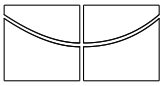


Universidade de Brasília

Instituto de Ciências Exatas

Departamento de Estatística

Dissertação de Mestrado

Testes de Independência "Distribution-Free"

por

Loyane Christina Soares Rocha

Orientador: Prof. Dr. Raul Yukihiro Matsushita

2014

Loyane Christina Soares Rocha

Testes de Independência "Distribution-Free"

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidadede de Brasília como requisito parcial à obtenção do título de Mestre em Estatística.

Universidade de Brasília

Brasília, 2014

Termo de Aprovação

Loyane Christina Soares Rocha

Testes de Independência "Distribution-Free"

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidadede de Brasília como requisito parcial à obtenção do título de Mestre em Estatística.

Data da defesa: 02 de julho de 2014

Orientador:

Prof. Raul Yukihiro Matsushita, Dr.
Departamento de Estatística, UnB

Comissão Examinadora:

Prof. Eraldo Sergio Barbosa da Silva, PhD
Departamento de Economia, UFSC

Prof. Peter Zörnig, PhD
Departamento de Estatística, UnB

Brasília, 2014

Ficha Catalográfica

ROCHA, LOYANE CHRISTINA SOARES

Testes de Independência "Distribution-Free", (UnB - IE, Mestre em Estatística, 2014).

Dissertação de Mestrado - Universidade de Brasília. Departamento de Estatística - Instituto de Ciências Exatas.

1. Dependência não linear 2. Teste *distribution-free*

Aos meus pais,

ao meu esposo,

e às minhas irmãs.

Agradecimentos

Agradeço a Deus pelos caminhos traçados e pela sua infinita bondade para comigo;
Aos meus pais, que sempre fizeram o possível e não mediram esforços para que eu obtivesse êxito em meus estudos;
Ao meu esposo, que com muita paciência tem sempre me apoiado, me encorajando e dando suporte naquilo que estava a seu alcance. Tem sido meu grande amigo, companheiro, transmitindo alegria, amor e paz;
Ao meu orientador, Raul, meus sinceros agradecimentos por sua grande compreensão, prestatividade, apoio e pelo seu excelente trabalho;
Meus agradecimentos aos demais professores da banca, que se dispuseram a contruibuir para o desenvolvimento do trabalho;
Não posso deixar de citar, também, os gestores do local em que trabalho, pela compreensão demonstrada.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

Este trabalho trata de testes estatísticos não paramétricos para a detecção de dependência não-linear entre duas variáveis. Estudos de Monte Carlo foram realizados para avaliar e comparar o desempenho de testes do tipo *distribution-free*. Foram considerados testes que se baseiam no critério de Cramér-von Mises e também uma variação do teste de Kolmogorov-Smirnov (KS). Os resultados mostram que o teste proposto por Matsushita et al. (2012) e o de KS apresentam bom poder para detecção de estruturas de dependência não-linear.

Palavras Chave: *Dependência não linear, Teste distribution-free.*

Abstract

This dissertation deals with nonparametric tests for the detection of nonlinear dependence between two variables. Monte Carlo studies were performed to evaluate and compare the performance of some distribution-free tests. Here, we considered test that are based on the criterion of Cramér-von Mises and also a variation of the Kolmogorov-Smirnov (KS) test. The results show that the our new suggested tests have good power to detect nonlinear bivariate dependence.

key words: *Nonlinear dependence, Distribution-free test.*

Abreviações e Siglas

a.a	amostra aleatória
BKR	estatística do teste de HBKR
CHI	estatística do teste com distribuição assintótica igual ao LR
Corr	correlação
$CORR$	estatística do teste baseado na correlação de Pearson
Cov	covariância
FC	função característica
FDA	função de distribuição acumulada
FDE	função de distribuição empírica
gl	graus de liberdade
H_0	hipótese nula
H_a	hipótese alternativa
HBKR	(teste de dependência de) Hoeffding, Blum, Kiefer e Rosenblatt
KAC	estatística do teste proposto por Kac (1951)
KS	estatística do teste de kolmogorov-Smirnov
LR	estatística do teste proposto por Matshushita et.al(2012)
N_c	n° pares concordantes
N_d	n° pares discordantes
ns	nível de significância
$P - value$	nível descritivo de um teste de hipóteses
sup	supremum
$v.a_s$	variáveis aleatórias
Var	variância

Lista de Símbolos e Notações

$\delta(x, y)$	índice de independência
D	medida D de Hoeffding
E(.)	esperança
f(x)	função densidade
F(x)	função de distribuição
R_{xy}	matriz de correlação
$\hat{F}(.)$	função de distribuição empírica
Σ	matriz de covariância
ρ	coeficiente de correlação populacional
ρ^2	coeficiente de determinação
ρ^*	coeficiente de correlação máxima
r	coeficiente de correlação amostral
r_s	coeficiente de Spearman
r_k	coeficiente de Kendall
R(.)	postos
σ	desvio-padrão
Φ^2	medida de Hoeffding
$\Phi_x(q; \theta)$	função característica
χ^2	medida de associação qui-quadrado
o_{ij}	frequência observada
e_{ij}	frequência esperada

Introdução

Os testes para avaliação de independência entre realizações de variáveis aleatórias (v.as.) absolutamente contínuas são fundamentais na análise estatística de dados. Eles comumente se aplicam, por exemplo, na fase de diagnósticos de uma análise de regressão ou de séries temporais — para a avaliação da independência residual, ou entre as covariáveis do modelo (Brockwell e Davis, 2006; Draper e Smith, 1998). Eles também são úteis para casos em que a própria estrutura de dependência seja objeto da modelagem, como em análise de séries temporais (Brockwell e Davis, 2006), análise de dados longitudinais (Diggle et al., 1996), estatística espacial (Cressie, 1993) e cópulas (Nelsen, 2006).

No ambiente gaussiano e linear — sob o qual muitos modelos se fundamentam — a dependência pode ser medida e avaliada com base na correlação linear de Pearson ou de suas variações, como a função de autocorrelação, a função de autocorrelação parcial e o semivariograma. Isso porque, nesse ambiente, a correlação nula implica independência. Porém, fora desse ambiente, em que a correlação nula não implica independência, outras medidas estatísticas são necessárias.

Embora o problema de se testar a independência multivariada entre m variáveis seja relativamente antigo, remetendo-nos a trabalhos a partir da década de 1950 (Ghoudi et al., 2001), a oferta desses testes multivariados ainda é pequena (Beran et al., 2006), principalmente aqueles que são aplicáveis para $m \geq 3$.

Assim, como objetivo geral deste trabalho, sentimo-nos desafiados a ajudar a preencher uma pequena parte dessa lacuna, mediante estudo de um novo teste de independência entre duas variáveis, que seja aplicável para qualquer tipo de distribuição.

Conceito Básico e Exemplo

Define-se como *dependência não linear* entre duas v.as. X e Y a situação em que $\text{Cov}(X,Y) = 0$ e $F_{(X,Y)}(x,y) \neq F_X(x)F_Y(y)$, $\forall (x,y) \in \mathbb{R}^2$, ou se $\text{Cov}(X,Y)$ inexiste.

Exemplo 0.1. *(Casella e Berger, 2010) Seja $Y = X^2 + Z$, em que $X \sim U[-1,1]$ e $Z \sim U[0,1/10]$ são variáveis aleatórias mutuamente independentes. Por causa da independência, podemos considerar que $Z|X = x \sim U[0,1/10]$, sendo x o parâmetro de locação da regressão $Y|X = x$. Portanto, $Y|X = x \sim U[x^2, x^2 + 1/10]$. A distribuição conjunta entre x e y é $f(x,y) = f(Y|X = x)f(X = x) = 5$, e a covariância entre X e Y pode ser obtida como:*

$$\begin{aligned} \text{Cov}(X,Y) &= \text{E}(XY) - \text{E}(X)\text{E}(Y) \\ &= \text{E}[X(X^2 + Z)] - \text{E}(X)\text{E}[X^2 + Z] \\ &= \text{E}[X^3] + \text{E}[XZ] - 0\text{E}[X^2 + Z] \\ &= 0 + \text{E}(X)\text{E}(Z) = 0\text{E}(Z) = 0 \end{aligned} \tag{1}$$

Assim, embora a Eq. (1) indique que a correlação entre as variáveis seja nula, a Fig. 1 mostra uma forte associação em forma de parábola (não linear) entre elas.

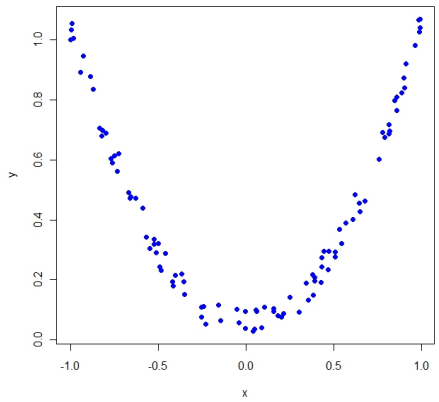


Figura 1: Função $f(x,y)$.

□

O exemplo acima, portanto, mostra uma situação em que a covariância nula não implica independência entre as variáveis em estudo. Há também situações em que

não se pode definir a covariância por causa da inexistência dos momentos, como é o caso, por exemplo, da distribuição de Cauchy.

É relevante destacar também a importância do diagrama de dispersão ao se trabalhar com variáveis quantitativas. Esse gráfico projeta a distribuição conjunta de duas variáveis, fornecendo uma idéia da forma de associação existente entre elas. Por exemplo, se as variáveis forem independentes e gaussianas, é possível notar a presença de uma nuvem de pontos aleatória (ou quando muito) um conjunto de pontos dispostos sobre uma reta horizontal. Na existência de uma correlação linear, a disposição dos pontos tende a ser uma reta. Mas na dependência não linear observam-se padrões às vezes muito complexos que nem sempre são claramente visíveis em diagramas de dispersão.

O caso de dependência linear pode ser avaliado com base em medidas bem conhecidas como a correlação linear de Pearson, os escores normais de Fisher-Yates ou os coeficientes de Spearman e de Kendall. Porém, para o caso não linear ainda há um campo aberto a ser estudado.

Formulação do problema e uma breve revisão

Considerando a amostra aleatória $(X_1, Y_1); \cdots, (X_n, Y_n)$, em que X e Y são vetores cujos elementos são v.as. absolutamente contínuas com função de distribuição acumulada (FDA), o propósito é testar a hipótese nula de independência bivariada

$$F_{(X,Y)}(x, y) = F_X(x)F_Y(y) \tag{2}$$

para todo x, y, em que $F_X(x)$ é a FDA marginal da v.a. X e $F_Y(y)$ é a FDA marginal da v.a. Y.

Por exemplo, com base em métodos clássicos, para se testar (2) calcula-se a distância entre as Tab. 1 e 2, ou seja, quão longe estão os valores esperados, sob a hipótese de independência, dos valores realmente observados.

Nessa formulação clássica, para se medir a distância entre as Tabelas 1 e 2 $\forall \mathbf{x}$,

Tabela 1: Distribuição bivariada para um par (x, y)

Eventos	$Y \leq y$	$Y > y$	Total
$X \leq x$	o_{11}	o_{10}	$n_{1\bullet}$
$X > x$	o_{01}	o_{00}	$n_{0\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 0}$	n

Tabela 2: Distribuição esperada para o par (x, y) sob a hipótese de independência

Eventos	$Y \leq y$	$Y > y$	Total
$X \leq x$	e_{11}	e_{10}	$n_{1\bullet}$
$X > x$	e_{01}	e_{00}	$n_{0\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 0}$	n

Blum et al. (1960) propuseram a estatística

$$B_{HBKR} = \sum_{i=1}^{n} [Q(x_i, y_i)]^2 \tag{3}$$

em que

$$Q(x_i, y_i) = \frac{o_{11}}{n} - \frac{n_{1\bullet}}{n} \cdot \frac{n_{\bullet 1}}{n} \tag{4}$$

Esse teste possui razoável poder estatístico para detectar diversas formas de dependência não linear (Matsushita et al., 2012), e é frequentemente considerado como base de comparação para novos testes. Csörgő (1985) propôs um teste equivalente ao de Blum, Kiefer e Rosenblatt (BKR) com base na função característica empírica. Skaug e Tjøstheim (1993) e Delgado (1996) aplicaram o teste de HBKR para detectar dependência serial. Esse último avaliou o desempenho do teste para pequenas amostras via método de Monte Carlo. De Wet (1980) propôs a modificação

$$B_W = \sum_{i=1}^{n} W(x_i, y_i)[Q(x_i, y_i)]^2, \tag{5}$$

em que $W(x_i, y_i)$ é uma função peso que permite elevar a eficiência do teste.

Outro teste clássico é o que se baseia no teste de Kolmogorov-Smirnov cuja estatística é (Parzen, 1962)

$$D_{KS} = \sqrt{n} \sup_{x,y \in \mathbb{R}^2} |Q(x, y)| \tag{6}$$

O problema desses testes é que o processo empírico $Q(x, y)$ não considera os eventos complementares mostrados nas Tabelas 1 e 2 e não são, portanto, a melhor medida para se avaliar independência em um dado ponto (x, y). Recentemente, Matsushita et al. (2012) propuseram um teste assintótico bivariado em que o processo empírico tem a forma

$$L^2(x, y) = 2 \sum_{k_1=0}^{1} \sum_{k_2=0}^{1} o_{k_1 k_2} \ln \left\{ \frac{o_{k_1 k_2}}{e_{k_1 k_2}} \right\} \tag{7}$$

ou

$$\chi^2(x, y) = \sum_{k_1=0}^{1} \sum_{k_2=0}^{1} \frac{(o_{k_1 k_2} - e_{k_1 k_2})^2}{e_{k_1 k_2}}. \tag{8}$$

Ambas as medidas são bastante conhecidas na literatura estatística. $L^2(x, y)$ é a estatística do teste da razão de verossimilhança generalizada para uma distribuição multinomial em uma tabela 2×2. $\chi^2(x, y)$ é o qui-quadrado de Pearson que, assintoticamente, é equivalente a $L^2(x, y)$. Assim, o teste pode ser efetuado com base nas estatísticas correspondentes

$$B_{LR} = \frac{1}{n} \sum_{i=1}^{n} L^2(x_i, y_i) \tag{9}$$

e

$$B_{\chi^2} = \frac{1}{n} \sum_{i=1}^{n} \chi^2(x_i, y_i). \tag{10}$$

Essas estatísticas são do tipo *distribution-free* e seus proponentes apresentaram suas distribuições exatas para o caso bivariado. Mostraram também, mediante exemplos, que esse teste apresenta maior poder que o teste de HBKR para detectar dependência não linear.

Objetivos

O principal objetivo deste trabalho, no que se refere ao problema de detecção de dependência não linear, consiste em comparar o teste de HBKR com o que foi proposto por Matsushita et al. (2012) mediante simulações de Monte Carlo. Aqui, serão acrescentadas a estatística de Kolmogorov-Smirnov e a que foi proposta por KAC (1951).

Neste trabalho, mostraremos que a estatística de Kolmogorov-Smirnov é uma alternativa tão boa como a que foi proposta por Matsushita et al. (2012), e que a de Kac é equivalente à que se baseia na estatística χ^2.

