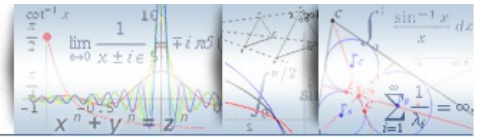
# O teste de uma hipótese univariada de normalidade por combinação de testes baseados na função característica empírica

Inês Alexandra Gonçalves Loureiro

DEPARTAMENTO DE MATEMÁTICA
FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE DE COIMBRA



# O teste de uma hipótese univariada de normalidade por combinação de testes baseados na função característica empírica

Inês Alexandra Gonçalves Loureiro

Dissertação para a obtenção do Grau de **Mestre em Matemática**

Área de Especialização em **Estatística, Otimização e Matemática Financeira**

**Júri**

| | |
|---|---|
| **Presidente:** | Prof. Dra. Nazaré Mendes Lopes |
| **Orientador:** | Prof. Dr. Carlos Tenreiro |
| **Vogal:** | Prof. Dra. Ana Cristina Rosa |

**Data: Agosto de 2014**

# Resumo

Em 1983, Epps e Pulley propuseram uma família de testes para uma hipótese univariada de normalidade baseada na distância ponderada $L_2$ entre a função característica empírica dos resíduos standardizados associados às observações e a função característica da distribuição normal standard, cuja ponderação depende de um parâmetro real $\beta$.

Neste trabalho, consideramos um procedimento de teste múltiplo que combina um número finito de estatísticas de teste de Epps e Pulley para escolhas extremas ($\beta \in \{0, \infty\}$) e não extremas ($\beta \in ]0, \infty[$) do parâmetro $\beta$. Para cada uma das estatísticas envolvidas no teste múltiplo bem como para este, estudamos as suas propriedades sob a hipótese nula de normalidade e estabelecemos a convergência dos testes associados.

Finalmente, apresentamos os resultados de um estudo de simulação realizado para analisar a potência do teste múltiplo e compará-la com outros testes de normalidade recomendados na literatura.

**Palavras Chave:** Normalidade univariada, função característica empírica, teste múltiplo, invariância para transformações afins, convergência.

# Abstract

In 1983, Epps and Pulley proposed a family of tests for assessing univariate normality based on a weighted $L_2$ distance between the empirical characteristic function of the scaled residuals and the characteristic function of the standard normal distribution, whose weight function depends on a real parameter $\beta$.

We consider a new multiple test procedure which combines a finite set of Epps and Pulley test statistics including extreme ($\beta \in \{0, \infty\}$) and non-extreme ($\beta \in ]0, \infty[$) choices of the tuning parameter $\beta$. For each of the statistics involved in the combination and for the multiple test, we study their main properties under the null hypothesis of normality and we establish the convergence of the associated tests.

Finally, we present the results of a simulation study carried out to analyze the power of the proposed multiple test and compare it with other highly recommended normality tests.

**Keywords:** Univariate normality, empirical characteristic function, multiple test, affine invariance, consistency.

# Agradecimentos

*Ao Professor Doutor Carlos Tenreiro, meu orientador nesta dissertação, agradeço por todo o conhecimento transmitido, pelas sugestões e críticas, pela total colaboração no solucionar de dúvidas que foram surgindo ao longo da realização deste trabalho e, essencialmente, pela sua paciência e disponibilidade.*

*Às minhas amigas, Patrícia, Sofia e Tatiana, agradeço pelo apoio, força e pelo constante interesse em diversos aspetos relacionados com este trabalho.*

*Ao André, agradeço pelos constantes desabafos, pelas palavras de motivação, pela confiança e por tentar melhorar os dias mais esmorecidos.*

*Por fim, mas não menos importante, agradeço aos meus pais pelo apoio incondicional e, principalmente, por me incentivarem perante os obstáculos. Agradeço ainda à minha avó pela constante preocupação mostrada ao longo de todo este processo.*

# Conteúdo

**Introdução** **ix**

**1 Testes de normalidade baseados em $D_{n,\beta}$, com $0 < \beta < \infty$** **1**
1.1 Aproximação assintótica para $nD_{n,\beta}$ . . . . . 1
1.2 A distribuição assintótica de $D_{n,\beta}$ sob $H_0$ . . . . . 5
1.3 A convergência do teste baseado em $D_{n,\beta}$ . . . . . 8

**2 Testes de normalidade baseados em $D_{n,\beta}$, com $\beta = 0$ e $\beta = \infty$** **11**
2.1 O caso $\beta \to 0$ . . . . . 11
2.2 O caso $\beta \to \infty$ . . . . . 15

**3 Um Teste Múltiplo de Normalidade** **21**
3.1 Introdução . . . . . 21
3.2 Um teste múltiplo de tipo Bonferroni . . . . . 22
3.3 Um teste múltiplo baseado nas estatísticas $D_{n,\beta}$ . . . . . 27

**4 Estudo de simulação** **31**
4.1 Os testes de Shapiro-Wilk e Anderson-Darling . . . . . 31
4.2 Distribuições alternativas . . . . . 32
4.3 Resultados de Potência . . . . . 32

**A Um teorema de convergência** **37**

**B Códigos em R** **39**

**Bibliografia** **45**

# Introdução

Os testes de ajustamento são procedimentos estatísticos que permitem inferir se a variável aleatória $X$ subjacente a um conjunto de observações é, ou não, bem modelada por uma dada distribuição ou família de distribuições fixas à partida. Representando por $\mathcal{F}_0$ essa família de distribuições e por $F$ a distribuição de $X$, pretendemos assim testar a hipótese nula $H_0 : F \in \mathcal{F}_0$ contra uma hipótese alternativa geral $H_a : F \notin \mathcal{F}_0$.

O teste de uma hipótese de normalidade é dos mais comuns na literatura e desde a invenção do teste de ajustamento do qui-quadrado por Pearson em 1900 (ver D'Agostino e Stephens, 1986, p. 63), diversos procedimentos para testar uma hipótese de normalidade têm sido desenvolvidos por diferentes autores. Entre estes procedimentos destacam-se os testes baseados nos coeficientes de assimetria e de curtose, assim como os testes de Shapiro-Wilk e de Anderson-Darling. Estes dois últimos testes são dos testes de normalidade mais recomendados (ver Thode, 2002, pp. 143–152).

Mais formalmente, se $X_1, \ldots, X_n$ são cópias independentes de uma variável aleatória absolutamente contínua $X$ com função densidade $f$ desconhecida, pretendemos assim testar a hipótese

$$H_0 : f \in \mathcal{N}$$

contra uma hipótese alternativa geral, onde $\mathcal{N}$ é a família das densidades de probabilidade normais sobre $\mathbb{R}$, isto é, $g \in \mathcal{N}$ se e só se

$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}, \ x \in \mathbb{R},$$

com $\mu \in \mathbb{R}$ e $\sigma > 0$.

Em 1983, Epps e Pulley introduziram uma família de estatísticas de teste para a hipótese $H_0$ baseadas na função característica empírica definida por

$$\Psi_n(t) = \frac{1}{n} \sum_{j=1}^{n} \exp(itY_j), \ t \in \mathbb{R},$$

onde

$$Y_j = \frac{X_j - \bar{X}_n}{s_n}, \tag{1}$$

são os resíduos standardizados associados a $X_1, \ldots, X_n$, e

$$\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j \quad \text{e} \quad s_n^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2$$

são a média e a variância amostrais, respetivamente. A estatística de teste de Epps e Pulley é definida através da distância quadrática ponderada entre a função característica empírica associada aos resíduos empíricos e a função característica da distribuição normal standard, isto é,

$$D_{n,\beta} = \int_{\mathbb{R}} \left| \Psi_n(t) - e^{-\frac{t^2}{2}} \right|^2 \varphi_\beta(t)dt,$$

onde $\varphi_\beta(t) = \frac{1}{\sqrt{2\pi\beta^2}} \exp\left(-\frac{t^2}{2\beta^2}\right), \beta > 0$, é a densidade da lei normal de média zero e variância $\beta^2$, que denotaremos por $\mathcal{N}(0, \beta^2)$. Esta função peso permite escrever a estatística de teste na forma simplificada

$$\begin{aligned} D_{n,\beta} = \frac{1}{n^2}\sum_{j,k=1}^{n} \exp\left(-\frac{\beta^2}{2}(Y_j - Y_k)^2\right) \\ - 2(1+\beta^2)^{-1/2}\frac{1}{n}\sum_{j=1}^{n} \exp\left(-\frac{\beta^2 Y_j^2}{2(1+\beta^2)}\right) + (1+2\beta^2)^{-1/2}. \end{aligned}$$

Da expressão anterior concluímos que a estatística $D_{n,\beta}$ é fácil de calcular e é ainda invariante para transformações afins dos dados, isto é,

$$D_{n,\beta}(aX_1 + b, \ldots, aX_n + b) = D_{n,\beta}(X_1, \ldots, X_n),$$

para todo o $a \in \mathbb{R} \setminus \{0\}$ e $b \in \mathbb{R}$.

A família de estatísticas $D_{n,\beta}$, $0 < \beta < \infty$, terá um papel fundamental no presente trabalho. O nosso objetivo não será o de considerar apenas um dos testes desta família para testar $H_0$. A ideia que desenvolvemos é a de tomar um número finito de tais estatísticas e combiná-las num único teste de normalidade.

No Capítulo 1 estudaremos a estatística de teste $D_{n,\beta}$, onde $\beta \in ]0, +\infty[$, fixo, obtendo o seu comportamento assintótico sob a hipótese nula e a convergência do teste nela baseado.

No Capítulo 2 identificamos as estatísticas que se obtêm de $D_{n,\beta}$ quando tomamos $\beta \to 0$ e $\beta \to \infty$, que denotaremos por $D_{n,0}$ e $D_{n,\infty}$, respetivamente. Tal como fizemos para $D_{n,\beta}$, com $0 < \beta < \infty$, obtemos a distribuição assintótica de $D_{n,0}$ e $D_{n,\infty}$ sob $H_0$ e estudamos a convergência dos testes de normalidade nelas baseados.

No Capítulo 3 é definido um procedimento de teste múltiplo para uma família de estatísticas quaisquer, sendo referidas as suas principais propriedades. Este procedimento é a seguir usado para combinar as estatísticas $D_{n,0}$ e $D_{n,\infty}$ correspondentes aos valores extremos $\beta = 0$ e $\beta = \infty$ do parâmetro $\beta$ e as estatísticas $D_{n,\beta_1}$ e $D_{n,\beta_2}$, com $0 < \beta_1 < \beta_2 < \infty$, correspondentes a dois valores não extremos do parâmetro $\beta$. Desta forma, esperamos que o teste combinado usufrua de boas propriedades de potência para um conjunto vasto de alternativas.

Por último, no Capítulo 4 apresentamos um breve estudo de simulação para avaliar a potência do teste múltiplo anterior, comparando-a com a de outros testes de normalidade existentes na literatura.

Durante este trabalho, denotaremos por $\xrightarrow{p}$ e $\xrightarrow{d}$ as convergências em probabilidade e em distribuição, respetivamente.

# Capítulo 1

# Testes de normalidade baseados em $D_{n,\beta}$, com $0 < \beta < \infty$

Neste capítulo, pretendemos estabelecer o comportamento assintótico da estatística

$$D_{n,\beta} = \int_{\mathbb{R}} \left| \Psi_n(t) - e^{-\frac{t^2}{2}} \right|^2 \varphi_\beta(t) dt$$

sob a hipótese nula de normalidade e a convergência do teste baseado nesta estatística, para valores fixos de $\beta$, $0 < \beta < \infty$.

Na Secção 1.1, começamos por apresentar aproximações assintóticas da estatística $D_{n,\beta}$ que nos permitirão obter a sua distribuição assintótica (Secção 1.2). Finalmente, na Secção 1.3, vamos estabelecer a convergência do teste baseado na estatística estudada nas secções anteriores.

Durante este trabalho admitiremos que $X_1, \ldots, X_n$ são variáveis reais, independentes e identicamente distribuídas (i.i.d.).

