Wilcoxon Rank Sum test (Mann Whitney U)

Μη Παραμετρική Στατιστική

Παναγιώτης Παπαστamoύλης
Επίκουρος Καθηγητής
Τμήμα Στατιστικής ΟΠΑ

papastamoulis@aueb.gr

Παρασκευή, 21/3/2020

Εισαγωγή

- Το τεστ που θα μελετήσουμε χρησιμοποιείται για ανίχνευση διαφορών στην κατανομή *ανεξάρτητων* παρατηρήσεων $(X_1, \ldots, X_m)$ και $(Y_1, \ldots, Y_n)$.
- Μελετήθηκε για πρώτη φορά από τον Wilcoxon (1945), για την ειδική περίπτωση $m = n$.
- Η περίπτωση $m \neq n$ μελετήθηκε από τους Mann and Whitney (1947) και από τον Wilcoxon (1949).
- Για τους λόγους αυτούς, ο έλεγχος αναφέρεται συχνά και ως Wilcoxon-Mann-Whitney.

Εισαγωγή

- Το τεστ που θα μελετήσουμε χρησιμοποιείται για ανίχνευση διαφορών στην κατανομή *ανεξάρτητων* παρατηρήσεων $(X_1, \ldots, X_m)$ και $(Y_1, \ldots, Y_n)$.
- Μελετήθηκε για πρώτη φορά από τον Wilcoxon (1945), για την ειδική περίπτωση $m = n$.
- Η περίπτωση $m \neq n$ μελετήθηκε από τους Mann and Whitney (1947) και από τον Wilcoxon (1949).
- Για τους λόγους αυτούς, ο έλεγχος αναφέρεται συχνά και ως Wilcoxon-Mann-Whitney.
- Η συνήθης περίπτωση είναι αυτή όπου ο ερευνητής έχει ανεξάρτητα δείγματα από δύο πληθυσμούς και ενδιαφέρεται να ελέγξει αν οι πληθυσμοί ταυτίζονται, πχ
  - τείνουν οι τιμές ενός πληθυσμού να είναι μεγαλύτερες;
  - έχουν οι δύο πληθυσμοί ίσες διαμέσους;
  - έχουν οι δύο πληθυσμοί ίσες μέσες τιμές;

Wilcoxon Rank Sum test (Mann Whitney U)

- Έστω $X = (X_1, \ldots, X_m)$ ανεξάρτητες και ισόνομες τ.μ. με συνάρτηση κατανομής F.
- Έστω $Y = (Y_1, \ldots, Y_n)$ ανεξάρτητες και ισόνομες τ.μ. με συνάρτηση κατανομής G.
- Υποθέτουμε οτι X_i και Y_j ανεξάρτητες για κάθε i, j.
- Ενδιαφέρει ο έλεγχος υποθέσεων για την ανιχνευση διαφορών μεταξύ F και G:

  Π_1 $H_0 : F(t) = G(t), \forall t$ έναντι $H_1 : F(t) \neq G(t)$ για τουλάχιστον ένα t
  Π_2 $H_0 : F(t) \leqslant G(t), \forall t$ έναντι $H_1 : F(t) > G(t)$ για τουλάχιστον ένα t
  Π_3 $H_0 : F(t) \geqslant G(t), \forall t$ έναντι $H_1 : F(t) < G(t)$ για τουλάχιστον ένα t

Wilcoxon Rank Sum test (Mann Whitney U)

Οι υποθέσεις αυτές είναι ισοδύναμες με

- Π_1 $H_0 : \mathrm{P}(X_1 < Y_1) = 1/2$ έναντι $H_1 : \mathrm{P}(X_1 < Y_1) \neq 1/2$
- Π_2 $H_0 : \mathrm{P}(X_1 < Y_1) \leqslant 1/2$ έναντι $H_1 : \mathrm{P}(X_1 < Y_1) > 1/2$
- Π_3 $H_0 : \mathrm{P}(X_1 < Y_1) \geqslant 1/2$ έναντι $H_1 : \mathrm{P}(X_1 < Y_1) < 1/2$

Ένα μοντέλο που περιγράφει αυτές τις υποθέσεις είναι όταν η F προέρχεται από μετατόπιση θέσης της G:

$$\exists \Delta : G(t) = F(t - \Delta) \quad \forall t \in \mathbb{R}.$$

Αυτό ισοδυναμεί με το ότι η X έχει την ίδια κατανομή με την $Y + \Delta$.

Στατιστική συνάρτηση αθροίσματος τάξεων (Wilcoxon Rank Sum)

$$W_{m,n} = \sum_{j=1}^{n} R_{N,j}$$

όπου

- $N = m + n$
- $R_{N,j} = \sum_{k=1}^{n} \mathrm{I}_{\{Y_k \leqslant Y_j\}} + \sum_{i=1}^{m} \mathrm{I}_{\{X_i < Y_j\}}$
- Με άλλα λόγια: το $R_{N,j}$ είναι η τάξη του Y_j στο ενιαίο δείγμα $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$, $j = 1, \ldots, n$.
- Περιπτώσεις ισοβαθμίας αντιμετωπίζονται κατά τον συνήθη τρόπο.
- Η ακριβής κατανομή της $W_{m,n}$ μπορεί να υπολογιστεί αναδρομικά (πχ: πίνακας 9, σημειώσεις Ξεκαλάκη, είτε μέσω της `pwilcox()` στην `R`)
- Για μεγάλα δείγματα χρησιμοποιούμε προσέγγιση από την κανονική κατανομή.