Este trabalho não considera formulações alternativas para se testar (2). Por exemplo, o teste poderia ser construído com base em densidades empíricas obtidas pelo método de suavização por Kernel (Gretton e Györfi, 2010; Skaug e Tjøstheim, 1993; Robinson, 1991; Chan e Tran, 1992). Em outro contexto, abordagens por cópulas foram feitas por Schweizer e Wolff (1981), Scaillet (2005), Schmid e Schmidt (2007). Outros processos empíricos ou critérios para mensuração da independência foram propostos por Ghoudi et al. (2001), Beran et al., (2006), Gieser e Randles (1997), Um e Randles (2001), Bakirov et al., (2006), Zhang (2008), entre outros.

Estrutura do trabalho

O Cap. 1 apresenta algumas medidas clássicas de dependência entre variáveis aleatórias. Elas são ferramentas importantes para a análise estatística de dados, e muitas delas se aplicam na elaboração de testes para a hipótese de independência. Por isso, o Cap. 1 também faz referência a alguns testes de independência. Esses testes foram divididos entre paramétricos e não-paramétricos, sendo adequados apenas para a identificação de dependência linear.

O Cap. 2 apresenta o teste de HBKR, que é um dos poucos testes com considerável poder estatístico para a detecção de estruturas de dependência não linear. Esse teste é do tipo Cramér-von Mises, ou seja, considera a distância ao quadrado entre a distribuição conjunta empírica e o produto das marginais empíricas.

No entanto, segundo Matsushita et al. (2012), é possível encontrar um teste assintótico com maior poder estatístico. Nesse caso, o teste também é do tipo Cramér von-Mises, mas é construído com base na estatística χ^2 da razão de verossimilhança (Cap. 3).

O Cap. 4 apresenta os resultados do estudo de Monte Carlo. O Cap. 5 mostra algumas aplicações dos testes utilizando taxas de câmbio de algumas moedas frente ao dólar americano. Por fim, tem-se as considerações finais que incluem, também, algumas perspectivas para trabalhos futuros.

Capítulo 1

Medidas e testes de dependência

Medidas de dependência são utilizadas para medir quanto de informação uma variável aleatória (v.a.), digamos X, pode ser explicada por outra Y. Essas medidas, quando expressas na forma de um escalar, são comumente denominadas índices. A dependência, por ser um problema de fundamental interesse, encontra aplicações em diferentes campos, tais como estatística, demografia, economia, física, processamento de sinais, epidemiologia, entre outros.

Em geral, para as variáveis qualitativas, a mensuração é realizada por medidas de associação. Quando as variáveis são quantitativas (foco deste trabalho), utilizam-se as medidas de correlação. No entanto, como veremos nos próximos capítulos, as medidas de associação também podem ser aplicadas para variáveis quantitativas.

Índices de dependência podem não assegurar se existe realmente independência entre as variáveis, mas fornecem alguma idéia quantitativa de proximidade. Em geral o índice varia de 0 a 1, indicando, respectivamente, ausência de dependência mútua e total dependência mútua.

Neste capítulo apresentamos duas medidas clássicas de dependência: coeficiente de correlação de Pearson e medida χ^2. O Anexo A apresenta outras medidas de dependência, tais como: coeficiente de correlação máxima e tetracórico e medidas de dependência de Cramér e Hoeffding Apresentamos, ainda, alguns testes de dependência que serão utilizados no Cap. 4 como base de comparação com o novo teste proposto por Matsushita et al. (2012).

1.1 Coeficiente de Correlação

O coeficiente de correlação de Person (ρ) é uma medida natural da dependência linear entre duas v.as. (X e Y) conjuntamente gaussianas, sendo definido como

$$\rho = \frac{\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)}{\sqrt{\sigma_X^2 \sigma_Y^2}}, \tag{1.1}$$

em que $|\rho| \leq 1$, e $\sigma_X^2 > 0$ e $\sigma_Y^2 > 0$ são as variâncias de X e Y, respectivamente. Quanto mais próximo dos extremos, maior a correlação (positiva ou negativa) entre essas variáveis.

Definição 1.1. *Se $F_{(X,Y)}(x, y) = F_X(x)F_Y(y)$ então $\rho = 0$, isto é, há independência entre X e Y. A recíproca, no entanto, não é verdadeira.*

O coeficiente ρ se relaciona com o conceito de regressão linear. A média condicional $\mathrm{E}[Y|X = x]$ denomina-se curva de regressão, e sua aproximação de primeira ordem, $\mathrm{E}[Y|X = x] \approx ax + b$, é chamada reta de regressão de Y em x. Nessa forma linear, $a = \rho\sigma_Y/\sigma_X$ é o coeficiente angular e b é o intercepto. Essa técnica permite explorar e inferir a relação de uma variável dependente (variável resposta) com variáveis independentes específicas (variáveis explicatórias). O percentual da variância de Y que pode ser explicado pela variância de X é dado pelo coeficiente de determinação (ρ^2).

Se X e Y são conjuntamente gaussianas, tem-se $\mathrm{E}[Y|X = x] = ax + b$. Mas fora desse ambiente, que é o caso não linear, a aproximação de primeira ordem nem sempre é boa o suficiente. Nesse contexto, $\rho = 0$ não necessariamente significa independência, pois é possível haver situações em que $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$, mesmo que haja dependência entre X e Y (veja exemplo 0.1).

Definição 1.2. *Em nosso trabalho, dizemos existir* dependência não linear *entre X e Y se* $\mathrm{Cov}(X, Y) = 0$ *e $F_{(X,Y)}(x, y) \neq F_X(x)F_Y(y)$, $\forall (x, y) \in \mathbb{R}^2$, ou se* $\mathrm{Cov}(X, Y)$ *inexiste. Por outro lado, define-se como* dependência linear *a situação em que $F_{(X,Y)}(x, y) \neq F_X(x)F_Y(y)$ implica, necessariamente,* $\mathrm{Cov}(X, Y) \neq 0$.

1.2 Qui-quadrado

Qui-quadrado (χ^2) é uma medida de associação que se baseia em comparações entre as frequências observadas e as esperadas em uma tabela de contingência. Ela se define como

$$\chi^2 = \sum_{i}^{n}\sum_{j}^{m}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}, \tag{1.2}$$

sendo o_{ij} a frequência observada para cada uma das caselas (linha i, coluna j) e e_{ij} a frequência esperada correspondente a cada o_{ij}, obtida pelo produto da frequência das marginais i,j dividido pela frequência global. Valores altos de χ^2 indicam forte associação entre as variáveis e, por outro lado, se a hipótese de não-associação for verdadeira, o valor para essa medida deve ser próximo de zero.

1.3 Testes de dependência

Testes estatísticos permitem a tomada de decisões sobre determinada hipótese, definindo-se objetivamente a probabilidade de se cometer erros estatísticos. Em linhas gerais, os testes podem ser classificados como paramétricos ou não-paramétricos.

Testes paramétricos são utilizados quando é possível identificar a distribuição teórica da variável em estudo. Esses testes são, em geral, mais rigorosos, exigindo condições muitas vezes restritivas para a sua utilização. Os testes não-paramétricos, também chamados de testes de distribuição livre, são baseados nas posições que os dados ordenados recebem. Os testes não-paramétricos não fazem suposição sobre a natureza ou forma das distribuições populacionais e são menos sensíveis a *outliers*.

Um dos testes de grande relevância é o de independência, para o qual a hipótese de independência bivariada pode ser escrita como

$$H_0 : F(x,y) = F(x)F(y), \quad \forall(x,y) \tag{1.3}$$

Para a realização desse teste, supõe-se que cada um dos pares de observação (x_i, y_i) sejam provenientes da mesma população bivariada, com função de distribuição contínua, havendo independência mútua entre todos os pares de observações considerados.

Dentre os testes não-paramétricos citam-se os de aderência, que em sua forma padrão são indicadores para verificar se uma distribuição se ajusta bem ou não aos dados amostrais. No entanto, adaptações podem ser feitas a fim de identificar a existência de dependência, conforme apresentado na Subseção 1.4.2.

A Seção 1.3.1 apresenta o teste paramétrico baseado no coeficiente de correlação de Pearson. Alguns testes não-paramétricos constam na Seção 1.4.

1.3.1 Teste baseado no coeficiente de correlação de Pearson

A estatística r apresentada em (1.4) é um estimador do parâmetro populacional ρ definido por (1.1).

$$r = \frac{\hat{C}ov(X,Y)}{\sqrt{\hat{V}ar(X)\hat{V}ar(Y)}} \tag{1.4}$$

No entanto, uma amostra aleatória de pontos, retirada de uma distribuição bivariada cujas variáveis não sejam correlacionadas, certamente produzirá $r \neq 0$. Por isso, é necessário testar se, de fato, a amostra foi colhida de uma população para qual o coeficiente de correlação é nulo.

Assumindo que (X, Y) possui distribuição normal bivariada, para testar as hipóteses abaixo descritas, utiliza-se a estatística do teste apresentada em (1.5)

$$\begin{cases} H_0 & : \quad \rho = 0 \\ H_a & : \quad \rho \neq 0 \end{cases}$$

$$T = r\sqrt{\frac{n-2}{1-r^2}}, \tag{1.5}$$

em que $T \sim t\ de\ Student_{(n-2)}$ e, portanto, rejeita-se H_0 se $|T| \geq t_{\alpha/2,(n-2)}$ quando o teste é bilateral. Para o caso unilateral, a região crítica é alterada conforme a hipótese alternativa.

A próxima seção apresenta os testes não-paramétricos χ^2 e uma adaptação do teste de ajustamento de Kolmogorov-Smirnov para avaliar dependência.

1.4 Testes não-paramétricos

Para se testar hipóteses estatísticas com teste não-paramétrico, nenhuma suposição sobre a forma da distribuição populacional é feita. De acordo com Hoeffding (1948), as propriedades de um bom teste não-paramétrico devem incluir a ausência de vício e a consistência. Para ele, um teste com hipótese nula (H_0) é consistente (com respeito a uma classe específica de hipóteses admissíveis) se a probabilidade de aceitar H_0, com o aumento do tamanho da amostra, tende a zero sempre que houver evidências para não se rejeitar a hipótese alternativa (H_a).

1.4.1 Teste baseado na estatística χ^2

Para o teste de dependência χ^2 utiliza-se a estatística (1.2). Esse teste mede o desvio entre os valores observados e os valores esperados (sob a hipótese de independência) em uma tabela de contingência.

O valor encontrado deve ser comparado com o valor tabelado (veja, por exemplo Tab. IV de Bussab e Morettin, 2006) de acordo com os graus de liberdade existentes e o nível de significância (α) escolhido. Para valores de χ^2 maiores que o valor tabelado, existem evidências de que a hipótese de independência deve ser rejeitada. O teste de χ^2 pode ser usado se as frequências esperadas forem maiores ou iguais a 5. Caso contrário, aplica-se a correção de Yates.

1.4.2 Testes de Kolmogorov para independência

A hipótese testada pelos testes de aderência tradicionais refere-se à forma da distribuição da população que a amostra em estudo representa. Ou seja, a hipótese nula afirma que determinada distribuição se ajusta bem ou não aos dados amostrais. A verificação é feita através da comparação das frequências amostrais com as frequências teóricas esperadas pelo modelo probabilístico que se está julgando válido para descrever os dados observados. Dentre os testes de aderência mais comuns cita-se o de Kolmogorov-Smirnov. Neste trabalho, a idéia consiste, portanto, em adaptar esse teste para verificar independência.

O teste de ajustamento de Kolmogorov-Smirnov segundo Conover (1999) e Parzen

(1962), em sua forma clássica, consiste em mensurar a maior distância entre a função de distribuição empírica $\hat{F}(x)$ e a função de distribuição teórica hipotetizada, isto porque $\hat{F}(x)$ pode ser um estimador de $F(x)$ (desconhecida). Esse teste pode ser dividido em várias classes: i) os testes com amostra única; ii) os testes com duas amostras e iii) os testes com 3 ou mais amostras.

Para a situação de duas amostras, $(X_1, \ldots, X_n)$ e $(Y_1, \ldots, Y_m)$, em que se deseja testar se as funções de distribuição de cada população são ou não iguais, tem se a seguinte estatística para o teste bilateral:

$$B_{KS} = \sup_{x,y} |\hat{F}(x) - \hat{F}(y)|. \tag{1.6}$$

Portanto, a hipótese nula deve ser rejeitada a um nível α de significância se a estatística do teste (T) for maior que o quantil $1 - \alpha$ tabelado em A20 de Conover (1999), se $n = m$, e em A21, se $n \neq m$.

Adaptando esse teste de aderência convencional com estatística do tipo Kolmogorov-Smirnov para identificar a existência de independência entre duas amostras, tem-se:

$$H_0 : \hat{F}(x, y) = \hat{F}(x)\hat{F}(y) \tag{1.7}$$

com a estatística Parzen (1962)

$$B_{KS} = \sqrt{n} \sup |\hat{F}(x, y) - \hat{F}(x)\hat{F}(y)|. \tag{1.8}$$

Nesse caso, porém, não dispomos de tabelas para os níveis críticos. Para o nosso trabalho, esses valores serão obtidos por simulação de Monte Carlo no Capítulo 5.

1.5 Considerações

As medidas de dependência e os testes χ^2 e o baseado no coeficiente de Pearson visam identificar a existência de dependência linear entre as variáveis. No entanto, o teste χ^2 também pode ser utilizado para avaliar dependência não linear, como apresentado no Cap. 3. A versão do teste de Kolmogorov-Smirnov (KS) apresentado nesse capítulo

visa a identificação de dependência não linear. Em geral, quando utilizados como testes de aderência, o teste de Kolmogorov-Smirnov (KS) é mais adequado que o qui-quadrado para tamanhos de amostra pequenos. O teste KS tende a ser mais poderoso que o qui-quadrado na maior parte das situações, mas não se sabe se esse padrão se mantém para testes de independência bivariada. O desempenho do teste KS se sobressai quando a função de distribuição hipotética é completamente especificada, ou seja, tanto a forma como seus parâmetros são conhecidos. Mas, para a nossa aplicação, não há especificação da forma funcional da distribuição sob a hipótese nula.

Por outro lado, o teste χ^2 requer grandes amostras, mas proporciona bons resultados para a detecção de dependência não linear (Matsushita et al., 2012).

Outros testes ou técnicas (como cópulas) poderiam ser considerados, mas preferimos restringir nossa atenção à classe dos testes discutidos aqui. Nos próximos capítulos serão apresentados testes capazes de identificar dependências não lineares. No Cap. 2 trataremos do teste de HBKR (Blum et al., 1960) e no Cap. 3 apresentamos a variação do teste de HBKR proposta por Matsushita et al. (2012).

Capítulo 2

O teste de HBKR

Em geral, a aplicação dos métodos estatísticos para a análise e modelagem de dados requer diagnósticos acerca da presença de dependência entre as variáveis envolvidas no estudo. Portanto, é importante que se disponha de um bom instrumento para a detecção da dependência linear e não linear entre duas ou mais variáveis aleatórias. Neste trabalho, nos restringimos ao caso bivariado.

Segundo Bakirov et al. (2006), entre os poucos métodos estatísticos que possuem poder estatístico suficiente para detectar as diversas formas de dependência não linear, encontra-se o teste de HBKR (Hoeffding, Blum, Kiefer e Rosenblatt) proposto por (Blum et al., 1960) como uma variação do teste de Hoeffding (1948). Esse teste é do tipo Cramér-von Mises (veja Genest, 2006) que considera um processo aleatório gerado pela diferença entre a distribuição conjunta empírica e o produto correspondente entre as marginais empíricas.

