Estatística não paramétrica
O problema geral de duas amostras

Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná

Abril de 2025

No caso de testes de pares combinados e postos sinalizados os dados consistiam em duas amostras, mas cada elemento em uma amostra estava vinculado a um elemento particular da outra amostra por alguma unidade de associação.

Esta situação de amostragem pode ser descrita como um caso de duas amostras dependentes ou, alternativamente, como uma única amostra de pares de uma população bivariada.

Quando as inferências a serem tiradas são relacionadas apenas para a população de diferenças das observações emparelhadas, o primeiro passo na análise geralmente é fazer as diferenças das observações emparelhadas; isso deixa apenas um único conjunto de observações.

Portanto, esse tipo de dado pode ser classificado como um problema de uma amostra. Agora trataremos de dados que consistem em duas amostras aleatórias mutuamente independentes, ou seja, amostras aleatórias obtidas independentemente de cada uma das duas populações.

Não apenas os elementos dentro de cada amostra são independentes, mas também todos os elementos da primeira amostra são independentes de cada elemento na segunda amostra.

O universo consiste em duas populações, com funções de distribuição F_X e F_Y. Temos uma amostra aleatória de tamanho m extraída de X e outra amostra aleatória de tamanho n obtida independentemente da população Y,

$$X_1, X_2, \cdots, X_m \qquad \text{e} \qquad Y_1, Y_2, \cdots, Y_n.$$

Em geral, a hipótese de interesse no problema de duas amostras é saber se as duas amostras são extraídas de populações idênticas, ou seja,

$$H_0 : F_Y(x) = F_X(x), \quad \text{para todo } x.$$

Se estivermos dispostos a fazer suposições paramétricas sobre as formas das populações e assumirmos que as diferenças entre as duas populações ocorrem apenas com relação a alguns parâmetros, como as médias ou as variâncias, é possível derivar o teste de Neyman-Pearson.

Se assumirmos que as populações são normalmente distribuídas, o teste *t*-Student de duas amostras para igualdade de médias e o teste F-Fisher para igualdade de variâncias são, respectivamente, os melhores testes. Os desempenhos destes dois testes são bem conhecidos.

No entanto, esses e outros testes clássicos paramétricos podem ser sensíveis a violações dos pressupostos fundamentais inerentes à derivação e à construção desses testes. Quaisquer conclusões obtidas com esses testes são tão válidas quanto as hipóteses feitas.

Se houver razão para suspeitar de uma violação de qualquer um desses postulados ou se informações suficientes para julgar sua validade não estiverem disponíveis ou se um teste completamente geral de igualdade para distribuições não especificadas for desejado, algum procedimento não paramétrico está recomendado.

Na prática, outras suposições são frequentemente feitas sobre a forma das populações. Uma suposição comum é chamada de modelo de locação.

Este modelo assume que as populações X e Y são as mesmas em todos os outros aspectos, exceto, possivelmente, por uma mudança na quantidade desconhecida de, digamos, θ, ou que

$$F_Y(x) = P(Y \leq x) = P(X \leq x - \theta) = F_X(x - \theta), \quad \forall x, \forall \theta \neq 0$$

Isso significa que $X + \theta$ e Y têm a mesma distribuição ou que X é distribuído como $Y - \theta$. A população Y é então a mesma que a população X se $\theta = 0$, é deslocada para a direita se $theta > 0$ e é deslocada para a esquerda se $\theta < 0$.

Sob a hipótese de mudança, as populações têm a mesma forma e a mesma variância, e a quantidade do deslocamento θ deve ser igual à diferença entre as médias populacionais, $\mu_Y - \mu_X$, as medianas populacionais, $M_Y - M_X$, e de fato a diferença entre quaisquer dois parâmetros de locação ou quantis da mesma ordem.

Outra suposição sobre a forma da população é chamado de modelo de escala, este assume que as populações de X e Y são as mesmas, exceto possivelmente para um fator de escala positivo θ, que não é igual a um.

O modelo de escala pode ser escrito como

$$F_Y(x) = P(Y \leq x) = P(X \leq \theta x) = F_X(\theta x), \quad \forall x, \ \forall \theta > 0, \ \theta \neq 1.$$

Isto significa que X/θ e Y têm a mesma distribuição para qualquer θ positivo ou que X é distribuído como θY. Além disso, a variância de X é θ^2 vezes a variância de Y e a média de X é θ vezes a média de Y.

Independentemente do modelo assumido, o problema geral de duas amostras talvez seja o problema mais discutido na estatísticas não-paramétrica.

A hipótese nula é quase sempre formulada como populações idênticas, com a distribuição comum completamente não especificada, exceto pela suposição de que é uma função de distribuição contínua.

Assim, sob o caso nulo, as duas amostras aleatórias podem ser consideradas uma única amostra aleatória de tamanho $N = m + n$ extraídas da população comum, contínua, mas não especificada. Então a configuração ordenada combinada das m variáveis aleatórias X e as n variáveis Y na amostra é um dos $\binom{m+n}{m}$ arranjos possíveis igualmente prováveis.

Muitos testes estatísticos são baseados em alguma função desse arranjo combinado. O tipo de função mais apropriado depende do tipo de diferença que se espera detectar o que é indicado pela hipótese alternativa.