Παράδειγμα υπολογισμού της $W_{m,n}$

- $x = (2, -1, 1)$ και $y = (3, -2, 0, 2)$
- $m = 3$ και $n = 4$
- $N = 3 + 4 = 7$
- Δημιουργώ το ενιαίο δείγμα $(2, -1, 1, 3, -2, 0, 2)$
- Διατάσσω το ενιαίο δείγμα $(-2, -1, 0, 1, 2, 2, 3)$
- Υπολογίζω τις τάξεις $(1, 2, 3, 4, 5.5, 5.5, 7)$
- $R_7 = (1, 3, 5.5, 7)$
- $W_{3,4} = \sum_{j=1}^{4} R_{7,j} = 1 + 3 + 5.5 + 7 = 16.5$

Διαίσθηση πίσω από την $W_{m,n}$

- Αν οι «κόκκινες» τιμές είναι πολύ μικρότερες σε σχέση με τις «μπλε» τιμές, τότε περιμένουμε οι τάξεις των Y στο ενιαίο δείγμα να είναι στριμωγμένες στα δεξιά

Διαίσθηση πίσω από την $W_{m,n}$

- Αν οι «κόκκινες» τιμές είναι πολύ μικρότερες σε σχέση με τις «μπλε» τιμές, τότε περιμένουμε οι τάξεις των Y στο ενιαίο δείγμα να είναι στριμωγμένες στα δεξιά
  - $R_N = \{1, 2, \ldots, m, m+1, m+2, \ldots, N\}$
  - Αυτό σημαίνει ότι το $W_{m,n}$ θα λάβει «μεγάλες» τιμές

Διαίσθηση πίσω από την $W_{m,n}$

- Αν οι «κόκκινες» τιμές είναι πολύ μικρότερες σε σχέση με τις «μπλε» τιμές, τότε περιμένουμε οι τάξεις των Y στο ενιαίο δείγμα να είναι στριμωγμένες στα δεξιά
  - $R_N = \{1, 2, \ldots, m, m+1, m+2, \ldots, N\}$
  - Αυτό σημαίνει ότι το $W_{m,n}$ θα λάβει «μεγάλες» τιμές
- Αν οι «κόκκινες» τιμές είναι πολύ μεγαλύτερες σε σχέση με τις «μπλε» τιμές, τότε περιμένουμε οι τάξεις των Y στο ενιαίο δείγμα να είναι στριμωγμένες στα αριστερά

Διαίσθηση πίσω από την $W_{m,n}$

- Αν οι «κόκκινες» τιμές είναι πολύ μικρότερες σε σχέση με τις «μπλε» τιμές, τότε περιμένουμε οι τάξεις των Y στο ενιαίο δείγμα να είναι στριμωγμένες στα δεξιά
  - $R_N = \{1, 2, \ldots, m, m+1, m+2, \ldots, N\}$
  - Αυτό σημαίνει ότι το $W_{m,n}$ θα λάβει «μεγάλες» τιμές
- Αν οι «κόκκινες» τιμές είναι πολύ μεγαλύτερες σε σχέση με τις «μπλε» τιμές, τότε περιμένουμε οι τάξεις των Y στο ενιαίο δείγμα να είναι στριμωγμένες στα αριστερά
  - $R_N = \{1, 2, \ldots, n, n+1, n+2, \ldots, N\}$
  - Αυτό σημαίνει ότι το $W_{m,n}$ θα λάβει «μικρές» τιμές

Διαίσθηση πίσω από την $W_{m,n}$

- Αν οι «κόκκινες» τιμές είναι πολύ μικρότερες σε σχέση με τις «μπλε» τιμές, τότε περιμένουμε οι τάξεις των Y στο ενιαίο δείγμα να είναι στριμωγμένες στα δεξιά
  - $R_N = \{1, 2, \ldots, m, m+1, m+2, \ldots, N\}$
  - Αυτό σημαίνει ότι το $W_{m,n}$ θα λάβει «μεγάλες» τιμές
- Αν οι «κόκκινες» τιμές είναι πολύ μεγαλύτερες σε σχέση με τις «μπλε» τιμές, τότε περιμένουμε οι τάξεις των Y στο ενιαίο δείγμα να είναι στριμωγμένες στα αριστερά
  - $R_N = \{1, 2, \ldots, n, n+1, n+2, \ldots, N\}$
  - Αυτό σημαίνει ότι το $W_{m,n}$ θα λάβει «μικρές» τιμές
- Αν οι «κόκκινες» τιμές δεν διαφέρουν «πολύ» από τις «μπλε» τιμές, τότε περιμένουμε οι τάξεις των Y στο ενιαίο (διατεταγμένο) δείγμα να εναλλάσσονται τυχαία με αυτές των X.

Διαίσθηση πίσω από την $W_{m,n}$

- Αν οι «κόκκινες» τιμές είναι πολύ μικρότερες σε σχέση με τις «μπλε» τιμές, τότε περιμένουμε οι τάξεις των Y στο ενιαίο δείγμα να είναι στριμωγμένες στα δεξιά
    - $R_N = \{1, 2, \ldots, m, m+1, m+2, \ldots, N\}$
    - Αυτό σημαίνει ότι το $W_{m,n}$ θα λάβει «μεγάλες» τιμές
- Αν οι «κόκκινες» τιμές είναι πολύ μεγαλύτερες σε σχέση με τις «μπλε» τιμές, τότε περιμένουμε οι τάξεις των Y στο ενιαίο δείγμα να είναι στριμωγμένες στα αριστερά
    - $R_N = \{1, 2, \ldots, n, n+1, n+2, \ldots, N\}$
    - Αυτό σημαίνει ότι το $W_{m,n}$ θα λάβει «μικρές» τιμές
- Αν οι «κόκκινες» τιμές δεν διαφέρουν «πολύ» από τις «μπλε» τιμές, τότε περιμένουμε οι τάξεις των Y στο ενιαίο (διατεταγμένο) δείγμα να εναλλάσσονται τυχαία με αυτές των X.
    - $R_N = \{1, 2, 3, 4, 5, \ldots, m+1, m+2, m+3, \ldots, N\}$
    - Αυτό σημαίνει ότι το $W_{m,n}$ θα λάβει τιμές που θα «μοιάζουν» με τυχαία δειγματοληψία n τιμών (χωρίς επανάθεση) από το σύνολο $\{1, 2, \ldots, N\}$

