2 - Teste de Mann-Whitney - Duas amostras independentes

O objetivo é comparar a distribuição de probabilidades associadas a duas populações independentes com base em duas amostras dessas populações.

Dados

$X_1 \dots X_n$ → amostra da pop. 1

$Y_1 \dots Y_m$ → amostra da pop. 2

Associa-se postos à amostra combinada (de $n+m$ elementos) da forma usual

$R(X_i) = \text{posto}(X_i)$

$R(Y_j) = \text{posto}(Y_j)$

Suposições:

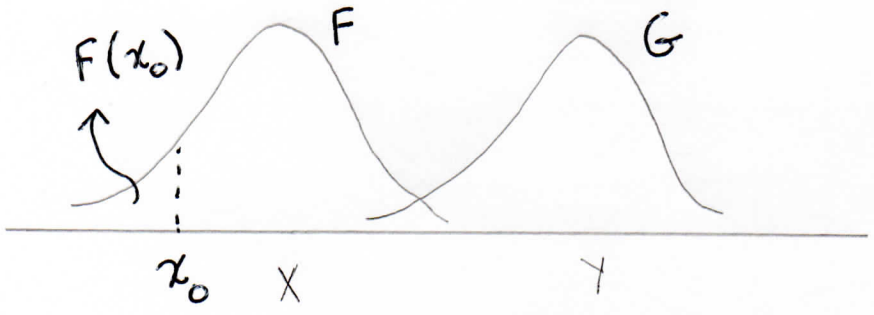
- As amostras são aleatórias e independentes entre si.
- A escala de medidas é pelo menos ordinal.

Sejam F - função distribuição de X

G - função distribuição de Y

- $F_X(x) = G_Y(x+c)$ $\quad c$ constante

Ex: X v.a. $\quad Y = X + c$



$$G(x_0) = 0$$

$$F_X(x) = P(X \leq x) = P(X + c \leq x + c) = P(Y \leq x + c) = G_Y(x + c)$$

$c > 0$ $\Rightarrow$ os valores de Y tendem a ser maiores que os de X.

$\Rightarrow G_Y(x) < F_X(x) \Rightarrow E(X) < E(Y)$.

$c < 0$ $\Rightarrow F_X(x) < G_Y(x) \Rightarrow E(X) > E(Y)$

A) Teste Bicaudal

$H_0: E(X) = E(Y) \iff c = 0 \iff F(x) = G(x)$

$H_a: E(X) \neq E(Y) \iff c \neq 0 \iff F(x) \neq G(x)$

B) Teste Unicaudal

$H_0: E(X) \geq E(Y) \iff c \leq 0 \iff F(x) \leq G(x)$

$H_a: E(X) < E(Y) \iff c > 0 \iff F(x) > G(x)$

C) Teste Unicaudal

$H_0: E(X) \leq E(Y) \iff c \geq 0 \iff F(x) \geq G(x)$

$H_a: E(X) > E(Y) \iff c < 0 \iff F(x) < G(x)$

Estatística de Teste:

$T = \sum_{i=1}^{n} R(X_i)$, se não existirem empates ou existirem poucos empates.

A Tabela A7 fornece os quantis de ordem 0,001, 0,005, 0,01, 0,025, 0,05 e 0,10 da distribuição de T sob H_0, para $n \leq 20$ e $m \leq 20$.

Os quantis superiores são

$$w_p = n(n+m+1) - w_{1-p}$$

Se $n > 20$ ou $m > 20$, usar

$$w_p = \frac{n(N+1)}{2} + z_p \sqrt{\frac{nm(N+1)}{12}}$$

z_p - quantil de ordem p da dist. $N(0,1)$, $N = n+m$.

Se existirem muitos empates, utilizar

$$T_1 = \frac{T - n\frac{N+1}{2}}{\sqrt{\frac{nm}{N(N-1)}\sum_{i=1}^{N} R_i^2 - \frac{nm(N+1)^2}{4(N-1)}}}$$

$\sum_{i=1}^{N} R_i^2$ - soma dos $n+m$ postos ao quadrado.

Neste caso, sob H_0, $T_1 \sim N(0,1)$

Regiões críticas

A) $T < w_{\alpha/2}$ ou $T > w_{1-\alpha/2}$

B) $T < w_{\alpha}$

C) $T > w_{1-\alpha}$.

Exemplo - Conover

Deseja-se testar se alunos que vivem no campo têm maior aptidão física que os que vivem na cidade.