Uma abundância de alternativas razoáveis para H_0 pode ser considerada, mas o tipo mais fácil de analisar usando técnicas de distribuição livre declara alguma relação funcional entre as distribuições. As alternativas bilaterais mais gerais são

$$H_1 : F_Y(x) \neq F_X(x), \quad \text{para algum } x$$

e a correspondente alternativa unilateral geral é

$$H_1 F_Y(x) \geq F_X(x), \ \forall x \quad \text{ou} \quad H_1 : F_Y(x) > F_X(x), \ \text{para algum } x.$$

Neste último caso, geralmente dizemos que a variável aleatória X é estocasticamente maior que a variável aleatória Y. Se a alternativa particular de interesse é simplesmente uma diferença na locação, usamos a alternativa de locação ou o modelo de locação

$$H_0 : F_Y(x) = F_X(x - \theta), \ \forall x \text{ e algum } \theta \neq 0.$$

Sob o modelo de locação, Y é distribuído como $X + \theta$, de modo que Y é estocasticamente maior ou menor que X se, e somente, $\theta > 0$ ou $\theta < 0$. Da mesma forma, se apenas uma diferença na escala é de interesse, usamos a alternativa de escala

Sejam dois conjuntos de variáveis aleatórias independentes

$$X_1, \cdots, X_m \qquad \text{e} \qquad Y_1, \cdots, Y_n$$

combinados em uma única sequência ordenada, do menor para o maior, acompanhando quais observações correspondem à amostra X e quais à Y.

Assumindo que as suas distribuições de probabilidade são contínuas, uma ordenação única é sempre possível, uma vez que teoricamente laços não existem. Por exemplo, com $m = 4$ e $n = 5$, o arranjo pode ser

$$XYYXXYXYY$$

que indica na amostra agrupada o menor elemento ser um X, o segundo menor um Y, etc., e o maior um Y.

Sob a hipótese nula de distribuições idênticas

$$H_0 : F_Y(x) = F_X(x), \quad \forall x,$$

esperamos que as variáveis aleatórias X e Y sejam bem misturadas na configuração ordenada, uma vez que as $m + n = N$ variáveis aleatórias constituem uma única amostra aleatória de tamanho N da população comum.

Com uma corrida; definida como uma sequência de letras idênticas precedido e seguido por uma letra diferente ou nenhuma letra, o número total de corridas ou execuções na amostra agrupada ordenada é indicativo do grau de mistura. Em nosso arranjo *XYYXXYXYY*, o número total de corridas é igual a 6, o que mostra uma boa mistura de X e Y.

Um padrão de arranjo com poucas corridas sugeriria que esse grupo de N não é uma amostra aleatória única, mas sim composto por duas amostras de duas populações distintas.

Por exemplo, se a disposição fosse $XXXXYYYYY$, todos os elementos da amostra X serão menores que todos os elementos da amostra Y, haveria apenas duas corridas.

Essa configuração específica pode indicar não apenas que as populações não são idênticas, mas também que os X são estocasticamente menores que os Y. No entanto, a ordenação reversa também contém apenas duas corridas e, portanto, um critério de teste baseado somente no número total de corridas não pode distinguir esses dois casos.

O teste de corridas é apropriado principalmente quando a alternativa é completamente geral e bilateral, como em

$$H_1 : F_Y(x) \neq F_X(x), \qquad \text{para algum } x.$$

Definimos a variável aleatória R como o número total de corridas no arranjo ordenado combinado de m observações da variável aleatória X e n observações de Y.

Uma vez que poucas corridas tendem a desacreditar a hipótese nula, o teste de Wald-Wolfowitz (1940) para o nível de significância α tem região de rejeição a cauda inferior como

$$R \leq c_\alpha.$$

A constante c_α é escolhida como sendo o maior número inteiro satisfazendo

$$P(R \leq c_\alpha \,|\, H_0) \neq \alpha.$$

O *p*valor para o teste de corridas é dado por

$$P(R \leq R_0 \,|\, H_0),$$

onde R_0 é o valor observado da estatística do teste de corridas R.

Como as observações X e Y são dois tipos de objetos dispostos em uma sequência completamente aleatória, se H_0 for verdadeira, a distribuição de R sob a hipóteses nula é exatamente a mesma encontrada para o teste de aleatoriedade.

A distribuição foi desenvolvida e aqui substituímos n_1 e n_2 por m e n, respectivamente, supondo que os X são chamados de objetos do tipo 1 e os Y chamados de objetos do tipo 2.

Outras propriedades de R discutidas, incluindo os momentos e a distribuição nula assintótica, também são inalteradas.

A única diferença aqui é que a região crítica apropriada para a alternativa de populações diferentes é observarmos pouquíssimas corridas.

Exemplo

A distribuição normal padrão e a qui-quadrado com grandes graus de liberdade podem ser aproximadas. Investigamos a concordância entre estas duas distribuições para moderados graus de liberdade. Duas amostras aleatórias independentes foram geradas, cada uma de tamanho 8, uma da distribuição normal padrão e a outra da distribuição qui-quadrado com $=$ 18 graus de liberdade.

Os dados resultantes são os seguintes:

```
> dadosNormal=c(-1.91,-1.22,-0.96,-0.72,0.14,0.82,1.45,1.86)
> dadosQui2=c(4.90,7.25,8.04,14.10,18.30,21.21,23.10,28.12)
> dadosQui2p = (dadosQui2 - 18)/6
> dadosQui2p
[1] -2.183 -1.791 -1.660 -0.650  0.050  0.535  0.850  1.686
```

Exemplo

Antes de testar a hipótese nula de distribuições iguais, os dados da amostra qui-quadrado devem ser padronizados subtraindo-se a média $\nu = 18$ e dividindo pelo desvio padrão $\sqrt{2\nu} = \sqrt{36} = 6$.

```
> RunsTest(dadosNormal, dadosQui2p, alternative = "two.sided")

Wald-Wolfowitz Runs Test

data:  dadosNormal and dadosQui2p
runs = 12, m = 8, n = 8, p-value = 0.2005
alternative hypothesis:
          true number of runs is not equal the expected number
```

Aceitamos a suposição de igualdade das distribuições.

