TESTES NÃO–PARAMÉTRICOS (para mediana/média)

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

Entretanto, nem sempre essas suposições são válidas. Mais ainda, se um método paramétrico é aplicado quando as suposições não são válidas, pode-se chegar a uma conclusão incorreta e o erro dessa estimativa poderá ser completamente desconhecido. Para tais situações foram desenvolvidos 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 as suposições 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, o teste de aderência 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 segue a descrição de algumas vantagens e desvantagens dos métodos não–paramétricos.

Vantagens dos métodos não–paramétricos

- As conclusões decorrentes dos testes não–paramétricos não dependem da forma da distribuição da variável na população.
- 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 a aplicação do Teorema Limite Central).
- 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*. Note que os testes paramétricos estudados nessa disciplina foram aplicados para dados quantitativos.

Desvantagens dos métodos não–paramétricos

- 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.
- Para situações mais complexas nem sempre existe um teste não–paramétrico equivalente a um teste paramétrico.
- Devido ao fato que os dados em geral serem transformados em postos, critica-se a possibilidade de haver *desperdício de informação* contida na amostra.
- 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.

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

TESTES PARA A MEDIANA

Nos testes paramétricos vistos em aula, a medida de tendência central considerada foi, muitas vezes, a média μ. Isso se deve ao fato que, em geral, supõe-se que os dados são oriundos de uma amostra de uma variável X (na população) com uma distribuição normal ou, que a distribuição da média amostral ($\overline{X}$) pode ser aproximada por uma distribuição 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, a seguir serão apresentados dois testes para a mediana de uma população.

(A) TESTE DO SINAL

Considere uma amostra aleatória $(X_1, \ldots, X_n)$ (isto é, considere 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:

$$\begin{cases} H_o: \ m = m_o \\ H_1: \ m > m_o \quad (m < m_o \text{ ou } m \neq m_o) \end{cases} \tag{1}$$

Note que, pela definição de mediana, tem–se que $P(X > m) = P(X \leq m) = 0,50$. Portanto, denotando-se $\theta = P(X > m_o)$, as hipóteses estabelecidas na expressão (1) são equivalentes a

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

(II) A ESTATÍSTICA DE TESTE

Faça 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 "−".

Denote 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. Então, no caso de H_o ser verdadeira, tem-se $K \sim \text{bin}(n; 0,5)$.

(III) REGIÃO CRÍTICA

Rejeita-se H_o, ao nível de significância α, se K for muito grande (muito pequeno, muito pequeno ou muito grande, respectivamente com a hipótese alternativa formulada), isto é, se $\{K \geq c\}$ (ou $\{K \leq c_1\}$, $\{K \leq c_2 \ \text{ ou } \ K \geq c_3\}$, respectivamente); em que c é um número inteiro satisfazendo

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

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

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

(IV) VALOR AMOSTRAL E CONCLUSÃO

Rejeita-se 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, será aplicado um teste não–paramétrico (*qual suposição está sendo assumida?*). A máquina precisa ser reajustada se a mediana dos diâmetros for diferente de 1,000. Portanto as hipóteses são:

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

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

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

A tabela abaixo é obtida atribuindo-se o sinal adequado, de acordo com o fato da observação ser maior ou menor do que 1,000.

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
+	+	+	−	+	+	+	+	+	−	−	+	+

Na amostra tem-se que $K_o = 10$. Como esse valor pertence à região crítica, rejeita-se H_o ao nível $\alpha = 0,10$; portanto a máquina deve ser reajustada.

==========

Note que o *teste do sinal* utiliza apenas o sinal da diferença entre cada observação e o valor hipotético (em H_o) da mediana m_o, de modo que a **magnitude** de cada diferença/observação é ignorada.

A seguir é apresentado um teste que leva em conta essas magnitudes, além do sinal, tentando extrair mais informação da amostra. No entanto, esse teste requer uma suposição extra.

(B) TESTE DE WILCOXON (dos postos sinalizados)

Analogamente ao teste anterior, considere uma amostra aleatória $(X_1, \ldots, X_n)$ de tamanho n extraída de uma variável aleatória X na população com distribuição qualquer, com mediana m, desconhecida. Adicionalmente suponha que a distribuição de X é <u>simétrica</u>[3].

