

Introdução

Vamos apresentar a parte dos Métodos não Paramétricos do capítulo 13 do livro do Bussab & Morettin. Nossa parte envolve a teoria de Populações não Normais (13.3.2 e 13.4.2).

Vou copiar descaradamente o que eles escreveram no capítulo 13: Inferência para Duas Populações mas acrescentarei algumas demonstrações bem como o uso do **R**.

13.1- Introdução:

Neste capítulo, abordaremos o tópico importante de comparar duas populações P_1 e P_2 , baseados em dados fornecidos por amostras dessa populações. como vimos, uma grande parte das técnicas usadas em Estatística supõe que as variáveis aleatórias envolvidas tenham distribuição normal. alguns testes que trataremos envolverão a normal. Contudo, se essa suposição de normalidade, for violada, procedimentos "mais robustos" têm de ser utilizados, e veremos exemplos de tal situação.

Uma pergunta que aparece frequentemente em Ciência é a seguinte: O método A é melhor do que o método B ? Em termos estatísticos, ela equivale a comparar dois conjuntos de informações, resultantes das medidas obtidas da aplicação dos dois métodos a dois conjuntos de objetos ou indivíduos.

Uma das dificuldades que enfrentamos é a de caracterizar adequadamente a "igualdade"ou "equivalência"de duas populações.

Por exemplo, suponha que estamos interessados em saber se alunos de duas regiões, A e B , tiveram desempenhos iguais em um mesmo teste nacional. Mas ainda, suponha que tenhamos os resultados do teste para "todos os alunos"das duas regiões, isto é, conhecemos, as duas populações. Suponha que cálculos posteriores revelem que as médias e os desvios padrões das duas populações sejam iguais, isto é,

$$\mu_A = \mu_B \text{ e } \sigma_A = \sigma_B.$$

Será que isso equivale a dizer que os desempenhos na duas regiões são equivalentes? Se uma análise mais cuidadosa não for feita, poderemos ser levados a responder afirmativamente a essa questão. Entretanto, observando a figura 13.1, vemos que é possível ter duas distribuições com os mesmos parâmetros acima, mas formas bastante distintas.

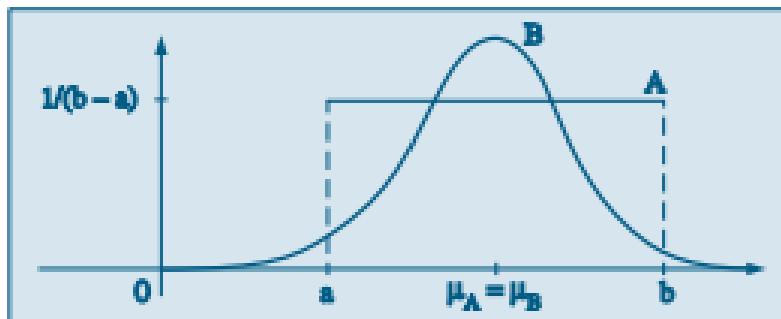
Esse fato nos remete à necessidade de também mencionarmos a forma da distribuição.

Especificada a forma, a igualdade dos parâmetros que identificam a curva implica a igualdade ou coincidência das duas populações. É bem pouco provável que um mesmo fenômeno obedeça a formas de distribuições distintas como no exemplo da Figura 13.1. Seguir uma mesma distribuição, porém com parâmetros distintos, é mais verossímil. Como a normal é um modelo importante e seguido por muitas variáveis de interesse prático, estaremos admitindo essa forma, a não ser quando uma análise dos dados nos diga o contrário.

Neste capítulo, trataremos de várias situações, que passamos a descrever.

1. Inferências para duas médias:amostras independentes

Figura 13.1: Distribuições das populações A e B, com $\mu_A = \mu_B = 4$, $\sigma_A = \sigma_B = 1,16$.



Aqui temos os dados na forma de duas amostras, extraídas independentemente de cada população. É muito comum em experimentos do tipo "controle"versus "tratamento", nos quais o interesse principal é verificar o efeito desse último. O caso típico é aquele de comparar um nova droga com uma padrão, usadas para o tratamento de uma doença.

Exemplo 13.1

- Um curso de Estatística é ministrado pela televisão para um grupo de alunos e ao vivo para outro grupo. Queremos testar a hipótese de que o curso ao vivo é mais eficaz que o curso por meio da televisão.
- Queremos comparar o efeito de duas rações, A e B, sobre o crescimento de porcos. Dois grupos de porcos em crescimento foram alimentados com as duas rações e após cinco semanas verificam-se quais foram os ganhos de peso dos porcos dos dois grupos.
- 20 canteiros foram plantados com uma variedade de milho. Em dez deles um novo tipo de fertilizante é aplicado e nos outros um fertilizante padrão. Examinando-se as produções dos dois grupos de dez canteiros, queremos saber se há diferenças significativas entre as produções

Na maioria das vezes fica claro o que chamamos de controle e tratamento. No exemplo acima, os canteiros tratados com o novo fertilizante seriam o grupo de tratamento enquanto os demais, tratados com o fertilizante usual, constituiriam o grupo controle. Mas nos exemplos (a) e (b) essa distinção é apenas convencional.

Formalmente, o modelo para o problema de duas amostras é o seguinte: As variáveis aleatórias X_1, X_2, \dots, X_m representam as respostas do grupo controle e são consideradas independentes., com a mesma distribuição, P_1 . Y_1, Y_2, \dots, Y_n representam as respostas do grupo tratamento e são consideradas independentes., com a mesma distribuição, P_2

A hipótese a ser testada é

$$H_0 : P_1 = P_2,$$

ou seja, queremos testar a homogeneidade das populações de onde as amostras foram extraídas. H_0 é chamado de homogeneidade.

O significado de (13.1) dependerá muito do interesse do pesquisador em considerar qual "tipo" de igualdade implicará a coincidência das duas distribuições. Admitamos que tanto P_1 e P_2 , sigam uma distribuição normal, ou seja,

$$P_1 \sim N(\mu_1, \sigma_1^2) \text{ e } P_2 \sim N(\mu_2, \sigma_2^2).$$

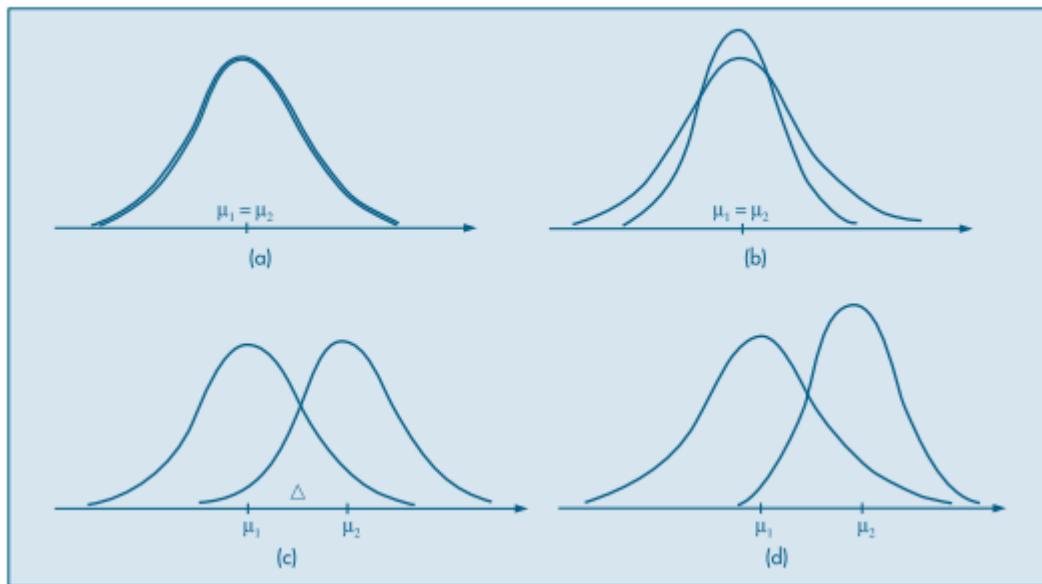
Na figura 13.2, temos as quatro situações possíveis. Observando os gráficos da figura 13.2 não temos dúvidas em reconhecer que as duas populações são iguais no caso **a** e diferentes no caso **d**. Já nos outros dois casos, podem existir situações em que elas podem ser consideradas iguais ou não. Por exemplo, uma pesquisa para verificar se o salário médio da região P_1 é o mesmo região P_2 aceita como resposta verdadeira tanto a situação **a** como a de **b**.

Outra pesquisa para verificar se dois processos produzem peças com a mesma qualidade em termos de dispersão aceita como verdadeira as situações **a** ou **c**. Assim, a estratégia para comparar duas populações, por meio de seus parâmetros, envolve suposições sobre a forma das distribuições, para depois testar médias e variâncias.

É comum estarmos interessados em testar apenas que P_1 e P_2 difiram em localização (ou posição), isto é, a alternativa a H_0 é que P_1 esteja à direita de P_2 , ou o contrário, mas que ambas tenham a mesma dispersão (caso $\mu_1 \neq \mu_2$ e $\sigma_1 = \sigma_2$). Nesse caso, H_0 será equivalente a

$$H_0 : \Delta = \mu_2 - \mu_1 = 0, \quad (13.2)$$

Figura 13.2: (a) $\mu_1 = \mu_2, \sigma_1 = \sigma_2$ (b) $\mu_1 = \mu_2, \sigma_1 \neq \sigma_2$ (c) $\mu_1 \neq \mu_2, \sigma_1 = \sigma_2$ (d) $\mu_1 \neq \mu_2, \sigma_1 \neq \sigma_2$.



Os testes t de Student e de Wilcoxon, descritos a seguir, são apropriados para este tipo de situação. O teste t é aplicável quando P_1 e P_2 supostas são normais, com médias $\mu_1 = \mu$ e $\mu_2 = \mu + \Delta$, respectivamente, e com mesma variância. O teste de Wilcoxon aplica-se para P_1 e P_2 quaisquer, mas supõe-se que a escala de medidas seja pelo menos ordinal.

A análise fica mais fácil quando a P_1 e P_2 são atribuídas distribuições de variáveis contínuas. Discutiremos a razão desta suposição adicional.

Outro caso de interesse é aquele em que queremos testar se as duas médias são iguais, mas as

variâncias são diferentes. Na figura 13.1, as duas curvas teriam dispersões diferentes ao redor de suas médias. Então, um teste preliminar de igualdade de variâncias seria necessário. O teste t de Student para o caso de populações normais será apresentado neste capítulo.

A hipótese (13.1) ou (13.2) nos diz que não há efeito de tratamento. A alternativa usual para H_0 é que o efeito do tratamento é o de aumentar as respostas. Isto é, P_2 gera valores maiores que P_1 , com maior frequência. Mas pode ocorrer o contrário: diminuir as respostas. Por exemplo, o "tratamento" visa diminuir o tempo para executar determinada tarefa.

1. Inferências para duas médias: amostras dependentes

Quando se comparam as médias de duas populações, pode ocorrer uma diferença significativa por causa de fatores externos não controlados. Por exemplo, no caso do Exemplo 13.4 abaixo, poderia ocorrer que um dos grupos tivesse vendedores mais experientes e habilidosos do que o outro. Logo, a diferença seria a esses fatos, e não ao mérito real da técnica de vendas. um modo de contornar esse problema é coletar as observações em pares, de modo que os dois elementos de cada par sejam homogêneos em todos os sentidos, exceto no que diz respeito ao fator que queremos comparar.

Por exemplo, no caso Exemplo 13.1(a), para testar os dois métodos de ensino poderíamos usar n pares de gêmeos, sendo que um elemento de cada par recebe aulas pela TV e outro ao vivo. Esse procedimento pretende controlar o maior número possível de fatores externos que possam afetar o aprendizado. Se houver diferenças no aprendizado, essa dever-se-á realmente ao método. Este

procedimento também é usado quando observações das duas amostras são feitas no mesmo indivíduo, por exemplo, medindo uma característica do indivíduo antes e depois de ele ser submetido a um tratamento.

O teste t de Student para as observações pareadas (ou emparelhadas), supondo normalidade, é apropriado para essas situações.

3. Inferências para duas variâncias: amostras independentes

Como vimos no item 1, podemos testar se duas amostras independentes provêm de duas populações com variâncias iguais, desconhecidas. Se essas variâncias forem diferentes, o teste tem que ser modificado. Esse teste, sob a suposição de normalidade das duas populações, usa uma estatística que tem uma distribuição especial chamada F de Snedecor.

Finalizando esta seção, ressaltamos que poderemos ter mais do que duas amostras, e técnicas semelhantes podem ser desenvolvidas.

13.2 Comparação das Variâncias de Duas Populações Normais

Uma das distribuições amostrais mais usadas, e que corresponde a uma distribuição F , resulta do seguinte problema. Suponha que temos duas amostras independentes, de tamanhos n_1 e n_2 , retiradas de duas populações normais com a mesma variância σ^2 . Indiquemos os estimadores de σ^2

obtidos das amostras por S_1^2 e S_2^2 , respectivamente. Já vimos que

$$U = \frac{(n_1 - 1)S_1^2}{\sigma^2} \sim \chi^2(n_1 - 1)$$

$$V = \frac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2(n_2 - 1)$$

e, portanto, a v.a.

$$\frac{U}{V} = \frac{\frac{(n_1 - 1)S_1^2}{\sigma^2}}{\frac{(n_2 - 1)S_2^2}{\sigma^2}} \rightarrow \frac{S_1^2}{S_2^2} = \frac{\frac{U}{n_1 - 1}}{\frac{V}{n_2 - 1}} \sim F(n_1 - 1, n_2 - 1) \quad 13.3$$

Essa variável será usada no teste desta seção.

Consideremos, agora, uma amostra X_1, \dots, X_n de uma população com distribuição $N(\mu_1, \sigma_1^2)$ e uma amostra Y_1, \dots, Y_m de uma população com distribuição $N(\mu_2, \sigma_2^2)$. Suponhamos que as duas amostras sejam independentes.

Queremos testar

$$\begin{aligned} H_0 : \sigma_1^2 &= \sigma_2^2 = \sigma^2 \\ H_1 : \sigma_1^2 &\neq \sigma_2^2. \end{aligned}$$

Chamemos de S_1^2 e S_2^2 as variâncias amostrais respectivas. De (13.3) e sob a suposição de H_0 ser verdadeira, isto é $\sigma_1^2 = \sigma_2^2$, temos que

$$W = S_1^2/S_2^2 \sim F(n - 1, m - 1). \quad 13.4$$

Fixado α , encontramos dois números f_1 e f_2 , da Tabela VI, tais que

$$P(W \in RC) = P(W < f_1 \text{ ou } W > f_2) = \alpha.$$

Os valores f_1 e f_2 são determinados de modo que $P(W < f_1) = \alpha/2 = P(W > f_2)$. Na prática, consideramos o quociente (13.4) de tal sorte que $S_1^2/S_2^2 > 1$.

Colhidas as amostras de n e m indivíduos, respectivamente, das duas populações, calculamos os valores observados s_{10}^2 e s_{20}^2 e o valor observado de W , ou seja,

$$w_0 = \frac{s_{10}^2}{s_{20}^2}.$$

Se w_0 pertencer à região crítica, rejeitamos H_0 ; caso contrário, a aceitamos.

