5.3.2. Teste de Wilcoxon / Mann-Whitney

5.3.2.1. Introdução

Este teste (*Wilcoxon rank sum test*) é o mais popular dos testes não-paramétricos feitos com duas amostras. Existe na literatura alguma incerteza sobre a nome deste teste; alguns livros e programas (como o R) o chamam de 'teste de Wilcoxon para soma de postos' (*Wilcoxon rank sum test*), outros o chamam de 'teste de Mann-Whitney', em homenagem aos pesquisadores que desenvolveram o teste, independentemente, no final dos anos 1950s. Há algumas diferenças nos modos como definiram as estatísticas de teste (chamada de *W* por Wilcoxon e de *U* por Mann-Whitney), mas os resultados obtidos são equivalentes. Usaremos nos exemplos a seguir o teste de Wilcoxon como implementado no R.

Este teste é um dos testes não-paramétricos de maior poder, e por isso é muito usado como alternativa ao teste *t* nos problemas em que as populações não são normais, ou as amostras são pequenas e as variáveis são qualitativas ordinais.A hipótese nula é a de que duas populações têm a mesma distribuição de probabilidades.

5.3.2.2. Testes com amostras pequenas

Para entender a lógica do teste de Wilcoxon / Mann-Whitney (e a dos testes não-paramétricos em geral), podemos analisar o que acontece quando usamos amostras muito pequenas, tomando como exemplo experimentos feitos com duas amostras de animais.

(i) Exemplo 1

Suponha que desejamos comparar dois tipos de ração e verificar qual deles leva ao maior ganho de peso, em gramas. Faremos para isto um experimento com duas amostras pequenas de ratos de laboratório, e os resultados serão analisados por meio de um teste de Wilcoxon bilateral, com $\alpha = 0{,}05$. Seis ratos são separados aleatoriamente em duas amostras de mesmo tamanho ($n = 3$), e têm seus pesos (em gramas) anotados. Os ratos da amostra A são alimentados com a ração A, os da amostra B com a ração B, durante um período determinado; ao fim deste período, todos os ratos são pesados, e o ganho de peso de cada um é anotado. Em seguida, estes pesos são ordenados em ordem decrescente, e recebem um *posto* de acordo com sua posição nesta ordem (estes postos são normalmente representados pela letra *r*, do inglês *rank*). O menor ganho recebe o posto 1; o segundo menor recebe o posto 2, e assim por diante; o maior ganho recebe o posto 6. A partir daí, a análise dos resultados é feita considerando apenas estes postos, e não mais os valores originais dos ganhos de peso. A estatística de teste, se trabalhamos com amostras pequenas, é simplesmente a *soma* dos postos obtidos pela menor amostra (menor *n*); esta soma é em geral representada por Σr_1, enquanto a soma dos postos da amostra maior é representada por Σr_2.

Suponha que os dois menores ganhos de peso (em gramas) tenham sido obtidos com a ração A, e também o quarto menor ganho; estes valores receberão os postos 1, 2 e 4. A ração B, teve em geral maiores ganhos de peso, e receberá os postos 3, 5 e 6. Os valores originais do ganho de peso e os postos obtidos são mostrados nas tabelas **1** e **2**. Os postos são em geral representados pela letra *r* (do inglês *rank*).

Neste exemplo, as duas amostras têm mesmo tamanho, e podemos usar qualquer delas para calcular a variável de teste; usaremos a amostra A. A soma dos postos da amostra A foi bem pequena ($\Sigma r_A = 1+2+4=7$), e bem inferior á soma da amostra B ($\Sigma r_B=3+5+6=14$); isto mostra que, em geral, a ração A levou a ganhos de peso menores do que a ração B. Podemos considerar este resultado uma evidência de que a ração B leva a maior ganho de peso, ou será que este resultado foi devido meramente ao acaso?

Tabela 1

amostra	ganhos de peso (g)		
A	67	73	84
B	108	103	81

→

Tabela 2

amostra	postos (r)		
A	1	2	4
B	6	5	3

No teste de Wilcoxon / Mann-Whitney, a hipótese nula é de que as duas populações *têm a mesma distribuição de probabilidades*, e a probabilidade de um posto ser obtido pela amostra A é idêntica à probabilidade de ele ser obtido pela amostra B; os postos altos (que indicam maiores ganhos de peso) estarão espalhados aleatoriamente entre as duas amostras, como por sorteio. A hipótese alternativa, num teste bilateral, é a de que as duas distribuições diferem entre si em termos de localização (tendência central) e a probabilidade de obtermos postos altos nas amostras de uma das populações é maior do que nas amostras da outra.