Neste capítulo, considerando que $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$ não necessariamente indica independência, e que os momentos da distribuição podem não existir, a Seção 2.1 apresenta resumidamente o teste de HBKR para o caso bivariado como ferramenta para a identificação de dependência não linear.

2.1 O teste de independência

Considere a amostra aleatória $(a.a)$ $(X_1, Y_1); \ldots; (X_n, Y_n)$ de vetores aleatórios bidimensionais com função de distribuição acumulada contínua (FDA) $F(x, y) \in \Re^2$.

O interesse consiste em testar a hipótese nula de independência:

$$H_0 \quad : \quad F(x,y) = F(x)F(y), \forall (x,y) \in \Re^2. \tag{2.1}$$

Ou seja, deseja-se testar se as variáveis que compõem a função F são ou não independentes. Isso ocorre quando as funções de distribuição são resultantes do produto de suas correspondentes marginais unidimensionais.

Para testar essa hipótese, utilizam-se as funções de distribuição amostral ou empíricas (FDE). A estimação das funções de densidades é uma alternativa viável que nos remete a outras abordagens, e por isso não será discutida neste trabalho. Considerando a amostra aleatória definida anteriormente, a FDE nesse caso pode ser obtida como:

$$\hat{F}_{X,Y}(x,y) \quad = \quad \frac{1}{n}\sum_{j=1}^{n} I_X(x) I_Y(y), \tag{2.2}$$

em que a função indicadora de X se define como

$$I_X(x) = \begin{cases} 1 & \text{se} \quad X \leq x \\ 0 & \text{se} \quad X > x. \end{cases}$$

Portanto, a Eq.(2.2) representa a proporção de valores observados das variáveis X e Y que são menores ou iguais a x e y. A representação da função indicadora de Y é feita de forma semelhante à de X, realizando apenas a troca de x por y.

Considere $\hat{F}_{X,Y}(x,y) = P(X \leq x, Y \leq y)$ como a função de distribuição conjunta empírica entre x e y. Assim, a estatística que mede a distância entre a distribuição conjunta e o produto de suas marginais é dada por

$$T_{X,Y}(x,y) = |\hat{F}_{X,Y}(x,y) - \hat{F}_X(x)\hat{F}_Y(y)|. \tag{2.3}$$

De acordo com Blum et al. (1960), testes baseados em (2.3) terão boas propriedades de poder. Os testes de Kolmogorov-Smirnov e de Cramér von-Mises satisfazem essas propriedades e são similares sob a hipótese de independência.

Blum et al. (1960) desenvolveram o teste utilizando a estatística do tipo Cramér-von Mises, conforme (2.4), e obtiveram os níveis críticos da distribuição assintótica de

nB_{HBKR} sob a hipótese de independência para o caso bivariado, conforme apresentado na Tab. 2.1.

$$B_{HBKR} = \sum_{i=1}^{n} [\hat{F}(x_i, y_i) - \hat{F}(x_i)\hat{F}(y_i)]^2, \tag{2.4}$$

sendo n a quantidade de observações das variáveis aleatórias do vetor bidimensional.

Tabela 2.1: Estatística $\frac{1}{2}\pi^4 nB_{HBKR}$: Valores ($\alpha$) e níveis críticos correspondentes

$\alpha(\%)$	0.1	1.0	2.0	5.0	10.0
b	6.32	4.23	3.62	2.84	2.29

Embora os resultados aqui apresentados se limitem ao caso bivariado, Blum et al. (1960) e também publicaram alguns resultados para o caso trivariado, e argumentam que, para os casos com dimensões superiores a 3, basta realizar as devidas modificações.

2.2 Considerações

Este capítulo apresentou o teste clássico para a identificação de dependência não linear conhecido como HBKR. Existem outros testes desenvolvidos para esse mesmo objetivo, mas a maior parte deles é equivalente ao teste de HBKR para o caso bivariado ou não é do tipo *distribution-free*.

Levando-se em consideração o princípio da máxima verossimilhança, o teste de HBKR apresenta características que não o tornam ótimo, pois não utiliza todas as informações disponíveis. Em 2012, foi desenvolvido por Matsushita et al. (2012) um teste que apresentou poder superior ao de HBKR, como apresentaremos a seguir.

Capítulo 3

Uma variação do teste de HBKR

Como discutimos no Cap. 2, o teste de HBKR é um dos poucos que possui bom desempenho para identificar estruturas de dependência não linear. E muitos dos outros testes são, essencialmente, equivalentes ao de HBKR para o caso bivariado (para mais detalhes consulte Bakirov et al. (2006); Beran et al. (2006); Bilodeau e Micheaux (2005) e Ghoudi et al. (2001)).

Em 2012, Matsushita et al. (2012) desenvolveram um novo teste de independência assintótico do tipo Cramér-von Mises, cujo processo empírico se baseia na estatística χ^2 da razão de verossimilhança (veja Mood et al. (1987)).

Este capítulo apresenta, de maneira sucinta, a idéia geral desse teste *distribution free* que, conforme demonstrado em Matsushita et al. (2012), possui poder estatístico superior ao do teste de HBKR para a detecção de diferentes tipos de dependência não linear, inclusive em situações com observações extremas.

A Seção 3.1 trata desse teste de independência para o caso bivariado ($m = 2$), apresentando um exemplo para ilustrar o cálculo da estatística do teste. No que se refere a distribuição amostral do novo teste, os valores para a média e a variância, assim como os níveis críticos encontrados constam na Seção 3.2.

3.1 O teste de independência

Considere novamente a amostra aleatória de vetores aleatórios bi-dimensionais (X_1, Y_1); $\ldots$; (X_n, Y_n) com FDA $F(x, y) \in \Re^2$. O propósito é testar a hipótese nula de inde-

pendência bivariada

$$H_0 \quad : \quad F(x,y) = F(x)F(y), \forall (x,y) \in \Re^2. \tag{3.1}$$

As Tabelas 3.1 e 3.2 representam, respectivamente, a distribuição bivariada observada e a distribuição esperada sob a hipótose de independência para um dado ponto $(x,y) \in \Re^2$. Os índices $11, 10, 01$ e 00 se relacionam com a função indicadora de X e Y, conforme descrito no Cap. 2.

Tabela 3.1: Distribuição bivariada para um par (x,y)

Eventos	$Y \leq y$	$Y > y$	Total
$X \leq x$	o_{11}	o_{10}	$n_{1\bullet}$
$X > x$	o_{01}	o_{00}	$n_{0\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 0}$	n

Tabela 3.2: Distribuição esperada para o par (x,y) sob a hipótese de independência

Eventos	$Y \leq y$	$Y > y$	Total
$X \leq x$	e_{11}	e_{10}	$n_{1\bullet}$
$X > x$	e_{01}	e_{00}	$n_{0\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 0}$	n

Os elementos da Tab. 3.2 são obtidos considerando a distribuição empírica da Tab. 3.1 sob a hipótese de independência, isto é,

$$e_{k_1 k_2} = \frac{n_{k_1 \bullet} \; n_{\bullet k_2}}{n}. \tag{3.2}$$

Os elementos que formam as tabelas acima são contagens dadas pelas equações abaixo, conforme Matsushita et al. (2012):

$$o_{k_1k_2} = \sum_{i=1}^{n} \big(1 - I(x - X_i)\big)^{1-k_1} \big(1 - I(y - Y_i)\big)^{1-k_2} I^{k_1}(x - X_i) I^{k_2}(y - Y_i); \tag{3.3}$$

$$n_{k_1\bullet} = \sum_{i=1}^{n} \big\{1 - I(x - X_i)\big\}^{1-k_1} I^{k_1}(x - X_i); \tag{3.4}$$

$$n_{\bullet k_2} = \sum_{i=1}^{n} \big\{1 - I(y - Y_i)\big\}^{1-k_2} I^{k_2}(y - Y_i); \tag{3.5}$$

em que k_1 e $k_2 = 0, 1$, e $I(y) = 1$, se $y \geq 0$; e $I(y) = 0$, caso contrário. Por exemplo, para o caso o_{11} tem-se

$$\begin{aligned} o_{11} &= \sum_{i=1}^{n} (1 - I(x - X_i))^{1-1} (1 - I(y - Y_i))^{1-1} (I^1(x - X_i))(I^1(y - Y_i)) \\ &= \sum_{i=1}^{n} \underbrace{I(x - X_i)}_{\delta_x} \underbrace{I(y - Y_i)}_{\delta_y}, \end{aligned} \tag{3.6}$$

sendo $\delta_x = I(\underbrace{x}_{\text{fixo}} - X_i) = 1 \quad \text{se} \quad X_i \leq x_i$.

Como já foi discutido, os testes do tipo Cramér-von Mises se definem com base no quadrado da distância entre a distribuição hipotética e a distribuição empírica. Portanto, diferentes testes podem ser obtidos a partir das Tabelas 3.1 e 3.2. A estatística de HBKR apresentada em (2.4), por exemplo, pode ser reescrita como

$$B_{HBKR} = \sum_{\forall(x,y)} \left[\frac{o_{11} - e_{11}}{\sqrt{n}} \right]^2, \tag{3.7}$$

em que a soma acima se dá sobre todos os pares (x, y) encontrados na amostra. É notório pela Eq. (3.7) que apenas os eventos $[X \leq x]$ e $[Y \leq y]$ são considerados. Como alternativa a esse teste, Matsushita et al. (2012) propuseram um novo teste que considera também os eventos complementares $[X > x]$ e $[Y > y]$ e possui como estatística de teste a seguinte expressão

$$L^2(x, y) = 2 \sum_{k_1=0}^{1} \sum_{k_2=0}^{1} o_{k_1k_2} \ln \frac{o_{k_1k_2}}{e_{k_1k_2}} \tag{3.8}$$

A estatística χ^2 de Pearson possui equivalência assintótica com $L^2(x,y)$ (Agresti, 2007), e portanto, foi considerada no estudo

$$\chi^2(x,y) = \sum_{k_1=0}^{1}\sum_{k_2=0}^{1}\frac{(o_{k_1k_2} - e_{k_1k_2})^2}{e_{k_1k_2}}. \tag{3.9}$$

As estatísticas $L^2(x,y)$ e $\chi^2(x,y)$ possuem propriedades ótimas relacionadas com o princípio da máxima verossimilhança (Roussas, 1997) e foram propostas por Matsushita et al. (2012) para o teste de dependência bivariado do tipo Cramér-von Mises como

$$B_{LR} = \frac{1}{n}\sum_{\forall(x,y)} L^2(x,y), \tag{3.10}$$

e

$$B_{\chi^2} = \frac{1}{n}\sum_{\forall(x,y)} \chi^2(x,y). \tag{3.11}$$

Matsushita et al. (2012) encontraram a função característica teórica da distribuição limite da estatística do teste conforme apresentado em (3.12)

$$\phi_B(q) = \prod_{j,k}\left\{1 - \frac{2iq}{j(j+1)k(k+1)}\right\}^{-\frac{1}{2}}. \tag{3.12}$$

em que $j = 1, 2, \ldots$

3.1.1 Uma representação alternativa do processo $\chi^2(x,y)$

Com base nas Tabelas 3.1 e 3.2, temos as relações

$$o_{10} = n_{1\bullet} - o_{11}, \tag{3.13}$$

$$o_{01} = n_{\bullet 1} - o_{11}, \tag{3.14}$$

$$\begin{aligned} o_{00} &= n - o_{11} - o_{10} - o_{01} \\ &= n - n_{\bullet 1} - n_{1\bullet} + o_{11}, \end{aligned} \tag{3.15}$$

e, de modo semelhante,

$$e_{10} = n_{1\bullet} - e_{11}, \tag{3.16}$$

$$e_{01} = n_{\bullet 1} - e_{11}, \tag{3.17}$$

$$e_{00} = n - n_{\bullet 1} - n_{1\bullet} + e_{11}. \tag{3.18}$$

Consequentemente, as diferenças entre as frequências observadas e as esperadas são todas iguais,

$$o_{10} - e_{10} = e_{11} - o_{11}, \tag{3.19}$$

$$o_{01} - e_{01} = e_{11} - o_{11}, \tag{3.20}$$

$$o_{00} - e_{00} = e_{11} - o_{11}. \tag{3.21}$$

Assim, para uma tabela de contingência 2×2, o processo empírico $\chi^2(x, y)$ definido na Eq. (3.9) pode ser reescrito como

$$\begin{aligned}
\chi^2(x,y) &= \frac{(o_{00}-e_{00})^2}{e_{00}} + \frac{(o_{01}-e_{01})^2}{e_{01}} + \frac{(o_{10}-e_{10})^2}{e_{10}} + \frac{(o_{11}-e_{11})^2}{e_{11}} \\
&= (o_{11}-e_{11})^2\left\{\frac{1}{e_{00}} + \frac{1}{e_{01}} + \frac{1}{e_{10}} + \frac{1}{e_{11}}\right\} \\
&= (o_{11}-e_{11})^2\left\{\frac{1}{n_{\bullet 0}n_{0\bullet}} + \frac{1}{n_{\bullet 0}n_{1\bullet}} + \frac{1}{n_{\bullet 1}n_{0\bullet}} + \frac{1}{n_{\bullet 1}n_{1\bullet}}\right\} \\
&= (o_{11}-e_{11})^2\left\{\frac{n_{\bullet 1}n_{1\bullet} + n_{\bullet 0}n_{1\bullet} + n_{\bullet 0}n_{1\bullet} + n_{\bullet 0}n_{0\bullet}}{n_{\bullet 0}n_{0\bullet}n_{\bullet 1}n_{1\bullet}}\right\} \\
&= \frac{(o_{11}-e_{11})^2 \cdot n}{n_{\bullet 0}n_{0\bullet}n_{\bullet 1}n_{1\bullet}} \\
&= \frac{(o_{11}-e_{11})^2 \cdot n \cdot n^4}{n_{\bullet 0}n_{0\bullet}n_{\bullet 1}n_{1\bullet} \cdot n^4} \\
&= \frac{(\hat{F}(x,y) - \hat{F}(x)\hat{F}(y))^2}{n \cdot (1-\hat{F}(x))(1-\hat{F}(y))\hat{F}(x)\hat{F}(y)},
\end{aligned}$$

em que $o_{11}/n = \hat{F}(x, y)$, $n_{\bullet 1}/n = \hat{F}(y)$ e $n_{1\bullet}/n = \hat{F}(x)$.

Portanto,

$$\chi(x,y) = \frac{|\hat{F}(x,y) - \hat{F}(x)\hat{F}(y)|}{\sqrt{n \cdot (1-\hat{F}(x))(1-\hat{F}(y))\hat{F}(x)\hat{F}(y)}}, \tag{3.22}$$

representa uma versão padronizada do processo empírico do teste de HBKR, tratado por Kac (1951). No entanto, sua distribuição amostral foi apresentada por Matsushita et al. (2012).

Ilustração

Apenas para ilustrar a operacionalização das estatísticas B_{HBKR} e B_{χ^2}, considere o seguinte exemplo.

