TESTES NÃO–PARAMÉTRICOS (para mediana/média)

Referências bibliográficas:

- G.E. Noether, 1983, *Introdução à Estatística, uma abordagem não–paramétrica*, 2a. edição, Editora Guanabara Dois.
- W.O. Bussab & P.A. Morettin, 2009, *Estatística Básica*, 6a. edição, Editora Saraiva.
- S. Siegel, 1975, *Estatística Não–Paramétrica*, McGraw–Hill.
- J.D. Gibbons, 1985, *Nonparametric Statistical Inference*, Marcel Dekker, 2nd edition.

Os métodos de estimação e testes de hipóteses, que estudamos até agora são chamados de **metódos paramétricos**, pois dizem respeito ao parâmetro da distribuição da população da qual se extraiu a amostra. Esses métodos baseiam-se em alguma suposição sobre a distribuição da variável de interesse na população. Geralmente, supõe-se que ela tenha distribuição normal.

Entretanto, nem sempre essas suposições são válidas e se aplicarmos um método paramétrico nesses casos, o erro da nossa estimativa poderá ser completamente desconhecido. Para tais situações temos os métodos denominados **metódos não–paramétricos**, nos quais nenhuma suposição sobre a distribuição da variável na população é feita. Naturalmente, há suposições básicas associadas à maioria dos testes não–paramétricos, mas essas suposições são em menor número e mais fracas que às associadas aos testes paramétricos.

A normalidade ou não da distribuição pode ser verificada fazendo-se o histograma dos dados e testes de aderência (quando se testa se uma variável tem distribuição normal, é chamado de teste de normalidade) disponíveis. Por um lado, um histograma assimétrico já é clara indicação que a distribuição pode não ser normal. Por outro lado, somente a simetria da distribuição não garante que ela seja normal. Por exemplo, se o histograma da variável for parecido com a distribuição uniforme ou com formato de uma parábola convexa, sua distribuição é simétrica, mas não normal.

As perguntas naturais que surgem são: *Por que não utilizar apenas os teste não–paramétricos ?* ou *Em que casos um método é melhor do que o outro ?* Para responder essas perguntas vejamos as vantagens e desvantagens dos métodos não–paramétricos.

Vantagens dos métodos não–paramétricos

1. As conclusões decorrentes dos testes não–paramétricos não dependem da forma da distribuição da variável na população.

2. Os métodos não–paramétricos são aplicáveis aos casos em que se desconheça a distribuição na população e o tamanho da amostra é pequeno, sendo, portanto, inviável de se aplicar o Teorema Limite Central.

3. Os métodos não–paramétricos podem ser aplicados tanto a dados quantitativos quanto a dados qualitativos[1] (ordinais ou mesmo nominais), visto que, em geral, os dados são transformados em *postos*, ou mesmos em *sinais*. Lembrem-se que os testes paramétricos que estudamos só se aplicam a dados quantitativos.

Desvantagens dos métodos não–paramétricos

1. Se todas as suposições associadas a um modelo paramétrico são satisfeitas, então o modelo paramétrico fornece estimativas mais precisas do parâmetro desconhecido, ou fornece menores probabilidades de erro nos testes de hipóteses, isto é, o teste paramétrico é mais poderoso (poder do teste melhor) quando comparado com um teste não–paramétrico equivalente, com mesmo tamanho de amostra.

[1] O teste de independência, visto em aula, é um exemplo de um teste não–paramétrico.

2. Para situações mais complexas nem sempre existe um teste não–paramétrico equivalente a um teste paramétrico.

3. Devido ao fato que os dados em geral são transformados em postos, critica-se a possibilidade de haver *desperdício de informação* contida na amostra.

4. Devido a particularidade dos teste não–paramétrico existe uma tabela de valores de significância para cada tipo de teste, e o acesso a essas tabelas nem sempre é fácil.

TESTES PARA A MEDIANA

Nos testes paramétricos que estudamos, a medida de tendência central considerada foi, muitas vezes, a média μ. Isso se deve ao fato que, em geral, estamos supondo que os dados provinham de uma distribuição normal ou podiam ser aproximados por uma normal (via Teorema Limite Central, para amostras grandes), e como a distribuição normal é simétrica em torno da média, a média e a mediana são iguais.

No entanto, quando a distribuição é assimétrica, uma medida de tendência central adequada é a mediana. Por isso, nos concentraremos em dois testes para a mediana de uma população.

(A) TESTE DO SINAL

Consideremos uma amostra aleatória $(X_1, \ldots, X_n)$ (isto é, consideramos que $(X_1, \ldots, X_n)$ são independentes e todas têm a mesma distribuição de probabilidade) de tamanho n extraída de uma população com distribuição qualquer, com mediana m, desconhecida.