Exemplo 13.2. Queremos verificar se duas máquinas produzem peças com a mesma homogeneidade quanto à resistência à tensão. Para isso, sorteamos duas amostras de seis peças de cada máquina, e obtivemos as seguintes resistências:

As hipóteses a serem testadas são:

$$H_0 : \sigma_A^2 = \sigma_B^2 = \sigma^2$$

Máquina A:	145	127	136	142	141	137
Máquina B:	143	128	132	138	142	132

$$H_1 : \sigma_A^2 \neq \sigma_B^2$$

Sob a suposição de normalidade das medidas de resistência à tensão, para as duas máquinas, temos que a *v.a.* W , definida por (13.4), tem uma distribuição $F(5,5)$. Fixando $\alpha = 0,10$ e consultando a Tabela VI, teremos:

$$RC =]0, (5,05)^{-1}[\cup]5,05, +\infty[.$$

Das amostras encontramos $s_A^2 = 40$ e $s_B^2 = 37$, portanto $w_0 = 1,08$. Como esse valor não pertence à região crítica, aceitamos H_0 , ou seja, as máquinas produzem com a mesma homogeneidade quanto à variabilidade.

Caso tivéssemos rejeitado a hipótese de igualdade das variâncias, seria conveniente obter um intervalo de confiança para o quociente das duas variâncias. De (13.3) podemos escrever, quando $\sigma_1^2 \neq \sigma_2^2$,

$$W = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{U/(n-1)}{V/(m-1)} \sim F(n-1, m-1)$$

e para um dado γ , $0 < \gamma < 1$, podemos encontrar dois valores f_1 e f_2 , tais que

$$P(f_1 < F(n-1, m-1) < f_2) = \gamma.$$

Dessa igualdade, segue-se que, com probabilidade γ ,

$$f_1 < \frac{S_1^2}{S_2^2} \frac{\sigma_2^2}{\sigma_1^2} < f_2$$

ou seja, o $IC(\sigma_2^2/\sigma_1^2; \gamma)$ será dado por

$$f_1 \frac{S_2^2}{S_1^2} < \frac{\sigma_2^2}{\sigma_1^2} < f_2 \frac{S_2^2}{S_1^2} \quad 13.5$$

Exemplo 13.3. Suponha que para outras seis medidas para as máquinas A e B do Exemplo 13.2 tivéssemos $S_A^2 = 85$ e $S_B^2 = 8$. Como $w_0 = 85/8 = 10,62$, rejeitariamós H_0 . Então, o IC dado por (13.5) ficaria, com $\gamma = 0,90$,

$$\frac{1}{5,05} \frac{8}{85} < \frac{\sigma_B^2}{\sigma_A^2} < 5,05 \frac{8}{85}$$

ou seja,

$$0,019 < \frac{\sigma_B^2}{\sigma_A^2} < 0,475.$$

Invertendo-se, obtemos, também,

$$2,10 < \frac{\sigma_A^2}{\sigma_B^2} < 52,6.$$

que indica a variação possível, no nível fixado, da razão entre as duas variâncias. Note que, sob H_0 , temos $\frac{\sigma_A^2}{\sigma_B^2} = 1$, que não pertence a esse intervalo.

Problemas

- Da população $X \sim N(50, 100)$ retirou-se uma amostra casual simples de $n = 10$ elementos. Da população $Y \sim N(60, 100)$ retirou-se uma amostra casual simples de $m = 6$ indivíduos, independente da primeira. Obtemos as variâncias amostrais S_1^2 e S_2^2 , respectivamente.
 - Encontre o valor de \mathbf{a} , tal que $P(S_1^2/S_2^2 < a) = 95\%$
 - Encontre o valor de \mathbf{b} , tal que $P(S_1^2/S_2^2 > b) = 95\%$
- Por que em (13.3) as *v.a.* \mathbf{U} e \mathbf{V} são independentes?
- Uma das maneiras de medir o grau de satisfação dos empregados de uma mesma categoria quanto à política salarial é por meio do desvio padrão de seus salários. A fábrica **A** diz ser mais coerente na política salarial do que a fábrica **B**. Para verificar essa afirmação, sorteou-se uma amostra de **10** funcionários não especializados de **A**, e **15** de **B**, obtendo-se os desvios padrões $s_A = 1.000$ reais e $s_B = 1.600$ reais. Qual seria a sua conclusão?
- Descreva-se comparar a qualidade de um produto produzido por duas fábricas. Essa qualidade será definida pela uniformidade com que o produto é produzido em cada fábrica. Tomaram-se duas amostras, uma de cada fábrica, medindo-se o comprimento dos produtos (o resumo dos resultados está no quadro abaixo). A qualidade das duas fábricas é a mesma? Caso a sua resposta seja negativa, dê um intervalo de confiança para indicar a intensidade dessa desigualdade.

Estatística	Fábrica A	Fábrica B
Amostra	21	17
Média	21,15	21,12
Variância	0,0412	0,1734

13.3 - Comparação de Duas Populações: Amostras Independentes

Nesta seção estudaremos o caso onde temos duas amostras independentes X_1, \dots, X_n e

Y_1, \dots, Y_m , de duas populações P_1 e P_2 , respectivamente. Estaremos interessados em comparar as médias dessas populações, verificando se elas podem ser consideradas iguais ou não. No caso de populações normais, teremos, preliminarmente, de usar o que aprendemos na seção anterior, para testar se as variâncias de P_1 e P_2 são iguais.

Consideraremos duas situações: na primeira, iremos supor que as populações sejam normais (reveja os Problemas 32,33 e 34 do Capítulo 10, os Problemas 31 e 32 do Capítulo 11 e o problema 29 do Capítulo 12); na segunda, essa suposição não é necessária.

13.3.1-Populações Normais

Aqui, $P_1 \sim N(\mu_1, \sigma_1^2)$ e $P_2 \sim N(\mu_2, \sigma_2^2)$

Queremos testar a hipótese (13.1), que aqui fica escrita na forma

$$H_0 : \mu_1 = \mu_2 = \mu.$$

Na situação da figura 13.2(c), a alternativa adequada é

$$H_1 : \mu_1 > \mu_2,$$

mas supondo variâncias iguais. Se estivermos apenas interessados se existe diferença entre as médias das duas populações, não importando a direção a direção, então a alternativa adequada é

$$H_1 : \mu_1 \neq \mu_2.$$

Para cada amostra calculamos os estimadores da média e da variância:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}, \quad S_1^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2;$$

$$\bar{Y} = \frac{\sum_{i=1}^m Y_i}{m}, \quad S_2^2 = \frac{1}{m-1} \sum_{i=1}^m (Y_i - \bar{Y})^2;$$

Sob a hipótese H_0 , isto é,

$$E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu - \mu = 0 \quad 13.6$$

$$Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m} \quad 13.7$$

Como $\bar{X} - \bar{Y}$ tem distribuição normal, se as variáveis fossem conhecidas, a estatística:

$$; \quad Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}}$$

Teria distribuição normal padrão, *sob a Hipótese Nula* H_0 , e poderia ser usada para testar H_0 contra H_1 . Contudo, nas situações de interesse prático, as variâncias não são conhecidas, devendo ser substituídas por estimadores convenientes. Aqui, a distribuição t de Student desempenha papel importante.

Notemos que, da definição de v.a t de Student, $t = \sqrt{n}(\bar{X} - \mu)/S$, podemos obter

$$t^2 = \frac{\left(\frac{\bar{X} - \mu}{\sigma\sqrt{n}}\right)^2}{[(n-1)S^2/\sigma^2]/(n-1)} \sim F(1, n-1) \quad 13.9$$

o que mostra uma relação entre as distribuições $t(n-1)$ e $F(1, n-1)$. Observe que o numerador de (13.9) é o quadrado de uma $N(0, 1)$ e, portanto, tem uma distribuição $\chi^2(1)$, e o denominador é o quociente de uma v.a. $\chi^2(n-1)$ por $(n-1)$.

Vamos considerar dois casos.

(a) Mesma Variância Desconhecida

Suponha que, ao testar a hipótese de igualdade de variâncias, esta não seja rejeitada, isto é, $\sigma_1^2 = \sigma_2^2 = \sigma^2$, porém essa variância comum é desconhecida. Como S_1^2 e S_2^2 são dois estimadores não-viesados de σ^2 , podemos combiná-los para obter um estimador comum:

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^m (Y_i - \bar{Y})^2}{n+m-2} \quad 13.10$$

que também é um estimador não-viesado de σ^2 . Mais ainda, cada parcela do numerador de (13.10), quando dividida por σ^2 , terá distribuição qui-quadrado, com $(n-1)$ e $(m-1)$ graus de liberdade, respectivamente. Logo, teremos que

$$\frac{(n+m-2)S_p^2}{\sigma^2} \sim \chi^2(n+m-2) \quad 13.11$$

Pelo Teorema 7.1, a estatística

$$T = \frac{\frac{\bar{X} - \bar{Y}}{\sigma\sqrt{1/n + 1/m}}}{S_p/\sigma} = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{1/n + 1/m}} \quad 13.12$$

terá uma distribuição t de Student, com $(n+m-2)$ graus de liberdade, sob a hipótese H_0 , isto é, se $\mu_1 = \mu_2$.

Exemplo 13.4: Duas técnicas de venda são aplicadas por dois grupos de vendedores: a técnica A, por 12 vendedores, e a técnica B, por 15 vendedores. Espera-se que a técnica B produza melhores resultados. No final de um mês, obtiveram-se os resultados da Tabela 1.

Dados	Vendas Técnica A	Vendas Técnica B
Média	68	76
Variância	50	75
Vendedores	12	15

Tabela 1: Dados para duas técnicas de Vendas

Vamos testar, para o nível de significância de 5%, se há diferenças significativas entre as vendas resultantes das duas técnicas. Informações adicionais permitem supor que as vendas sejam normalmente distribuídas, com uma variância comum σ^2 , desconhecida.

As hipóteses a serem testadas ficam

$$H_0 : \mu_A = \mu_B$$

$$H_1 : \mu_A < \mu_B.$$

Pelas suposições acima, podemos usar a estatística (13.12), com $n = 12$, $m = 15$ e

$$S_p^2 = \frac{11S_A^2 + 14S_B^2}{25}.$$

Da Tabela V obtemos

$$RC =]1, 708; +\infty[.$$

Da Tabela 13.1 calculamos:

$$S_p^2 = \frac{11(50) + 14(75)}{25} = 64,$$

$$t_0 = \frac{76 - 68}{8\sqrt{1/12 + 1/15}} = 2,56.$$

Como $t_0 \in RC$, rejeitamos H_0 , ou seja, existe evidência de que a técnica B produz melhores resultados do que a técnica A.

Encontrada diferença entre os métodos, a continuação natural é construir um intervalo de confiança para a diferença $\Delta = \mu_B - \mu_A$. Do resultado (13.12) é fácil verificar que:

$$IC(\Delta; \gamma) = (\bar{x}_0 - \bar{y}_0) \pm t_\gamma S_p \sqrt{1/n + 1/m}$$

Para o nosso exemplo, com $\gamma = 0,95$, esse intervalo reduz-se a

$$\begin{aligned} IC(\Delta; 0,95) &= 8 \pm (2,06)(8) \sqrt{1/12 + 1/15} \\ &= 8 \pm 6,38 =]1,62; 14,38[. \end{aligned}$$

(b) Variâncias Desiguais, Desconhecidas

Quando a hipótese de igualdade de variâncias for rejeitada, devemos usar a estatística:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{S_1^2/n + S_2^2/m}} \quad 13.13$$

Pode-se provar que, sob a veracidade de H_0 , a v.a. T aproxima-se de uma distribuição t de Student, com o número de graus de liberdade dado aproximadamente por:

$$v = \frac{(A + B)^2}{A^2/(n-1) + B^2/(m-1)} \quad 13.14$$

na qual

$$A = s_1^2/n, \quad B = s_2^2/m.$$

Como esse valor é geralmente fracionário, arredonde para o inteiro mais próximo para obter o número de graus de liberdade.

Exemplo 13.5: Queremos testar as resistências de dois tipos de vigas de aço, A e B. Tomando-se $n = 15$ vigas do tipo A e $m = 20$ vigas do tipo B, obtemos os valores na Tabela 2. Usando um teste F com nível $\alpha = 10\%$ rejeitamos a hipótese de variâncias iguais.

Tipo	Média	Variância
A	70,5	81,6
B	84,3	161,5

Tabela 2: Médias e variâncias para dois tipos de vigas de aço

Consideremos as hipóteses:

$$H_0 : \mu_A = \mu_B$$

$$H_1 : \mu_A \neq \mu_B.$$

A estatística a ser usada é (13.13), com $v = (182,66)/(2,11 + 3,43) = 32,9$, logo tomamos $v = 33$. Com $\alpha = 0,05$, obtemos da Tabela V que $RA =]-2,0345; 2,0345[$. Com os dados da Tabela 2, temos $t_0 = (-13,8)/3,68 = -3,75$.

Como $t_0 \in RC$, rejeitamos H_0 , ou seja, há evidências de que os dois tipos de vigas têm resistências médias diferentes.

Problemas

5. Num estudo comparativo do tempo médio de adaptação, uma amostra aleatória, de 50 homens e 50 mulheres de um grande complexo industrial, produziu os seguintes resultados:

Estatística	Homens	Mulheres
Médias	3,2 anos	3,7 anos
Desvios padrões	0,8 anos	0,9 anos

Que conclusões você poderia tirar para a população de homens e mulheres dessa indústria?
(Indique as suposições feitas para resolver o problema.)

6. Diversas políticas em relação às filiais de uma rede de supermercados estão associadas ao gasto médio dos clientes em cada compra. Deseja-se comparar esse parâmetro para duas novas filiais, por meio de duas amostras de 50 clientes cada. As médias obtidas foram 62 e 71, respectivamente. Sabe-se que o desvio padrão, em ambos os casos, deve ser da ordem de 20 unidades. É possível afirmar que o gasto médio nas duas filiais seja o mesmo? Caso contrário, dê um intervalo de confiança para a diferença.

7. Uma fábrica de embalagens para produtos químicos está estudando dois processos para combater a corrosão de suas latas especiais. Para verificar o efeito dos tratamentos, foram usadas amostras cujos resultados estão no quadro abaixo (em porcentagem de corrosão eliminada). Qual seria a conclusão sobre os dois tratamentos?

Método	Amostra	Média	Desvio Padrão
A	15	48	10
B	12	52	15

8. No **Problema 4**, teste a hipótese de que as médias dos comprimentos do produto produzido pelas duas fábricas são iguais.

9. Para investigar a influência da opção profissional sobre o salário inicial de recém-formados, investigaram-se dois grupos de profissionais: um de liberais em geral e outro de formados em Administração de Empresas. Com os resultados abaixo, expressos em salários mínimos, quais seriam suas conclusões?