## 1.1. Aproximação assintótica para $nD_{n,\beta}$

Tal como Henze e Wagner (1997), para estabelecer a convergência em distribuição da estatística $nD_{n,\beta}$, começamos por representá-la da seguinte forma:

**Proposição 1.1.** *Se $X_1, \ldots, X_n$ são variáveis aleatórias reais quaisquer, então*

$$nD_{n,\beta} = \int Z_n(t)^2 \varphi_\beta(t) dt,$$

*onde*

$$Z_n(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left\{ \cos(tY_j) + \sin(tY_j) - e^{-\frac{t^2}{2}} \right\}.$$

*Demonstração.* Temos

$$\begin{aligned} & n\left|\Psi_n(t) - e^{-\frac{t^2}{2}}\right|^2 \\ &= n\left|\frac{1}{n}\sum_{j=1}^{n}\left\{\cos(tY_j) + i\sin(tY_j) - e^{-\frac{t^2}{2}}\right\}\right|^2 \\ &= \frac{1}{n}\sum_{j,k=1}^{n}\left\{\cos(tY_j) - e^{-\frac{t^2}{2}} + i\sin(tY_j)\right\}\left\{\cos(tY_j) - e^{-\frac{t^2}{2}} - i\sin(tY_j)\right\} \\ &= \frac{1}{n}\sum_{j,k=1}^{n}\left\{\left(\cos(tY_j) - e^{-\frac{t^2}{2}}\right)\left(\cos(tY_k) - e^{-\frac{t^2}{2}}\right) + \sin(tY_j)\sin(tY_k)\right\} \end{aligned}$$

e

$$\begin{aligned} Z_n(t)^2 &= \frac{1}{n}\left(\sum_{j=1}^{n}\left\{\cos(tY_j) - e^{-\frac{t^2}{2}} + \sin(tY_j)\right\}\right)^2 \\ &= \frac{1}{n}\sum_{j,k=1}^{n}\left\{\left(\cos(tY_j) - e^{-\frac{t^2}{2}}\right)\left(\cos(tY_k) - e^{-\frac{t^2}{2}}\right)\right. \\ &\qquad \left. + 2\left(\cos(tY_j) - e^{-\frac{t^2}{2}}\right)\sin(tY_k) + \sin(tY_j)\sin(tY_k)\right\} \end{aligned}$$

Basta agora ter em conta que

$$\int\left(\cos(tx) - e^{-\frac{t^2}{2}}\right)\sin(ty)\varphi_\beta(t)dt = 0, \text{ para todo o } x, y \in \mathbb{R}.$$

□

Para simplificar a notação que usaremos nos resultados seguintes escreveremos

$$Z_n = Z_n^* + o_{2,p}(1) \tag{1.1}$$

sempre que se verificar

$$\int \left(Z_n(t) - Z_n^*(t)\right)^2 \varphi_\beta(t)dt \xrightarrow{p} 0,$$

onde $Z_n(t) = Z_n(\omega; t)$ e $Z_n^*(t) = Z_n^*(\omega; t)$ são funções mensuráveis e limitadas definidas em $\Omega \times \mathbb{R}$ com valores em $\mathbb{R}$. Reparemos que sempre que $Z_n = Z_n^* + o_{2,p}(1)$ então

$$\int Z_n(t)^2\varphi_\beta(t)dt = \int Z_n^*(t)^2\varphi_\beta(t) + o_p(1). \tag{1.2}$$

Atendendo a que $D_{n,\beta}$ é invariante para transformações afins dos dados podemos, sem perda de generalidade, assumir que $X_1, \ldots, X_n$ são variáveis aleatórias de média 0 e variância 1. Tendo em conta a Proposição 1.1, o resultado seguinte permite obter uma primeira aproximação para a estatística $D_{n,\beta}$.

**Lema 1.2.** *Se $X_1, \ldots, X_n$ são variáveis aleatórias de média 0 e variância 1, então é válida a aproximação*

$$Z_n = Z_n' + o_{2,p}(1),$$

*onde*

$$Z_n'(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left\{ \cos(tX_j) + \sin(tX_j) + t(Y_j - X_j)(\cos(tX_j) - \sin(tX_j)) - e^{-\frac{t^2}{2}} \right\}.$$

*Demonstração.* Usando a fórmula de Taylor com resto de segunda ordem, temos

$$\cos(tY_j) = \cos(tX_j) - t\sin(tX_j)(Y_j - X_j) - \frac{t^2}{2}\cos(t\tilde{X}_j)(Y_j - X_j)^2$$

e

$$\sin(tY_j) = \sin(tX_j) + t\cos(tX_j)(Y_j - X_j) - \frac{t^2}{2}\sin(t\tilde{\tilde{X}}_j)(Y_j - X_j)^2$$

onde $\tilde{X}_j$ e $\tilde{\tilde{X}}_j$ pertencem ao intervalo de extremidades $X_j$ e $Y_j$, o que nos leva a concluir que

$$Z_n(t) - Z_n'(t) = -\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{t^2}{2}(Y_j - X_j)^2 \left\{ \cos(t\tilde{X}_j) + \sin(t\tilde{\tilde{X}}_j) \right\}.$$

Atendendo agora a que

$$Y_j - X_j = (s_n^{-1} - 1)X_j - s_n^{-1}\bar{X}_n, \tag{1.3}$$

ficamos com

$$\begin{aligned} \left| Z_n(t) - Z_n'(t) \right| &\leq \frac{1}{\sqrt{n}} \sum_{j=1}^{n} t^2 (Y_j - X_j)^2 \\ &= \frac{t^2}{\sqrt{n}} \sum_{j=1}^{n} \left\{ (s_n^{-1} - 1)^2 X_j^2 + s_n^{-2}\bar{X}_n^2 - 2(s_n^{-1} - 1)s_n^{-1} X_j \bar{X}_n \right\} \\ &= t^2 R_n, \end{aligned}$$

onde

$$R_n = \sqrt{n} \left\{ (s_n^{-1} - 1)^2 \frac{1}{n} \sum_{j=1}^{n} X_j^2 + s_n^{-2}\bar{X}_n^2 - 2(s_n^{-1} - 1)s_n^{-1}\bar{X}_n^2 \right\}.$$

Para concluir a demonstração, basta ter em conta que

$$R_n \xrightarrow{p} 0, \ n \to \infty,$$

uma vez que $X_1, \ldots, X_n$ são variáveis centradas e reduzidas. □

Uma segunda aproximação para a estatística $D_{n,\beta}$ pode ser obtida a partir do lema seguinte.

**Lema 1.3.** *Se $X_1, \ldots, X_n$ são variáveis aleatórias normais de média 0 e variância 1, então é válida a aproximação*

$$Z'_n = Z''_n + o_{2,p}(1),$$

*onde*

$$Z''_n(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left\{ \cos(tX_j) + \sin(tX_j) - \left( \frac{1}{2}(X_j^2 - 1)t^2 + X_j t + 1 \right) e^{-\frac{t^2}{2}} \right\}.$$

*Demonstração.* Atendendo a (1.3), concluímos que

$$\begin{aligned} Z'_n(t) &= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \Big\{ \cos(tX_j) + \sin(tX_j) + t(s_n^{-1} - 1)X_j\{\cos(tX_j) - \sin(tX_j)\} \\ &\qquad -ts_n^{-1}\bar{X}_n\{\cos(tX_j) - \sin(tX_j)\} - e^{-\frac{t^2}{2}} \Big\} \\ &= Z_n^*(t) + \frac{1}{\sqrt{n}}(s_n^{-1} - 1)t \sum_{j=1}^{n} X_j\{\cos(tX_j) - \sin(tX_j)\} \\ &\qquad - \frac{1}{\sqrt{n}} s_n^{-1}\bar{X}_n t \sum_{j=1}^{n} \{\cos(tX_j) - \sin(tX_j)\} \\ &= Z_n^*(t) + \sqrt{n}(s_n^{-1} - 1)tA_n(t) - \sqrt{n}s_n^{-1}\bar{X}_n t B_n(t), \end{aligned} \tag{1.4}$$

onde $Z_n^*$ é definida por

$$Z_n^*(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left\{ \cos(tX_j) + \sin(tX_j) - e^{-\frac{t^2}{2}} \right\},$$

$$A_n(t) = \frac{1}{n} \sum_{j=1}^{n} X_j\{\cos(tX_j) - \sin(tX_j)\}$$

e

$$B_n(t) = \frac{1}{n} \sum_{j=1}^{n} \{\cos(tX_j) - \sin(tX_j)\}.$$

Vamos verificar que

$$tA_n(t) = -t^2 e^{-\frac{t^2}{2}} + o_{2,p}(1) \tag{1.5}$$

e que

$$tB_n(t) = te^{-\frac{t^2}{2}} + o_{2,p}(1). \tag{1.6}$$

Demonstremos apenas a primeira das igualdades anteriores estabelecendo que

$$E\left[ \int \left( tA_n(t) + t^2 e^{-\frac{t^2}{2}} \right)^2 \varphi_\beta(t) dt \right] \xrightarrow{p} 0.$$

Com efeito, usando o Teorema de Tonelli temos

$$\begin{aligned} & E\left[\int \left(tA_n(t)+t^2e^{-\frac{t^2}{2}}\right)^2 \varphi_\beta(t)dt\right] \\ &= \int t^2 E\left(A_n(t)+te^{-\frac{t^2}{2}}\right)^2 \varphi_\beta(t)dt \\ &= \frac{1}{n}\int \left(1-t^2e^{-t^2}\right) t^2\varphi_\beta(t)dt \xrightarrow{p} 0, n\to\infty. \end{aligned}$$

Da igualdade (1.5) e usando o facto de

$$s_n^{-1}-1 = -\frac{1}{2n}\sum_{j=1}^{n}(X_j^2-1)+O_p(n^{-1}),$$

temos

$$\sqrt{n}(s_n^{-1}-1)tA_n(t) = \frac{1}{2\sqrt{n}}\sum_{j=1}^{n}(X_j^2-1)t^2e^{-\frac{t^2}{2}}+o_{2,p}(1). \tag{1.7}$$

Da igualdade (1.6) temos também que

$$\sqrt{n}s_n^{-1}\bar{X}_ntB_n(t) = \sqrt{n}\bar{X}_nte^{-\frac{t^2}{2}}+o_{2,p}(1). \tag{1.8}$$

Usando agora (1.7) e (1.8) em (1.4), obtemos

$$\begin{aligned} Z_n'(t) &= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left\{\cos(tX_j)+\sin(tX_j)-e^{-\frac{t^2}{2}}\right\} \\ &\quad + \frac{1}{2\sqrt{n}}\sum_{j=1}^{n}(X_j^2-1)t^2e^{-\frac{t^2}{2}} - \frac{1}{\sqrt{n}}\sum_{j=1}^{n}X_jte^{-\frac{t^2}{2}}+o_{2,p}(1) \\ &= Z_n''(t)+o_{2,p}(1), \end{aligned}$$

onde $Z_n''(t)$ é definida no Lema 1.3. □

Atendendo aos resultados anteriores podemos finalmente obter a anunciada aproximação assintótica.

**Teorema 1.4.** *Se $X_1,\ldots,X_n$ são variáveis aleatórias normais, centradas e reduzidas, então é válida a aproximação assintótica*

$$nD_{n,\beta} = \int Z_n''(t)^2\varphi_\beta(t)dt+o_p(1),$$

*onde $Z_n''(t)$ é definida no Lema 1.3.*

## 1.2. A distribuição assintótica de $D_{n,\beta}$ sob $H_0$

Neste momento é possível mostrar que $nD_{n,\beta}$ é assintoticamente equivalente a uma V-estatística degenerada, o que nos permitirá obter a sua distribuição assintótica.

**Teorema 1.5.** *Se $X_1, \ldots, X_n$ são variáveis aleatórias normais centradas e reduzidas, temos*

$$nD_{n,\beta} = \frac{1}{n}\sum_{i,j=1}^{n} h_*(X_i, X_j) + o_p(1),$$

*onde o núcleo $h_*$ é dado por*

$$h_*(x,y) = \int h(x;t)h(y;t)\varphi_\beta(t)dt, \ \ com\ x, y, t \in \mathbb{R}$$

*e*

$$h(x;t) = \cos(tx) + \sin(tx) + \left(\frac{1}{2}(x^2-1)t^2 - xt - 1\right)e^{-\frac{t^2}{2}}.$$

*Além disso, tem-se $E(h_*(x,X)) = 0$.*

*Demonstração.* Atendendo à definição de $h(x;t)$, concluímos então pelo Teorema 1.4 que

$$\begin{aligned} nD_{n,\beta} &= \int Z_n''(t)^2\varphi_\beta(t)dt + o_p(1) = \int \left\{\frac{1}{\sqrt{n}}\sum_{j=1}^{n} h(X_j;t)\right\}^2 \varphi_\beta(t)dt + o_p(1) \\ &= \frac{1}{n}\sum_{i,j=1}^{n}\int h(X_i;t)h(X_j;t)\varphi_\beta(t)dt + o_p(1) \\ &= \frac{1}{n}\sum_{i,j=1}^{n} h_*(X_i,X_j) + o_p(1). \end{aligned}$$

Finalmente, para concluirmos que $nD_{n,\beta}$ tem a estrutura de uma V-estatística degenerada, isto é, $E(h_*(x,X)) = 0$, basta provar que $E(h(X;t)) = 0$.

Aplicando a linearidade da esperança matemática, vem que

$$\begin{aligned} E(h(X;t)) = E(\cos(tX)) + E(\sin(tX)) + \frac{1}{2}(E(X^2)-1)t^2e^{-\frac{t^2}{2}} \\ - E(X)te^{-\frac{t^2}{2}} - e^{-\frac{t^2}{2}} \end{aligned}$$

Basta apenas ter em conta que

$$E(\cos(tX)) = e^{-\frac{t^2}{2}}$$

e

$$E(\sin(tX)) = 0.$$

□

Sabendo que a estatística de teste $nD_{n,\beta}$ é aproximada por uma V-estatística degenerada, é agora possível conhecer a sua distribuição limite.

**Teorema 1.6.** *Se $X_1, \ldots, X_n$ são variáveis aleatórias normais i.i.d. temos*

$$nD_{n,\beta} \xrightarrow{d} \sum_{k\geq 1} \lambda_k Z_k^2,$$

*onde $Z_j$ são variáveis aleatórias i.i.d. que seguem uma lei normal standard e $\lambda_k, k \geq 1$, são os valores próprios não nulos correspondentes ao operador integral $A : L^2(\mathbb{R}, \phi) \to L^2(\mathbb{R}, \phi)$ definido por*

$$Ag(x) = \int h_*(x, y)g(y)\phi(y)dy, \ x \in \mathbb{R},$$

*com $g \in L^2(\mathbb{R}, \phi)$, $\phi$ a densidade normal standard e $L^2(\mathbb{R}, \phi)$ o espaço das funções $g$, reais de variável real, tais que $\int g^2(y)\phi(y)dy < \infty$.*

*Demonstração.* Sendo $D_{n,\beta}$ invariante para transformações afins dos dados vamos admitir que $X_1, \ldots, X_n$ são variáveis normais centradas e reduzidas.

Comecemos por provar que

$$E|h_*(X_1, X_1)| < \infty$$

e

$$E(h_*^2(X_1, X_2)) < \infty.$$

Com efeito,

$$E|h_*(X_1, X_1)| \leq \int E|h^2(X; t)|\varphi_\beta(t)dt$$

e

$$E(h_*^2(X_1, X_2)) \leq \int [E(h^2(X; t))]^2\varphi_\beta(t)dt,$$

onde

$$E(h^2(X; t)) = 1 - e^{-t^2}(t^4 + t^2 + 1), \ t \in \mathbb{R},$$

o que prova o pretendido.