Exemplo 3.1. *Considere a Tab.3.3 com os possíveis valores para x e y*

id	x	y
1	5	6
2	4	7
3	6	3
4	7	8

Tabela 3.3: Possíveis valores para x e y

1. *Escolher um par (x, y) para ser o valor fixo de referência. Aqui será feita na ordem em que aparecem na Tab. 3.3*

2. *Montar as tabelas de contingência*

3. *Calcular as estatísticas e tomar a decisão de rejeitar ou não a hipótese de independência*

As tabelas 3.4 a 3.7 são as tabelas de contingência montadas a partir dos pares escolhidos para ser os valores de referência, em que é feita a comparação com todos os demais elementos da amostra.

x/y	$y \leq 6$	$y > 6$	Total
x ≤ 5	1 (1/4)	1 (1/4)	2 (2/4)
x > 5	1 (1/4)	1 (1/4)	2 (2/4)
Total	2 (2/4)	2 (2/4)	4 (1)

Tabela 3.4: Primeiro par (5,6) - valores observados e (frequências relativas)

x/y	$y \leq 7$	$y > 7$	Total
x ≤ 4	1 (1/4)	0	2 (1/4)
x > 4	2 (2/4)	1 (1/4)	2 (3/4)
Total	3 (3/4)	1 (1/4)	4 (1)

Tabela 3.5: Segundo par (4,7) - valores observados e frequências relativas

x/y	$y \leq 3$	$y > 3$	Total
x ≤ 6	1 (1/4)	0	2 (1/4)
x > 6	2 (2/4)	1 (1/4)	2 (3/4)
Total	3 (3/4)	1 (1/4)	4 (1)

Tabela 3.6: Terceiro par (6,3) - valores observados e frequências relativas

x/y	$y \leq 8$	$y > 8$	Total
x ≤ 7	4 (1)	0	4 (1)
x > 7	0	0	0
Total	4 (1)	0	4 (1)

Tabela 3.7: Quarto par (7,8) - valores observados e frequências relativas

A estatística de HBKR para este exemplo é

$$\begin{aligned} B_{HBKR} &= \left(\frac{1}{4} - \frac{1}{2}\cdot\frac{1}{2}\right)^2 + \left(\frac{1}{4} - \frac{3}{4}\cdot\frac{1}{4}\right)^2 + \left(\frac{1}{4} - \frac{1}{4}\cdot\frac{3}{4}\right)^2 + \left(\frac{4}{4} - \frac{4}{4}\cdot\frac{4}{4}\right)^2 \\ &= (0)^2 + \left(\frac{1}{16}\right)^2 + \left(\frac{1}{16}\right)^2 + (0)^2 \\ &= 0.012 \end{aligned}$$

e, utilizando a representação alternativa,

$$\begin{aligned} B_{\chi^2} &= \frac{\left(\frac{1}{4} - \frac{1}{2}\cdot\frac{1}{2}\right)^2}{\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}} + \frac{\left(\frac{1}{4} - \frac{3}{4}\cdot\frac{1}{4}\right)^2}{\frac{1}{4}\cdot\frac{3}{4}\cdot\frac{3}{4}\cdot\frac{1}{4}} + \frac{\left(\frac{1}{4} - \frac{1}{4}\cdot\frac{3}{4}\right)^2}{\frac{3}{4}\cdot\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{3}{4}} \\ &= 0,22 \end{aligned}$$

Consultando as Tab. 2.1 e 3.9 para B_{HBKR} e B_{χ^2}, respectivamente, para qualquer α escolhido, não há evidência para se rejeitar a hipótese de independência.

3.2 Valores críticos assintóticos

Como consta em Matsushita et al. (2012), a Eq. (3.12) sugere que a estatística B_{χ^2} seja uma soma de v.as. independentes gama com parâmetros de forma iguais a 1/2, mas com parâmetros de escala iguais a $2/\{j(j+1)k(k+1)\}$, $j, k \geq 1$. Os autores, para efetuar a soma numericamente, utilizaram o software SAS 9.3 para gerar 100.000 replicações de amostras tamanho 200 com correção de vício decorrente do truncamento da soma.

A Tab. 3.8 mostra que a média e a variância dos dados simulados são bem próximos dos valores obtidos pela teoria assintótica.

Tabela 3.8: Estatística B_{LR} simulada: média e variância empíricas, e seus valores teóricos correspondentes

distribuição	média	variância
empírica	0,998	0,1671
teórica	1,000	0,1680

Com base na distribuição empírica, a hipótese nula de independência (3.1) deve ser rejeitada se $B > b$, onde b é o valor crítico relativo ao nível de significância desejado ns. Os valores críticos b, após a correção do vício, para os níveis críticos ns $= 0, 1\%, 1\%, 2\%, 2, 5\%, 5\%$, e 10% estão apresentados na Tab. 3.9

Tabela 3.9: Estatística B_{LR}: níveis de significância ns e os valores críticos correspondentes b para se testar a hipótese de independência

$ns(\%)$	0,1	1,0	2,0	2,5	5,0	10,0
b	3,527	2,468	2,164	2,072	1,778	1,507

A validação e o poder do teste em Matsushita et al. (2012) também foram obtidos por meio de experimentos de Monte Carlo, comparando o teste proposto com o teste de HBKR e o teste para o coeficiente de correlação ρ. A partir dos resultados é possível verificar que o teste proposto é consistente e poderoso em identificar estruturas de dependência não linear de variáveis aleatórias bivariadas não gaussianas.

3.3 Considerações

Esse capítulo apresentou o teste proposto por Matsushita et al. (2012) que, embora seja do tipo Cramér-von Mises (assim como o teste de HBKR do Cap. 2) é baseado na razão de verossimilhança da estatística clássica χ^2 de Pearson. Os resultados apresentados são todos baseados na teoria assintótica. O Cap. 4 apresentará resultados empíricos para o teste proposto com diversos tamanhos amostrais.

Capítulo 4

Simulações

O Cap. 3 apresentou os resultados teóricos obtidos com o auxílio computacional para o teste proposto por Matsushita et al. (2012), com base na função característica encontrada para o caso bivariado. Os estudos de validação e poder do teste, realizados por simulações de Monte Carlo, mostraram, a partir de comparações com o teste do coeficiente de Pearson e o de HBKR que o teste proposto possui melhor desempenho.

Neste capítulo é feita a validação do teste e a avaliação do poder do teste comparando-se a estatística do teste proposto baseado na razão de verossimilhança (LR) com a estatística clássica χ^2 de Pearson (CHI), com a estatística do teste de HBKR (BKR), com a estatística de Kolmogorov-Smirnov (KS), com a correlação de Pearson (CORR) e com a estatística KAC proposta por KAC (1951). No entanto, na Seção 3.1.1, observamos que a estatística KAC equivale à estatística CHI, o que justifica o fato de os resultados para essas duas estatísticas serem semelhantes.

A Seção 4.1 define as estatísticas utilizadas no estudo. A validação e o poder do teste estão descritos na Seção 4.2.

Todas as simulações deste capítulo foram realizadas em linguagem Fortran com compilador g95, por possibilitar maior velocidade de execução. Os cálculos e gráficos, no entanto, foram realizados no software livre R (R Core Team, 2014).

4.1 Estatísticas utilizadas na simulação

$$CORR = \sum_{i=1}^{n} \frac{E(x_i y_i) - E(x_i)E(y_i)}{\sqrt{(Var(x_i)Var(y_i))}} \tag{4.1}$$

$$BKR = \sum_{i=1}^{n} [\hat{F}(x_i, y_i) - \hat{F}(x_i)\hat{F}(y_i)]^2 \tag{4.2}$$

$$CHI = \frac{1}{n} \sum_{\forall(x,y)} \left(\frac{(o_{11} - e_{11})^2}{e_{11}} + \ldots + \frac{(o_{00} - e_{00})^2}{e_{00}} \right) \tag{4.3}$$

$$KAC = \sum_{i=1}^{n} \frac{[\hat{F}(x_i, y_i) - \hat{F}(x_i)\hat{F}(y_i)]^2}{\hat{F}(x_i)\hat{F}(y_i)(1 - \hat{F}(x_i))(1 - \hat{F}(y_i))} \tag{4.4}$$

$$LR = \frac{1}{n} \sum_{\forall(x,y)} 2 \left[\left(\frac{o_{11} \ln o_{11}}{e_{11}} + \ldots + \frac{o_{00} \ln o_{00}}{e_{00}} \right) \right] \tag{4.5}$$

$$KS = \max |\hat{F}(x_i, y_i) - \hat{F}(x_i)\hat{F}(y_i)| \tag{4.6}$$

4.2 Validação do teste

Para a validação do teste foram geradas 50.000 replicações de duas distribuições uniformes independentes (x e y) de tamanhos variáveis. Isto é, replicações de amostras aleatórias supondo que a hipótese (3.1) seja verdadeira.

A Tab. 4.1 mostra que os valores convergem para o valor teórico, no entanto, são ligeiramente piores que os valores empíricos encontrados por Matsushita et al. (2012) (Tab. 3.8), apresentando erro apenas na terceira casa decimal para amostras de tamanho iguais ou superiores a 2.500.

A Tab. 4.2 apresenta os valores críticos encontrados via simulação de Monte Carlo para a estatística KS, segundo níveis de significância (α) para diferentes tamanhos de amostra. Por ser um teste bilateral, as letras a e b indicam, respectivamente o valor à esquerda e o valor à direita da distribuição.

Os valores da Tab. 4.3 são referentes ao percentual de réplicas para o qual o teste não aceitou a hipótese de independência ao comparar os valores das estatísticas (4.2) a (4.4) com os respectivos níveis críticos esperados para determinado valor de α. Para todas as estatísticas, os valores de α nominais estão flutuando em torno dos valores teóricos, exceto para o KS onde, por desconhecimento dos níveis críticos, após adaptar o teste de Kolmogorov-Smirnov para testar independência, optou-se por utilizar os

Tabela 4.1: Média da estatística B_{LR} simulada para amostras variando, sob H_0

Tamanho da amostra	Média
500	0.9602720
1000	0.9786637
1375	0.9852231
1750	0.9858690
2125	0.9883423
2500	0.9904258
3125	0.9900133
3750	0.9903289
4375	0.9910121
5000	0.9921837
5625	0.9922280
6250	0.9921639
6875	0.9925639
7500	0.9933274
8125	0.9940462
8750	0.9943572
9375	0.9938466
10000	0.9941286

Tabela 4.2: Estatística KS: níveis de significância (ns) e valores críticos bilaterais para se testar a hipótese de independência com diferentes tamanhos de amostra

ns(%)		Tamanho da amostra				
		1000	2500	5000	7500	10000
0.1	a	0.009	0.006	0.004	0.004	0.003
	b	0.033	0.022	0.015	0.012	0.011
1.0	a	0.009	0.007	0.005	0.004	0.004
	b	0.029	0.019	0.013	0.011	0.009
2.0	a	0.010	0.007	0.005	0.004	0.004
	b	0.027	0.018	0.013	0.010	0.009
2.5	a	0.010	0.007	0.005	0.004	0.004
	b	0.027	0.017	0.012	0.010	0.009
5.0	a	0.011	0.007	0.005	0.004	0.004
	b	0.025	0.016	0.012	0.010	0.008
10.0	a	0.008	0.007	0.006	0.005	0.004
	b	0.024	0.015	0.011	0.009	0.008

Tabela 4.3: Validação (%) para tamanhos de amostra variáveis

α(%)	Estatística	Tamanho da amostra				
		1.000	2.500	5.000	7.500	10.000
0.1	CORR	0.102	0.112	0.086	0.082	0.112
	BKR	0.11	0.104	0.088	0.092	0.102
	CHI	0.106	0.098	0.088	0.08	0.122
	KAC	0.106	0.098	0.088	0.08	0.122
	LR	0.092	0.094	0.084	0.074	0.12
	KS	0.1	0.1	0.1	0.1	0.1
1.0	CORR	0.992	1.062	0.972	0.944	0.986
	BKR	1.012	1.058	0.932	0.912	0.984
	CHI	1.13	1.15	0.986	0.964	1.068
	KAC	1.13	1.15	0.986	0.964	1.068
	LR	0.984	1.066	0.962	0.936	1.048
	KS	1.0	1.0	1.0	1.0	1.0
2.0	CORR	2.084	2.11	1.938	1.894	1.994
	BKR	2.062	2.12	1.96	1.886	1.932
	CHI	2.316	2.264	2.04	1.982	1.994
	KAC	2.316	2.264	2.04	1.982	1.994
	LR	2.000	2.116	1.974	1.914	1.946
	KS	2.0	2.0	2.0	2.0	2.0
2.5	CORR	2.586	2.588	2.35	2.394	2.414
	BKR	2.638	2.67	2.416	2.396	2.456
	CHI	2.878	2.776	2.57	2.454	2.408
	KAC	2.878	2.776	2.57	2.454	2.408
	LR	2.448	2.604	2.474	2.382	2.362
	KS	2.5	2.5	2.5	2.5	2.5
5.0	CORR	5.114	5.134	4.892	4.982	4.854
	BKR	5.082	5.074	4.85	4.814	4.784
	CHI	5.792	5.468	4.998	5.092	5.008
	KAC	5.792	5.468	4.998	5.092	5.008
	LR	5.01	5.114	4.812	4.966	4.896
	KS	5.0	5.0	5.0	5.0	5.0
10.0	CORR	10.07	10.058	9.942	9.872	9.824
	BKR	10.072	10.1	9.8	9.848	9.834
	CHI	11.26	10.74	10.124	10.134	10.094
	KAC	11.26	10.74	10.124	10.134	10.094
	LR	9.616	10.04	9.686	9.828	9.864
	KS	10.0	10.0	10.0	10.0	10.0

quantis da própria distribuição simulada.

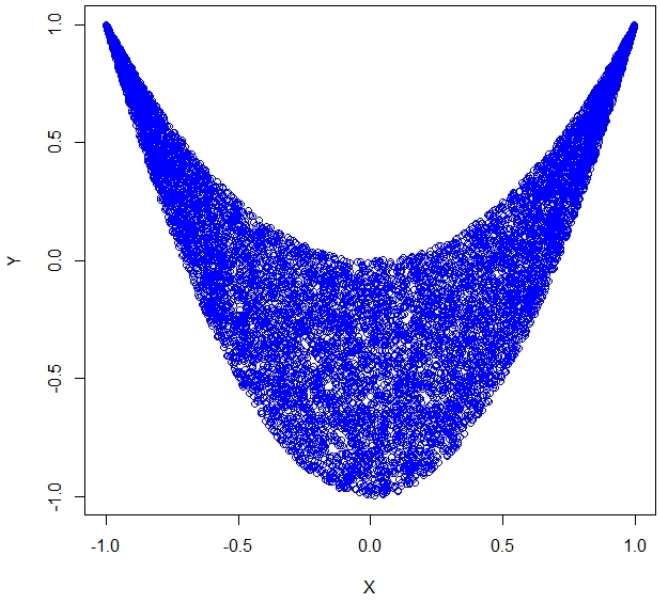


Figura 4.1: Valores gerados para a verificação de poder do teste

4.3 Poder do teste

Para avaliar o poder do teste foram realizadas simulações para dois exemplos distintos: para o primeiro caso, a forma da distribuição imita um sorriso e, no segundo caso, possui semelhança com um cata-vento.

Exemplo 4.1. *1. Considere as funções*

$$\begin{aligned} f(x) &= \frac{1}{2} \quad se \quad |x| \leq 1 \\ f(y|x) &= \frac{1}{1-x^2} \quad para \quad 2x^2 - 1 \leq y \leq x^2 \end{aligned} \tag{4.7}$$

Portanto, foram gerados dois vetores $x = 2u - 1$ *e* $y = (u-1)(1-x^2) + x^2$*, em que u possui distribuição uniforme[0,1]*

A Fig. 4.1 mostra a forma da distribuição gerada e a Tab. 4.4 mostra os resultados. Dentre os testes semelhantes (LR,CHI e KAC) todos apresentaram excelente desempenho em detectar a estrutura de dependência não linear. Para esse exemplo, o teste de BKR e KS também apresentaram bom desempenho. No entanto, o CORR não apresentou bons resultados, confirmando não ser o teste ideal para a identificação de estruturas de dependência não linear.