Liberais	6,6	10,3	10,8	12,9	9,2	12,3	7,0	
Administradores	8,1	9,8	8,7	10,0	10,2	8,2	8,7	10,1

Populações Não-Normais

Passamos, agora a descrever um teste que não faz suposições a respeito da forma das distribuições P_1 e P_2 , a não ser que as variáveis envolvidas tenham uma escala de medida pelo menos ordinal. Ou seja, podemos abordar o caso de variáveis qualitativas ordinais e variáveis quantitativas. Este teste (chamado de Wilcoxon ou de Mann-Whitney) pertence a uma categoria de procedimentos chamados não-paramétricos ou livres de distribuição.

Teremos para análise amostras independentes de duas populações e queremos testar a hipótese(13.1), isto é,

$$H_0 : P_1 = P_2,$$

contra a alternativa de que as distribuições diferem em localização: estaremos interessados em saber se uma população tende a ter valores maiores do que a outra, ou se elas têm a mesma mediana ou média.

O teste de Wilcoxon é baseado nos **postos** dos valores obtidos combinando-se as duas amostras. Isto é feito ordenando esses valores, do menor para a maior, independentemente do fato de qual população cada valor provém.

A estatística do teste é a soma dos **postos** associados aos valores amostrados de uma população, P_1 , por exemplo. Se essa soma for grande, isso é, uma indicação de que os valores desta população tendem a ser maiores do que os valores de P_2 , e, rejeitamos a hipótese nula (13.1).

No caso de termos uma v.a. qualitativa ordinal, comumente associamos números às diversas categorias (ou classes, ou atributos,) segundo as quais a variável é classificada. por exemplo, podemos ter 1 para bom, 2 para muito bom e 3 para ótimo. Vemos, então, que esses valores são os **postos**, nesse caso, e em outras situações é preferível trabalhar com postos do que valores arbitrários associados à variável qualitativa.

Quando trabalhamos com variáveis quantitativas podemos ter valores repetidos nas amostras. Veremos então como associar postos nesse caso. Para evitar esses **empates**, uma possibilidade é supor que a v.a. seja contínua, de modo que se X for uma tal variável

$$P(X = x_0) = 0.$$

Essa suposição é eventualmente necessária para o desenvolvimento teórico do teste, mas na prática quer X seja discreta ou contínua, valores repetidos poderão aparecer.

Vamos trabalhar inicialmente com observações distintas:

Suponha que tenhamos N observações Z_1, Z_2, \dots, Z_N . Ordenando-as da menor para a maior obtemos as estatísticas de ordem,

$$Z_{(1)} \leq Z_{(2)} \leq \dots \leq Z_{(N)}.$$

Inicialmente suponha que não haja observações coincidentes, de modo os sinais \leq são substituídos por $<$, isto é,

$$Z_{(1)} < Z_{(2)} < \dots < Z_{(N)}.$$

Então, associamos números (normalmente $1, 2, \dots, N$), chamados **postos**, que correspondem às posições das observações na ordenação. O posto de Z_i é igual a $1 + (\text{número de } Z_j < Z_i)$. Assim dadas as observações

$$Z_1 = 0,3 \ Z_2 = 1,5 \ Z_3 = -0,5, \ Z_4 = 2,$$

Ordenado do menor para o maior valor temos:

$$Z_{(1)} = -0,5 \ Z_{(2)} = 0,3 \ Z_{(3)} = 1,5, \ Z_{(4)} = 2.$$

Os postos de Z_1, Z_2, Z_3, Z_4 serão, respectivamente

$$R_1 = 2, R_2 = 3, R_3 = 1, R_4 = 4,$$

já que a ordenação resulta em

$$-0,5 < 0,3 < 1,5 < 2, \text{ ou } Z_3 < Z_1 < Z_2 < Z_4.$$

Vamos utilizar o pacote *R* para trabalhar este exemplo.

```
> ## Amostra
>
> Z=c(0.3,1.5,-0.5,2);Z
[1] 0.3 1.5 -0.5 2.0
> N=length(Z);N
[1] 4
>
> ## Amostra Ordenada do menor para o maior.
>
> Z_o=sort(Z);Z_o
[1] -0.5 0.3 1.5 2.0
>
>
>
> ### Atribuir os postos.
>
> Z_R=rank(Z);Z_R
[1] 2 3 1 4
>
>
> ####Posição dos postos no vetor original
Z_OP= order(Z);Z_OP
[1] 3 1 2 4
>
>
> tab=cbind(Z,Z_o,Z_R,Z_OP);tab
      Z    Z_o  R  Z_OP
[1,] 0.3 -0.5 2   3
```

```
[2,] 1.5 0.3 3 1
[3,] -0.5 1.5 1 2
[4,] 2.0 2.0 4 4
>
```

Exemplo 13.6

Num estudo sobre um novo método para ensinar Matemática Elementar, foram selecionadas 5 crianças. Destas, 3 são escolhidas ao acaso e ensinadas segundo o novo método, enquanto as outras duas funcionaram como controle e receberam instrução por um método tradicional.

Após um período de 5 semanas é feito um teste e as crianças são ordenadas segundo seu desempenho: a criança que tiver a menor nota recebe posto 1, etc., até a criança que tiver maior nota receber posto 5.

O método de ensino será considerado eficaz se as três crianças que recebem o novo método tiverem postos altos nesta ordenação combinada das cinco crianças.

Seja H_0 a hipótese nula que especifica que o tratamento (novo método) não tem efeito, isto é, a nota da criança não é afetada se ela for ou não ensinada pelo novo método. Se H_0 for verdadeira, o posto atribuído a cada criança é determinado somente pela sua inteligência, ou seja a ordenação das crianças não depende de qual recebe tratamento e qual funciona como controle.

A tabela 13.3 mostra todos os casos possíveis para a ordenação onde **C** indica controle e **T**, tratamento.

1	2	3	4	5	W_s
C	C	T	T	T	3+4+5=12
C	T	C	T	T	2+4+5=11
T	C	C	T	T	1+4+5=10
C	T	T	C	T	2+3+5=10
T	C	T	C	T	1+3+5=9
C	T	T	T	C	2+3+4=9
T	C	T	T	C	1+3+4=8
T	T	C	T	C	1+2+4=7
T	T	T	C	C	1+2+3=6
T	T	C	C	T	1+2+5=8

Vemos que as crianças e seus postos podem ser divididos em dois grupos (tratados e controles) de

$$\binom{5}{3} = 10$$

maneiras diferentes. A suposição de que as três crianças recebendo o tratamento são selecionadas ao acaso e de que os métodos são equivalentes, implica que todas as dez possibilidades têm a mesma probabilidade $\frac{1}{10}$.

Consideremos a estatística

$$W_s = S_1 + S_2 + S_3$$

em que S_1, S_2 e S_3 são os postos das crianças que receberam o tratamento na amostra combinada.

Poderíamos considerar com regra de decisão para rejeitar H_0 a ocorrência de $W_s = 12$, correspondendo a ocorrência de $CCTT$, clara superioridade do tratamento. Qual a probabilidade de que esse evento ocorreu por mero acaso, ou seja, quando os dois métodos são equivalentes?

Nesse caso teremos

$$P(W_s = 12 \mid H_0 \text{ verdadeira}) = 0,10,$$

que é a probabilidade do erro do tipo I, ou seja o nível de significância do teste. mas, como vimos antes, usualmente procedemos de maneira oposta , ou seja, fixamos α e não a regra de decisão.

Como vimos acima, rejeitamos H_0 para valores grandes de W_s ou seja $W_s \geq c$, em que c é uma constante determinada a partir do nível de significância do teste, α .

Obtemos o teste de Wilcoxon:

Rejeite H_0 se $W_s \geq c$ em que c é determinada por

$$P(W_s \geq c \mid H_0 \text{ verdadeira}) = \alpha.$$

A distribuição nula (isto é, sob H_0) de W_s é obtida da tabela 13.3 e está na tabela 13.4.

w	6	7	8	9	10	11	12
$P(W_s = w)$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

A distribuição de W_s é simétrica ao redor do valor 9 que, como veremos, representa a média de W_s que é dada por:

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

Com $N = n + m$ e m é o tamanho do grupo tratamento e n é o tamanho do grupo controle.

A acumulada de W_s é dada por:

w	6	7	8	9	10	11	12
$P(W_s \leq w)$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{4}{10}$	$\frac{6}{10}$	$\frac{8}{10}$	$\frac{9}{10}$	1

Note que

$$P(W_s \leq 8) = 0,4 \quad e \quad P(W_s \leq 9) = 0,6$$

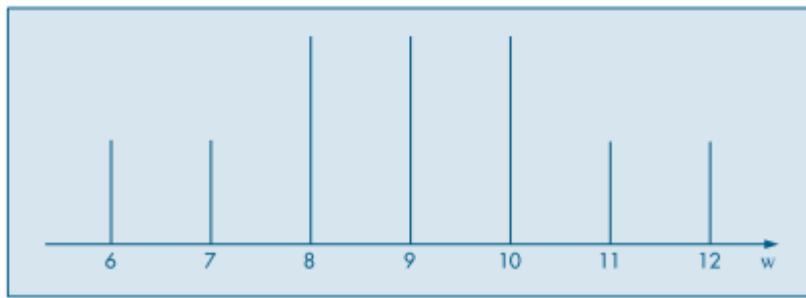
Note ainda:

$$P(W_s \leq 8) < 0,5 \quad e \quad P(W_s \geq 9) > 0,5;$$

assim $Md = 9$ é a mediana de W_s .

Vamos representar graficamente a lei de probabilidade de W_s :

Figura 13.3: Distribuição de W_s para o Exemplo 13.6.



Da tabela 13.4 temos:

$$E(W_s) = \frac{n(N+1)}{2} = \frac{3(5+1)}{2} = 9.$$

Vamos calcular diretamente da tabela 13.4:

$$E(W_s) = \frac{6}{10} + \frac{7}{10} + \frac{16}{10} + \frac{18}{10} + \frac{20}{10} + \frac{11}{10} + \frac{12}{10} = \frac{90}{10} = 9.$$

Usando a fórmula dada temos:

A variância de W_s é dada por:

$$Var(W_s) = \frac{nm(N+1)}{12}.$$

Vamos calcular diretamente da tabela 13.4:

$$Var(W_s) = E[((W_s - 9)^2)] = \frac{9}{10} + \frac{4}{10} + \frac{2}{10} + 0 + \frac{2}{10} + \frac{4}{10} + \frac{9}{10} = \frac{30}{10} = 3.$$

Poderíamos ter calculado:

$$E(W_s^2) = \frac{36}{10} + \frac{49}{10} + \frac{128}{10} + \frac{162}{10} + \frac{200}{10} + \frac{121}{10} + \frac{144}{10} = \frac{840}{10} = 84.$$

$$Var(W_s) = E(W_s^2) - E^2(W_s) = 84 - 81 = 3.$$

Usando a fórmula temos :

$$Var(W_s) = \frac{nm(N+1)}{12} = \frac{3 \times 2 \times 6}{12} = 3.$$

Se por exemplo, $\alpha = 0,05$, não existe valor satisfazendo

$$P(W_s \geq c | H_0 \text{ verdadeira}) = 0,05.$$

Podemos encontrar c para valores de $\alpha = 0,1; 0,2; 0,4$ etc. Por exemplo, se $\alpha = 0,1$, então

$$P(W_s \geq 12) = 0,1 \quad \text{e} \quad c = 12.$$

Vamos supor que a nossa amostra foi:

$$CTCTT$$

que tem

$$W_s = 2 + 4 + 5 = 11.$$

O nível descritivo do teste será dado por:

$$nd = P(W_s \geq 11) = \frac{2}{10} = 0,2.$$

Consideremos agora a situação geral. Queremos testar (13.1)

$$H_0 : P_1 = P_2.$$

Temos duas amostras independentes X_1, X_2, \dots, X_n de P_1 e Y_1, Y_2, \dots, Y_m de P_2 .

Seja $N = n + m$ e combinamos as duas amostras numa só, ordenamos os N valores do menor para o maior e chamemos de

$$S_1 < S_2 < \dots < S_m,$$

os postos dos Y_i (tratamentos) e

$$R_1 < R_2 < \dots < R_n,$$

os postos dos X_i (controles). Estamos supondo que não haja empates. Seja

$$W_s = \sum_{i=1}^m S_i, \tag{13.16}$$

a soma dos postos dos tratamentos. Rejeitamos H_0 se

$$W_s \geq c.$$

No caso bilateral, rejeite H_0 se

$$W_s < c_1 \text{ ou } W_s > c_2,$$

para dado α .

Não é difícil verificar que, se a distribuição de P_1 for contínua, então

$$P(S_1 = s_1, \dots, S_m = s_m) = \frac{1}{\binom{N}{m}}, \tag{13.17}$$

em que $s_1 < s_2 < \dots < s_m$ e $s_i \in \{1, 2, \dots, N\}$, $N = n + m$.

Observação: Por (13.17) vemos que a distribuição dos postos e portanto de W_s não depende de P_1 . Isso não ocorrerá se P_1 não for contínua. Se as distribuições de P_1 e P_2 forem contínuas, há ausência de empates (isto é, coincidência entre valores de X e Y).

Isso significa que poderíamos considerar nossas medidas de X e Y de tal sorte que coincidências deveriam ser evitadas. Na prática, contudo, as medidas são feitas em geral com o mesmo número de casas decimais, de modo que empates podem ocorrer. Essa situação é analisada abaixo.

A distribuição sob H_0 de W_s pode ser encontrada como no exemplo 13.6. Para dado valor de w , verificamos quantas amostras de tamanho m , retiradas de $P = \{1, 2, \dots, N\}$ fornecem o valor w . Se $\#(w; n, m)$ indicar esse número, então, por (13.17)

$$P(W_s = w \mid H_0 \text{ verdadeira}) = \frac{\#(w; n, m)}{\binom{N}{m}}, \quad (13.18)$$

Pode-se provar o seguinte resultado (veja por exemplo, Lehmann, 1975):

Teorema 13.1 Para a estatística W_s temos:

$$E(W_s) = \frac{m(N+1)}{2} \quad (13.19),$$

$$Var(W_s) = \frac{nm(N+1)}{12} \quad (13.20).$$

Vamos agora mostrar algumas propriedades de W_s :

a. O valor mínimo de W_s ocorre quando os postos dos m tratamentos $1, 2, \dots, m$ e portanto

$$W_{min} = 1 + 2 + \dots + m = \frac{m(m+1)}{2}.$$

b. O valor máximo de W_s ocorre quando os postos dos m tratamentos $n+1, n+2, \dots, n+m$ e portanto

$$W_{max} = (n+1) + (n+2) + \dots + (n+m) = mn + \frac{m(m+1)}{2}.$$

c. Podemos representar W_s como:

$$W_s = \sum_{i=1}^N i I_i,$$

em que $I_i = 1$ se o elemento com posto i pertence ao grupo tratamento e $I_i = 0$ caso contrário, $i = 1, 2, \dots, N$.

Só para exemplificar considere o exemplo 13.6 e a amostra *CCTTT*.

Aqui temos $m = 3$, $n = 2$ e $N = 5$. Logo os indivíduos com postos 3, 4 e 5 pertencem a nossa amostra.

$$W_s = \sum_{i=1}^5 i I_i = 1 \times I_1 + 2 \times I_2 + 3 \times I_3 + 4 \times I_4 + 5 \times I_5,$$

$$W_s = 1 \times 0 + 2 \times 0 + 3 \times 1 + 4 \times 1 + 5 \times 1 = 3 + 4 = 5 = 12.$$

d. $I_i, i = 1, 2, \dots, N$ tem distribuição de Bernoulli de parâmetro $p = \frac{m}{N}$.