Σχέση με Mann-Whitney U

- Ισχύει ότι: $R_{N,j} = \sum_{k=1}^{n} \mathrm{I}_{\{Y_k \leqslant Y_j\}} + \sum_{i=1}^{m} \mathrm{I}_{\{X_i < Y_j\}}$
- Συνεπώς

$$\begin{aligned} W_{m,n} &= \sum_{j=1}^{n}\sum_{k=1}^{n} \mathrm{I}_{\{Y_k \leqslant Y_j\}} + \sum_{j=1}^{n}\sum_{i=1}^{m} \mathrm{I}_{\{X_i < Y_j\}} \\ &= \frac{n(n+1)}{2} + U_{m,n} \end{aligned}$$

  όπου $U_{m,n} = \sum_{j=1}^{n}\sum_{i=1}^{m} \mathrm{I}_{\{X_i < Y_j\}}$
- Η στατιστική συνάρτηση $U_{m,n}$ ονομάζεται **Mann-Whitney U**.
- Οι $W_{m,n}$ και $U_{m,n}$ είναι ισοδύναμες όσον αφορά τη συμπερασματαλογία
- Θα χρησιμοποιούμε την $W_{m,n}$

Παράδειγμα υπολογισμού του $U_{m,n}$

- $x = (2, -1, 1)$ και $y = (3, -2, 0, 2)$
- $m = 3$ και $n = 4$
- $N = 3 + 4 = 7$
- Δημιουργώ το ενιαίο δείγμα $(2, -1, 1, 3, -2, 0, 2)$
- Διατάσσω το ενιαίο δείγμα $(-2, -1, 0, 1, 2, 2, 3)$
- Υπολογίζω τις τάξεις $(1, 2, 3, 4, 5.5, 5.5, 7)$
- $R_7 = (1, 3, 5.5, 7)$
- $W_{3,4} = \sum_{j=1}^{4} R_{7,j} = 1 + 3 + 5.5 + 7 = 16.5$
- $U_{3,4} = W_{3,4} - 4 \times 5/2 = 6.5$

Προσεγγιστική κατανομή για $W_{m,n}$

Πρόταση : Υπό την H_0 ισχύει ότι

$$\mathrm{E}W_{m,n} = \frac{n(N+1)}{2} \tag{1}$$

$$\mathrm{Var}W_{m,n} = \frac{n(N+1)(N-n)}{12}. \tag{2}$$

- Απόδειξη : επόμενη σελίδα.
- Κανονική προσέγγιση

$$\begin{aligned} Z &= \frac{W_{m,n} - \mathrm{E}W_{m,n}}{\sqrt{\mathrm{Var}W_{m,n}}} \\ &= \frac{W_{m,n} - n(N+1)/2}{\sqrt{n(N+1)(N-n)/12}} \overset{m,n\to\infty}{\to} \mathcal{N}(0,1) \end{aligned}$$

Απόδειξη των (1) και (2)

- Το πρόβλημα είναι ισοδύναμο με την εύρεση της μέσης τιμής και διασποράς της τ.μ. που ισούται με το άθροισμα n ακεραίων που έχουν επιλεγεί τυχαία και χωρίς επανάθεση από το σύνολο $\{1, 2, \ldots, N\}$

Απόδειξη των (1) και (2)

- Το πρόβλημα είναι ισοδύναμο με την εύρεση της μέσης τιμής και διασποράς της τ.μ. που ισούται με το άθροισμα n ακεραίων που έχουν επιλεγεί τυχαία και χωρίς επανάθεση από το σύνολο $\{1, 2, \ldots, N\}$
- Έστω $Z_1, Z_2, \ldots, Z_n$ τ.μ που προκύπτουν τυχαία και χωρίς επανάθεση από το σύνολο $\{1, 2, \ldots, N\}$.

Απόδειξη των (1) και (2)

- Το πρόβλημα είναι ισοδύναμο με την εύρεση της μέσης τιμής και διασποράς της τ.μ. που ισούται με το άθροισμα n ακεραίων που έχουν επιλεγεί τυχαία και χωρίς επανάθεση από το σύνολο $\{1, 2, \ldots, N\}$
- Έστω $Z_1, Z_2, \ldots, Z_n$ τ.μ που προκύπτουν τυχαία και χωρίς επανάθεση από το σύνολο $\{1, 2, \ldots, N\}$.
- Προφανώς $W_{m,n} \overset{H_0}{=} \sum_{j=1}^{n} Z_j$

Απόδειξη των (1) και (2)

- Το πρόβλημα είναι ισοδύναμο με την εύρεση της μέσης τιμής και διασποράς της τ.μ. που ισούται με το άθροισμα n ακεραίων που έχουν επιλεγεί τυχαία και χωρίς επανάθεση από το σύνολο $\{1, 2, \ldots, N\}$
- Έστω $Z_1, Z_2, \ldots, Z_n$ τ.μ που προκύπτουν τυχαία και χωρίς επανάθεση από το σύνολο $\{1, 2, \ldots, N\}$.
- Προφανώς $W_{m,n} \overset{H_0}{=} \sum_{j=1}^{n} Z_j$
- Συνεπώς

$$\begin{aligned} \mathrm{E}W_{m,n} &= \mathrm{E}\left(\sum_{j=1}^{n} Z_j\right) = \sum_{j=1}^{n} \mathrm{E}Z_j = n\mathrm{E}Z_1 \\ &= n\sum_{j=1}^{N} j\,\frac{1}{N} = \frac{n}{N}\,\frac{N(N+1)}{2} = \frac{n(N+1)}{2} \end{aligned}$$

και η απόδειξη της (1) ολοκληρώθηκε.