O teste de corridas de Wald-Wolfowitz é extremamente geral e consistente contra todos os tipos de diferenças nas populações (Wald e Wolfowitz, 1940).

A própria generalidade do teste sinaliza seu desempenho em relação a alternativas específicas. O poder assintótico pode ser avaliado usando a distribuição normal com momentos apropriados sob a alternativa, que são dados em Wolfowitz (1949).

Sua principal utilidade é em análises preliminares dos dados em que nenhuma forma particular de alternativa é formulada. Então, se a hipótese for rejeitada, estudos adicionais podem ser feitos com outros testes, na tentativa de classificar o tipo de diferença entre as populações.

A estatística Kolmogorov-Smirnor é outro teste de uma amostra que pode ser adaptado ao problema de duas amostras. Lembre-se de que, como critério de bondade de ajuste, esse teste comparou a função de distribuição empírica de uma amostra aleatória com uma distribuição hipotética.

Neste caso; a comparação é feita entre as funções de distribuição empíricas das duas amostras.

As estatísticas de ordem correspondentes às duas amostras, de tamanhos m e n das populações contínuas F_X e F_Y, são

$$X_{(1)}, X_{(2)}, \cdots, X_{(m)} \qquad \text{e} \qquad Y_{(1)}, Y_{(2)}, \cdots, Y_{(n)}.$$

Suas respectivas funções de distribuição empírica, denotadas por $\widehat{F}_m(x)$ e $\widehat{F}_n(x)$, são definidas como

$$\widehat{F}_m(x) = \begin{cases} 0, & \text{se} \quad x < X_{(1)} \\ \dfrac{k}{m}, & \text{se} \quad X_{(k)} \leq x < X_{(k+1)}, \quad k = 1, 2, \cdots, m-1 \\ 1, & \text{se} \quad x \geq X_{(m)} \end{cases}$$

e

$$\widehat{F}_n(x) = \begin{cases} 0, & \text{se} \quad x < Y_{(1)} \\ \dfrac{k}{n}, & \text{se} \quad Y_{(k)} \leq x < Y_{(k+1)}, \quad k = 1, 2, \cdots, n-1 \\ 1, & \text{se} \quad x \geq Y_{(n)} \end{cases}$$

Em um arranjo ordenado combinado das $m + n$ observações, $\widehat{F}_m(x)$ e $\widehat{F}_n(x)$, são as respectivas proporções de observações X e Y que não excedem o valor especificado x.

Se a hipótese nula

$$H_0 : F_Y(x) = F_X(x), \qquad \forall x$$

for verdadeira, as distribuições populacionais são idênticas e temos duas amostras da mesma população.

As distribuição empíricas são estimativas razoáveis das respectivas funções de distribuição. Portanto, permitindo a variação da amostragem, deve haver concordância razoável entre as duas distribuições empíricas se, de fato, H_0 for verdadeira. Caso contrário, os dados sugerem que H_0 não é verdadeira e, portanto, deve ser rejeitada.

Essa é a lógica intuitiva por trás da maioria dos testes de duas amostras e o problema é definir o que é uma concordância razoável entre as duas funções de distribuição empíricas.

Em outras palavras, quão próximas as duas distribuições empíricas devem estar, de modo que possam ser vistas como não significativamente diferentes, levando-se em consideração a variabilidade da amostragem.

Esta abordagem requer necessariamente uma definição de proximidade.

O teste de duas amostras Kolmogorov-Smirnov bilateral, denotado por $D_{m,n}$, é baseado na diferença absoluta máxima entre as duas distribuições empíricas

$$D_{m,n} = \max_x |\widehat{F}_m(x) - \widehat{F}_n(x)|.$$

Uma vez que aqui apenas as grandezas, e não as direções dos desvios são consideradas, $D_{m,n}$ é apropriado para uma alternativa geral bilateral

$$H_1 : F_Y(x) \neq F_X(x), \qquad \text{para algum } x$$

e a região de rejeição está na cauda superior, definida por $D_{m,n} \geq c_\alpha$ onde $P(D_{m,n} \geq c_\alpha \,|\, H_0) \leq \alpha$.

Por causa do teorema de Gilvenko-Cantelli, o teste é consistente para esta alternativa. O *p*valor é

$$P(D_{m,n} \geq D_0 \,|\, H_0),$$

onde D_0 é o valor observado da estatística de teste.

Como com a estatística do teste de Kolmogorov-Smirnov de uma amostra, $D_{m,n}$ é completamente de distribuição livre para qualquer distribuição contínua da população comum já que a ordem é preservada sob uma transformação monótona. Isso é, se fizermos $z = F(x)$ para a função de distribuição F comum, temos $\widehat{F}_m(z) = \widehat{F}_m(x)$ e $\widehat{F}_n(z) = \widehat{F}_n(x)$, em que a variável aleatória Z, correspondente para z, tem distribuição uniforme no intervalo unitário.

A derivação da distribuição nula exata de $D_{m,n}$ é geralmente atribuída à escola russa, particularmente Gnedenko (1954) e Korolyuk (1961); mas os artigos de Massey (1951, 1952) também são importantes.

Vários métodos de cálculo são possíveis, geralmente envolvendo fórmulas recursivas. Drion (1952) derivou uma expressão fechada para as probabilidades exatas no caso $m = n$, aplicando técnicas de reamostragem. Diversas abordagens estão resumidas em Hodges (1958).