Atendendo agora ao Teorema 1.5 e ao Teorema 2.1 de Gregory (1977) deduzimos que

$$\begin{aligned} nD_{n,\beta} &= \frac{1}{n}\sum_{i=1}^{n} h_*(X_i, X_i) + \frac{1}{n}\sum_{i\neq j} h_*(X_i, X_j) \\ &\xrightarrow{d} E(h_*(X_1, X_1)) + \sum_{k=1}^{\infty} \lambda_k(Z_k^2 - 1). \end{aligned}$$

Para concluir a demonstração, basta agora ter em conta que

$$E(h_*(X_1, X_1)) = \sum_{k=1}^{\infty} \lambda_k$$

(cf. Serfling, 1980, p. 227). □

## 1.3. A convergência do teste baseado em $D_{n,\beta}$

Seguidamente, vamos enunciar um resultado que nos permitirá estabelecer a convergência do teste baseado na estatística $D_{n,\beta}$.

**Teorema 1.7.** *Sejam $X_1, \ldots, X_n$ cópias independentes de uma variável real $X$ com média $\mu$, variância $\sigma^2$ e função característica $\Phi$. Então*

$$D_{n,\beta} = \int \left| \Psi_n(t) - e^{-\frac{t^2}{2}} \right|^2 \varphi_\beta(t)dt \xrightarrow{p} \int \left| \Phi_0(t) - e^{-\frac{t^2}{2}} \right|^2 \varphi_\beta(t)dt,$$

*onde $\Phi_0(t) = e^{it\frac{\mu}{\sigma}}\Phi\left(\frac{t}{\sigma}\right)$ é a função característica da variável $\frac{X-\mu}{\sigma}$.*

*Demonstração.* Sendo $D_{n,\beta}$ invariante para transformações afins dos dados, basta provar o resultado para uma variável aleatória $X$ de média $\mu = 0$ e variância $\sigma^2 = 1$.

Atendendo a que

$$\Phi_n(t) = \frac{1}{n}\sum_{j=1}^{n} e^{itX_j} \xrightarrow{p} \Phi_0(t)$$

para todo o $t \in \mathbb{R}$, concluímos pelo Teorema A.1 que

$$\int \left| \Phi_n(t) - e^{-\frac{t^2}{2}} \right|^2 \varphi_\beta(t)dt \xrightarrow{p} \int \left| \Phi_0(t) - e^{-\frac{t^2}{2}} \right|^2 \varphi_\beta(t)dt.$$

Basta então demonstrar que

$$\int \left| \Psi_n(t) - e^{-\frac{t^2}{2}} \right|^2 \varphi_\beta(t)dt = \int \left| \Phi_n(t) - e^{-\frac{t^2}{2}} \right|^2 \varphi_\beta(t)dt + o_p(1). \tag{1.9}$$

Atendendo à Proposição 1.1, temos

$$\int \left| \Psi_n(t) - e^{-\frac{t^2}{2}} \right|^2 \varphi_\beta(t)dt = \frac{1}{n}\int Z_n(t)^2\varphi_\beta(t)dt$$

e

$$\int \left| \Phi_n(t) - e^{-\frac{t^2}{2}} \right|^2 \varphi_\beta(t)dt = \frac{1}{n}\int Z_n^*(t)^2\varphi_\beta(t)dt,$$

onde

$$Z_n(t) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left\{\cos(tY_j) + \sin(tY_j) - e^{-\frac{t^2}{2}}\right\}$$

e

$$Z_n^*(t) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left\{\cos(tX_j) + \sin(tX_j) - e^{-\frac{t^2}{2}}\right\}.$$

Para estabelecer (1.9), vamos de seguida mostrar que

$$\frac{Z_n}{\sqrt{n}} = \frac{Z_n^*}{\sqrt{n}} + o_{2,p}(1).$$

Usando a fórmula de Taylor com resto de primeira ordem, isto é,

$$\cos(tY_j) = \cos(tX_j) - t\sin(t\tilde{X}_j)(Y_j - X_j)$$

e

$$\sin(tY_j) = \sin(tX_j) + t\cos(t\tilde{\tilde{X}}_j)(Y_j - X_j)$$

onde $\tilde{X}_j$ e $\tilde{\tilde{X}}_j$ pertencem ao intervalo de extremidades $X_j$ e $Y_j$, temos

$$\begin{aligned}\left|\frac{Z_n}{\sqrt{n}} - \frac{Z_n^*}{\sqrt{n}}\right| &\leq \frac{1}{n}\sum_{j=1}^{n}\left\{|t||\sin(t\tilde{X}_j)| + |t||\cos(t\tilde{\tilde{X}}_j)|\right\}|Y_j - X_j| \\ &\leq 2|t|\left\{(s_n^{-1} - 1)\frac{1}{n}\sum_{j=1}^{n}|X_j| + s_n^{-1}\bar{X}_n\right\}\end{aligned}$$

o que permite concluir o pretendido. □

Estamos agora em condições de enunciar a convergência do teste baseado em $D_{n,\beta}$ para toda a distribuição não normal.

**Teorema 1.8.** *Se $X_1, \ldots, X_n$ são variáveis aleatórias i.i.d. que verificam $E(X_1^2) < \infty$, então o teste de região crítica*

$$\{nD_{n,\beta} > c_{n,\alpha}\},$$

*onde $c_{n,\alpha}$ é o quantil de ordem $1-\alpha$ da distribuição de $nD_{n,\beta}$ sob $H_0$, é convergente.*

*Demonstração.* Sejam $X_1, \ldots, X_n$ variáveis aleatórias cuja distribuição não é normal, isto é,

$$\Phi_0(t) \neq e^{-\frac{t^2}{2}}, \text{ para algum } t \in \mathbb{R}.$$

Pelo Teorema 1.7, sabemos que

$$D_{n,\beta} \xrightarrow{d} \int \left|\Phi_0(t) - e^{-\frac{t^2}{2}}\right|^2 \varphi_\beta(t)dt.$$

Uma vez que

$$\int \left|\Phi_0(t) - e^{-\frac{t^2}{2}}\right|^2 \varphi_\beta(t)dt > 0$$

concluímos que

$$nD_{n,\beta} \xrightarrow{p} +\infty,$$

e assim

$$P_f\big(nD_{n,\beta} > c_{n,\alpha}\big) \to 1, \text{ quando } n \to \infty.$$

□

# Capítulo 2

# Testes de normalidade baseados em $D_{n,\beta}$, com $\beta = 0$ e $\beta = \infty$

Contrariamente ao Capítulo 1, é mais conveniente, a partir deste momento, considerarmos a forma simplificada da estatística $D_{n,\beta}$, ou seja,

$$D_{n,\beta} = \frac{1}{n^2}\sum_{j,k=1}^{n}\exp\left(-\frac{\beta^2}{2}(Y_j - Y_k)^2\right) - 2(1+\beta^2)^{-1/2}\frac{1}{n}\sum_{j=1}^{n}\exp\left(-\frac{\beta^2 Y_j^2}{2(1+\beta^2)}\right) + (1+2\beta^2)^{-1/2}. \tag{2.1}$$

Os testes para a normalidade multivariada cujas estatísticas de teste são obtidas tomando $\beta \to 0$ e $\beta \to \infty$ em (2.1) foram propostos e estudados por Henze (1997). No que se segue, pretendemos identificar cada uma das estatísticas limite, $D_{n,0}$ e $D_{n,\infty}$, e descrever as suas propriedades. Assim, dividiremos este capítulo em duas secções, em que estudaremos a distribuição assintótica de cada uma das estatísticas $D_{n,0}$ e $D_{n,\infty}$ e estabeleceremos a convergência dos testes de normalidade nelas baseados.

## 2.1. O caso $\beta \to 0$

No teorema seguinte identificamos a estatística que se obtém de $D_{n,\beta}$ tomando $\beta \to 0$.

**Teorema 2.1.** *Se $X_1, \ldots, X_n$ são variáveis aleatórias quaisquer, então*

$$\lim_{\beta\to 0}\beta^{-6}D_{n,\beta} = \frac{5}{12}b_1,$$

*onde*

$$b_1 = \frac{1}{n^2}\sum_{j,k=1}^{n}(Y_jY_k)^3.$$

*Demonstração.* Atendendo a que

$$\frac{1}{n}\sum_{j=1}^{n}Y_j = \frac{1}{n}\sum_{j=1}^{n}\frac{X_j - \bar{X}_n}{s_n} = \frac{1}{s_n}\left(\frac{1}{n}\sum_{j=1}^{n}X_j - \bar{X}_n\right) = 0$$

e

$$\frac{1}{n}\sum_{j=1}^{n}Y_j^2 = \frac{1}{n}\sum_{j=1}^{n}\frac{(X_j - \bar{X}_n)^2}{s_n^2} = 1,$$

podemos concluir que

$$\frac{1}{n^2}\sum_{j,k=1}^{n}(Y_j - Y_k)^2 = 2,$$

$$\frac{1}{n^2}\sum_{j,k=1}^{n}(Y_j - Y_k)^4 = \frac{2}{n}\sum_{j=1}^{n}Y_j^4 + 6$$

e

$$\frac{1}{n^2}\sum_{j,k=1}^{n}(Y_j - Y_k)^6 = -20b_1 + \frac{30}{n}\sum_{j=1}^{n}Y_j^4 + \frac{2}{n}\sum_{j=1}^{n}Y_j^6.$$

Usando a fórmula de Taylor com resto de ordem 4, podemos obter os desenvolvimentos seguintes para os dois primeiros termos da estatística $D_{n,\beta}$ dada em (2.1):

$$\begin{aligned}&\frac{1}{n^2}\sum_{j,k=1}^{n}\exp\left(-\frac{\beta^2}{2}(Y_j - Y_k)^2\right)\\ &\quad= \frac{1}{n^2}\sum_{j,k=1}^{n}\left\{1 - \frac{\beta^2}{2}(Y_j - Y_k)^2 + \frac{\beta^4}{8}(Y_j - Y_k)^4 - \frac{\beta^6}{48}(Y_j - Y_k)^6\right\} + O(\beta^8) \quad (2.2)\end{aligned}$$

e

$$\begin{aligned}&\frac{1}{n}\sum_{j=1}^{n}\exp\left(-\frac{\beta^2 Y_j^2}{2(1+\beta^2)}\right)\\ &\quad= \frac{1}{n}\sum_{j=1}^{n}\left\{1 - \frac{\beta^2 Y_j^2}{2(1+\beta^2)} + \frac{\beta^4 Y_j^4}{8(1+\beta^2)^2} - \frac{\beta^6 Y_j^6}{48(1+\beta^2)^3}\right\} + O(\beta^8) \quad (2.3)\end{aligned}$$

Temos ainda

$$2(1+\beta^2)^{-1/2} = 2 - \beta^2 + \frac{3}{4}\beta^4 - \frac{5}{8}\beta^6 + O(\beta^8) \quad (2.4)$$

e

$$(1+2\beta^2)^{-1/2} = 1 - \beta^2 + \frac{3}{2}\beta^4 - \frac{5}{2}\beta^6 + O(\beta^8) \quad (2.5)$$

Desenvolvendo a expressão (2.2) e fazendo os cálculos necessários usando ainda (2.3), (2.4) e (2.5), obtém-se

$$D_{n,\beta} = \frac{5}{12}\beta^6 b_1 + O(\beta^8).$$

Portanto,

$$\lim_{\beta\to 0}\beta^{-6}D_{n,\beta} = \frac{5}{12}b_1.$$

□

A estatística de teste $D_{n,0}$ não é definida diretamente a partir de (2.1) mas sim a partir do limite apresentado, a menos das constantes. Nesse sentido, a estatística

que se obtém quando $\beta \to 0$ é a estatística utilizada por Mardia (1970) para estimar o coeficiente de assimetria da distribuição, isto é,

$$D_{n,0} = b_1.$$

Para determinarmos a distribuição assintótica da estatística de teste $D_{n,0}$, comecemos por escrevê-la na seguinte forma:

$$D_{n,0} = \left( \frac{\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^3}{\left(\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2\right)^{\frac{3}{2}}} \right)^2$$

De seguida, vamos estabelecer um resultado que nos será útil para determinarmos a distribuição limite de $D_{n,0}$ quando $n \to \infty$.

**Teorema 2.2.** *Se $X_1, \ldots, X_n$ são variáveis aleatórias i.i.d. de média $\mu$ e variância $\sigma^2$, que verificam $E(X_1^6) < \infty$, temos*

$$\sqrt{n}\left\{\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^3 - \mu_3\right\} \xrightarrow{d} \mathcal{N}(0, \nu^2),$$

*onde*

$$\nu^2 = E\left[X_1^3 - \mu_3 - 3X_1\right]^2$$

*e $\mu_3 = E[X_1 - \mu]^3$ é o momento centrado de ordem 3.*

*Demonstração.* Atendendo a que $D_{n,0}$ é invariante para transformações afins dos dados, basta considerar que $X_1, \ldots, X_n$ são variáveis aleatórias de média $\mu = 0$ e variância $\sigma^2 = 1$. Comecemos por notar que

$$\begin{aligned}
\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^3 &= \frac{1}{n}\sum_{j=1}^{n}\left\{X_j^3 + 3X_j^2\bar{X}_n + 3\bar{X}_n^2 X_j + \bar{X}_n^3\right\} \\
&= \frac{1}{n}\sum_{j=1}^{n}X_j^3 + 3\bar{X}_n + 3\bar{X}_n\left(\frac{1}{n}\sum_{j=1}^{n}X_j^2 - 1\right) + o_p(1) \\
&= E(X^3) + \frac{1}{n}\sum_{j=1}^{n}\left\{X_j^3 - E(X^3) - 3X_j\right\} + o_p(1) \\
&= \mu_3 + \frac{1}{n}\sum_{j=1}^{n}Z_j + o_p(1),
\end{aligned}$$

onde $Z_j = X_j^3 - \mu_3 - 3X_j$.