Suposição: $P(X = m) = 0$.

(I) HIPÓTESES

As hipóteses a serem testadas são:

$$\begin{cases} H_o : \ m = m_o \\ H_1 : \ m > m_o \quad (m < m_o \text{ ou } m \neq m_o) \end{cases}$$

Note que pela definição de mediana, temos que $P(X > m) = P(X \leq m) = 0,50$, então, chamando $\theta = P(X > m_o)$, (1) é equivalente a

$$\begin{cases} H_o : \ \theta = 0,5 \\ H_1 : \ \theta > 0,5 \quad (\theta < 0,5, \ \theta \neq 0,5, \text{ respectivamente}). \end{cases}$$

(II) A ESTATÍSTICA DE TESTE

Façamos a seguinte transformação nos dados: as observações que forem maiores que m_o recebem o sinal "+" e as observações que forem menores que m_o recebem o sinal "−"

Denotemos por K o número de observações maiores que m_o, isto é, o número de "+". Portanto, como as observações são independentes, K tem distribuição [2] binomial (n, θ).

Se a hipótese nula H_o for verdadeira, isto é, se m_o for realmente a mediana, então $\theta = 0,5$, o que significa que a metade dos valores observados estariam abaixo de m_o e a outra metade acima dele. E nesse caso,
$K \sim \text{bin}(n; 0,5)$.

(III) REGIÃO CRÍTICA

Rejeitaremos H_o, ao nível de significância α, se K for muito grande (muito pequeno, muito pequeno ou muito grande, respectivamente com as hipóteses formuladas), isto é, se $\{K \geq c\}$ (ou $\{K \leq c_1\}$, $\{K \leq c_2 \text{ ou } K \geq c_3\}$, respectivamente); onde c é um número inteiro satisfazendo

$$P(X \geq c) = \sum_{k=c}^{n} \binom{n}{k} 0,5^n \leq \alpha \, .$$

[2] Note que K é uma variável aleatória pois seu valor varia de amostra para amostra.

As mudanças de acordo com a hipótese alternativa são as usuais.

(IV) VALOR AMOSTRAL E CONCLUSÃO

Rejeitamos H_o ao nível de significância α, se o valor numérico K_o de sinais "+" observados na amostra pertence à região crítica. Caso contrário, não há evidências suficientes para a rejeição de H_o.

EXEMPLO:
Uma máquina deve produzir arames com diâmetro de 1 milímetro. Para verificar se a máquina está ajustada de maneira adequada, 13 peças por ela produzidas são selecionadas e medidas com os seguintes resultados:

1,017 1,001 1,008 0,995 1,006 1,011 1,009 1,009 1,003 0,998 0,990 1,007 1,002

Com base nessa amostra, você diria que essa máquina precisa ser reajustada?

Resolução:
Como não se tem conhecimento algum sobre a distribuição do diâmetro dos arames produzidos pela máquina, usaremos um teste não–paramétrico. A máquina precisaria ser reajustada se a mediana dos diâmetros for diferente de 1,000. Portanto as hipóteses são:

$$\begin{cases} H_o: \ m = 1,000 \\ H_1: \ m \neq 1,000 \end{cases} \Longleftrightarrow \begin{cases} H_o: \ \theta = 0,5 \\ H_1: \ \theta \neq 0,5 \end{cases} .$$

onde $\theta = P(X > 1,000)$.

Se H_o for verdadeira, $K \sim \text{bin}(13; 0,5)$. Portanto, para $\alpha = 0,10$, a região crítica é $\{K \leq 3\}$ ou $\{K \geq 10\}$ (ver tabela da binomial).

Atribuindo os sinais, de acordo com as observações serem maiores ou menores do que 1,000 nós obtemos a tabela abaixo:

1,017	1,001	1,008	0,995	1,006	1,011	1,009	1,009	1,003	0,998	0,990	1,007	1,002
+	+	+	−	+	+	+	+	+	−	−	+	+

Então, na amostra, $K_o = 10$. Como K_o pertence à região crítica, rejeitamos H_o ao nível $\alpha = 0,1$, portanto a máquina deve ser reajustada.

==

Desde que o *teste do sinal* utiliza apenas os sinais das diferenças entre cada observação e o valor hipotético (em H_o) da mediana m_o, as magnitudes dessas observações são ignoradas. É razoável esperar que um teste que leve em conta essas magnitudes seja melhor que o teste de sinal. Isso de fato ocorre, porém necessitamos de uma suposição extra como veremos a seguir.

