1 Zadanie 6. (Test Wilcoxona dla dwóch prób.)

Na dwóch grupach osób o liczebnościach m i n porównujemy skuteczność dwóch terapii A i B, chcąc wykazać, że B działa lepiej niż A. Niech X i Y oznaczają reakcję na działanie terapii A i B, odpowiednio. Im X (Y) większe tym terapia A (B) zadziałała lepiej. Znając wartości przyjęte przez zmienne losowe $X_1, \ldots, X_m$ (reakcje pacjentów na terapię A) i $Y_1, \ldots, Y_n$ (reakcje pacjentów na terapię B) chcemy zweryfikować:

$$\begin{array}{ll} H_0: & \text{terapie A i B działają tak samo,} \\ H_1: & \text{terapia B działa lepiej niż A.} \end{array} \tag{1}$$

1.1 Statystyka testowa

Model. Obserwujemy dwie niezależne próby losowe

$$\begin{array}{ll} X_1, \ldots, X_m & \text{i.i.d. } F, \\ Y_1, \ldots, Y_n & \text{i.i.d. } G. \end{array}$$

Zakładamy, że F i G są dystrybuantami rozkładów ciągłych oraz $G(x) = F(x-\theta)$ dla pewnego nieznanego $\theta \in \mathbb{R}$. Przy takich założeniach, problem testowania (1) można zapisać w równoważnej postaci:

$$\begin{array}{ll} H_0: & \theta = 0 \\ H_1: & \theta > 0, \end{array}$$

Aby opisać test Wilcoxona musimy wprowadzić pojęcie rangi.

Definition 1 *Ranga obserwacji a_k w dowolnej próbie $a_1, \ldots, a_n$ to liczba tych obserwacji w tej próbie, które są mniejsze lub równe a_k*

$$r_k \stackrel{def}{=} \#\{j \,:\, a_j \leq a_k\}.$$

Aby nadać rangi obserwacjom $a_1, \ldots, a_n$, ustawiamy je w kolejności od najmniejszej do największej (tworzymy *statystystyki porządkowe*). Ranga każdej obserwacji to numer miejsca, który zajmuje ona w tym uporządkowanym ciągu. Najmniejsza z obserwacji ma więc rangę 1, a największa rangę n.

Uwaga. Taki sposób przypisywania rang ma sens, gdy w próbie nie ma powtarzających się obserwacji. Oczywiście, jeśli próba pochodzi z rozkładu ciągłego, to z prawdopodobieństwem 1 nie ma w niej powtarzających się obserwacji.

Opis testu Wilcoxona sumy rang (ang. Wilcoxon rank-sum test). Niech $s_1 < s_2 < \ldots < s_n$ oznaczają uporządkowane rosnąco rangi y-ów w **połączonej** próbie $x_1, \ldots, x_m, y_1, \ldots, y_n$. Można pokazać, że jeśli hipoteza H_0 jest prawdziwa, to statystyka

$$W = \sum_{i=1}^{n} S_i$$

ma rozkład o średniej i wariancji

$$\mu_W = \frac{n(n+m+1)}{2}, \quad \sigma_W^2 = \frac{mn(n+m+1)}{12}.$$

Wartości W znacznie różniące się od μ_W **są nietypowe** dla H_0. Co więcej, duże wartości W są znacznie bardziej prawdopodobne dla H_1 niż dla H_0. Jeśli bowiem kuracja B jest lepsza od kuracji A, to rangi $s_1, \ldots, s_n$ przyjmują duże wartości, bo $y_1, \ldots, y_n$ mają tendencję do przyjmowania większych wartości od $x_1, \ldots, x_m$. Test Wilcoxona odrzuca więc H_0 na rzecz H_1, gdy statystyka W przyjmie wartość w znacznie większą od μ_W.