Para concluir basta aplicar o Teorema do Limite Central à soma $\frac{1}{n}\sum_{j=1}^{n}Z_j$. □

Neste momento, é mais simples estabelecermos a distribuição assintótica da estatística de teste $D_{n,0}$.

**Corolário 2.3.** *Nas condições do Teorema 2.2, assumindo ainda que as variáveis $X_1, \ldots, X_n$ seguem uma distribuição normal, temos*

$$nD_{n,0} \xrightarrow{d} 6\chi^2_1.$$

*Demonstração.* Sendo $D_{n,0}$ invariante para transformações afins dos dados, vamos assumir que as variáveis aleatórias $X_1, \ldots, X_n$ são centradas e reduzidas. Atendendo a que $\mu_3 = 0$, pelo Teorema 2.2, sabemos que

$$\sqrt{n}\left\{\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^3\right\} \xrightarrow{d} \mathcal{N}(0, \nu^2),$$

onde

$$\nu^2 = E\left[X_1^3 - \mu_3 - 3X_1\right]^2 = \mu_6 + \mu_3^2 + 9 - 6\mu_4 = 6.$$

Pela lei dos grandes números, temos ainda

$$\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2 \xrightarrow{p} 1.$$

Assim,

$$\sqrt{nD_{n,0}} = \frac{\sqrt{n}\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^3}{\left(\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2\right)^{\frac{3}{2}}} \xrightarrow{d} \mathcal{N}(0, 6),$$

de onde concluímos que

$$nD_{n,0} \xrightarrow{d} 6\chi^2_1.$$

□

Finalmente, podemos enunciar a convergência do teste baseado na estatística $D_{n,0}$.

**Teorema 2.4.** *Se $X_1, \ldots, X_n$ são variáveis aleatórias i.i.d. de média $\mu$ e variância $\sigma^2$, que verificam $E(X_1^6) < \infty$, então o teste de região crítica*

$$\{nD_{n,0} > c_{n,\alpha}\},$$

*onde $c_{n,\alpha}$ é o quantil de ordem $1-\alpha$ da distribuição de $nD_{n,0}$ sob $H_0$, é convergente se e só se $\mu_3 = E[X_1 - \mu]^3 \neq 0$.*

*Demonstração.* Comecemos por supor que $\mu_3 \neq 0$. Temos então

$$\sqrt{n}\left\{\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^3\right\} = \sqrt{n}\left\{\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^3 - \mu_3\right\} + \sqrt{n}\mu_3$$

e, portanto,

$$n\left(\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^3\right)^2 \xrightarrow{p} +\infty.$$

Assim,

$$nD_{n,0} \xrightarrow{p} +\infty,$$

o que permite concluir que

$$P_f\big(nD_{n,0} > c_{n,\alpha}\big) \to 1,\ n \to \infty.$$

Suponhamos agora que

$$P_f\big(nD_{n,0} > c_{n,\alpha}\big) \to 1,\ n \to \infty \tag{2.6}$$

e admitamos por absurdo que $\mu_3 = 0$. Neste caso, atendendo ao Teorema 2.2 teríamos

$$\frac{\sqrt{n}\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^3}{\left(\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2\right)^{\frac{3}{2}}} \xrightarrow{d} \mathcal{N}\left(0, \frac{\nu^2}{\sigma^6}\right)$$

e portanto

$$nD_{n,0} \xrightarrow{d} \frac{\nu^2}{\sigma^6}\chi^2_1$$

o que contradiz (2.6). □

## 2.2. O caso $\beta \to \infty$

Escrevendo a expressão da estatística de teste $D_{n,\beta}$ dada em (2.1) na forma

$$D_{n,\beta} = (1+2\beta^2)^{-1/2} - 2(1+\beta^2)^{-1/2}\frac{1}{n}\sum_{j=1}^{n}\exp\left(-\frac{\beta^2 Y_j^2}{2(1+\beta^2)}\right) + \frac{1}{n}, \tag{2.7}$$

podemos identificar a estatística que se obtém de $D_{n,\beta}$ tomando o limite quando $\beta \to \infty$.

**Teorema 2.5.** *Se $X_1, \ldots, X_n$ são variáveis aleatórias quaisquer, então*

$$\lim_{\beta\to\infty} \beta\left(D_{n,\beta} - \frac{1}{n}\right) = \frac{1}{\sqrt{2}} - \frac{2}{n}\sum_{j=1}^{n}\exp\left(-Y_j^2/2\right).$$

*Demonstração.* Usando a fórmula de Taylor concluímos que

$$(1+2\beta^2)^{-1/2} = \left(\beta\sqrt{2}\right)^{-1} + O(\beta^{-3})$$

e

$$(1+\beta^2)^{-1/2} = \beta^{-1} + O(\beta^{-3}),$$

quando $\beta \to \infty$.

Substituindo estas expressões em (2.7), obtemos

$$D_{n,\beta} = \frac{1}{\beta\sqrt{2}} - \frac{2}{\beta n}\sum_{j=1}^{n} \exp\left(-\frac{\beta^2 Y_j^2}{2(1+\beta^2)}\right) + \frac{1}{n} + O(\beta^{-3}).$$

Por último, basta subtrair por $\frac{1}{n}$, multiplicar por $\beta$ e aplicar o limite quando $\beta \to \infty$ (note-se que, ao aplicar o limite, $\frac{\beta^2}{1+\beta^2}$ tende para 1), ou seja

$$\begin{aligned}\lim_{\beta\to\infty} \beta\left(D_{n,\beta} - \frac{1}{n}\right) &= \frac{1}{\sqrt{2}} - \frac{2}{n}\sum_{j=1}^{n} \exp\left(-\frac{\beta^2 Y_j^2}{2(1+\beta^2)}\right) + \frac{1}{n} + O(\beta^{-2}) \\ &= \frac{1}{\sqrt{2}} - \frac{2}{n}\sum_{j=1}^{n} \exp\left(-Y_j^2/2\right).\end{aligned}$$

□

Tal como $D_{n,0}$, a estatística de teste $D_{n,\infty}$ é definida a partir do limite estabelecido no teorema anterior, isto é,

$$D_{n,\infty} = \frac{1}{n}\sum_{j=1}^{n} \exp\left(-Y_j^2/2\right).$$

A estatística de teste obtida quando $\beta \to \infty$ é de certa forma semelhante à estatística utilizada por Mardia (1970) para estimar o coeficiente de curtose

$$b_2 = \frac{1}{n}\sum_{j=1}^{n} Y_j^4,$$

uma vez que só usa os valores de $Y_j^2$, para $j = 1, \ldots, n$.

De seguida, determinemos a distribuição limite de $D_{n,\infty}$, quando $n \to \infty$.

**Teorema 2.6.** *Se $X_1, \ldots, X_n$ são variáveis aleatórias i.i.d. que verificam $E(X_1^2) < \infty$, temos*

$$\sqrt{n}\left(D_{n,\infty} - E\left[\exp\left(-X_1^2/2\right)\right]\right) \xrightarrow{d} \mathcal{N}(0, \nu^2),$$

*onde*

$$\nu^2 = Var\left(\exp\left(-X_1^2/2\right) + \frac{1}{2}BX_1^2 + bX_1\right), \tag{2.8}$$

$$B = E\left[X_1^2 \exp\left(-X_1^2/2\right)\right],$$

*e*

$$b = E\left[X_1 \exp\left(-X_1^2/2\right)\right].$$

*Demonstração.* Uma vez que $D_{n,\infty}$ é invariante para transformações afins dos dados, vamos assumir que $X_1, \ldots, X_n$ são variáveis aleatórias centradas e reduzidas. Seja $\xi = E\left[\exp\left(-\frac{X_1^2}{2}\right)\right]$. Usando um desenvolvimento de Taylor da função $f(u) = \exp\left(-\frac{u}{2}\right)$, podemos escrever

$$\begin{aligned}\sqrt{n}(D_{n,\infty} - \xi) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left(\exp\left(-X_j^2/2\right) - \xi\right) - \frac{1}{2\sqrt{n}} \sum_{j=1}^{n} \exp\left(-X_j^2/2\right) \left\{Y_j^2 - X_j^2\right\} \\ + \frac{1}{2} \frac{1}{\sqrt{n}} \sum_{j=1}^{n} f''(\tilde{X}_j) \left(Y_j^2 - X_j^2\right)^2, \qquad (2.9)\end{aligned}$$

onde $\tilde{X}_j$ pertence ao intervalo de extremidades $Y_j$ e $X_j$ e $\left|f''(\tilde{X}_j)\right| \leq 1/4$.

Além disso,

$$\begin{aligned}\frac{1}{n} \sum_{j=1}^{n} \left(Y_j^2 - X_j^2\right)^2 &= (s_n^{-2} - 1)^2 \frac{1}{n} \sum_{j=1}^{n} X_j^4 - 4 s_n^{-2} (s_n^{-2} - 1) \bar{X}_n \frac{1}{n} \sum_{j=1}^{n} X_j^3 \\ &\quad + 2(s_n^{-2} - 1)^2 s_n^{-2} \bar{X}_n^2 \frac{1}{n} \sum_{j=1}^{n} X_j^2 + 4 s_n^{-4} \bar{X}_n^2 \frac{1}{n} \sum_{j=1}^{n} X_j^2 - 3 s_n^{-4} \bar{X}_n^4 \\ &= O_p(n^{-1}).\end{aligned}$$

Concluímos assim que o termo residual da expressão (2.9) converge em probabilidade para zero, ou seja, é um $o_p(1)$. Portanto, temos

$$\begin{aligned}\sqrt{n}(D_{n,\infty} - \xi) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left(\exp\left(-X_j^2/2\right) - \xi\right) \\ - \frac{1}{2\sqrt{n}} \sum_{j=1}^{n} \exp\left(-X_j^2/2\right) \left\{Y_j^2 - X_j^2\right\} + o_p(1).\end{aligned}$$

Atendendo a que

$$\begin{aligned}&-\frac{1}{2\sqrt{n}} \sum_{j=1}^{n} \exp\left(-X_j^2/2\right) \left\{Y_j^2 - X_j^2\right\} \\ &\quad = -\frac{1}{2\sqrt{n}} \sum_{j=1}^{n} \exp\left(-X_j^2/2\right) \left\{\frac{X_j^2 - 2X_j\bar{X}_n + \bar{X}_n^2}{s_n^2}\right\} + \frac{1}{2} \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \exp\left(-X_j^2/2\right) X_j^2,\end{aligned}$$

substituindo

$$s_n^{-2} = 1 - n^{-1/2} A_n + O_p(n^{-1}), \text{ onde } A_n = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (X_j^2 - 1),$$

vamos obter

$$\begin{aligned}
&-\frac{1}{2\sqrt{n}}\sum_{j=1}^{n}\exp\left(-X_j^2/2\right)\left\{Y_j^2 - X_j^2\right\} \\
&\quad= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\exp\left(-X_j^2/2\right)X_j\bar{X}_n - \frac{1}{2\sqrt{n}}\sum_{j=1}^{n}\exp\left(-X_j^2/2\right)\bar{X}_n^2 \\
&\qquad+ \frac{A_n}{2n}\sum_{j=1}^{n}\exp\left(-X_j^2/2\right)X_j^2 - \frac{A_n}{n}\sum_{j=1}^{n}\exp\left(-X_j^2/2\right)X_j\bar{X}_n \\
&\qquad+ \frac{A_n}{2n}\sum_{j=1}^{n}\exp\left(-X_j^2/2\right)\bar{X}_n^2 + o_p(1) \\
&\quad= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\exp\left(-X_j^2/2\right)X_j\bar{X}_n + \frac{A_n}{2n}\sum_{j=1}^{n}\exp\left(-X_j^2/2\right)X_j^2 + o_p(1). \qquad (2.10)
\end{aligned}$$

Vamos verificar que

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\exp\left(-X_j^2/2\right)X_j\bar{X}_n = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}bX_j + o_p(1)$$

e que

$$\frac{A_n}{2n}\sum_{j=1}^{n}\exp\left(-X_j^2/2\right)X_j^2 = \frac{B}{2\sqrt{n}}\sum_{j=1}^{n}(X_j^2 - 1) + o_p(1).$$

Demonstremos apenas a primeira igualdade, sendo a outra análoga. Temos então

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\exp\left(-X_j^2/2\right)X_j\bar{X}_n \\
&\quad= \frac{\sqrt{n}\bar{X}_n}{n}\sum_{j=1}^{n}\left\{\exp\left(-X_j^2/2\right)X_j - E\left(X\exp\left(-X_j^2/2\right)\right)\right\} \\
&\qquad+ \sqrt{n}\bar{X}_n E\left(X\exp\left(-X_j^2/2\right)\right) \\
&\quad= \sqrt{n}\left(\frac{1}{n}\sum_{j=1}^{n}X_j\right)b + o_p(1) \\
&\quad= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}bX_j + o_p(1)
\end{aligned}$$

Fazendo as substituições necessárias na expressão (2.10), obtemos

$$\sqrt{n}(D_{n,\infty} - \xi) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}V_j + o_p(1),$$

onde

$$V_j = \exp\left(-X_j^2/2\right) - \xi + \frac{B}{2}(X_j^2 - 1) + bX_j.$$

Pelo Teorema do Limite Central, temos

$$\sqrt{n}(D_{n,\infty} - \xi) \xrightarrow{d} \mathcal{N}(0, \nu^2),$$

onde

$$\nu^2 = \text{ Var}\left(\exp\left(-X_1^2/2\right) + \frac{B}{2}X_1^2 + bX_1\right).$$

□

**Corolário 2.7.** *Nas condições do Teorema 2.6, assumindo ainda que as variáveis $X_1, \ldots, X_n$ seguem uma distribuição normal, temos*

$$\sqrt{n}\left(D_{n,\infty} - 2^{-1/2}\right) \xrightarrow{d} \mathcal{N}(0, 3^{-1/2} - 2^{-1} - 2^{-4}).$$

*Demonstração.* Uma vez que $X_1, \ldots, X_n$ são variáveis aleatórias normais, $E\left[\exp\left(-X_1^2/2\right)\right] = 2^{-1/2}$, $B = 2^{-3/2}$ e $b = 0$.

