UNIVERSIDADE D
COIMBRA

Mariana Gama Mendes Lopes

UM TESTE DE AJUSTAMENTO A UMA FAMÍLIA DE DISTRIBUIÇÕES DE LOCALIZAÇÃO E ESCALA BASEADO NA DISTÂNCIA DE ENERGIA

Dissertação no âmbito do Mestrado em Matemática, Ramo de Estatística, Otimização e Matemática Financeira orientada pelo Professor Doutor Carlos Manuel Rebelo Tenreiro da Cruz e apresentada ao Departamento de Matemática da Faculdade de Ciências e Tecnologia.

Junho de 2023

Um teste de ajustamento a uma família de distribuições de localização e escala baseado na distância de energia

Mariana Gama Mendes Lopes

UNIVERSIDADE D
COIMBRA

Mestrado em Matemática

Master in Mathematics

Dissertação de Mestrado | MSc Dissertation

Junho 2023 | June 2023

Agradecimentos

Ao meu orientador, Professor Doutor Carlos Tenreiro, agradeço por todos os conhecimentos transmitidos, pelas opiniões e críticas, pelo apoio incansável, paciência e disponibilidade constante que contribuíram muito para a realização desta dissertação.

À minha família, em particular aos meus Pais, agradeço por estarem sempre presentes e pelo incentivo e preocupação constantes. À minha irmã, Carolina, por ser desde sempre um exemplo para mim e desde cedo me ter cultivado o gosto pela Matemática.

Ao Diogo, por todo o amor, companheirismo e ajuda constante. Sem ele o meu caminho teria sido muito mais difícil.

A todos os meus amigos que acompanharam o meu percurso académico, em particular, à Beatriz, pela amizade e por estar sempre presente.

Resumo

Em 2005, Székely e Rizzo propuseram um teste de ajustamento baseado numa distância estatística entre distribuições de vetores aleatórios a que chamam de distância de energia. Nesta dissertação estudamos este teste focando a nossa atenção no caso univariado. Começamos por fazer uma breve introdução ao conceito de distância de energia e estudar as suas principais propriedades, seguindo-se a respetiva demonstração. Posteriormente, obtemos um estimador da distância de energia entre duas variáveis aleatórias reais, o qual irá admitir duas representações alternativas, tendo em conta as propriedades da distância de energia, o que nos permite desenvolver o estudo do teste de ajustamento baseado na distância de energia. Como ponto de partida, analisamos o caso particular do teste de ajustamento a uma distribuição fixa, generalizando de seguida a uma família de distribuições de localização e escala. Tanto nos casos em que a hipótese nula é simples como composta, definimos a estatística de teste envolvente, com base no estimador da distância de energia anteriormente determinado, e estabelecemos a sua distribuição assintótica sob a hipótese nula. Além disso, estudamos o nível de significância do teste bem como a convergência do mesmo. Para finalizar, a potência empírica do teste em estudo será avaliada através de um estudo de simulação tanto no caso do teste de ajustamento a uma distribuição fixa à partida como no teste de ajustamento a uma família de distribuições de localização e escala.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Os testes de ajustamento são procedimentos estatísticos que nos permitem averiguar, com base na observação de realizações de variáveis $X_1, \ldots, X_n$, que supomos serem cópias independentes de uma variável aleatória real (v.a.r.) X com função de distribuição desconhecida F, se a distribuição F pertence a uma família de funções de distribuição fixa à partida. Mais precisamente, denotando por $\mathscr{F}_0$ uma tal família de funções de distribuição, pretendemos testar a hipótese

$$H_0 : F \in \mathscr{F}_0,$$

contra a hipótese alternativa

$$H_a : F \notin \mathscr{F}_0.$$

Para diferentes famílias de distribuições $\mathscr{F}_0$ são vários os testes de ajustamento existentes na literatura, sendo os mais conhecidos o teste clássico do qui-quadrado proposto por Karl Pearson em 1900 (Moore, 1986) e os testes de Kolmogorov-Smirnov, Cramer-von Mises e Anderson-Darling, estes últimos baseados na função de distribuição empírica associada a $X_1, \ldots, X_n$ (Stephens, 1986).

Com o intuito inicial de testar o ajustamento a uma distribuição normal multivariada Székely e Rizzo (2005) propõem um novo teste de ajustamento baseado numa distância estatística entre distribuições de vetores aleatórios a que chamam distância de energia pela sua relação com a noção de energia potencial de Newton (Székely e Rizzo, 2013, p. 1267). Apesar de estarem principalmente interessados no teste duma hipótese univariada ou multivariada de normalidade, que continuam a estudar em Móri et al. (2021), estes autores referem que testes de ajustamento baseados na distância de energia podem ser desenvolvidos para famílias de distribuições $\mathscr{F}_0$ não necessariamente normais, sendo esta a motivação inicial para o presente trabalho. Nesta dissertação centramos a nossa atenção no caso das observações univariadas, em que abordamos as situações em que H_0 é uma hipótese simples, onde estudamos o teste de ajustamento a uma distribuição fixa à partida, e em que H_0 é uma hipótese composta, onde estudamos o teste de ajustamento a uma família de distribuições de localização e escala.

Uma vez que a estatística de teste é baseada na distância de energia, no Capítulo 2 faremos uma breve introdução a este conceito, bem como a apresentação e respetiva demonstração de algumas das suas propriedades. Além disso, deduzimos um estimador da distância de energia entre duas variáveis aleatórias reais, onde, atendendo às propriedades enunciadas, concluímos que este admite duas representações

alternativas: uma baseada na distância de energia e outra baseada no desvio quadrático entre a função de distribuição empírica associada à amostra $X_1,\ldots,X_n$ e a função de distribuição comum a estas variáveis.

Apesar do teste proposto baseado na distância de energia ter sido introduzido para um problema de ajustamento a uma família de distribuições, achamos oportuno estudar, também, a situação em que a hipótese nula é simples, isto é, $\mathscr{F}_0 = \{F_0\}$, onde F_0 é uma função de distribuição fixa à partida. Assim, no Capítulo 3 estamos interessados no teste das hipóteses

$$H_0 : F = F_0 \quad vs. \quad H_a : F \neq F_0. \tag{1.1}$$

Neste capítulo determinamos, em primeiro lugar, a distribuição assintótica do estimador da distância de energia entre duas v.a.r., onde, com este resultado, definimos a estatística a utilizar no teste de energia (Teorema 3.1.1). De seguida, estudamos o nível de significância do teste (Teorema 3.2.1) e estabelecemos a convergência do mesmo (Teorema 3.3.1). Terminamos este capítulo apresentando um breve estudo de simulação com o objetivo de avaliar a potência empírica do teste de ajustamento em causa a uma distribuição normal standard. Tomaremos o teste de Anderson-Darling como referência, podendo assim comparar os resultados de potência deste teste com o teste de energia em análise.

O Capítulo 4 é o mais importante desta dissertação, onde nos debruçamos sobre o teste de ajustamento às famílias de distribuições da forma

$$\mathscr{F}_0 = \left\{G(\,\cdot\,;\theta_1,\theta_2) : \theta_1 \in \mathbb{R} \text{ e } \theta_2 \in \mathbb{R}^+\right\},$$

com

$$G(x;\theta_1,\theta_2) = F_0\Big(\frac{x-\theta_1}{\theta_2}\Big),$$

onde F_0 é uma função de distribuição conhecida e $(\theta_1,\theta_2) \in \mathbb{R}\times\mathbb{R}^+$ é o vetor dos parâmetros de localização e escala da família $\mathscr{F}_0$. Tal como no Capítulo 3, começamos por estudar a distribuição assintótica da estatística de teste sob a hipótese nula. Uma vez que os parâmetros θ_1 e θ_2 são desconhecidos, a necessidade de os estimar conduz a uma distribuição assintótica diferente da obtida sob a hipótese simples (Teorema 4.3.1). Sob condições adicionais sobre os estimadores considerados de θ_1 e θ_2, estabelecemos a invariância da estatística de teste sob a hipótese nula, o que permite concluir que as distribuições exata e assintótica da estatística de teste não dependem da distribuição F tomada em $\mathscr{F}_0$ (Teorema 4.4.1). A propriedade de invariância é importante na implementação do procedimento de teste, uma vez que os quantis da estatística de teste são calculados simulando valores da estatística de teste sob a hipótese nula que é agora uma hipótese composta. Terminamos este capítulo estudando o nível de significância do teste (Teorema 4.5.1) e estabelecendo a convergência do mesmo (Teorema 4.6.1).

O quinto e último capítulo é dedicado à apresentação de um estudo de simulação que tem como objetivo principal analisar a potência empírica do teste apresentado no quarto capítulo, comparando-a, mais uma vez, com a do teste de Anderson-Darling. Para tal, vamos concentrar-nos nos testes de ajustamento às famílias de distribuições normais e exponenciais. Este estudo, bem como o realizado no final do Capítulo 3, foi desenvolvido usando código escrito em linguagem R (R Core Team, 2021).

No decurso desta dissertação, o conceito de integrabilidade deve ser entendido no sentido de Lebesgue. Adotamos ainda as seguintes notações: $\xrightarrow{d}$, $\xrightarrow{p}$ e $\xrightarrow{q.c.}$ representam as convergências em distribuição, em probabilidade e quase certa, respetivamente, quando $n \to +\infty$; indicamos $X \sim Y$ quando X e Y possuem a mesma distribuição e $X \sim F$ quando X tem função de distribuição F; representamos por X' uma v.a.r.

independente de X com $X \sim X'$; finalmente, denotaremos por $o_P(1)$ uma sucessão de variáveis aleatórias reais que converge para zero em probabilidade e por $O_P(1)$ uma sucessão de variáveis aleatórias reais limitada em probabilidade (sobre as propriedades destes símbolos estocásticos seguimos van der Vaart, 2000, p. 12).

Capítulo 2

Distância de energia

Este capítulo será dedicado à exposição do conceito de distância de energia e à demonstração de propriedades a esta associada. Além disso, é deduzido um estimador da distância de energia entre duas variáveis aleatórias reais.

2.1 Definição e propriedades

Sendo X uma variável aleatória real (v.a.r.), denotamos por X' uma v.a.r. com a mesma distribuição de X, isto é, $X \sim X'$, e independente de X.

Definição 2.1.1. *Sendo X e Y v.a.r. independentes e integráveis, chamamos distância de energia entre X e Y a*

$$\mathscr{E}(X,Y) = 2E|X-Y| - E|X-X'| - E|Y-Y'|.$$

A propriedade seguinte relaciona a distância de energia $\mathscr{E}(X,Y)$ com a norma L_2 da diferença entre as funções de distribuição de X e Y. Além disso, a distância de energia sob a forma integral vai ter particular importância no Capítulo 4.

Proposição 2.1.1. *Se X e Y são v.a.r. independentes e integráveis com funções de distribuição F e G, respetivamente, então*

$$\mathscr{E}(X,Y) = 2\int_{-\infty}^{+\infty} \{F(t)-G(t)\}^2 \, dt.$$

Demonstração. Atendendo a que F e G são as funções de distribuição de X e Y, respetivamente, e que estas variáveis são independentes, para $t \in \mathbb{R}$ temos (Székely e Rizzo, 2005)

$$\begin{aligned}
&\{F(t)-G(t)\}^2 \\
&= F(t)^2 - 2F(t)G(t) + G(t)^2 \\
&= F(t) - F(t)G(t) + G(t) - F(t)G(t) - F(t) + F(t)^2 - G(t) + G(t)^2 \\
&= F(t)\,(1-G(t)) + G(t)\,(1-F(t)) - F(t)\,(1-F(t)) - G(t)\,(1-G(t)) \\
&= P(X \le t)P(Y > t) + P(Y \le t)P(X > t) - P(X \le t)P(X' > t) - P(Y \le t)P(Y' > t) \\
&= P(X \le t < Y) + P(Y \le t < X) - P(X \le t < X') - P(Y \le t < Y').
\end{aligned} \tag{2.1}$$

Uma vez que

$$P(X \le t < Y) = E\left(\mathbb{1}_{\{X \le t < Y\}}\right),$$

pelo Teorema de Fubini temos

$$\int_{-\infty}^{+\infty} P(X \leq t < Y)\ dt = E\left(\int_{-\infty}^{+\infty} \mathbb{1}_{\{X \leq t < Y\}}\ dt\right) = E\left(\int_{X}^{Y} dt\ \mathbb{1}_{\{X<Y\}}\right) = E\left((Y-X)\mathbb{1}_{\{X<Y\}}\right).$$

Procedendo da mesma forma com as restantes parcelas de (2.1) temos

$$\begin{aligned}
&\int_{-\infty}^{+\infty} \{F(t)-G(t)\}^2\ dt \\
&= E\left((Y-X)\mathbb{1}_{\{X<Y\}}\right) + E\left((X-Y)\mathbb{1}_{\{Y<X\}}\right) - E\left((X'-X)\mathbb{1}_{\{X<X'\}}\right) - E\left((Y'-Y)\mathbb{1}_{\{Y<Y'\}}\right) \\
&= E|X-Y| - E\left((X'-X)\mathbb{1}_{\{X<X'\}}\right) - E\left((Y'-Y)\mathbb{1}_{\{Y<Y'\}}\right) \\
&= E|X-Y| - \frac{1}{2}E|X-X'| - \frac{1}{2}E|Y-Y'| \\
&= \frac{1}{2}\mathscr{E}(X,Y),
\end{aligned}$$

pois

$$\begin{aligned}
E|X-X'| &= E\left(\left(X-X'\right)\mathbb{1}_{\{X'<X\}} + \left(X'-X\right)\mathbb{1}_{\{X<X'\}}\right) \\
&= E\left(\left(X-X'\right)\mathbb{1}_{\{X'<X\}}\right) + E\left(\left(X'-X\right)\mathbb{1}_{\{X<X'\}}\right) \\
&= 2E\left(\left(X'-X\right)\mathbb{1}_{\{X<X'\}}\right),
\end{aligned}$$

uma vez que $X \sim X'$ e X e X' são independentes. □

Decorre diretamente da definição de $\mathscr{E}(X,Y)$ que esta é simétrica, isto é, $\mathscr{E}(X,Y) = \mathscr{E}(Y,X)$. Provamos a seguir que, além de não negativa, a distância de energia entre duas variáveis com distribuições distintas é sempre estritamente positiva.

Corolário 2.1.1. *Nas condições da Proposição 2.1.1, $\mathscr{E}(X,Y) \geq 0$, e*

$$\mathscr{E}(X,Y) = 0 \qquad sse \qquad X \sim Y.$$

Demonstração. Da Proposição 2.1.1 vem $\mathscr{E}(X,Y) \geq 0$ e se $X \sim Y$, então $F = G$ e, portanto, $\mathscr{E}(X,Y) = 0$.

Suponhamos agora que $\mathscr{E}(X,Y) = 0$, o que implica que $\int_{-\infty}^{+\infty} (F(t)-G(t))^2\ dt = 0$. Decorre desta igualdade que se t é um ponto de continuidade de $F-G$ então $F(t) = G(t)$. Uma vez que o conjunto dos pontos de continuidade de $F-G$ é denso em $\mathbb{R}$ e $F-G$ é contínua à direita, podemos concluir que $F(t) = G(t)$ para todo o $t \in \mathbb{R}$, ou seja, $X \sim Y$.

□

Apesar de chamarmos distância a $\mathscr{E}(X,Y)$, a aplicação $\mathscr{E}$ não verifica a desigualdade triangular. Como provamos a seguir, tal propriedade é satisfeita por $\mathscr{E}^{1/2}$. Uma vez que $\mathscr{E}^{1/2}$ é simétrica e também satisfaz a propriedade expressa no Corolário 2.1.1, $\mathscr{E}^{1/2}$ é efetivamente uma distância definida no conjunto das v.a.r. independentes e integráveis.

Corolário 2.1.2. *Se X, Y e Z são v.a.r. independentes e integráveis então*

$$\mathscr{E}^{1/2}(X,Z) \leq \mathscr{E}^{1/2}(X,Y) + \mathscr{E}^{1/2}(Y,Z).$$

Demonstração. Sendo F, G e H as funções de distribuição de X, Y e Z, respetivamente, pela desigualdade de Minkowski temos

$$\begin{aligned}\mathscr{E}^{1/2}(X,Z) &= \sqrt{2}\left(\int_{-\infty}^{+\infty}\{F(t)-H(t)\}^2\,dt\right)^{1/2}\\ &= \sqrt{2}\left(\int_{-\infty}^{+\infty}\{(F(t)-G(t))+(G(t)-H(t))\}^2\,dt\right)^{1/2}\\ &\leq \sqrt{2}\left(\int_{-\infty}^{+\infty}\{F(t)-G(t)\}^2\,dt\right)^{1/2}+\sqrt{2}\left(\int_{-\infty}^{+\infty}\{G(t)-H(t)\}^2\,dt\right)^{1/2}\\ &= \mathscr{E}^{1/2}(X,Y)+\mathscr{E}^{1/2}(Y,Z).\end{aligned}$$

□

2.2 Estimação da distância de energia

Sendo X_0 uma v.a.r. independente de X, com função de distribuição F_0, isto é, $X_0 \sim F_0$, e X e X_0 integráveis, decorre do Corolário 2.1.1 que o problema de teste (1.1) é equivalente ao problema do teste das hipóteses

$$H_0 : \mathscr{E}(X,X_0) = 0 \quad vs. \quad H_a : \mathscr{E}(X,X_0) \neq 0.$$

Pretendendo desenvolver um procedimento estatístico para testar as hipóteses anteriores, será natural basear tal procedimento num estimador da distância de energia $\mathscr{E}(X,X_0)$. Atendendo a que

$$\mathscr{E}(X,X_0) = 2E|X-X_0| - E|X-X'| - E|X_0-X_0'|,$$

seguindo a abordagem de Móri et al. (2021), vamos estimar $\mathscr{E}(X,X_0)$ através da estatística $\mathscr{E}_n(X_1,\ldots,X_n)$, onde

$$\mathscr{E}_n(x_1,\ldots,x_n) = \frac{2}{n}\sum_{i=1}^{n} E|x_i-X_0| - \frac{1}{n^2}\sum_{i,j=1}^{n}|x_i-x_j| - E|X_0-X_0'|, \tag{2.2}$$

para $x_1,\ldots,x_n \in \mathbb{R}$.

Reparemos que $\mathscr{E}_n(X_1,\ldots,X_n)$ tem a estrutura de uma V-estatística associada ao núcleo simétrico, definido para $x,y \in \mathbb{R}$, por

$$h(x,y) = E|x-X_0| + E|y-X_0| - |x-y| - E|X_0-X_0'|, \tag{2.3}$$

uma vez que,

$$\mathscr{E}_n(X_1,\ldots,X_n) = \frac{1}{n^2}\sum_{i,j=1}^{n} h(X_i,X_j).$$

Proposição 2.2.1. *Sendo $X_1,\ldots,X_n$ v.a.r. integráveis, independentes e identicamente distribuídas (i.i.d.) com X, então*

$$\mathscr{E}_n(X_1,\ldots,X_n) \xrightarrow{q.c.} \mathscr{E}(X,X_0).$$

Demonstração. Atendendo a que

$$\mathscr{E}_n(X_1,\ldots,X_n) = \frac{1}{n}\frac{1}{n}\sum_{i=1}^{n} h(X_i,X_i) + \frac{n-1}{n}\frac{2}{n(n-1)}\sum_{1\leq i<j\leq n} h(X_i,X_j),$$

onde

$$|h(x,y)| \leq 2(|x|+|y|+2E|X_0|),$$

concluímos que $E|h(X,X)| \leq 4\,(E|X|+E|X_0|) < \infty$ e $E|h(X,X')| \leq 2\,(E|X|+E|X'|+2E|X_0|) < \infty$, pois X, X' e X_0 são integráveis. Assim, como $E|h(X,X)| < \infty$, $E|h(X,X')| < \infty$, pela lei dos grandes números de Kolmogorov e pelo Teorema A.1.1 temos

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i,X_i) \xrightarrow{q.c.} E(h(X,X))$$

e

$$\frac{2}{n(n-1)}\sum_{1\leq i<j\leq n} h(X_i,X_j) \xrightarrow{q.c.} E(h(X,X')),$$

e portanto

$$\mathscr{E}_n(X_1,\ldots,X_n) \xrightarrow{q.c.} E(h(X,X')).$$

Para concluir a demonstração basta notar que

$$\begin{aligned}
E(h(X,X')) &= \iint_{\mathbb{R}^2} h(x,y)\; dF(x)dF(y) \\
&= \iint_{\mathbb{R}^2} |x-x_0|\; dF_0(x_0)dF(x) + \iint_{\mathbb{R}^2} |y-x_0|\; dF_0(x_0)dF(y) - E|X-X'| - E|X_0-X_0'| \\
&= 2\iint_{\mathbb{R}^2} |x-x_0|\; dF_0(x_0)dF(x) - E|X-X'| - E|X_0-X_0'| \\
&= 2E|X-X_0| - E|X-X'| - E|X_0-X_0'| \\
&= \mathscr{E}(X,X_0).
\end{aligned}$$

□

Como decorre da demonstração anterior, um estimador alternativo da distância de energia $\mathscr{E}(X,X_0)$ é dado por

$$\mathscr{E}_n'(X_1,\ldots,X_n) = \frac{2}{n(n-1)}\sum_{1\leq i<j\leq n} h(X_i,X_j).$$

Contrariamente ao estimador $\mathscr{E}_n$, $\mathscr{E}_n'$ é estimador cêntrico de $\mathscr{E}(X,X_0)$.