Para a distribuição nula assintótica, ou seja, quando m e n se aproximam ao infinito de tal forma que m/n permaneça constante, Smirnov (1939) provou que

$$\lim_{m,n\to\infty} P\left(\sqrt{\frac{mn}{m+n}}D_{m,n} \leq d\right) = L(d),$$

onde

$$L(d) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2d^2}.$$

Note que a distribuição assintótica de $\sqrt{mn/(m+n)}D_{m,n}$ é exatamente a mesma que a distribuição assintótica de $\sqrt{N}D_N$. Isso não é surpreendente, já que sabemos do Teorema de Glivenco-Cantelli que, quando $n \to \infty$, $\widehat{F}_n(x)$ converge para $F_Y(x)$, que pode ser remarcada como $F_X(x)$. Então a única diferença aqui é no fator de normalização $\sqrt{mn/(m+n)}$, que substitui $\sqrt{N}$.

Exemplo

Mesmos dados utilizados no exemplo do teste de Wald-Wolfowitz.

```
> library(dgof)
> ks.test(dadosNormal, dadosQui2p)

Two-sample Kolmogorov-Smirnov test

data:  dadosNormal and dadosQui2p
D = 0.25, p-value = 0.9801
alternative hypothesis: two-sided
```

Acontece que *p*valores exatos não estão disponíveis para o caso de duas amostras se a alternativa for unilateral ou na presença de empates.

Exemplo

Se exact = NULL (o padrão), um *p*valor exato é calculado se o tamanho da amostra for menor que 100 no caso de uma amostra e não houver empates e se o produto dos tamanhos das amostras for menor que 10000 no caso de duas amostras. Caso contrário, distribuições assintóticas são usadas cujas aproximações podem ser imprecisas em pequenas amostras.

Os testes de Kolmogorov-Smirnov são úteis principalmente para as alternativas gerais, uma vez que o teste estatístico é sensível a todos os tipos de diferenças entre as funções de distribuição. Sua aplicação principal deve ser para estudos preliminares de dados. Os testes de Kolmogorov-Smirnov são mais poderosos do que os testes de corridas quando comparados para grandes tamanhos de amostra.

Para testar a hipótese nula de populações idênticas com duas amostras independentes, o teste de duas amostras de Kolmogorov-Smirnov compara as proporções de observações de cada amostra que não excede um número x para todos os números reais x.

O critério do teste foi a diferença máxima absoluta ou unidirecional entre os duas distribuições empíricas, que são definidas para todos os x. Suponha que em vez de usar todas as diferenças possíveis, escolhemos algumas arbitrárias mas num número específico δ e comparamos apenas as proporções de observações de cada amostra que são estritamente menores que δ. Como antes, as duas amostras independentes são denotadas por

$$X_1, X_2, \cdots, X_m \qquad \text{e} \qquad Y_1, Y_2, \cdots, Y_n.$$

Cada uma das $m + n = N$ observações deve ser classificada de acordo se é menor que δ ou não. Sejam U e V os respectivos números de observações X e Y menores que δ. Desde que as variáveis aleatórias em cada amostra foram dicotomizadas, U e V seguem a mesma distribuição binomial com parâmetros

$$p_X = P(X \leq \delta) \qquad \text{e} \qquad p_Y = P(Y < \delta),$$

e número de tentativas m e n, respectivamente.

Para duas amostras independentes, a distribuição conjunta de U e V é então

$$P(U = u, V = u) = \binom{m}{u}\binom{n}{v} p_X^u p_Y^v (1 - p_X)^{m-u}(1 - p_Y)^{n-v},$$

para $u = 0, 1, \cdots, m$ e $\nu = 0, 1, \cdots, n$.

As variáveis aleatórias U/m e V/n são estimativas pontuais não viciadas dos parâmetros p_X e p_Y, respectivamente. A diferença $U/m - V/n$ então é apropriada para testar a hipótese nula

$$H_0 : p_X - p_Y = 0.$$

A distribuição nula exata de $U/m - V/n$ pode ser encontrada e, para m e n grandes, sua distribuição pode ser aproximada pela normal.

A estatística de teste em qualquer um dos casos depende no valor comum $p = p_X = p_Y$, mas o teste pode ser realizado substituindo p por sua estimativa não viciada $(u + v)/(m + n)$. Caso contrário, não há dificuldade em construir um teste, embora aproximado, baseado no critério de diferença de proporções de observações menor que δ.

Este é essencialmente um teste de sinal modificado para duas amostras independentes, com a hipótese de que δ é o p-ésimo quantil em ambas as populações, onde p não é especificado, mas estimado a partir dos dados.

Quando as populações são consideradas idênticas mas não especificadas, não podemos escolher p e, em seguida, determinar o δ correspondente. Ainda δ deve ser conhecido pelo menos de forma posicional para classificar cada observação da amostra como menor do que δ ou não. Portanto, suponha que decidimos controlar a posição de δ em relação às magnitudes das observações da amostra.

Se as quantidades U e V forem fixadas pelo experimentador antes da amostragem p é, até certo ponto, controlada desde que $(u + v)/(m + n)$ é uma estimativa de p comum.

Se p denota a probabilidade de que qualquer observação seja menor que δ, a distribuição de $T = U + V$ é

$$P(T = t) = \binom{m+n}{t} p^t (1-p)^{m+n-t}, \quad t = 0, 1, 2, \cdots, m+n\cdot$$

A distribuição condicional de $U|T = t$ pode ser encontrada utilizando as expressões acima e, no caso nulo, quando $p = p_X = p_Y$, temos por resultado

$$P_{U|T}(u \,|\, t) = \frac{\binom{m}{u}\binom{n}{t-u}}{\binom{m+n}{t}}, \quad u = \max(0, t-n), 1, \cdots, \max(m, t),$$

a qual é a distribuição hipergeométrica.

