Μάθημα 7ο

3/4/19

Το τεστ των Wilcoxon-Mann-Whitney (W-M-W)

Έστω τα ανεξ. τ.δ. $X_{11},\dots,X_{1n_1}$ και $X_{21},\dots,X_{2n_2}$ από 2 ανεξ. πληθυσμούς X_1 και X_2 αντίστοιχα (διατάξιμη κλίμακα). Το τεστ των W-M-W μπορεί να ελέγξει αν τα δύο δείγματα προέρχονται από τον ίδιο πληθυσμό.

Ο έλεγχος ενδιαφέρει:

H_0: οι τ.μ. X_1 και X_2 έχουν την ίδια κατανομή

H_a: οι τ.μ. X_1 και X_2 έχουν διαφορετική κατανομή

διατυπώνεται ισοδύναμα ως:

$H_0: P(X_1 > X_2) = \frac{1}{2}$ ν $H_a: P(X_1 > X_2) \neq \frac{1}{2}$

$H_\alpha: P(X_1 > X_2) < \frac{1}{2}$

$H_\alpha: P(X_1 > X_2) > \frac{1}{2}$

Έστω $R(X_{ij})$, $i=1,2$, $j=1,\dots,n_i$ οι τάξεις των μετρήσεων στο σύνολο $n=n_1+n_2$ μετρήσεων και $R_i=\sum_{j=1}^{n_i} R(X_{ij})$, $i=1,2$ στο ενοποιημένο δείγμα

(σε αύξουσα τάξη μεγέθους)

Αν αληθεύει η H_0, αναμένουμε οι μέσοι όροι των τάξεων των δειγμάτων 1 και 2 να είναι "περίπου" ίσοι. Δηλ. $\frac{R_1}{n_1} = \frac{R_2}{n_2} \left(= \frac{R_1+R_2}{n_1+n_2} = \frac{\frac{n(n+1)}{2}}{n} \right.$

$\left. = \frac{n+1}{2} \right)$ // $R_1 + R_2 = \frac{n(n+1)}{2}$, R_1, R_2 δεν είναι ανεξ/τα

$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$

Πρόταση: (σελ. 66 - Μπατσίδης)

Υπό την H_0, και δεν υπάρχουν ισοτιμίες, μεταξύ των δεδομένων

(i) $R_1 \underset{\text{προσ.}}{\overset{H_0}{\sim}} N(ER_1, Var R)$:

$R_1 = \sum_{j=1}^{n_1} R(x_{1j})$: άθροισμα n_1 ισότομων τ.μ.

$\left(\begin{array}{ll} \text{Δυνατές τιμές} & 1, \dots, n \\ \text{πιθανότητες} & 1/n, \dots, 1/n \end{array} \right)$

(ii) $E(R_i) = n_i \frac{n+1}{2}$ $\quad \left[E(R_1) = n_1 \frac{n+1}{2} \right]$

$E(R_1) = \sum_{j=1}^{n_1} E\, R(x_{1j}) = 1 \cdot \frac{1}{n} + \dots + n \frac{1}{n}$

(iii) $Var(R_i) = n_i \frac{(n+1)(n-n_i)}{12}$

$\left[Var(R_1) = \frac{n_1 n_2 (n+1)}{12} \right]$

Το τεστ των W-M-W χρησιμοποιεί το στατιστικό
U_1 = αριθμός των φορών που μια X_1 παρατήρηση ακολουθεί μια X_2 παρατήρηση (= άθροισμα αριθμών των παρατηρήσεων X_1 που ακολουθούν - είναι μεγαλύτερες - κάθε παρατήρηση X_2)

π.χ. (τάξεις)

1	2	3	4	5	6	7	8	9	10	11
x_2	x_2	x_1	x_1	x_2	x_1	x_2	x_1	x_2	x_2	x_1
		2	2		3		4			6

$n_1 = 5, n_2 = 6$
$n = 11$

$U_1 = 2+2+3+4+6 = 17$ ή $U_1 = R_1 - \frac{n_1(n_1+1)}{2} = 17.$

$R_1 = 3+4+6+8+11 = 32$

και $U_2 = R_2 - \frac{n_2(n_2+1)}{2}$

και $U = \min\{U_1, U_2\}$

όπου $\left.\begin{aligned} U_1+U_2 &= \underbrace{R_1+R_2}_{\frac{n(n+1)}{2} = \frac{(n_1+n_2)(n_1+n_2+1)}{2}} - \frac{n_1(n_1+1)}{2} - \frac{n_2(n_2+1)}{2} \end{aligned}\right\} \Rightarrow$