Para isso foram avaliadas amostras de $n = 12$ alunos moradores no campo e $m = 36$ alunos moradores na cidade.

Para cada aluno, foi obtido um escore de aptidão física tal que baixo escore indica pobre condição física.

X_i (campo)	Y_i (cidade)
14,8	12,7
⋮	⋮
2,7	4,0
$n=12$	$m=36$

H_0: Os alunos que vivem no campo não têm melhor aptidão física.

H_a: Os alunos que vivem no campo têm melhor aptidão.

H_0: $E(X) \leq E(Y)$

H_a: $E(X) > E(Y)$

→ Escore médio populacional dos alunos moradores no campo.

X	Y	Posto	X	Y	Posto	X	Y	Posto
	1,0	1	4,2		10		7,6	23
	1,8	2		5,0	11		7,9	24
	2,1	3		5,6	13		8,3	25
	2,4	4		5,6	13	9,0		26
	2,6	5	5,6		13		9,1	27
2,7		6		5,9	15		9,9	28
	3,2	7		6,1	16		10,6	30,5
	3,6	8		6,2	17		10,6	30,5
	4,0	9	6,3		18		10,6	30,5
				6,4	19	10,6		30,5
				6,7	20,5			
				6,7	20,5			
			7,3		22			

$$T = \sum_{i=1}^{12} R(X_i) = 6 + 10 + 13 + \dots + 45 = 321$$

RC: $T > w_{1-\alpha}$

$m > 20 \quad \alpha = 0{,}05$

$$w_{0,95} = \frac{12.36}{2} + 1{,}64 \sqrt{\frac{12.36.49}{12}} \simeq 285$$

$T < w_{0,95}$, não rejeita-se H_0.

Ao nível 0,05, não rejeitamos a hipótese de igualdade das aptidões médias de alunos moradores no campo e na cidade.

Distribuição exata de T:

Sob H_0 $F(x) = G(x)$ (teste bicaudal) X e Y são identicamente distribuídas.

No teste unicaudal, verifica-se que

$$\sup P(T \in RC \mid H_0) = P(T \in RC \mid F(x) = G(x))$$

A distribuição exata de T pode ser obtida assumindo que X e Y são identicamente distribuídas ($\forall$ Ho casos A, B ou C) e sob essa hipótese, os postos são distribuídos de forma aleatória.

X e Y ident. distribuídas $\Rightarrow$ todos os arranjos de X's e Y's na amostra ordenada combinada são equiprováveis.

$$P(\text{Cada arranjo}) = \frac{1}{\binom{n+m}{n}}$$

$$P(T=k) = \frac{\#(n,m,k)}{\binom{n+m}{n}} \quad \text{onde}$$

$\#(n,m,k)$ = nº de arranjos para os quais $T=k$.

Distribuição exata de T para $n=3$, $m=2$ sob H_0

1	2	3	4	5	$T = \sum R(X_i)$
Y	Y	X	X	X	12
Y	X	Y	X	X	11
Y	X	X	Y	X	10
Y	X	X	X	Y	9
X	Y	Y	X	X	10
X	Y	X	Y	X	9
X	Y	X	X	Y	8
X	X	Y	Y	X	8
X	X	Y	X	Y	7
X	X	X	Y	Y	6

Sob H_0, cada uma das 10 distribuições de postos tem mesma probabilidade $\frac{1}{10}$

$$10 = \binom{3+2}{2}$$

Dist. exata de T sob H_0

T	6	7	8	9	10	11	12
$P(T=k)$	1/10	1/10	2/10	2/10	2/10	1/10	1/10

RC do teste bicaudal: $T \leq 6$ ou $T \geq 12$

$\alpha = \frac{2}{10} = 0,2$ menor α possível

Problema: n e m pequenos e T discreta

Obs:

- O correspondente teste paramétrico para detectar diferenças de duas médias é o teste t, que assume normalidade.

- Se houver normalidade, o teste de Mann-Whitney é quase tão poderoso quanto o teste t.

- Sob normalidade, a eficiência relativa assintótica do teste de Mann-Whitney (com relação ao t) é 0,955. Se a dist. é uniforme, a ef. relativa assintótica é 1,0 e pode ser muito maior para outros tipos de distribuição.

- Se as duas populações diferem apenas em locação ($F(x) = G(x+c)$), verifica-se que ARE $\geq 0{,}864$

Leitura recomendada:

Seção 13.3.2 Populações Não-Normais
Bussab e Morettin
Na 7ª edição pag 372 a 381.