Pelo Teorema 2.6, sabemos que

$$\sqrt{n}\left(D_{n,\infty} - 2^{-1/2}\right) \xrightarrow{d} \mathcal{N}(0, \nu^2),$$

onde

$$\begin{aligned}
\nu^2 &= \text{Var}\left(\exp\left(-X_1^2/2\right) + 2^{-5/2}X_1^2\right) \\
&= E\left(\exp\left(-X_1^2/2\right) + 2^{-5/2}X_1^2\right)^2 - \left[E\left(\exp\left(-X_1^2/2\right) + 2^{-5/2}X^2\right)\right]^2 \\
&= E\left(\exp(-X_1^2) + 2^{-3/2}X_1^2\exp\left(-X_1^2/2\right) + 2^{-5}X_1^4\right) - \left[2^{-1/2} + 2^{-5/2}\right]^2 \\
&= 3^{-1/2} + 2^{-3} + 3 \times 2^{-5} - 2^{-1} - 2^{-2} - 2^{-5} \\
&= 3^{-1/2} - 2^{-1} - 2^{-4}.
\end{aligned}$$

□

Por último, vamos estabelecer a convergência do teste baseado na estatística $D_{n,\infty}$.

**Teorema 2.8.** *Se $X_1, \ldots, X_n$ são variáveis aleatórias i.i.d. de média $\mu$ e variância $\sigma^2$, que verificam $E(X_1^2) < \infty$, então o teste de região crítica*

$$\left\{\sqrt{n}\left|D_{n,\infty} - 2^{-1/2}\right| > c_{n,\alpha}\right\},$$

*onde $c_{n,\alpha}$ é o quantil de ordem $1 - \frac{\alpha}{2}$ da distribuição de $\sqrt{n}\left(D_{n,\infty} - 2^{-1/2}\right)$ sob $H_0$, é convergente se e só se*

$$\mu(f) = E\left[\exp\left(-\frac{(X_1 - \mu)^2}{2\sigma}\right)\right] \neq 2^{-1/2}.$$

*Demonstração.* Sabemos do Teorema 2.6 que

$$\sqrt{n}\,(D_{n,\infty} - \mu(f)) \xrightarrow{d} \mathcal{N}(0, \nu^2), \tag{2.11}$$

onde $\nu^2$ é definido por (2.8).

Comecemos por supor que $f$ é tal que $\mu(f) \neq 2^{-1/2}$. Temos então

$$\sqrt{n}(D_{n,\infty} - 2^{-1/2}) = \sqrt{n}(D_{n,\infty} - \mu(f)) + \sqrt{n}(\mu(f) - 2^{-1/2}).$$

Portanto, da convergência (2.11) deduzimos que

$$\sqrt{n}\left|D_{n,\infty} - 2^{-1/2}\right| \xrightarrow{p} +\infty,$$

o que permite concluir que

$$P_f\left(\sqrt{n}\left|D_{n,\infty} - 2^{-1/2}\right| > c_{n,\alpha}\right) \to 1,\ n \to +\infty.$$

Reciprocamente, suponhamos que

$$P_f\left(\sqrt{n}\left|D_{n,\infty} - 2^{-1/2}\right| > c_{n,\alpha}\right) \to 1,\ n \to +\infty$$

e admitamos, por absurdo, que $\mu(f) = 2^{-1/2}$. Neste caso, teríamos

$$P_f\left(\sqrt{n}\,|D_{n,\infty} - \mu(f)| > c_{n,\alpha}\right) \to 1,\ n \to +\infty,$$

o que é falso atendendo à convergência em distribuição (2.11). □

# Capítulo 3

# Um Teste Múltiplo de Normalidade

## 3.1. Introdução

Comecemos por recordar a forma da estatística de teste de Epps e Pulley (1983):

$$D_{n,\beta} = \int_{\mathbb{R}} \left| \Psi_n(t) - e^{-\frac{t^2}{2}} \right|^2 \varphi_\beta(t) dt,$$

onde $0 < \beta < \infty$ e $\varphi_\beta(t)$ é a densidade da lei $\mathcal{N}(0, \beta^2)$. De acordo com o Capítulo 1 sabemos que, $nD_{n,\beta}$ possui como distribuição assintótica uma soma ponderada de qui-quadrados independentes. Vimos ainda que, contrariamente à maioria dos testes de normalidade considerados na literatura, o teste associado a esta estatística de teste é convergente para toda a distribuição alternativa fixa com momentos de segunda ordem finitos. Apesar da convergência do teste ser independente do valor tomado para o parâmetro $\beta > 0$, a sua potência depende fortemente da escolha deste parâmetro (ver Henze e Wagner, 1997; Henze e Zirkler, 1990; Tenreiro, 2009). Para valores pequenos de $\beta$, o teste baseado em $D_{n,\beta}$ é sensível a alternativas com caudas pesadas uma vez que a função $\varphi_\beta$ coloca maior parte do seu peso numa vizinhança da origem de $\mathbb{R}$ e o comportamento da cauda de uma distribuição de probabilidade é reflectido através do comportamento da sua função característica na origem. Por outro lado, para valores grandes de $\beta$, o teste é mais sensível para alternativas com caudas leves. Como casos limite das situações anteriores temos os casos extremos $\beta = 0$ e $\beta = \infty$.

Tendo como objetivo obter um teste que possua uma potência razoável para um conjunto vasto de distribuições alternativas, vamos neste capítulo considerar um teste que resulta da combinação de testes baseados nas estatísticas $D_{n,\beta}$ para alguns valores extremos e não extremos do parâmetro $\beta$. Esperamos que o teste combinado usufrua das boas propriedades dos testes envolvidos na combinação.

Na Secção 3.2, vamos usar um método proposto por Fromont e Laurent (2006), que pode ser visto como uma generalização do método de Bonferroni, para definir

um tal teste. Relativamente a este teste, damos a conhecer as condições para as quais tem nível de significância menor ou igual a $\alpha$ e é convergente, resultados esses apresentados em Tenreiro (2011). Na Secção 3.3, vamos explicitar quais os testes que queremos usar na combinação e verificar se são satisfeitas todas as condições referidas na secção anterior, para que possamos aplicar o procedimento de teste múltiplo definido nessa secção.

## 3.2. Um teste múltiplo de tipo Bonferroni

Consideremos

$$T_{n,h} = T_{n,h}(X_1, \ldots, X_n),\ h \in H,$$

uma família finita de estatísticas para testar a hipótese de normalidade $H_0 : f \in \mathcal{N}$, que assumimos serem invariantes para transformações afins dos dados, isto é,

$$T_{n,h}(aX_1 + b, \ldots, aX_n + b) = T_{n,h}(X_1, \ldots, X_n),$$

para todo o $b \in \mathbb{R}$ e $a \in \mathbb{R}$ não nulo, e sejam

$$R_h = \{T_{n,h} > c_{n,h}(\alpha)\},$$

as respetivas regiões críticas, onde $c_{n,h}(\alpha)$ é o quantil de ordem $1-\alpha$ da distribuição de $T_{n,h}$.

O teste múltiplo que vamos considerar corresponde a rejeitar a hipótese nula se pelo menos um dos testes envolvidos rejeitar essa hipótese e, por isso, para que o novo teste tenha nível de significância menor do que $\alpha$, é necessário corrigir os quantis associados a cada um dos testes parciais. Assim, a região crítica do teste múltiplo será dada por

$$R = \bigcup_{h \in H} \{T_{n,h} > c_{n,h}(\alpha^*)\},\ \text{onde } \alpha^* \leq \alpha.$$

Se considerarmos

$$\alpha^* = \frac{\alpha}{|H|},$$

onde $|H|$ denota a cardinalidade do conjunto $H$, o procedimento de teste múltiplo obtido é o método clássico de Bonferroni. No entanto, pretendemos que o nível de significância do teste múltiplo seja tão próximo quanto possível de $\alpha$, aumentando o conjunto de valores de cada uma das estatísticas de teste que nos levam a rejeitar a hipótese nula. Assim, é definido um teste múltiplo de normalidade invariante para transformações afins dos dados com nível de significância igual a $\alpha$, sugerido por

Fromont e Laurent (2006), que pode ser interpretado como uma melhoria do método de Bonferroni clássico. De seguida, descrevemos a correção dos quantis associados a cada um dos testes parciais que darão origem ao novo teste múltiplo.

Considerando $u \in ]0,1[$ e $h \in H$, denotemos por $c_{n,h}(u)$ o quantil de ordem $1-u$ da estatística de teste $T_{n,h}$ sob a hipótese $H_0$ e consideremos a estatística corrigida

$$T_n(u) = \max_{h \in H} \left( T_{n,h} - c_{n,h}(u) \right). \tag{3.1}$$

Notemos que o quantil $c_{n,h}(u)$ não depende da distribuição considerada sob a hipótese nula e ainda que a invariância de cada uma das estatísticas implica a invariância de $T_n(u)$, para cada $u \in ]0,1[$. Desta forma, a região crítica do novo teste será então

$$R = \bigcup_{h \in H} \{T_{n,h} > c_{n,h}(u_{n,\alpha})\} = \{T_n(u_{n,\alpha}) > 0\},$$

onde

$$u_{n,\alpha} = \sup I_{n,\alpha}, \tag{3.2}$$

com

$$I_{n,\alpha} = \{u \in ]0,1[ \colon P_\phi \left( T_n(u) > 0 \right) \leq \alpha\}$$

e $\phi$ a densidade gaussiana standard. Na prática, o valor de $u_{n,\alpha}$ é estimado recorrendo a simulações de Monte Carlo sob a hipótese nula, sendo aqui fundamental o facto das estatísticas $T_{n,h}$, $h \in H$, serem invariantes sob $H_0$.

Denotando por $F_{T_{n,h}}$ a função distribuição e $F^{-1}_{T_{n,h}}$ a função quantil da estatística $T_{n,h}$ sob $H_0$, temos

$$\begin{aligned} P_\phi \left( T_n \left( \frac{\alpha}{|H|} \right) > 0 \right) &\leq \sum_{h \in H} P_\phi \left( T_{n,h} > c_{n,h} \left( \frac{\alpha}{|H|} \right) \right) \\ &\leq \sum_{h \in H} \left\{ 1 - F_{T_{n,h}} \left( F^{-1}_{T_{n,h}} \left( 1 - \frac{\alpha}{|H|} \right) \right) \right\} \leq \alpha \end{aligned}$$

Portanto,

$$\frac{\alpha}{|H|} \in I_{n,\alpha} \quad \text{e} \quad \frac{\alpha}{|H|} \leq u_{n,\alpha},$$

o que nos mostra que o teste múltiplo proposto é pelo menos tão potente quanto o procedimento de Bonferroni.

Sob determinadas condições na distribuição das estatísticas $T_{n,h}, h \in H$ podemos mostrar que o teste obtido por reunião das regiões críticas terá nível de significância menor ou igual a $\alpha$, isto é,

$$P_\phi(R) \leq \alpha.$$

Sendo $T_n(u)$ invariante para cada $u \in ]0,1[$, este resultado depende essencialmente das propriedades de continuidade da função

$$\psi(u) = P_\phi\left(T_n(u) > 0\right)$$

definida no intervalo $]0,1[$, apresentadas no seguinte lema.

**Lema 3.1.** *Para $n \in \mathbb{N}$, a função $\psi$ é crescente com $\lim_{u\downarrow 0}\psi(u) = 0$. Além disso, $\psi$ satisfaz:*

*a) Se $F_{T_{n,h}}$ é estritamente crescente para todo o $h \in H$, $\psi$ é contínua à esquerda.*

*b) Se $F_{T_{n,h}}$ é contínua para todo o $h \in H$, $\psi$ é contínua à direita.*

*Demonstração.* Comecemos por provar que $\phi$ é uma função crescente. Sejam $u, v \in ]0,1[$ tais que $u < v$. Para todo o $h \in H$, temos $c_{n,h}(u) \geq c_{n,h}(v)$ e portanto

$$T_n(u) = \max_{h\in H}\left(T_{n,h} - c_{n,h}(u)\right) \leq \max_{h\in H}\left(T_{n,h} - c_{n,h}(v)\right) = T_n(v).$$

Isto implica que

$$\psi(u) = P_\phi\left(T_n(u) > 0\right) \leq P_\phi\left(T_n(v) > 0\right) = \psi(v),$$

o que prova o pretendido.