Απόδειξη των (1) και (2)

Για τη διασπορά έχουμε

$$\begin{aligned}\mathrm{Var}W_{m,n} &= \mathrm{Var}\left(\sum_{j=1}^{n} Z_j\right) \\ &= \sum_{j=1}^{n}\mathrm{Var}Z_j + 2\sum_{j=1}^{n}\sum_{i<j}\mathrm{Cov}(Z_i, Z_j) \qquad (3)\end{aligned}$$

Απόδειξη των (1) και (2)

Για τη διασπορά έχουμε

$$\begin{aligned}\mathrm{Var}W_{m,n} &= \mathrm{Var}\left(\sum_{j=1}^{n} Z_j\right) \\ &= \sum_{j=1}^{n}\mathrm{Var}Z_j + 2\sum_{j=1}^{n}\sum_{i<j}\mathrm{Cov}(Z_i, Z_j) \qquad (3)\end{aligned}$$

- Υπολογισμός $\mathrm{Var}Z_j$

$$\begin{aligned}\mathrm{Var}Z_j &= \mathrm{E}(Z_j^2) - (\mathrm{E}Z_j)^2 = \sum_{j=1}^{N} j^2\frac{1}{N} - \left(\frac{N+1}{2}\right)^2 \\ &= \frac{1}{N}\frac{N(N+1)(2N+1)}{6} - \left(\frac{N+1}{2}\right)^2 = \dots \\ &= \frac{N^2-1}{12} \qquad (4)\end{aligned}$$

Απόδειξη των (1) και (2)

Για τη συνδιασπορά $\mathrm{Cov}(Z_i, Z_j)$ έχουμε ($i \neq j$):

$$\begin{aligned}
\mathrm{Cov}(Z_i, Z_j) &= \mathrm{E}(Z_i Z_j) - (\mathrm{E}Z_i)(\mathrm{E}Z_j) \\
&= \sum_{i=1}^{N} \sum_{j \neq i} ij \frac{1}{N} \frac{1}{N-1} - \left(\frac{N+1}{2}\right)^2 \\
&= \frac{1}{N(N-1)} \left(\sum_{i=1}^{N} \sum_{j=1}^{N} ij - \sum_{i=1}^{N} i^2 \right) - \frac{(N+1)^2}{4} \\
&= \frac{1}{N(N-1)} \left(\sum_{i=1}^{N} i \sum_{j=1}^{N} j - \sum_{i=1}^{N} i^2 \right) - \frac{(N+1)^2}{4} = \ldots \\
&= -\frac{N+1}{12} \qquad (5)
\end{aligned}$$

Αντικαθιστώντας τις (4) και (5) στην (3) προκύπτει η (2), και η αποδειξη ολοκληρώθηκε.

Παράδειγμα υπολογισμού του Z

- $x = (2, -1, 1)$ και $y = (3, -2, 0, 2)$
- $m = 3$ και $n = 4$
- $N = 3 + 4 = 7$
- Δημιουργώ το ενιαίο δείγμα $(2, -1, 1, 3, -2, 0, 2)$
- Διατάσσω το ενιαίο δείγμα $(-2, -1, 0, 1, 2, 2, 3)$
- Υπολογίζω τις τάξεις $(1, 2, 3, 4, 5.5, 5.5, 7)$
- $R_7 = (1, 3, 5.5, 7)$
- $W_{3,4} = \sum_{j=1}^{4} R_{7,j} = 1 + 3 + 5.5 + 7 = 16.5$
- $U_{3,4} = W_{3,4} - 4 \times 5/2 = 6.5$

Παράδειγμα υπολογισμού του Z

- $x = (2, -1, 1)$ και $y = (3, -2, 0, 2)$
- $m = 3$ και $n = 4$
- $N = 3 + 4 = 7$
- Δημιουργώ το ενιαίο δείγμα $(2, -1, 1, 3, -2, 0, 2)$
- Διατάσσω το ενιαίο δείγμα $(-2, -1, 0, 1, 2, 2, 3)$
- Υπολογίζω τις τάξεις $(1, 2, 3, 4, 5.5, 5.5, 7)$
- $R_7 = (1, 3, 5.5, 7)$
- $W_{3,4} = \sum_{j=1}^{4} R_{7,j} = 1 + 3 + 5.5 + 7 = 16.5$
- $U_{3,4} = W_{3,4} - 4 \times 5/2 = 6.5$
- Υπολογισμός Z:
  - $\mathrm{E}W_{3,4} = 4(7+1)/2 = 16$
  - $\mathrm{Var}W_{3,4} = 4(7+1)(7-4)/12 = 8$
  - $Z = \frac{W_{m,n} - \mathrm{E}W_{m,n}}{\sqrt{\mathrm{Var}W_{m,n}}} = \frac{16.5-16}{\sqrt{8}} \approx 0.177$

(Προσεγγιστική) συμπερασματολογία για τις υποθέσεις

Π_1 $H_0 : F(t) = G(t), \forall t$ έναντι $H_1 : F(t) \neq G(t)$ για τουλάχιστον ένα t

Π_2 $H_0 : F(t) \leqslant G(t), \forall t$ έναντι $H_1 : F(t) > G(t)$ για τουλάχιστον ένα t

Π_3 $H_0 : F(t) \geqslant G(t), \forall t$ έναντι $H_1 : F(t) < G(t)$ για τουλάχιστον ένα t

όπου

- z_α το άνω α ποσοστιαίο σημείο της $\mathcal{N}(0,1)$, $0 < \alpha < 1$
- z η παρατηρηθείσα τιμή της Z στο δείγμα
- $\Phi(\cdot)$ η συνάρτηση κατανομής της $\mathcal{N}(0,1)$

(Προσεγγιστική) συμπερασματολογία για τις υποθέσεις

Π_1 $H_0 : F(t) = G(t), \forall t$ έναντι $H_1 : F(t) \neq G(t)$ για τουλάχιστον ένα t

- Περιοχή απόρριψης σε επίπεδο σημαντικότητας α: $|Z| > z_{\alpha/2}$
- p-value: $2\mathrm{P}(Z > |z|) = 2(1 - \Phi(|z|))$

Π_2 $H_0 : F(t) \leqslant G(t), \forall t$ έναντι $H_1 : F(t) > G(t)$ για τουλάχιστον ένα t