Prova: O elemento com posto i pertence a nossa amostra. Vamos fixá-lo e escolher os $(m-1)$ restantes do grupo dos tratamentos dentre os $(N-1)$ remanescentes. Isto pode ser feita de

$$\binom{N-1}{m-1},$$

maneiras. Logo

$$p = P(I_1 = 1) = \frac{\binom{N-1}{m-1}}{\binom{N}{m}} = \frac{m}{N}.$$

A média de I_i é dada por:

$$E(I_i) = \frac{m}{N},$$

e sua variância por:

$$Var(I_i) = p(1-p) = \frac{m}{N} \times \frac{N-m}{N} = \frac{mn}{N^2}.$$

e. $E(W_s) = \frac{n(N+1)}{2}$.

Prova

$$\begin{aligned}
 E(W_s) &= E\left(\sum_{i=1}^N i I_i\right) \\
 &= \sum_{i=1}^N E(i I_i) \\
 &= \sum_{i=1}^N i E(I_i) \\
 &= \sum_{i=1}^N i \frac{m}{N} \\
 &= \frac{m}{N} \sum_{i=1}^N i \\
 &= \frac{m}{N} \frac{N(N+1)}{2} \\
 &= \frac{m(N+1)}{2}.
 \end{aligned}$$

f. $I_i \times I_j, i, j = 1, 2, \dots, N, j \neq i$ tem distribuição de Bernoulli de parâmetro $p = \frac{m(m-1)}{N(N-1)}$.

Prova Os elementos com postos i e j pertencem a nossa amostra. Vamos fixá-los e escolher os $(m-2)$ restantes do grupo dos tratamentos dentre os $(N-2)$ remanescentes. Isto pode ser feita de

$$\binom{N-2}{m-2},$$

maneiras. Logo

$$p = P(I_1 \times I_j = 1) = P(I_1 = 1, I_j = 1) = \frac{\binom{N-2}{m-2}}{\binom{N}{m}} = \frac{m(m-1)}{N(N-1)}.$$

A média de $I_i \times I_j$ é dada por:

$$E(I_i \times I_j) = \frac{m(m-1)}{N(N-1)}.$$

g. A covariância entre I_i e I_j , $i \neq j$ é dada por:

$$COV(I_i, I_j) = -\frac{mn}{N^2(N-1)}.$$

Prova

$$\begin{aligned}
 COV(I_i, I_j) &= E(I_i I_j) - E(I_i)E(I_j) \\
 &= \frac{m(m-1)}{N(N-1)} - \frac{m^2}{N^2} \\
 &= \frac{m}{N} \left[\frac{m-1}{N-1} - \frac{m}{N} \right] \\
 &= \frac{m}{N} \left[\frac{Nm - N - Nm + m}{(N-1)N} \right] \\
 &= -\frac{m}{N} \frac{N-m}{N(N-1)} \\
 &= -\frac{mn}{N^2(N-1)}.
 \end{aligned}$$

h.

$$Var(W_s) = \frac{nm(N+1)}{12}.$$

Prova

$$\begin{aligned}
 Var(W_s) &= Var \left(\sum_{i=1}^N i I_i \right) \\
 &= \sum_{i=1}^N Var(i I_i) + \sum_{i=1}^N \sum_{j=1, j \neq i}^N Cov(i I_i, j I_j) \\
 &= \sum_{i=1}^N i^2 Var(I_i) + \sum_{i=1}^N \sum_{j=1, j \neq i}^N ij Cov(I_i, I_j) \\
 &= Soma_1 + Soma_2.
 \end{aligned}$$

Mas,

$$\begin{aligned}
 Soma_1 &= \sum_{i=1}^N Var(i I_i) \\
 &= \sum_{i=1}^N i^2 Var(I_i) \\
 &= \sum_{i=1}^N i^2 \frac{mn}{N^2} \\
 &= \frac{mn}{N^2} \sum_{i=1}^N i^2 \\
 &= \frac{mn}{N^2} \frac{N(N+1)(2N+1)}{6} \\
 &= \frac{mn(N+1)(2N+1)}{6N}.
 \end{aligned}$$

$$\begin{aligned}
 Soma_2 &= \sum_{i=1}^N \sum_{j=1, j \neq i}^N ij \operatorname{Cov}(I_i, I_j) \\
 &= - \sum_{i=1}^N \sum_{j=1, j \neq i}^N ij \frac{mn}{N^2(N-1)} \\
 &= - \frac{mn}{N^2(N-1)} \sum_{i=1}^N i \sum_{j=1, j \neq i}^N j \\
 &= - \frac{mn}{N^2(N-1)} \sum_{i=1}^N i \left[\sum_{j=1}^N j - i \right] \\
 &= - \frac{mn}{N^2(N-1)} \sum_{i=1}^N i \left[\frac{N(N+1)}{2} - i \right] \\
 &= - \frac{mn}{N^2(N-1)} \left[\frac{N(N+1)}{2} \sum_{i=1}^N i - \sum_{i=1}^N i^2 \right] \\
 &= - \frac{mn}{N^2(N-1)} \left[\frac{N^2(N+1)^2}{4} - \frac{N(N+1)(2N+1)}{6} \right] \\
 &= - \frac{mn(N+1)}{N(N-1)} \left[\frac{N(N+1)}{4} - \frac{2N+1}{6} \right] \\
 &= - \frac{mn(N+1)}{N(N-1)} \frac{(3N^2 - N - 2)}{12} \\
 &= - \frac{mn(N+1)}{N(N-1)} \frac{(N-1)(3N+2)}{12} \\
 &= - \frac{mn(N+1)(3N+2)}{12N}.
 \end{aligned}$$

Finalmente,

$$\begin{aligned}
 Var(W_s) &= Soma_1 + Soma_2 \\
 &= \frac{mn(N+1)(2N+1)}{6N} - \frac{mn(N+1)(3N+2)}{12N} \\
 &= \frac{mn(N+1)}{N} \left[\frac{2N+1}{6} - \frac{3N+2}{12} \right] \\
 &= \frac{mn(N+1)}{N} \left[\frac{4N+2 - 3N-2}{12} \right] \\
 &= \frac{mn(N+1)}{12}.
 \end{aligned}$$

- i. A distribuição de W_s pode ser aproximada pela distribuição normal quando $n, m \rightarrow \infty$, a v.a.

$$Z = \frac{W_s - E(W_s)}{\sqrt{Var(W_s)}},$$

tem uma distribuição normal $N(0, 1)$.

Uma estatística equivalente a W_s é

$$U_s = W_s - \frac{m(m+1)}{2},$$

chamada de Mann-Whitney. Há duas vantagens em se usar U_s :

- a. A distribuição de U_s para $n = n_1$ e $m = m_1$ é a mesma que a distribuição de U_s quando os tamanhos são invertidos, isto é, quando $n = m_1$ e $m = n_1$. Isto não acontece com W_s .
- b. O valor mínimo de U_s é zero, para quaisquer valores de m e n simplificando a construção de tabelas. A tabela VIII do apêndice dá os valores $P(U_s \leq u)$.

Teorema 13.2 A média e a variância de U_s são dada por:

$$E(U_s) = \frac{nm}{2} \tag{13.23},$$

$$Var(U_s) = \frac{nm(N+1)}{12} \tag{13.24},$$

respectivamente. Além disso, a distribuição de U_s pode também ser aproximada por uma normal.

A distribuição U aparece no R com o nome de Distribuição da Soma de Postos de Wilcoxon. Com o comando `?dwilcox` entre no R e estude.

Vamos apresentar no **R**, a lei de probabilidade, a Função de distribuição, a média e a variância de U do exemplo 13.6:

```
> ##### A distribuição de Wilcoxon no R.
>
> ### Na realidade traz a distribuição U de Mann-Whitney
> #####Suporte A={0,1,...,mn}
>
>
> #####?dwilcox
>
>
> m=3;n=2;N=m+n;N
[1] 5
> u=0:(m*n)
>
```

```

>
> f_U= dwilcox(u,m,n)
> F_U=pwilcox(u,m,n)
>
> EU= (m*n)/2; EU
[1] 3
> VU=(m*n*(N+1))/12; VU
[1] 3
>
> Tab_1=cbind(u,f_U,F_U);Tab_1
  u f_U F_U
[1,] 0 0.1 0.1
[2,] 1 0.1 0.2
[3,] 2 0.2 0.4
[4,] 3 0.2 0.6
[5,] 4 0.2 0.8
[6,] 5 0.1 0.9
[7,] 6 0.1 1.0
>

```

É fácil usando o R apresentar a distribuição W de Wilcoxon

$$W = U + \frac{m * (m + 1)}{2} = U + 6.$$

Logo,

```

## Distribuição W de Wilcoxon
>
> w=(m*(m+1))/2 +u
> f_W=f_U
> F_W=F_U
>
> EW=(m*(N+1))/2;EW
[1] 9
> VW=(m*n*(N+1))/12; VW
[1] 3
> Tab_2=cbind(w,f_W,F_W);Tab_2
  w f_W F_W
[1,] 6 0.1 0.1
[2,] 7 0.1 0.2
[3,] 8 0.2 0.4
[4,] 9 0.2 0.6
[5,] 10 0.2 0.8
[6,] 11 0.1 0.9
[7,] 12 0.1 1.0
>
>

```

Exemplo 13.7 Suponha que $m = n = 10$ e queremos calcular $P(W_s \leq 87)$. O valor tabelado é 0,0952, que é encontrado na tabela VIII com $m = n = 10$ e levando em conta que:

$$U_s = 87 - \frac{10 \times 11}{2} = 87 - 55 = 32,$$

e portanto

$$P(U_s \leq 32) = 0,0952.$$

Por outro lado usando a aproximação normal temos:

Note que $N = 10 + 10 = 20$.

$$E(W_s) = \frac{10 \times 21}{2} = 105,$$

e

$$Var(W_s) = \frac{10 \times 10 \times 21}{12} = 175.$$

Logo

$$P(W_s \leq 87) \approx P(Z \leq \frac{87 - 105}{\sqrt{175}}) = \Phi(-1,36) \approx 0,087,$$

que está bem próximo do valor encontrado na tabela.

A aproximação pode ser melhorada usando-se a correção de continuidade pois estamos aproximando a distribuição de uma variável discreta W_s por uma distribuição de uma variável contínua (normal).

Assim:

$$P(W_s \leq 87) \approx P(W_s \leq 87,5) = P(Z \leq \frac{87,5 - 105}{\sqrt{175}}) = \Phi(-1,32) \approx 0,0934.$$

Vamos fazer no R.

```
> ###Exemplo 13.7
>
> m=10;n=10
> N=m+n;N
[1] 20
> w=87
> a=m*(m+1)/2;a
[1] 55
> u=w-a;u
[1] 32
>
> ###P=P(W_s<=87)=P(U<=32)
>
> p=pwilcox(32,m,n);p;round(p,4)
[1] 0.09515794
[1] 0.0952
>
>
> #####Aproximação para Normal
>
>
> EW=(m*(N+1))/2;EW
[1] 105
```

```
> VW=(m*n*(N+1))/12; VW
[1] 175
> DW=sqrt(VW);DW
[1] 13.22876
>
> z=(87-EW)/DW;z;round(z,2)
[1] -1.360672
[1] -1.36
>
> pa=pnorm(z);round(pa,3)
[1] 0.087
>
>
> #####Probabilidade aproximada com fator de correção de continuidade.
>
> zc=(87.5-EW)/DW;z;round(z,2);zc;round(zc,2)
[1] -1.360672
[1] -1.36
[1] -1.322876
[1] -1.32

>pnorm(-1.32);round(pnorm(-1.32),4)
[1] 0.09341751
[1] 0.0934
>
```

Vamos apresentar a tabela VIII do livro que traz a distribuição de U para alguns valores de m e n .

		Tabela VIII – Distribuição de Mann-Whitney U_s O corpo da tabela dá as probabilidades $P(U_s \leq u)$.							
m	u	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n = 7$	$n = 8$	$n = 9$	$n = 10$
3	0	0500	0286	0179	0119	0083	0061	0045	0035
	1	1000	0571	0357	0238	0167	0121	0091	0070
	2	2000	1143	0714	0476	0333	0242	0182	0140
	3	3500	2000	1250	0833	0583	0424	0318	0245
	4	5000	3143	1964	1319	0917	0667	0500	0385
	5	6500	4286	2857	1905	1333	0970	0727	0559
	6	8000	5714	3929	2738	1917	1394	1045	0804
	7	9000	6857	5000	3571	2583	1879	1409	1084
	8	9500	8000	6071	4524	3333	2485	1864	1434
	9	1,0000	8857	7143	5476	4167	3152	2409	1853
	10		9429	8036	6429	5000	3879	3000	2343
	11		9714	8750	7262	5833	4606	3636	2867
	12		1,0000	9286	8095	6667	5394	4318	3462
	13			9643	8690	7417	6121	5000	4056
	14			9821	9167	8083	6848	5682	4685
	15				1,0000	9524	8667	7515	6364
4	0		0143	0079	0048	0030	0020	0014	0010
	1		0286	0159	0095	0061	0040	0028	0020
	2		0571	0317	0190	0121	0081	0056	0040
	3		1000	0556	0333	0212	0141	0098	0070
	4		1714	0952	0571	0364	0242	0168	0120
	5		2429	1429	0857	0545	0364	0252	0180
	6		3429	2063	1286	0818	0545	0378	0270
	7		4429	2778	1762	1152	0768	0531	0380
	8		5571	3651	2381	1576	1071	0741	0529
	9		6571	4524	3048	2061	1414	0993	0709
	10		7571	5476	3810	2536	1838	1301	0939
	11		8286	6349	4571	3242	2303	1650	1199
	12		9000	7222	5429	3939	2848	2070	1518
	13		9429	7937	6190	4636	3414	2517	1868
	14		9714	8571	6952	5364	4040	3021	2268
	15		9857	9048	7619	6061	4667	3552	2697
	16		1,0000	9444	8238	6758	5333	4126	3177
	17			9683	8714	7364	5960	4699	3666
	18			9841	9143	7939	6586	5301	4196
	19			9921	9429	8424	7152	5874	4725
	20				1,0000	9667	8848	7607	6148

Obs.: Todas as entradas (com exceção de 1,0000) devem ser precedidas de 0.