(B) TESTE DOS POSTOS SINALIZADOS DE WILCOXON

Analogamente ao teste anterior, consideremos uma amostra aleatória $(X_1, \ldots, X_n)$ de tamanho n extraída de uma população com distribuição qualquer, com mediana m, desconhecida. Adicionalmente supomos que a distribuição é simétrica[3].

Como queremos levar em conta a magnitude das observações em relação ao valor hipotético m_o da mediana, vamos considerar as diferenças $D_i = X_i - m_o$, $i = 1, \ldots, n$. Peguemos as diferenças em valores absolutos $|D_1|, |D_2|, \ldots, |D_n|$ e vamos ordená-las da menor para a maior diferença absoluta atribuindo os **postos** $1, 2, \ldots, n$. Mantenha registrado o sinal da diferença.

[3]Note que se a distribuição for simétrica, a média e a mediana são iguais, logo o Teste de Wilcoxon é na realidade um **teste para a média**. Portanto, esse teste é um concorrente ao teste *t-student* para a média de uma população com variância desconhecida. A aplicação do teste *t-student* ou do teste de *Wilcoxon* dependerá da suposição de normalidade ser satisfeita ou não.

Seja

$$\begin{cases} T^+ = & \text{a soma dos postos das diferenças positivas e} \\ T^- = & \text{a soma dos postos das diferenças negativas.} \end{cases} \tag{1}$$

Com a suposição de simetria, se H_o for verdadeira, as diferenças D_i são simetricamente distribuídas em torno do zero, de maneira que as diferenças positivas e negativas de mesma magnitude em valor absoluto tem a mesma probabilidade de ocorrência. Portanto, se m_o for a verdadeira mediana, T^+ e T^- devem ser aproximadamente iguais. Baseado nessa discussão apresentamos o teste em seguida.

Suposições:
(1) A distribuição na população é qualquer, com mediana m (desconhecida) e $P(X = m) = 0$;
(2) A distribuição na população é simétrica.

(I) HIPÓTESES

$$\begin{array}{l} H_o: \ m = m_o \\ H_1: \ m > m_o \quad (m < m_o \text{ ou } m \neq m_o). \end{array}$$

(II) A ESTATÍSTICA DE TESTE

A estatística de teste é T^+ ou T^- definidos em (1). Note que a soma de ambos é constante, isto é,

$$T^+ + T^- = 1 + 2 + \cdots + n = \sum_{k=1}^{n} k = \frac{n(n+1)}{2} \ .$$

Por isso podemos usar indiferentemente T^+ ou T^-, já que um pequeno valor de um implica num valor grande do outro (naturalmente entre 0 e $n(n+1)/2$).

(III) REGIÃO CRÍTICA

Se H_o for verdadeira, ambos devem ser aproximadamente iguais. Portanto, rejeitaremos H_o se encontrarmos um valor amostral de T^+ (ou T^-) muito grande ou pequeno (de acordo com a hipótese alternativa).

Por conveniência vamos considerar apenas valores pequenos, de modo que as respectivas regiões (críticas) de rejeição de H_o são:

$$\begin{array}{ll} T^- \leq t_\alpha & \text{para } H_1: m > m_o \ ; \\ T^+ \leq t_\alpha & \text{para } H_1: m < m_o \ ; \\ T^+ \leq t_{\alpha/2} \text{ ou } T^- \leq t_{\alpha/2} & \text{para } H_1: m \neq m_o \ . \end{array}$$

em que t_α é o número tal que $P(T \leq t_\alpha) = \alpha$ para $T = T^+$ ou T^-.

Os valores de t_α podem ser obtidos através da Tabela IX (de Bussab & Morettin, *Estatística Básica*, 6a.edição) em anexo. Na Tabela IX, n representa o tamanho da amostra, p corresponde aos possíveis valores de α (0,005; 0,01; 0,025; 0,05 e 0,10) e o valor de t_α encontra-se na respectiva coluna w_p, desde que estamos considerando apenas os valores pequenos para rejeição).

Por exemplo, para tamanho de amostra $n = 9$,

$$\begin{array}{lcl} \alpha = 0,01 & \Longrightarrow & t_\alpha = 4 \\ \alpha = 0,05 & \Longrightarrow & t_\alpha = 9 \\ \alpha = 0,15 & \Longrightarrow & t_\alpha = 11 \end{array}$$

(IV) VALOR AMOSTRAL E CONCLUSÃO

Obtenha na amostra os valores de T^+ e T^-. Se o correspondente valor amostral pertence à região crítica, então rejeitamos H_o ao nível de significância α, isto é, a mediana da distribuição não é m_o; caso contrário, não rejeitamos H_o.