Como em qualquer teste estatístico, para tomar uma decisão precisamos analisar o que seria mais provável ocorrer nas amostras se a hipótese nula fosse verdadeira. No exemplo, é fácil calcular as probabilidades de cada valor da soma de postos, dada a hipótese nula – como as amostras são pequenas, basta *enumerar* todos os resultados possíveis, e construir a partir daí tabelas para os valores críticos. Para amostras de tamanhos $n_1=n_2=3$, há $C_6^3=20$ maneiras equiprováveis de combinarmos os 6 postos em uma mesma amostra. Estas combinações resultam em 10 somas Σr_1 que variam de 6 a 15, conforme mostradas na Tab. **3**.

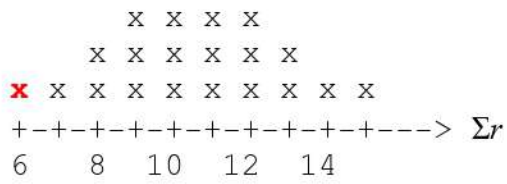
Tabela 3

Σr_1	combinações de postos que levam à mesma soma Σr_1,			quantidade	probab.
6	1-2-3			1	0,05
7	1-2-4			1	0,05
8	1-2-5	1-3-4		2	0,10
9	1-2-6	1-3-5	2-3-4	3	0,15
10	1-3-6	1-4-5	2-3-5	3	0,15
11	1-4-6	2-3-6	2-4-5	3	0,15
12	1-5-6	2-4-6	3-4-5	3	0,15
13	2-5-6	3-4-6		2	0,10
14	3-5-6			1	0,05
15	4-5-6			1	0,05
				20	1

Analisando a tabela, vemos que não há valores *críticos*, se fazemos um teste bilateral; não é possível construir intervalos em cada extremo da curva de modo que sua soma de probabilidades seja igual a $\alpha = 0,05$. Não há portanto nenhum valor de Σr_1 que leve à rejeição da hipótese nula; o valor de $\Sigma r_1=7$ encontrado na amostra é não-significativo.

Se tivéssemos feito um teste unilateral à esquerda, a região de rejeição iria conter apenas o $\Sigma r_1= 6$, como representado no gráfico de pontos da Fig. **1**; o valor $\Sigma r_1=7$ encon-

trado na amostra continuaria a ser não-significativo. Se o teste fosse unilateral à direita, a região de rejeição iria conter apenas o valor Σr_1=15.



(ii) Exemplo 2

Suponha agora que tenhamos feito o mesmo teste, mas usando amostras maiores, de $n_1 = 3$ e $n_2 = 7$, e obtivemos os resultados mostrados na Tab. **4**, e os postos na Tab. **5**.

Tabela 4

amostra	Ganhos de peso ao final do tratamento (g)						
A	67	73	84				
B	108	103	81	121	94	118	124

Tabela 5

amostra	postos							Σr
A	1	2	4					$\Sigma r_1 = 7$
B	7	6	3	9	5	8	10	$\Sigma r_2 = 48$

Para amostras com $n_1 = 3$ e $n_2 = 7$, podemos calcular por análise combinatória que existirão 120 combinações diferentes dos 10 postos 3 a 3:

$$C_{10}^{3} = 120$$

Estas 120 combinações serão equiprováveis, cada uma com probabilidade

$$p = 1/120 = 0{,}0083$$

As probabilidades de cada valor da soma Σr_1 dos postos da amostra menor ($n = 3$) são as mostradas na Tab. **6**.