Esse resultado também poderia ter sido argumentado diretamente da seguinte forma. Cada uma das $m + n$ observações é dicotomizada de acordo com δ, ou seja, se é menos ou não do que δ. Entre todas as observações, se $p = p_X = p_Y$, cada um dos $\binom{m+n}{t}$ conjuntos de números de t números é igualmente susceptível de compreender o grupo dos menores do que δ.

O número de conjuntos que tem exatamente u elementos da amostra X é $\binom{m}{u}\binom{n}{t-u}$. Como U/m é uma estimativa de p_X, se a hipótese $p = p_X = p_Y$ for verdadeira, u/m deve estar perto de $t/(m + n)$.

Um critério de teste pode ser encontrado usando a distribuição condicional de U para qualquer t escolhido.

O fato de que δ não pode ser determinado antes que as amostras sejam obtidas pode ser perturbador, pois implica que δ deve ser tratado como uma variável aleatória.

Ao derivar a distribuição condicional de $U|T$ tratamos δ como uma constante, mas o mesmo resultado é obtido para δ definido como a mediana da amostra. Vamos denotar por Z a mediana da amostra combinada e por F_X e F_Y as funções de distribuição de X e Y, respectivamente, e assumamos que N seja ímpar.

A mediana Z pode ser de uma das variáveis aleatórias X ou Y, e essas possibilidades são mutuamente exclusivas.

A função de densidade conjunta de U e Z para t observações menores que a mediana amostral onde $t = (N-1)/2$ é o limite, quando Δz se aproxima de zero, da soma das probabilidades de que

(1) os X estão divididos em três classificações, u menores que z, um entre z e $z + \Delta z$ e os restantes maiores que $z + \Delta z$ e os Y são divididos de tal forma que $t - u$ são menores que z e

(2) exatamente u dos X sejam menores que z e os Y sejam divididos de tal forma que $t - u$ sejam menores que z, um entre z e $z + \Delta z$ e os restantes sejam maiores que $z + \Delta z$.

O resultado então é

$$
\begin{aligned}
f_{U,Z}(u,z) \quad = \quad & \binom{m}{u,1,m-1-u} F_X^u(z) f_X(z) \big(1-F_X(z)\big)^{m-1-u} \\
& \binom{n}{t-u} F_Y^{t-u}(z) \big(1-F_Y(z)\big)^{n-t+u} + \\
& + \binom{m}{u} F_X^u(z) \big(1-F_X(z)\big)^{m-u} \binom{n}{t-u,1,n-t+u-1}^{t-u} \\
& F_Y^{t-u}(z) f_Y(z) \big(1-F_Y(z)\big)^{m-t+u-1}.
\end{aligned}
$$

A densidade marginal de U é obtida pela integração da expressão sobre todo z.

Por causa desse resultado podemos dizer que antes da amostragem, ou seja, antes que o valor de δ seja determinado, a estatística do teste da mediana é apropriado para a hipótese geral de populações idênticas, e depois que as amostras são obtidas, a hipótese testada é que δ seja o p-ésimo quantil em ambas as populações, onde p é um número próximo a 0.5.

As distribuições nulas da estatística de teste são as mesmas para ambas hipóteses, no entanto. O procedimento para duas amostras de medições independentes consiste em organizar as amostras combinadas em ordem crescente de magnitude e determinar a mediana amostral δ, a observação com classificação $(N+1)/2$ se N for ímpar e qualquer número entre as observações com classificações $N/2$ e $(N+2)/2$ caso N par.

Um total de t observações são, então menores que δ, onde $t = (N1)/2$ ou $N/2$ conforme N seja ímpar ou par. Seja U o número de observações X menores que δ.

Se as duas amostras são extraídas de populações contínuas idênticas, a função de probabilidade de U para t fixo é

$$f_U(u) = \frac{\binom{m}{u}\binom{n}{t-u}}{\binom{m+n}{t}},$$

onde $u = \max(0, t-n), \cdots, \min(m, t)$, $t = [N/2]$ sendo que $[x]$ denota o maior número inteiro que não excede o valor x. Se a hipótese nula é verdadeira, então $P(X < \delta) = P(Y < \delta)$ para todos os δ e em particular as duas populações têm mediana comum, que é estimada por δ.

Como U/m é um estimador de $P(X < \delta)$, que é aproximadamente metade sob H_0, um teste baseado no valor de U será mais sensível a diferenças de locação.

Se U for muito maior que $m/2$, a maior parte dos valores de X serão menores do que a maioria dos valores de Y. Isso dá credibilidade à relação $P(X < \delta) > P(Y < \delta)$, que são os X estocasticamente menores que os Y, de modo que a mediana da população X é menor que a mediana da população Y, ou que $\theta > 0$.

Se U é muito pequena em relação a $m/2$, a conclusão oposta está implícita. As regiões de rejeição apropriadas e os p-valores para o nível de significância nominal são então os seguintes:

Alternativa	**Região de rejeição**	p**-valor**
$Y \overset{\text{ST}}{>} X$, $\theta > 0$ ou $M_Y > M_X$	$U \geq c'_\alpha$	$P(U \geq U_0)$
$Y \overset{\text{ST}}{<} X$, $\theta < 0$ ou $M_Y < M_X$	$U \leq c'_\alpha$	$P(U \leq U_0)$
$\theta \neq 0$ ou $M_Y \neq M_X$	$U \leq c$ ou $U \geq c'$	$2\min\left(P(U \geq U_0), P(U \leq U_0)\right)$

onde c_α e c'_α são, respectivamente, os maiores e menores inteiros tais que $P(U \leq c_\alpha \,|\, H_0) \leq \alpha$ e $P(U \geq c_\alpha \,|\, H_0) \leq \alpha$, c_α e c'_α são dois inteiros $c_\alpha < c'_\alpha$ tais que

$$P(U \leq c \,|\, H_0) + P(U \geq c' \,|\, H_0) \leq \alpha$$

e U_0 é o valor observado de U, a estatística do teste da mediana.