Figura 1:

Tabela VIII – Distribuição de Mann-Whitney (continuação)													
<i>m</i>	<i>u</i>	<i>n</i> = 5	<i>n</i> = 6	<i>n</i> = 7	<i>n</i> = 8	<i>n</i> = 9	<i>n</i> = 10	<i>m</i>	<i>u</i>	<i>n</i> = 7	<i>n</i> = 8	<i>n</i> = 9	<i>n</i> = 10
5	0	0040	0022	0013	0008	0005	0003	7	0	0003	0002	0001	0001
1		0079	0043	0025	0016	0010	0007	1		0006	0003	0002	0001
2		0159	0087	0051	0031	0020	0013	2		0012	0006	0003	0002
3		0278	0152	0088	0054	0035	0023	3		0020	0011	0006	0004
4		0476	0260	0152	0093	0060	0040	4		0035	0019	0010	0006
5		0754	0411	0240	0148	0095	0063	5		0055	0030	0017	0010
6		1111	0628	0356	0225	0145	0097	6		0087	0047	0026	0015
7		1548	0887	0530	0326	0210	0140	7		0131	0070	0039	0023
8		2103	1234	0745	0466	0300	0200	8		0189	0103	0058	0034
9		2738	1645	1010	0637	0415	0276	9		0265	0145	0082	0048
10		3452	2143	1338	0855	0559	0376	10		0364	0200	0115	0068
11		4206	2684	1717	1111	0734	0496	11		0487	0270	0156	0093
12		5000	3312	2159	1422	0949	0646	12		0641	0361	0209	0125
13		5794	3961	2652	1772	1199	0823	13		0825	0469	0274	0165
14		6548	4654	3194	2176	1489	1032	14		1043	0603	0356	0215
15		7262	5346	3775	2618	1818	1272	15		1297	0760	0454	0277
16		7897	6039	4381	3108	2188	1548	16		1588	0946	0571	0351
17		8452	6688	5000	3621	2592	1855	17		1914	1159	0708	0439
18		8889	7316	5619	4165	3032	2198	18		2279	1405	0879	0544
19		9246	7857	6225	4716	3497	2567	19		2675	1678	1052	0665
20		9524	8355	6806	5284	3986	2970	20		3100	1984	1261	0806
								21		3552	2317	1496	0976
6	0	0011	0006	0003	0002	0001		22		4024	2679	1755	1148
1		0022	0012	0007	0004	0002		23		4508	3063	2039	1349
2		0043	0023	0013	0006	0005		24		5000	3472	2349	1574
3		0076	0041	0023	0014	0009		25		5492	3894	2680	1819
4		0130	0070	0040	0024	0015		26		5976	4333	3032	2087
5		0206	0111	0063	0038	0024		27		6448	4775	3403	2374
6		0325	0175	0100	0060	0037		28		6900	5225	3788	2681
7		0465	0256	0147	0088	0055		29		7325	5667	4185	3004
8		0660	0367	0213	0128	0080		30		7721	6106	4591	3345
9		0898	0507	0296	0180	0112							
10		1201	0688	0406	0248	0156							
11		1548	0903	0539	0332	0210							
12		1970	1171	0709	0440	0280							
13		2424	1474	0906	0567	0363							
14		2944	1830	1142	0723	0467							
15		3496	2226	1412	0905	0589							
16		4091	2669	1725	1119	0736							
17		4686	3141	2068	1361	0903							
18		5314	3654	2454	1638	1999							
19		5909	4178	2864	1942	1317							
20		6504	4726	3310	2280	1566							
21		7056	5274	3773	2643	1838							
22		7576	5822	4259	3035	2139							
23		8030	6346	4749	3445	2461							
24		8452	6859	5251	3878	2811							
25		8799	7331	5741	4320	3177							
26		9102	7774	6227	4773	3564							
27		9340	8170	6690	5227	3962							
28		9535	8526	7136	5680	4374							
29		9675	8829	7546	6122	4789							
30		9794	9097	7932	6555	5211							

Figura 2:

Tabela VIII – Distribuição de Mann-Whitney (continuação)

<i>m</i>	<i>u</i>	<i>n</i> = 8	<i>n</i> = 9	<i>n</i> = 10	<i>m</i>	<i>u</i>	<i>n</i> = 9	<i>n</i> = 10	<i>m</i>	<i>u</i>	<i>n</i> = 10
8	0	0001	0000	0000	9	0	0000	0000	10	0	0000
1		0002	0001	0000	1		0000	0000	1		0000
2		0003	0002	0001	2		0001	0000	2		0000
3		0005	0003	0002	3		0001	0001	3		0000
4		0009	0005	0003	4		0002	0001	4		0001
5		0015	0008	0004	5		0004	0002	5		0001
6		0023	0012	0007	6		0006	0003	6		0002
7		0035	0019	0010	7		0009	0005	7		0002
8		0052	0028	0015	8		0014	0007	8		0004
9		0074	0039	0022	9		0020	0011	9		0005
10		0103	0056	0031	10		0028	0015	10		0008
11		0141	0076	0043	11		0039	0021	11		0010
12		0190	0103	0058	12		0053	0028	12		0014
13		0249	0137	0078	13		0071	0038	13		0019
14		0325	0180	0103	14		0094	0051	14		0026
15		0415	0232	0133	15		0122	0066	15		0034
16		0524	0296	0171	16		0157	0086	16		0045
17		0653	0372	0217	17		0200	0110	17		0057
18		0803	0464	0273	18		0252	0140	18		0073
19		0974	0570	0338	19		0313	0175	19		0093
20		1172	0694	0416	20		0385	0217	20		0116
21		1393	0836	0506	21		0470	0267	21		0144
22		1641	0998	0610	22		0567	0326	22		0177
23		1911	1179	0729	23		0680	0394	23		0216
24		2209	1383	0864	24		0807	0474	24		0262
25		2527	1606	1015	25		0951	0564	25		0315
26		2869	1852	1185	26		1112	0667	26		0376
27		3227	2117	1371	27		1290	0782	27		0446
28		3605	2404	1577	28		1487	0912	28		0526
29		3992	2707	1800	29		1701	1055	29		0615
30		4392	3029	2041	30		1933	1214	30		0716
31		4796	3365	2299	31		2181	1388	31		0827
32		5204	3715	2574	32		2447	1577	32		0952
33		5608	4074	2863	33		2729	1781	33		1088
34		6008	4442	3167	34		3024	2001	34		1237
35		6395	4813	3482	35		3332	2235	35		1399
36		6773	5187	3809	36		3652	2483	36		1575
37		7131	5558	4143	37		3981	2745	37		1763
38		7473	5926	4484	38		4317	3019	38		1965
39		7791	6285	4827	39		4657	3304	39		2179
40		8089	6635	5173	40		5000	3598	40		2406

Figura 3:

(b). Observações Não Todas Distintas. Consideremos, agora, a situação em que haja observações coincidentes, ou empates. Suponha, por exemplo, que $n = 3$ e $m = 2$ e as observações são:

$$1,3 ; 1,5 ; 1,5 ; 2,2 ; 2,5.$$

Nesse caso, usamos postos médios. Associamos o posto 1 à observação 1,3; às duas observações empatadas 1,5 associamos a média dos postos 2 e 3, que seriam atribuídas se as observações fossem distintas, ou seja, atribuímos o posto $(2+3)/2=2,5$; à observação 2,1 atribuímos o posto 4 e à observação 2,5 atribuímos o posto 5.

Veja como ficaria no R:

```
> x=c(1.3,1.5,1.5,2.2,2.5)
> x_rank=rank(x)
> tab=rbind(x,x_rank);tab
      [,1] [,2] [,3] [,4] [,5]
x     1.3  1.5  1.5  2.2  2.5
x_rank 1.0  2.5  2.5  4.0  5.0
>
```

Embora a atribuição de postos seja diferente nesse caso, continuaremos a usar a mesma notação anterior para os postos das observações X_i e Y_i . A distribuição da estatística W_s não é dada mais por (13.17), pois os valores de S_1, \dots, S_m não são mais os anteriores. Retomemos o exemplo dado. Temos que a distribuição conjunta dos postos S_1 e S_2 será:

$$P(S_1 = 1, S_2 = 2, 5) = 0,2 ; \quad P(S_1 = 1, S_2 = 4) = 0,1,$$

$$P(S_1 = 1, S_2 = 5) = 0,1 ; \quad P(S_1 = 2, 5, S_2 = 2, 5) = 0,1,$$

$$P(S_1 = 2, 5, S_2 = 4) = 0,2 ; \quad P(S_1 = 1, S_2 = 5) = 0,2,$$

$$P(S_1 = 4, S_2 = 5) = 0,1,$$

pois ainda cada uma das $\binom{5}{2}$ escolhas de dois dos postos médios como S_1 e S_2 são igualmente prováveis. Portanto a distribuição de $W_s = S_1 + S_2$ é dada pela tabela 13.5.

w	3,5	5,0	6,0	6,5	7,5	9
$P(W_s = w)$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{1}{10}$

Observe que a distribuição da v.a. W_s não é simétrica. Será simétrica em torno de $\frac{m(N+1)}{2}$ se $n = m$.

Mostre que:

$$E(W_s) = 6 \quad e \quad \text{Var}(W_s) = 2,85.$$

```

> w=c(3.5,5,6,6.5,7.5,9);w
[1] 3.5 5.0 6.0 6.5 7.5 9.0
> pw=c(2,2,1,2,2,1)/10;pw
[1] 0.2 0.2 0.1 0.2 0.2 0.1
> sum(pw)
[1] 1
>
> EW_s=sum(w*pw);EW_s
[1] 6
>
> m=2;n=3;N=m+n;N
[1] 5
>
> mu=m*(N+1)/2;mu
[1] 6
>
>
> EW2_s=sum(w^2*pw);EW2_s
[1] 38.85
>
>
> VarW_s=EW2_s-(EW_s)^2;VarW_s
[1] 2.85
>

```

Genericamente, o teste de Wilcoxon, no caso de observações empatadas, rejeita H_0 usando a mesma regra de decisão que no caso de observações não empatadas, exceto que a distribuição de W_s vai depender de m , n e dos números de observações empatadas em cada valor, ao contrário da situação de não empates, para a qual a distribuição de W_s depende somente de m e n .

Exemplo 13.8

Supondo $n = 3$ e $m = 2$, as observações dos controles são $1, 3; 1, 5; 2, 1$ e as observações dos tratamentos são $1,5$ e $2,5$. Então,

$$S_1 = 2, 5 ; S_2 = 5 ; 1, 5 ; R_1 = 1, R_2 = 2, 5 ; R_3 = 4,$$

e

$$W_s = S_1 + S_2 = 7, 5.$$

Pelo que vimos acima, o valor-p será:

$$\hat{\alpha} = P(W_s \geq 7, 5) = 0, 3,$$

logo não rejeitamos H_0 nos níveis usuais.

Suponha que d_1 observações empatadas no menor valor, d_2 observações empatadas no segundo menor valor etc. até d_e observações empatadas no maior valor, onde e é o número de valores distintos.

Denominamos $(e; d_1, d_2, \dots, d_e)$ de configuração de empates, e a distribuição de W_s dependerá dessa configuração.

Assim sendo, tabelas teriam de ser construídas para cada configuração de empates, o que não é prático. O que se faz é o seguinte: se o número de empates for pequeno continue a usar a Tabela VIII. Caso contrário use a aproximação normal. Nesse caso, a média de W_s é a mesma anterior, mas a variância é igual a anterior menos uma correção devida aos empates:

$$Var(W_s) = \frac{mn(N+1)}{12} - \frac{mn}{12N(N-1)} \sum_{i=1}^e (d_i^3 - d_i). \quad (13.25)$$

A aproximação normal será adequada se m e n forem relativamente grandes, e as proporções d_i/N não forem próximas de 1.

Para facilitar o cálculo da variância com empates vamos usar:

$$Var(W_s) = aux_1 - aux_2 * aux_3,$$

com

$$aux_1 = \frac{mn(N+1)}{12}, \quad aux_2 = \frac{mn}{12N(N-1)}, \quad aux_3 = \sum_{i=1}^e (d_i^3 - d_i).$$

Vamos calcular usando o *R*:

```
> ###No caso de empates a média não muda!!!!
>
> mu=m*(N+1)/2;mu
[1] 6
> #####Variância com empates
>
> aux_1=(m*n*(N+1))/12;aux_1
[1] 3
>
> aux_2=(m*n)/(12*N*(N-1));aux_2
[1] 0.025
>
> ## Na amostra temos 4 valores distintos= 1,3;1,5;2,1;2,5.. Assim temos duas observações empata
>
> d_1=1;d_2=2,d_3=1;d_4=1
>
e=c(d_1,d2,d_3,d_4)
>
> aux_3= sum(e^3-e);aux_3
[1] 6
>
> sigma2=aux_1-aux_2*aux_3;sigma2
[1] 2.85
>
```

Exemplo 13.9 Em aparelhos dentários são usados grampos de dois tipos: um modelo em **T** e outro circunferencial, **C**. O objetivo é verificar se a resistência à remoção dos grampos em **T** é a mesma do modelo **C**. Foram usados 40 corpos de provas (dente-grampo), sendo 20 para o modelo *T* e 20 para o modelo *C*, com cinco leituras para cada corpo de prova, num total de 100 observações para cada modelo. As figuras 13.4 e 13.5 mostram os histogramas para os dois modelos, a resistência sendo medida em kg.

Figura 13.4: Resistência à remoção, em kg, para o modelo **C**.

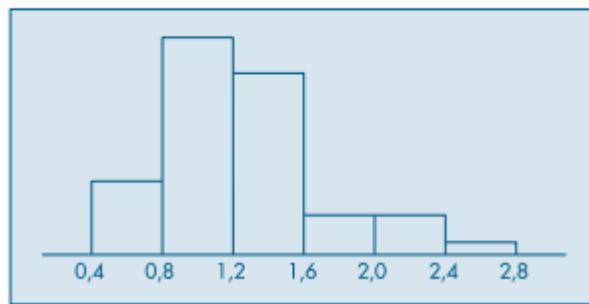
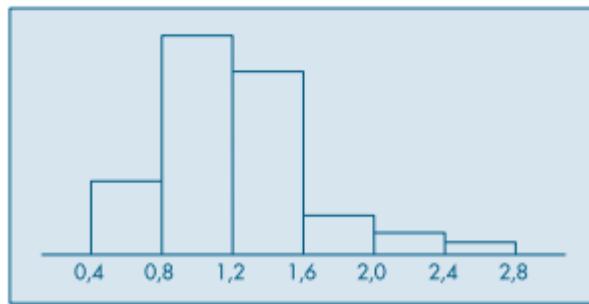


Figura 13.5: Resistência à remoção, em kg, para o modelo **T**.



Vemos que há assimetrias nos histogramas, sugerindo que a aplicação do teste t de Student não é adequada nessa situação. A tabela 13.6 mostra as médias das 5 leituras para cada corpo de provas, para o modelo *T* e para o modelo *C* (em ordem crescente)

Admitamos que o grupo de controle seja aquele em os grampos seja do *T*, e grampos do tipo *C* constituem os tratamentos. Ordenando as médias da tabela 13.6 e atribuindo postos obtemos a tabela 13.7

Aqui $m = n = 20$ e queremos testar

H_0 : a resistência à remoção é a mesma para os dois tipos de grampos;

H_1 : O tipo *C* apresenta menor resistência à remoção do que o tipo *T*.