Tabela 4.4: Poder do teste (%) para tamanhos de amostra variáveis

α(%)	Estatística	Tamanho da amostra				
		1.000	2.500	5.000	7.500	10.000
0.1	CORR	0.55	0.528	0.454	0.5	0.504
	BKR	100	100	100	100	100
	CHI	100	100	100	100	100
	KAC	100	100	100	100	100
	LR	100	100	100	100	100
	KS	100	100	100	100	100
1.0	CORR	2.794	2.764	2.692	2.8	2.76
	BKR	100	100	100	100	100
	CHI	100	100	100	100	100
	LR	100	100	100	100	100
	KAC	100	100	100	100	100
	KS	100	100	100	100	100
2.0	CORR	4.698	4.784	4.758	4.71	4.62
	BKR	100	100	100	100	100
	CHI	100	100	100	100	100
	LR	100	100	100	100	100
	KAC	100	100	100	100	100
	KS	100	100	100	100	100
2.5	CORR	5.508	5.684	5.622	5.554	5.518
	BKR	100	100	100	100	100
	CHI	100	100	100	100	100
	LR	100	100	100	100	100
	KAC	100	100	100	100	100
	KS	100	100	100	100	100
5.0	CORR	9.4	9.556	9.598	9.274	9.328
	BKR	100	100	100	100	100
	CHI	100	100	100	100	100
	LR	100	100	100	100	100
	KAC	100	100	100	100	100
	KS	100	100	100	100	100
10.0	CORR	15.732	16.266	16.2	16.038	15.952
	BKR	100	100	100	100	100
	CHI	100	100	100	100	100
	LR	100	100	100	100	100
	KAC	100	100	100	100	100
	KS	100	100	100	100	100

Exemplo 4.2. *Considere a densidade* $f(x,y) = f(x)f(y) + g(x)h(y) - g(y)h(x)$, *sendo:*

$$
\begin{aligned}
f(x) &= \frac{1}{2} \quad se \quad |x| \leq 1 \\
f(y) &= \frac{1}{2} \quad se \quad |y| \leq 1 \\
g(x) &= A \times sign(x) \\
h(x) &= x
\end{aligned}
$$

em que $sign(x) = +1$*, se* $x > 0$*, e* -1*, se* $x \leq 0$*. Portanto;*

$$
\begin{aligned}
f(x,y) &= \frac{1}{4} + A \times [sign(x)y - sign(y)x] \quad se \quad |x| \quad e \quad |y| \leq 1 \\
f(x,y) &\geq 0 \quad \rightarrow \quad 0 \leq A \leq 1/4
\end{aligned} \tag{4.8}
$$

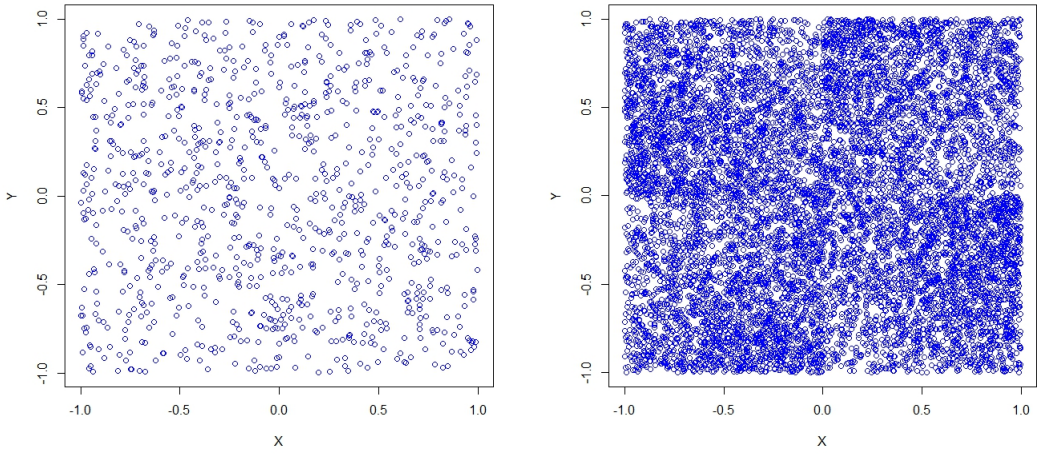
Para a geração dos dados, consideramos o tamanho da amostra igual a 5.000 e a quantidade de réplicas em 10.000. O parâmetro A *sofreu variação de* 0 *a* 0.25 *com incremento de* 0.05*, sendo que* $A = 0$ *indica ausência de dependência e* $A = 0.25$ *indica dependência extrema. A Fig. 4.2 mostra três cenários para distintos valores de* A*.*

A Tab. 4.5 reúne os resultados para o poder do teste sob essas condições. Nota-se que os testes semelhantes (LR,CHI e KAC) apresentam melhor desempenho, em identificar estruturas de dependência, que o teste clássico de HBKR. Os resultados para o KS (conforme realizado) foram muito bons, sendo mais sensíveis em pontos próximos da mediana da distribuição do que nas caudas.

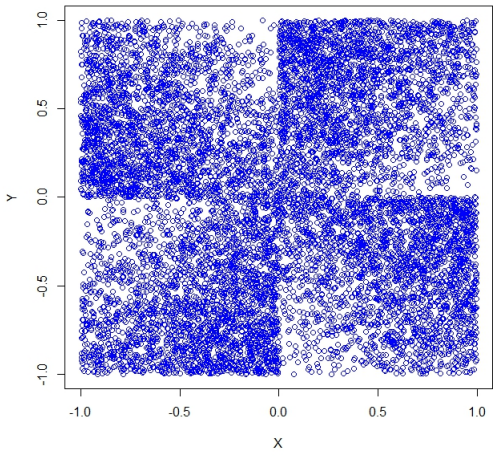
A estatística CORR, assim como no Ex. 4.1, não foi capaz de identificar estruturas de dependência não linear. Entre as demais estatísticas, todas tiveram 100% de acerto para valores de $A \geq 0.15$*. Para valores da constante menores que 0.15, as estatísticas semelhantes LR,CHI e KAC registraram valores um pouco melhores que os de KS.*

4.4 Considerações

Este capítulo apresentou um estudo de validação e poder do teste proposto por Matsushita et al. (2012), realizando comparações com outros testes: o que se baseia no coeficiente de correlação (e, portanto, não ideal para a detecção de dependência não



(a) A = 0 **(b)** A=0.13



(c) A=0.25

Figura 4.2: Representações gráficas para os dados gerados no exemplo do Cata-vento, sendo a) ausência de dependência, b) dependência moderada e c) dependência extrema

Tabela 4.5: Poder empírico (%) para diferentes valores de A

α(%)	Estatística	Variações do parâmetro					
		$A = 0$	$A = 0.05$	$A = 0.10$	$A = 0.15$	$A = 0.20$	$A = 0.25$
0.1	CORR	0.102	0.04	0.04	0.05	0.06	0.06
	BKR	0.11	0.5	49.28	100	100	100
	CHI	0.106	0.78	72.23	100	100	100
	KAC	0.106	0.78	72.23	100	100	100
	LR	0.092	0.76	71.62	100	100	100
	KS	0.1	2.98	79.9	100	100	100
1.0	CORR	0.992	1.02	1.01	1.05	1.1	1.1
	BKR	1.012	5.48	97.87	100	100	100
	CHI	1.13	8.99	99.25	100	100	100
	KAC	1.13	8.99	99.25	100	100	100
	LR	0.984	8.68	99.19	100	100	100
	KS	1.0	13.01	97.77	100	100	100
2.0	CORR	2.084	1.94	1.96	1.9	1.87	1.96
	BKR	2.062	11.84	99.71	100	100	100
	CHI	2.316	18.71	99.88	100	100	100
	KAC	2.316	18.71	99.88	100	100	100
	LR	2.000	18.09	99.87	100	100	100
	KS	2.0	19.64	99.44	100	100	100
2.5	CORR	2.586	2.52	2.56	2.55	2.47	2.45
	BKR	2.638	15.61	99.85	100	100	100
	CHI	2.878	22.99	99.94	100	100	100
	KAC	2.878	22.99	99.94	100	100	100
	LR	2.448	22.27	99.94	100	100	100
	KS	2.5	22.4	99.63	100	100	100
5.0	CORR	5.114	5.01	4.93	4.89	4.88	4.78
	BKR	5.082	32.29	100	100	100	100
	CHI	5.792	42.18	100	100	100	100
	KAC	5.792	42.18	100	100	100	100
	LR	5.01	41.04	100	100	100	100
	KS	5.0	34.12	99.94	100	100	100
10.0	CORR	10.07	9.79	9.81	9.73	9.68	9.81
	BKR	10.072	58.6	100	100	100	100
	CHI	11.26	65.01	100	100	100	100
	KAC	11.26	65.01	100	100	100	100
	LR	9.616	63.94	100	100	100	100
	KS	10.0	50.49	99.99	100	100	100

linear); teste clássico de HBKR (que possui desempenho um pouco inferior ao teste proposto). Além disso, buscou-se comparar, ainda, o desempenho da adaptação feita ao teste de Kolmogorov-Smirnov a fim de testar a hipótese de independência entre duas variáveis.

Para a validação, gerou-se duas distribuições uniformes independentes considerando diferentes tamanhos de amostra. Ao aplicar os testes, como esperado, os valores nominais flutuaram em torno dos valores teóricos, com exceção dos valores para a estatística KS que tiveram valores coincidentes, devido o desconhecimento dos níveis críticos.

Para o poder do teste foram utilizados dois exemplos. No primeiro caso, em que o gráfico da função imita um sorriso, todos os testes foram capazes de identificar a estrutura de dependência não linear, exceto o teste baseado no coeficiente de correlação linear (ideal apenas para identificar dependências lineares). A distribuição dos dados no segundo exemplo possui semelhança com um cata-vento, representando uma estrutura de dependência não linear mais complexa: a formulação da função apresenta uma contante de calibração do grau de dependência existente entre as variáveis. Dentre os testes utilizados neste trabalho, novamente o teste baseado no coeficiente de correlação de Pearson não foi capaz de identificar a estrutura de dependência não linear. Os demais testes apresentaram bons resultados, no entanto, o teste proposto por Matsushita et al. (2012) foi superior aos demais, mesmo que por diferenças mínimas.

Capítulo 5

Ilustrações

Como ilustração dos testes apresentados, foram utilizadas algumas taxas de câmbio das principais moedas em relação ao dólar americano.

Taxa de câmbio é definida como o preço da moeda de um país em relação à moeda de outro país. Em geral, essa taxa é determinada no mercado de câmbio e as transações, com o exterior, influenciam diretamente as variações na mesma. São participantes do mercado de câmbio: importadores/exportadores; turistas/empresas de turismo, investidores estrangeiros, participantes de empréstimos, entre outros.

A taxa de câmbio possui papel importantíssimo para o comércio internacional. Ela estabelece a competitividade entre os produtos e, portanto, os governos buscam mantê-la de forma que possa trazer benefícios para o país. Assim, se a taxa está elevada, é mais vantajoso importar que exportar. Caso a taxa esteja baixa, ocorre o inverso. Outra relação ocorre com a taxa de juros: quanto maior for a taxa de juros interna, em relação à externa (mantendo a taxa de câmbio fixa), maior a chance da entrada de capitais externos no país.

Intervenções na moeda local para determinada direção por causa de uma outra moeda, cria uma espécie de dependência entre elas. Por exemplo, a ocorrência de recessão em um país tende a aumentar a dependência entre pares de moedas. Se o dólar fortalece, os investidores passam a adquirí-lo mais, causando depreciação nas outras moedas e gerando, portanto, dependência entre as mesmas. Essas dependências, em geral, não pertencem a um ambiente gaussiano e, de tal forma, não podem ser corretamente mensuradas com medidas que captam apenas o grau de dependência

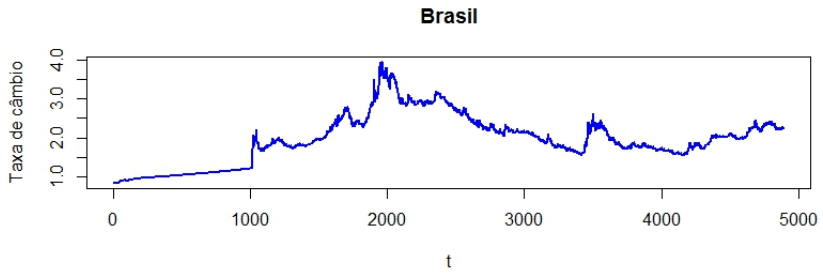


Figura 5.1: Evoluções diárias das taxas de câmbio X_t do real em relação ao dólar americano

linear.

Neste capítulo não abordaremos o caso de dependência entre as moedas. Em vez disso, abordaremos o caso de dependência entre os retornos de *lag* 1.

A Tab. 5.1 apresenta informações sobre diversas moedas, das quais utilizaremos algumas para fazer a demonstraçao dos testes estudados. Esses dados de taxas de câmbio, cotadas pelo Federal Reserv Bank of New York, indicam o valor necessário de cada moeda para adquirir um dólar ao meio-dia em Nova York. Por exemplo, se a taxa de câmbio em determinado dia é equivalente a $R\$1.5$, então são necessários 1 real e 50 centavos para adquirir 1 dólar nessa data.

Indicaremos por X_t o valor registrado como a taxa de câmbio, sendo t o tempo de referência para cada taxa, ignorando períodos sem registros. O retorno logarítmico, portanto, será representado por $R_t = \ln(X_t) - \ln(X_{t-1})$.

Tabela 5.1: Séries de taxas de câmbio

País	Moeda	Data início	Data fim	Tamanho da amostra
África do Sul	Rand	04-01-1971	20-06-2014	11.341
Austrália	Dólar australiano	04-01-1971	20-06-2014	11.340
Brasil	Real	02-01-1995	20-06-2014	6.385
Canadá	Dólar canadense	04-01-1971	20-06-2014	11.340
México	Peso	08-11-1993	20-06-2014	5.380
Suiça	Franco Suíço	04-01-1971	20-06-2014	10.343
Taiwan	Dólar de Taiwan	30-10-1983	20-06-2014	8.015
Zona do Euro	Euro	04-01-1999	20-06-2014	6.385

As Fig. 5.1 e 5.2 representam, respectivamente, as séries referentes à taxa de câmbio do real e o retorno logarítmico dessa taxa.