Como o objetivo é levar em conta a magnitude das observações em relação ao valor hipotético m_o da mediana, considera-se as diferenças $D_i = X_i - m_o$, $i = 1, \ldots, n$. Em seguida, ordena-se as diferenças

[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 de X estar satisfeita ou não.

em valores absolutos $|D_1|, |D_2|, \ldots, |D_n|$, da menor para a maior diferença absoluta, e atribui-se os **postos** $1, 2, \ldots, n$; mas mantendo o registro do sinal da diferença.

Considere então as estatísticas

$$\begin{cases} T^+ = \text{ a soma dos postos das diferenças positivas e} \\ T^- = \text{ a soma dos postos das diferenças negativas.} \end{cases} \tag{2}$$

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, têm a mesma probabilidade de ocorrência. Portanto, se m_o for a verdadeira mediana, T^+ e T^- devem ser aproximadamente iguais. Baseado nessa discussão o teste de Wilcoxon tem a seguinte metodologia.

Suposições:

(i) A distribuição da variável X na população é simétrica (X pode ter qualquer distribuição);
(ii) A distribuição da variável X na população tem mediana m (desconhecida) e $P(X = m) = 0$.

(I) HIPÓTESES

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

(II) A ESTATÍSTICA DE TESTE

A estatística de teste é T^+ ou T^- definidos na expressão (2). 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 pode-se usar indiferentemente T^+ ou T^-, já que um valor pequeno 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, rejeita-se H_o se o valor amostral de T^+ (ou T^-) for muito grande ou pequeno (de acordo com a hipótese alternativa).

Por conveniência, pode-se 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 valor tal que $P(T \leq t_\alpha) = \alpha$ para $T = T^+$ ou T^- (sem perda de generalidade, considera-se 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 se está considerando apenas os valores pequenos para rejeição).

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

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

(IV) VALOR AMOSTRAL E CONCLUSÃO

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

EXEMPLO:

Considere o mesmo exemplo do teste anterior, supondo, adicionalmente, que a distribuição de X é **simétrica.**

Portanto as hipóteses são:

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

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

Calculando as diferenças $D_i = X_i - 1,000$ obtém-se 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
$\lvert D_i \rvert$	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

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% conclui-se que a máquina deve ser reajustada.

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

Considere o caso de duas variáveis aleatórias X e Y, cujas amostras são observações pareadas, isto é, tem-se $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ como sendo amostras aleatórias de X e de Y, de tamanho n, em que os pares são independentes, mas há dependência entre cada par. Considerando então a variável aleatória $D = X - Y$, tem-se 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 normal com variância desconhecida*). A aplicação do teste t-Student ou do teste de *Wilcoxon* dependerá da suposição de normalidade de D ser satisfeita ou não.

Suposições:

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

As possíveis hipóteses nesse caso são dadas por

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

A estatística de teste (T^+ ou T^-) e as respectivas regiões críticas, de acordo com a hipótese alternativa, são as mesmas do teste de Wilcoxon.

EXEMPLO: (este é o exercício 3 da lista 7 - que você deve ter resolvido através do teste $t-$Student)

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 $D = X - Y$, e as suposições apropriadas, as hipóteses adequadas são

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

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

Calculando as diferenças e atribuindo os postos, a seguinte tabela é obtida

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

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

==========

Para várias outras situações há testes não-paramétricos adequados. Abaixo encontram-se algumas referências sobre o assunto.

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.

Tabela extraída da página 506 do livro *Estatística Básica* de W.O. Bussab & P.A. Morettin, 2009, 6a. edição, Editora Saraiva.

Tabela IX – Distribuição de Wilcoxon T^+ (sob H_0)