Tabela 6

Σr_1	combinações possíveis				quantidade de combinações que levam à mesma Σr_1	Probabilidade	
6	1-2-3				1	1 × 0,0083	= 0,0083
7	1-2-4				1	1 × 0,0083	= 0,0083
8	1-2-5	1-3-4			2	2 × 0,0083	= 0,0167
9	1-2-6	1-3-5	2-3-4		3	3 × 0,0083	= 0,0250
10	1-2-7	1-3-6	1-4-5	2-3-5	4	4 × 0,0083	= 0,0333
...	...	...	...	...	...		...
23	4-9-10	5-8-10	6-7-10	6-8-9	4	4 × 0,0083	= 0,0333
24	5-9-10	6-8-10	7-8-9		3	3 × 0,0083	= 0,0250
25	6-9-10	7-8-10			2	2 × 0,0083	= 0,0167
26	7-9-10				1	1 × 0,0083	= 0,0083
27	8-9-10				1	1 × 0,0083	= 0,0083
					120	1	

Esta tabela é simétrica; para testes bilaterais, os valores críticos serão os da Tab. 7.

Tabela 7. Valores críticos para testes bilaterais (n_1=3, n_2=7)

α	combinações possíveis de postos na área de rejeição	Probab	área de rejeição	área de aceitação
0,010	-	-	-	-
0,025	1-2-3 e 8-9-10	2 × 0,0083 = 0,0167	$\Sigma r_1 < 7$, $\Sigma r_1 > 26$	$7 \leq \Sigma r_1 \leq 26$
0,050	1-2-3, 1-2-4 e 7-9-10, 8-9-10	4 × 0,0083 = 0,0334	$\Sigma r_1 < 8$, $\Sigma r_1 > 25$	$8 \leq \Sigma r_1 \leq 25$

Para testes unilaterais à esquerda, os valores críticos serão os da Tab. 8.

Tabela 8 – Valores críticos para testes unilaterais à esquerda (n_1=3, n_2=7)

α	combinações possíveis de postos na área de rejeição	Probab	área de rejeição	área de aceitação
0,010	1-2-3	0,0083	$\Sigma r_1 = 6$	$\Sigma r_1 \geq 7$
0,025	1-2-3, 1-2-4	2×0,0083 = 0,0167	$\Sigma r_1 \leq 7$	$\Sigma r_1 \geq 8$
0,050	1-2-3, 1-2-4, 1-2-5,1-3-4	4×0,0083 = 0,0334	$\Sigma r_1 \leq 8$	$\Sigma r_1 \geq 9$

Note que a soma encontrada na amostras, Σr_1=7 está entre os valores extremos da distribuição mostrada na Tab. 7, e o resultado portanto é *significativo*, e leva à rejeição da hipótese de que os ganhos de pesos nas duas populações (ratos alimentados com a ração A e ratos alimentados com a ração B) tenham a mesma distribuição de probabilidades. Ou seja, a probabilidade de um animal obter grandes ganhos de peso deve ser maior se ele for alimentado com a ração B, do que se for com a ração A.

Fazendo no R, usando a função `wilcox.test`, obtemos:

```
x=c(67,73,84)
y=c(108,103,81,121,94,118,124)
wilcox.test(x,y,alternative = c("two.sided"))
```

```
        Wilcoxon rank sum exact test
        data:  x and y
        W = 10, p-value = 0.03333
        alternative hypothesis: true location shift is not equal to 0
```

O valor p = 0,03333 confirma a conclusão a que já tínhamos chegado antes: existe diferença significativa entre os ganhos de pesos obtidos com as duas dietas.

5.3.2.3. Testes com amostras grandes

Para amostras grandes, não seria possível usar métodos de análise combinatória para calcular as probabilidades associadas a cada valor da soma de postos, como feito acima. Wilcoxon e Mann-Whitney sugeriram usar duas estatísticas de teste definidas de formas diferentes, ambas com distribuição normal.

(i) Teste de Mann-Whitney

Mann e Whitney sugeriram usar como estatística de teste o valor U, definido como

U = (soma de pontos da amostra menor) – (menor soma teoricamente possível)

ou,

$$U = \sum r_1 - \frac{n_1(n_1+1)}{2} \qquad (1)$$

onde Σr_1 é a soma dos postos da amostra de menor tamanho, e a parcela $n_1(n_1+1)/2$ calcula o valor mínimo da soma de postos para uma amostra de tamanho n_1