Exemplo

O gerente de produção de uma pequena empresa que fabrica um determinado componente eletrônico acredita que tocar música contemporânea na área de produção ajudará a reduzir o número de itens não conformes produzidos. Um grupo de trabalhadores com antecedentes (treinamento, experiência, etc.) são selecionados e cinco deles são atribuídos, ao acaso, para trabalhar na área enquanto a música é tocada. Então, do restante do grupo, quatro trabalhadores são aleatoriamente designados para trabalhar da maneira usual sem música. O número de itens não conformes produzidos pelos trabalhadores durante um determinado período de tempo são dados abaixo. Teste para ver se o número mediano de itens não-conformes produzidos quando a música é tocada é menor do que quando nenhuma música é tocada.

Amostra sem música				Amostra com música				
3	4	9	10	1	2	5	7	8

Denotemos a amostra acima sem música como X e com música por Y, respectivamente.

Assuma o modelo de turnos e suponha que a hipótese nula a ser testada é $M_X = M_Y$ contra a alternativa $M_Y < M_X$. Então, o p-valor para o teste da mediana está na cauda esquerda. Como $N = 9$ é ímpar, $t = (91)/2 = 4$. A mediana da amostra combinada é igual a 5 e, portanto, $U = 2$. Usando f_U, o p-valor exato para o teste da mediana é

$$P(U \leq 2 \,|\, H_0) = \frac{\binom{4}{0}\binom{5}{4} + \binom{4}{1}\binom{5}{3} + \binom{4}{2}\binom{5}{2}}{\binom{9}{4}} = \frac{105}{126} = 0.8333.$$

Não há evidências suficientes em favor da alternativa H_1 e não rejeitamos H_0.

```
> phyper(2, 4, 5, 4)
[1] 0.8333333
```

Se m e n forem tão grandes que o cálculo para encontrar valores críticos não é viável, uma aproximação normal à distribuição hipergeométrica pode ser usada.

Usando expressões para a média e a variância da distribuição hipergeométrica e a distribuição f_U, a média e a variância de U são encontradas como sendo

$$\mathrm{E}(U \mid t) = \frac{mt}{N} \quad \text{e} \quad \mathrm{Var}(U \mid t) = \frac{mnt(N-t)}{N^2(N-1)}.$$

Se m e n crescerem ao infinito de tal forma que m/n permaneça constante, esta distribuição hipergeométrica se aproxima da distribuição binomial para t tentativas com parâmetro m/N, que por sua vez se aproxima a distribuição normal.

Para N grande, a variância de U é aproximadamente

$$\text{Var}(U \mid t) = \frac{mnt(N-t)}{N^3}$$

e assim a distribuição assintótica de

$$Z = \frac{U - \frac{mt}{N}}{\sqrt{\frac{mnt(N-t)}{N^3}}}$$

é aproximadamente normal padrão.

Uma correção de continuidade de 0.5 pode ser utilizada para melhorar a aproximação. Por exemplo, quando a alternativa é $\theta < 0$ ou $M_Y < M_X$, o p-valor aproximado com uma correção de continuidade é dado por

$$\Phi\left(\frac{U_0 + 0.5 - mt/N}{\sqrt{mnt(N-t)/N^3}}\right).$$

Exemplo

Gerente de produção no exemplo anterior. Continuação.

```
> Z = (2+0.5-4*4/9)/sqrt((5*4*4*(9-4))/9^3)
> Z
[1] 0.975
> pnorm(Z)
[1] 0.8352199
```

O p-valor é 0.8352199, levando à mesma conclusão.

Como o teste de corridas de Wald-Wolfowitz, o teste U de Mann-Whitney (Mann and Whitney, 1947) baseia-se na ideia de que o padrão particular exibido quando as variáveis aleatórias X e Y estão dispostas juntas em ordem crescente de magnitude fornece informações sobre a relação entre suas populações.

No entanto, em vez de medir a tendência de agrupar pelo número total de corridas, o critério de Mann-Whitney é baseado nas magnitudes de, digamos, os Y em relação aos X, ou seja, a posição dos Y na sequência combinada ordenada. Um padrão de arranjo de amostra onde a maioria dos Y é maior que a maioria dos X ou vice-versa, ou ambos seria uma evidência contra uma mistura aleatória e, assim, tenderia a desacreditar a hipótese nula de distribuições idênticas.

A estatística do teste U de Mann-Whitney é definida como o número de vezes que um Y precede um X no arranjo ordenado combinado das duas amostras aleatórias independentes

$$X_1, X_2, \cdots, X_m \qquad \text{e} \qquad Y_1, Y_2, \cdots, Y_n$$

em uma única sequência de $m + n = N$ variáveis aumentando em magnitude.

Vamos assumir que ambas amostras são extraídas de distribuições contínuas, de modo que a possibilidade de que $X_i = Y_j$ para alguns i e j não precisa ser considerada.

Se as mn variáveis aleatórias indicadoras forem definidas como

$$D_{ij} = \begin{cases} 1, & \text{se} \quad Y_j < X_i, \\ 0, & \text{se} \quad Y_j > X_i \end{cases} \quad i = 1, 2, \cdots, m; \; j = 1, 2, \cdots, n$$

a representação simbólica da estatística U de Mann-Whitney é

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} D_{ij}.$$

A região de rejeição lógica para a alternativa unilateral que os Y são estocasticamente maiores que os X,

$$H_1 : F_Y(x) \leq F_X(x),$$

com a desigualdade estrita para alguns x, seria claramente valores pequenos de U.