Notemos que podemos, ainda, estimar $\mathscr{E}(X,X_0)$ de outra forma. Uma vez que F pode ser estimada pela função de distribuição empírica associada à amostra $X_1,\ldots,X_n$ definida por

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

de acordo com a Proposição 2.1.1, um estimador natural de $\mathscr{E}(X,X_0)$ é dado por

$$\mathscr{E}_n''(X_1,\ldots,X_n) = 2\int_{\mathbb{R}} \{F_n(t)-F_0(t)\}^2\; dt. \tag{2.4}$$

Como mostramos a seguir este estimador coincide com $\mathscr{E}_n(X_1,\ldots,X_n)$. Com efeito, para $x_1,\ldots,x_n \in \mathbb{R}$, sejam F_n a função de distribuição associada a $x_1,\ldots,x_n$ e Y_n a v.a.r. com função de distribuição F_n. Assim, pela Proposição 2.1.1 temos

$$\begin{aligned}
\mathscr{E}_n''(x_1,\ldots,x_n) &= \mathscr{E}(Y_n,X_0)\\
&= 2E|Y_n - X_0| - E|Y_n - Y_n'| - E|X_0 - X_0'|\\
&= 2\iint_{\mathbb{R}^2} |x-y|\, dF_n(x)dF_0(y) - \iint_{\mathbb{R}^2} |x-y|\, dF_n(x)dF_n(y) - E|X_0 - X_0'|\\
&= 2\int_{\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} |x_i - y|\, dF_0(y) - \int_{\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} |x_i - y|\, dF_n(y) - |X_0 - X_0'|\\
&= \frac{2}{n}\sum_{i=1}^{n} E|x_i - X_0| - \frac{1}{n^2}\sum_{i,j=1}^{n} |x_i - x_j| - |X_0 - X_0'|\\
&= \mathscr{E}_n(x_1,\ldots,x_n).
\end{aligned}$$

Assim,

$$\mathscr{E}_n(X_1,\ldots,X_n) = 2\int_{\mathbb{R}} \{F_n(t) - F_0(t)\}^2\, dt, \tag{2.5}$$

de onde se conclui que o núcleo (2.3) admite a representação

$$h(x,y) = 2\int_{\mathbb{R}} \{\mathbb{1}_{]-\infty,t]}(x) - F_0(t)\}\{\mathbb{1}_{]-\infty,t]}(y) - F_0(t)\}\, dt. \tag{2.6}$$

Capítulo 3

Teste de ajustamento a uma distribuição fixa

No presente capítulo, focamos o nosso estudo no desenvolvimento de um teste de ajustamento a uma distribuição fixa à partida baseado na estatística $\mathscr{E}_n(X_1,\ldots,X_n)$. Em particular, determinamos a distribuição assintótica da estatística de teste a ser utilizada e estudamos, ainda, o nível de significância do teste e a convergência do mesmo.

3.1 Distribuição assintótica de $\mathscr{E}_n$

Pretendendo desenvolver um teste de ajustamento para testar

$$H_0 : F = F_0 \quad \text{contra} \quad H_a : F \neq F_0,$$

a um nível de significância α, $\alpha \in]0,1[$, baseado na estatística

$$\mathscr{E}_n(X_1,\ldots,X_n) = \frac{1}{n^2}\sum_{i,j=1}^{n} h(X_i,X_j), \tag{3.1}$$

onde

$$h(x,y) = E|x-X_0| + E|y-X_0| - |x-y| - E|X_0 - X_0'|, \tag{3.2}$$

para $x,y \in \mathbb{R}$, começamos por determinar a distribuição assintótica de $\mathscr{E}_n(X_1,\ldots,X_n)$ sob a hipótese nula.

Teorema 3.1.1. *Sendo $X_1,\ldots,X_n$ v.a.r. de quadrado integrável e i.i.d. com X_0, onde $X \sim F_0$, então*

$$n\mathscr{E}_n(X_1,\ldots,X_n) \xrightarrow{d} E(h(X_0,X_0)) + \sum_{k=1}^{+\infty} \lambda_k (Z_k^2 - 1),$$

onde Z_k, $k \geq 1$, são v.a.r. i.i.d. com lei normal standard e $\lambda_k \geq 0$, $k \geq 1$, com $\sum_{k=1}^{n} \lambda_k^2 < \infty$, são os valores próprios associados ao operador A_h definido de $L^2(\mathbb{R},F_0)$ em $L^2(\mathbb{R},F_0)$ por

$$A_h g(x) = \int_{\mathbb{R}} h(x,y)g(y)\, dF_0(y),\ x \in \mathbb{R}, \tag{3.3}$$

para $g \in L^2(\mathbb{R}, F_0)$, onde $L^2(\mathbb{R}, F_0)$ é o espaço das funções reais de variável real g tais que $\int_{\mathbb{R}} g(y)^2\, dF_0(y) < +\infty$.

Demonstração. Da demonstração da Proposição 2.2.1 sabemos que

$$n\mathscr{E}_n(X_1,\ldots,X_n) = \frac{1}{n}\sum_{i=1}^{n} h(X_i,X_i) + (n-1)U_n$$

onde

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i,X_i) \xrightarrow{q.c.} E(h(X_0,X_0)),$$

e U_n é a U-estatística definida por

$$U_n = \frac{2}{n(n-1)} \sum_{1\le i<j\le n} h(X_i,X_j).$$

Para estabelecer o resultado enunciado, basta agora estudar a convergência em distribuição de nU_n.

Com o intuito de aplicar o Teorema A.1.2 sobre a convergência em distribuição de uma U-estatística, vamos verificar que o núcleo simétrico definido em (3.2) satisfaz as condições $E(h(X_0,X_0')^2) < +\infty$ e $E(h(x,X_0)) = 0$, para todo o $x \in \mathbb{R}$.

Tendo em conta que $(a+b)^2 \le 2(a^2+b^2)$, para todo o $a,b \in \mathbb{R}$, e da demonstração da Proposição 2.2.1 temos

$$\begin{aligned} |h(x,y)|^2 &\le 4\left(2\left(|x|+|y|\right)^2 + 2\left(2E|X_0|\right)^2\right) \\ &\le 4\left(4x^2+4y^2+8\left(E|X_0|\right)^2\right) \\ &= 16\left(x^2+y^2+2\left(E|X_0|\right)^2\right). \end{aligned}$$

Logo

$$E(h(X_0,X_0')^2) \le 32\left(E(X_0^2) + (E|X_0|)^2\right) < +\infty,$$

uma vez que, por hipótese, X_0 é de quadrado integrável.

Por outro lado,

$$\begin{aligned} E(h(x,X_0)) &= \int_{\mathbb{R}} h(x,y)\, dF_0(y) \\ &= \int_{\mathbb{R}^2} |x-x_0|\, dF_0(x_0)dF_0(y) + \int_{\mathbb{R}^2} |y-x_0|\, dF_0(x_0)dF_0(y) \\ &\quad - \int_{\mathbb{R}} |x-y|\, dF_0(y) - \int_{\mathbb{R}} E|X_0-X_0'|\, dF_0(y) \\ &= \int_{\mathbb{R}} |x-x_0|\, dF_0(x_0) + E|X_0-X_0'| - \int_{\mathbb{R}} |x-y|\, dF_0(y) - E|X_0-X_0'| \\ &= 0. \end{aligned}$$

Verificadas as condições anteriores podemos concluir, pelo Teorema A.1.2, que

$$nU_n \xrightarrow{d} \sum_{k=1}^{+\infty} \lambda_k(Z_k^2-1),$$

onde Z_k, $k \geq 1$, são v.a.r. i.i.d. com lei normal standard e λ_k, $k \geq 1$, são os valores próprios associados ao operador A_h definido em (3.3).

Para terminar a demonstração, mostramos que $\lambda_k \geq 0$, para todo o $k \geq 1$.

Usando a forma integral (2.6) para h temos, para todo o $g \in L^2(\mathbb{R}, F_0)$,

$$\begin{aligned}
&\int_{\mathbb{R}} g(x)(A_h g)(x)\, dF_0(x) \\
&= \iint_{\mathbb{R}^2} g(x)h(x,y)g(y)\, dF_0(y)dF_0(x) \\
&= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} g(x)(\mathbb{1}_{]-\infty,t]}(x) - F_0(t))\, dF_0(x) \int_{\mathbb{R}} g(y)(\mathbb{1}_{]-\infty,t]}(y) - F_0(t))\, dF_0(y) \right) dt \\
&= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} g(x)(\mathbb{1}_{]-\infty,t]}(x) - F_0(t))\, dF_0(x) \right)^2 dt \geq 0,
\end{aligned}$$

o que, de acordo com o Teorema A.1.2, conclui a demonstração. □

3.2 Região crítica do teste e nível de significância

Atendendo ao Teorema 3.1.1 e tomando $T_n = n\mathscr{E}_n(X_1, \ldots, X_n)$ como a estatística de teste, onde $\mathscr{E}_n$ é definido por (3.1), vamos considerar o teste de região crítica dada por

$$\{T_n > c_n(\alpha)\}, \tag{3.4}$$

onde, para $\alpha \in]0,1[$, $c_n(\alpha) = F^{-1}_{T_n}(1-\alpha)$, em que F^{-1} denota a função quantil de F definida por

$$F^{-1}(t) = \inf\{x \in \mathbb{R} : F(x) \geq t\}, \quad \text{para} \quad 0 < t < 1.$$

Uma vez que o quantil $c_n(\alpha)$ não é conhecido, na implementação prática do procedimento de teste anterior, $c_n(\alpha)$ será estimado pelo método de Monte Carlo simulando valores da estatística de teste sob a hipótese nula. Voltaremos a este ponto mais à frente.

De seguida apresentamos resultados importantes cujas demonstrações se podem encontrar em van der Vaart (2000, pp. 12, 304, 305) e em Shorack e Wellner (1986, p. 8).

Proposição 3.2.1. *Sendo F^{-1} a função quantil da função de distribuição F temos*

(a) $F(F^{-1}(t)) \geq t$, para todo o $t \in]0,1[$ e $F(F^{-1}(t)) = t$ se F é contínua em $F^{-1}(t)$;

(b) F é contínua se e só se F^{-1} é estritamente crescente e F é estritamente crescente (em $]0,1[$) se e só se F^{-1} é contínua.

Proposição 3.2.2. *Para qualquer sucessão (X_n) de v.a.r. temos*

(a) $X_n \xrightarrow{d} X$ se e só se $F^{-1}_{X_n}(t) \longrightarrow F^{-1}_X(t)$, para todo o ponto t de continuidade de F^{-1}_X;

(b) Se $X_n \xrightarrow{d} X$ e F_X é contínua então

$$\sup_{x \in \mathbb{R}} |F_{X_n}(x) - F_X(x)| \longrightarrow 0.$$

O resultado seguinte relativo a propriedades da função de distribuição da distribuição assintótica da estatística de teste obtida no Teorema 3.1.1 será importante no estudo do nível de significância e da convergência do teste (3.4).

Lema 3.2.1. *Se*

$$Y_\infty = \sum_{k=1}^{+\infty} \lambda_k (Z_k^2 - 1),$$

onde Z_k, $k \geq 1$, são v.a.r. i.i.d. com lei normal standard e $\lambda_k \geq 0$, $k \geq 1$, são tais que $\sum_{k=1}^{+\infty} \lambda_k^2 < \infty$, então F_{Y_∞} é contínua e estritamente crescente (em $]-\sum_{k=1}^{+\infty} \lambda_k, +\infty[$).

Demonstração. Vamos assumir, sem perda de generalidade, que $\lambda_k > 0$, para $k \geq 1$. Começamos por mostrar que Y_∞ está bem definida. Atendendo ao critério de convergência quase certa (Métivier, 1972, p. 201), uma vez que as v.a.r. $\lambda_k(Z_k^2 - 1)$ são independentes com média zero e

$$\begin{aligned} \sum_{k=1}^{+\infty} E(\lambda_k (Z_k^2 - 1))^2 &= E(Z^2 - 1)^2 \sum_{k=1}^{+\infty} \lambda_k^2 \\ &= 2 \sum_{k=1}^{+\infty} \lambda_k^2 < \infty, \end{aligned}$$

concluímos que a série $\sum_{k=1}^{+\infty} \lambda_k (Z_k^2 - 1)$ converge quase certamente.

De seguida provamos que F_{Y_∞} é contínua. Para tal vamos provar que a função característica ϕ_{Y_∞} é integrável (à Lebesgue), o que, de acordo com o teorema de inversão de uma função característica (Métivier, 1972, p. 168), podemos concluir que F_{Y_∞} é absolutamente contínua.

Atendendo a que $Y_N \xrightarrow{d} Y_\infty$, onde $Y_N = \sum_{k=1}^{N} \lambda_k (Z_k^2 - 1)$, pelo teorema da continuidade de Lévy (Métivier, 1972, p. 178), $|\phi_{Y_N}(t)| \longrightarrow |\phi_{Y_\infty}(t)|$, para todo o $t \in \mathbb{R}$. Para $N \geq 3$ temos

$$\begin{aligned} |\phi_{Y_N}(t)| &= \left| \prod_{k=1}^{N} \phi_{Z_k^2 - 1}(\lambda_k t) \right| = \prod_{k=1}^{N} \left| \phi_{Z_k^2}(\lambda_k t) \right| \\ &\leq \prod_{k=1}^{3} (1 + 4t^2 \lambda_k^2)^{-1/4} \leq (1 + 4t^2 \lambda^2)^{-3/4}, \end{aligned}$$

onde $t \mapsto (1 + 4t^2\lambda^2)^{-3/4}$, com $\lambda = \min(\lambda_1, \lambda_2, \lambda_3)$, é integrável. Pelo teorema da convergência dominada de Lebesgue (Métivier, 1972, p. 71) concluímos que $|\phi_{Y_\infty}(t)|$ é integrável.

Resta provar que F_{Y_∞} é estritamente crescente (em $]-\sum_{k=1}^{+\infty} \lambda_k, +\infty[$). Sejam $a < b$ neste intervalo e consideremos $\varepsilon > 0$ tal que $a + \varepsilon < b - \varepsilon$. Para $N \in \mathbb{N}$ e $R_N = Y_\infty - Y_N$ temos

$$\begin{aligned} P(Y_\infty \in]a, b]) &= P(Y_N + R_N \in]a, b]) \\ &\geq P(Y_N + R_N \in]a, b], |R_N| < \varepsilon) \\ &\geq P(Y_N \in]a + \varepsilon, b - \varepsilon], |R_N| < \varepsilon). \end{aligned}$$

Usando agora a independência de Y_N e R_N concluímos que

$$P(Y_\infty \in]a, b]) = (F_{Y_N}(b - \varepsilon) - F_{Y_N}(a + \varepsilon)) P(|R_N < \varepsilon|).$$

Uma vez que F_{Y_N} é estritamente crescente em $]-\sum_{k=1}^{N}\lambda_k,+\infty[$ e $R_N \xrightarrow{p} 0$, então escolhendo $N \in \mathbb{N}$ tal que

$$-\sum_{k=1}^{N}\lambda_k < a+\varepsilon < b-\varepsilon$$

e

$$P(|R_N| < \varepsilon) > 0,$$

concluímos que

$$P(Y_\infty \in]a,b]) > 0,$$

o que termina a demonstração. □

Teorema 3.2.1. *Nas condições do Teorema 3.1.1, o teste de região crítica (3.4) é um teste com nível de significância inferior ou igual a α, isto é,*

$$P_{F_0}(T_n > c_n(\alpha)) \leq \alpha,$$

onde P_{F_0} representa a probabilidade calculada sob a hipótese H_0. Além disso, o teste é de nível assintótico α, ou seja,

$$\lim_{n\to+\infty} P_{F_0}(T_n > c_n(\alpha)) = \alpha.$$

Demonstração. Tendo em conta que

$$P_{F_0}(T_n > c_n(\alpha)) = 1 - P_{F_0}(T_n \leq c_n(\alpha)) = 1 - F_{T_n}(c_n(\alpha)) = 1 - F_{T_n}(F_{T_n}^{-1}(1-\alpha)),$$

pela Proposição 3.2.1 (a) temos

$$P_{F_0}(T_n > c_n(\alpha)) \leq 1 - (1-\alpha) = \alpha.$$

Para concluir a demonstração, tomando $T = E(h(X_0,X_0)) + \sum_{k=1}^{+\infty}\lambda_k(Z_k^2-1)$, resta provar que

$$F_{T_n}(F_{T_n}^{-1}(1-\alpha)) \longrightarrow F_T(F_T^{-1}(1-\alpha)), \quad n \longrightarrow +\infty, \tag{3.5}$$

uma vez que, pelo Lema 3.2.1, F_T contínua em $\mathbb{R}$, então $F_T(F_T^{-1}(1-\alpha)) = 1-\alpha$, pela Proposição 3.2.1 (a).

Ora,

$$\begin{aligned}
&\left|F_{T_n}(F_{T_n}^{-1}(1-\alpha)) - F_T(F_T^{-1}(1-\alpha))\right| \\
=\ &\left|F_{T_n}(F_{T_n}^{-1}(1-\alpha)) - F_T(F_{T_n}^{-1}(1-\alpha)) + F_T(F_{T_n}^{-1}(1-\alpha)) - F_T(F_T^{-1}(1-\alpha))\right| \\
\leq\ &\left|F_{T_n}(F_{T_n}^{-1}(1-\alpha)) - F_T(F_{T_n}^{-1}(1-\alpha))\right| + \left|F_T(F_{T_n}^{-1}(1-\alpha)) - F_T(F_T^{-1}(1-\alpha))\right| \\
\leq\ &\sup_{x\in\mathbb{R}}|F_{T_n}(x) - F_T(x)| + \left|F_T(F_{T_n}^{-1}(1-\alpha)) - F_T(F_T^{-1}(1-\alpha))\right|.
\end{aligned} \tag{3.6}$$

Sabendo pelo Teorema 3.1.1 que $T_n \xrightarrow{d} T$ sob H_0 e que F_T é contínua em $\mathbb{R}$, então, pela Proposição 3.2.2 (b) ,

$$\sup_{x\in\mathbb{R}} |F_{T_n}(x) - F_T(x)| \longrightarrow 0, \quad n \to +\infty. \tag{3.7}$$

Por outro lado, como, pelo Lema 3.2.1, F_T é estritamente crescente, então F_T^{-1} é contínua em $]0,1[$, pela Proposição 3.2.1 (b). Deste modo, pela Proposição 3.2.2 (a) temos que

$$F_{T_n}^{-1}(1-\alpha) \longrightarrow F_T^{-1}(1-\alpha), \quad n \longrightarrow +\infty$$

e portanto

$$F_T(F_{T_n}^{-1}(1-\alpha)) \longrightarrow F_T(F_T^{-1}(1-\alpha)), \quad n \longrightarrow +\infty. \tag{3.8}$$

Isto conclui a demonstração, bastando usar (3.6), (3.7) e (3.8) para obter (3.5). □

3.3 Convergência do teste

De seguida, passamos ao estudo da convergência do teste em estudo cuja região crítica foi definida em (3.4). Vamos denotar por $\mathscr{F}$ o conjunto das funções de distribuição F tais que $\int_{\mathbb{R}} |x|\, dF(x) < \infty$.

Teorema 3.3.1. *O teste de região crítica (3.4) é convergente para qualquer distribuição alternativa em $\mathscr{F}$, isto é,*

$$\lim_{n\longrightarrow+\infty} P_F(T_n > c_n(\alpha)) = 1, \text{ para todo o } F \in \mathscr{F} \setminus \{F_0\}.$$

Demonstração. Sejam $X_1,\ldots,X_n$ v.a.r. i.i.d. com $X \sim F \in \mathscr{F} \setminus \{F_0\}$. Pela Proposição 2.2.1 temos que

$$\mathscr{E}_n(X_1,\ldots,X_n) \xrightarrow{q.c.} \mathscr{E}(X,X_0),$$

onde, pelo Corolário 2.1.1, $\mathscr{E}(X,X_0) > 0$, uma vez que, por hipótese, X não é identicamente distribuído com X_0. Assim,

$$\frac{T_n}{n} = \mathscr{E}_n(X_1,\ldots,X_n) \xrightarrow{q.c.} \mathscr{E}(X,X_0) > 0,$$

e portanto

$$T_n \xrightarrow{q.c.} +\infty.$$

Isto conclui a demonstração uma vez que, como vimos na demonstração do Teorema 3.2.1, $c_n(\alpha) = F_{T_n}^{-1}(1-\alpha) \longrightarrow F_T^{-1}(1-\alpha)$, quando $n \longrightarrow +\infty$.

□

3.4 Estudo de simulação para a hipótese simples de normalidade

Terminamos este capítulo com um estudo de simulação de carácter exploratório que tem como objetivo avaliar a potência do teste de energia para diferentes distribuições alternativas e tamanhos de amostras. Iremos comparar os resultados de potência do teste de energia com os teste Anderson-Darling, tomado, assim, como teste de referência.

3.4.1 Estatística de teste

Neste estudo, pretendemos testar

$$H_0 : F = F_0 \quad vs. \quad H_a : F \neq F_0,$$

onde F_0 é a função de distribuição da lei $\mathcal{N}(\mu, \sigma^2)$, com μ e σ fixos à partida.

Como decorre do resultado seguinte, no caso do teste para uma hipótese simples de normalidade, o cálculo da estatística de teste $T_n = n\mathscr{E}_n(X_1, \ldots, X_n)$, onde $\mathscr{E}_n(X_1, \ldots, X_n)$ é definido em (2.2), não levanta problemas de maior uma vez que os termos $E|x - X_0|$ e $E|X_0 - X_0'|$ podem ser calculados de forma simples.