No exemplo, n_1=3, e a menor soma possível pode ser calculada como

$$\frac{n_1(n_1+1)}{2} = \frac{3(3+1)}{2} = 6$$

Esta soma somente será obtida se a amostra menor contiver os três menores postos, 1, 2 e 3. Mann e Whitney mostraram que, se a variável medida nas duas amostras for contínua, a estatística U terá distribuição que tenderá para a normal, com a média e variância dados em (2).

$$\mu_U = \frac{n_1 n_2}{2} \qquad \sigma_U^2 = \frac{n_1 n_2 (n_1+n_2+1)}{12} \qquad (2)$$

(ii) Teste de Wilcoxon

Na versão do teste desenvolvida por Wilcoxon, a estatística de teste W tem seu valor dado simplesmente pela soma dos postos da amostra menor:

$$W = \sum r_1 \qquad (3)$$

Wilcoxon mostrou que, se a variável medida nas duas amostras for contínua, a estatística W terá distribuição que tenderá para a normal, com média e variância:

$$\mu_W = \frac{n_1(n_1+n_2+1)}{2} \qquad \sigma_W^2 = \frac{n_1 n_2 (n_1+n_2+1)}{12} \qquad (4)$$

As fórmulas de Mann-Whitney e de Wilcoxon resultam o mesmo valor da estatística Z, e no mesmo valor-p. Note que voltamos, depois de tudo, a cair na distribuição normal, e o teste agora exige que a variável medida seja contínua (o teste por análise combinatória visto nos dois exemplos acima não faz nenhuma exigência, e pode ser usado também para variáveis ordinais). O teste de Wilcoxon / Mann-Whitney continua porém a ser *não-paramétrico*, porque o objetivo é comparar as formas de duas distribuições, não seus parâmetros.

(iii) Exemplo 3

A capacidade máxima de absorção de oxigênio durante atividade física, por unidade de peso corporal, é chamada de "capacidade aeróbica" e serve como medida da capacidade de trabalho físico de uma pessoa. Para um estudo comparativo, foram feitas medidas da capacidades aeróbicas de 10 habitantes nativos das terras altas do Peru (Amostra 1) e de 15 habitantes nativos das cidades litorâneas, mas já aclimatizados às terras altas. Os dados são mostrados na Tab. 1. Estes valores indicam que existe uma diferença significativa entre a capacidade aeróbica destes dois grupos?

Tabela 1. Valores observados nas amostras

Amostra 1	Amostra 2
40.2 46.8 54.1	37.3 36.1 37.9
43.4 51.4 45.5	30.8 37.4 39.9
50.1 41.2 53.7	42.3 47.8 36.7
50.2	41.7 49.2 38.7
	53.0 39.0 44.3

→

Tabela 2. Postos atribuídos

Amostra 1	Amostra 2
10 17 25	04 02 06
14 22 16	01 05 09
20 11 24	13 18 03
21	12 19 07
	23 08 15
Σr_1=180	Σr_2=145

Os postos atribuídos a cada observação são mostrados na Tab. **2**. Faremos primeiro os cálculos usando a estatística de teste *W*, usando as equações em (**3**) e (**4**). O valor de *W* será:

$$W = \sum r_1 = 180$$

Os parâmetros de sua distribuição amostral (normal) serão:

$$\mu_W = \frac{n_1(n_1+n_2+1)}{2} = \frac{10(10+15+1)}{2} = 130$$

$$\sigma_W^2 = \frac{n_1 n_2(n_1+n_2+1)}{12} = \frac{10\times 15(10+15+1)}{12} = 325$$

Padronizando a estatística amostral, obtemos:

$$z = \frac{U-\mu_W}{\sigma_W} = \frac{180-130}{\sqrt{325}} = 2{,}7735$$

Na tabela da distribuição normal, vemos que para z = 2,77

$$P(Z > 2{,}77) = 0{,}5 - 0{,}4972 = 0{,}0028$$

Como o teste é bilateral,

$$\text{valor-p} = 2 \times 0{,}0028 = 0{,}0056$$

A hipótese nula pode então ser rejeitada; há forte evidência de que as distribuições da capacidade aeróbica nas duas populações não sejam idênticas.