Π_3 $H_0 : F(t) \geqslant G(t), \forall t$ έναντι $H_1 : F(t) < G(t)$ για τουλάχιστον ένα t

όπου

- z_α το άνω α ποσοστιαίο σημείο της $\mathcal{N}(0,1)$, $0 < \alpha < 1$
- z η παρατηρηθείσα τιμή της Z στο δείγμα
- $\Phi(\cdot)$ η συνάρτηση κατανομής της $\mathcal{N}(0,1)$

(Προσεγγιστική) συμπερασματολογία για τις υποθέσεις

Π_1 $H_0 : F(t) = G(t), \forall t$ έναντι $H_1 : F(t) \neq G(t)$ για τουλάχιστον ένα t

- Περιοχή απόρριψης σε επίπεδο σημαντικότητας α: $|Z| > z_{\alpha/2}$
- p-value: $2\mathrm{P}(Z > |z|) = 2(1 - \Phi(|z|))$

Π_2 $H_0 : F(t) \leqslant G(t), \forall t$ έναντι $H_1 : F(t) > G(t)$ για τουλάχιστον ένα t

- Περιοχή απόρριψης σε επίπεδο σημαντικότητας α: $Z > z_\alpha$
- p-value: $\mathrm{P}(Z > z) = 1 - \Phi(z)$

Π_3 $H_0 : F(t) \geqslant G(t), \forall t$ έναντι $H_1 : F(t) < G(t)$ για τουλάχιστον ένα t

όπου

- z_α το άνω α ποσοστιαίο σημείο της $\mathcal{N}(0,1)$, $0 < \alpha < 1$
- z η παρατηρηθείσα τιμή της Z στο δείγμα
- $\Phi(\cdot)$ η συνάρτηση κατανομής της $\mathcal{N}(0,1)$

(Προσεγγιστική) συμπερασματολογία για τις υποθέσεις

Π_1 $H_0 : F(t) = G(t), \forall t$ έναντι $H_1 : F(t) \neq G(t)$ για τουλάχιστον ένα t

- Περιοχή απόρριψης σε επίπεδο σημαντικότητας α: $|Z| > z_{\alpha/2}$
- p-value: $2\mathrm{P}(Z > |z|) = 2(1 - \Phi(|z|))$

Π_2 $H_0 : F(t) \leqslant G(t), \forall t$ έναντι $H_1 : F(t) > G(t)$ για τουλάχιστον ένα t

- Περιοχή απόρριψης σε επίπεδο σημαντικότητας α: $Z > z_\alpha$
- p-value: $\mathrm{P}(Z > z) = 1 - \Phi(z)$

Π_3 $H_0 : F(t) \geqslant G(t), \forall t$ έναντι $H_1 : F(t) < G(t)$ για τουλάχιστον ένα t

- Περιοχή απόρριψης σε επίπεδο σημαντικότητας α: $Z < -z_\alpha$
- p-value: $\mathrm{P}(Z < z) = \Phi(z)$

όπου

- z_α το άνω α ποσοστιαίο σημείο της $\mathcal{N}(0,1)$, $0 < \alpha < 1$
- z η παρατηρηθείσα τιμή της Z στο δείγμα
- $\Phi(\cdot)$ η συνάρτηση κατανομής της $\mathcal{N}(0,1)$

Εφαρμογή

Οι βαθμοί στο μάθημα Στατιστικής σε δύο διαφορετικά Τμήματα είναι:

- τμήμα 1: $(7, 8, 6, 9, 4, 2, 6, 5, 8, 9, 10, 4, 5, 7, 4)$
- τμήμα 2: $(6, 5, 3, 7, 9, 8, 6, 9, 10, 5, 2, 3, 7, 8, 9, 7, 4)$.

Να ελεγχθεί σε επίπεδο σημαντικότητας $\alpha = 5\%$ αν οι βαθμοί στο Τμήμα 2 τείνουν να είναι μεγαλύτεροι από αυτούς των φοιτητών στο Τμήμα 1.

Εφαρμογή

- Έστω $X = (X_1, \ldots, X_m)$ ανεξάρτητες και ισόνομες τ.μ. που περιγράφουν την βαθμολογία των φοιτητών στο τμήμα 1 ($m = 15$).

Εφαρμογή

- Έστω $X = (X_1, \ldots, X_m)$ ανεξάρτητες και ισόνομες τ.μ. που περιγράφουν την βαθμολογία των φοιτητών στο τμήμα 1 $(m = 15)$.
- Έστω $Y = (Y_1, \ldots, Y_n)$ ανεξάρτητες και ισόνομες τ.μ. που περιγράφουν την βαθμολογία των φοιτητών στο τμήμα 2 $(n = 17)$.

Εφαρμογή

- Έστω $X = (X_1, \ldots, X_m)$ ανεξάρτητες και ισόνομες τ.μ. που περιγράφουν την βαθμολογία των φοιτητών στο τμήμα 1 $(m = 15)$.
- Έστω $Y = (Y_1, \ldots, Y_n)$ ανεξάρτητες και ισόνομες τ.μ. που περιγράφουν την βαθμολογία των φοιτητών στο τμήμα 2 $(n = 17)$.
- Υποθέτουμε οτι X_i και Y_j ανεξάρτητες για κάθε i, j.

Εφαρμογή

- Έστω $X = (X_1, \ldots, X_m)$ ανεξάρτητες και ισόνομες τ.μ. που περιγράφουν την βαθμολογία των φοιτητών στο τμήμα 1 $(m = 15)$.
- Έστω $Y = (Y_1, \ldots, Y_n)$ ανεξάρτητες και ισόνομες τ.μ. που περιγράφουν την βαθμολογία των φοιτητών στο τμήμα 2 $(n = 17)$.
- Υποθέτουμε οτι X_i και Y_j ανεξάρτητες για κάθε i, j.
- Έστω $F(t)$ η συνάρτηση κατανομής του X_i, $i = 1, \ldots, m$.