O fato de ser um critério de teste consistente pode ser mostrado investigando a convergência de U/mn para um determinado parâmetro, onde H_0 pode ser escrito como uma declaração sobre o valor desse parâmetro.

Para esse propósito, definimos

$$p = P(Y < X) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_Y(y) f_X(x) \mathrm{d}y \mathrm{d}x = \int_{-\infty}^{\infty} F_Y(x) \mathrm{d}F_X(x)$$

e o problema do teste de hipóteses pode ser redefinido em termos do parâmetro p. Se $H_0 : F_Y(x) = F_X(x)$ para todo x for verdadeira, então

$$p = \int_{-\infty}^{\infty} F_X(x) \mathrm{d}F_X(x) = 0.5.$$

Se, por exemplo, a hipótese alternativa for $H_1 : F_Y(x) \leq F_X(x)$, isto é, $Y \overset{ST}{>} X$, então $H_1 : p \leq 0.5$ para todo x e $p < 0.5$ para algum x. Assim, a hipótese nula de distribuições idênticas pode ser parametrizada para $H_0 : p = 0.5$ e a hipótese alternativa para $H_1 : p < 0.5$.

As mn variáveis aleatórias indicadoras são variáveis Bernoulli com momentos

$$\mathrm{E}(D_{ij}) = \mathrm{E}(D_{ij}^2) = p \qquad \text{e} \qquad \mathrm{Var}(D_{ij}) = p(1-p).$$

Para os momentos de conjuntos, notamos que essas variáveis aleatórias não são independentes sempre que os subscritos X ou os subscritos Y são comuns, de modo que

$$\mathrm{Cov}(D_{ij}, D_{hk}) = 0 \qquad \text{para } i \neq h \quad \text{e} \quad j \neq k$$

$$\text{Cov}(D_{ij}, D_{ik}) = p_1 - p^2 \; j \neq k \quad \text{e} \quad \text{Cov}(D_{ij}, D_{hj}) = p_2 - p^2 \; i \neq h,$$

sendo que os parâmetros adicionais introduzidos são

$$p_1 = P(Y_j < X_i \cap Y_k < X_i) = P(Y_j \text{ e } Y_k < X_i) = \int_{-\infty}^{\infty} F_y^2(x) \mathrm{d}F_X(x)$$

e

$$\begin{aligned} p_2 &= P(X_i > Y_j \cap X_h > Y_j) = P(X_i \text{ e } X_h > Y_j) \\ &= \int_{-\infty}^{\infty} \left(1 - F_X(y)\right)^2 \mathrm{d}F_Y(y). \end{aligned}$$

Como U foi definida como uma combinação linear de mn variáveis aleatórias, a média e variância de U são

$$\mathrm{E}(U) = \sum_{i=1}^{m}\sum_{j=1}^{n}\mathrm{E}(D_{ij}) = mnp,$$

e

$$\begin{aligned}\mathrm{Var}(U) = & \sum_{i=1}^{m}\sum_{j=1}^{n}\mathrm{Var}(D_{ij}) + \sum_{i=1}^{m}\sum_{1\le j\neq k\le n}\sum\mathrm{Cov}(D_{ij},D_{ik}) + \\ & + \sum_{j=1}^{n}\sum_{1\le i\neq h\le m}\sum\mathrm{Cov}(D_{ij},D_{hj}) + \\ & + \sum_{1\le i\neq h\le m}\sum\sum_{1\le j\neq k\le n}\sum\mathrm{Cov}(D_{ij},D_{hk}).\end{aligned}$$

Substituindo os valores correspondentes à variância, temos que $\mathrm{E}(U/mn) = p$ e $\lim_{m,n\to\infty} \mathrm{Var}(U/mn) = 0$ do qual concluímos que U/mn é um estimador consistente de p.

Com bases nestes resultados o teste de Mann-Whitney é consistente nas seguintes situações:

Alternativa		Região de rejeição
$p < 0.5$	$F_Y(x) \leq F_X(x)$	$U - mn/2 < k_1$
$p > 0.5$	$F_Y(x) \geq F_X(x)$	$U - mn/2 > k_2$
$p \neq 0.5$	$F_Y(x) \neq F_X(x)$	$U - mn/2 > k_3$

Para determinar o tamanho das regiões críticas do teste U de Mann-Whitney, devemos agora encontrar a distribuição de probabilidade nula de U.

Sob H_0, cada um dos $\binom{m+n}{m}$ arranjos das variáveis aleatórias em uma sequência combinada ocorre com igual probabilidade, de modo que

$$f_U(u) = P(U = u) = \frac{r_{m,n}(u)}{\binom{m+n}{m}},$$

onde $r_{m,n}(u)$ é o número de arranjos distinguíveis das m variáveis aleatórias X e n variáveis aleatórias Y, de modo que em cada sequência o número de vezes que um Y precede um X é exatamente u.

Os valores de u para os quais $f_U(u)$ é diferente de zero entre zero e mn, para as duas ordenações mais extremas em que cada x precede cada y e todo y precede cada x, respectivamente.

Primeiro notamos que a distribuição de probabilidade de U é simétrica em relação à média $mn/2$ sob a hipótese nula. Esta propriedade pode ser discutida da seguinte forma.

Para cada disposição particular z das m letras x e as n letras y, defina o arranjo conjugado z como a sequência z escrita para trás. Em outras palavras, se z denota um conjunto de números escritos do menor para o maior para o maior, z denota os mesmos números escritos do maior para o menor.