A soma dos postos de tratamentos é

$$W_s = S_1 + S_2 + \dots + S_{20} = 406,5.$$

A esperança de W_s é dada por:

$$E(W_s) = \frac{20 \times 41}{2} = 410.$$

Para calcular a variância temos que levar em conta os empates. Assim

Há na amostra 34 valores distintos. assim $e = 34$. Logo

$$d_1 = d_2 = d_3 = d_4 = 1, d_5 = 2, d_6 = 2, d_7 = \dots = d_{12} = 1,$$

$$d_{13} = 2, d_{14} = 2, d_{15} = d_{16} = 1, d_{17} = d_{18} = 2,$$

$$d_{19} = d_{34} = 1.$$

Para $d_i = 1$ temos

$$d_i^3 - d_i = 0$$

Para $d_i = 2$ temos

$$d_i^3 - d_i = 6$$

Assim,

$$\sum_{i=1}^e (d_i^3 - d_i) = 6 \times 6 = 36.$$

$$Var(W_s) = \frac{20 \times 20 \times 41}{12} - \frac{20 \times 20}{12 \times 40 \times 39} [36],$$

$$Var(W_s) = 1366,667 - 2,857 = 1.363,810.$$

Vamos padronizar

$$z = \frac{406,5 - 410}{36,93} = -0,095.$$

Como rejeitaremos H_0 se $W_s \leq c$, no nível $\alpha = 0,05$, devemos comparar esse valor com o valor -1,64 da normal padrão, portanto não rejeitamos h_0

Usando o R temos:

Vemos que o valor-p do teste é

$$nd = P(W_s \leq 406,5) \approx P(Z \leq -0,095) = 0,46,$$

que é uma indicação de que a hipótese H_0 deve ser aceita.

Observação Comparação entre o Teste t e o Teste de Wilcoxon.

O teste t baseia-se na suposição de que as populações P_1 e P_2 sejam normais. Uma violação dessa altera a distribuição da estatística usada no teste e muda as probabilidades do erro do tipo I e II. Dizemos que um teste é robusto contra a violação de uma suposição se suas probabilidades de erro do tipo I e II não são afetadas de forma apreciável pela violação. Pode-se mostrar que o teste t é pouco sensível à heterogeneidade de variâncias se $m = n$, mas ele será afetado se as variâncias forem diferentes e $m \neq n$.

Os teste t e de Wilcoxon são comparados através de seus poderes em termos de uma quantidade chamada eficiência relativa assintótica, mas não entraremos em detalhes aqui sobre o assunto. Mas podemos resumir a situação da seguinte maneira:

- a. O teste t é mais poderoso quando temos populações normais, mas a perda de eficiência do teste de Wilcoxon é pequena (menos de 5%) nesse caso.
- b. haverá pouca diferença entre os dois testes para distribuições próximas da normal.
- c. O teste de Wilcoxon é mais eficiente para distribuições que têm caudas mais pesadas do que a normal.

Para se ter uma ideia do que significa mais pesada, observamos que as distribuições t e Cauchy têm distribuições com caudas mais pesadas do que a normal. Se P_1 e P_2 forem ambas uniformes, pode-se provar que os dois testes são igualmente eficientes e se P_1 e P_2 forem ambas exponenciais, o teste de Wilcoxon é três vezes mais eficiente.

Problemas

1. (P10) Vinte canteiros foram plantados com milho. Em dez deles um novo tipo de fertilizante foi aplicado, obtendo-se as produções abaixo. Há diferenças significativas entre as produções? A alternativa é que o novo fertilizante tende a produzir valores maiores. Tome $\alpha = 0,05$. Calcule o nível descritivo.

Controle	7,1	6,0	8,0	7,0	6,6	7,4	7,0	7,0	6,9	6,8
Tratamento	6,9	6,8	7,5	6,8	6,9	6,8	6,8	6,8	6,7	6,6

2. (P11) Obtenha a distribuição nula de W_s para os casos:

$$(a) m = n = 2 \quad (b) m = 2, n = 4 \quad (c) m = n = 3.$$

Faça também no R.

3. (P12) Calcule as seguintes probabilidades, usando a tabela VIII e a aproximação Normal.

- a. $m=6, n=7$, $P(W_s \leq 48)$.
- b. $m=8, n=10$, $P(W_s \leq 65)$.
- c. $m=10, n=10$, $P(W_s \geq 63)$.
- d. Agora faça tudo no R.

4. (P13) Encontre a distribuição nula de W_s no caso de empates, para os casos:

- a. $m=n=3$, $d_1 = d_2 = 1$, $d_3 = 2$, $d_4 = d_5 = 1$.
- b. $m=n=3$, $d_1 = d_2 = d_3 = 2$.
- c. $m=2, n=3$, $d_1 = d_2 = 1$, $d_3 = 3$
- d. Agora faça tudo no R.

5. (P14) Faça os histogramas para W_s nos problemas 11 e 13.

6. (P15) Suponha que as observações dos tratamentos sejam 3, 3, 5 e 7, e as observações dos controles sejam 1, 4 e 8, e que o teste de Wilcoxon rejeite para valores grandes de W_s . Calcule $\hat{\alpha} = P(W_s \geq w)$, em que w é o valor observado de W_s .

7. (P37) Em um estudo para comparar os efeitos de duas dietas ,**A** e **B**, sobre o crescimento, 6 ratos foram submetidos à dieta **A**, e 9 ratos à dieta **B**. Após 5 semanas, os ganhos em peso foram:

A	15	18	12	11	14	15			
B	11	11	12	16	12	13	8	10	13

- a. Admitindo que temos duas amostras independentes de populações normais , teste a hipótese de que não há diferença entre as duas dietas, contra a alternativa que a dieta **A** é mais eficaz, usando o teste t de Student, no nível de significância de 0,01. Calcule o nível descritivo.
- b. Efetue o teste usando a estatística de Wilcoxon com nível de significância de 0,01. Calcule o nível descritivo.
8. (P40) Seja $W_R = R_1 + R_2 + \dots + R_n$ a soma dos postos dos controles. Qual o valor de $W_R + W_S$?
9. (P41) Se $n = 4$ e $m = 6$, prove que $P(W_S \geq 35) = P(W_S \leq 31)$, usando o fato que W_S é simétrica em torno de $\frac{m(N+1)}{2}$.
10. (P42) Se $n = 4$ e $m = 6$, prove que $P(W_S \geq 35) = P(W_S \leq 20)$

13.4-Comparação de Duas Populações: Amostras Dependentes

Na seção 13.1 já discutimos essa situação. Aqui, temos duas amostras X_1, \dots, X_n e Y_1, \dots, Y_n , só que agora as observações são pareadas, isto é, podemos considerar que temos na realidade uma amostra de pares $(X_1, Y_1), \dots, (X_n, Y_n)$. Se definirmos a *v.a.* $D = X - Y$, teremos a amostra D_1, D_2, \dots, D_n , resultante das diferenças entre os valores de cada par.

Observe que reduzimos a um problema com uma única população, conforme estudado nos capítulos anteriores.

Consideraremos dois casos: no primeiro, supomos que a população das diferenças é normal; no segundo, supomos que essa população é simétrica.

13.4.1: População Normal

Nessa situação, faremos a seguinte suposição: a *v.a.* D tem distribuição normal $N(\mu_D, \sigma_D^2)$. Podemos deduzir daqui que

$$\bar{D} = \frac{1}{n} \sum_{i=1}^n D_i = \frac{1}{n} (X_i - Y_i) = \bar{X} - \bar{Y} \quad 13.27$$

terá distribuição $N(\mu_D, \sigma_D^2/n)$.

Considere

$$S_D^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2 \quad 13.28$$

Pelo Teorema 7.1, a estatística

$$T = \frac{\sqrt{n}(\bar{D} - \mu_D)}{S_D} \quad 13.29$$

terá distribuição *t* de Student, com $(n - 1)$ graus de liberdade.

Como

$$\mu_D = E(D) = E(X - Y) = E(X) - E(Y) = \mu_1 - \mu_2.$$

qualquer afirmação sobre o $\mu_1 - \mu_2$ corresponde a uma afirmação sobre μ_D .

Exemplo 13.10. Cinco operadores de certo tipo de máquina são treinados em máquinas de duas marcas diferentes, A e B. Mediú-se o tempo que cada um deles gasta na realização de uma mesma tarefa, e os resultados estão na Tabela 13.8.

Operador	Marca A	Marca B
1	80	75
2	72	70
3	65	60
4	78	72
5	85	78

Tabela 3: Tempos para realização de tarefa para cinco operadores.

Com o nível de significância de 10%, poderíamos afirmar que a tarefa realizada na máquina A demora mais do que na máquina B?

Estamos interessados em testar

$$H_0 : \mu_A = \mu_B$$

$$H_1 : \mu_A > \mu_B$$

Essas hipóteses são equivalentes a

$$H_0 : \mu_D = 0$$

$$H_1 : \mu_D > 0$$

Como é o mesmo operador que realiza a tarefa nas duas máquinas, estamos diante do caso em que se pode usar variáveis emparelhadas. Vamos admitir que, sob H_0 , a diferença de tempo segue uma distribuição normal $N(0, \sigma_D^2)$. Logo, usamos a estatística (13.29). Para determinar a região crítica, note que, devido à forma de H_1 , devemos encontrar t_c tal que $P(T > t_c) = 0,10$, sendo que T tem distribuição $t(4)$. Usando a Tabela V, obtemos

$$RC =]1,54; +\infty[$$

Da Tabela 3 obtemos os valores de D:

$$d_i = 5, 2, 5, 6, 7$$

E, por tanto,

$$\bar{d} = 5 \quad e \quad S_D^2 = 3,5$$

O valor observado da estatística T é $t_0 = (5/1,87)(\sqrt{5}) = 5,98$. Segue-se que rejeitamos H_0 , ou seja, demora-se mais para realizar a tarefa com a máquina A.

Podemos construir um intervalo de confiança para μ_D ; para $\gamma = 0,90$,

$$IC(\mu_A - \mu_B; 0,90) = IC(\mu_D; 0,90) = 5 \pm (2,13)(1,87)/\sqrt{5}.$$

Ou seja,

$$IC(\mu_D; 0,90) =]3,22; 6,78[.$$

13.4.1: População Não-Normal

Vamos considerar, agora, um teste baseado nos postos das diferenças D_i : o chamado *teste dos postos sinalizados de Wilcoxon*.

Para esse teste, supomos que a escala das diferenças seja pelo menos intervalar e que os pares (X_i, Y_i) constituam uma AAS.

Isso implica, em particular, que os D_i são independentes, com a mesma mediana. Suponha, ainda, que cada D_i tenha uma distribuição simétrica. Ou seja, as médias e medianas coincidem.

Exemplo 13.11. Suponha que se possa simular um modelo por meio de duas linguagens computacionais, que chamaremos A e B. Supostamente, o tempo usando B é menor que o tempo usando A. Cinco pares de alunos são selecionados para o teste, de modo que cada membro de um par tenha a mesma habilidade computacional nas duas linguagens do que o outro. Um membro de cada par é escolhido ao acaso e este vai usar a linguagem B; o outro usará A.

O tempo de simulação (em segundos) de cada linguagem é anotado, obtendo-se a Tabela 4:

Par	1	2	3	4	5
Tempo de B(X)	300	410	420	410	400
Tempo de A(Y)	350	309	490	435	440
$D = X - Y$	-50	20	-70	-25	-40
Posto de $ D $	4	1	5	2	3
Posto Sinalizado	-4	+1	-5	-2	-3

Tabela 4: Tempos de simulação (em segundos) para as linguagens A e B.

Queremos testar a hipótese de que os tempos são semelhantes contra a hipótese de que os tempos de B são menores. Ou, ainda,

$$H_0 : \mu_B - \mu_A = \mu_D = 0$$

$$H_1 : \mu_B - \mu_A = \mu_D < 0$$

Na quarta linha da Tabela 4 estão apresentadas as diferenças D_i , e os postos são calculados a partir das variáveis $|D_i|$, ou seja, os módulos (ou valores absolutos) dos D_i (quinta linha). A sexta linha, "posto sinalizado", é obtida atribuindo-se ao posto de $|D_i|$ o sinal correspondente de D_i . Por exemplo, para a primeira observação, $D_1 = 300 - 350 = -50$, com $|D_1| = 50$, que tem posto 4 e, portanto, posto sinalizado -4.

Notamos que só há um posto positivo, +1. Se indicarmos por T^+ a soma dos postos positivos, rejeitaremos H_0 se T^+ for pequeno.

É claro que podemos trabalhar com os postos negativos também, e considerar

$$T^- = -(\text{soma dos postos negativos}).$$

No exemplo,

$$T^+ = 1 \quad e \quad T^- = 14.$$

Usando T^- , rejeitaremos H_0 se esta for "grande"?

Note que

$$T^+ + T^- = 1 + 14 = 15 = \frac{5(5+1)}{2},$$

que é a soma de todos os postos dos $|D_i|$, que, por sua vez, é $n(n+1)/2$, sendo $n = 5$ o número de pares.

Em geral, devemos usar a menor soma.

Trabalhemos com T^+ . Para conduzir o teste, devemos obter a distribuição dessa estatística, sob a hipótese nula H_0 .

Para isso, note que, se H_0 for verdadeira, cada posto tem a mesma probabilidade de ser associado com um sinal + ou com um sinal -. Logo, a sequência de postos sinalizados é uma de todas as possíveis combinações de $\pm 1, \pm 2, \dots, \pm 5$.

Há

$$2^5 = 32$$

de tais combinações, todas equiprováveis sob H_0 , ou seja, com probabilidade $\frac{1}{32}$.

Na Tabela 5 temos todas as possibilidades juntamente com o valor de T^+ .

Na Tabela 6 temos a distribuição de T^+ . Note que a distribuição de T^+ é simétrica, com média e mediana iguais a 7,5.

1	2	3	4	5	T^+	1	2	3	4	5	T^+
+	+	+	+	+	15	+	+	-	+	-	7
-	+	+	+	+	14	-	+	-	-	+	7
+	-	+	+	+	13	-	-	+	+	-	7
+	+	-	+	+	12	+	-	-	-	+	6
-	-	+	+	+	12	+	+	+	-	-	6
+	+	+	-	+	11	-	+	-	+	-	6
-	+	-	+	+	11	+	-	-	+	-	5
+	+	+	+	-	10	-	+	+	-	-	5
-	+	+	-	+	10	-	-	-	-	+	5
+	-	-	+	+	10	+	-	+	-	-	4
-	+	+	+	-	9	-	-	-	+	-	4
-	-	-	+	+	9	+	+	-	-	-	3
+	-	+	-	+	9	-	-	+	-	-	3
+	+	-	-	+	8	-	+	-	-	-	2
+	-	+	+	-	8	+	-	-	-	-	1
-	-	+	-	+	8	-	-	-	-	-	0

Tabela 5: Sinais possíveis para os postos, Exemplo 13.10

O valor-p do teste é

$$P(T^+ \leq 1 | H_0) = 2/32 = 0,06,$$

T^+	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Frequência	1	1	1	2	2	3	3	3	3	3	3	2	2	1	1	1

Tabela 6: Distribuição de T^+ sob H_0

usando a Tabela 6.

Ou seja, há indicação de que o tempo de simulação usando a linguagem B é menor do que o tempo de A. Observe que temos poucos pares, e o valor $\hat{\alpha} = 0,06$ não é tão pequeno. Mas como temos somente um posto positivo dentre cinco, somos levados a duvidar da validade de H_0 .