O corpo da tabela dá os valores w_p tais que $P(T^+ < w_p) = p$

	$w_{0,005}$	$w_{0,01}$	$w_{0,025}$	$w_{0,05}$	$w_{0,10}$		$w_{0,005}$	$w_{0,01}$	$w_{0,025}$	$w_{0,05}$	$w_{0,10}$
$n = 4$	0	0	0	0	1	$n = 27$	84	94	108	120	135
5	0	0	0	1	3	28	92	102	117	131	146
6	0	0	1	3	4	29	101	111	127	141	158
7	0	1	3	4	6	30	110	121	138	152	170
8	1	2	4	6	9	31	119	131	148	164	182
9	2	4	6	9	11	32	129	141	160	176	195
10	4	6	9	11	15	33	139	152	171	188	208
11	6	8	11	14	18	34	149	163	183	201	222
12	8	10	14	18	22	35	160	175	196	214	236
13	10	13	18	22	27	36	172	187	209	228	251
14	13	16	22	26	32	37	184	199	222	242	266
15	16	20	26	31	37	38	196	212	236	257	282
16	20	24	30	36	43	39	208	225	250	272	298
17	24	28	35	42	49	40	221	239	265	287	314
18	28	33	41	48	56	41	235	253	280	303	331
19	33	38	47	54	63	42	248	267	295	320	349
20	38	44	53	61	70	43	263	282	311	337	366
21	44	50	59	68	78	44	277	297	328	354	385
22	49	56	67	76	87	45	292	313	344	372	403
23	55	63	74	84	95	46	308	329	362	390	423
24	62	70	82	92	105	47	324	346	379	408	442
25	69	77	90	101	114	48	340	363	397	428	463
26	76	85	99	111	125	49	357	381	416	447	483
						50	374	398	435	467	504

Para $n > 50$, usa-se a aproximação normal, isto é T^+ tem distribuição aproximadamente normal com média e variância, respectivamente, dadas por

$$E(T^+) = \frac{n(n+1)}{4} \qquad \text{e} \qquad Var(T^+) = \frac{n(n+1)(2n+1)}{24}\,.$$

EXERCÍCIOS

1. Os resultados obtidos por dez estudantes numa prova foram os seguintes:

   72 95 79 83 93 80 91 74 70 86

   Teste a hipótese de que a nota média nessa prova é 75. Utilize dois testes (um paramétrico e outro não–paramétrico) diferentes e escreva as suposições feitas para aplicar cada teste.

2. Um laboratório fez sete determinações da porcentagem de gordura de salsichas de uma determinada marca. Os resultados foram os seguintes:

   19,9 18,5 19,8 19,2 20,5 19,1 19,6

   (a) Determine uma estimativa pontual do conteúdo mediano de gordura das salsichas desta marca.
   (b) Se o padrão de qualidade exige que o conteúdo mediano de gordura seja no máximo 19,25 %, você acha que as salsichas desta marca se encontra dentro dos padrões?

3. Uma companhia de processamento de dados mantém registros dos tempos totais gastos em tarefas individuais. Os tempos (em minutos) de 14 tarefas, selecionadas ao acaso, foram as seguintes:

   6 4,5 15 4 7 28 5,5 52 24,5 12 5 6,5 40 65

   Teste a hipótese de que o tempo mediano é menor que 10 min.

4. Uma companhia de testes verifica que as durações (em km) de 16 pneus de uma determinada marca foram as seguintes:

27.900	35.100	29.800	27.700	26.700	30.700	26.900	32.400
24.800	27.400	24.900	33.300	31.600	24.300	28.300	27.600

   Você acha que esses resultados estão de acordo com a afirmação de que os pneus dessa marca duram, em média, mais de 30.000 km ?

5. Dez lâmpada de um certo tipo são selecionadas ao acaso num grande carregamento de lâmpadas. Os tempos de duração (em horas) dessas lâmpadas são:

   220 2352 451 377 1561 257 1329 111 876 525

   (a) Estime o tempo mediano de vida das lâmpadas deste tipo.
   (b) Você acha que os dados estão de acordo com a afirmação do fabricante de que o tempo mediano de vida é maior do que 1000 horas?

6. Os diretores de uma seguradora estão considerando a possibilidade de oferecer um certo tipo de seguro. Porém esse projeto só se justifica se a renda familiar mediana[4] da população alvo for superior a 10.000 reais. Uma amostra de 18 famílias que se encontram na população alvo foi coletada e as correspondentes rendas se encontram na tabela abaixo.

9.300	50.000	15.500	7.000	12.700	9.600	15.000	35.000	11.800
7.600	21.000	19.500	17.500	6.500	14.100	9.800	12.400	8.000

   Com base nesses dados, você recomendaria aos diretores que o projeto fosse implantado?

7. Refaça os exercícios da Lista 5 (exerc. 3 e 4) e da Lista 7 (exerc. 5, 6 e 7) da disciplina, usando outro teste. Atente para as suposições necessárias para a resolução dos problemas, elas diferem das suposições feitas quando foram resolvidas nas respectivas listas?

[4]Em problemas ligados a renda considera-se a renda mediana em vez da renda média devido ao fato da distribuição de renda ser, em geral, bem assimétrica.