Proposição 3.4.1. *Se X_0 é uma v.a.r. com função de distribuição $F_0 = F_{\mathcal{N}(\mu,\sigma^2)}$ e $x \in \mathbb{R}$, então*

$$E|x - X_0| = (x - \mu)\left(2F_{\mathcal{N}(\mu,\sigma^2)}(x) - 1\right) + 2\sigma^2 f_{\mathcal{N}(\mu,\sigma^2)}(x)$$

e

$$E|X_0 - X_0'| = \frac{\sigma}{\sqrt{\pi}}.$$

Demonstração. Comecemos pelo caso particular em que $\mu = 0$ e $\sigma = 1$. Tendo em conta que $Z \sim F_{\mathcal{N}(0,1)}$ e para $a \in \mathbb{R}$ temos

$$\begin{aligned}
E|a - Z| &= \int_{\mathbb{R}} |a - z| f_{\mathcal{N}(0,1)}(z)\, dz \\
&= \int_{\mathbb{R}} |z - a| f_{\mathcal{N}(0,1)}(z)\, dz \\
&= \int_{-\infty}^{a} -(z-a) f_{\mathcal{N}(0,1)}(z)\, dz + \int_{a}^{+\infty} (z-a) f_{\mathcal{N}(0,1)}(z)\, dz \\
&= \int_{-\infty}^{a} -z f_{\mathcal{N}(0,1)}(z)\, dz + a\int_{-\infty}^{a} f_{\mathcal{N}(0,1)}(z)\, dz + \int_{a}^{+\infty} z f_{\mathcal{N}(0,1)}(z)\, dz - a\int_{a}^{+\infty} f_{\mathcal{N}(0,1)}(z)\, dz \\
&= \int_{-\infty}^{a} f'_{\mathcal{N}(0,1)}(z)\, dz + aF_{\mathcal{N}(0,1)}(a) - \int_{a}^{+\infty} f'_{\mathcal{N}(0,1)}(z)\, dz - a\left(1 - F_{\mathcal{N}(0,1)}(a)\right) \\
&= f_{\mathcal{N}(0,1)}(a) + 2aF_{\mathcal{N}(0,1)}(a) + f_{\mathcal{N}(0,1)}(a) - a \\
&= a\left(2F_{\mathcal{N}(0,1)}(a) - 1\right) + 2f_{\mathcal{N}(0,1)}(a).
\end{aligned}$$

Sendo $X_0 = \mu + \sigma Z$ pelo resultado anterior temos que

$$\begin{aligned}
E|x - X_0| &= \sigma E\left|\frac{x-\mu}{\sigma} - Z\right| \\
&= \sigma\left(\frac{x-\mu}{\sigma}\left(2F_{\mathcal{N}(0,1)}\left(\frac{x-\mu}{\sigma}\right) - 1\right) + 2f_{\mathcal{N}(0,1)}\left(\frac{x-\mu}{\sigma}\right)\right) \\
&= (x-\mu)\left(2F_{\mathcal{N}(\mu,\sigma^2)}(x) - 1\right) + 2\sigma^2 f_{\mathcal{N}(\mu,\sigma^2)}(x),
\end{aligned} \tag{3.9}$$

uma vez que $F_{\mathcal{N}(\mu,\sigma^2)}(x) = F_{\mathcal{N}(0,1)}\left(\frac{x-\mu}{\sigma}\right)$ e $f_{\mathcal{N}(\mu,\sigma^2)}(x) = \frac{1}{\sigma} f_{\mathcal{N}(0,1)}\left(\frac{x-\mu}{\sigma}\right)$.

Sabendo que $X_0 = \mu + \sigma Z$ e tomando $X_0' = \mu + \sigma Z'$, com $Z,\ Z' \sim F_{\mathcal{N}(0,1)}$ então

$$E|X_0 - X_0'| = \sigma E|Z - Z'|,$$

onde $Z - Z' \sim F_{\mathcal{N}(0,2)}$. Assim, tomando $x = 0$, $\mu = 0$ e $\sigma^2 = 2$ em (3.9), obtemos

$$E|Z - Z'| = 4 f_{\mathcal{N}(0,2)}(0) = \frac{2}{\sqrt{\pi}},$$

e portanto

$$E|X_0 - X_0'| = \frac{2\sigma}{\sqrt{\pi}}.$$

□

No que se segue vamos considerar o caso do teste da hipótese $H_0 : F = F_{\mathcal{N}(0,1)}$, sendo a estatística de teste associada ao teste de energia dada por

$$T_n = n\mathscr{E}_n(X_1, \ldots, X_n) = 2\sum_{i=1}^{n} \left(X_i \left(2F_{\mathcal{N}(0,1)}(X_i) - 1 \right) + 2 f_{\mathcal{N}(0,1)}(X_i) \right) - \frac{1}{n} \sum_{i,j=1}^{n} |X_i - X_j| - \frac{2n}{\sqrt{\pi}}. \quad (3.10)$$

3.4.2 Distribuições alternativas

Neste estudo de simulação vamos considerar distribuições alternativas, centradas e reduzidas, baseadas em três famílias de distribuições: Student, qui-quadrado e lognormal.

1. Relativamente à distribuição de Student com ν graus de liberdade, t_ν, consideramos as transformações lineares da forma $\frac{t_\nu}{\sqrt{\frac{\nu}{\nu-2}}}$, com $\nu = 3$, $\nu = 2.75$ e $\nu = 2.5$, que denotaremos, respetivamente, por $ST1$, $ST2$ e $ST3$. É de realçar que estas distribuições são simétricas e de caudas mais pesadas que a normal.

2. Quanto à distribuição do qui-quadrado com k graus de liberdade, $\chi^2(k)$, consideramos as transformações afins da forma $\frac{\chi^2(k)-k}{\sqrt{2k}}$, com $k = 4$, $k = 3$ e $k = 2$, que denotaremos, respetivamente, por $QQ1$, $QQ2$ e $QQ3$. Estas distribuições são assimétricas e apresentam coeficientes de assimetria 1.4142, 1.6330 e 2 e de curtose 6, 7 e 9, respetivamente.

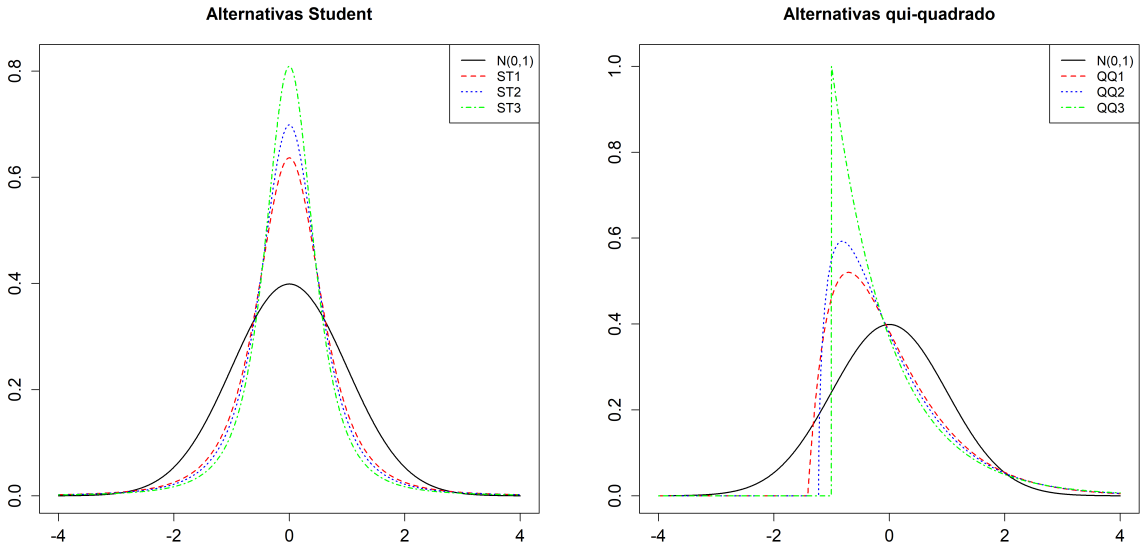
3. Relativamente à distribuição lognormal com parâmetros μ e σ, $\mathrm{LN}(\mu, \sigma^2)$, consideramos as transformações afins da forma $\frac{LN(\mu,\sigma^2) - \exp\left(\mu + \frac{\sigma^2}{2}\right)}{\sqrt{(\exp(\sigma^2)-1)\exp(2\mu+\sigma^2)}}$, com $\mu = 0$ para as três distribuições e $\sigma = 0.3$, $\sigma = 0.5$ e $\sigma = 0.7$, que denotaremos, respetivamente, por $LN1$, $LN2$ e $LN3$. Tal como no caso anterior, estas distribuições são assimétricas e apresentam coeficientes de assimetria 0.9495, 1.7502 e 2.8883 e de curtose 4.6449, 8.8984 e 20.7912, respetivamente.

A função densidade de cada uma das distribuições alternativas mencionadas está representada na Figura 3.1 conjuntamente com a função densidade de $\mathcal{N}(0,1)$.

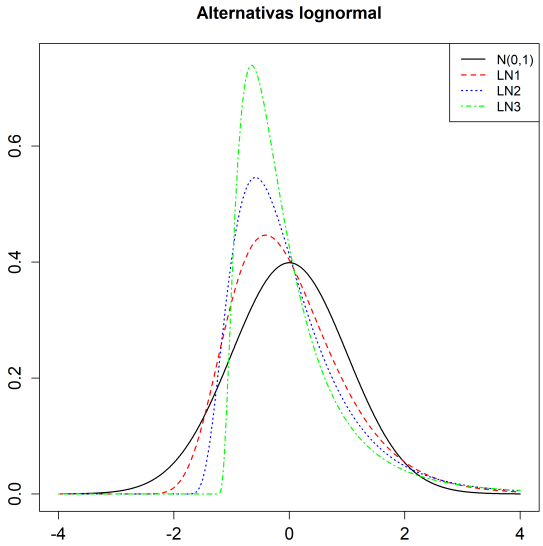
3.4.3 Resultados de potência

Com o objetivo de avaliar a potência do teste de energia, vamos comparar os resultados com o teste de Anderson-Darling, tomado como teste de referência.

A estatística de teste proposta por Anderson e Darling, A_n^2, é baseada na diferença quadrática ponderada entre a função de distribuição empírica associada à amostra $X_1, \ldots, X_n$, F_n, e F_0 (Moore, 1986, p. 100),



(a) (b)



(c)

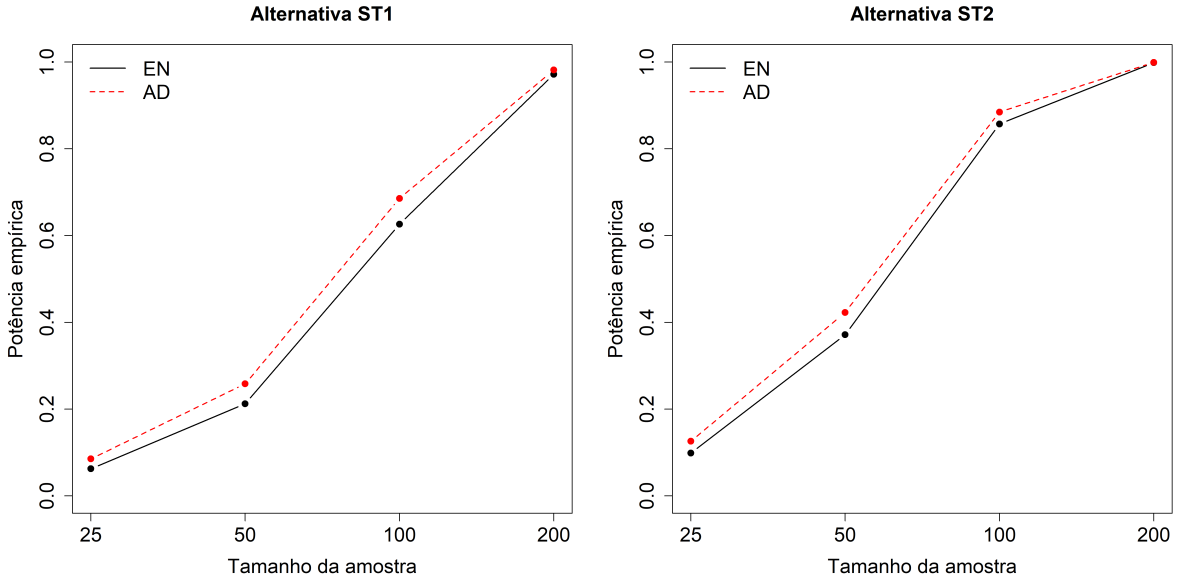
Figura 3.1 Densidades de probabilidade de cada conjunto de alternativas Student (a), qui-quadrado (b) e lognormal (c) conjuntamente com a de $\mathcal{N}(0,1)$.

isto é,

$$A_n^2 = n\int_{-\infty}^{+\infty} (F_n(x) - F_0(x))^2 \frac{1}{F_0(x)(1-F_0(x))}\, dF_0(x).$$

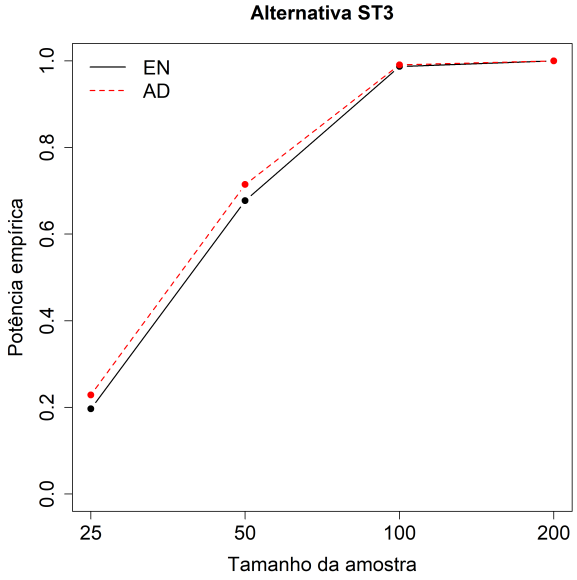
Assim, o nosso objetivo é estimar a potência do teste de energia comparando-a com a do teste de Anderson-Darling, onde, por uma questão de simplicidade, passamos a designar por teste EN e teste AD, respetivamente.

Recordamos que a região crítica associada ao teste EN é dada por $\{T_n > c_n(\alpha)\}$, onde $c_n(\alpha) = F_{T_n}^{-1}(1-\alpha)$ é o quantil de ordem $1-\alpha$ de T_n, dada por (3.10), sob a hipótese nula. Uma vez que este



(a)

(b)



(c)

Figura 3.2 (a), (b) e (c) Resultados das potências das respetivas alternativas Student.

quantil não é conhecido, na implementação prática deste procedimento de teste, vamos estimar o quantil $c_n(\alpha)$ pelo método de Monte Carlo, simulando valores da estatística de teste sob a hipótese nula. Como tal, calculamos o valor da estatística de teste T_n para 10^5 amostras de tamanho n geradas com distribuição $F_0 = F_{\mathcal{N}(0,1)}$ e, através da função $\texttt{quantile}(\cdot, \texttt{type} = 7)$ do R, é determinado o valor do respetivo quantil empírico que tomamos como aproximação para $c_n(\alpha)$.

Para este estudo consideramos os tamanhos de amostras $n = 25, 50, 100, 200$ e ainda os níveis de significância $\alpha = 0.01, 0.02, 0.05$. De modo análogo, obtivemos, também, os quantis empíricos para a estatística A_n^2. Apresentamos na Tabela 3.1 os valores estimados dos quantis de T_n e A_n^2 obtidos para as diversas combinações de n e α.

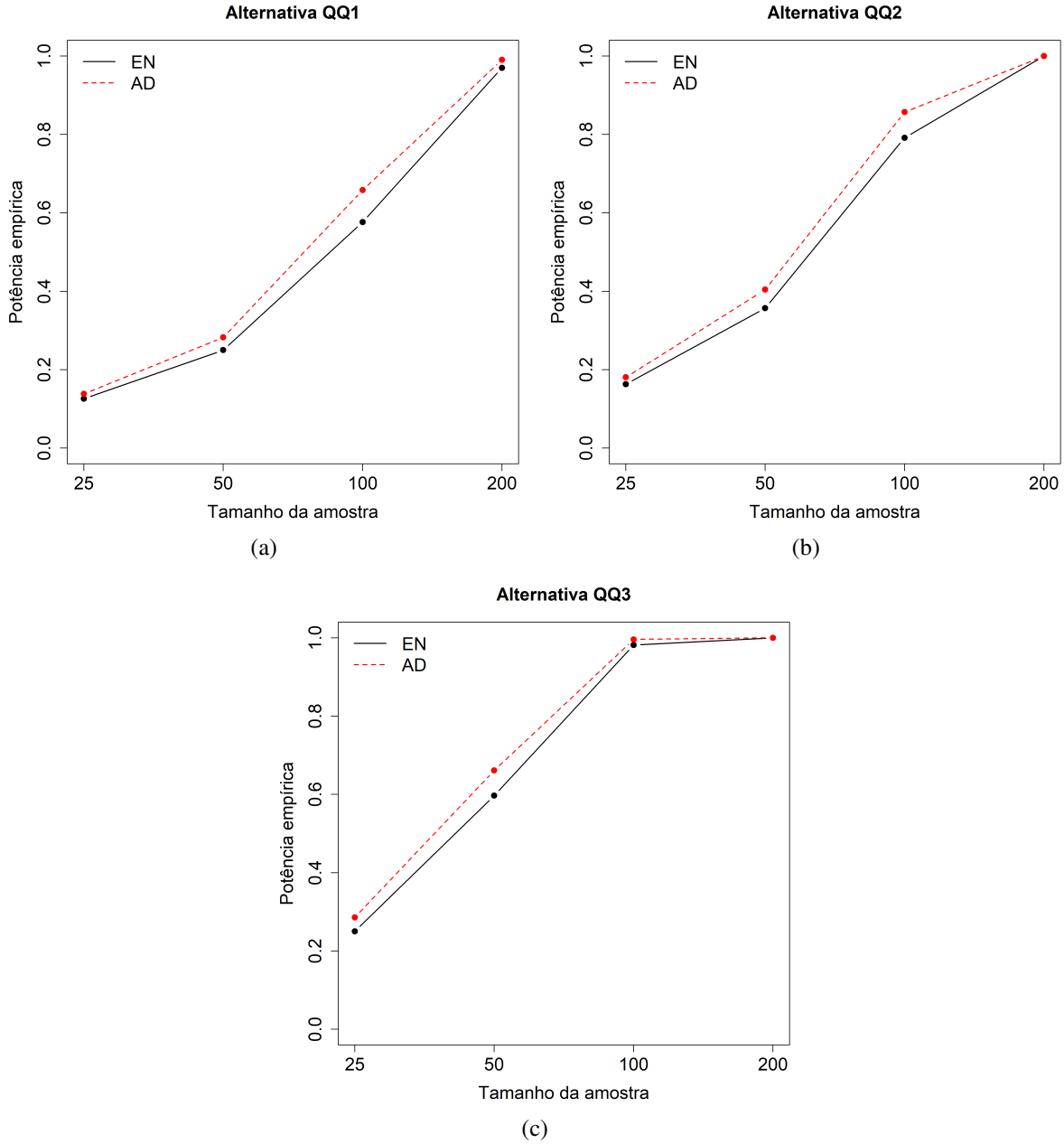


Figura 3.3 (a), (b) e (c) Resultados das potências das respetivas alternativas qui-quadrado.

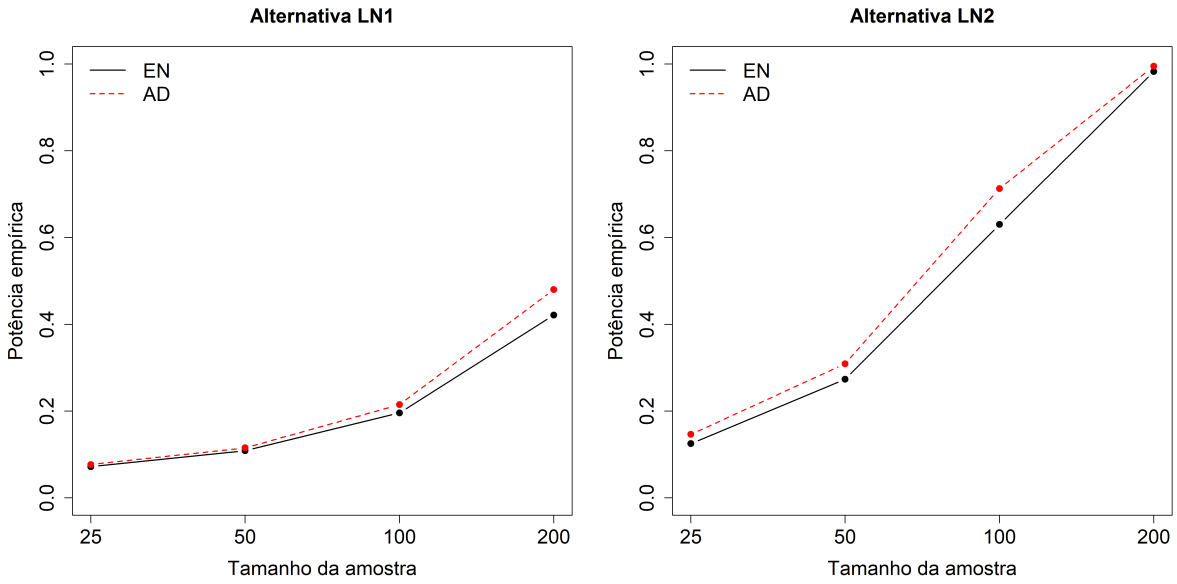
$n\backslash\alpha$	0.01	0.02	0.05
25	4.5038	3.8367	2.9121
50	4.5146	3.8064	2.9025
100	4.4518	3.7514	2.8819
200	4.5081	3.7886	2.8883

(a)

$n\backslash\alpha$	0.01	0.02	0.05
25	3.9006	3.3015	2.5064
50	3.8594	3.2660	2.4950
100	3.8783	3.2605	2.4720
200	3.8909	3.2907	2.4970

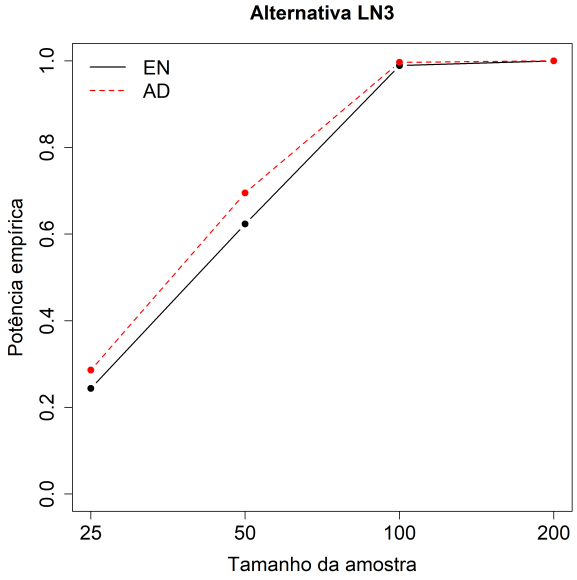
(b)

Tabela 3.1 (a) Quantis empíricos de T_n; (b) Quantis empíricos de A_n^2.