Ao plotar o gráfico de dispersão entre os retornos R_t e R_{t-1} da taxa de câmbio bra-

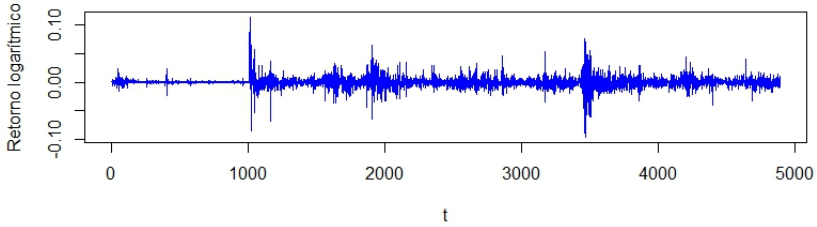


Figura 5.2: Evoluções diárias dos retornos logarítmicos R_t do real

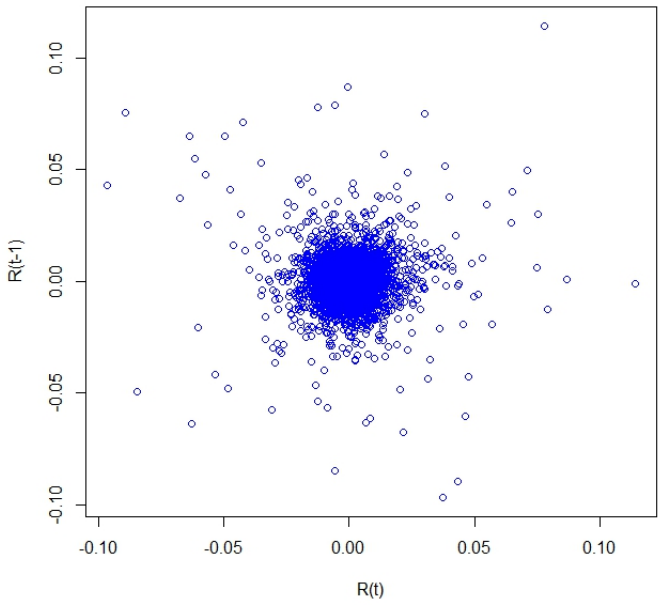


Figura 5.3: Retornos R_t e R_{t-1} para o real

sileira é possível ver que as autocorrelações de primeira ordem, isto é, $\rho = Corr(R_t, R_{t-1})$ não apresentam nenhum padrão, como mostra a Fig. 5.3

O teste baseado no coeficiente de correlação não é estatisticamente significativo, apresentando estatística CORR=0.0136. No entanto, como é sabido, a ausência de correlação linear não implica independência entre as variáveis, pois pode existir, ainda, alguma espécie de associação não linear entre elas. Para efeito de comparação, nesse exemplo foi considerada a série de taxa de câmbio brasileira truncada (utilizando os dados até a data de 16 de março de 2012), assim como apresentado em Matsushita et al. (2012).

Portanto, embora o teste de correlação continue a não rejeitar a hipótese nula de

Figura 5.4: Evoluções diárias das taxas de câmbio X_t do euro em relação ao dólar americano

independência, ao aplicar os testes de dependência não linear, foram obtidos resultados significativos. A estatística BKR registrou o valor 0.69 (valor superior à 0.12 -valor crítico a 1% de significância). As estatísticas semelhantes CHI e KAC registraram valores próximo a 25 e a estatística LR também. Todos esses valores são altos e indicam que a hipótese de independência deve ser rejeitada.

As Fig. 5.4 e 5.5 são referentes à taxa de câmbio e seu retorno logarítmico para a Zona do euro. Assim como para o real, a Fig. 5.6 também não apresenta nenhum padrão que indique algum tipo de dependência linear. De fato, o teste baseado na correlação de Pearson (CORR=0.009) não apresenta evidências para se rejeitar a hipótese nula de independência. Para essa moeda, os testes capazes de identificar dependência não linear também não apresentaram resultados significativos. As estatísticas LR, CHI, KAC e BKR registraram valores muito baixos, inferiores aos níveis críticos dessas estatísticas para todos os níveis críticos citados nas Tab. 3.9 e 2.1, respectivamente. Resultado semelhante foi registrado por Matsushita et al. (2012) para um tamanho de amostra menor.

Analisando os dados para a moeda de Taiwan, a Fig. 5.9 também não apresenta nenhum padrão de dependência linear entre os retornos (CORR= -0.027). Porém, ao aplicar os testes citados, capazes de identificar estruturas de dependência não linear entre as observações, foi constatado que há fortes indícios para se rejeitar a hipótese de independência. O valor para as estatísticas semelhantes foi 16.69 e, para o teste de Matsushita et al. (2012), a estatística LR apresentou valor igual a 16.55. Já a estatística do teste de HBKR registrou o valor 0.75. Enfim, todos esses resultados

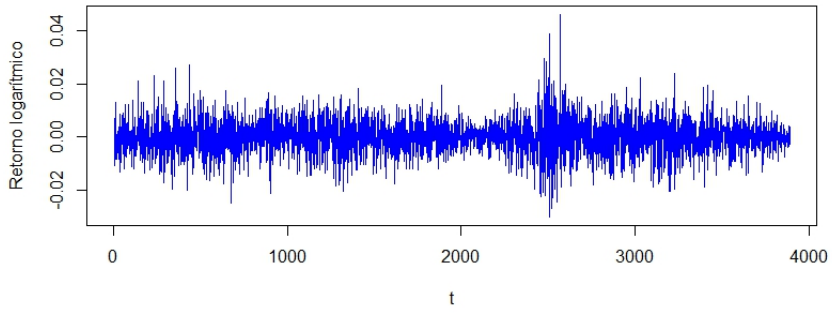


Figura 5.5: Evoluções diárias dos retornos logarítmicos R_t do euro

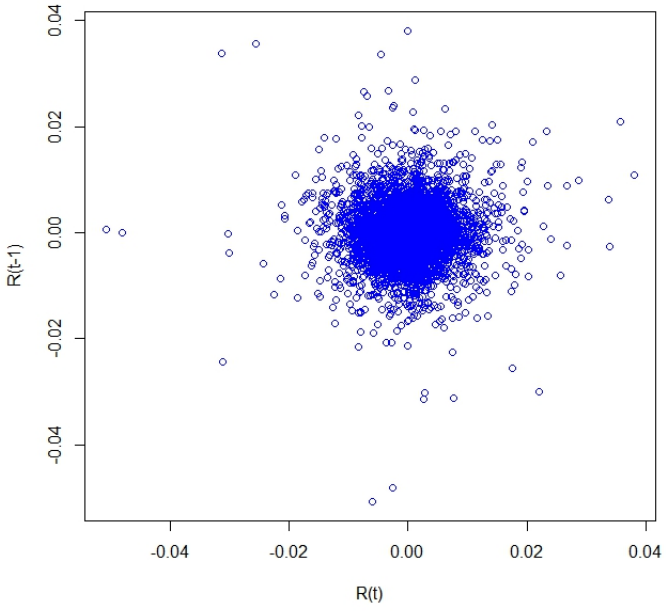


Figura 5.6: Retornos R_t e R_{t-1} do euro

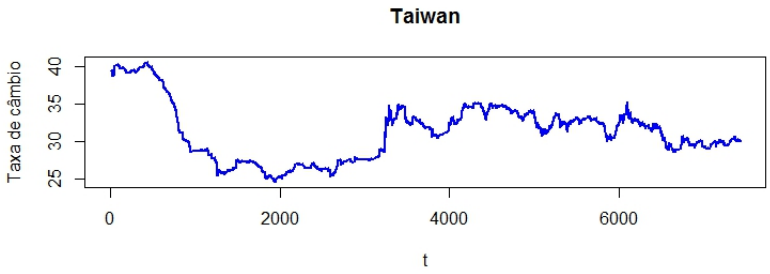


Figura 5.7: Evoluções diárias das taxas de câmbio X_t do dólar de Taiwan em relação ao dólar americano

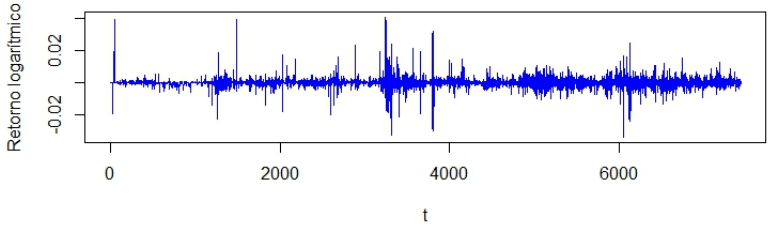


Figura 5.8: Evoluções diárias dos retornos logarítmicos R_t do dólar de Taiwan

são grandes o suficiente para se rejeitar a hipótese de independência bivariada.

Outra moeda para o qual se identificou a existência de dependência não linear, mesmo na ausência de dependência linear, como mostra a Fig. 5.12, foi o Dólar Canadense. A estatística CORR foi igual a 0.019, valor bem menor que os das outras moedas - indicando que não se deve rejeitar a hipótese nula de independência. Já os demais testes, apresentaram valores ainda maiores para as estatísticas em estudo, variando de 0.44, para a estatística BKR, a 92.96, para as estatísticas semelhantes (CHI e KAC). A estatística LR também apresentou valor grande o suficiente para indicar a existência de dependência não linear entre os retornos de $lag1$.

5.1 Considerações

Este capítulo apresentou ilustrações, com um exemplo real utilizando o *software* R, ao buscar identificar a existência de dependência não linear em dados relacionados à taxa de câmbio. Os exemplos comprovam que mesmo na ausência de dependência linear,

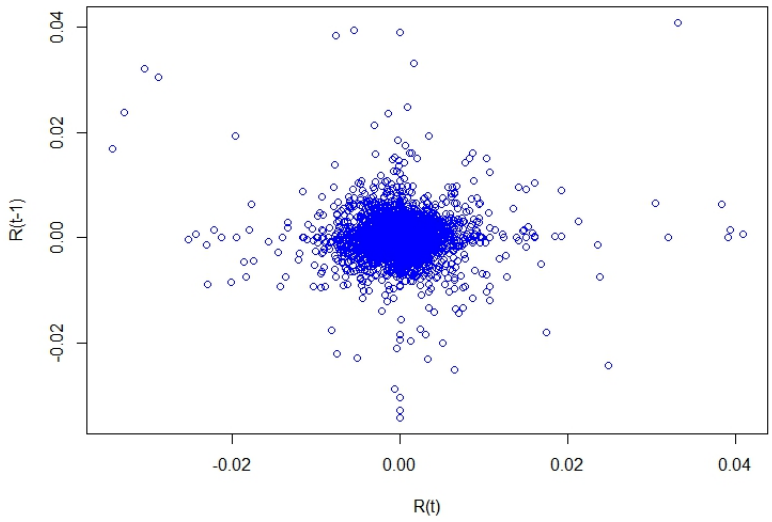


Figura 5.9: Retornos R_t e R_{t-1} do dólar Taiwan

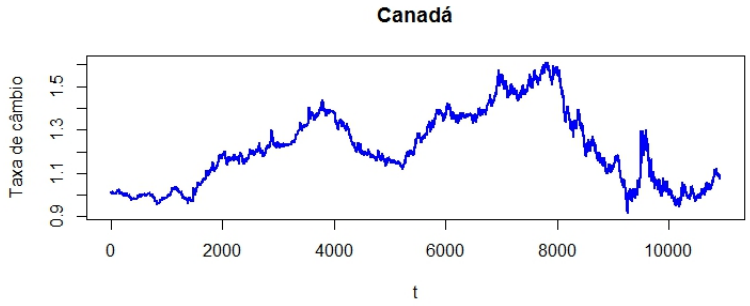


Figura 5.10: Evoluções diárias das taxas de câmbio X_t do dólar canadense em relação ao dólar americano

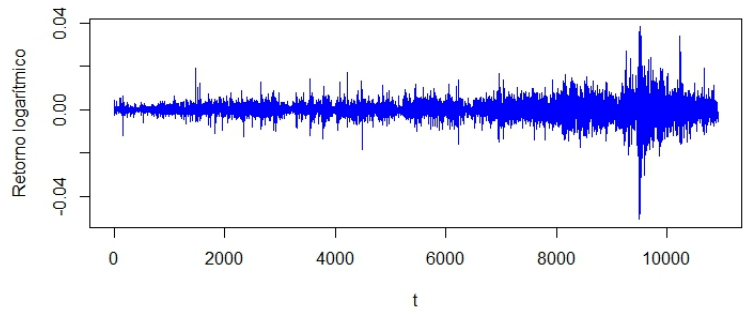


Figura 5.11: Evoluções diárias dos retornos logarítmicos R_t do dólar canadense

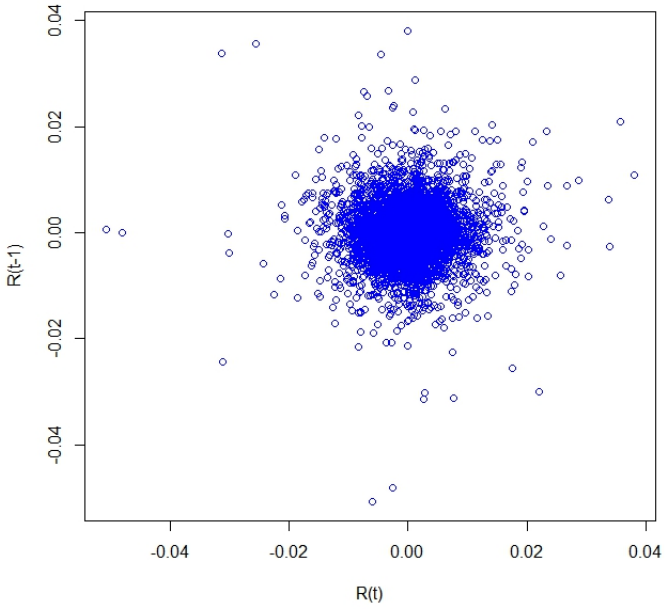


Figura 5.12: Retornos R_t e R_{t-1} do dólar canadense

os dados ainda podem apresentar dependência não linear. Ou seja, o teste baseado no coeficiente de correlação linear de Pearson não é suficiente para determinar se duas variáveis são ou não independentes.

Todos os demais testes, baseados nas estatísticas BKR, CHI, KAC e LR se mostraram eficientes em identificar estruturas de dependência não linear, quando de fato existiam. Embora a estatística KS tenha mostrado bons resultados nas simulações apresentadas no Cap. refcap:simu, optou-se por não utilizá-la neste capítulo, devido ao desconhecimento de seus níveis críticos teóricos, após a adaptação da estatística de Kolmogorov-Smirnov para avaliação de estruturas de dependência.

Considerações Finais

Este trabalho abordou métodos estatísticos para a detecção de dependência não linear. Estudamos, mediante simulações de Monte Carlo, o desempenho de testes não paramétricos do tipo *distribution-free*. Consideramos duas classes de testes; uma se baseia na forma de Crámer-von Mises, e a outra é uma variação do teste de Kolmogorov-Smirnov (KS).

Discutimos que há diversos testes que permitem identificar a dependência linear entre as observações, mas aqueles que se destinam à detecção do caso não linear são mais escassos. Neste trabalho nos restringimos ao teste de HBKR e a uma variação dele. Ambos são capazes de detectar a presença de dependência não linear com poder de teste significativo. No entanto, o segundo, proposto por Matsushita et.al (2012), possui poder estatístico ainda maior que o de HBKR.

O teste proposto por Matsushita et al. (2012), apresentado no Cap. 3, está formalizado para o caso bivariado, pois a função de covariância é conhecida. A nossa contribuição foi considerar a estatística de KS, e buscar outros exemplos de estruturas de dependência para o estudo do poder dos testes.

Porém, não há conhecimento da distribuição da estatística de KS para o teste de independência bivariada. Sugere-se que estudos posteriores sejam realizados com o propósito de determiná-la. A utilização dos quantis das distribuições simuladas apresenta resultados menos confiáveis do que a dos níveis críticos fornecidos pela distribuição assintótica teórica da estatística de Kolmogorov-Smirnov sob H_0.