Εφαρμογή

- Έστω $X = (X_1, \ldots, X_m)$ ανεξάρτητες και ισόνομες τ.μ. που περιγράφουν την βαθμολογία των φοιτητών στο τμήμα 1 $(m = 15)$.
- Έστω $Y = (Y_1, \ldots, Y_n)$ ανεξάρτητες και ισόνομες τ.μ. που περιγράφουν την βαθμολογία των φοιτητών στο τμήμα **2** $(n = 17)$.
- Υποθέτουμε οτι X_i και Y_j ανεξάρτητες για κάθε i, j.
- Έστω $F(t)$ η συνάρτηση κατανομής του X_i, $i = 1, \ldots, m$.
- Έστω $G(t)$ η συνάρτηση κατανομής του Y_j, $j = 1, \ldots, n$.

Εφαρμογή

- Έστω $X = (X_1, \dots, X_m)$ ανεξάρτητες και ισόνομες τ.μ. που περιγράφουν την βαθμολογία των φοιτητών στο τμήμα 1 ($m = 15$).
- Έστω $Y = (Y_1, \dots, Y_n)$ ανεξάρτητες και ισόνομες τ.μ. που περιγράφουν την βαθμολογία των φοιτητών στο τμήμα 2 ($n = 17$).
- Υποθέτουμε οτι X_i και Y_j ανεξάρτητες για κάθε i, j.
- Έστω $F(t)$ η συνάρτηση κατανομής του X_i, $i = 1, \dots, m$.
- Έστω $G(t)$ η συνάρτηση κατανομής του Y_j, $j = 1, \dots, n$.
- $H_0 : F(t) \leqslant G(t), \forall t$ έναντι της $H_1 : F(t) > G(t)$, για τουλάχιστον ένα t (πρόβλημα Π_2)

Εφαρμογή

- Έστω $X = (X_1, \ldots, X_m)$ ανεξάρτητες και ισόνομες τ.μ. που περιγράφουν την βαθμολογία των φοιτητών στο τμήμα 1 $(m = 15)$.
- Έστω $Y = (Y_1, \ldots, Y_n)$ ανεξάρτητες και ισόνομες τ.μ. που περιγράφουν την βαθμολογία των φοιτητών στο τμήμα 2 $(n = 17)$.
- Υποθέτουμε οτι X_i και Y_j ανεξάρτητες για κάθε i, j.
- Έστω $F(t)$ η συνάρτηση κατανομής του X_i, $i = 1, \ldots, m$.
- Έστω $G(t)$ η συνάρτηση κατανομής του Y_j, $j = 1, \ldots, n$.
- $H_0 : F(t) \leqslant G(t), \forall t$ έναντι της $H_1 : F(t) > G(t)$, για τουλάχιστον ένα t (πρόβλημα Π_2)
- Ισοδυναμεί με $H_0 : \mathrm{P}(X_1 < Y_1) \leqslant 1/2$ έναντι της $H_1 : \mathrm{P}(X_1 < Y_1) > 1/2$

Εφαρμογή

- Έστω $X = (X_1, \ldots, X_m)$ ανεξάρτητες και ισόνομες τ.μ. που περιγράφουν την βαθμολογία των φοιτητών στο τμήμα 1 $(m = 15)$.
- Έστω $Y = (Y_1, \ldots, Y_n)$ ανεξάρτητες και ισόνομες τ.μ. που περιγράφουν την βαθμολογία των φοιτητών στο τμήμα 2 $(n = 17)$.
- Υποθέτουμε οτι X_i και Y_j ανεξάρτητες για κάθε i, j.
- Έστω $F(t)$ η συνάρτηση κατανομής του X_i, $i = 1, \ldots, m$.
- Έστω $G(t)$ η συνάρτηση κατανομής του Y_j, $j = 1, \ldots, n$.
- $H_0 : F(t) \leqslant G(t), \forall t$ έναντι της $H_1 : F(t) > G(t)$, για τουλάχιστον ένα t (πρόβλημα Π_2)
- Ισοδυναμεί με $H_0 : \mathrm{P}(X_1 < Y_1) \leqslant 1/2$ έναντι της $H_1 : \mathrm{P}(X_1 < Y_1) > 1/2$
- Η H_0 σημαίνει ότι οι βαθμολογίες στο τμήμα 1 (X_i) είναι στοχαστικά μεγαλύτερες-ίσες (*stochastically dominant*) από τις βαθμολογίες στο τμήμα 2 (Y_j)

Εφαρμογή

	βαθμός	τμήμα	τάξη	$R_{N,i}$
1	2.00	1.00	1.50	—
2	2.00	2.00	1.50	1.50
3	3.00	2.00	3.50	3.50
4	3.00	2.00	3.50	3.50
5	4.00	1.00	6.50	—
6	4.00	1.00	6.50	—
7	4.00	1.00	6.50	—
8	4.00	2.00	6.50	6.50
9	5.00	1.00	10.50	—
10	5.00	1.00	10.50	—
11	5.00	2.00	10.50	10.50
12	5.00	2.00	10.50	10.50
13	6.00	1.00	14.50	—
14	6.00	1.00	14.50	—
15	6.00	2.00	14.50	14.50
16	6.00	2.00	14.50	14.50
17	7.00	1.00	19.00	—
18	7.00	1.00	19.00	—
19	7.00	2.00	19.00	19.00
20	7.00	2.00	19.00	19.00
21	7.00	2.00	19.00	19.00
22	8.00	1.00	23.50	—
23	8.00	1.00	23.50	—
24	8.00	2.00	23.50	23.50
25	8.00	2.00	23.50	23.50
26	9.00	1.00	28.00	—
27	9.00	1.00	28.00	—
28	9.00	2.00	28.00	28.00
29	9.00	2.00	28.00	28.00
30	9.00	2.00	28.00	28.00
31	10.00	1.00	31.50	—
32	10.00	2.00	31.50	31.50