Vamos calcular a média e a variância de T^+ usando o **R**:

```

> #####n=5
>
> t=0:15;t
[1] 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
>
> freq=c(rep(1,3),rep(2,2),rep(3,6),rep(2,2),rep(1,3));freq
[1] 1 1 1 2 2 3 3 3 3 3 2 2 1 1 1
>
> pt=freq/32;pt
[1] 0.03125 0.03125 0.03125 0.06250 0.06250 0.09375 0.09375 0.09375 0.09375
[10] 0.09375 0.09375 0.06250 0.06250 0.03125 0.03125 0.03125
> sum(pt)
[1] 1
> ETmais=sum(t*pt);ETmais
[1] 7.5
>
> n=5
>
> mu=n*(n+1)/4;mu
[1] 7.5
>
>
> ET2mais=sum(t^2*pt);ET2mais
[1] 70
>
> VarTmais=ET2mais-ETmais^2;VarTmais
[1] 13.75
>
> sigma2=n*(n+1)*(2*n+1)/24;sigma2
[1] 13.75
>
>

```

Vejamos, agora, o caso geral. Tomemos os valores absolutos das diferenças, ou seja

$$|D_i| = |X_i - Y_i|, \quad i = 1, \dots, m.$$

Quando $X_i = Y_i$ omitir a diferença correspondente e seja n o número de diferenças estritamente diferentes de zero.

Associemos a cada par (X_i, Y_i) o posto do módulo de D_i correspondente. Use postos médios, se houver D_i coincidentes.

A hipótese a ser testada é que a média (ou a mediana) das diferenças seja igual a zero contra a alternativa que não seja. Testes unilaterais podem, também, ser considerados. Ou seja, dada a simetria da distribuição dos D_i , iremos testar

$$H_0 : \mu_D = 0$$

$$H_1 : \mu_D \neq 0$$

onde μ_D representa, como antes, a média das diferenças.

Considere

$$R_i = \begin{cases} R(X_i, Y_i), & \text{se } D_i > 0 \\ -R(X_i, Y_i), & \text{se } D_i < 0 \end{cases} \quad 13.30$$

onde $R(X_i, Y_i)$ é o posto associado a (X_i, Y_i) .

Temos dois casos a tratar:

(A) Se não houver empates, use a estatística:

$$T^+ = \sum(R_i \text{ com } D_i > 0) \quad 13.31$$

ou seja, a soma dos postos positivos. Use a Tabela IX, pág. 506, para obter os quantis w_p da estatística, ou seja, o valor, tal que $P(T^+ < w_p) \leq p$ e $P(T^+ > w_p) \leq 1 - p$, se H_0 for verdadeira.

Para $n > 50$ use a aproximação normal, com média e variância dados no teorema abaixo. Para $p > 0,5$ o quantil é dado por:

$$w_p = \frac{n(n+1)}{2} - w_{1-p}$$

(B) Se houver empates, use a estatística:

$$V = \frac{\sum_{i=1}^N R_i}{\sqrt{\sum_{i=1}^N R_i^2}} \quad 13.32$$

que tem uma distribuição aproximadamente $N(0, 1)$, sob a hipótese nula.

Teorema 13.3. A média e variância de T^+ são dadas por

$$E(T^+) = \frac{n(n+1)}{4} \quad 13.33$$

e

$$VAR(T^+) = \frac{n(n+1)(2n+1)}{24}, \quad 13.34$$

Prova: Lembre que T^+ representa o número de diferenças $D = X - Y$ positivas. Se H_0 é verdade temos que lembrando que X e Y são variáveis aleatórias contínuas

$$P(X < Y) = P(X > Y) = P(X - Y > 0) = P(D > 0) = \frac{1}{2}.$$

Considere a variável indicadora I_i que assume o valor 1 se $X_i - Y_i > 0$, $i = 1, 2, \dots, n$. Se $X_i - Y_i > 0$ o posto terá sinal positivo.

Logo

$$I_i \sim Ber(1/2), \quad i = 1, 2, \dots, n.$$

Note que:

$$E(I_i) = \frac{1}{2} \quad e \quad V(I_i) = \frac{1}{4}.$$

Vamos escrever T^+ da seguinte maneira:

$$T^+ = \sum_{i=1}^n i I_i.$$

Perceba que:

$$\begin{aligned} E(T^+) &= \sum_{i=1}^n E(i I_i) = \sum_{i=1}^n i E(I_i) = \frac{1}{2} \sum_{i=1}^n i. \\ E(T^+) &= \frac{1}{2} \times \frac{n(n+1)}{2} = \frac{n(n+1)}{4}. \end{aligned}$$

Por outro lado:

$$\begin{aligned} V(T^+) &= \sum_{i=1}^n V(i I_i) = \sum_{i=1}^n i^2 V(I_i) = \frac{1}{4} \sum_{i=1}^n i^2. \\ V(T^+) &= \frac{1}{4} \times \frac{n(n+1)(2n+1)}{6} = \frac{n(n+1)(2n+1)}{24}. \end{aligned}$$

Exemplo 13.11. (continuação)

Obtivemos aqui $T^+ = 1$.

A região crítica é unilateral à esquerda, logo rejeitamos H_0 se $T^+ < w_\alpha$, onde w_α é o quantil dado pela Tabela IX.

Se fixarmos $\alpha = 0,025$ ou $\alpha = 0,01$, obteremos $w_\alpha = 0$, com $n = 5$, e, portanto, aceitaremos H_0 . Se $\alpha = 0,05$, então $w_\alpha = 1$, e o valor observado estará na fronteira da região crítica e teremos dúvidas em aceitar ou rejeitar H_0 . Como salientamos antes, a decisão, nesse caso, dependerá de uma análise cuidadosa dos resultados, dado o pequeno valor de n .

Vamos apresentar a tabela IX do livro:

Vamos fazer usando o **R** :

Tabela IX – Distribuição de Wilcoxon T^+ O corpo da tabela dá os valores w_p tais que $P(T^+ < w_p) = p$									
$w_{0.005}$					$w_{0.01}$				
$w_{0.025}$					$w_{0.05}$				
$n = 4$	0	0	0	0	1	28	94	108	120
5	0	0	0	1	3	29	92	102	117
6	0	0	1	3	4	30	101	111	127
7	0	1	3	4	6	31	110	121	138
8	1	2	4	6	9	32	119	131	148
9	2	4	6	9	11	33	129	141	160
10	4	6	9	11	15	34	139	152	171
11	6	8	11	14	18	35	149	163	183
12	8	10	14	18	22	36	160	175	196
13	10	13	18	22	27	37	172	187	209
14	13	16	22	26	32	38	184	199	222
15	16	20	26	31	37	39	196	212	236
16	20	24	30	36	43	40	208	225	250
17	24	28	35	42	49	41	221	239	265
18	28	33	41	48	56	42	235	253	280
19	33	38	47	54	63	43	248	267	295
20	38	44	53	61	70	44	263	282	311
21	44	50	59	68	78	45	277	297	328
22	49	56	67	76	87	46	292	313	344
23	55	63	74	84	95	47	308	329	362
24	62	70	82	92	105	48	324	346	379
25	69	77	90	101	114	49	340	363	397
26	76	85	99	111	125	50	357	381	416
							374	398	435
								467	504

Figura 4:

```

> #####Bussab & Morettin-Exemplo 13.11
>
>
>
> #### Suponha que se possa um modelo por meio de duas linguagens computacionais
> ##A e B. Supostamente o tempo usando B é menor que o tempo usando A. Vamos usar cinco pares de
> ##habilidade computacional. As linguagens serão sorteadas as acaso uma para cada elemento do p
> ##Y=tempo usando A,-----X=tempo usando B
>
>
> ##### Seja teta_A o tempo mediano populacional usando a linguagem A.
>
> ##### Seja teta_B o tempo mediano populacional usando a linguagem B.
>
>
> #####H_0: teta_B=teta_A vs H_1: teta_B < teta_A ou
>
> #####H_0: teta_B-teta_A=teta_D=0 vs H_1: teta_B - teta_A=teta_D <0
>
> ##teta_D= tempo mediano populacional da diferença D=X-Y
>
> #####Os tempos de simulação em segundos obtidos.
>
> n=5
> X=c(300,410,420,410,400)
>
> Y=c(350,390,490,435,440)
>

```

```
> D=X-Y
> MD=abs(D)
> Z=rank(MD)
> Posto_s=sign(D)*Z #####Posto Sinalizado.
>
> sign(4);sign(-pi);sign(0)
[1] 1
[1] -1
[1] 0
>
> tab13.9=cbind(X,Y,D,MD,Z,Posto_s);tab13.9
X   Y   D MD Z Posto_s
[1,] 300 350 -50 50 4      -4
[2,] 410 390  20 20 1       1
[3,] 420 490 -70 70 5      -5
[4,] 410 435 -25 25 2      -2
[5,] 400 440 -40 40 3      -3
>
>
> #####TM= T+=soma dos postos positivos----Tm=T-= -soma dos postos negativos
>
> #####TM+Tm=n(n+1)/2
>
>
>
> TM=1;Tm=4+5+2+3
>
> TM+Tm;n*(n+1)/2
[1] 15
[1] 15
>
> #####Vamos obter a distribuição amostral de T+ usando O R:
>
> ##### signrank
>
> #####t um valor de T+ no suporte A={0,1,2,...,n*(n+1)/2}.
>
> t=0: (n*(n+1)/2);t
[1] 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
>
> pt=dsignrank(t,n) #####P(T+=t)=A_t/2^n
>
> Pt=psignrank(t,n) #####P(T+ <=t)
> #####?dsignrank
>
> A_t=32*pt
>
> tab13_11=cbind(t,A_t,pt,Pt);tab13_11
t A_t      pt      Pt
[1,] 0 1 0.03125 0.03125
[2,] 1 1 0.03125 0.06250
[3,] 2 1 0.03125 0.09375
```

```
[4,] 3 2 0.06250 0.15625
[5,] 4 2 0.06250 0.21875
[6,] 5 3 0.09375 0.31250
[7,] 6 3 0.09375 0.40625
[8,] 7 3 0.09375 0.50000
[9,] 8 3 0.09375 0.59375
[10,] 9 3 0.09375 0.68750
[11,] 10 3 0.09375 0.78125
[12,] 11 2 0.06250 0.84375
[13,] 12 2 0.06250 0.90625
[14,] 13 1 0.03125 0.93750
[15,] 14 1 0.03125 0.96875
[16,] 15 1 0.03125 1.00000
>
> #####Cálculo do nível descritivo do teste
>
> ####nd=P(T+<= TM)=P(T+<=1)
>
> nd=psignrank(TM,n);nd
[1] 0.0625
>
> #####Direto no R
>
> #####Testando se a mediana da diferença é zero.
>
> wilcox.test(X,Y,paired=T,alternative="less")
```

Wilcoxon signed rank exact test

```
data: X and Y
V = 1, p-value = 0.0625
alternative hypothesis: true location shift is less than 0
```

```
>
>
```

13.5 Comparação de Proporções em Duas Populações

Nosso objetivo agora é a comparação das proporções de duas populações P_1 e P_2 . Sendo mais explícitos, queremos comparar as proporções populacionais p_1 e p_2 , por meio dos estimadores \hat{p}_1 e \hat{p}_2 obtidos de amostras independentes de tamanhos n_1 e n_2 respectivamente. Das seções 10.9 e 12.6 temos:

$$\hat{p}_1 \sim N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right), \quad \hat{p}_2 \sim N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right)$$

Comparando com o resultado da seção 13.3.1, e também do Problema 10.32, obtemos

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

e portanto, a estatística de decisão, tanto para a construção de intervalos de confiança como para testes de hipóteses, será

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0, 1) \quad 13.35$$

Exemplo 13.12: Para lançamento da nova embalagem do sabonete SEBO a divisão de criação estuda duas propostas:

- A: amarela com letras vermelhas, ou
- B: preta com letras douradas.

Eles acreditam que a proposta A chama a atenção em pelo menos 5% a mais do que a proposta B. Para verificar a validade de tal informação conduziu-se o seguinte experimento: em cada um de dois supermercados similares, foram colocados sabonetes com cada tipo de embalagem, e a clientes selecionados aleatoriamente foi perguntado se tinham notado o sabonete e que descrevessem qual a embalagem. Abaixo estão os resultados.

Proposta	Notaram	Não Notaram	Total
A	168	232	400
B	180	420	600
Total	348	652	1000

Queremos testar a hipótese que os dois tipos de embalagem são igualmente prováveis. Aqui, consideramos

$$H_0 : p_A - p_B = 0 \quad H_1 : p_A - p_B \neq 0$$

Da tabela obtemos: $\hat{p}_A = 0,42$ e $\hat{p}_B = 0,30$, e aplicando a fórmula (13.35) obtemos:

$$Z = \frac{(0,42 - 0,30) - 0,05}{\sqrt{\frac{(0,42)(0,58)}{400} + \frac{(0,30)(0,70)}{600}}} = 2,26.$$

Consultando a Tabela III, encontramos o valor-p $\hat{\alpha} = 1,19\%$, o que leva a rejeição de H_0 . O passo seguinte seria a construção de um Intervalo de Confiança, e novamente aplicado a expressão (13.35), obtém-se:

$$IC(p_A - p_B; 95\%) = (0,42 - 0,30) \pm 1,96 \sqrt{\frac{(0,42)(0,58)}{400} + \frac{(0,30)(0,70)}{600}}$$

$$IC(p_A - p_B; 95\%) = 0,12 \pm 0,036 = [0,084 ; 0,156].$$

Para testar a hipótese de igualdade de proporções, $p_1 = p_2$, e usando as mesmas argumentações apresentadas na seção 13.3.1(a), deve-se usar uma estimativa comum das variâncias dada por \hat{p}_c ($1 - \hat{p}_c$), onde $\hat{p}_c = (n_1 p_1 + n_2 p_2) / (n_1 + n_2)$ resultando no teste:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}_c(1 - \hat{p}_c) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim N(0, 1) \quad 13.36$$

Exemplo 13.12 (continuação): Voltando ao problema do sabonete SEBO, suponha que eles não sabem se uma embalagem é ou não mais atraente do que outra, e a pesquisa foi feita para responder a essa questão. Portanto o teste agora será:

$$H_0 : p_A = p_B, \quad H_1 : p_A \neq p_B.$$

Da tabela obtemos $\hat{p}_c = (348/1000) = 0,348$, substituindo em (13.36), obtemos:

$$Z = \frac{0,42 - 0,30}{\sqrt{0,348(0,652) \left(\frac{1}{400} + \frac{1}{600} \right)}}$$

Consultando a Tabela III, encontramos valor- p próximo de zero, o que leva a rejeição de H_0 . Como esse resultado mostra que as variâncias também são diferentes, a construção do Intervalo de Confiança é obtida do mesmo modo acima.