Outra contribuição deste trabalho foi a demonstração de semelhança entre as estatísticas CHI e KAC, possibilitando assim um ganho no tempo de processamento computacional. A utilização da linguagem de programação Fortran também proporcionou maior velocidade na execução das simulações.

As simulações de Monte Carlo desenvolvidas neste trabalho confirmam os resultados obtidos por Matsushita et al. (2012): o teste baseado na razão de verossimilhança apresenta melhor desempenho que o teste clássico de HBKR, e confirma também a equivalência assintótica entre as estatísticas CHI e LR.

As ilustrações a partir de dados de taxa de câmbio de algumas moedas em relação ao dólar americano torna visivelmente mais clara a idéia defendida de que a não existência de correlação linear não implica independência entre as variáveis.

Em virtude da escassez de bons testes para a identificação de dependência entre três ou mais variáveis, tem-se grande interesse em estender esse teste e, caso a função de covariância não seja encontrada, o estudo da distribuição amostral da estatística LR poderá ser feito empiricamente mediante simulações de Monte Carlo.

Referências Bibliográficas

Agresti, A. *An introduction to categorical data analysis*. John Wiley & Sons, 2007.

Bakirov, N. K., Rizzo, M. L., and Székely, G. J. A multivariate nonparametric test of independence. *Journal of Multivariate Analysis 97* (2006), 1742–1756.

Beran, R., Bilodeau, M., and de Micheaux, P. Nonparametric tests of independence between random vectors. *Journal of Multivariate Analysis 98* (2006), 1805–1824.

Bilodeau, M., and Micheaux, P. A multivariate empirical characteristic function test of independence with normal marginals. *Journal of Multivariate Analysis* (2005).

Blum, J. R., Kiefer, J., and Rosenblatt, M. Distribution free test of independence based on the sample distribution function. In *The annals of mathematical statistics* (1960).

Brockwell, P., and Davis, R. *Time Series: theory and methods*, 2nd ed. Springer, 2006.

Bussab, W. O., and Morettin, P. A. *Estatística básica*, 5ª ed. Saraiva, 2006.

Casella, G., and Berger, R. L. *Inferência estatística*, tradução da 2ª edição norte-americana ed. Cengage Learning, 2010.

Chan, N. H., and Tran, L. Nonparametric tests for serial dependence. *J. Time Ser. Anal. 13* (1992), 19–28.

Conover, J. W. *Practical nonparametric statistics*, 3ªed ed. Wiley, 1999.

Cressie, N. *Statistics for spatial Data.* Wiley, 1993.

Csörgő, S. Testing for independence by the empirical characteristic function. *Journal of Multivariate Analysis 16*, 3 (1985), 290–299.

De Wet, T. Cramér-von mises tests for independence. *Journal of Multivariate Analysis 10* (1980), 38–50.

Delgado, M. A. Testing serial independence using the sample distribution function. *Journal of Time Series Analysis 17* (1996), 271–286.

Diggle, P., Liang, K., and Zeger, S. *Analysis of longitudinal data.* Oxford, 1996.

Draper, N., and Smith, H. *Applied regression analysis.* Wiley, 1998.

Genest, C., Quessy, J., and Rémillard, B. Local efficiency of a cramér-von mises test of independence. *Journal of Multivariate Analysis 97* (2006), 274–294.

Ghoudi, K., Kulperger, R. J., and Rémillard, B. A nonparametric test of serial independence for time series and residuals. *Journal of Multivariate Analysis 79* (2001), 191–218.

Gieser, P. W., and Randles, R. H. A nonparametric test of independence between two vectors. *Journal of the American Statistical Association 92* (1997), 561–567.

Gnedenko, B. The theory of probability. *The theory of probability* (1973).

Gretton, A., and Györfi, L. Consistent nonparametric tests of independence. *Journal of Machine Learning Research 11* (2010), 1391–1423.

Hoeffding, W. A non-parametric test of independence. In *The annals of Mathematical Statistics* (1948), vol. 19, pp. 546–557.

Johnson, R. A., and Wichern, D. W. *Applied multivariate statistical analysis.*, 6ª ed. Pearson International Edition., 2007.

Kac, M. On some connections between probability theory and differential and integral equations. In *Proceedings of the Second Berkeley Symposium of Mathematical Statistics and Probability* (1951), pp. 180–215.

Matsushita, R., Figueiredo, A., and Da Silva, S. A suggested statistical test for measuring bivariate nonlinear dependence. *Physica A 391* (2012), 4891–4898.

Mood, A., Graybill, F., and Boes, D. *Introduction to the theory of statistics*, 3rd edition ed. McGraw-Hill, 1987.

Nelsen, R. *An Introduction to Copulas*, 2nd ed. Springer, 2006.

Papadatos, N., and Xifara, T. A simple method for obtaining the maximal correlation coefficient and related characterizations. *Journal of Multivariate Analysis 118* (2013), 102–114.

Parzen, E. *Stochastic processes*. Holden-Day, 1962.

R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

Robinson, P. M. Consistent nonparametric entropy - based testing. *The Review of Economic Studies 58* (1991), 437–453.

Roussas, G. *A course in mathematical statistics*, 2nd edition ed. Academic Press, 1997.

Santos, S. Estudo comparativo de medidas de dependência e aplicações em dados de expressão gênica. Master's thesis, Instituto de Matemática e Estatística - USP, 2012.

Scaillet, O. A kolmogorov-smirnov type test for positive quadrant dependence. *The Canadian Journal of Statistics 33* (2005), 415–427.

Schmid, F., and Schmidt, R. Multivariate extension of spearman's rho and related statistics. *Statistics & Probability Letters* (2007).

Schweizer, B., and Wolff, E. F. On nonparametric measures of dependence for random variables. *The Annals of Statistics Vol. 9, No. 4* (1981), 879–885.

Skaug, H. J., and Tjøstheim, D. A nonparametric test of serial independence based on the empirical distribution function. *Biometrika 80*, 3 (1993), 591–602.

Um, Y., and Randles, R. A multivariate nonparametric test of independence among many vectors. *Nonparametric Statist. 13* (2001), 699–708.

Zhang, Z. Quotient correlation: a sample based alternative to pearson's correlation. *The annals of statistics 36* (2008), 1007–1030.

Apêndice A

Medidas e Índices de Dependência

A.1 Condições para índices de dependência

1. O índice $\delta(X, Y)$ pode ser obtido para qualquer par de v.as., isto é, nenhum dos termos deve ser uma constante

2. $\delta(X, Y) = \delta(Y, X)$ representa a propriedade simétrica de independência mútua. No entanto, a dependência completa considera a direção da associação. Por exemplo, $Y = \pm X$ se $X \sim N(0, 1)$

3. O intervalo de valores para o índice é definido como $0 \leq \delta(X, Y) \leq 1$

4. $\delta(X, Y) = 0$ se, e somente se, X e Y forem mutuamente independentes. Se o índice assume o outro valor extremo ($\delta(X, Y) = 1$), então essas variáveis são completamente associadas

5. Caso X e Y sejam conjuntamente normais então a não-correlação implica independência, portanto $\delta(X, Y) = |\rho|$

6. Se as funções mensuráveis Borel $g(X)$ e $h(Y)$ são funções de $\Re$ em $\Re$, 1-1 (biunívocas) então $\delta(g(X), h(Y)) = \delta(X, Y)$. Ou seja, o índice permanece invariante sob a transformação proposta.

A.2 Índices

A.2.1 Correlação máxima

O coeficiente de correlação máxima foi introduzido por Gebelein como $\rho^*(X,Y) = \sup \text{Corr}(g_1(X), g_2(Y))$. O supremo é calculado sobre todas as funções Borel mensuráveis g_1 e $g_2 : \Re \to \Re$ ambas com variância positiva e finita.

O coeficiente de correlação máxima possui fundamental importância em diferentes áreas da estatística, tais como: obter transformações ótimas para a análise de regressão, além de estar relacionado à teoria de convergência do algoritmo amostrador de Gibbs em análise bayesiana (Papadatos e Xifara, 1994).

Segundo Johnson e Wichern (2007) a análise de correlação canônica foi inicialmente desenvolvida por H. Hotelling e a correlação máxima coincide com o primeiro par de variáveis canônicas. Esse par é composto pelas v.as. $U = a'X$ e $V = b'Y$ tais que U e V maximizem a correlação $\rho = \text{Corr}(U, V)$.

Apesar de sua importância, o cálculo do coeficiente de forma explícita não é trivial, como afirmou Papadatos e Xifara.

A.2.2 Correlação tetracórica

O coeficiente de correlação tetracórico foi introduzido por Karl Pearson e pode ser considerado como uma estimativa do coeficiente de correlação entre duas variáveis latentes[1], ambas originalmente contínuas e normais, porém efetivamente observadas como variáveis dicotômicas, ou seja, o coeficiente é uma medida de associação para variáveis contínuas, porém transformadas em tabela 2×2.

O cálculo do coeficiente é facilitado pelo uso de tabelas de contingência que permite utilizar o valor das frequências para cada par de modalidade possível entre as variáveis em estudo.

As suposições básicas quanto à utilização desse coeficiente são: as variáveis latentes devem ser contínuas e normalmente distribuídas, com relação linear entre si; as variáveis devem ser dicotomizadas (ao serem medidas) o mais próximo possível à

[1] Variável que não pode ser mensurada diretamente, tais como inteligência, ansiedade, conhecimento. Assim, constroem-se construtos capazes de serem medidos

mediana.

O Coeficiente de Correlação Tetracórico é menos confiável (possui erro padrão maior) do que o de Pearson, pois sua variabilidade é aproximadamente 50% maior quando o coeficiente de Pearson é igual a zero. Assim, para obter a mesma confiabilidade que a obtida no Coeficiente de Correlação de Pearson, é necessário o dobro do tamanho da amostra.

A.2.3 Cramér

O índice de dependência de Cramér é definido por:

$$\mu = \sigma_x \sigma_y \int \int (f(x,y) - r(x)c(y))^2 dxdy \tag{A.1}$$

em que $r(x)$ e $c(y)$ são funções contínuas escolhidas para tornar o valor da Eq. (A.1) mínimo.

Esse índice pode ser utilizado para avaliar a capacidade de ajustamento de uma função de distribuição acumulada $F(x)$ comparada com uma função de distribuição empírica $S(x)$ (amostra única), para se comparar duas distribuições empíricas (2 amostras) ou até mesmo em algoritmos que buscam a estimação de distância mínima.

A.2.4 Hoeffding

Uma medida de dependência para variáveis retangularmente distribuídas[2] foi definida por Hoeffding como:

$$\Phi^2 = 90 \int_0^1 \int_0^1 (F(x,y) - xy)^2 dxdy \tag{A.2}$$

No caso em que as variáveis não possuem essa distribuição é necessário utilizar uma transformação monótona para obtê-la. Essa é uma medida consistente de dependência, pois $\Phi^2 = 1$ se, e somente se, as variáveis forem perfeitamente dependentes.

[2] Apresenta a mesma densidade para todos os valores dentro do intervalo $[\mu - a; \mu + a]$ e zero fora do intervalo.

Uma outra forma de definir essa medida, como consta em Santos (2012), é conhecida como medida D de Hoefding - medida de associação entre duas $v.a_s$ que é calculada a partir das amostras x e y .

$$D = \frac{(n-2)(n-3)D_1 + D_2 - 2(n-2)D_3}{n(n-1)(n-2)(n-3)(n-4)} \tag{A.3}$$

sendo:

$D_1 = \sum_i = 1^n Q_i(Q_i - 1)$

Q_i o número de pontos com ambos valores de x e y menores que o i-ésimo ponto

$D_2 = \sum_i = 1^n (R_i - 1)(R_i - 2)(S_i - 1)(S_i - 2)$

R_i o posto de x_i

S_i o posto de y_i

$D_3 = \sum_i = 1^n (R_i - 2)(S_i - 2)Q_i$

D é uma medida da distância entre $F(XY)$ e $F(X)F(Y)$. Caso essa distância seja nula, tem-se a independência entre as variáveis.

Apêndice B

Programas

B.1 Validação do teste

```
c       1         2         3         4         5         6         7
c234567890123456789012345678901234567890123456789012345678901234567890123456789012
```

```
      program BKRv5
      use ieee_arithmetic
      integer,parameter :: seed = 81348530
      integer i,j,k,l
      integer n,m
      integer seed2
      real T0,TF
      real deltax,deltay
      real C1, C2, C3, C4, Chisq
      real L1, L2, L3, L4, LL
      real KAK
      real n00,n11,n10,n01,e00,e01,e10,e11,nx,ny
      real MeanX, MeanY, MeanXY, MeanX2, MeanY2,VarX, VarY
      real fx,fy,fxy,fcx,fcy,fcxy,fxcy,fcxcy
      real,dimension(10000) :: x,y,KS1,KS2,KS3,KS4
```

```
real,dimension(50000) :: BKR,CHI,LR,KS,KSM,CORR,KAC
character*25 output
call system('cls')
print*, 'BKR versao 1.5 ----------------------------------------'
print*, 'R. Matsushita/L.Rocha                                2014'
print*, '------------------------------------------------------'
print*, 'Simulacao das estatisticas de testes de independencia '
print*, 'para amostras IID'
print*, '------------------------------------------------------'
print*, '------------------------------------------------------'
print *, ' '
print *, '(1) entre com o tamanho da amostra (max=10000)'
read *, n
print *, ' '
print *, '(2) entre com a quantidade de replicacoes (max=50000)'
read *, m
print *, ' '
print *, '(3) entre com a semente (no. inteiro)'
read *, seed2
print *, ' '
print *, '(4) nome do arquivo de saida (max=25 char)'
read *, output
print *, ' '
print*, 'calculando...'
do i = 1,10000,1
KS1(i)=0
KS2(i)=0
KS3(i)=0
KS4(i)=0
end do
call cpu_time(T0)
do l = 1,m,1
```

```
call srand(seed+seed2+l)
MeanX = 0
MeanY = 0
MeanXY= 0
MeanX2 = 0
MeanY2 = 0
do i = 1,n,1
x(i) = rand()
y(i) = rand()
MeanX = MeanX+x(i)
MeanY = MeanY+y(i)
MeanXY= MeanXY+x(i)*y(i)
MeanX2 = MeanX2+x(i)*x(i)
MeanY2 = MeanY2+y(i)*y(i)
end do
MeanX = MeanX/n
MeanY = MeanY/n
MeanXY= MeanXY/n
MeanX2 = MeanX2/n
MeanY2 = MeanY2/n
VarX  = MeanX2-MeanX**2
VarY  = MeanY2-MeanY**2
Corr(l)=(MeanXY-MeanX*MeanY)/(sqrt(VarX*VarY))
BKR(l)  = 0
KAC(l)  = 0
CHI(l)  = 0
LR(l)   = 0
do i = 1,n,1
nx     = 0
ny     = 0
n11    = 0
n01    = 0
```

```
n10    = 0
n00    = 0
do k = 1,n,1
deltax=0
deltay=0
if(x(k) .le. x(i)) deltax=1
if(y(k) .le. y(i)) deltay=1
nx      = nx  + deltax
ny      = ny  + deltay
n11     = n11 + deltax*deltay
n00     = n00 + (1-deltax)*(1-deltay)
n10     = n10 + deltax*(1-deltay)
n01     = n01 + (1-deltax)*deltay
end do
fx      = nx/n
fy      = ny/n
fxy     = n11/n
fcx     = 1 - fx
fcy     = 1 - fy
fcxy    = fy - fxy
fxcy    = fx - fxy
fcxcy   = 1 - fxy - fxcy - fcxy
e11     = n*fx*fy
e00     = n*(1-fx)*(1-fy)
e10     = n*fx*(1-fy)
e01     = n*(1-fx)*fy
BKR(l)  = BKR(l) + (fxy-fx*fy)**2
KAK     = ((fxy-fx*fy)**2)/((1-fx)*(1-fy)*fx*fy)
if (ieee_is_nan(KAK) ) KAK=0
KAC(l)  = KAC(l) + KAK
C1 = ((n11-e11)**2)/e11
C2 = ((n10-e10)**2)/e10
```

```
      C3 = ((n01-e01)**2)/e01
      C4 = ((n00-e00)**2)/e00
      L1 = n11*ALOG(n11/e11)
      L2 = n10*ALOG(n10/e10)
      L3 = n01*ALOG(n01/e01)
      L4 = n00*ALOG(n00/e00)
      Chisq = C1 + C2 + C3 + C4
      LL    = L1 + L2 + L3 + L4
      if (ieee_is_nan(Chisq) ) Chisq=0
      if (ieee_is_nan(LL) )    LL=0
      CHI(l) = CHI(l) + Chisq
      LR(l)  = LR(l)  + 2*LL
      KS1(i)    = abs(fxy-fx*fy)
      KS2(i)    = abs(fcxy-fcx*fy)
      KS3(i)    = abs(fxcy-fx*fcy)
      KS4(i)    = abs(fcxcy-fcx*fcy)
      end do
      KS(l)  = maxval(KS1)
      KSM(l) = max(maxval(KS1),maxval(KS2),maxval(KS3),maxval(KS4))
      CHI(l) = CHI(l)/n
      LR(l)  = LR(l)/n
c      print *, 'rep', l, ' de ', m
       write(*,"(a)",advance="no") '.'
      end do
      print *, ' --- ok ---'
      call cpu_time(TF)
      print *, 'tempo de execucao ', TF-T0, 'segundos.'
      print *, 'gravando...'
      open(unit=10,file=output,status='UNKNOWN')
      write(10,*) 'BKR  ','CHI  ','LR   ','KS   ','KSM  ','Corr  ','KAC'
      do l = 1,m,1
      write(10,*) BKR(l),CHI(l),LR(l),KS(l),KSM(l),Corr(l),KAC(l)
```

```
      end do
      close(10)
c      open(unit=20,file='lastdata.dat',status='UNKNOWN')
c      do i = 1,n,1
c      write(20,*) x(i), y(i)
c      end do
c      close(20)
      print *, 'fim.'
      stop
      end
```