(a)

(b)



(c)

Figura 3.4 (a), (b) e (c) Resultados das potências das respetivas alternativas lognormal.

De seguida, passamos à estimação da potência do teste EN e do teste AD. Para o efeito foram geradas 10^4 amostras de tamanho n de cada uma das alternativas.

Os resultados de potência apresentados nas Figuras 3.2, 3.3 e 3.4 são relativos ao nível de significância $\alpha = 0.05$. Vamos analisar as alternativas em três grupos: distribuições baseadas nas leis de Student, nas leis do qui-quadrado e nas leis lognormal. Cada gráfico indica-nos a evolução da potência empírica de cada alternativa considerada para os testes EN e AD à medida que o tamanho da amostra aumenta.

De um modo geral, podemos concluir que o teste AD apresenta melhores resultados de potência do que o teste EN, apesar de, para alguns tamanhos de amostra, não haver grandes diferenças entre os dois testes.

Verificamos, também, que a potência aumenta à medida que o tamanho da amostra aumenta, como seria de esperar tendo em conta o Teorema 3.3.1.

Reparemos, ainda, que a potência aumenta à medida que a distribuição alternativa vai sofrendo um maior afastamento em relação à distribuição normal standard. Atentemos no caso das alternativas Student e na Figura 3.1 (a), onde vai havendo um afastamento cada vez maior em relação à distribuição normal standard, resultando em valores de potência cada vez maiores. Relativamente às distribuições alternativas qui-quadrado e lognormal, observando as Figuras 3.1 (b) e 3.1 (c) e, ainda, tendo em conta os valores do coeficiente de assimetria e da curtose, verificamos que a potência também aumenta à medida que a distribuição alternativa se vai afastando da distribuição normal standard.

Os valores dos resultados de potência e o código em linguagem R estão apresentados nos Anexos B e C, respetivamente.

Capítulo 4

Teste de ajustamento a uma família de distribuições

Neste capítulo estamos interessados no teste de ajustamento a uma família de distribuições de localização e escala. Analogamente ao capítulo anterior, vamos estabelecer a estatística de teste e a sua distribuição assintótica e, por fim, estudar o nível de significância do teste e a convergência do mesmo. No entanto, como veremos, será necessário, agora, impor algumas condições e vários resultados auxiliares.

4.1 Estatística de teste T_n

Neste capítulo apresentamos a estatística de teste associada ao teste de hipóteses

$$H_0 : F \in \mathscr{F}_0 \quad vs. \quad H_a : F \notin \mathscr{F}_0, \tag{4.1}$$

onde $\mathscr{F}_0$ é a família de funções de distribuição de localização e escala, isto é,

$$\mathscr{F}_0 = \left\{ G(\,\cdot\,;\theta_1,\theta_2) : \theta_1 \in \mathbb{R} \text{ e } \theta_2 \in \mathbb{R}^+ \right\}, \tag{4.2}$$

com

$$G(x;\theta_1,\theta_2) = F_0\Big(\frac{x-\theta_1}{\theta_2}\Big),$$

onde F_0 é uma função de distribuição conhecida e $(\theta_1,\theta_2) \in \mathbb{R}\times\mathbb{R}^+$ é o vetor dos parâmetros de localização e escala da família $\mathscr{F}_0$.

Sejam $X_1,\ldots,X_n$ v.a.r. i.i.d. com função de distribuição F_X. Uma vez que

$$F_X \in \mathscr{F}_0 \text{ se e só se } F_Y = F_0,$$

onde $Y = \frac{X-\theta_1}{\theta_2}$, para algum $(\theta_1,\theta_2) \in \mathbb{R}\times\mathbb{R}^+$, pretendendo testar as hipóteses (4.1) com base numa amostra $X_1,\ldots,X_n$ é natural considerar uma estatística de teste análoga à considerada no Capítulo 3, onde as variáveis $X_1,\ldots,X_n$ são substituídas por $\hat{Y}_1,\ldots,\hat{Y}_n$ com

$$\hat{Y}_i = \hat{Y}_i(X_1,\ldots,X_n) = \frac{X_i-\hat{\theta}_1}{\hat{\theta}_2},$$

sendo $\hat{\theta}_1$ e $\hat{\theta}_2$ estimadores dos parâmetros desconhecidos θ_1 e θ_2, respetivamente. Neste sentido, estudamos neste capítulo o teste das hipóteses (4.1) baseado na estatística de teste

$$T_n = T_n(X_1,\ldots,X_n) = \frac{1}{n}\sum_{i,j=1}^{n} h(\hat{Y}_i,\hat{Y}_j), \tag{4.3}$$

onde, recordando (2.3) e (2.6),

$$\begin{aligned} h(x,y) &= E|x-X_0| + E|y-X_0| - |x-y| - E|X_0-X_0'| \\ &= 2\int_{\mathbb{R}} \{\mathbb{1}_{]-\infty,t]}(x) - F_0(t)\}\{\mathbb{1}_{]-\infty,t]}(y) - F_0(t)\}\, dt, \end{aligned}$$

para todo o $x,y \in \mathbb{R}$.

Atendendo à representação (2.6) do núcleo h e de forma análoga ao que vimos na parte final do Capítulo 2, a estatística T_n é dada por

$$T_n = 2n\int_{\mathbb{R}} \{\hat{F}_n(t) - F_0(t)\}^2\, dt,$$

onde

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{]-\infty,t]}(\hat{Y}_i).$$

A estatística T_n pode, ainda, ser expressa em termos da função de distribuição empírica associada à amostra $X_1,\ldots,X_n$,

$$\begin{aligned} T_n &= 2n\int_{\mathbb{R}} \{F_n(\hat{\theta}_2 t + \hat{\theta}_1) - F_0(t)\}^2\, dt \\ &= 2n\int_{\mathbb{R}} \Big\{F_n(s) - F_0\Big(\frac{s-\hat{\theta}_1}{\hat{\theta}_2}\Big)\Big\}^2 \frac{1}{\hat{\theta}_2}\, ds \\ &= \frac{2n}{\hat{\theta}_2}\int_{\mathbb{R}} \{F_n(s) - G(s;\hat{\theta}_1,\hat{\theta}_2)\}^2\, ds. \end{aligned} \tag{4.4}$$

4.2 Comportamento assintótico de T_n sob H_0

Com o objetivo de obter o comportamento assintótico da estatística de teste T_n sob a hipótese nula é necessário impor condições sobre F_0 e os estimadores $\hat{\theta}_1$ e $\hat{\theta}_2$:

(C1) Seja F_0 duas vezes diferenciável em $\mathbb{R}$ com

$$\int_R F_0'(x)^2\, dx < \infty, \quad \int_R x^2 F_0'(x)\, dx < \infty,$$

$$\int_R F_0''(x)^2\, dx < \infty \quad \text{e} \quad \int_R x^4 F_0''(x)^2\, dx < \infty.$$

(C2) Para todo o $F = G(\,\cdot\,,\theta_1,\theta_2) \in \mathscr{F}_0$, tem-se

$$\sqrt{n}(\hat{\theta}_i - \theta_i) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} \Psi_i(X_j;\theta_1,\theta_2) + o_P(1),$$

para $i = 1,2$, onde as funções Ψ_i, $i = 1,2$, são tais que

$$E(\Psi_i(X;\theta_1,\theta_2)) = 0 \text{ e } E(\Psi_i(X;\theta_1,\theta_2)^2) < +\infty,$$

onde $X \sim F$.

Esta condição estabelece que o estimador $\hat{\theta}_i$ é um estimador convergente em probabilidade de θ_i, com $i = 1,2$, e, além disso, $\sqrt{n}(\hat{\theta}_i - \theta_i)$ é assintoticamente normal. Sob certas condições de regularidade, esta hipótese é satisfeita pelos estimadores dos momentos e da máxima verosimilhança.

Como mostramos a seguir, sob as condições (C1) e (C2) e certas condições de regularidade sobre a função F_0, o comportamento assintótico da estatística de teste T_n pode ser obtido, tal como no capítulo anterior, a partir do Teorema A.1.2 que estabelece a distribuição limite duma U-estatística degenerada.

No que se segue assumimos que $X_1,\ldots,X_n$ são v.a.r. independentes com distribuição $F = G(\,\cdot\,,\theta_1,\theta_2)$ para algum $\theta_1 \in \mathbb{R}$ e $\theta_2 \in \mathbb{R}^+$.

Começamos por utilizar a fórmula de Taylor de segunda ordem (Lima, 1992, p. 220). Assim, para $t \in \mathbb{R}$ temos

$$\begin{aligned} G(t;\hat{\theta}_1,\hat{\theta}_2) = G(t;\theta_1,\theta_2) &+ \frac{\partial G}{\partial\theta_1}(t;\theta_1,\theta_2)(\hat{\theta}_1-\theta_1) + \frac{\partial G}{\partial\theta_2}(t;\theta_1,\theta_2)(\hat{\theta}_2-\theta_2) + \frac{1}{2}\frac{\partial^2 G}{\partial\theta_1^2}(t;\tilde{\theta}_1,\tilde{\theta}_2)(\hat{\theta}_1-\theta_1)^2 \\ &+ \frac{\partial^2 G}{\partial\theta_1\partial\theta_2}(t;\tilde{\theta}_1,\tilde{\theta}_2)(\hat{\theta}_1-\theta_1)(\hat{\theta}_2-\theta_2) + \frac{1}{2}\frac{\partial^2 G}{\partial\theta_2^2}(t;\tilde{\theta}_1,\tilde{\theta}_2)(\hat{\theta}_2-\theta_2)^2, \end{aligned} \tag{4.5}$$

para algum $\tilde{\theta}_i$ entre θ_i e $\hat{\theta}_i$, $i = 1,2$, o que permite escrever

$$\begin{aligned} T_n &= \frac{2n}{\hat{\theta}_2}\int_{\mathbb{R}}\{F_n(t) - G(t;\hat{\theta}_1,\hat{\theta}_2)\}^2\,dt \\ &= \frac{2n}{\hat{\theta}_2}\int_{\mathbb{R}}\Big\{F_n(t) - G(t;\theta_1,\theta_2) - \frac{\partial G}{\partial\theta_1}(t;\theta_1,\theta_2)(\hat{\theta}_1-\theta_1) - \frac{\partial G}{\partial\theta_2}(t;\theta_1,\theta_2)(\hat{\theta}_2-\theta_2) \\ &\quad - \frac{1}{2}\frac{\partial^2 G}{\partial\theta_1^2}(t;\tilde{\theta}_1,\tilde{\theta}_2)(\hat{\theta}_1-\theta_1)^2 - \frac{\partial^2 G}{\partial\theta_1\partial\theta_2}(t;\tilde{\theta}_1,\tilde{\theta}_2)(\hat{\theta}_1-\theta_1)(\hat{\theta}_2-\theta_2) - \frac{1}{2}\frac{\partial^2 G}{\partial\theta_2^2}(t;\tilde{\theta}_1,\tilde{\theta}_2)(\hat{\theta}_2-\theta_2)^2\Big\}^2\,dt \\ &= \frac{2}{\hat{\theta}_2}\Big(n\int_{\mathbb{R}}A_n(t)^2\,dt - 2n\int_R A_n(t)B_n(t)\,dt + n\int_R B_n(t)^2\,dt\Big), \end{aligned} \tag{4.6}$$

onde

$$A_n(t) = F_n(t) - G(t;\theta_1,\theta_2) - \frac{\partial G}{\partial\theta_1}(t;\theta_1,\theta_2)(\hat{\theta}_1-\theta_1) - \frac{\partial G}{\partial\theta_2}(t;\theta_1,\theta_2)(\hat{\theta}_2-\theta_2)$$

e

$$B_n(t) = \frac{1}{2}\frac{\partial^2 G}{\partial\theta_1^2}(t;\tilde{\theta}_1,\tilde{\theta}_2)(\hat{\theta}_1-\theta_1)^2 + \frac{\partial^2 G}{\partial\theta_1\partial\theta_2}(t;\tilde{\theta}_1,\tilde{\theta}_2)(\hat{\theta}_1-\theta_1)(\hat{\theta}_2-\theta_2) + \frac{1}{2}\frac{\partial^2 G}{\partial\theta_2^2}(t;\tilde{\theta}_1,\tilde{\theta}_2)(\hat{\theta}_2-\theta_2)^2.$$

Uma vez que $G(t;\theta_1,\theta_2) = F_0\Big(\frac{t-\theta_1}{\theta_2}\Big)$, temos

$$\frac{\partial G}{\partial\theta_1}(t;\theta_1,\theta_2) = -\frac{1}{\theta_2}F_0'\Big(\frac{t-\theta_1}{\theta_2}\Big),$$

$$\begin{aligned}
\frac{\partial G}{\partial \theta_2}(t;\theta_1,\theta_2) &= -\frac{t-\theta_1}{\theta_2^2}F_0'\Big(\frac{t-\theta_1}{\theta_2}\Big),\\
\frac{\partial^2 G}{\partial \theta_1^2}(t;\theta_1,\theta_2) &= \frac{1}{\theta_2^2}F_0''\Big(\frac{t-\theta_1}{\theta_2}\Big),\\
\frac{\partial^2 G}{\partial \theta_2^2}(t;\theta_1,\theta_2) &= \frac{2(t-\theta_1)}{\theta_2^3}F_0'\Big(\frac{t-\theta_1}{\theta_2}\Big)+\Big(\frac{t-\theta_1}{\theta_2^2}\Big)^2F_0''\Big(\frac{t-\theta_1}{\theta_2}\Big),\\
\frac{\partial G}{\partial \theta_1\partial \theta_2}(t;\theta_1,\theta_2) &= \frac{1}{\theta_2^2}F_0'\Big(\frac{t-\theta_1}{\theta_2}\Big)+\frac{t-\theta_1}{\theta_2^3}F_0''\Big(\frac{t-\theta_1}{\theta_2}\Big).
\end{aligned} \tag{4.7}$$

Estamos, agora, em condições de analisar os termos de (4.6), o que faremos nos lemas seguintes.

Lema 4.2.1. *Nas condições (C1) e (C2) tem-se que*

$$n\int_{\mathbb{R}} B_n(t)^2\,dt = O_P(n^{-1})$$

Demonstração. Tendo em conta que $(a+b+c)^2 \le 3(a^2+b^2+c^2)$, para todo o $a,b,c\in\mathbb{R}$, temos

$$\begin{aligned}
n\int_{\mathbb{R}} B_n(t)^2\,dt \le{}& \frac{3n}{4}(\hat\theta_1-\theta_1)^4\int_{\mathbb{R}}\frac{\partial^2 G}{\partial\theta_1^2}(t;\tilde\theta_1,\tilde\theta_2)^2\,dt+3n(\hat\theta_1-\theta_1)^2(\hat\theta_2-\theta_2)^2\int_{\mathbb{R}}\frac{\partial^2 G}{\partial\theta_1\partial\theta_2}(t;\tilde\theta_1,\tilde\theta_2)^2\,dt\\
&+\frac{3n}{4}(\hat\theta_2-\theta_2)^4\int_{\mathbb{R}}\frac{\partial^2 G}{\partial\theta_2^2}(t;\tilde\theta_1,\tilde\theta_2)^2\,dt,
\end{aligned}$$

onde

$$\int_{\mathbb{R}}\frac{\partial^2 G}{\partial\theta_1^2}(t;\tilde\theta_1,\tilde\theta_2)^2\,dt=\frac{1}{\tilde\theta_2^3}\int_{\mathbb{R}}F_0''(x)^2\,dx,$$

$$\int_{\mathbb{R}}\frac{\partial^2 G}{\partial\theta_2^2}(t;\tilde\theta_1,\tilde\theta_2)^2\,dt\le\frac{2}{\tilde\theta_2^3}\Big(4\int_{\mathbb{R}}x^2F_0'(x)^2\,dx+\int_{\mathbb{R}}x^4F_0''(x)^2\,dx\Big)$$

e

$$\int_{\mathbb{R}}\frac{\partial^2 G}{\partial\theta_1\partial\theta_2}(t;\tilde\theta_1,\tilde\theta_2)^2\,dt\le\frac{2}{\tilde\theta_2^3}\Big(\int_{\mathbb{R}}F_0'(x)^2\,dx+\int_{\mathbb{R}}x^2F_0''(x)^2dx\Big).$$

Para concluir a demonstração basta notar que $\hat\theta_i-\theta_i=O_P(n^{-1/2})$. □

Lema 4.2.2. *Nas condições (C1) e (C2) tem-se*

$$n\int_{\mathbb{R}}A_n(t)B_n(t)\,dt=O_P(n^{-1/2}).$$

Demonstração. Pela desigualdade de Cauchy-Schwarz sabemos que

$$\Big|n\int_{\mathbb{R}}A_n(t)B_n(t)\,dt\Big|\le\Big(n\int_{\mathbb{R}}A_n(t)^2\,dt\Big)^{1/2}\Big(n\int_{\mathbb{R}}B_n(t)^2\,dt\Big)^{1/2}, \tag{4.8}$$

onde

$$\begin{aligned}
n\int_{\mathbb{R}}A_n(t)^2\,dt\le{}&3n\int_{\mathbb{R}}\{F_n(t)-G(t;\theta_1,\theta_2)\}^2\,dt+3n(\hat\theta_1-\theta_1)^2\int_{\mathbb{R}}\frac{\partial G}{\theta_1}(t;\theta_1,\theta_2)^2\,dt\\
&+3n(\hat\theta_2-\theta_2)^2\int_{\mathbb{R}}\frac{\partial G}{\theta_2}(t;\theta_1,\theta_2)^2\,dt,
\end{aligned}$$

com

$$\int_{\mathbb{R}} \frac{\partial G}{\partial \theta_1}(t;\theta_1,\theta_2)^2\, dt = \frac{1}{\theta_2}\int_{\mathbb{R}} F_0'(x)^2\, dx \tag{4.9}$$

e

$$\int_{\mathbb{R}} \frac{\partial G}{\partial \theta_2}(t;\theta_1,\theta_2)^2\, dt = \frac{1}{\theta_2}\int_{\mathbb{R}} x^2 F_0'(x)^2\, dx. \tag{4.10}$$

Como $\hat{\theta}_i - \theta_i = O_P(n^{-1/2})$ e pelo Teorema 3.1.1

$$n\int_{\mathbb{R}} \{F_n(t) - G(t;\theta_1,\theta_2\}^2\, dt = O_P(1),$$

concluímos que

$$n\int_{\mathbb{R}} A_n(t)^2\, dt = O_P(1). \tag{4.11}$$

Finalmente, pelo Lema 4.2.1 e pela desigualdade (4.8) obtemos o resultado enunciado. □

Relembrando a igualdade (4.6), pelos lemas anteriores concluímos que

$$\begin{aligned} T_n &= \frac{2n}{\hat{\theta}_2}\int_{\mathbb{R}} A_n(t)^2\, dt + O_P(n^{-1/2}) \\ &= \frac{2n}{\theta_2}\int_{\mathbb{R}} A_n(t)^2\, dt + o_P(1), \end{aligned} \tag{4.12}$$

onde esta última igualdade é consequência da convergência em probabilidade de $\hat{\theta}_2$ para θ_2 e de (4.11).

4.3 Distribuição assintótica de T_n sob H_0

Nesta secção determinamos a distribuição assintótica da estatística T_n sob a hipótese nula. Para tal vamos mostrar, no resultado seguinte, que o primeiro termo de (4.12) admite uma representação assintótica como uma V-estatística, o que permitirá, pelo Teorema A.1.2, obter a distribuição assintótica de T_n.

Lema 4.3.1. *Nas condições (C1) e (C2) tem-se*

$$\frac{2n}{\theta_2}\int_{\mathbb{R}} A_n(t)^2\, dt = \frac{1}{n}\sum_{i,j=1}^{n} H(X_i,X_j) + o_P(1),$$

onde

$$H(x,y) = \frac{2}{\theta_2}\int_{\mathbb{R}} \Phi(x,t;\theta_1,\theta_2)\Phi(y,t;\theta_1,\theta_2)\, dt \tag{4.13}$$

e

$$\Phi(x,t;\theta_1,\theta_2) = \mathbb{1}_{]-\infty,t]}(x) - G(t;\theta_1,\theta_2) - \frac{\partial G}{\partial \theta_1}(t;\theta_1,\theta_2)\Psi_1(x;\theta_1,\theta_2) - \frac{\partial G}{\partial \theta_2}(t;\theta_1,\theta_2)\Psi_2(x;\theta_1,\theta_2),$$

para todo o $x, y \in \mathbb{R}$.