Εφαρμογή

	βαθμός	τμήμα	τάξη	$R_{N,i}$
1	2.00	1.00	1.50	–
2	2.00	2.00	1.50	1.50
3	3.00	2.00	3.50	3.50
4	3.00	2.00	3.50	3.50
5	4.00	1.00	6.50	–
6	4.00	1.00	6.50	–
7	4.00	1.00	6.50	–
8	4.00	2.00	6.50	6.50
9	5.00	1.00	10.50	–
10	5.00	1.00	10.50	–
11	5.00	2.00	10.50	10.50
12	5.00	2.00	10.50	10.50
13	6.00	1.00	14.50	–
14	6.00	1.00	14.50	–
15	6.00	2.00	14.50	14.50
16	6.00	2.00	14.50	14.50
17	7.00	1.00	19.00	–
18	7.00	1.00	19.00	–
19	7.00	2.00	19.00	19.00
20	7.00	2.00	19.00	19.00
21	7.00	2.00	19.00	19.00
22	8.00	1.00	23.50	–
23	8.00	1.00	23.50	–
24	8.00	2.00	23.50	23.50
25	8.00	2.00	23.50	23.50
26	9.00	1.00	28.00	–
27	9.00	1.00	28.00	–
28	9.00	2.00	28.00	28.00
29	9.00	2.00	28.00	28.00
30	9.00	2.00	28.00	28.00
31	10.00	1.00	31.50	–
32	10.00	2.00	31.50	31.50

- Wilcoxon rank sum:

$$
\begin{aligned}
W_{m,n} &= \sum_{i=1}^{n} R_{N,i} \\
&= 1.50 + 3.50 + \ldots + 31.5 = 284.5
\end{aligned}
$$

Εφαρμογή

	βαθμός	τμήμα	τάξη	$R_{N,i}$
1	2.00	1.00	1.50	–
2	2.00	2.00	1.50	1.50
3	3.00	2.00	3.50	3.50
4	3.00	2.00	3.50	3.50
5	4.00	1.00	6.50	–
6	4.00	1.00	6.50	–
7	4.00	1.00	6.50	–
8	4.00	2.00	6.50	6.50
9	5.00	1.00	10.50	–
10	5.00	1.00	10.50	–
11	5.00	2.00	10.50	10.50
12	5.00	2.00	10.50	10.50
13	6.00	1.00	14.50	–
14	6.00	1.00	14.50	–
15	6.00	2.00	14.50	14.50
16	6.00	2.00	14.50	14.50
17	7.00	1.00	19.00	–
18	7.00	1.00	19.00	–
19	7.00	2.00	19.00	19.00
20	7.00	2.00	19.00	19.00
21	7.00	2.00	19.00	19.00
22	8.00	1.00	23.50	–
23	8.00	1.00	23.50	–
24	8.00	2.00	23.50	23.50
25	8.00	2.00	23.50	23.50
26	9.00	1.00	28.00	–
27	9.00	1.00	28.00	–
28	9.00	2.00	28.00	28.00
29	9.00	2.00	28.00	28.00
30	9.00	2.00	28.00	28.00
31	10.00	1.00	31.50	–
32	10.00	2.00	31.50	31.50

- Wilcoxon rank sum:

$$W_{m,n} = \sum_{i=1}^{n} R_{N,i}$$
$$= 1.50 + 3.50 + \ldots + 31.5 = 284.5$$

- Mann-Whitney U:

$$U_{m,n} = W_{m,n} - \frac{n(n+1)}{2} = 131.5$$

Εφαρμογή

	βαθμός	τμήμα	τάξη	$R_{N,i}$
1	2.00	1.00	1.50	–
2	2.00	2.00	1.50	1.50
3	3.00	2.00	3.50	3.50
4	3.00	2.00	3.50	3.50
5	4.00	1.00	6.50	–
6	4.00	1.00	6.50	–
7	4.00	1.00	6.50	–
8	4.00	2.00	6.50	6.50
9	5.00	1.00	10.50	–
10	5.00	1.00	10.50	–
11	5.00	2.00	10.50	10.50
12	5.00	2.00	10.50	10.50
13	6.00	1.00	14.50	–
14	6.00	1.00	14.50	–
15	6.00	2.00	14.50	14.50
16	6.00	2.00	14.50	14.50
17	7.00	1.00	19.00	–
18	7.00	1.00	19.00	–
19	7.00	2.00	19.00	19.00
20	7.00	2.00	19.00	19.00
21	7.00	2.00	19.00	19.00
22	8.00	1.00	23.50	–
23	8.00	1.00	23.50	–
24	8.00	2.00	23.50	23.50
25	8.00	2.00	23.50	23.50
26	9.00	1.00	28.00	–
27	9.00	1.00	28.00	–
28	9.00	2.00	28.00	28.00
29	9.00	2.00	28.00	28.00
30	9.00	2.00	28.00	28.00
31	10.00	1.00	31.50	–
32	10.00	2.00	31.50	31.50

- Wilcoxon rank sum:

$$\begin{aligned} W_{m,n} &= \sum_{i=1}^{n} R_{N,i} \\ &= 1.50 + 3.50 + \ldots + 31.5 = 284.5 \end{aligned}$$

- Mann-Whitney U:

$$U_{m,n} = W_{m,n} - \frac{n(n+1)}{2} = 131.5$$

- Προσεγγιστική ελεγχοσυνάρτηση:

$$\begin{aligned} z &= \frac{W_{m,n} - n(N+1)/2}{\sqrt{n(N+1)(N-n)/12}} \\ &= \frac{284.5 - 280.5}{\sqrt{701.25}} \approx 0.15 < z_{0.05} \approx 1.64 \end{aligned}$$