EXEMPLO:
Tomemos o mesmo exemplo do teste anterior, *supondo que a distribuição é simétrica.*

Portanto as hipóteses são:

$$\begin{cases} H_o: \ m = 1,000 \\ H_1: \ m \neq 1,000 \end{cases}$$

Calculando as diferenças $D_i = X_i - 1,000$ obtemos a seguinte tabela:

	1,017	1,001	1,008	0,995	1,006	1,011	1,009	1,009	1,003	0,998	0,990	1,007	1,002
sinal	+	+	+	−	+	+	+	+	+	−	−	+	+
D_i	0,017	0,001	0,008	-0,005	0,006	0,011	0,009	0,009	0,003	-0,002	-0,010	0,007	0,002
$\|D_i\|$	0,017	0,001	0,008	0,005	0,006	0,011	0,009	0,009	0,003	0,002	0,010	0,007	0,002
posto	13	1	8	5	6	12	9,5	9,5	4	2,5	11	7	2,5

Para $\alpha = 0,10$ e pela Tabela IX ($n = 13, p = 0,05 = \alpha/2$) temos que a região crítica é dada por: $\{T^+ \leq 22$ ou $T^- \leq 22\}$.

Pelos resultados acima, o valor amostral de $T^- = 18,5$ e $T^+ = 72,5$, logo pertence à região de rejeição. Portanto, ao nível de significância de 10% concluímos que a máquina deve ser reajustada.

(C) TESTE DE WILCOXON - comparação de duas populações: amostras dependentes

Consideremos duas amostras nas quais as observações são pareadas, isto é, temos $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ amostras aleatórias de X e de Y, de tamanho n, em que os pares são independentes, mas há dependência entre cada par. Se definirmos a variável aleatória $D = X - Y$, teremos a amostra $D_1, D_2, \ldots, D_n$, correspondendo às diferenças entre os valores de cada par. Assim, o problema fica reduzido ao caso de uma amostra.

Se a distribuição de D for simétrica, então, novamente o Teste de Wilcoxon é na realidade um **teste para a média** (da diferença). Portanto, esse teste é um concorrente ao teste *t-student* para amostras pareadas - teste para a média de uma população com variância desconhecida. A aplicação do teste *t-student* ou do teste de *Wilcoxon* dependerá da suposição de normalidade ser satisfeita ou não.

Suposições:
(1) A distribuição da variável $D = X - Y$ população é qualquer, com mediana m (desconhecida) e $P(X = m) = 0$;
(2) A distribuição da variável $D = X - Y$ na população é simétrica, de modo que $m = \mu_D$.

As possíveis hipóteses nesse caso são dadas por

$$\begin{cases} H_o: \ \mu_X = \mu_Y \\ H_1: \ \mu_X \neq \mu_Y \ (\mu_X < \mu_Y \text{ ou } \mu_X > \mu_Y) \end{cases} \Longleftrightarrow \begin{cases} H_o: \ \mu_D = 0 \\ H_1: \ \mu_D \neq 0 \ (\mu_D < 0 \text{ ou } \mu_D > 0) \end{cases}$$

A estatística de teste (T^+ ou T^-) e região crítica são as mesmas do teste de Wilcoxon.

EXEMPLO:
Uma empresa deseja estudar o efeito de uma pausa de dez minutos para um cafezinho sobre 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 melhora na produtividade?

Operário	1	2	3	4	5	6
Sem intervalo (X)	23	35	29	33	43	32
Com intervalo (Y)	28	38	30	37	42	30

Resolução:
Para esse caso, considerando $D = X - Y$, as hipóteses adequadas são

$$\begin{cases} H_o: \ \mu_D = 0 \text{ (pausa não melhora a produtividade média)} \\ H_1: \ \mu_D < 0 \text{ (pausa melhora a produtividade média)} \end{cases}$$

Calculando as diferenças e atribuindo os postos, obtemos a seguinte tabela

Operário	1	2	3	4	5	6
$D_i = X_i - Y_i$	-5	-3	-1	-4	1	2
$\|D_i\|$	5	3	1	4	1	2
sinal	$-$	$-$	$-$	$-$	$+$	$+$
posto	6	4	1,5	5	1,5	3

Para $\alpha = 0,05$ e pela Tabela IX ($n = 6, p = 0,05$) temos que a região crítica é dada por: $\{T^+ \leq 3$ ou $T^- \leq 3\}$.

Pelos resultados acima, o valor amostral de $T^- = 16,5$ e $T^+ = 4,5$, logo T^+ não pertence à região de rejeição. Portanto, ao nível de significância de 5%, não há evidências suficiente para concluirmos que a pausa para café melhore a produtividade média.