = 6.5$

$H_1: m_x \neq 6.5$

$$T = \frac{\sum R_i}{\sqrt{\sum R_i^2}} = \frac{13}{\sqrt{1007,5}} = 0,41$$

Απορρίπτω την H_0 αν $T < -z_{0.025}$ ή αν $T > z_{0.975}$
(σε ε.σ.σ. $\alpha = 5\%$)

$-z_{0.025} = -1.96$, $\quad z_{0.975} = -z_{0.025}$

Όμως, $T = 0.41$ άρα δεν απορρίπτουμε την H_0. Συνεπώς, τα δεδομένα <u>δεν</u> παρέχουν σημαντικές ενδείξεις απόρριψης της H_0 άρα η βαθμολογία των νέων φοιτητών δεν φαίνεται να διαφέρει στατιστικά σημαντικά από τη βαθμολογία των φοιτητών των προηγούμενων ετών.

Λύση με την R

Χρησιμοποιώντας την library(ggpubr) φτιάχνουμε ένα boxplot, στο οποίο φαίνεται η συμμετρία.

Με το correct=TRUE επιλέγουμε την προσέγγιση με κανονική κατανομή.

Κάνουμε τον έλεγχο με την εντολή: wilcox.test

Επισυνάπτονται και οι εντολές όπως αυτές διαμορφώνονται για τους μονόπλευρους ελέγχους.

Δίνεται και το output της R για τον αμφίπλευρο έλεγχο.

$V = 59 = T$ = άθροισμα θετικών τάξεων μεγέθους.

υψηλή τιμή p-value (ενδείξεις μη-απόρριψης H_0).

```
x=c(4,4.5,1,9.5,2.5,6,7,9,9.5,8,6.5,10,9,6,7)
summary(x)
library(ggpubr)
boxplot(x, ylab = "Bathmologies(x)", xlab = FALSE)
# One-sample wilcoxon test
res <- wilcox.test(x, mu = 6.5, correct=TRUE)
# Printing the results
res
wilcox.test(x, mu = 6.5,
              alternative = "less")
wilcox.test(x, mu = 6.5,
              alternative = "greater")
```

```
Wilcoxon signed rank test with continuity correction

data:  x
V = 59, p-value = 0.7054
alternative hypothesis: true location is not equal to 6.5
```