Além disso, atendendo à Proposição 1 de Shorack e Wellner (1986, p.5),

$$P_\phi\left(T_n(u) > c_{n,h}(u)\right) = 1 - F_{T_{n,h}}(c_{n,h}(u)) = 1 - F_{T_{n,h}}\left(F^{-1}_{T_{n,h}}(1-u)\right) \leq 1-(1-u) = u$$

e

$$\lim_{u\downarrow 0} P_\phi\left(T_n(u) > c_{n,h}(u)\right) = 0, \text{ para todo o } h \in H.$$

Portanto,

$$\lim_{u\downarrow 0}\psi(u) \leq \sum_{h\in H}\lim_{u\downarrow 0} P_\phi\left(T_{n,h} > c_{n,h}(u)\right) = 0.$$

a) Para $u \in ]0,1[$ fixo, seja $u_m$ uma sucessão com $u_m \uparrow u$. Sabendo por hipótese que a função de distribuição de $T_{n,h}$ sob $H_0$ é estritamente crescente para todo o $h \in H$, temos que a função $F^{-1}_{T_{n,h}}$ é contínua à direita para cada $h \in H$, o que nos leva a

$$c_{n,h}(u_m) = F^{-1}_{T_{n,h}}(1-u_m) \downarrow F^{-1}_{T_{n,h}}(1-u) = c_{n,h}(u), \text{ para todo o } h \in H.$$

Portanto,

$$T_n(u_m) \uparrow T_n(u) \quad \text{e} \quad \psi(u_m) = P_\phi\left(T_n(u_m) > 0\right) \uparrow P_\phi\left(T_n(u) > 0\right) = \psi(u).$$

b) Para $u \in ]0,1[$ fixo, seja $u_m$ uma sucessão com $u_m \downarrow u$. Pela continuidade à esquerda de $F^{-1}_{T_{n,h}}$ temos

$$c_{n,h}(u_m) = F^{-1}_{T_{n,h}}(1-u_m) \uparrow F^{-1}_{T_{n,h}}(1-u) = c_{n,h}(u), \text{ para todo o } h \in H$$

e

$$T_n(u_m) \downarrow T_n(u).$$

Sendo $u_m$ uma sucessão decrescente, isto é, $u_m \geq u_{m+1}$, temos $c_{n,h}(u_m) \leq c_{n,h}(u_{m+1})$ e ainda $T_n(u_m) \geq T_n(u_{m+1})$. Logo,

$$\{T_n(u) > 0\} \subset \bigcap_m \{T_n(u_m) > 0\} \subset \{T_n(u) \geq 0\}.$$

Finalmente, $\psi(u) \leq \lim_m \psi(u_m) \leq \psi(u) + P_\phi(T_n(u) = 0)$, onde

$$P_\phi(T_n(u) = 0) \leq \sum_{h \in H} P_\phi(T_{n,h} = c_{n,h}(u)) = 0,$$

pela continuidade de $T_{n,h}$ sob $H_0$ para todo o $h \in H$.

□

No que se segue, veremos quais as condições suficientes para que o teste tenha nível inferior ou igual a $\alpha$.

**Teorema 3.2.** *Se para todo o $h \in H$, $F_{T_{n,h}}$ é estritamente crescente (no conjunto $\{t : 0 < F_{T_{n,h}}(t) < 1\}$), então para todo o $f \in \mathcal{N}$, tem-se*

$$P_f(R) \leq \alpha, \text{ para } 0 < \alpha < 1.$$

*Além disso, se $F_{T_{n,h}}$ é contínua para todo o $h \in H$ então $u_{n,\alpha} \leq \alpha$ e para todo o $f \in \mathcal{N}$ tem-se*

$$P_f(R) = \alpha.$$

*Demonstração.* Usando o facto de $\psi$ ser uma função crescente, deduzimos que $I_{n,\alpha} = ]0, u_{n,\alpha}[$ ou $I_{n,\alpha} = ]0, u_{n,\alpha}]$, uma vez que $u_{n,\alpha} = \sup I_{n,\alpha}$. Tomando $u_m \in I_{n,\alpha}$ tal que $u_m \uparrow u_{n,\alpha}$, pela parte a) do Lema 3.1, concluímos que

$$\psi(u_{n,\alpha}) = \psi\left(\lim_m u_m\right) = \lim_m \psi(u_m) = \lim_m P_\phi(T_n(u_m) > 0) \leq \alpha,$$

o que prova que o nível de significância do teste múltiplo é no máximo $\alpha$, quando $F_{T_{n,h}}$ é estritamente crescente para todo o $h \in H$.

Adicionalmente, assumindo que $F_{T_{n,h}}$ é contínua para todo o $h \in H$, pela parte (b) do Lema 3.1 e para uma sucessão $u_m$ tal que $u_m \downarrow u_{n,\alpha}$, temos $\psi(u_m) > \alpha$ (pela definição de $u_{n,\alpha}$) e

$$\psi(u_{n,\alpha}) = \psi\left(\lim_m u_m\right) = \lim_m \psi(u_m) \geq \alpha.$$

Portanto, $\psi(u_{n,\alpha}) = \alpha$, o que prova que o teste múltiplo tem um nível de significância igual a $\alpha$.

Finalmente, iremos provar que $u_{n,\alpha} \leq \alpha$, usando o facto de existir um $h \in H$ tal que $F_{T_{n,h}}$ é contínua. Para este $h$ e para $u \in ]0,1[$, temos

$$\{T_{n,h} > c_{n,h}(u)\} \subset \left\{\max_{h\in H}(T_{n,h} > c_{n,h}(u))\right\} = \{T_n(u) > 0\}$$

e então

$$\{u \in ]0,1[ \colon P_\phi(T_n(u) > 0) \leq \alpha\} \subset \{u \in ]0,1[ \colon P_\phi(T_{n,h} > c_{n,h}(u)) \leq \alpha\}.$$

Sabendo que $F_{T_{n,h}} \circ F^{-1}_{T_{n,h}}$ é a função identidade e pelo Lema 3.1 obtemos

$$\begin{aligned} u_{n,\alpha} &= \sup\{u \in ]0,1[ \colon P_\phi(T_n(u) > 0) \leq \alpha\} \\ &\leq \sup\{u \in ]0,1[ \colon P_\phi(T_{n,h} > c_{n,h}(u)) \leq \alpha\} \\ &\leq \sup\{u \in ]0,1[ \colon 1 - F_{T_{n,h}}\left(F^{-1}_{T_{n,h}}(1-u)\right) \leq \alpha\} \\ &= \sup\{u \in ]0,1[ \colon 1 - (1-u) \leq \alpha\} \\ &= \sup\{u \in ]0,1[ \colon u \leq \alpha\} = \alpha. \end{aligned}$$

□

Por último, iremos estabelecer um resultado que determina a convergência do teste múltiplo.

Sob as condições referidas no lema e teorema anteriores, para uma alternativa fixa $f$, a potência do teste múltiplo

$$P_f(T_n(u_{n,\alpha}) > 0)$$

satisfaz a dupla desigualdade

$$\max_{h\in H} P_f(T_{n,h} > c_{n,h}(u_{n,\alpha})) \leq P_f(T_n(u_{n,\alpha}) > 0) \leq \sum_{h\in H} P_f(T_{n,h} > c_{n,h}(\alpha)). \quad (3.3)$$

As principais características do teste múltiplo são apresentadas em (3.3), isto é, o novo teste apresenta potência inferior para alternativas que mostram potência inferior

para cada um dos testes baseados em $T_{n,h}$, $h \in H$. No entanto, a potência do novo teste é sempre superior à do melhor dos testes realizados ao nível $u_{n,\alpha}$ envolvidos na combinação.

Sob certas condições, o procedimento de teste múltiplo é convergente para toda a distribuição alternativa $f$, como se pode verificar no teorema seguinte.

**Teorema 3.3.** *Seja $f$ uma densidade de probabilidade não normal e assumimos que existe um $h \in H$ tal que $T_{n,h} \xrightarrow{p} +\infty$, sob $f$. Se*

$$T_{n,h} \xrightarrow{d} T_{\infty,h}$$

*sob $H_0$, onde a função de distribuição de $T_{\infty,h}$ é estritamente crescente, então*

$$P_f(T_n(u_{n,\alpha}) > 0) \to 1, n \to \infty.$$

*Demonstração.* Seja $f \notin \mathcal{N}$ e seja $h \in H$ tal que

$$T_{n,h} \xrightarrow{p} +\infty. \tag{3.4}$$

Uma vez que $c_{n,h}(u_{n,\alpha}) \leq c_{n,h}\left(\frac{\alpha}{|H|}\right)$, temos

$$P_f(T_n(u_{n,\alpha}) > 0) \geq P_f(T_{n,h} > c_{n,h}(u_{n,\alpha})) \geq P_f\left(T_{n,h} > c_{n,h}\left(\frac{\alpha}{|H|}\right)\right). \tag{3.5}$$

Atendendo a que $F^{-1}_{T_{\infty,h}}$ é contínua em $]0,1[$ (uma vez que $F_{T_{\infty,h}}(t)$ é estritamente crescente), a convergência em distribuição $T_{n,h} \xrightarrow{d} T_{\infty,h}$ implica a convergência $F^{-1}_{T_{n,h}}(t) \to F^{-1}_{T_{\infty,h}}(t)$, para todo o $t \in ]0,1[$ (ver Shorack e Wellner, 1986, p.10). Assim, temos

$$c_{n,h}\left(\frac{\alpha}{|H|}\right) = F^{-1}_{T_{n,h}}\left(1 - \frac{\alpha}{|H|}\right) \to F^{-1}_{T_{\infty,h}}\left(1 - \frac{\alpha}{|H|}\right). \tag{3.6}$$

Finalmente, de (3.4), (3.5) e (3.6) deduzimos que

$$P_f(T_n(u_{n,\alpha}) > 0) \geq P_f\left(T_{n,h} > \sup_{n\in\mathbb{N}} c_{n,h}\left(\frac{\alpha}{|H|}\right)\right) \to 1.$$

□

## 3.3. Um teste múltiplo baseado nas estatísticas $D_{n,\beta}$

Nesta secção, vamos definir o teste múltiplo que iremos considerar, explicitando os testes usados na combinação, e verificar que o teste resultante dessa combinação satisfaz as condições enunciadas na secção anterior.

Iremos assim considerar o teste múltiplo baseado nas estatísticas

$$T_{n,1} = D_{n,0}, \quad T_{n,2} = nD_{n,\beta_1}, \quad T_{n,3} = nD_{n,\beta_2} \quad \text{e} \quad T_{n,4} = D_{n,\infty},$$

com $0 < \beta_1 < \beta_2 < \infty$, que tem como região crítica

$$\{T_n(u_{n,\alpha}) > 0\},$$

onde $T_n$ e $u_{n,\alpha}$ são definidos em (3.1) e (3.2), respetivamente.

De acordo com a breve referência que foi feita na secção 3.1, o teste múltiplo que consideramos é baseado na estatística de teste $D_{n,\beta}$, escolhendo valores extremos $\beta = 0$ e $\beta = \infty$ e dois valores intermédios $\beta_1 = 0.725$ e $\beta_2 = 1.57$. As escolhas dos valores de $\beta_1$ e $\beta_2$ são as sugeridas em Tenreiro (2009), valores esses obtidos a partir de um estudo de simulação. A partir desse estudo sabemos que o teste baseado em $D_{n,\beta_1}$ é adequado para distribuições simétricas ou com caudas leves enquanto que o teste baseado em $D_{n,\beta_2}$ é apropriado para distribuições com caudas pesadas.

O teorema seguinte estabelece que o teste múltiplo tem um nível de significância que é menor ou igual a $\alpha$.

**Teorema 3.4.** *Para $n > 1$ e $\alpha \in ]0,1[$ temos*

$$P_f(T_n(u_{n,\alpha}) > 0) \leq \alpha.$$

*Demonstração.* Tendo em conta o Teorema 3.2, basta mostrar que a função de distribuição de $T_{n,h}$ é estritamente crescente, para todo o $h \in H$. Notemos que as estatísticas $T_{n,h}$, $h \in H$, são contínuas e definidas no subconjunto aberto de $\mathbb{R}^n$ dado por

$$\mathcal{D} = \{x = (x_1, \ldots, x_n) \in \mathbb{R}^n : s_n^2(x) > 0\} \text{ com } P_\phi(\mathcal{D}) = 1,$$

onde

$$s_n^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2 \quad \text{e} \quad \bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j.$$

Pela continuidade de $T_{n,h}$, $h \in H$, para todo o $s < t$ com $0 < F_{T_{n,h}}(s) \leq F_{T_{n,h}}(t) < 1$, concluímos que $T_{n,h}^{-1}(]s,t[)$ é um subconjunto não vazio e aberto de $\mathbb{R}^n$.

Suponhamos por absurdo que $F_{T_{n,h}}(s) = F_{T_{n,h}}(t)$. Isto implica que $P_\phi(T_{n,h} \leq s) = P_\phi(T_{n,h} \leq t)$ e, portanto,

$$P_\phi\left(T_{n,h}^{-1}(]s,t[)\right) = 0,$$

o que é falso uma vez que o subconjunto $T_{n,h}^{-1}(]s,t[)$ é não vazio. □

Na prática, o valor $u_{n,\alpha}$, o nível a que cada um dos testes baseados em $T_{n,h}$, $h \in H$, é executado, é estimado recorrendo a simulações de Monte Carlo sob a hipótese nula (ver Fromont e Laurent, 2006). Em primeiro lugar, foram geradas

100000 valores de cada uma das quatro estatísticas de teste sob $H_0$. Usando os primeiros 50000 valores, foram estimados os quantis $c_{n,h}(u)$, onde $u$ varia na malha regular do intervalo $]0,1[$ definida por

$$u_{i+1} = u_i + p, \text{ onde } u_1 = p \quad \text{e} \quad p = 0.0001.$$

Os restantes 50000 valores foram usados para estimar as probabilidades $P_\phi(T_n(u) > 0)$. Por último, tomámos o maior dos valores de $u$ que satisfaz a desigualdade $P_\phi(T_n(u) > 0) \leq \alpha$ como uma aproximação para $u_{n,\alpha}$ definido por (3.2).

Considerando os níveis de significância $\alpha = 0.01$ e $\alpha = 0.05$, e ainda diferentes tamanhos de amostras $n$, apresentamos na Tabela 3.1 o nível $u_{n,\alpha}$ em que cada teste é realizado.

| $\alpha$ | Tamanho da amostra | | | | |
|---|---|---|---|---|---|
| | 25 | 50 | 100 | 200 | 400 |
| 0.01 | 3.7e-03 | 3.4e-03 | 3.5e-03 | 3.6e-03 | 3.5e-03 |
| 0.05 | 2.08e-02 | 2.04e-02 | 1.94e-02 | 1.85e-02 | 1.85e-02 |

Tabela 3.1: Valores estimados para $u_{n,\alpha}$, onde $\alpha = 0.01$ e $\alpha = 0.05$, tendo em conta uma malha regular de tamanho 0.0001 no intervalo $]0,1[$. O número de repetições para cada fase do processo de estimação é 50000.