B.2 Poder do teste: Smile

```
c        1         2         3         4         5         6         7
c2345678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
```

```
      program PowerSmilev1
      use ieee_arithmetic
      integer,parameter :: seed = 81398984
      integer i,j,k,l
      integer n,m
      integer seed2
      real T0,TF
      real deltax,deltay
      real C1, C2, C3, C4, Chisq
      real L1, L2, L3, L4, LL
      real KAK
      real n00,n11,n10,n01,e00,e01,e10,e11,nx,ny
      real MeanX, MeanY, MeanXY, MeanX2, MeanY2,VarX, VarY
      real fx,fy,fxy,fcx,fcy,fcxy,fxcy,fcxcy
      real,dimension(10000) :: x,y,KS1,KS2,KS3,KS4
```

```
real,dimension(50000) :: BKR,CHI,LR,KS,KSM,CORR,KAC
character*25 output
call system('cls')
print*, 'PowerSmile versao 1.0-----------------------------------'
print*, 'R. Matsushita/L. Rocha                                2014'
print*, '-------------------------------------------------------'
print*, 'Simulacao das estatisticas de testes de independencia '
print*, 'para amostras dependentes (modelo smile)'
print*, '-------------------------------------------------------'
print*, '-------------------------------------------------------'
print *, ' '
print *, '(1) entre com o tamanho da amostra (max=10000)'
read *, n
print *, ' '
print *, '(2) entre com a quantidade de replicacoes (max=50000)'
read *, m
print *, ' '
print *, '(3) entre com a semente (no. inteiro)'
read *, seed2
print *, ' '
print *, '(4) nome do arquivo de saida (max=25 char)'
read *, output
print *, ' '
print*, 'calculando...'
do i = 1,10000,1
KS1(i)=0
KS2(i)=0
KS3(i)=0
KS4(i)=0
end do
call cpu_time(T0)
do l = 1,m,1
```

```
call srand(seed+seed2+l)
MeanX = 0
MeanY = 0
MeanXY= 0
MeanX2 = 0
MeanY2 = 0
do i = 1,n,1
x(i) = 2*rand()-1
y(i) = (rand()-1)*(1-x(i)*x(i))+ x(i)*x(i)
MeanX = MeanX+x(i)
MeanY = MeanY+y(i)
MeanXY= MeanXY+x(i)*y(i)
MeanX2 = MeanX2+x(i)*x(i)
MeanY2 = MeanY2+y(i)*y(i)
end do
MeanX = MeanX/n
MeanY = MeanY/n
MeanXY= MeanXY/n
MeanX2 = MeanX2/n
MeanY2 = MeanY2/n
VarX  = MeanX2-MeanX**2
VarY  = MeanY2-MeanY**2
Corr(l)=(MeanXY-MeanX*MeanY)/(sqrt(VarX*VarY))
BKR(l)  = 0
KAC(l)  = 0
CHI(l)  = 0
LR(l)   = 0
do i = 1,n,1
nx     = 0
ny     = 0
n11    = 0
n01    = 0
```

```
n10    = 0
n00    = 0
do k = 1,n,1
deltax=0
deltay=0
if(x(k) .le. x(i)) deltax=1
if(y(k) .le. y(i)) deltay=1
nx      = nx  + deltax
ny      = ny  + deltay
n11     = n11 + deltax*deltay
n00     = n00 + (1-deltax)*(1-deltay)
n10     = n10 + deltax*(1-deltay)
n01     = n01 + (1-deltax)*deltay
end do
fx      = nx/n
fy      = ny/n
fxy     = n11/n
fcx     = 1 - fx
fcy     = 1 - fy
fcxy    = fy - fxy
fxcy    = fx - fxy
fcxcy   = 1 - fxy - fxcy - fcxy
e11     = n*fx*fy
e00     = n*(1-fx)*(1-fy)
e10     = n*fx*(1-fy)
e01     = n*(1-fx)*fy
BKR(l)  = BKR(l) + (fxy-fx*fy)**2
KAK     = ((fxy-fx*fy)**2)/((1-fx)*(1-fy)*fx*fy)
if (ieee_is_nan(KAK) ) KAK=0
KAC(l)  = KAC(l) + KAK
C1 = ((n11-e11)**2)/e11
C2 = ((n10-e10)**2)/e10
```

```
C3 = ((n01-e01)**2)/e01
C4 = ((n00-e00)**2)/e00
L1 = n11*ALOG(n11/e11)
L2 = n10*ALOG(n10/e10)
L3 = n01*ALOG(n01/e01)
L4 = n00*ALOG(n00/e00)
Chisq = C1 + C2 + C3 + C4
LL    = L1 + L2 + L3 + L4
if (ieee_is_nan(Chisq) ) Chisq=0
if (ieee_is_nan(LL) )    LL=0
CHI(l) = CHI(l) + Chisq
LR(l)  = LR(l)  + 2*LL
KS1(i)    = abs(fxy-fx*fy)
KS2(i)    = abs(fcxy-fcx*fy)
KS3(i)    = abs(fxcy-fx*fcy)
KS4(i)    = abs(fcxcy-fcx*fcy)
end do
KS(l)  = maxval(KS1)
KSM(l) = max(maxval(KS1),maxval(KS2),maxval(KS3),maxval(KS4))
CHI(l) = CHI(l)/n
LR(l)  = LR(l)/n
write(*,"(a)",advance="no") '.'
end do
print *, ' --- ok ---'
call cpu_time(TF)
print *, 'tempo de execucao ', TF-T0, 'segundos.'
print *, 'gravando...'
open(unit=10,file=output,status='UNKNOWN')
write(10,*) 'BKR  ','CHI  ','LR   ','KS   ','KSM  ','Corr  ','KAC'
do l = 1,m,1
write(10,*) BKR(l),CHI(l),LR(l),KS(l),KSM(l),Corr(l),KAC(l)
end do
```

```
close(10)
open(unit=20,file='lastdata.dat',status='UNKNOWN')
do i = 1,n,1
write(20,*) x(i), y(i)
end do
close(20)
print *, 'fim.'
stop
end
```

B.3 Poder do teste: Cata-vento

```
c        1         2         3         4         5         6         7

c234567890123456789012345678901234567890123456789012345678901234567890123456789012
```

```
      program PowerKolmogorov1
      use ieee_arithmetic
      integer,parameter :: seed = 81398984
      integer i,j,k,l
      integer n,m
      integer seed2
      real T0,TF
      real deltax,deltay
      real C1, C2, C3, C4, Chisq
      real L1, L2, L3, L4, LL
      real KAK
      real sgny,U,A,delta,F0x,Ix,Iy,beta2,beta1,beta0,ypos,yneg
      real n00,n11,n10,n01,e00,e01,e10,e11,nx,ny
      real MeanX, MeanY, MeanXY, MeanX2, MeanY2,VarX, VarY
      real fx,fy,fxy,fcx,fcy,fcxy,fxcy,fcxcy
      real,dimension(10000) :: x,y,KS1,KS2,KS3,KS4, xaux
```

```
real,dimension(50000) :: BKR,CHI,LR,KS,KSM,CORR,KAC
character*25 output
call system('cls')
print*, 'PowerSmile versao 1.0-----------------------------------'
print*, 'R. Matsushita/L.Rocha                                2014'
print*, '-------------------------------------------------------'
print*, 'Simulacao das estatisticas de testes de independencia '
print*, 'para amostras dependentes com densidade conjunta'
print*, 'f(x,y) = 0.25 + A*[sgn(x)*y - sgn(y)*x]'
print*, '-------------------------------------------------------'
print*, '-------------------------------------------------------'
print *, ' '
print *, '(0) entre com o parametro A (0<A<=0.25)'
read *, A
print *, ' '
print *, '(1) entre com o tamanho da amostra (max=10000)'
read *, n
print *, ' '
print *, '(2) entre com a quantidade de replicacoes (max=50000)'
read *, m
print *, ' '
print *, '(3) entre com a semente (no. inteiro)'
read *, seed2
print *, ' '
print *, '(4) nome do arquivo de saida (max=25 char)'
read *, output
print *, ' '
print*, 'calculando...'
do i = 1,10000,1
KS1(i)=0
KS2(i)=0
KS3(i)=0
```

```
KS4(i)=0
end do
call cpu_time(T0)
do l = 1,m,1
call srand(seed+seed2+l)
MeanX = 0
MeanY = 0
MeanXY= 0
MeanX2 = 0
MeanY2 = 0
do i = 1,n,1
x(i)  = 2*rand()-1
Ix = 1
Iy = 1
if(x(i) < 0) Ix = 0
F0x = (0.5 - sign(1.0,x(i))*A + 2*A*x(i))
U = rand()
if(U < F0x) Iy = 0
beta2 = A*(2*Ix-1)
beta1 = 2*A*x(i)-4*A*x(i)*Iy+0.5
beta0 = 0.5-A*sign(1.0,x(i))+2*x(i)*A - U
delta=beta1**2-4*beta0*beta2
ypos = (-beta1 + sqrt(delta))/(2*beta2)
yneg = (-beta1 - sqrt(delta))/(2*beta2)
y(i) = ypos
if(abs(ypos)>1) y(i) = yneg
MeanX = MeanX+x(i)
MeanY = MeanY+y(i)
MeanXY= MeanXY+x(i)*y(i)
MeanX2 = MeanX2+x(i)*x(i)
MeanY2 = MeanY2+y(i)*y(i)
end do
```

```
MeanX = MeanX/n
MeanY = MeanY/n
MeanXY= MeanXY/n
MeanX2 = MeanX2/n
MeanY2 = MeanY2/n
VarX  = MeanX2-MeanX**2
VarY  = MeanY2-MeanY**2
Corr(l)=(MeanXY-MeanX*MeanY)/(sqrt(VarX*VarY))
BKR(l)  = 0
KAC(l)  = 0
CHI(l)  = 0
LR(l)   = 0
do i = 1,n,1
nx     = 0
ny     = 0
n11    = 0
n01    = 0
n10    = 0
n00    = 0
do k = 1,n,1
deltax=0
deltay=0
if(x(k) .le. x(i)) deltax=1
if(y(k) .le. y(i)) deltay=1
nx      = nx  + deltax
ny      = ny  + deltay
n11     = n11 + deltax*deltay
n00     = n00 + (1-deltax)*(1-deltay)
n10     = n10 + deltax*(1-deltay)
n01     = n01 + (1-deltax)*deltay
end do
fx      = nx/n
```

```
fy      = ny/n
fxy     = n11/n
fcx     = 1 - fx
fcy     = 1 - fy
fcxy    = fy - fxy
fxcy    = fx - fxy
fcxcy   = 1 - fxy - fxcy - fcxy
e11     = n*fx*fy
e00     = n*(1-fx)*(1-fy)
e10     = n*fx*(1-fy)
e01     = n*(1-fx)*fy
BKR(l)  = BKR(l) + (fxy-fx*fy)**2
KAK     = ((fxy-fx*fy)**2)/((1-fx)*(1-fy)*fx*fy)
if (ieee_is_nan(KAK) ) KAK=0
KAC(l)  = KAC(l) + KAK
C1 = ((n11-e11)**2)/e11
C2 = ((n10-e10)**2)/e10
C3 = ((n01-e01)**2)/e01
C4 = ((n00-e00)**2)/e00
L1 = n11*ALOG(n11/e11)
L2 = n10*ALOG(n10/e10)
L3 = n01*ALOG(n01/e01)
L4 = n00*ALOG(n00/e00)
Chisq = C1 + C2 + C3 + C4
LL    = L1 + L2 + L3 + L4
if (ieee_is_nan(Chisq) ) Chisq=0
if (ieee_is_nan(LL) )    LL=0
CHI(l) = CHI(l) + Chisq
LR(l)  = LR(l)  + 2*LL
KS1(i)    = abs(fxy-fx*fy)
KS2(i)    = abs(fcxy-fcx*fy)
KS3(i)    = abs(fxcy-fx*fcy)
```

```
KS4(i)    = abs(fcxcy-fcx*fcy)
end do
KS(l)  = maxval(KS1)
KSM(l) = max(maxval(KS1),maxval(KS2),maxval(KS3),maxval(KS4))
CHI(l) = CHI(l)/n
LR(l)  = LR(l)/n
write(*,"(a)",advance="no") '.'
end do
print *, ' --- ok ---'
call cpu_time(TF)
print *, 'tempo de execucao ', TF-T0, 'segundos.'
print *, 'gravando...'
open(unit=10,file=output,status='UNKNOWN')
write(10,*) 'BKR  ','CHI  ','LR   ','KS   ','KSM  ','Corr  ','KAC'
do l = 1,m,1
write(10,*) BKR(l),CHI(l),LR(l),KS(l),KSM(l),Corr(l),KAC(l)
end do
close(10)
open(unit=20,file='lastdata.dat',status='UNKNOWN')
do i = 1,n,1
write(20,*) x(i), y(i)
end do
close(20)
print *, 'fim.'
stop
end
```