TESTE DA MEDIANA

UMA PERSPECTIVA VEROSSIMILHANCISTA

Esley Caminhas Ferreira Renan Carlos da Silva

Introdução

O **teste da mediana** é um teste não paramétrico para comparar medianas de dois grupos indepentes, útil para os casos em que são violadas as suposições de normalidade, há presença de *outliers* ou pequenas amostras.

Nesta apresentação, nosso objetivo é desenvolver o teste a partir de uma abordagem focada na verossimilhança, modelando a probabilidade de uma observação estar acima da mediana global utilizando o modelo binomial.

Artigos de inspiração

Nos baseamos sobretudo em 2 artigos:

- "*On the Asymptotic Efficiency of Certain Nonparametric Two-Sample Tests*" (Mood, 1950);
- "*The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypoteses*" (Wilks, 1938).

O primeiro formaliza o teste da mediana como um método não paramétrico, estabelecendo suas propriedades assintóticas, enquanto o último fala sobre a equivalência assintótitca entre o teste da razão de verossimilhanças e a distribuição χ^2.

Definição Formal da Verossimilhança

Considere dois grupos independentes A e B com observações:

- Grupo A: $Y_{A,1}, \dots, Y_{A,n_A}$
- Grupo B: $Y_{B,1}, \dots, Y_{B,n_B}$

Para cada grupo k, $k \in \{A, B\}$, definimos:

$$ \begin{aligned} p_k &= P(Y_k > \theta) \\ \gamma_k &= P(Y_k = \theta) \\ 1 - p_k - \gamma_k &= P(Y_k < \theta) \end{aligned} $$

onde θ é a mediana global dos dados combinados.

Hipóteses

Hipóteses:

- $H_0: \theta_A = \theta_B$ (As medianas populacionais são iguais)
- $H_1: \theta_A \neq \theta_B$

Isto é, sob H_0 temos:

$p_A = p_B = p = {N-m \over 2N}$,

onde m é o número de observações onde $Y = \theta$ e $N = n_A + n_B$.

Desenvolvendo a verossimilhança

Para cada observação $Y_{k,i}$, a contribuição para a verossimilhança é:

$$ \begin{split} L_{k,i}(p_k, \gamma_k) = & \ p_k \cdot I(Y_{k,i} > \theta) \cdot \gamma_k \cdot I(Y_{k,i} = \theta) \cdot \\ & (1-p_k-\gamma_k) \cdot I(Y_{k,i} < \theta) \end{split} $$

Para o grupo k:

$L_k(p_k, \gamma_k) = \prod_{i=1}^{n_k} L_{k,i} = (p_k)^{a_k} \cdot (\gamma_k)^{c_k} \cdot (1-p_k-\gamma_k)^{b_k}$

Desenvolvendo a Verossimilhança

onde:

$$ \begin{align*} a_k &= \sum_{i=1}^{n_k} I(Y_{k,i} > M) \\ c_k &= \sum_{i=1}^{n_k} I(Y_{k,i} = M) \\ b_k &= \sum_{i=1}^{n_k} I(Y_{k,i} < M) \end{align*} $$

Verossimilhança Conjunta

$$L(p_A, p_B, \gamma_A, \gamma_B) = L_A(p_A, \gamma_A) \cdot L_B(p_B, \gamma_B)$$

EMV: Com 4 parâmetros, teríamos certo trabalho para encontrar os estimadores de máxima verossimilhança. Após alguns cálculos, obtemos os estimadores que são intuitivos, isto é:

$$\hat{p}_k = {a_k \over n_k} \quad \text{e} \quad \hat{\gamma}_k = {c_k \over n_k}$$

Estimadores de Máxima Verossimilhança

Sob H_0:

$$\hat{p} = {a_A + a_B \over N} \quad \text{e} \quad \hat{\gamma} = {c_A + c_B \over N}$$

Teste da Razão de Verossimilhanças

$$ \Lambda = -2\ln \bigg({L_0(\hat{p}, \hat{\gamma}) \over L_1(\hat{p}_A, \hat{p}_B, \hat{\gamma}_A, \hat{\gamma}_B)} \bigg) \stackrel{a}{\sim} \chi_2^2 $$

Os graus de liberdade, nesse caso, são 2 porque é a diferença entre os 4 parâmetros que estimamos na verossimilhança irrestrita e os 2 que estimamos sob H_0.