Demonstração. Usando a condição (C2) temos

$$\begin{aligned}\frac{2n}{\theta_2}\int_{\mathbb{R}} A_n(t)^2\,dt &= \frac{2n}{\theta_2}\int_{\mathbb{R}}\Big\{F_n(t)-G(t;\theta_1,\theta_2)-\frac{\partial G}{\partial\theta_1}(t;\theta_1,\theta_2)\Big(\frac{1}{n}\sum_{i=1}^{n}\Psi_1(X_i;\theta_1,\theta_2)+\varphi_1\Big)\\ &\qquad -\frac{\partial G}{\partial\theta_2}(t;\theta_1,\theta_2)\Big(\frac{1}{n}\sum_{i=1}^{n}\Psi_2(X_i;\theta_1,\theta_2)+\varphi_2\Big)\Big\}^2\,dt\\ &= \frac{2n}{\theta_2}\int_{\mathbb{R}}\Big\{\frac{1}{n}\sum_{i=1}^{n}\Phi(X_i,t;\theta_1,\theta_2)-\frac{\partial G}{\theta_1}(t;\theta_1,\theta_2)\varphi_1-\frac{\partial G}{\theta_2}(t;\theta_1,\theta_2)\varphi_2\Big\}^2\,dt\\ &= \frac{1}{n}\sum_{i,j=1}^{n}H(X_i,X_j)-\frac{4n}{\theta_2}\int_{\mathbb{R}}A_{1,n}(t)A_{2,n}(t)\,dt+\frac{2n}{\theta_2}\int_{\mathbb{R}}A_{2,n}(t)^2\,dt, \qquad (4.14)\end{aligned}$$

onde $\varphi_i = o_P(n^{-1/2}),\ i=1,2,$

$$A_{1,n}(t)=\frac{1}{n}\sum_{i=1}^{n}\Phi(X_i,t;\theta_1,\theta_2)$$

e

$$A_{2,n}(t)=\frac{\partial G}{\partial\theta_1}(t;\theta_1,\theta_2)\varphi_1+\frac{\partial G}{\partial\theta_2}(t;\theta_1,\theta_2)\varphi_2.$$

Analisando o último termo da equação (4.14), pelas igualdades (4.9) e (4.10) e pela desigualdade $(a+b)^2\le 2(a^2+b^2)$, para todo o $a,b\in\mathbb{R}$, concluímos que

$$n\int_{\mathbb{R}}A_{2,n}(t)^2\,dt\le 2n\varphi_1^2\int_{\mathbb{R}}\frac{\partial G}{\partial\theta_1}(t;\theta_1,\theta_2)^2\,dt+2n\varphi_2^2\int_{\mathbb{R}}\frac{\partial G}{\partial\theta_2}(t;\theta_1,\theta_2)^2\,dt=o_P(1). \qquad (4.15)$$

Relativamente ao segundo termo de (4.14), pela desigualdade de Cauchy-Schwarz sabemos que

$$\Big|n\int_{\mathbb{R}}A_{1,n}(t)A_{2,n}(t)\,dt\Big|\le\Big(n\int_{\mathbb{R}}A_{1,n}(t)^2\,dt\Big)^{1/2}\Big(n\int_{\mathbb{R}}A_{2,n}(t)\,dt\Big)^{1/2}, \qquad (4.16)$$

onde

$$\begin{aligned}n\int_{\mathbb{R}}A_{1,n}(t)^2\,dt &\le 3n\int_{\mathbb{R}}\{F_n(t)-G(t;\theta_1,\theta_2)\}^2\,dt+3n\Big(\frac{1}{n}\sum_{i=1}^{n}\Psi_1(X_i;\theta_1,\theta_2)\Big)^2\int_{\mathbb{R}}\frac{\partial G}{\partial\theta_1}(t;\theta_1,\theta_2)^2\,dt\\ &\quad+3n\Big(\frac{1}{n}\sum_{i=1}^{n}\Psi_2(X_i;\theta_1,\theta_2)\Big)^2\int_{\mathbb{R}}\frac{\partial G}{\partial\theta_2}(t;\theta_1,\theta_2)^2\,dt.\end{aligned}$$

Como $\frac{1}{n}\sum_{i=1}^{n}\Psi_j(X_i;\theta_1,\theta_2)=O_P(n^{-1/2}),\ j=1,2,$ e pelo Teorema 3.1.1

$$n\int_{\mathbb{R}}\{F_n(t)-G(t;\theta_1\theta_2)\}^2\,dt=O_P(1),$$

concluímos que

$$n\int_{\mathbb{R}}A_{1,n}(t)^2\,dt=O_P(1).$$

Assim, por (4.15) e pela desigualdade (4.16) concluímos que

$$n\int_{\mathbb{R}}A_{1,n}(t)A_{n_2}(t)\,dt=o_P(1),$$

o que termina a demonstração. □

O resultado seguinte, permite, finalmente, estabelecer a distribuição limite da estatística de teste T_n sob a hipótese nula.

Teorema 4.3.1. *Sendo $X_1,\ldots,X_n$ v.a.r. i.i.d. com X, onde $X \sim F = G(\,\cdot\,,\theta_1,\theta_2) \in \mathscr{F}_0$, então*

$$T_n(X_1,\ldots,X_n) \xrightarrow{d} E(H(X,X)) + \sum_{k=1}^{+\infty} \lambda_k (Z_k^2 - 1),$$

onde Z_k, $k \geq 1$, são v.a.r i.i.d. com lei normal standard e $\lambda_k \geq 0$, $k \geq 1$, com $\sum_{k=1}^{n} \lambda_k^2 < \infty$, são os valores próprios associados ao operador A_H definido de $L^2(\mathbb{R},F)$ em $L^2(\mathbb{R},F)$ por

$$A_H g(x) = \int_{\mathbb{R}} H(x,y) g(y)\, dF(y),\ x \in \mathbb{R}, \tag{4.17}$$

para $g \in L^2(\mathbb{R},F)$, onde H é a função definida em (4.13) e $L^2(\mathbb{R},F)$ é o espaço das funções reais de variável real g tais que $\int_{\mathbb{R}} g(y)^2\, dF(y) < +\infty$.

Demonstração. Notemos que podemos escrever

$$T_n(X_1,\ldots,X_n) = \frac{1}{n}\sum_{i=1}^{n} H(X_i,X_i) + (n-1)U_n + o_P(1),$$

onde U_n é a U-estatística definida por

$$U_n = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} H(X_i,X_j).$$

Em primeiro lugar, provemos que $E|H(X,X)| < \infty$. Para todo o $a,b,c \in \mathbb{R}$, $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ temos

$$\begin{aligned} E|H(X,X)| &= E\Big(\frac{2}{\theta_2}\int_{\mathbb{R}} \Phi(X,t;\theta_1,\theta_2)^2\, dt\Big) \\ &\leq \frac{2}{\theta_2}\Big\{3E\Big(\int_{\mathbb{R}} \{\mathbb{1}_{]-\infty,t]}(X) - G(t;\theta_1,\theta_2)\}^2\, dt\Big) + 3\int_{\mathbb{R}} \frac{\partial G}{\partial \theta_1}(t;\theta_1,\theta_2)^2 E(\Psi_1(X;\theta_1,\theta_2)^2)\, dt \\ &\qquad + 3\int_{\mathbb{R}} \frac{\partial G}{\partial \theta_2}(t;\theta_1,\theta_2)^2 E(\Psi_2(X;\theta_1,\theta_2)^2)\, dt\Big\}. \end{aligned}$$

Tendo em conta a demonstração do Teorema 3.1.1, pelas igualdades (4.9) e (4.10) e pela condição (C2) concluímos que

$$E|H(X,X)| < \infty, \tag{4.18}$$

o que, pela lei dos grandes números de Kolmogorov, permite afirmar que

$$\frac{1}{n}\sum_{i=1}^{n} H(X_i,X_i) \xrightarrow{q.c.} E(H(X,X)).$$

De modo análogo à demonstração feita no Teorema 3.1.1, vamos verificar que o núcleo simétrico H satisfaz as condições $E(H(X,X')^2) < +\infty$ e $E(H(x,X)) = 0$, para todo o $x \in \mathbb{R}$, com o intuito de aplicar o Teorema A.1.2.

A primeira destas condições é consequência de (4.18) uma vez que da desigualdade de Cauchy-Schwarz e da independência de X e X' temos

$$\begin{aligned} E(H^2(X,X')) &= E\Big(\frac{2}{\theta_2}\int_R \Phi(X,t;\theta_1,\theta_2)\Phi(X',t;\theta_1,\theta_2)\,dt\Big)^2 \\ &\leq E\Big(\Big\{\frac{2}{\theta_2}\int_{\mathbb{R}} \Phi(X,t;\theta_1,\theta_2)^2\,dt\Big\}^{1/2}\Big\{\frac{2}{\theta_2}\int_{\mathbb{R}} \Phi(X',t;\theta_1,\theta_2)^2\,dt\Big\}^{1/2}\Big)^2 \\ &= \Big(E\Big(\frac{2}{\theta_2}\int_{\mathbb{R}} \Phi(X,t;\theta_1,\theta_2)^2\,dt\Big)\Big)^2 \\ &= \big(E|H(X,X)|\big)^2. \end{aligned}$$

A segunda das condições anteriores é consequência da hipótese (C2) uma vez que

$$\begin{aligned} E(H(x,X)) &= E\Big(\frac{2}{\theta_2}\int_{\mathbb{R}} \Phi(x,t;\theta_1,\theta_2)\Phi(X,t;\theta_1,\theta_2)\,dt\Big) \\ &= \frac{2}{\theta_2}\int_{\mathbb{R}} \Phi(x,t;\theta_1,\theta_2)E(\Phi(X,t;\theta_1,\theta_2))\,dt \end{aligned}$$

e

$$\begin{aligned} &E(\Phi(X,t;\theta_1,\theta_2)) \\ &= E\Big(\mathbb{1}_{]-\infty,t]}(X) - G(t;\theta_1,\theta_2) - \frac{\partial G}{\partial\theta_1}(t;\theta_1,\theta_2)\Psi_1(X;\theta_1,\theta_2) - \frac{\partial G}{\partial\theta_2}(t;\theta_1,\theta_2)\Psi_2(X;\theta_1,\theta_2)\Big) \\ &= G(t;\theta_1,\theta_2) - G(t;\theta_1,\theta_2) - \frac{\partial G}{\partial\theta_1}(t;\theta_1,\theta_2)E(\Psi_1(X;\theta_1,\theta_2)) - \frac{\partial G}{\partial\theta_2}(t;\theta_1,\theta_2)E(\Psi_2(X;\theta_1,\theta_2)) \\ &= 0. \end{aligned}$$

Verificadas as condições anteriores podemos concluir, pelo Teorema A.1.2, que

$$nU_n \xrightarrow{d} \sum_{k=1}^{+\infty} \lambda_k(Z_k^2 - 1),$$

onde Z_k, $k \geq 1$, são v.a.r. i.i.d. com lei normal standard e λ_k, $k \geq 1$, são os valores próprios associados ao operador integral A_H definido em (4.17).

Para terminar a demonstração, mostramos que $\lambda_k \geq 0$, para todo o $k \geq 1$. Para todo o $g \in L^2(\mathbb{R},F)$ temos

$$\begin{aligned} &\int_{\mathbb{R}} g(x)(A_H g)(x)\,dF(x) \\ =&\iint_{\mathbb{R}^2} g(x)H(x,y)g(y)\,dF(y)dF(x) \\ =&\frac{2}{\theta_2}\int_{\mathbb{R}}\Big(\int_{\mathbb{R}} g(x)\Phi(x,t;\theta_1,\theta_2)\,dF(x)\int_{\mathbb{R}} g(y)\Phi(y,t;\theta_1,\theta_2)\,dF(y)\Big)\,dt \\ =&\frac{2}{\theta_2}\int_{\mathbb{R}}\Big(\int_{\mathbb{R}} g(x)\Phi(x,t;\theta_1,\theta_2)\,dF(x)\Big)^2\,dt \geq 0, \end{aligned}$$

o que, de acordo com o Teorema A.1.2, conclui a demonstração. □

4.4 Invariância de T_n sob H_0

O resultado expresso no Teorema 4.3.1 é válido para uma estatística T_n que não é necessariamente invariante para transformações de localização e escala dos dados. No entanto, essa propriedade de invariância é útil na implementação do procedimento de teste uma vez que, tal como no teste de ajustamento a uma distribuição fixa, os quantis da estatística de teste são calculados simulando valores da estatística de teste sob a hipótese nula que é agora uma hipótese composta.

No resultado seguinte mostramos que sob certas condições sobre os estimadores $\hat{\theta}_1$ e $\hat{\theta}_2$, a estatística T_n é invariante para transformações de localização e escala dos dados.

Proposição 4.4.1. *Sejam $\mathscr{F}_0$ a família definida em (4.2) e $\hat{\theta}_1$ e $\hat{\theta}_2$ estimadores de θ_1 e θ_2, respetivamente, tais que*

$$\hat{\theta}_1(aX_1+b,\ldots,aX_n+b) = a\hat{\theta}_1(X_1,\ldots,X_n)+b$$

e

$$\hat{\theta}_2(aX_1+b,\ldots,aX_n+b) = a\hat{\theta}_2(X_1,\ldots,X_n),$$

para todo o $a \in \mathbb{R}^+$ e $b \in \mathbb{R}$. Nestas condições, T_n é invariante para transformações de localização e escala dos dados, isto é,

$$T_n(aX_1+b,\ldots,aX_n+b) = T_n(X_1,\ldots,X_n),$$

para todo o $a \in \mathbb{R}^+$ e $b \in \mathbb{R}$.

Demonstração. Atendendo a que $T_n = n\mathscr{E}_n(\hat{Y}_1,\ldots,\hat{Y}_n)$ basta mostrar que $\hat{Y}_i$ é invariante para transformações de localização e escala dos dados. Com efeito, para $a \in \mathbb{R}^+$ e $b \in \mathbb{R}$ temos

$$\begin{aligned}
\hat{Y}_i(aX_1+b,\ldots,aX_n+b) &= \frac{aX_i+b-\hat{\theta}_1(aX_1+b,\ldots,aX_n+b)}{\hat{\theta}_2(aX_1+b,\ldots,aX_n+b)} \\
&= \frac{aX_i+b-a\hat{\theta}_1(X_1,\ldots,X_n)-b}{a\hat{\theta}_2(X_1,\ldots,X_n)} \\
&= \frac{a(X_i-\hat{\theta}_1(X_1,\ldots,X_n))}{a\hat{\theta}_2(X_1,\ldots,X_n)} \\
&= \frac{X_i-\hat{\theta}_1(X_1,\ldots,X_n)}{\hat{\theta}_2(X_1,\ldots,X_n)} \\
&= \hat{Y}_i(X_1,\ldots,X_n).
\end{aligned}$$

□

O resultado seguinte é agora consequência da proposição anterior e do Teorema 4.3.1.

Teorema 4.4.1. *Sob as condições do Teorema 4.3.1, se os estimadores $\hat{\theta}_1$ e $\hat{\theta}_2$ satisfizerem as condições impostas na Proposição 4.4.1, então*

$$T_n(X_1,\ldots,X_n) \xrightarrow{d} E(H(X_0,X_0)) + \sum_{k=1}^{+\infty} \lambda_k(Z_k^2-1),$$

onde $X_0 \sim F_0$, Z_k, $k \geq 1$, *são v.a.r. i.i.d. com lei normal standard e* $\lambda_k \geq 0$, $k \geq 1$, *com* $\sum_{k=1}^{n} \lambda_k^2 < \infty$, *são os valores próprios associados ao operador* A_H *definido de* $L^2(\mathbb{R}, F_0)$ *em* $L^2(\mathbb{R}, F_0)$ *por*

$$A_H g(x) = \int_{\mathbb{R}} H(x,y) g(y)\, dF_0(y),\ x \in \mathbb{R},$$

para $g \in L^2(\mathbb{R}, F_0)$, *onde*

$$H(x,y) = 2 \int_{\mathbb{R}} \Phi_0(x,t)\Phi_0(y,t)\, dt,$$

com

$$\Phi_0(x,t) = \mathbb{1}_{]-\infty,t]}(x) - F_0(t) + F_0'(t)\Psi_1(x;0,1) + tF_0'(t)\Psi_2(x;0,1)$$

e $L^2(\mathbb{R}, F_0)$ *é o espaço das funções reais de variável real g tais que* $\int_{\mathbb{R}} g(y)^2\, dF_0(y) < +\infty$.

Demonstração. Atendendo à Proposição 4.4.1, a estatística de teste T_n é invariante sob a hipótese nula. Assim, a distribuição de T_n para $F \in \mathscr{F}_0$ coincide com a distribuição de T_n para $F = F_0 = G(\,\cdot\,;0,1)$.

O resultado enunciado é agora consequência do Teorema 4.3.1 notando que

$$\Phi(x,t;0,1) = \Phi_0(x,t),$$

uma vez que $\frac{\partial G}{\partial \theta_1}(t;0,1) = -F_0'(t)$ e $\frac{\partial G}{\partial \theta_2}(t;0,1) = -tF_0'(t)$. □

4.5 Região crítica e nível de significância

De modo análogo ao que foi feito no caso do teste de ajustamento a uma distribuição fixa, vamos considerar o teste de região crítica

$$\{T_n > c_n(\alpha)\},$$

onde $c_n(\alpha)$ é o quantil de ordem $1-\alpha$ da distribuição de T_n sob a hipótese nula. Tal como provamos anteriormente, a estatística T_n é invariante sob certas condições relativamente aos estimadores $\hat{\theta}_1$ e $\hat{\theta}_2$, as quais supomos verdadeiras, pelo que $c_n(\alpha)$ não depende da distribuição considerada sob H_0.

Teorema 4.5.1. *Nas condições do Teorema 4.4.1, o teste de região crítica (3.4) é um teste com nível de significância inferior ou igual a* α, *isto é, para todo o* $F \in \mathscr{F}_0$

$$P_F(T_n > c_n(\alpha)) \leq \alpha.$$

Além disso, o teste é de nível assintótico α, *ou seja,*

$$\lim_{n \longrightarrow +\infty} P_F(T_n > c_n(\alpha)) = \alpha,$$

para todo o $F \in \mathscr{F}_0$.

Demonstração. Atendendo à invariância de T_n, para $F \in \mathscr{F}_0$, temos

$$P_F(T_n > c_n(\alpha)) = P_{F_0}(T_n > c_n(\alpha)),$$

pelo que a demonstração segue os mesmos passos do caso de hipótese nula simples (ver Teorema 3.2.1). □

4.6 Convergência do teste

De seguida, passamos ao estudo da convergência do teste cuja região crítica foi definida em (3.4), onde será necessário impor a seguinte condição sobre os estimadores $\hat{\theta}_1$ e $\hat{\theta}_2$:

(C3) Para todo o $F \in \mathscr{F} \setminus \mathscr{F}_0$, existem $\theta_1 \in \mathbb{R}$ e $\theta_2 \in \mathbb{R}^+$ tais que

$$\hat{\theta}_i \xrightarrow{p} \theta_i,$$

para $i = 1,2$.

Os seguintes resultados estabelecem a convergência do teste (3.4).

Lema 4.6.1. *Para todo o $F \in \mathscr{F} \setminus \mathscr{F}_0$ e sob as condições (C2) e (C3) tem-se que*

$$\frac{T_n}{n} = \frac{2}{\hat{\theta}_2} \int_R \{F_n(t) - G(t;\hat{\theta}_1,\hat{\theta}_2)\}^2 \, dt \xrightarrow{p} \frac{1}{\theta_2} \int_{\mathbb{R}} \{F(t) - G(t;\theta_1,\theta_2)\}^2 \, dt.$$

Demonstração. Começamos por utilizar a fórmula de Taylor de primeira ordem. Assim, para $t \in \mathbb{R}$ e $\tilde{\theta}_i$ entre θ_i e $\hat{\theta}_i$, para $i = 1,2$, temos

$$G(t;\hat{\theta}_1,\hat{\theta}_2) = G(t;\theta_1,\theta_2) + \frac{\partial G}{\partial \theta_1}(t;\tilde{\theta}_1,\tilde{\theta}_2)(\hat{\theta}_1 - \theta_1) + \frac{\partial G}{\partial \theta_2}(t;\tilde{\theta}_1,\tilde{\theta}_2)(\hat{\theta}_2 - \theta_2),$$

onde $\frac{\partial G}{\partial \theta_1}(t;\tilde{\theta}_1,\tilde{\theta}_2)$ e $\frac{\partial G}{\partial \theta_2}(t;\tilde{\theta}_1,\tilde{\theta}_2)$ são dados por (4.7).

Assim,

$$\begin{aligned}
\frac{T_n}{n} &= \frac{2}{\hat{\theta}_2} \int_{\mathbb{R}} \{F_n(t) - G(t;\hat{\theta}_1,\hat{\theta}_2)\}^2 \, dt \\
&= \frac{2}{\hat{\theta}_2} \int_R \left\{ F_n(t) - G(t;\theta_1,\theta_2) - \frac{\partial G}{\partial \theta_1}(t;\tilde{\theta}_1,\tilde{\theta}_2)(\hat{\theta}_1 - \theta_1) - \frac{\partial G}{\partial \theta_2}(t;\tilde{\theta}_1,\tilde{\theta}_2)(\hat{\theta}_2 - \theta_2) \right\}^2 dt \\
&= \frac{2}{\hat{\theta}_2} \left\{ \int_{\mathbb{R}} C_n(t)^2 \, dt - 2 \int_{\mathbb{R}} C_n(t) D_n(t) \, dt + \int_{\mathbb{R}} D_n(t)^2 \, dt \right\}, \qquad (4.19)
\end{aligned}$$

onde

$$C_n(t) = F_n(t) - G(t;\theta_1,\theta_2)$$

e

$$D_n(t) = \frac{\partial G}{\partial \theta_1}(t;\tilde{\theta}_1,\tilde{\theta}_2)(\hat{\theta}_1 - \theta_1) + \frac{\partial G}{\partial \theta_2}(t;\tilde{\theta}_1,\tilde{\theta}_2)(\hat{\theta}_2 - \theta_2).$$

Analisando o último termo de (4.19), pela condição (C3), pelas igualdades (4.9) e (4.10) e tendo em conta que $(a+b)^2 \leq 2(a^2+b^2)$, para todo o $a,b \in \mathbb{R}$, temos

$$\int_{\mathbb{R}} D_n(t)^2 \, dt \leq 2(\hat{\theta}_1 - \theta_1)^2 \int_{\mathbb{R}} \frac{\partial G}{\partial \theta_1}(t;\tilde{\theta}_1,\tilde{\theta}_2)^2 \, dt + 2(\hat{\theta}_2 - \theta_2)^2 \int_{\mathbb{R}} \frac{\partial G}{\partial \theta_2}(t;\tilde{\theta}_1,\tilde{\theta}_2)^2 \, dt = o_P(1). \qquad (4.20)$$

Relativamente ao segundo termo de (4.19), pela desigualdade de Cauchy-Schwarz sabemos que

$$\left| \int_{\mathbb{R}} C_n(t) D_n(t) \, dt \right| \leq \left(\int_{\mathbb{R}} C_n(t)^2 \, dt \right)^{1/2} \left(\int_{\mathbb{R}} D_n(t)^2 \, dt \right)^{1/2}, \qquad (4.21)$$

onde por (2.5) e pelas Proposições 2.1.1 e 2.2.1,

$$\int_{\mathbb{R}} C_n(t)^2 \, dt = \frac{1}{2}\mathscr{E}_n(X_1,\ldots,X_n) \xrightarrow{q.c.} \frac{1}{2}\int_{\mathbb{R}} \{F(t) - G(t;\theta_1,\theta_2)\}^2 \, dt, \tag{4.22}$$

isto é,

$$\int_{\mathbb{R}} C_n(t)^2 \, dt = O_P(1).$$

Deste modo, pela igualdade (4.20), concluímos que

$$\int_{\mathbb{R}} C_n(t) D_n(t) \, dt = o_P(1).$$

Assim,

$$\frac{T_n}{n} = \frac{2}{\hat{\theta}_2}\int_{\mathbb{R}} C_n(t)^2 \, dt + o_P(1),$$

onde, pela convergência (4.22) e tendo em conta que $\hat{\theta}_i \xrightarrow{p} \theta_i$, $i = 1,2$, concluímos que

$$\frac{T_n}{n} \xrightarrow{p} \frac{1}{\theta_2}\int_{\mathbb{R}} \{F(t) - G(t;\theta_1,\theta_2)\}^2 \, dt.$$

□

Teorema 4.6.1. *O teste de região crítica (3.4) é convergente para qualquer distribuição alternativa $F \in \mathscr{F} \setminus \mathscr{F}_0$, isto é,*

$$\lim_{n\longrightarrow+\infty} P_F(T_n > c_n(\alpha)) = 1, \text{ para todo o } F \in \mathscr{F} \setminus \mathscr{F}_0.$$

Demonstração. A demonstração é consequência do lema anterior e segue os mesmos passos do caso da hipótese nula simples (ver Teorema 3.3.1). □

Capítulo 5

Estudo de simulação

Esta dissertação termina com um estudo de simulação de carácter exploratório com o principal objetivo de avaliar a potência do teste estudado no capítulo anterior. Para tal, vamos abordar o caso das famílias de localização e escala normal e exponencial.