Todo y que precede um x em z segue então aquele x em z, de modo que se u é o valor da estatística de Mann-Whitney para z, mnu é o valor para z. Portanto, sob H_0 temos,

$$\begin{aligned} P\left(U - \frac{mn}{2} = u\right) &= P\left(U = \frac{mn}{2} + u\right) \\ &= P\left(U - \frac{mn}{2} = -u\right). \end{aligned}$$

Devido a essa propriedade de simetria, somente os valores críticos da cauda inferior precisam ser encontrados para um teste de um ou dois lados. Definimos a variável aleatória U como o número de vezes que um X precede um Y ou

$$U' = \sum_{i=1}^{m} \sum_{j=1}^{n} (1 - D_{ij})$$

e redefinimos as regiões de rejeição para testes de tamanho α correspondentes ao seguinte:

Alternativa		Região de rejeição
$p < 0.5$	$F_Y(x) \leq F_X(x)$	$U \leq c_\alpha$
$p > 0.5$	$F_Y(x) \geq F_X(x)$	$U' \leq c_\alpha$
$p \neq 0.5$	$F_Y(x) \neq F_X(x)$	$U \leq c_{\alpha/2}$ ou $U' \leq c_{\alpha/2}$

Para determinar o número c_α para qualquer m e n, podemos enumerar os casos começando com $u = 0$ e trabalhar até que, pelo menos, $\alpha\binom{m+n}{m}$ casos sejam contados.

Embora seja relativamente fácil adivinhar quais ordenamentos levarão aos menores valores de u, $(m + nm)$ aumenta rapidamente à medida que m e n aumentam. Algum método mais sistemáticos de geração de valores críticos é necessário para eliminar a possibilidade de ignorar alguns arranjos com u pequeno e aumentar a faixa viável de tamanhos de amostras e níveis de significância.

Uma relação de recorrência particularmente simples e útil pode ser derivada para a estatística de Mann-Whitney.

Considere uma sequência de $m+n$ letras sendo construídas adicionando uma letra à direita de uma sequência de $m+n-1$ letras. Se as $m+n-1$ letras consistirem em m letras x e $n-1$ letras y, a letra extra deve ser y. Mas se y for adicionado à direita, o número de vezes que y precede um x não é alterado.

Se a letra adicional é um x, o que seria o caso das $m1$ letras x e n letras y na sequência original, todos os y precedem este novo x e há n deles, de modo que u é aumentado por n. Essas duas possibilidades são mutuamente exclusivas. Usando a notação do numerador em f_U novamente, esta relação de recorrência pode ser expressa como

$$r_{m,n}(u) = r_{m,n-1}(u) + r_{m-1,n}(u-n)$$

e

$$
\begin{aligned}
f_U(u) &= p_{m,n}(u) = \frac{r_{m,n-1}(u) + r_{m-1,n}(u-n)}{\binom{m+n}{m}} \\
&= \frac{n}{m+n}\frac{r_{m,n-1}(u)}{\binom{m+n-1}{n-1}} + \frac{m}{m+n}\frac{r_{m-1,n}(u-n)}{\binom{m+n-1}{m-1}}
\end{aligned}
$$

ou

$$
(m+n)p_{m,n}(u) = np_{m,n-1}(u) + p_{m-1,n}(u-n)\cdot
$$

Esta relação recursiva vale para todos os $u = 0, 1, 2, \cdots, mn$ e todos valores inteiros m e n com condições iniciais e de fronteira.

Exemplo

Consideremos a mesma situação do gerente de produção no exemplo anterior. Vamos utilizar o teste U de Mann-Whitney.

```
> library(coin)
> dados = data.frame(Obs = c(3, 4, 9, 10, 1, 2, 5, 7, 8),
                     Musica = factor(c(rep("N",4), rep("S", 5))))
> wilcox_test(Obs ~ Musica, data = dados, distribution = "exact")

Exact Wilcoxon-Mann-Whitney Test

data:  Obs by Musica (N, S)
Z = 0.9798, p-value = 0.4127
alternative hypothesis: true mu is not equal to 0
```

Quando m e n são grandes, as distribuição assintótica pode ser usada. Como U é a soma de variáveis aleatórias distribuídas identicamente, embora dependentes, uma generalização do Teorema Central do Limite nos permite concluir que a distribuição nula de U padronizada se aproxima da normal padrão quando $m, n \to \infty$ de tal maneira que m/n permanece constante (Mann & Whitney, 1947).

Para fazer uso dessa aproximação, a média e a variância de U sob a hipótese nula deve ser determinada. Obtemos assim que

$$\mathrm{E}(U \,|\, H_0) = \frac{mn}{2} \qquad \text{e} \qquad \mathrm{Var}(U \,|\, H_0) = \frac{mn(N+1)}{2}.$$

A estatística de teste em amostras grandes é então

$$Z = \frac{U - mn/2}{\sqrt{mn(N+1)/12}},$$

cuja distribuição é aproximadamente normal padrão.

Esta aproximação foi encontrada razoavelmente precisa para tamanhos de amostra iguais tão pequenos quanto 6. Como U pode assumir apenas valores inteiros, uma correção de continuidade de 0.5 pode ser usada.

Exemplo

Continuação do exemplo anterior.

```
> wilcox_test(Obs ~ Musica, data = dados,
                        distribution = "asymptotic")

Asymptotic Wilcoxon-Mann-Whitney Test

data:  Obs by Musica (N, S)
Z = 0.9798, p-value = 0.3272
alternative hypothesis: true mu is not equal to 0
```