Teste da Razão de Verossimilhanças

Sabemos que, como temos em geral dados contínuos (geralmente discretizados, em muitos casos) e, à medida que \(N \to \infty\), o seguinte resultado se aplica:

\[\gamma \stackrel{a}{\approx} 0 \]

Isso facilitará nosso procedimento, à medida que agora temos um caso binomial. Isto é: agora sem o termo \(\gamma\), a verossimilhança conjunta se reduz a:

\[L(p_A, p_B) = L(p_A) \cdot L(p_B) \]

Teste da Razão de Verossimilhanças

Portanto, o teste da razão de verossimilhanças fica:

$$ \Lambda = -2\ln \bigg(\frac{L_0(\hat{p})}{L_1(\hat{p}_A, \hat{p}_B)}\bigg) $$

Sendo que:

$$ \Lambda \stackrel{a}{\sim} \chi_1^2 $$

Observação

É importante notar que omitir \(\gamma\) torna o estimador viesado, mas ele ainda é assintoticamente não viesado.

Esta manipulação está em consonância com o teste tradicional, que ignora os valores que são iguais à mediana.

Exemplo prático

Para um caso prático, vamos usar microdados do ENEM 2023, disponíveis no portal de dados abertos do gov.com.br.

Neste exemplo, vamos considerar as médias das notas da prova de matemática por município, em 2 grupos de escolas: públicas e particulares.

Consideramos apenas alunos que compareceram a todos os dias de prova e não tiveram nenhum problema com o resultado (como redação anulada ou eliminação da prova, por exemplo). Também desconsideramos as escolas federais.

O objetivo é verificar se a nota média das escolas particulares é superior à das escolas públicas.

O conjunto de Dados

Primeiras 5 observações:

tipo_escola	municipio	media_matematica
Pública	Afonso Cláudio	507,6
Pública	Água Doce do Norte	425,1
Pública	Águia Branca	489,8
Pública	Alegre	454,5
Pública	Alfredo Chaves	432,7
Pública	Alto Rio Novo	447,4
Pública	Anchieta	494,3

tipo_escola	municipio	media_matematica
Pública	Apiacá	431,6
Pública	Aracruz	504,9
Pública	Atílio Vivácqua	450,0
Pública	Baixo Guandu	526,9
Pública	Barra de São Francisco	470,4
Pública	Boa Esperança	487,3
Pública	Bom Jesus do Norte	403,5
Pública	Brejetuba	420,2
Pública	Cachoeiro de Itapemirim	474,8

tipo_escola	municipio	media_matematica
Pública	Cariacica	473,9
Pública	Castelo	513,7
Pública	Colatina	471,3
Pública	Conceição da Barra	455,1
Pública	Conceição do Castelo	519,6
Pública	Divino de São Lourenço	467,7
Pública	Domingos Martins	533,0
Pública	Dores do Rio Preto	464,2
Pública	Ecoporanga	474,4
Pública	Fundão	485,7

tipo_escola	municipio	media_matematica
Pública	Governador Lindenberg	480,8
Pública	Guaçuí	473,8
Pública	Guarapari	476,4
Pública	Ibatiba	454,9
Pública	Ibiraçu	483,0
Pública	Ibitirama	438,8
Pública	Iconha	465,7
Pública	Irupi	508,8
Pública	Itaguaçu	447,0

tipo_escola	municipio	media_matematica
Pública	Itapemirim	463,0
Pública	Itarana	462,5
Pública	Iúna	489,7
Pública	Jaguaré	502,4
Pública	Jerônimo Monteiro	386,3
Pública	João Neiva	532,2
Pública	Laranja da Terra	469,0
Pública	Linhares	500,0
Pública	Mantenópolis	485,3
Pública	Marataízes	507,5