Εφαρμογή

	βαθμός	τμήμα	τάξη	$R_{N,i}$
1	2.00	1.00	1.50	–
2	2.00	2.00	1.50	1.50
3	3.00	2.00	3.50	3.50
4	3.00	2.00	3.50	3.50
5	4.00	1.00	6.50	–
6	4.00	1.00	6.50	–
7	4.00	1.00	6.50	–
8	4.00	2.00	6.50	6.50
9	5.00	1.00	10.50	–
10	5.00	1.00	10.50	–
11	5.00	2.00	10.50	10.50
12	5.00	2.00	10.50	10.50
13	6.00	1.00	14.50	–
14	6.00	1.00	14.50	–
15	6.00	2.00	14.50	14.50
16	6.00	2.00	14.50	14.50
17	7.00	1.00	19.00	–
18	7.00	1.00	19.00	–
19	7.00	2.00	19.00	19.00
20	7.00	2.00	19.00	19.00
21	7.00	2.00	19.00	19.00
22	8.00	1.00	23.50	–
23	8.00	1.00	23.50	–
24	8.00	2.00	23.50	23.50
25	8.00	2.00	23.50	23.50
26	9.00	1.00	28.00	–
27	9.00	1.00	28.00	–
28	9.00	2.00	28.00	28.00
29	9.00	2.00	28.00	28.00
30	9.00	2.00	28.00	28.00
31	10.00	1.00	31.50	–
32	10.00	2.00	31.50	31.50

- Wilcoxon rank sum:

$$\begin{aligned} W_{m,n} &= \sum_{i=1}^{n} R_{N,i} \\ &= 1.50 + 3.50 + \ldots + 31.5 = 284.5 \end{aligned}$$

- Mann-Whitney U:

$$U_{m,n} = W_{m,n} - \frac{n(n+1)}{2} = 131.5$$

- Προσεγγιστική ελεγχοσυνάρτηση:

$$\begin{aligned} z &= \frac{W_{m,n} - n(N+1)/2}{\sqrt{n(N+1)(N-n)/12}} \\ &= \frac{284.5 - 280.5}{\sqrt{701.25}} \approx 0.15 < z_{0.05} \approx 1.64 \end{aligned}$$

- p-value: $\mathrm{P}(Z > 0.15) \approx 0.44 > \alpha = 0.05$

Εφαρμογή

	βαθμός	τμήμα	τάξη	$R_{N,i}$
1	2.00	1.00	1.50	–
2	2.00	2.00	1.50	1.50
3	3.00	2.00	3.50	3.50
4	3.00	2.00	3.50	3.50
5	4.00	1.00	6.50	–
6	4.00	1.00	6.50	–
7	4.00	1.00	6.50	–
8	4.00	2.00	6.50	6.50
9	5.00	1.00	10.50	–
10	5.00	1.00	10.50	–
11	5.00	2.00	10.50	10.50
12	5.00	2.00	10.50	10.50
13	6.00	1.00	14.50	–
14	6.00	1.00	14.50	–
15	6.00	2.00	14.50	14.50
16	6.00	2.00	14.50	14.50
17	7.00	1.00	19.00	–
18	7.00	1.00	19.00	–
19	7.00	2.00	19.00	19.00
20	7.00	2.00	19.00	19.00
21	7.00	2.00	19.00	19.00
22	8.00	1.00	23.50	–
23	8.00	1.00	23.50	–
24	8.00	2.00	23.50	23.50
25	8.00	2.00	23.50	23.50
26	9.00	1.00	28.00	–
27	9.00	1.00	28.00	–
28	9.00	2.00	28.00	28.00
29	9.00	2.00	28.00	28.00
30	9.00	2.00	28.00	28.00
31	10.00	1.00	31.50	–
32	10.00	2.00	31.50	31.50

- Wilcoxon rank sum:

$$W_{m,n} = \sum_{i=1}^{n} R_{N,i}$$
$$= 1.50 + 3.50 + \ldots + 31.5 = 284.5$$

- Mann-Whitney U:

$$U_{m,n} = W_{m,n} - \frac{n(n+1)}{2} = 131.5$$

- Προσεγγιστική ελεγχοσυνάρτηση:

$$z = \frac{W_{m,n} - n(N+1)/2}{\sqrt{n(N+1)(N-n)/12}}$$
$$= \frac{284.5 - 280.5}{\sqrt{701.25}} \approx 0.15 < z_{0.05} \approx 1.64$$

- p-value: $\mathrm{P}(Z > 0.15) \approx 0.44 > \alpha = 0.05$
- Δεν απορριπτω την H_0 έναντι της H_1: οι βαθμοί στο τμήμα 2 δεν είναι μεγαλύτεροι από το τμήμα 1 ($\alpha = 0.05$).

Εφαρμογή: Στην R

- α' τρόπος: με δικές μας εντολές
- β' τρόπος: μέσω της base function: `wilcox.test()`
- δείτε το αρχείο `wilcoxon_sum_rank.R`
- Παρατήρηση
  - Η R υπολογίζει το Mann-Whitney U ($U_{m,n}$) και όχι το Wilcoxon rank sum ($W_{m,n}$)
  - Το p-value του ελέγχου δεν επηρεάζεται από αυτό

Εφαρμογή: ακριβής κατανομή

- Υπολογισμός ποσοστιαίου σημείου της ακριβούς κατανομής του $U_{m,n}$
  - `qwilcox(p = 0.05, m = 17, n = 15, lower.tail = F)`
  - 171
- Προηγουμένως όμως υπολογίσαμε ότι $U_{m,n} = 131.5$
- Επειδή $131.5 < 171$, δεν απορρίπτουμε την H_0, έναντι της H_1.
- Υπολογισμός ακριβούς p-value:
  - `pwilcox(q = 131.5, m = 15, n = 17,lower.tail = F)`
  - `0.4408149`
- Παρατηρήστε ότι το προσεγγιστικό p-value που υπολογίσαμε πριν είναι αρκετά κοντά στο (ακριβές) p-value
- Προσοχή: οι παραπάνω συναρτήσεις μπορούν να κρασάρουν την `R` αν τουλάχιστον ένα εκ των m, n είναι «μεγάλο».