Z symetrii wynika że dla $\binom{N}{n}$ możliwych wyborów liczb naturalnych $1 \leq s_1 < \ldots < s_n \leq N = n + m$ zachodzi

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

To umożliwia wyznaczenie kwantyli rozkładu tej statystyki testowej dla H_0. Mamy bowiem

$$P_{H_0}(W \geq w) = \sum_{\left\{ \substack{1 \leq s_1 < \ldots < s_n \leq n \\ s_1 + \ldots + s_n \geq w} \right\}} \frac{1}{\binom{N}{n}}$$

Ponadto, można udowodnić, że dla H_0 zachodzi zbieżność

$$W^* := \frac{W - \frac{1}{2}n(N+1)}{\sqrt{mn(N+1)/12}} \xrightarrow{D} N(0,1)$$

gdy n i m dążą do nieskończoności.

1.2 Test Wilcoxona

1. **Test dokładny:** dla małych n, m rozkład W dla H_0 został stablicowany, więc można z tablic odczytać

   $$\text{p-value} = P_{H_0}(W \geq w).$$

   Odrzucamy H_0, gdy p-value $\leq \alpha$.

2. **Test asymptotyczny:** Dla dużych m, n odrzucamy H_0, gdy

   $$W^* := \frac{W - \frac{1}{2}n(N+1)}{\sqrt{mn(N+1)/12}} \geq z_\alpha,$$

   gdzie z_α jest kwantylem rzędu $1 - \alpha$ rozkładu $N(0,1)$. Taki test ma asymptotyczny poziom istotności α, tzn.

   $$\lim_{m,n \to \infty} P_{H_0}\left(\frac{W - \frac{1}{2}n(N+1)}{\sqrt{mn(N+1)/12}} \geq z_\alpha \right) \to \alpha.$$

1.3 Zadanie.

W pierwszej części zadania oblicz dokładną wartość prawdopodobieństwa $P_{H_0}(W^* \geq z_\alpha)$ dla wskazanych (m, n) i α (generując kombinacje za pomocą odpowiedniej funkcji, np. **combinations** z pakietu R).

(n,m)	$P_H(W^* \geq z_\alpha)$			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.005$
$(5,5)$				
$(8,5)$				
$(10,5)$				

Czy dla małych (m,n) aproksymacja rozkładu statystyki W^* rozkładem normalnym jest sensowna?

Z aproksymacji $N(0,1)$ wynika, że dla każdego $w \in R$,

$$P_{H_0}(W \geq w) = P_H\left(W^* \geq \frac{w - \frac{1}{2}n(N+1)}{\sqrt{mn(N+1)/12}}\right) \approx 1 - \Phi\left(\frac{w - \frac{1}{2}n(N+1)}{\sqrt{mn(N+1)/12}}\right)$$

Jak dokładna jest ta aproksymacja? Oznaczmy:

$\alpha_0(w) = P_{H_0}(W \geq w)$ - wartość dokładna (funkcja **combinations** z R);
$\alpha_1(w) = 1 - \Phi\left(\frac{w - \frac{1}{2}n(N+1)}{\sqrt{mn(N+1)/12}}\right)$ - aproksymacja;
$\alpha_2(w) = 1 - \Phi\left(\frac{w - \frac{1}{2} - \frac{1}{2}n(N+1)}{\sqrt{mn(N+1)/12}}\right)$ - aproksymacja z poprawką na ciągłość. Tę poprawkę stosujemy wtedy, gdy wykorzystujemy CTG do aproksymacji rozkładu dyskretnego.

W drugiej części zadania wyznacz $\alpha_0(w), \alpha_1(w), \alpha_2(w)$ dla wskazanych (m,n) i w.

	$m = 6,\ n = 3$				
	$w = 9$	$w = 12$	$w = 15$	$w = 18$	$w = 21$
$\alpha_0(w)$					
$\alpha_1(w)$					
$\alpha_2(w)$					

	$m = 6,\ n = 6$				
	$w = 27$	$w = 33$	$w = 39$	$w = 45$	$w = 51$
$\alpha_0(w)$					
$\alpha_1(w)$					
$\alpha_2(w)$					

Czy poprawka na ciągłość polepsza jakość aproksymacji?