tipo_escola	municipio	media_matematica
Pública	Marechal Floriano	545,6
Pública	Marilândia	464,0
Pública	Mimoso do Sul	452,4
Pública	Montanha	466,9
Pública	Mucurici	431,7
Pública	Muniz Freire	489,5
Pública	Muqui	482,1
Pública	Nova Venécia	483,6
Pública	Pancas	447,4
Pública	Pedro Canário	445,1

tipo_escola	municipio	media_matematica
Pública	Pinheiros	398,3
Pública	Piúma	482,9
Pública	Ponto Belo	406,1
Pública	Presidente Kennedy	465,7
Pública	Rio Bananal	507,9
Pública	Rio Novo do Sul	508,4
Pública	Santa Leopoldina	380,7
Pública	Santa Maria de Jetibá	527,3
Pública	Santa Teresa	540,0
Pública	São Domingos do Norte	378,1

tipo_escola	municipio	media_matematica
Pública	São Gabriel da Palha	509,0
Pública	São José do Calçado	467,2
Pública	São Mateus	478,0
Pública	São Roque do Canaã	408,7
Pública	Serra	479,8
Pública	Sooretama	460,4
Pública	Vargem Alta	466,9
Pública	Venda Nova do Imigrante	478,3
Pública	Viana	475,5

tipo_escola	municipio	media_matematica
Pública	Vila Pavão	583,8
Pública	Vila Valério	504,6
Pública	Vila Velha	489,2
Pública	Vitória	499,2
Particular	Anchieta	415,2
Particular	Aracruz	578,1
Particular	Cachoeiro de Itapemirim	589,7
Particular	Cariacica	555,0
Particular	Castelo	563,0

tipo_escola	municipio	media_matematica
Particular	Colatina	585,6
Particular	Guaçuí	636,8
Particular	Guarapari	568,2
Particular	Linhares	573,0
Particular	Marataízes	561,7
Particular	Marilândia	302,7
Particular	Santa Maria de Jetibá	384,3
Particular	Santa Teresa	635,6
Particular	São Gabriel da Palha	440,9
Particular	São Mateus	598,5

Tabela de Contingência

Tipo de Escola	Acima	Abaixo	Total
Particular	15	4	19
Pública	33	44	77
Total	48	48	**96**

Estimadores de Máxima Verossimilhança

Estimador	Estimativa
$\hat{p}_{particular}$	0.789
$\hat{p}_{pública}$	0.428
$\hat{p}_{global}$	0.5

Teste da Razão de Verossimilhanças

```r
# Log-verossimilhança sob H0
log_L0 <- sum(tabela$acima) * log(p_hat_geral) +
  (sum(tabela$total) - sum(tabela$acima)) * log(1 - p_hat_geral)

# Log-verossimilhança sob H1
log_L1 <- tabela$acima[1] * log(p_hat_particular) +
  (tabela$total[1] - tabela$acima[1]) * log(1 - p_hat_particular) +
  tabela$acima[2] * log(p_hat_publica) +
  (tabela$total[2] - tabela$acima[2]) * log(1 - p_hat_publica)
```

```r
Lambda <- -2 * (log_L0 - log_L1)
p_valor <- pchisq(Lambda, df = 1, lower.tail = FALSE)

cat("Estatística do TRV =", round(Lambda, 4), "\np-valor =", round(p_valor,
```

```
Estatística do TRV = 8.3596
p-valor = 0.0038
```

Comparação com o teste tradicional

```r
teste_chi2 <- chisq.test(matrix(c(tabela$acima, tabela$abaixo), nrow = 2),
print(teste_chi2)
```

```
    Pearson's Chi-squared test

data:  matrix(c(tabela$acima, tabela$abaixo), nrow = 2)
X-squared = 7.9398, df = 1, p-value = 0.004836
```

Método	Estatística de Teste	P-valor
TRV	8.3596	0.0038
\(\chi^2\)	7.9398	0.0048