5.1 Teste de ajustamento a uma família normal de distribuições

Nesta secção, vamos abordar o teste de ajustamento a uma família das distribuições normais dada por

$$\mathscr{F}_0 = \left\{ F_0\left(\frac{\cdot - \mu}{\sigma}\right) : \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+ \right\},$$

onde F_0 representa a função de distribuição de uma lei normal standard. Para esta família prova-se que a condição (C1) é satisfeita.

Vamos considerar os estimadores obtidos pelo método da máxima verosimilhança como estimadores para μ e σ, ou seja,

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n$$

e

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}.$$

Notemos que estes estimadores têm um erro quadrático médio inferior aos estimadores do método dos momentos e, além disso, verificam a condição (C2).

Para verificar a condição (C2) para $\hat{\mu}$,

$$\sqrt{n}(\hat{\mu} - \mu) = \sqrt{n}(\overline{X}_n - \mu) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \Psi_1(X_i; \mu, \sigma),$$

onde, para todo o $x \in \mathbb{R}$,

$$\Psi_1(x; \mu, \sigma) = x - \mu$$

é tal que

$$E(\Psi_1(X; \mu, \sigma)) = 0 \text{ e } E(\Psi_1(X; \mu, \sigma)^2) < +\infty,$$

onde $X \sim F \in \mathscr{F}_0$.

Resta verificar que o estimador $\hat{\sigma}$ satisfaz a condição (C2). Pelo teorema do valor médio (Lima, 1992, Teorema 7, p. 213) aplicado à função $g(x) = \sqrt{x}$, temos

$$\sqrt{n}(\hat{\sigma} - \sigma) = \sqrt{n}(g(\hat{\sigma}^2) - g(\sigma^2)) = \frac{1}{2\tilde{\sigma}}\sqrt{n}(\hat{\sigma}^2 - \sigma^2), \tag{5.1}$$

para $\tilde{\sigma}$ entre σ e $\hat{\sigma}$.

Por outro lado,

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu + \mu - \overline{X}_n)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{2}{n}\sum_{i=1}^{n}(X_i - \mu)(\overline{X}_n - \mu) + (\overline{X}_n - \mu)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\overline{X}_n - \mu)^2,
\end{aligned}$$

o que, pelo teorema do limite central, implica que

$$\sqrt{n}(\hat{\sigma}^2 - \sigma^2) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}((X_i - \mu)^2 - \sigma^2) + o_P(1).$$

Desta igualdade e de (5.1), temos

$$\sqrt{n}(\hat{\sigma} - \sigma) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{2\sigma}((X_i - \mu)^2 - \sigma^2) + o_P(1),$$

e portanto,

$$\Psi_2(x;\mu,\sigma) = \frac{1}{2\sigma}((x-\mu)^2 - \sigma^2),$$

o que satisfaz

$$E(\Psi_2(X;\mu,\sigma)) = 0 \text{ e } E(\Psi_2(X;\mu,\sigma)^2) < +\infty.$$

Concluímos, assim, que os estimadores considerados verificam a condição (C2).

Atendendo à Proposição 3.9, a estatística de teste (4.3) é dada por

$$T_n = 2\sum_{i=1}^{n}\left(\hat{Y}_i\left(2F_0(\hat{Y}_i) - 1\right) + 2f_0(\hat{Y}_i)\right) - \frac{1}{n}\sum_{i,j=1}^{n}|\hat{Y}_i - \hat{Y}_j| - \frac{2n}{\sqrt{\pi}}, \tag{5.2}$$

com

$$\hat{Y}_i = \frac{X_i - \hat{\mu}}{\hat{\sigma}}.$$

5.1.1 Distribuições alternativas

De modo análogo ao teste de ajustamento a uma distribuição normal standard, vamos optar por escolher distribuições alternativas Student, qui-quadrado e lognormal, sendo que, ainda, optamos por testar uma distribuição alternativa logística. Assim, incluímos no nosso estudo, por um lado, distribuições

simétricas como as Student e a logística e, por outro lado, distribuições assimétricas como as qui-quadrado e as lognormal.

No caso das distribuições alternativas Student consideramos duas: uma com 10 graus de liberdade, t_{10}, e outra com 3 graus de liberdade, t_3.

Relativamente à distribuição alternativa logística apenas consideramos uma, uma vez que sendo os seus parâmetros de localização e escala, parâmetros diferentes não iriam produzir diferenças nos resultados de potência. Por isso, optamos por escolher parâmetros de modo que a média e o desvio padrão fossem 0 e 1, respetivamente. Assim, os parâmetros considerados são $\mu = 0$ e $s = \sqrt{3}/\pi$.

Por sua vez, nas distribuições alternativas qui-quadrado consideramos uma distribuição com 10 graus de liberdade, $\chi^2(10)$, e outra com 4 graus de liberdade, $\chi^2(4)$. Estas distribuições são assimétricas e apresentam coeficientes de assimetria 0.8944 e 1.4142 e de curtose 4.2 e 6, respetivamente.

Finalmente, relativamente às distribuições alternativas lognormal consideramos, em ambas, $\mu = 0$, variando apenas o parâmetro σ, considerando numa delas $\sigma = 0.2$, $LN(0,0.2)$, e noutra $\sigma = 0.5$, $LN(0,0.5)$. Tal como no caso anterior, estas distribuições são assimétricas e apresentam coeficientes de assimetria 0.6143 e 1.7502 e de curtose 3.6784 e 8.8984, respetivamente.

A função densidade de cada uma destas alternativas está representada na Figura 5.1 conjuntamente com a de F_0.

5.1.2 Resultados de potência

Atendendo à expressão da região crítica (3.4), uma vez que a distribuição exata de T_n sob a hipótese nula é desconhecida, foi necessário, em primeiro lugar, estimar os valores dos quantis $c_n(\alpha)$. Para tal, consideramos os níveis de significância usuais, 0.01 e 0.05, e os tamanhos de amostras $n = 25, 50, 100, 200, 400$ e calculamos o valor da estatística T_n, dada por (5.2), para 10^5 amostras desses tamanhos, geradas com distribuição F_0 onde determinamos uma aproximação de $c_n(\alpha)$ através da função `quantile(·,type=7)`. Por sua vez, tal como mencionado, a estimação dos parâmetros μ e σ foi feita com o método da máxima verosimilhança.

Tal como feito anteriormente, vamos tomar como referência o teste de ajustamento baseado na estatística de teste proposta por Anderson e Darling que designaremos, mais uma vez, por teste AD. Por outro lado, designaremos o teste em estudo por teste EN. De modo análogo ao teste EN, estimamos, para o teste AD, os quantis empíricos da estatística A_n^2.

Os resultados obtidos para os quantis empíricos das estatísticas T_n e A_n^2 estão apresentados na tabela seguinte.

$n\backslash\alpha$	0.01	0.05
25	1.1437	0.8293
50	1.1595	0.8386
100	1.1637	0.8411
200	1.1662	0.8427
400	1.1703	0.8439

(a)

$n\backslash\alpha$	0.01	0.05
25	1.0070	0.7391
50	1.0161	0.7488
100	1.0210	0.7458
200	1.0302	0.7515
400	1.0281	0.7530

(b)

Tabela 5.1 (a) Quantis empíricos de T_n; (b) Quantis empíricos de A_n^2.

Os resultados de potência apresentados nas Figuras 5.2 a 5.5, obtidos gerando 10^4 amostras de tamanho n de cada uma das alternativas, são relativos aos níveis de significância $\alpha = 0.01$ (círculo vazio)

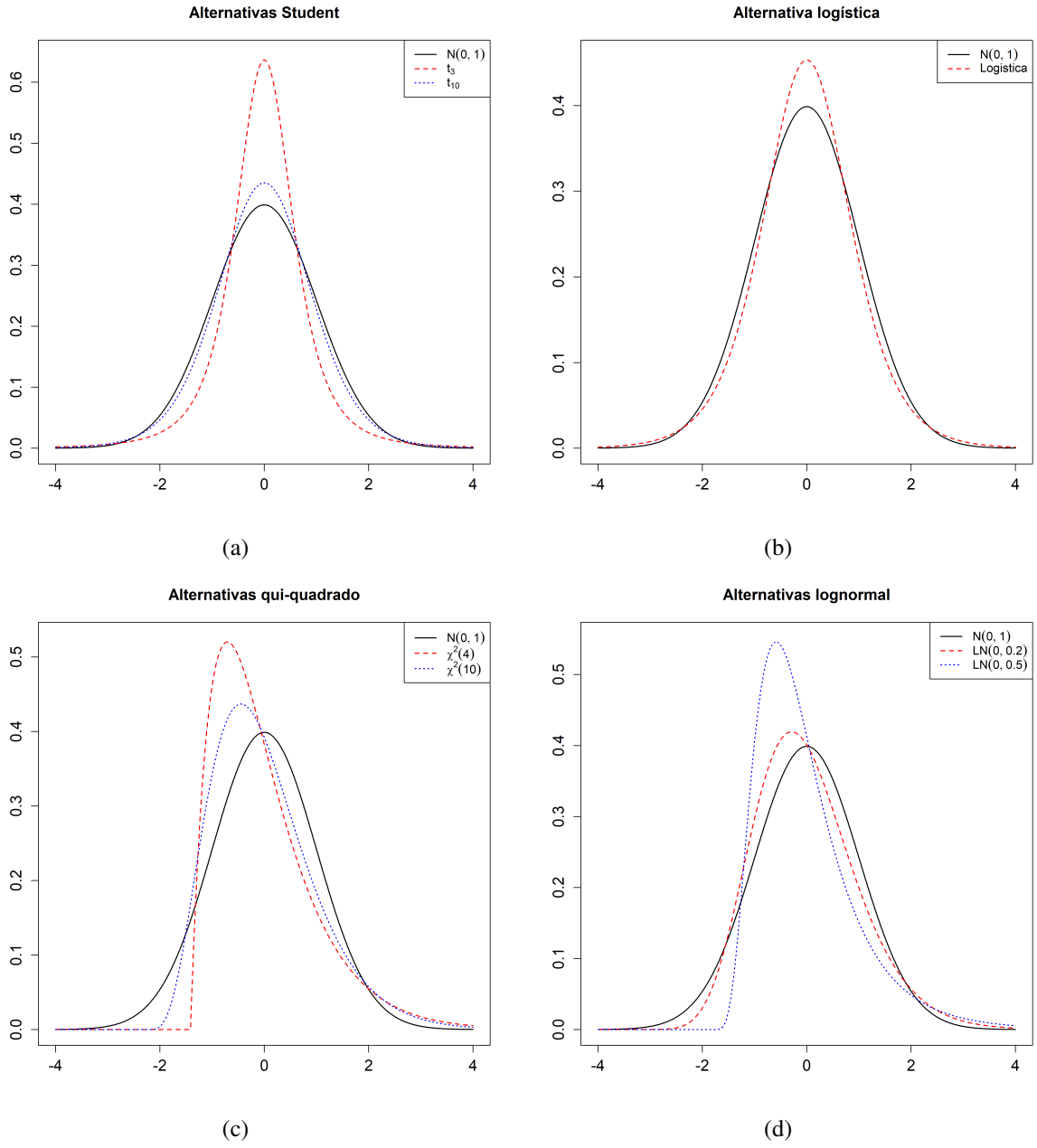
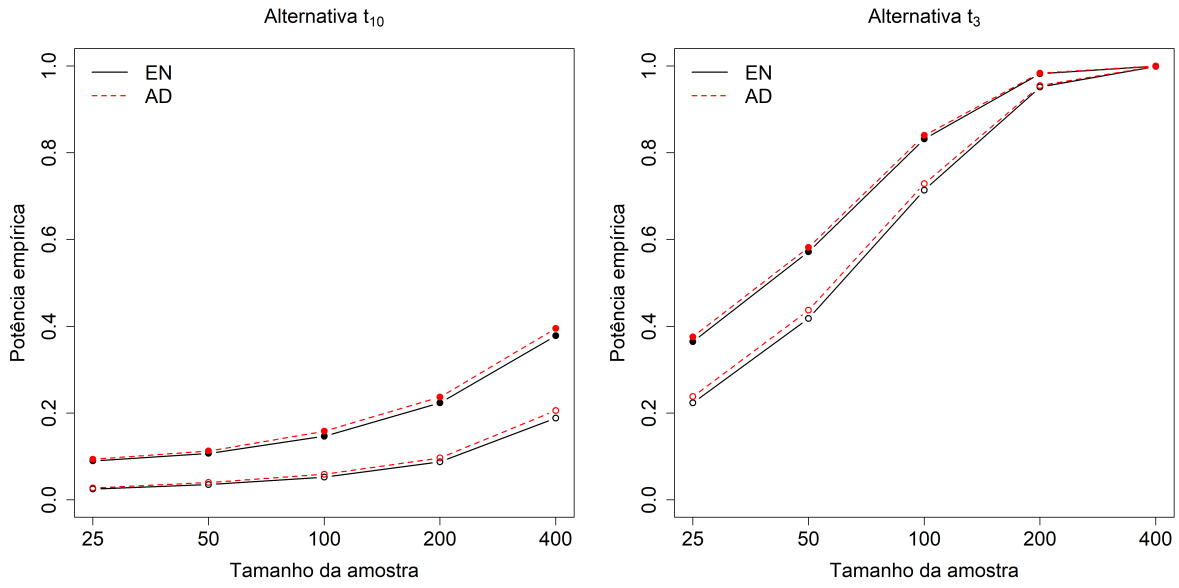


Figura 5.1 Densidades de probabilidade de cada distribuição alternativa considerada conjuntamente com a de F_0.

e $\alpha = 0.05$ (círculo cheio). Cada gráfico indica-nos a evolução da potência empírica de cada distribuição alternativa considerada para os testes EN e AD (tomado como referência) à medida que n aumenta. As Figuras 5.2 e 5.3 dizem respeito às alternativas simétricas.

Observando a Figura 5.2 nota-se grande diferença entre os resultados de potência das duas alternativas Student consideradas, onde concluímos que a alternativa t_3 apresenta resultados de potência bastante superiores à alternativa t_{10}, uma vez que a alternativa t_3 está bastante "mais afastada" que a alternativa t_{10} em relação à normal, como podemos observar pela Figura 5.1 (a).



(a) (b)

Figura 5.2 Resultados de potência para as distribuições alternativas Student ($\circ$: $\alpha = 0.01$; $\bullet$: $\alpha = 0.05$).

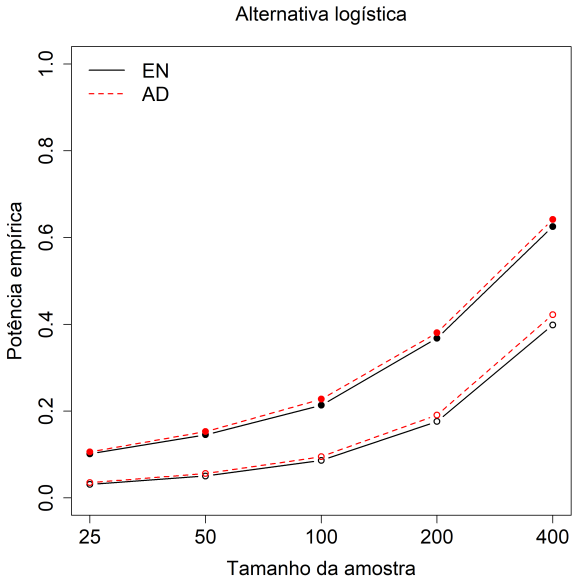
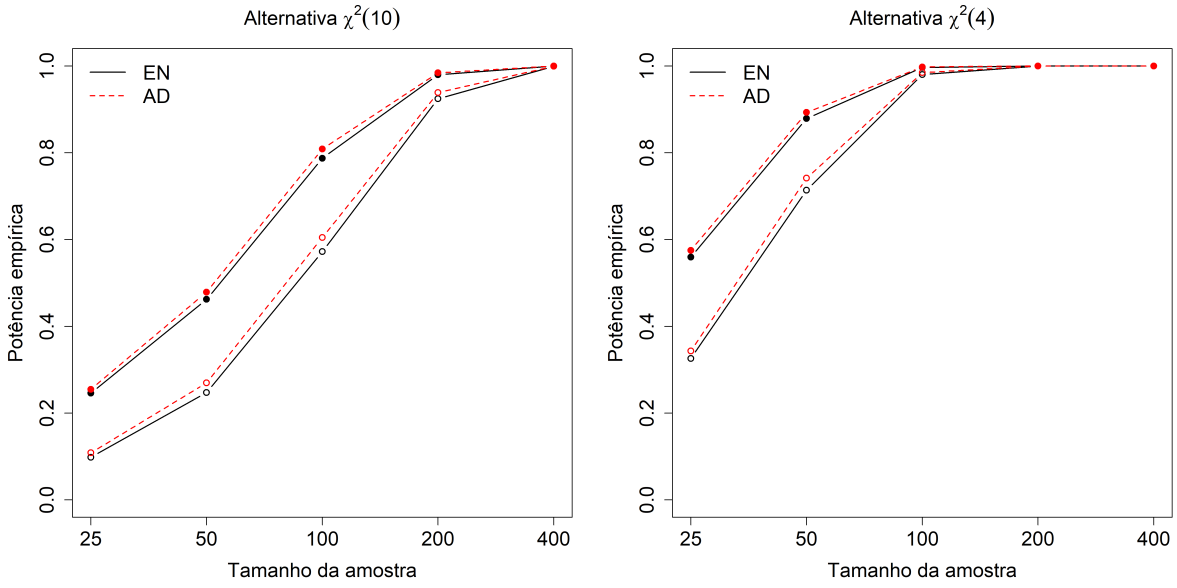


Figura 5.3 Resultados de potência para a distribuição alternativa logística ($\circ$: $\alpha = 0.01$; $\bullet$: $\alpha = 0.05$).

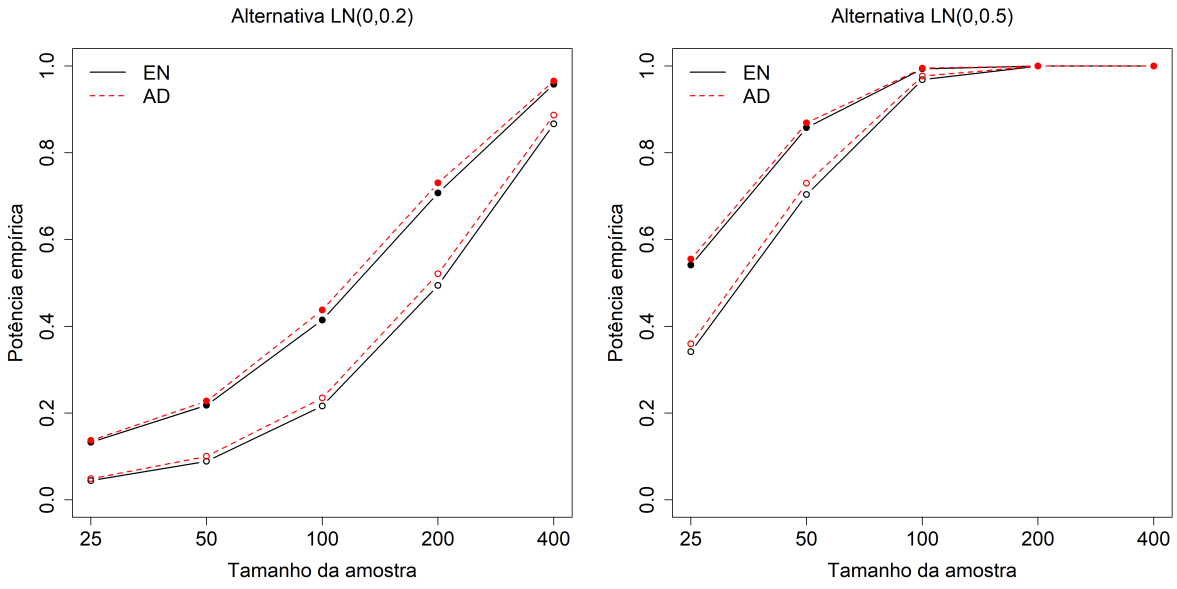
Relativamente à Figura 5.3, verificamos que a distribuição alternativa logística apresenta resultados de potência empírica não muito elevados, isto porque esta distribuição é relativamente próxima da distribuição normal, facto que observamos na Figura 5.1 (b).