Problemas

16. Para investigar a lealdade de consumidores a um determinado produto, sorteou-se uma amostra de 200 homens e 200 mulheres. Foram classificados como tendo alto grau de fidelidade 100 homens e 120 mulheres. Os dados trazem evidências de diferença de grau de fidelidade entre os sexos? Em caso afirmativo construa um intervalo de confiança para a diferença.
17. Em uma amostra de 500 famílias da cidade A, constatou-se que 298 haviam comprado, durante os últimos 30 dias, o refrigerante Meca-Mela em sua nova versão incolor. Na cidade B esse número foi de 147 em 300 famílias entrevistadas. Na cidade A foi feita uma campanha publicitária através da rádio local, e não na cidade B. Os resultados trazem evidências de que as campanhas locais aumentam as vendas?
18. Um partido afirma que a porcentagem de votos masculinos a seu favor será 10% a mais que a de votos femininos. Em uma pesquisa feita entre 400 homens, 170 votariam no partido, enquanto que entre 625 mulheres, 194 lhe seriam favoráveis. A afirmação do partido é verdadeira ou não? Caso rejeite a igualdade, dê um IC para a diferença.
19. Para investigar os resultados do segundo turno de uma eleição estadual tomaram-se duas amostras de 600 eleitores cada: uma da capital e outra do interior. Da primeira, 276 disseram que votariam no candidato A, enquanto que 312 eleitores do interior também o fariam.
- (a) Estime a proporção de eleitores da capital que votariam em A. Dê um IC. (b) Existe diferença nas proporções entre capital e interior? (c) Que tamanho igual deveriam ter ambas as amostras para que a diferença entre as proporções fosse estimada com erro inferior a 2%? (d) Qual a proporção esperada de votos que irá receber o candidato A no estado? (e) De
- uma amostra de 120 indivíduos da classe A e B, 69 são favoráveis a eleição em dois turnos, enquanto que em uma amostra de 100 indivíduos da classe C, 48 são favoráveis. Existe evidência e diferenças de opiniões em relação à classe social?
20. Para verificar a importância de um cartaz nas compras de certo produto, procedeu-se do seguinte modo: (a) formaram-se sete pares de lojas; (b) os pares foram formados de modo que tivessem as mesmas características quanto à localização, ao tamanho e ao volume de vendas; (c) num dos elementos do par, colocou-se o cartaz; no outro, não; (d) as vendas semanais foram registradas, e os resultados estão a seguir. Qual seria a sua conclusão sobre a eficiência do cartaz? Use o teste t, fazendo as suposições necessárias.
21. Resolva o problema anterior, usando o teste dos postos sinalizados de Wilcoxon.

Pares	Vendas Sem Cartaz	Vendas com Cartaz
1	13	16
2	18	24
3	14	18
4	16	14
5	19	26
6	12	17
7	22	29

22. Aplique o teste de Wilcoxon para os dados do Exemplo **13.10**.
23. Os dados abaixo referem-se a medidas de determinada variável em 19 pessoas antes e depois de uma cirurgia. Verifique se as medidas pré e pós-operatórias apresentam a mesma média. Que suposições você faria para resolver o problema? Faça gráficos apropriados para verificar suas suposições.

Pessoas	Pré	Pós	Pessoas	Pré	Pós
1	50	42	10	40	50
2	50	42	11	50	48
3	50	78	12	75	52
4	87,5	33	13	92,5	74
5	32,5	96	14	38	47,5
6	35	82	15	46,5	49
7	40	44	16	50	58
8	45	31	17	30	42
9	62,5	87	18	35	60
			19	39,4	28

13-7:Problemas Suplementares:

24. Uma empresa deseja estudar a produtividade de seus trabalhadores. Para isso, sorteou seis operários, e contou o número de peças produzidas durante uma semana sem intervalo e uma semana com intervalo.

Os resultados sugerem se há ou não melhora na produtividade? Caso haja melhora, qual deve ser o acréscimo médio de produção para todos os trabalhadores da fábrica?

Operário	1	2	3	4	5	6
Sem Intervalo	23	35	29	33	43	32
Com Intervalo	28	38	29	37	42	30

26. Numa indústria, deseja-se testar se a produtividade média dos operários do período diurno é igual à produtividade média dos operários do período noturno.

Para isso, colheram-se duas amostras, uma de cada período, observando- se a produção de cada operário. Os resultados obtidos foram os seguintes:

Período	n	$\sum x_i$	$\sum x_i^2$
Diurno	15	180	2.660
Noturno	15	150	2.980

De acordo com esses resultados, quais seriam suas conclusões?

26. Num levantamento feito com os operários da indústria mecânica, chegou-se aos seguintes números: salário médio = 3,64 salários mínimos e desvio padrão = 0,85 salário mínimo. Suspeita-se que os salários da subclasse formada pelos torneiros mecânicos são diferentes dos salários do conjunto todo, tanto na média como na variância. Que conclusões você obteria se uma amostra de 25 torneiros apresentasse salário médio igual a 4,22 salários mínimos e desvio padrão igual a 1,25 salário mínimo?
27. Os dados abaixo representam a porcentagem do orçamento gasto com pessoal para 50 pequenos municípios de uma certa região.

69,5	71,6	73,0	68,9	68,9	70,0	72,6	66,2	68,1	72,4
67,6	73,2	67,6	69,7	71,0	69,4	71,5	73,8	69,6	69,6
68,2	69,9	71,4	70,7	69,7	71,0	66,0	70,3	71,7	69,2
69,8	68,4	69,5	68,2	72,1	70,8	72,2	69,2	71,7	65,6
69,6	70,1	69,9	70,5	68,0	70,2	69,0	66,3	69,4	67,1

- (a) Analise estatisticamente os dados.
- (b) Com base na sua análise, e sabendo que na região considerada existem, ao todo, 200 municípios, em quantos deles você acha que o gasto com pessoal é maior que 70% do orçamento?
- (c) Em outra região, sabe-se que o gasto médio com pessoal é de 65%, e o desvio padrão é de 20%. Qual das duas regiões é mais homogênea em relação a essa variável? Por quê?
28. Uma amostra de 100 trabalhadores de uma fábrica grande demora, em média, 12 minutos para completar uma tarefa, com um desvio padrão de dois minutos. Uma amostra de 50 trabalhadores de uma outra fábrica demora, em média, 11 minutos para completar a mesma tarefa, com desvio padrão igual a três minutos.
- (a) Construa um IC de 95% para a diferença entre as duas médias populacionais.
- (b) Deixe bem claro quais as suposições feitas para a solução apresentada.
29. Deseja-se testar se dois tipos de ensino profissional são igualmente eficazes. Para isso, sortearam-se duas amostras de operários; a cada uma, deu-se um dos tipos de treinamento e, no final, submeteram-se os dois grupos a um mesmo teste. Que tipo de conclusão você poderia tirar, baseando-se nos resultados abaixo?

Amostra	Número de elementos	Média	Desvio padrão
Tipo I	12	75	5
Tipo II	10	74	10

30. Numa discussão sobre reajuste salarial, entre empresários e o sindicato dos empregados, chegou-se a um impasse. Os empresários dizem que o salário médio da categoria é 7,6 salários mínimos (SM), e os empregados dizem que é 6,5 SM. Para eliminar dúvidas, cada um dos grupos resolveu colher uma amostra independente. Os empresários, com uma amostra de 90 operários, observaram um salário médio de 7,0 SM, com um desvio padrão igual a 2,9 SM. Já a amostra do sindicato, com 60 operários, apresentou média igual a 7,10 SM e desvio padrão de 2,4 SM.

- (a) As amostras colhidas servem para justificar as respectivas afirmações dos dois grupos?
- (b) De posse dos dois resultados, qual é o seu parecer?

31. A Torrefação Guarany está querendo comprar uma nova ensacadora de café. Após consultar o mercado, ficou indecisa entre comprar a de marca A ou a de marca B. Quanto ao custo, facilidade de pagamento, tamanho etc. elas são equivalentes. O fator que decidirá a compra será a precisão em encher os pacotes (medido pela variância). Deseja-se, na realidade, testar hipótese

$$\sigma_A^2 = \sigma_B^2$$

, por meio da estatística

$$F = \frac{S_A^2}{S_B^2}.$$

Podem-se construir regiões críticas bilaterais, unilaterais à direita ou à esquerda, dependendo do objetivo. Indique qual seria a região crítica mais favorável às seguintes pessoas: (Justifique.)

- (a) proprietário da torrefação;
- (b) fabricante de A; e
- (c) fabricante de B.

32. Um médico deseja saber se uma certa droga reduz a pressão arterial média. Para isso, mediu a pressão arterial em cinco voluntários, antes e depois da ingestão da droga, obtendo os dados do quadro abaixo. Você acha que existe evidência estatística de que a droga realmente reduz a pressão arterial média? Que suposições você fez para resolver o problema?

Voluntário	A	B	C	D	E
Antes	68	80	90	72	80
Depois	60	71	88	74	76

33. Uma amostra de 100 lâmpadas elétricas produzidas pela fábrica A indica uma vida média de 1.190 horas, com desvio padrão de 90 horas. Uma amostra de 75 lâmpadas produzidas pela fábrica B indica uma vida média de 1.230 horas, com desvio padrão de 120 horas. Admitindo que as variâncias populacionais sejam diferentes, você acha que existe diferença entre as vidas médias populacionais das lâmpadas produzidas pelas fábricas A e B?

34. Queremos comparar dois métodos de ensino A e B. Dispomos de 40 crianças. Podemos proceder de duas maneiras:

(i) Sorteamos 20 crianças para compor uma classe, e as restantes formam outra classe. Aplicamos um método a cada classe e, depois, fazemos uma avaliação para todas as crianças a respeito do assunto ensinado.

(ii) Aplicamos inicialmente um teste de inteligência às 40 crianças. Numeramos as crianças de 1 a 40, segundo o resultado do teste.

Consideramos os 20 pares

$$(1, 2), (3, 4), \dots, (39, 40)$$

, e de cada par sorteamos uma criança para cada classe.

Obtemos, assim, duas classes de 20 crianças, homogêneas quanto à inteligência. Aplicamos um método a cada classe e depois avaliamos todas as crianças.

- (a) Qual a variável de observação em cada procedimento?
- (b) Quais as hipóteses estatísticas adequadas?
- (c) Qual o teste estatístico de decisão em cada caso?
- (d) Qual dos dois procedimentos você preferiria? Por quê?

35. De 400 moradores sorteados de uma grande cidade industrial, 300 são favoráveis a um projeto governamental, e de uma amostra de 160 moradores de uma cidade cuja principal atividade é o turismo, 120 são contra.

- (a) Você diria que a diferença de opiniões nas duas cidades é estatisticamente significante?
- (b) Qual seria um IC de 90% para a proporção de favoráveis ao projeto nas duas cidades? (Suponha que o número de pessoas nas duas cidades seja aproximadamente igual.)

36. Para verificar o grau de adesão de uma nova cola para vidros, preparam-se dois tipos de montagem: cruzado (A), em que a cola é posta em forma de X, e quadrado (B), em que a cola é posta apenas nas quatro bordas. Os resultados da resistência para duas amostras de 10 cada estão abaixo. Que tipo de conclusão poderia ser tirada?

Método A	16	14	19	18	19	20	15	18	17	18
Método B	13	19	14	17	21	24	10	14	13	15

37. Em um estudo para comparar os efeitos de duas dietas, A e B, sobre o crescimento, 6 ratos foram submetidos à dieta A, e 9 ratos à dieta B. Após 5 semanas, os ganhos em peso foram:

A	15	18	12	11	14	15			
B	11	11	12	16	12	13	8	10	13

- (a) Admitindo que temos duas amostras independentes de populações normais, teste a hipótese de que não há diferença entre as duas dietas, contra a alternativa que a dieta A é mais eficaz, usando o teste t de Student, no nível de $\alpha = 0,01$. Calcule $\hat{\alpha}$.
- (b) Efetue o teste usando a estatística de Wilcoxon, com $\alpha = 0,01$. Calcule $\hat{\alpha}$.

38. As amostras (X_1, \dots, X_{10}) e (Y_1, \dots, Y_{10}) de duas populações normais com médias μ_1 e μ_2 e mesma variância σ^2 forneceram as estatísticas:

$$\bar{X} = 80, S_1^2 = 16; \bar{Y} = 83, S_2^2 = 18.$$

Teste, com o nível $\alpha = 0,05$, a hipótese $H_0 : \mu_1 = \mu_2$ contra a alternativa $H_1 : \mu_1 < \mu_2$.

39. Em um estudo sobre um novo método para ensinar Matemática a alunos do primeiro grau, dez crianças foram selecionadas ao acaso de um grupo de 20 e ensinadas pelo novo método, enquanto as outras dez serviram como controle e ensinadas pelo método tradicional. Após dez semanas o desempenho dos alunos em um teste foi avaliado e obtiveram-se as seguintes notas:

Novo método	8,5	7,5	9,0	9,5	10,0	7,0	6,5	8,0	8,5	7,0
Controle	7,5	10,0	6,5	5,0	8,0	7,5	4,5	9,5	6,5	7,5

Teste, com nível $\alpha = 0,05$, a hipótese de que o novo método é mais eficaz, utilizando o teste t e o teste Wilcoxon. Obtenha $\hat{\alpha}$ em cada caso.

40. Seja $W_R = R_1 + \dots + R_n$ a soma dos postos dos controles. Qual o valor de $W_R + W_S$?
41. Se $n = 4$ e $m = 6$, prove que $P(W_S \geq 35) = P(W_S \leq 31)$, usando o fato que W_S é simétrica em torno de
- $$\mu = \frac{m(N + 1)}{2}.$$
42. Se $n = 4$ e $m = 6$, prove que $P(W_S \geq 35) = P(W_R \leq 20)$.
43. Para o CD-Placa, teste se a escova convencional é eficaz para remover a placa bacteriana. Calcule o valor-p do teste.
44. Para o CD-Temperaturas, teste se a temperatura média de Cananeia é igual à temperatura média de Ubatuba (suponha que as observações para cada cidade sejam independentes, embora saibamos que elas não são, pois temos dados de séries temporais).
45. Numa pesquisa sobre a opinião dos moradores de duas cidades, A e B, com relação a um determinado projeto, obteve-se:

Cidade	A	B
Número de entrevistados	400	600
Número de favoráveis	180	350

Construa um IC para a diferença de proporções de opiniões nas duas cidades.

46. Duas máquinas A e B, são usadas para empacotar pó de café. A experiência passada garante que o desvio padrão para ambas é de 10 g. Porém, suspeita-se que elas têm médias diferentes. Para verificar, sortearam-se duas amostras: uma com 25 pacotes da máquina A e outra com 16 pacotes da máquina B. As médias foram, respectivamente,

$$\bar{x}_A = 502,74g \quad e \quad \bar{x}_B = 496,60g.$$

Com esses números, e com o nível de 5%, qual seria a conclusão do teste

$$H_0 : \mu_A = \mu_B ?$$

47. Na região sul da cidade, 60 entre 400 pessoas preferem a bebida Meca-Mela entre as demais similares. Na região norte, a proporção é de 40 entre 225 entrevistados. Baseado no resultado dessa amostra, você diria que a proporção de todos os moradores nas duas regiões é a mesma? Use $\alpha = 0,05$.
48. Uma pesquisa mercadológica sobre fidedignidade a um produto foi realizada em dois anos consecutivos, com duas amostras independentes de 400 donas de casa em cada uma delas. A preferência pela marca em questão foi de 33% e 29%, respectivamente. Os resultados trazem alguma evidência de mudança de preferência?
49. No exemplo 13.12, suponha que se acredite que a proposta A chame a atenção em pelo menos 5% a mais do que a proposta B. Teste a validade de tal afirmação, isto é, teste

$$H_0 : p_A - p_B = 0,05 \text{ contra } H_1 : p_A - p_B > 0,05.$$