$\Rightarrow U_1 + U_2 = n_1 n_2$

Άρα $U = \min\{U_1, n_1 n_2 - U_1\}$

Απορρ. H_0 αν $P(U \leq u) < \alpha$ (μονόπλευρο)
$P(U \leq u) < \frac{\alpha}{2}$ (δίπλευρος έλεγχος)

Προσεγγιστικά:

$Z = \frac{R_1 - \frac{n_1(n+1)}{2}}{\sqrt{\frac{n_1 n_2 (n+1)}{12}}} \overset{H_0}{\underset{\text{προσ.}}{\sim}} N(0,1)$

κρ. περιοχή
$|Z| \geq Z_{\frac{a}{2}}$
$Z \leq -Z_a$
$Z \geq Z_a$

Παράδειγμα 3 (4.1., σ.70- Μπατσίδης)

$R_1 \overset{H_0}{\sim}$; $\quad n_1=2$, $n_2=3$ ΣΑΔ. X_{11}, X_{12} και X_{21}, X_{22}, X_{23}

R_1 = αθρ. των τάξεων των τιμών της X_1.

$\binom{5}{2\ 3} = 10$ δυνατές περιπτώσεις

1	2	3	4	5	10 περιπτώσεις $(R(X_{11}), R(X_{12}))$	R_1
X_{11}	X_{12}	X_{21}	X_{22}	X_{23}	(1, 2)	3
x	y	x	y	y	(1, 3)	4
x	y	y	x	y	(1, 4)	5
y	x	x	y	y	(2, 3)	5
x	y	y	y	x	(1, 5)	6
y	x	y	x	y	(2, 4)	6
y	x	y	y	x	(2, 5)	7
y	y	x	x	y	(3, 4)	7
y	y	x	y	x	(3, 5)	8
y	y	y	x	x	(4, 9)	9

$$P(R_1 = r) = \begin{cases} 1/10 & \text{για } r = 3, 4, 8, 9 \\ 1/5 & \text{για } r = 5, 6, 7 \end{cases}$$

Παράδειγμα 1 (7.9)

X_1 : 9, 11, 15, και X_2 : 6, 8, 10, 13

$H_0 : F_{X_1}(x_1) = F_{X_2}(x_2)$ v $H_a : F_{X_1}(x_1) \neq F_{X_2}(x_2)$

	x_2	x_2	x_1	x_2	x_1	x_2	x_1	
Αναμενόμενο διατεταγμ. δείγμα :	6	8	9	10	11	13	15	$n_1 = 3$, $n_2 = 4$
Τάξεις $R(x_{ij})$:	1	2	3	4	5	6	7	$n = 7$

$U_1 = R_1 - \frac{n_1(n_1+1)}{2} = 3+5+7 - \frac{3(3+1)}{2} = 9$

$U = \min\{U_1, n_1 n_2 - U_1\} = \min\{9, 12-9\} = 3$

Επειδή $P(U \leq 3) = 0.20$

και $P = 2 \cdot 0.20 = 0.40 > 0.05$ δεν απορρ. H_0

Παράδειγμα 2

Να ελεγχθεί κατά πόσο τα δεδομένα:

X_1 : 26, 27, 25, 31

X_2 : 28, 35, 32, 29

προέρχονται από τον ίδιο πληθ.
(αριθμός βακτηριδίων ανά μονάδα όγκου για δύο τύπους καλλιεργειών X_1 κ X_2).

	X_1	X_1	X_1	X_2	X_2	X_1	X_2	X_2	
Αναμ. Διατ. Δείγμα:	25	26	27	28	29	31	32	35	$n_1=4$
$R(X_{ij})$:	1	2	3	4	5	6	7	8	$n_2=4$

$$U_1 = 1+2+3+6 - \frac{4(4+1)}{2} = 12-10 = 2$$

$$U = \min\{2, 16-2\} = 2$$

Για $n_1=4$, $n_2=4$, $P(U \leq 2) = 0.057$

$p = 2 * 0.057 > 0.05$ δεν απορ. η H_0.