Apesar de não termos conseguido provar que a função de distribuição das estatísticas de teste $T_{n,h}$ é contínua para todo o $h \in H$, ou seja, que o teste múltiplo tem nível de significância exatamente igual a $\alpha$, observamos na tabela anterior que o nível $u_{n,\alpha}$ é claramente maior que $\alpha/4$, o nível de cada um dos testes baseados em $T_{n,h}$, $h \in H$, considerado no método de Bonferroni.

Considerando os mesmos valores de $\alpha$ e $n$, a Tabela 3.2 apresenta uma estimação para o nível de significância do teste múltiplo, baseada em 100000 simulações sob a hipótese nula.

| $\alpha$ | Tamanho da amostra | | | | |
|---|---|---|---|---|---|
| | 25 | 50 | 100 | 200 | 400 |
| 0.01 | 9.0e-03 | 8.3e-03 | 9.1e-03 | 9.5e-03 | 1.1e-02 |
| 0.05 | 4.94e-02 | 4.96e-02 | 4.94e-02 | 5.04e-02 | 5.28e-02 |

Tabela 3.2: Valores estimados para o nível de significância do teste múltiplo, para $\alpha = 0.01$ e $\alpha = 0.05$. O número de repetições para cada caso é 100000.

Na maioria dos casos, a Tabela 3.2 mostra que o teste múltiplo que estamos a considerar apresenta um nível de significância bastante próximo de $\alpha$.

Por último, vamos verificar que o teste múltiplo é convergente para cada alternativa fixa.

**Teorema 3.5.** *Para $\alpha \in ]0,1[$, temos*

$$P_f\left(T_n\left(u_{n,\alpha} > 0\right)\right) \to 1,\ n \to \infty,$$

*para todo a densidade $f \notin \mathcal{N}$.*

*Demonstração.* Dada uma densidade $f$ não normal, da demonstração do Teorema 1.8, sabemos que

$$T_{n,2} = nD_{n,\beta_1} \xrightarrow{p} +\infty \text{ sob } f.$$

Pelo Teorema 1.6, sabemos que a distribuição limite de $T_{n,2}$ sob $H_0$, que denotamos por $T_{\infty,2}$, é uma soma ponderada de qui-quadrados independentes. Assim, $F_{T_{\infty,2}}$ é estritamente crescente $\left(\text{no conjunto } \left\{t : 0 < F_{T_{\infty,2}} < 1\right\}\right)$, sendo agora possível aplicar o Teorema 3.3 para provar o pretendido. □

# Capítulo 4

# Estudo de simulação

Neste capítulo apresentamos um breve estudo de simulação para avaliar a potência do teste múltiplo proposto na Secção 3.3 e compará-la com a de outros testes de normalidade recomendados na literatura, como são os casos dos testes de Shapiro e Wilk (1965) e de Anderson e Darling (1954).

## 4.1. Os testes de Shapiro-Wilk e Anderson-Darling

Nesta secção, vamos fazer uma breve descrição dos testes de normalidade usados para além do teste múltiplo (TM).

A estatística de Shapiro-Wilk (SW) é definida por

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2},$$

onde $X_{(i)}$ é a i-ésima estatística associada à amostra $X_1, \ldots, X_n$, $\bar{X}$ é a média da amostra e as constantes $a_i$ são dadas por

$$a' = (a_1, \ldots, a_n) = \frac{m'V^{-1}}{(m'V^{-1}V^{-1}m)^{1/2}},$$

onde $m' = (m_1, \ldots, m_n)$ é o vetor dos valores esperados das estatísticas ordenadas e $V = (v_{ij})$ é a correspondente matriz de covariância. Através do estudo de simulação efetuado por D'Agostino e Stephens (1986, pp. 403–406), sabemos que este teste apresenta boas propriedades de potência para todos os tipos de distribuições e tamanhos de amostra nele considerados.

A estatística de Anderson-Darling (AD) é definida pela distância quadrática ponderada entre a função de distribuição empírica da amostra e a função de distribuição de uma distribuição normal e é dada por

$$W_n^2 = n \int \frac{[F_n(x) - F_0(x)]^2}{F_0(x)\,(1 - F_0(x))}\, dF_0(x),$$

onde $F_0$ é a função de distribuição da normal standard e $F_n$ é a função de distribuição empírica associada aos resíduos standardizados (1) dada por

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{]-\infty,x[}(Y_i).$$

Referindo novamente o estudo de simulação de D'Agostino e Stephens (1986, pp. 403–406), este teste possui propriedades de potência semelhantes às do teste SW.

## 4.2. Distribuições alternativas

O conjunto de distribuições que consideramos no nosso estudo inclui algumas distribuições anteriormente consideradas em outros estudos de simulação, tais como os de Epps e Pulley (1983), Noughabi e Arghami (2011), Razali e Wah (2011), Yap e Sim (2011). Vamos começar por considerar quatro distribuições: a uniforme no intervalo $[0, 1]$, que denotaremos por $U(]0, 1[)$, a beta de parâmetros 2 e 1, que denotaremos por $B(2, 1)$, a Student com 6 graus de liberdade, que denotaremos por $t_6$, e a lognormal de parâmetros 0 e 0.5, que denotaremos por $\mathrm{LN}(0, 0.5)$. As duas primeiras são distribuições de caudas leves, enquanto que as outras duas são distribuições com caudas pesadas onde, em cada tipo de distribuição, existe uma simétrica e uma assimétrica. Vamos ainda considerar duas outras distribuições tais como a $t_{10}$ e a exponencial generalizada (ver Jonhson et al., 1980) de parâmetros 0, 1, 0.21 e 0.15, que denotaremos apenas por GEP. Estas distribuições podem ser consideradas mais próximas da distribuição normal do que as outras. Em particular, a distribuição GEP tem momentos de ordens 3 e 4 exatamente iguais aos da distribuição normal.

## 4.3. Resultados de Potência

Os resultados que apresentamos nas Tabelas 4.1 a 4.5 baseiam-se em 10000 amostras de tamanhos $n = 25, 50, 100, 200, 400$, para o conjunto de distribuições considerado, usando os níveis de significância habituais $\alpha = 0.01$ e $\alpha = 0.05$.

O teste múltiplo mostra em geral bons resultados de potência, com exceção das distribuições de caudas leves em que é claramente inferior aos testes AD e SW. Para as restantes alternativas, o teste múltiplo nunca é o pior dos testes considerados, o que revela que este teste possui uma potência razoável para as alternativas consideradas.

Relativamente às distribuições $t_6$ e $\mathrm{LN}(0, 0.5)$ (caudas pesadas), o teste AD é aquele que mostra piores resultados sendo superado pelos outros dois testes. Somos

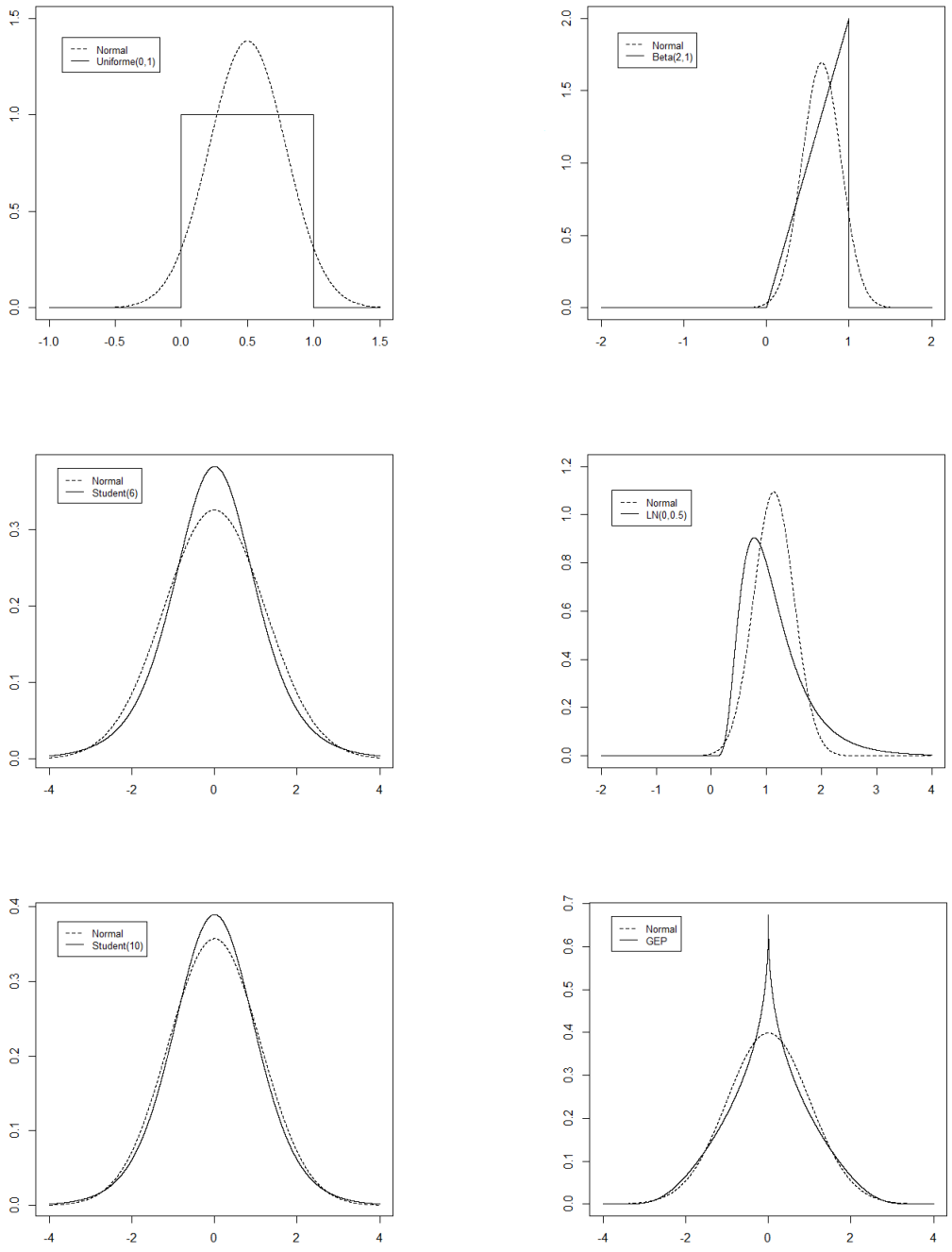



Figura 4.1: Distribuições usadas no estudo de simulação.

ainda levados a concluir que quanto mais pesadas são as caudas das distribuições, mais sensível são os testes na deteção destas alternativas.

Para distribuições mais próximas da distribuição normal, tais como a $t_{10}$ e a GEP, o teste múltiplo mostra resultados ligeiramente melhores do que os testes AD e SW, respetivamente. Podemos ainda realçar que qualquer um dos três testes mostra

| Testes | Distribuições Alternativas | | | | | |
|---|---|---|---|---|---|---|
| | $U([0,1])$ | $B(2,1)$ | $t_6$ | LN$(0,0.5)$ | $t_{10}$ | GEP |
| | $\alpha = 0.01$ | | | | | |
| AD | 6.1 | 12.9 | 6.6 | 35.7 | 2.9 | 1.5 |
| SW | 5.2 | 13.2 | 9.1 | 41.9 | 4.1 | 1.0 |
| TM | 3.1 | 7.8 | 10.0 | 35.8 | 4.5 | 1.2 |
| | $\alpha = 0.05$ | | | | | |
| AD | 23.1 | 34.4 | 16.2 | 56.6 | 9.6 | 6.2 |
| SW | 28.5 | 41.1 | 18.2 | 63.3 | 11.2 | 4.8 |
| TM | 13.6 | 24.3 | 21.2 | 56.9 | 11.9 | 5.4 |

Tabela 4.1: Valores estimados para a potência dos três testes relativamente a seis distribuições alternativas, tendo em conta uma amostra de tamanho $n = 25$.

| Testes | Distribuições Alternativas | | | | | |
|---|---|---|---|---|---|---|
| | $U([0,1])$ | $B(2,1)$ | $t_6$ | LN$(0,0.5)$ | $t_{10}$ | GEP |
| | $\alpha = 0.01$ | | | | | |
| AD | 27.2 | 43.1 | 11.3 | 73.5 | 4.2 | 1.8 |
| SW | 35.9 | 52.7 | 17.2 | 81.6 | 7.6 | 0.7 |
| TM | 16.9 | 28.5 | 18.4 | 74.7 | 8.0 | 1.1 |
| | $\alpha = 0.05$ | | | | | |
| AD | 58.1 | 71.7 | 23.6 | 87.7 | 12.5 | 8.2 |
| SW | 74.7 | 84.1 | 28.0 | 92.5 | 16.3 | 4.8 |
| TM | 42.5 | 56.8 | 33.8 | 89.0 | 19.5 | 6.6 |

Tabela 4.2: Valores estimados para a potência dos três testes relativamente a seis distribuições alternativas, tendo em conta uma amostra de tamanho $n = 50$.

baixa potência para as alternativas referidas. O contrário também é comprovado, ou seja, para distribuições "mais afastadas" da normal é visível uma potência bastante elevada, o que nos leva a concluir que qualquer um dos testes é muito sensível na deteção destas distribuições.