As Figuras 5.4 e 5.5 dizem respeito às alternativas assimétricas qui-quadrado e lognormal, respetivamente. Observando a Figura 5.4 nota-se diferença entre os resultados de potência das duas alternativas qui-quadrado consideradas, onde concluímos que a alternativa $\chi^2(4)$ apresenta resultados de potência superiores à alternativa $\chi^2(10)$, uma vez que a alternativa $\chi^2(4)$ está "mais afastada" que a alternativa $\chi^2(10)$ em relação à normal, como podemos confirmar pela Figura 5.1 (c) e pelos valores do coeficiente de assimetria e da curtose para cada distribuição.



(a) (b)

Figura 5.4 Resultados de potência para as distribuições alternativas qui-quadrado ($\circ$: $\alpha = 0.01$; $\bullet$: $\alpha = 0.05$).



(a) (b)

Figura 5.5 Resultados de potência para as distribuições alternativas lognormal ($\circ$: $\alpha = 0.01$; $\bullet$: $\alpha = 0.05$).

Relativamente à Figura 5.5, verificamos que acontece algo análogo, isto é, à medida que a distribuição alternativa se vai afastando da distribuição normal, os resultados de potência vão aumentando, como podemos observar pela Figura 5.1 (d) e pelos valores do coeficiente de assimetria e da curtose das alternativas consideradas.

Por fim, observando desde a Figura 5.2 à 5.5 concluímos que o teste AD apresenta sempre valores de potência empírica superiores ao teste EN.

5.2 Teste de ajustamento a uma família exponencial de distribuições

Nesta secção vamos abordar o teste de ajustamento a uma família das distribuições exponenciais dada por

$$\mathscr{F}_0 = \left\{ F_0\left(\frac{\cdot - \mu}{\sigma}\right) : \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+ \right\},$$

onde, para todo o $x \in \mathbb{R}$,

$$F_0(x) = 1 - e^{-x}.$$

Para esta família a condição (C1) é satisfeita. Além disso, notemos que, se $X \sim F \in \mathscr{F}_0$, $E(X) = \mu + \sigma$ e $E((X - E(X))^2) = \sigma^2$.

Sejam $\hat{\mu}$ e $\hat{\sigma}$ os estimadores de μ e σ definidos por Johnson et al. (1994), isto é,

$$\hat{\mu} = \frac{nX_{(1)} - \overline{X}_n}{n-1}$$

e

$$\hat{\sigma} = \frac{n(\overline{X}_n - X_{(1)})}{n-1},$$

onde $X_{(1)}$ representa o mínimo da amostra $X_1, \ldots, X_n$. Estes estimadores são cêntricos de μ e σ, de variância mínima e verificam as condições (C1).

Provamos que os estimadores considerados verificam a condição (C2). Por um lado,

$$\sqrt{n}(\hat{\mu} - \mu) = \sqrt{n}(\hat{\mu} - E(\hat{\mu})) = \frac{\sqrt{n}}{n-1}\left(n(X_{(1)} - E(X_{(1)})) - (\overline{X}_n - E(\overline{X}_n))\right).$$

Uma vez que $\text{var}(X_{(1)}) = \sigma^2/n^2$ (Johnson et al. , 1994, p. 507) concluímos que

$$X_{(1)} - E(X_{(1)}) = O_P(n^{-1}).$$

Assim, pelo teorema do limite central

$$\sqrt{n}(\hat{\mu} - \mu) = O_P(n^{-1/2}) = o_P(1),$$

o que permite concluir que, para todo o $x \in \mathbb{R}$,

$$\Psi_1(x; \mu, \sigma) = 0,$$

verificando-se a condição (C2) para $\hat{\mu}$.

Para verificar a condição (C2) para o estimador $\hat{\sigma}$ notemos que

$$\begin{aligned}
\sqrt{n}(\hat{\sigma} - \sigma) &= \sqrt{n}(\hat{\sigma} - E(\hat{\sigma})) \\
&= \frac{n\sqrt{n}}{n-1}(\overline{X}_n - E(\overline{X}_n) - (X_{(1)} - E(X_{(1)})) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - (\mu + \sigma)) + o_P(1),
\end{aligned}$$

onde, para todo o $x \in \mathbb{R}$,

$$\Psi_2(x;\mu,\sigma) = x - (\mu + \sigma)$$

verificando-se

$$E(\Psi_2(X;\mu,\sigma)) = 0 \text{ e } E(\Psi_2(X;\mu,\sigma)^2) < +\infty.$$

Tal como no teste de ajustamento a uma família de distribuições normais, o cálculo da estatística T_n não levanta problemas, uma vez que os termos $E|x-X|$ e $E|X-X'|$ podem ser calculados de forma simples.

Proposição 5.2.1. *Seja X uma v.a.r. com função densidade $f_X(x) = \frac{1}{\sigma}\exp\left(-\frac{x-\mu}{\sigma}\right)\mathbb{1}_{[\mu,+\infty[}(x)$, com $\sigma \in \mathbb{R}^+$, então, para todo o $x \in \mathbb{R}$,*

$$E|x-X| = x - \mu + \sigma(1-2F_X(x))$$

e

$$E|X-X'| = \sigma.$$

Demonstração. Por um lado,

$$\begin{aligned}
E|x-X| &= \int_\mu^{+\infty} |x-y| f_X(y)\, dy \\
&= \int_\mu^x (x-y) f_X(y)\, dy + \int_x^{+\infty} (y-x) f_X(y)\, dy \\
&= \int_\mu^x x f_X(y)\, dy - \int_\mu^x y f_X(y)\, dy + \int_x^{+\infty} y f_X(y)\, dy - \int_x^{+\infty} x f_X(y)\, dy \\
&= xF_X(x) - \int_\mu^x y \frac{1}{\sigma}\exp\left(-\frac{y-\mu}{\sigma}\right) dy + \int_x^{+\infty} y \frac{1}{\sigma}\exp\left(-\frac{y-\mu}{\sigma}\right) dy - x(1-F_X(x)) \\
&= x(2F_X(x)-1) + \left(x\exp\left(-\frac{x-\mu}{\sigma}\right) - \mu - \int_\mu^x \exp\left(-\frac{y-\mu}{\sigma}\right) dy\right) \\
&\quad + \left(x\exp\left(-\frac{x-\mu}{\sigma}\right) + \int_x^{+\infty} \exp\left(-\frac{y-\mu}{\sigma}\right) dy\right) \\
&= x\left(1-2\exp\left(-\frac{x-\mu}{\sigma}\right)\right) + x\exp\left(-\frac{x-\mu}{\sigma}\right) - \mu + \sigma\exp\left(-\frac{x-\mu}{\sigma}\right) - \sigma \\
&\quad + x\exp\left(-\frac{x-\mu}{\sigma}\right) + \sigma\exp\left(-\frac{x-\mu}{\sigma}\right) \\
&= x - \mu + 2\sigma\exp\left(-\frac{x-\mu}{\sigma}\right) - \sigma \\
&= x - \mu + \sigma(1-2F_X(x)).
\end{aligned}$$

Por outro lado,

$$\begin{aligned}
E|X-X'| &= \int_\mu^{+\infty} \left(\int_\mu^{+\infty} |x-y| f_X(y)\, dy\right) f_X(x)\, dx \\
&= \int_\mu^{+\infty} (x - \mu + \sigma(1-2F_X(x))) f_X(x)\, dx
\end{aligned}$$

$$
\begin{aligned}
&= \int_{\mu}^{+\infty} x f_X(x)\, dx - \mu \int_{\mu}^{+\infty} f_X(x)\, dx + \sigma \int_{\mu}^{+\infty} f_X(x)\, dx - \sigma \int_{\mu}^{+\infty} 2F_X(x) f_X(x)\, dx \\
&= \mu + \sigma - \mu + \sigma - \sigma \\
&= \sigma.
\end{aligned}
$$

□

Atendendo à proposição anterior, a estatística de teste (4.3) é dada por

$$T_n = 2\sum_{i=1}^{n}(\hat{Y}_i + 1 - 2F_0(\hat{Y}_i)) - \frac{1}{n}\sum_{i,j=1}^{n}|\hat{Y}_i - \hat{Y}_j| - n, \tag{5.3}$$

com

$$\hat{Y}_i = \frac{X_i - \hat{\mu}}{\hat{\sigma}}.$$

5.2.1 Distribuições alternativas

O conjunto de distribuições alternativas escolhido teve por base as escolhas de Henze e Meintanis (2002), uma vez que se trata, também, de um teste de ajustamento a uma distribuição exponencial. Assim, optamos por escolher alternativas Weibull, gamma, lognormal e halfnormal.

No caso das distribuições alternativas Weibull, $W(k,\lambda)$, consideramos o parâmetro $\lambda = 1$, por ser de escala, variando apenas o parâmetro de forma, considerando numa delas $k = 1.25$, $W(1.25,1)$, e noutra $k = 1.8$, $W(1.8,1)$.

De modo análogo, nas alternativas gamma, mantivemos o parâmetro $\lambda = 1$, por ser de escala, variando o parâmetro de forma, considerando numa delas $k = 1.2$, $\Gamma(1.2,1)$, e noutra $k = 1.6$, $\Gamma(1.6,1)$.

Relativamente às alternativas lognormal consideramos, em ambas, $\mu = 0$, variando apenas o parâmetro σ, considerando numa delas $\sigma = 0.8$, $LN(0,0.8)$, e noutra $\sigma = 1.5$, $LN(0,1.5)$.

Por fim, consideramos a alternativa halfnormal, com o parâmetro $\theta = 1$, uma vez que é de escala, cuja densidade é dada por

$$f(x) = \sqrt{\frac{2}{\pi}} \exp\left(-\frac{x^2}{2}\right) \mathbb{1}_{[0,+\infty[}(x).$$

A função densidade de cada uma destas alternativas está representada na Figura 5.6 conjuntamente com a de F_0.

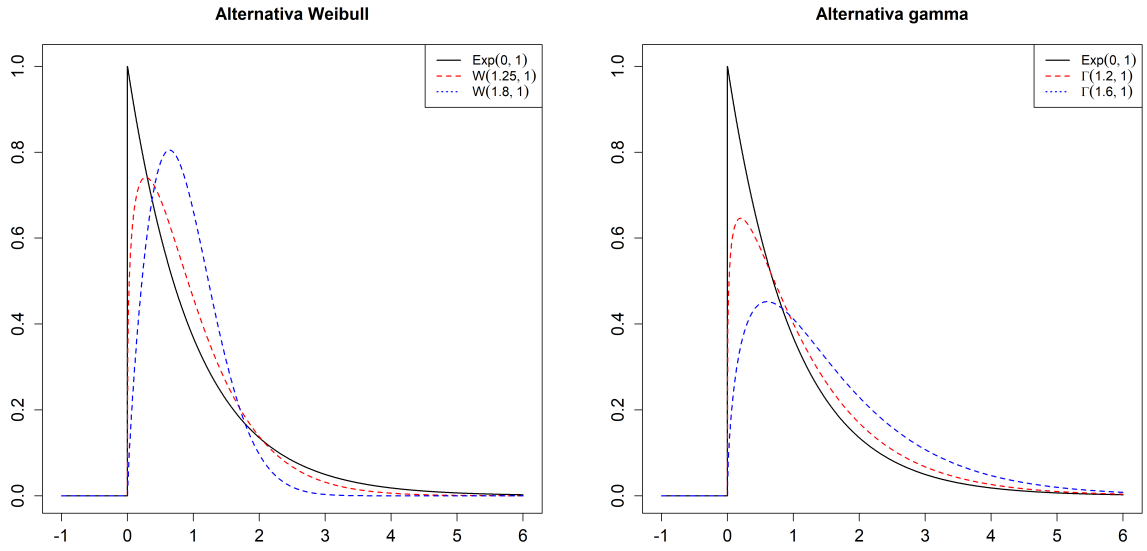
5.2.2 Resultados de potência

Com o objetivo de calcular a potência e de modo análogo à secção anterior, vamos em primeiro lugar estimar os valores dos quantis $c_n(\alpha)$, considerando os mesmos valores para os níveis de significância e para os tamanhos de amostra. Calculamos o valor da estatística T_n, dada por (5.3), para 10^5 amostras desses tamanhos, geradas com distribuição F_0 e determinamos uma aproximação para $c_n(\alpha)$. Por sua vez, estimamos μ e σ tal como referido no início desta secção.

De modo análogo ao teste EN, estimamos, para o teste AD, os quantis empíricos para a estatística A_n^2.

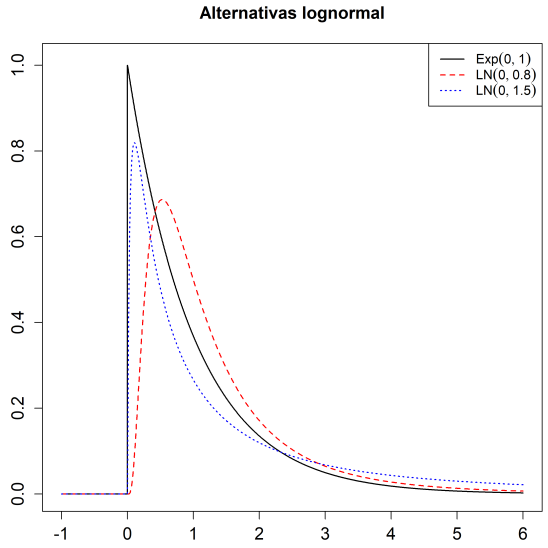
Os resultados obtidos para os quantis empíricos das estatísticas T_n e A_n^2 estão apresentados na Tabela 5.2.

Os resultados de potência apresentados nas Figuras 5.7 a 5.10, obtidos gerando 10^4 amostras de tamanho n de cada uma das alternativas, são relativos aos níveis de significância $\alpha = 0.01$ (círculo vazio)

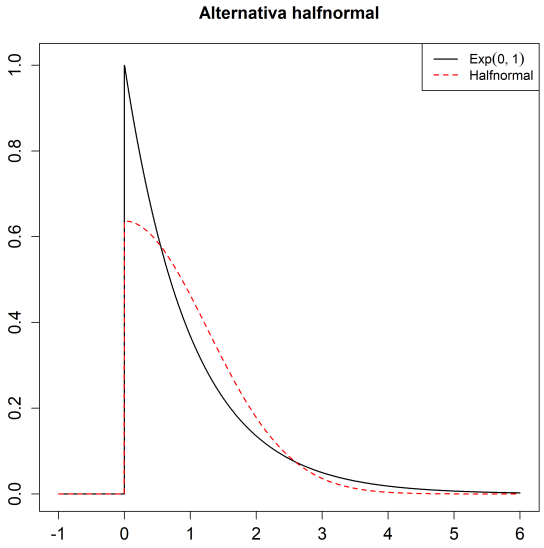


(a)

(b)



(c)



(d)

Figura 5.6 Densidades de probabilidade de cada distribuição alternativa considerada conjuntamente com a de F_0 .

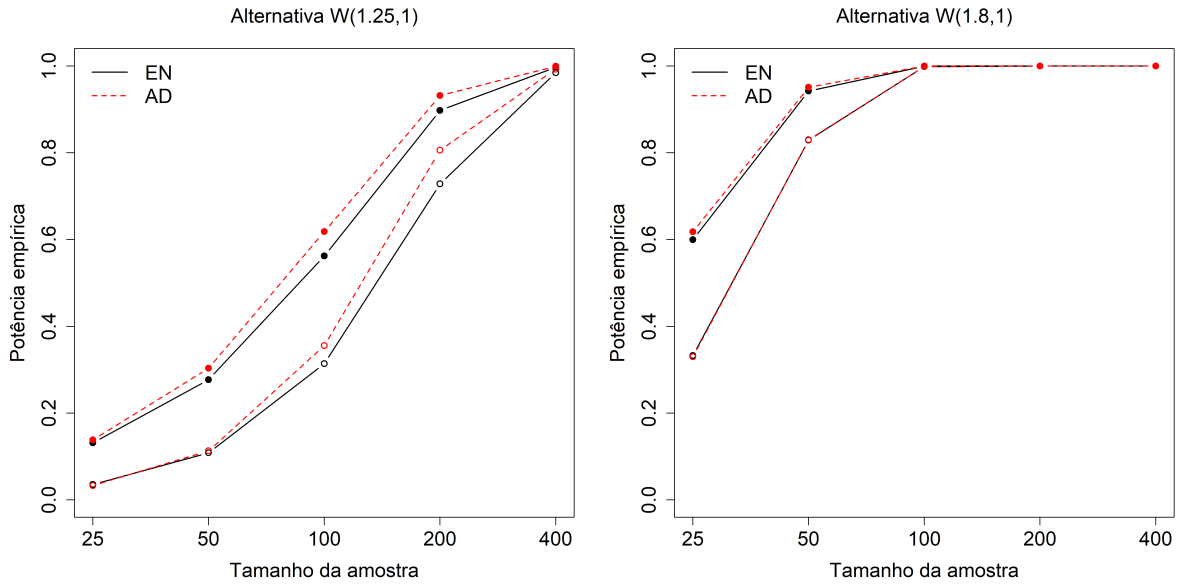
$n\backslash\alpha$	0.01	0.05
25	1.4816	1.0339
50	1.5266	1.0596
100	1.5660	1.0758
200	1.5832	1.0832
400	1.5713	1.0818

(a)

$n\backslash\alpha$	0.01	0.05
25	1.6373	1.1064
50	1.7749	1.1944
100	1.8632	1.2487
200	1.8948	1.2786
400	1.9377	1.3014

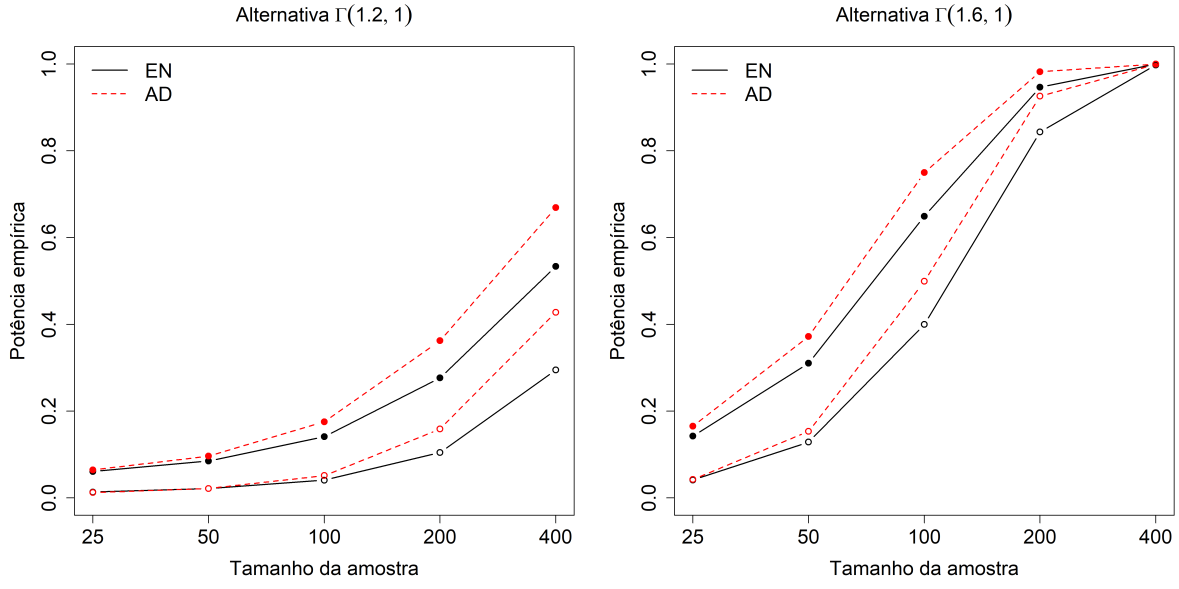
(b)

Tabela 5.2 (a) Quantis empíricos para T_n; (b) Quantis empíricos para A_n^2.



(a) (b)

Figura 5.7 Resultados de potência para as distribuições alternativas Weibull ($\circ$: $\alpha = 0.01$; $\bullet$: $\alpha = 0.05$).

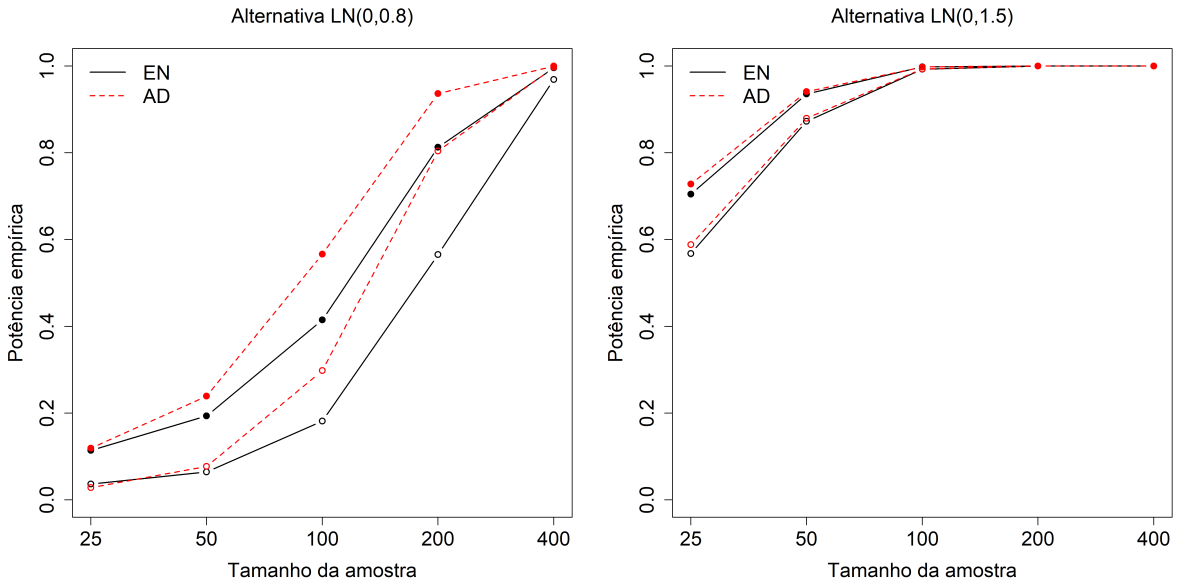


(a) (b)

Figura 5.8 Resultados de potência para as distribuições alternativas gamma ($\circ$: $\alpha = 0.01$; $\bullet$: $\alpha = 0.05$).

e $\alpha = 0.05$ (círculo cheio). Cada gráfico indica-nos a evolução da potência empírica de cada distribuição alternativa considerada para os testes EN e AD (tomado como referência) à medida que o tamanho da amostra aumenta.