O programa R tem a função `wilcox.test`, que faz o teste da soma de postos de Wilcoxon / Mann-Whitney (*Wilcoxon rank sum test*). Os comandos necessários para fazer o teste são:

```
x=c(40.2,46.8,54.1,43.4,51.4,45.5,50.1,41.2,53.7,50.2)
y=c(37.3,36.1,37.9,30.8,37.4,39.9,42.3,47.8,36.7,41.7,
49.2,38.7,53.0,39.0,44.3)
wilcox.test(x,y,alternative = c("two.sided"))
```

Os resultados do teste são dados abaixo:

```
Wilcoxon rank sum exact test
data:  x and y
W = 125, p-value = 0.004435
alternative hypothesis:
true location shift is not equal to 0
```

O valor-*p* calculado é muito significativo, e leva à rejeição da hipótese de que as duas amostras tenham vindo de populações com a mesma distribuição. No R a hipótese nula é expressa em termos de deslocamento (*shift*) da posição da distribuição; a conclusão é que as duas distribuição não tem a mesma localização.

Note que o valor-p encontrado (p=0,0044) é um pouco diferente do valor que obtivemos fazendo os cálculos a partir da distribuição normal (p=0,0056); isto ocorre porque, para estes tamanhos de amostra (n_1=10 e n_2=15), o R faz o teste *exato*, no qual as probabilidades são calculadas por meio de Análise Combinatória (como no Exemplo **2**), e não apenas aproximadas pela distribuição normal. A diferença entre os dois valores porém é muito pequena, e não tem nenhuma importância na prática. (Note que o R chama o teste de *Wilcoxon rank sum test* e representa a estatística de teste por *W*; no entanto, usa a estatística e os parâmetros definidos em (**1**) e (**2**) para o teste que a maioria dos livros chama de *teste de Mann-Whitney*. Não há acordo entre os autores sobre esta denominação, mas isto não faz diferença, em termos dos resultados.)

Para verificar se o teste t também poderia ter sido usado neste problema, fazemos a seguir testes de Shapiro para a normalidade nas duas populações:

```
shapiro.test(x)
        Shapiro-Wilk normality test
        data:  x
        W = 0.93468, p-value = 0.4954
shapiro.test(y)
        Shapiro-Wilk normality test
        data:  y
        W = 0.94257, p-value = 0.4158
```

Estes resultados não levam à rejeição da hipótese de normalidade das populações, e o teste t também pode portanto ser usado:

```
t.test(x,y)
        Welch Two Sample t-test
        data:  x and y
        t = 3.1581, df = 21.246, p-value = 0.004696
        alternative hypothesis:
        true difference in means is not equal to 0
        95 percent confidence interval:
              2.343561 11.363106
        sample estimates:
        mean of x mean of y
        47.66000  40.80667
```

Note que as hipóteses alternativas do teste de Wilcoxon e do teste *t* são diferentes. No teste *t*, a hipótese é que há diferença entre as médias; no de Wilcoxon, a hipótese não faz menção aos parâmetros, mas diz apenas que há um deslocamento da posição (*location shift*) da distribuição, sem mencionar como esta posição é medida (pela média ou pela mediana?).

5.3.2.4. Comentários finais

O valor-p obtido nos dois testes no Exemplo **3** foram praticamente iguais, o que mostra que o teste de Wilcoxon / Mann-Whitney tem poder próximo do teste *t*, quando as amostras não são demasiado pequenas.

Um problema que pode surgir no teste de Wilcoxon / Mann-Whitney, ou em qualquer outro teste baseado em postos, é ocorrerem observações que tenham o mesmo valor (empates). Neste caso, se estes empates não forem muito numerosos, e ocorrerem dentro de uma mesma amostra, a solução mais simples é a de atribuir a estes valores o valor médio dos postos que ocupam.

Por exemplo, se temos esta amostra de 10 observações já ordenadas:

40	41	43	45	46	50	50	51	53	54

O valor 50 aparece duas vezes, nos postos 6 e 7; atribuímos então a este número o posto 6,5, e os demais postos são atribuídos normalmente, da forma:

1	2	3	4	5	6,5	6,5	8	9	10

Se contudo o número de empates for muito grande, e houver muitos casos em que observações de uma amostra empatam com observações de outra amostra, o poder do teste pode ser pode reduzido. Existem algumas maneiras de fazer correções no cálculo da estatística de teste; estes maneiras não serão discutidas neste texto, mas em geral já estão implementadas nas funções do R.