Os três testes apresentam uma potência razoável para a generalidade das alternativas consideradas. Apesar das limitações do estudo que apresentamos, este parece indicar que o teste múltiplo considerado possui uma potência empírica que é comparável à dos testes AD e SW, testes estes que estão entre os mais recomendados testes de normalidade.

| Testes | Distribuições Alternativas | | | | | |
|---|---|---|---|---|---|---|
| | $U([0,1])$ | $B(2,1)$ | $t_6$ | LN$(0,0.5)$ | $t_{10}$ | GEP |
| | $\alpha = 0.01$ | | | | | |
| AD | 79.1 | 90.4 | 19.5 | 97.7 | 6.3 | 3.4 |
| SW | 94.7 | 97.5 | 30.3 | 99.3 | 12.4 | 1.1 |
| TM | 61.5 | 77.7 | 33.7 | 98.4 | 13.5 | 1.7 |
| | $\alpha = 0.05$ | | | | | |
| AD | 95.1 | 98.1 | 36.0 | 99.4 | 16.4 | 12.4 |
| SW | 99.6 | 99.9 | 44.5 | 99.9 | 23.6 | 5.9 |
| TM | 85.8 | 93.7 | 51.8 | 99.6 | 28.4 | 7.7 |

Tabela 4.3: Valores estimados para a potência dos três testes relativamente a seis distribuições alternativas, tendo em conta uma amostra de tamanho $n = 100$.

| Testes | Distribuições Alternativas | | | | | |
|---|---|---|---|---|---|---|
| | $U([0,1])$ | $B(2,1)$ | $t_6$ | LN$(0,0.5)$ | $t_{10}$ | GEP |
| | $\alpha = 0.01$ | | | | | |
| AD | 99.8 | 100 | 37.8 | 100 | 10.2 | 7.4 |
| SW | 100 | 100 | 54.2 | 100 | 21.5 | 1.7 |
| TM | 98.7 | 99.8 | 59.9 | 100 | 23.9 | 3.7 |
| | $\alpha = 0.05$ | | | | | |
| AD | 100 | 100 | 57.4 | 100 | 23.9 | 22.3 |
| SW | 100 | 100 | 68.9 | 100 | 35.7 | 10.0 |
| TM | 99.9 | 100 | 75.4 | 100 | 41.2 | 12.1 |

Tabela 4.4: Valores estimados para a potência dos três testes relativamente a seis distribuições alternativas, tendo em conta uma amostra de tamanho $n = 200$.

| Testes | Distribuições Alternativas | | | | | |
|---|---|---|---|---|---|---|
| | $U([0,1])$ | $B(2,1)$ | $t_6$ | LN$(0,0.5)$ | $t_{10}$ | GEP |
| | $\alpha = 0.01$ | | | | | |
| AD | 100 | 100 | 69.6 | 100 | 19.9 | 20.7 |
| SW | 100 | 100 | 83.1 | 100 | 39.2 | 5.9 |
| TM | 100 | 100 | 87.6 | 100 | 42.2 | 10.1 |
| | $\alpha = 0.05$ | | | | | |
| AD | 100 | 100 | 84.6 | 100 | 40.2 | 44.3 |
| SW | 100 | 100 | 91.2 | 100 | 56.9 | 24.0 |
| TM | 100 | 100 | 94.9 | 100 | 63.3 | 25.4 |

Tabela 4.5: Valores estimados para a potência dos três testes relativamente a seis distribuições alternativas, tendo em conta uma amostra de tamanho $n = 400$.

# Apêndice A

# Um teorema de convergência

**Teorema A.1.** *Suponhamos que $X_n(\omega;t)$, $X(\omega;t)$ e $Y(\omega;t)$ são funções mensuráveis definidas em $\Omega \times \Omega_0$ e sejam $P$ uma probabilidade em $\mathcal{B}$ e $\mu$ uma medida finita em $\mathcal{B}_0$, com $\mathcal{B}$ e $\mathcal{B}_0$ $\sigma$-álgebras em $\Omega$ e $\Omega_0$, respetivamente. Se*

$$X_n(\omega;t) \xrightarrow{p} X(\omega;t), \mu \text{ } q.c.$$

$$|X_n(\omega;t)| \leq Y(\omega;t)$$

*e*

$$\int\int |Y(\omega;t)|\, dP \otimes \mu < \infty$$

*então*

$$\int |X_n(\omega;t) - X(\omega;t)|\, d\mu(t) \xrightarrow{p} 0$$

*e ainda*

$$\int X_n(\omega;t) d\mu(t) \xrightarrow{p} \int X(\omega;t) d\mu(t).$$

# Apêndice B

# Códigos em R

```
,
################################################
# ESTATÍSTICAS DE TESTE
# (para observações standardizadas)

Dn.z = function(y)  # Estatística D_{n,0}
   {
    (mean(y^3)/(mean(y^2))^(3/2))^2
   }

Dn.b = function(y,beta)  # Estatística n*D_{n,beta}
   {
    b <- (1+beta^2)^(-1/2)
    bb <- (1+2*beta^2)^(-1/2)
    dif <- outer(y,y,"-")
    length(y)*(mean(exp(-beta^2*dif^2/2)) - 2*b*mean(exp(-(beta*b)^2*y
        ^2/2)) + bb)
   }

Dn.i = function(y)  # Estatística D_{n,infinito}
   {
    mean(exp(-y^2/2))
   }

# Valores de beta_1 e beta_2
b1 = 1/(sqrt(2)*0.975)
b2 = 1/(sqrt(2)*0.45)


################################################
# VALORES DAS ESTATÍSTICAS SOB N(0,1)

null.values = function(n,rep=100000)
    {
```

```
      set.seed(458755, kind = NULL)
      test.stat<-array(dim=c(rep,4))

      for (i in 1:rep)
      {
       x <- rnorm(n)
       y <- (x-mean(x))/sd(x)

       test.stat[i,1]<-Dn.z(y)
       test.stat[i,2]<-Dn.i(y)
       test.stat[i,3]<-Dn.b(y,beta=b1)
       test.stat[i,4]<-Dn.b(y,beta=b2)
      }

      #test.stat
      if (n<100) nn <- paste(0,n,sep="") else nn <- n
      write.table(format(test.stat,scientific=TRUE),file=paste("valores
          _n",nn,".tex",sep=""),col.names=FALSE,row.names=FALSE)
     }



#####################################################################
# CORREÇÕES

# 50000 observações são usadas para o cálculo dos quantis e outras
    50000
# para o cálculo de psi(u)

alphas = c(0.01,0.05)
lalphas = length(alphas)

correccoes <- function(n, n.obs = 50000)
     {
      passo <- 0.0001
      u <- seq(passo,1,passo)
      lu <- length(u)

      c.alpha <- array(dim=c(5,lalphas))
      #linhas 1-4 = correções c_{n,h}(u_{n,alpha}) para cada um dos 4
          testes
```

```
#linha 5 - corresponde ao nível u_{n,alpha} em que cada teste é
    usado

quantiles.u <- array(dim=c(4,lu))
T.u <- array(dim=c(4,lu))

T <- array(dim=lu)

if (n < 100) nn <- paste(0,n,sep="") else  nn <- n

# As estatísticas D0, Di, D1, D2 surgem nas colunas de valores
    (100000 linhas)

null.values <- as.matrix(read.table(paste("valores_n",nn,".tex",
    sep="")))

# Linhas 1:n.obs usadas para estimar os quantis

for (j in 1:4) quantiles.u[j,]<-quantile(null.values[1:n.obs,j
    ],1-u,names=FALSE)

# Linhas n.obs+1:n.obs+n.obs usadas para estimar as
    probabilidades
# psi(u) = P_0( T_n(u) > 0 ) onde T_n(u)= max_h (T_nh - c_nh(u))

T <- 0
for (i in 1:n.obs)
{
 for (j in 1:4) T.u[j,] <- null.values[n.obs+i,j] - quantiles.u[j
     ,]
 T <- T + 1*(apply(T.u,MARGIN=2,FUN=max)>0)
}

psi <- T/n.obs

for (j in 1:lalphas)
{
 ind <- sum(psi<= alphas[j])
 for (i in 1:4) c.alpha[i,j] <- quantiles.u[i,max(ind,1)]
 c.alpha[5,j] <- u[max(ind,1)]
}
```

```
     write.table(format(c.alpha,scientific=TRUE),file=paste("
         correccoes_n",nn,".tex",sep=""),col.names=FALSE,row.names=
         FALSE)
    }


#####################################################################
# DISTRIBUIÇÕES ALTERNATIVAS

# Distribuição Exponencial Generalizada

rGEP = function(n,mean,sd,alpha,tau)
     {
      W <- sqrt(3*gamma(alpha)/gamma(alpha+2*tau))*(rgamma(n,alpha))^
          tau
      return(sd*W*(2*runif(n)-1)+mean)
     }

alternativa = function(dist)
     {
      if (dist==0) gerador = function(n){return(rnorm(n))} else
      if (dist==1) gerador = function(n){return(runif(n))} else
      if (dist==2) gerador = function(n){return(rbeta(n,2,1))} else
      if (dist==3) gerador = function(n){return(rt(n,6))} else
      if (dist==4) gerador = function(n){return(rlnorm(n,0,0.5))} else
      if (dist==5) gerador = function(n){return(rt(n,10))} else
      if (dist==6) gerador = function(n){return(rGLD(n,0,1,0.21,0.15))
          }
     }


#####################################################################
# POTÊNCIA

potencia = function(dist,n,rep)
    {
     if (dist<10) dd <- paste(0,dist,sep="") else  dd <- dist
     if (n<100) nn <- paste(0,n,sep="") else  nn <- n
     correccoes <- as.matrix(read.table(paste("correccoes_n",nn,".tex"
         ,sep="")))

     rejeicao5<-array(dim=c(rep,3))
```

```
rejeicao1<-array(dim=c(rep,3))

pvalor<-array(dim=c(rep))

gerador<-alternativa(dist)

#set.seed(878555, kind = NULL)

for (i in 1:rep)
{
 x <- gerador(n)
 y <- (x-mean(x))/sd(x)

 # Teste combinado
 C1 <- Dn.z(y) - correccoes[1,]
 C2 <- Dn.i(y) - correccoes[2,]
 C3 <- Dn.b(y,beta=b1) - correccoes[3,]
 C4 <- Dn.b(y,beta=b2) - correccoes[4,]
 rejeicao1[i,1]<-1*(max(c(C1[1],C2[1],C3[1],C4[1]))>0)
 rejeicao5[i,1]<-1*(max(c(C1[2],C2[2],C3[2],C4[2]))>0)

 # Teste de Shapiro-Wilk
 pvalueSW <- shapiro.test(y)$p.value
 rejeicao1[i,2]<-1*(pvalueSW <= 0.01)
 rejeicao5[i,2]<-1*(pvalueSW <= 0.05)

 # Teste de Anderson-Darling
 pvalueAD <- ad.test(y)$p.value
 pvalor[i] <- pvalueAD
 if (pvalor[i]=='NaN') pvalor[i]=0
 rejeicao1[i,3]<-1*(pvalor[i] <= 0.01)
 rejeicao5[i,3]<-1*(pvalor[i] <= 0.05)

 }

potencia1<-apply(rejeicao1,MARGIN=2,mean)
potencia5<-apply(rejeicao5,MARGIN=2,mean)

potencia <- matrix(c(potencia1,potencia5),ncol=3,byrow=TRUE)

escrever<-paste("pot_dist",dd,"_n",nn,".tex",sep="")
```

```
    write.table(format(potencia,scientific=TRUE),file=escrever,col.
        names=FALSE,row.names=FALSE)

}
```

# Bibliografia


Anderson, T.W., Darling D.A., 1954. A Test of Goodness of Fit. *Journal of the American Statistical Association* 49, 268, 765–769.

D'Agostino, R.B., Stephens, M.A., 1986. *Goodness-of-Fit Techniques.* New York: Marcel Dekker.

Epps, T.W., Pulley L.B., 1983. A test for normality based on the empirical characteristic function. *Biometrika* 70, 3, 723–726.

Fromont, M., Laurent, B., 2006. Adaptative goodness-of-fit tests in a density model. *Ann. Statist.* 34, 680–720.

Gregory, G.G., 1977. Large sample theory for U-statistics and test of fit. *Ann. Statist.* 5, 110–123.

Henze, N., 1997. Extreme smoothing and testing for multivariate normality. *Comm. Stat. Theory Methods* 19, 3595–3617.

Henze, N., Wagner T., 1997. A new approach to the BHEP tests for multivariate normality. *Journal of Multivariate Analysis* 62, 1–23.

Henze, N., Zirkler,B., 1990. A class of invariante consistent tests for multivariate normality. *Statistics & Probability Letters* 35, 203–213.

Johnson, M.E., Tietjen, G.L., Beckman, R.J., 1980. A new family of probability distributions with applications to Monte Carlo studies. *J. Amer. Statist. Assoc.* 75, 276—279.

Mardia, K.V., 1970. Measures of Multivariate Skewness and Kurtosis with Applications. *Biometrika* 57, 3, 519–530.

Noughabi, H.A., Arghami, N.R., 2011. Monte Carlo comparison of seven normality tests. *Journal of Statistical Computation and Simulation* 81, 8, 965–972.

Razali, N.M., Wah, Y.B., 2011. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics* 2, 1, 21–33.

Serfling, R.J., 1980. *Approximation theorems of mathematical statistics.* New York: John Wiley.

Shapiro, S.S., Wilk, M.B., 1965. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* 52, 3/4, 591–611.

Shorack, G.R., Wellner, J.A., 1986. *Empirical Processes with Applications to Statistics.* New York: John Wiley.

Tenreiro, C., 2009. On the choice of the smoothing parameter for the BHEP goodness-of-fit test. *Computational Statistics and Data Analysis* 53, 1038–1053.

Tenreiro, C., 2011. An affine invariant multiple test procedure for assessing multivariate normality. *Computational Statistics and Data Analysis* 55, 1980–1992.

Thode, Jr., H.C., 2002. *Testing for normality.* New York: Marcel Dekker.

Yap, B.W., Sim, C.H., 2011. Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation* 81, 12, 2141–2155.