Observando a Figura 5.7, nota-se diferença entre os resultados de potência das duas alternativas Weibull consideradas, onde concluímos que a alternativa $W(1.8,1)$ apresenta resultados de potência empírica superiores à alternativa $W(1.25,1)$, uma vez que a alternativa $W(1.8,1)$ está "mais afastada" em relação à exponencial, como podemos observar pela Figura 5.6 (a).



(a) (b)

Figura 5.9 Resultados de potência para as distribuições alternativas lognormal ($\circ$: $\alpha = 0.01$; $\bullet$: $\alpha = 0.05$).

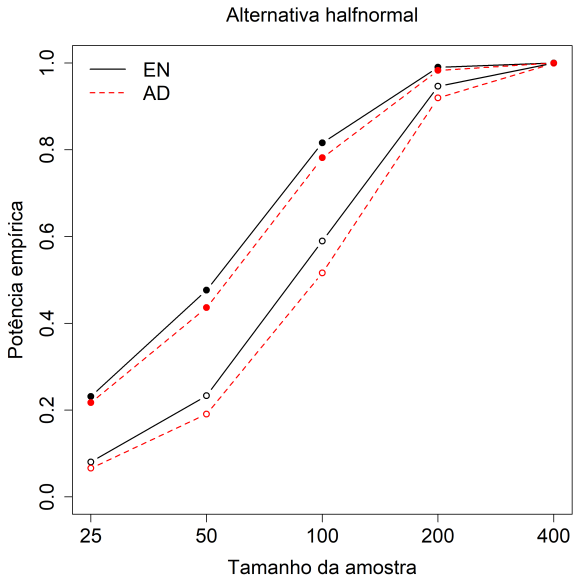


Figura 5.10 Resultados de potência para a distribuição alternativa halfnormal ($\circ$: $\alpha = 0.01$; $\bullet$: $\alpha = 0.05$).

Por sua vez, observando a Figura 5.8 vemos algo análogo a acontecer. A alternativa $\Gamma(1.6,1)$ apresenta resultados de potência empírica superiores à alternativa $\Gamma(1.2,1)$ uma vez que, pela Figura 5.6 (b), a alternativa $\Gamma(1.6,1)$ está "mais afastada" da exponencial do que $\Gamma(1.2,1)$.

Relativamente à Figura 5.9, a alternativa $LN(0,1.5)$ apresenta resultados de potência empírica superiores em relação à alternativa $LN(0,0.8)$. Basta notar pela Figura 5.6 (c) que a alternativa $LN(0,1.5)$ está "mais afastada" da exponencial do que a alternativa $LN(0,0.8)$.

Por fim, observando 5.10, além dos expectáveis resultados de potência que vão aumentando à medida que o tamanho da amostra aumenta, vemos que o teste EN apresenta, para todos os tamanhos de amostra,

resultados de potência superiores ao teste AD, sendo a única distribuição alternativa considerada onde tal acontece.

5.3 Discussão dos resultados e conclusão

Tal como seria de esperar pelo Teorema 4.6.1, para todas as distribuições alternativas consideradas relativamente aos testes de ajustamento às famílias de distribuições normais e exponenciais, verificamos que a potência aumenta à medida que o tamanho da amostra aumenta.

Notamos, ainda, que à medida que uma distribuição alternativa se vai afastando da distribuição da hipótese nula considerada, a potência também aumenta.

Observando os resultados obtidos, tanto no teste de ajustamento a uma família normal de distribuições como no teste de ajustamento a uma família exponencial de distribuições, concluímos que o teste AD é, regra geral, o que apresenta melhores resultados de potência, apesar de, para alguns tamanhos de amostra, não haver grandes diferenças entre os dois testes.

De facto, relativamente ao teste de ajustamento a uma família de distribuições normais, o teste AD apresenta sempre resultados de potência empírica superiores ao teste EN. Por sua vez, relativamente ao teste de ajustamento a uma família de distribuições exponenciais, apenas na alternativa halfnormal, o teste EN produz, para todos os tamanhos de amostra, resultados de potência empírica superiores ao teste AD. Contudo, reparemos que para algumas outras distribuições alternativas e para o tamanho de amostra $n = 25$, verificamos que o teste EN pode apresentar resultados de potência superiores ao teste AD, no entanto, à medida que o tamanho da amostra aumenta, o teste AD supera os valores da potência empírica do teste EN.

Os valores dos resultados de potência obtidos para os testes de ajustamento a famílias de distribuições normais e exponenciais estão apresentados no Anexo B.

Bibliografia

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

Henze, N., Meintanis, S. (2002). Tests of fit for exponentiality based on the empirical laplace transform. *Statistics* 36, 147–161.

Johnson, N., Kotz, S., Balakrishnan, N. (1994). *Continuous univariate distributions*. New York: John Wiley & Sons.

Lee, A.J. (1990). *U-statistics, theory and practice*. New York: Marcel Dekker.

Lima, E.L. (1992). *Curso de análise, vol. 1*. Rio de Janeiro: Projeto Euclides.

Métivier, M. (1972). *Notions foundamentales de la théorie des probabilités*. Paris: Dunod.

Moore, D.S. (1986). Tests of chi-squared type. *In Goodness-of-fit techniques*, D'Agostino, R.B., Stephens, M.A. (Eds), pp. 63–95. New York: Marcel Dekker.

Móri, T.F., Székely, G.J., Rizzo, M.L. (2021). On energy tests for normality. *J. Statist. Plann. Inference* 213, 1–15.

Shorack, G.R., Wellner, J. (1986). *Empirical processes with applications to statistics*. New York: John Wiley & Sons.

Stephens, M.A. (1986). Tests based on EDF statistics. *In Goodness-of-fit techniques*, D'Agostino, R.B., Stephens, M.A. (Eds), pp. 97–194, New York: Marcel Dekker.

Székely, G.J., Rizzo, M.L. (2005). A new test for multivariate normality. *J. Multivariate Anal.* 93, 58–80.

Székely, G.J., Rizzo, M.L. (2005). Hierarchical clustering via joint between-within distances: extending Ward's minimum variance method. *J. Classification* 22, 151–183.

Székely, G.J., Rizzo, M.L. (2013). Energy statistics: a class of statistics based on distances *J. Statist. Plann. Inference* 143, 1249–1272.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL *https://www.R-project.org/*.

van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge: Cambridge University Press.

Anexo A

Resultados auxiliares

Sejam $X_1, \ldots, X_n$ v.a.r. identicamente distribuídas com X com função de distribuição F. Consideremos a U-estatística de núcleo h definida por

$$U_n = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} h(X_i, X_j)$$

Nos resultados seguintes estabelecemos a convergência das U-estatísticas. As suas demonstrações encontram-se em Lee (1990, pp. 80 e 122) e em Gregory (1977).

Teorema A.1.1. *Se* $E|h(X,X')| < \infty$*, então*

$$U_n \xrightarrow{q.c.} E(h(X,X')).$$

Para estabelecer a convergência em distribuição da U-estatística U_n são necessárias as seguintes condições:

- $h(x,y) = h(y,x)$, para todo o $x, y \in \mathbb{R}$;
- $E(h(X,X')^2) = \iint_{\mathbb{R}^2} h(x,y)^2 \, dF(x) dF(y) < \infty$;
- $E(h(x,X)) = \int_{\mathbb{R}} h(x,y) \, dF(y) = 0.$

Se a U-estatística verificar esta última condição dizemos que U_n é uma U-estatística degenerada.

Teorema A.1.2. *Nas condições anteriores temos que*

$$nU_n \xrightarrow{d} \sum_{k=1}^{+\infty} \lambda_k (Z_k^2 - 1),$$

onde Z_k*,* $k \geq 1$*, são v.a.r. i.i.d. com lei normal standard e* λ_k*,* $k \geq 1$*, com* $\sum_{k=1}^{+\infty} \lambda_k^2 < \infty$*, são os valores próprios associados ao operador* A_h *definido de* $L^2(\mathbb{R}, F)$ *em* $L^2(\mathbb{R}, F)$ *por*

$$A_h g(x) = \int_{\mathbb{R}} h(x,y) g(y) \, dF(y), \; x \in \mathbb{R},$$

para $g \in L^2(\mathbb{R}, F)$, onde $L^2(\mathbb{R}, F)$ é o espaço das funções reais de variável real g, tais que $\int_{\mathbb{R}} g(y)^2 \, dF(y) < +\infty$. Além disso, se $\int_{\mathbb{R}} g(x)(A_h g)(x) \, dF(x) \geq 0$, para todo o $g \in L^2(\mathbb{R}, F)$, então os valores próprios λ_k são não negativos.

Anexo B

Potência empírica dos testes EN e AD

Nas tabelas que se seguem apresentamos os resultados de potência dos testes EN e AD obtidos para cada alternativa considerada nas secções 3.4, 5.1 e 5.2.

1. Teste de ajustamento a uma distribuição normal standard

n	25		50		100		200	
Lei\Teste	EN	AD	EN	AD	EN	AD	EN	AD
$\mathcal{N}(0,1)$	0.0485	0.05	0.0507	0.0510	0.0553	0.0553	0.0516	0.0513
$ST1$	0.0625	0.0853	0.2124	0.2583	0.6264	0.6857	0.9717	0.9816
$ST2$	0.0986	0.1260	0.3715	0.4228	0.8572	0.8847	0.9986	0.9991
$ST3$	0.1969	0.2286	0.6774	0.7148	0.9870	0.9908	1	1
$QQ1$	0.1260	0.1383	0.2499	0.2824	0.5765	0.6582	0.9699	0.9902
$QQ2$	0.1628	0.1809	0.3568	0.4045	0.7916	0.8569	0.9998	1
$QQ3$	0.2503	0.2859	0.5970	0.6612	0.9816	0.9959	1	1
$LN1$	0.0719	0.0769	0.1088	0.1153	0.1958	0.2150	0.4213	0.4801
$LN2$	0.1249	0.1463	0.2734	0.3090	0.6303	0.7131	0.9826	0.9950
$LN3$	0.2438	0.2862	0.6233	0.6951	0.9892	0.9967	1	1

Tabela B.1 Resultados de potência para $\alpha = 0.05$.

2. Teste de ajustamento a uma família normal de distribuições

n	25		50		100		200		400	
Lei\Teste	EN	AD	EN	AD	EN	AD	EN	AD	EN	AD
$\mathcal{N}(0,1)$	0.011	0.011	0.010	0.010	0.010	0.011	0.010	0.010	0.011	0.011
t_{10}	0.025	0.028	0.035	0.040	0.053	0.059	0.088	0.097	0.188	0.206
t_3	0.224	0.238	0.418	0.437	0.714	0.729	0.951	0.955	0.999	0.999
$Logistica$	0.031	0.035	0.050	0.056	0.087	0.095	0.176	0.191	0.399	0.422
$\chi^2(10)$	0.098	0.109	0.248	0.267	0.572	0.605	0.925	0.939	0.999	1
$\chi^2(4)$	0.326	0.343	0.714	0.7412	0.980	0.985	1	1	1	1
$LN(0,0.2)$	0.044	0.049	0.089	0.100	0.217	0.235	0.494	0.521	0.867	0.887
$LN(0,0.5)$	0.342	0.360	0.704	0.730	0.969	0.977	1	1	1	1

Tabela B.2 Resultados de potência para $\alpha = 0.01$.

n	25		50		100		200		400	
Lei\Teste	EN	AD	EN	AD	EN	AD	EN	AD	EN	AD
$\mathcal{N}(0,1)$	0.053	0.053	0.049	0.049	0.052	0.053	0.054	0.054	0.050	0.051
t_{10}	0.090	0.094	0.107	0.113	0.147	0.158	0.224	0.237	0.378	0.395
t_3	0.365	0.376	0.572	0.582	0.832	0.840	0.982	0.984	1	1
$Logistica$	0.101	0.106	0.145	0.153	0.213	0.228	0.368	0.381	0.626	0.642
$\chi^2(10)$	0.246	0.255	0.462	0.479	0.788	0.809	0.980	0.984	1	1
$\chi^2(4)$	0.559	0.575	0.879	0.893	0.997	0.998	1	1	1	1
$LN(0,0.2)$	0.133	0.137	0.218	0.228	0.415	0.438	0.707	0.731	0.958	0.965
$LN(0,0.5)$	0.541	0.555	0.858	0.869	0.993	0.995	1	1	1	1

Tabela B.3 Resultados de potência para $\alpha = 0.05$.

3. Teste de ajustamento a uma família exponencial de distribuições

n	25		50		100		200		400	
Lei\Teste	EN	AD	EN	AD	EN	AD	EN	AD	EN	AD
$\mathcal{E}(0,1)$	0.010	0.010	0.011	0.009	0.009	0.011	0.011	0.012	0.010	0.010
$W(1.25,1)$	0.035	0.033	0.109	0.113	0.314	0.356	0.728	0.806	0.985	0.993
$W(1.8,1)$	0.333	0.300	0.829	0.830	0.998	0.999	1	1	1	1
$\Gamma(1.2,1)$	0.013	0.012	0.022	0.022	0.041	0.051	0.105	0.159	0.295	0.428
$\Gamma(1.6,1)$	0.041	0.042	0.129	0.153	0.400	0.499	0.844	0.926	0.998	0.999
$LN(0,0.8)$	0.037	0.028	0.064	0.077	0.182	0.298	0.565	0.804	0.969	0.997
$LN(0,1.5)$	0.568	0.588	0.873	0.880	993	0.993	1	1	1	1
$Halfnormal$	0.080	0.066	0.233	0.191	0.590	0.516	0.946	0.920	1	0.999

Tabela B.4 Resultados de potência para $\alpha = 0.01$.

n	25		50		100		200		400	
Lei\Teste	EN	AD	EN	AD	EN	AD	EN	AD	EN	AD
$\mathcal{E}(0,1)$	0.046	0.050	0.050	0.050	0.051	0.049	0.051	0.049	0.049	0.049
$W(1.25,1)$	0.132	0.138	0.277	0.303	0.532	0.619	0.898	0.932	0.997	0.999
$W(1.8,1)$	0.600	0.618	0.942	0.951	0.999	0.999	1	1	1	1
$\Gamma(1.2,1)$	0.061	0.065	0.085	0.096	0.141	0.178	0.277	0.362	0.533	0.669
$\Gamma(1.6,1)$	0.143	0.165	0.310	0.372	0.649	0.750	0.947	0.982	0.999	1
$LN(0,0.8)$	0.114	0.119	0.194	0.239	0.415	0.566	0.813	0.936	0.996	1
$LN(0,1.5)$	0.705	0.728	0.936	0.941	0.998	0.998	1	1	1	1
$Halfnormal$	0.232	0.217	0.477	0.436	0.816	0.782	0.990	0.983	1	1

Tabela B.5 Resultados de potência para $\alpha = 0.05$.

Anexo C

Códigos R usados no estudo de simulação

1. Código para a estimação dos quantis:

```
# Calculo da estatistica de teste
F1 = function(x,mu,sig)
{
  2*(x-mu)*pnorm(x,mean = mu, sd = sig)+2*sig^2*dnorm(x,mean = mu, sd = sig)-(x
      -mu)
}

Tn = function(s,mu,sig)
{
  n <- length(s)
  2*sum(F1(s,mu,sig)) - sum(abs(outer(s,s,"-")))/n - n*2*sig/sqrt(pi)
}


# N(mu,sig^2)
mu <- 0
sig <- 1

#Valores da estistica sob H0
valoresEN=function(n,rep=10^5)
{
  val <- array(dim=rep)

  for (i in 1:rep)
  {
    # Amostra proveniente de uma lei normal (0,1) de tamanho n
    s <- rnorm(n)
    val[i] <- Tn(s,mu,sig)
  }


  if (n<100) nn<- paste(0,n,sep="") else nn<-n
  escrever<-paste("Estatistica_EN_n", nn, ".txt", sep="")
  write.table(format(val,scientific=TRUE,digits=10),file=escrever,col.names=
      FALSE,  row.names=FALSE)
}
```

```
#Tamanho da amostra
ns<-c(25,50,100,200)
lns<-length(ns)

#Niveis de significancia a avaliar
alpha<-c(0.05,0.02,0.01)
lalpha <- length(alpha)
q5<-array(dim=lns)
q2<-array(dim=lns)
q1<-array(dim=lns)

for (i in 1:lns)
{
  if (ns[i]<100) nn<- paste(0, ns[i], sep= "") else nn <- ns[i]
  valores <-read.table(paste("Estatistica_EN_n", nn, ".txt", sep=""))

  q<-quantile(valores[,1], 1-alpha, names=FALSE)
  q5[i]<-q[1]
  q2[i]<-q[2]
  q1[i]<-q[3]
}

write.table(q5, file=paste("quantis5_EN.txt", sep=""), col.names=FALSE, row.
    names=FALSE)
write.table(q2, file=paste("quantis2_EN.txt", sep=""), col.names=FALSE, row.
    names=FALSE)
write.table(q1, file=paste("quantis1_EN.txt", sep=""), col.names=FALSE, row.
    names=FALSE)


# Estatistica de AD (caso Uniforme)

ad.stat = function(z,n)
{
  z <- sort(z)
  -n-sum( (2*(1:n)-1)*log(z) + (2*n+1-2*(1:n))*log(1-z) )/n
}

valoresAD=function(n,rep=10^5)
{
  valAD<-array(dim=rep)

  for (i in 1:rep)
  {
    # Amostra proveniente de uma lei normal (0,1) de tamanho n
    s <- rnorm(n)
    z<-pnorm(s, mean=0, sd=1)
    valAD[i] <- ad.stat(z,n)
  }

  if(n<100) nn<- paste(0, n, sep="") else nn<-n
  escrever<- paste("Estatistica_AD_n_", nn, ".txt", sep="")
```

```
  write.table(format(valAD, scientific=TRUE, digits=10), file=escrever, col.
      names=FALSE, row.names=FALSE)
}

q5AD<-array(dim=lns)
q2AD<-array(dim=lns)
q1AD<-array(dim=lns)

for (i in 1:lns)
{
  if (ns[i]<100) nn<- paste(0, ns[i], sep="") else nn <- ns[i]
  valores<-read.table(paste("Estatistica_AD_n_", nn, ".txt", sep=""))

        #Valor dos quantis
  q<- quantile(valores[,1], 1-alpha, names=FALSE)
  q5AD[i]<-q[1]
  q2AD[i]<-q[2]
  q1AD[i]<-q[3]
}

write.table(q5AD, file=paste("quantis5_AD.txt", sep=""), col.names=FALSE, row.
    names=FALSE)
write.table(q2AD, file=paste("quantis2_AD.txt", sep=""), col.names=FALSE, row.
    names=FALSE)
write.table(q1AD, file=paste("quantis1_AD.txt", sep=""), col.names=FALSE, row.
    names=FALSE)
```

2. Código para o cálculo das potências de cada teste:

```
#Calculo da potencia do teste em estudo e do teste AD

#Funcao que contempla todas as distribuicoes em estudo
potencia = function(d,n,rep=10^4)
{
  if (d==0) alternativa <- function(n){return(rnorm(n))}
  if (d==1) alternativa <- function(n){return(rt(n,df=2.5)/sqrt(5))}
  if (d==2) alternativa <- function(n){return(rt(n,df=2.75)/sqrt(11/3))}
  if (d==3) alternativa <- function(n){return(rt(n,df=3)/sqrt(3))}
  if (d==4) alternativa <- function(n){return((rchisq(n,df=2)-2)/sqrt(4))}
  if (d==5) alternativa <- function(n){return((rchisq(n,df=3)-3)/sqrt(6))}
  if (d==6) alternativa <- function(n){return((rchisq(n,df=4)-4)/sqrt(8))}
  if (d==7) alternativa <- function(n){return((rlnorm(n, meanlog = 0, sdlog =
      0.3)-exp(0.3^2/2))/sqrt((exp(0.3^2)-1)*exp(0.3^2)))}
  if (d==8) alternativa <- function(n){return((rlnorm(n, meanlog = 0, sdlog =
      0.5)-exp(0.5^2/2))/sqrt((exp(0.5^2)-1)*exp(0.5^2)))}
  if (d==9) alternativa <- function(n){return((rlnorm(n, meanlog = 0, sdlog =
      0.7)-exp(0.7^2/2))/sqrt((exp(0.7^2)-1)*exp(0.7^2)))}

  if (n<100) nn <- paste(0,n,sep="") else nn <- n
  if (d<10) dd <- paste(0,d,sep="") else dd <- d

  q5EN <- read.table(paste("quantis5_EN.txt", sep=""))
  q5AD <- read.table(paste("quantis5_AD.txt", sep=""))
```

```
22
23    ns <- c(25,50,100,200)
24    ind=1
25    while(ns[ind] != n) ind <- ind + 1
26
27    q5nEN <- q5EN[ind,]
28    q5nAD <- q5AD[ind,]
29
30    set.seed(2000, kind=NULL)
31
32    VTn<-array(dim=rep)
33    VAD<-array(dim=rep)
34
35    for (i in 1:rep)
36    {
37      s <- alternativa(n)
38      z<-pnorm(s, mean=0, sd=1)
39      VTn[i] <- Tn(s,mu=0,sig=1)
40      VAD[i]<-ad.stat(z,n)
41    }
42
43    potenciaEN <- sum(VTn > q5nEN)/rep  #Potencia teste EN
44    potenciaAD <- sum(VAD > q5nAD)/rep  #Potencia teste AD
45
46    pot<-matrix(c(potenciaEN,potenciaAD))
47    escrever <- paste("potencia_d", dd, "n", nn, ".txt", sep="")
48    write.table(format(pot, scientific=TRUE), file=escrever, col.names=FALSE, row
          .names=FALSE)
49  }
```