Luís Augusto Correia Roque

# Métodos Inferenciais Para o Coeficiente de Correlação $\rho_\omega$


Departamento de Matemática Aplicada
Faculdade de Ciências da Universidade do Porto
Maio de 2003

Luís Augusto Correia Roque

# Métodos Inferenciais Para o Coeficiente de Correlação $\rho_\omega$


Tese submetida á Faculdade de Ciências da Universidade do Porto para obtenção do grau de Mestre em Estatística

Departamento de Matemática Aplicada
Faculdade de Ciências da Universidade do Porto
Maio de 2003

À minha família, de modo especial,
aos meus pais e irmãos

# Agradecimentos

Apresentando desde já um pedido de desculpa àqueles que a memória não guardou, gostaria de agradecer a uma série de pessoas que em muito contribuíram para a realização deste trabalho:

ao meu orientador Doutor Joaquim Pinto da Costa, pela disponibilidade que sempre demonstrou para me aconselhar, corrigir e orientar;

ao Dr. Carlos Soares, do Centro de Informática da Universidade do Porto, pela colaboração prestada no domínio da programação;

ao Doutor Craig Borkowf, membro da Divisão de Epidemiologia e Aplicações Clínicas, *National Heart, Lung, and Blood Institute ( **NHLBI**), USA*, pela cedência do código do programa usado para o estudo do coeficiente de correlação de Spearman;

e a todos os professores, colegas de mestrado, amigos e familiares por todo o apoio que me prestaram.

# Resumo

O coeficiente de correlação de ordens, $\rho_w$, é uma estatística proposta recentemente (Costa § Soares, 2002) para dar resposta a determinados problemas em que o uso do coeficiente de correlação de Spearman, $\rho_s$, não é muito adequado. Neste documento elaboramos um estudo sobre a distribuição de $r_w$, estimador de $\rho_w$, usando alguns métodos não paramétricos de estimação. Apresentamos o método "empirical bivariate quantile-partitioned" (EBQP) adaptado para a estimação da variância de $r_w$ em amostras finitas. São efectuadas simulações que nos permitem apresentar resultados para a variância de $r_w$ e intervalos de confiança a 90% construídos para $\rho_w$. Comparamos os resultados fornecidos pelo método EBQP com os resultados fornecidos pelos algoritmos de bootstrap e jackknife. No caso de independência dos dados, estes resultados demonstram que a adaptação do método EBQP para a estimação da variância de $r_w$ é bem sucedida. Ao longo deste documento, paralelamente ao estudo desenvolvido para $\rho_w$, enunciamos as conclusões de um estudo similar relativo ao coeficiente de correlação de Spearman (Borkowf, 2000).

# Abstract

Weighted rank correlation, $\rho_w$, has been proposed recently as a statistic (Costa § Soares, 2002) in order to answer some problems where the use of the Spearman rank correlation , $\rho_s$, is not very adequate. We study the distribution of $r_w$ , estimator of $\rho_w$ , using some nonparametric methods. We propose the "empirical bivariate quantile-partitioned" (EBQP) method adapted for estimating the variance of $r_w$ for finite samples. We present extensive simulations to study the estimation of the sample variance of $r_w$. We compare the results for the EBQP method with those for the bootstrap and jackknife algorithms. In the case of independence of the variables, these results demonstrate that EBQP method can be successfully adapted for the estimation of the sample variance $r_w$. We construct 90% confidence intervals for $\rho_w$ . Simultaneously to the study developed to $\rho_w$, we also present conclusions of a similar study for the Spearman rank correlation coefficient (Borkowf, 2000).

# Conteúdo

# INTRODUÇÃO

Em dois artigos, (Spearman,1904) e (Spearman,1906), Spearman introduziu o coeficiente de correlação de ordens (rankings), $\rho_s$, o qual veio a tornar-se uma das estatísticas não paramétricas mais usadas em estudos nas mais diversas áreas como a Medicina, Epidemiologia, Biologia, etc.

Fundamentalmente os coeficientes de correlação de ordens são medidas da tendência para os valores de duas sequências de dados crescerem ou decrescerem conjuntamente (coeficiente de correlação positivo), ou para crescerem numa das sequências e decrescerem na outra (coeficiente de correlação negativo).

Os coeficientes de correlação de ordens fornecem informação complementar e, por vezes, alternativa à que é fornecida pelo coeficiente de correlação teórico de Pearson. Utiliza-se quando os pares de variáveis apresentam medidas erradas, o relacionamento entre as variáveis tende a ser monótono mas não linear, as variáveis tomam valores de diferentes ordens de grandeza ou são de origem ordinal.

Ao longo do século passado até aos dias que correm, o coeficiente de correlação de Spearman tem permitido resolver uma grande variedade de problemas, mas é uma estatística na qual todos as ordens (ranks) assumem a mesma importância.

Nos últimos anos surgiram problemas em áreas como Meta-Aprendizagem, "Information Retrieval", Recomendação de Sistemas, entre outras. Nestes problemas é necessário atribuir maior importância às ordens mais baixas, o que faz com que a utilização do coeficiente de correlação de Spearman $\rho_s$ possa ser questionada. Assim, Costa & Soares (2002) propuseram um coeficiente de correlação, designado por $\rho_w$, com funções lineares que constituem pesos para penalizarem os erros nas ordens mais baixas.

A presente dissertação constitui, numa primeira fase, uma compilação de alguns resultados conhecidos e aspectos mais importantes relacionados com os coeficientes $\rho_s$, $\rho_w$, e as variâncias dos seus estimadores $r_s$ e $r_w$, respectivamente. Numa fase posterior, num

estudo baseado em simulações, estimamos a variância de $r_w$ e construímos intervalos de confiança a 90% para $\rho_w$. Este estudo é similar ao realizado por Borkowf (2000) para o coeficiente de correlação de Spearman $\rho_s$ e do qual nós expômos algumas conclusões, de forma resumida, a finalizar os terceiro e quarto capítulos.

Assim, no primeiro capítulo, é definido $r_s$ para a situação em que existem empates de ordens e para o caso contrário. Seguem-se, de forma resumida, algumas situações em que é conhecida a variância de $r_s$.

No início do segundo capítulo é dado um exemplo, no campo da Meta-Aprendizagem, em que se pretende a recomendação de uma sequência de algoritmos com vista a resolver determinados problemas. Neste exemplo, a aplicação do coeficiente $\rho_w$ é mais adequada do que a aplicação do coeficiente $\rho_s$. Posteriormente, dedica-se especial atenção à distribuição de $r_w$ sob a hipótese dos dois vectores de ordens serem independentes ($H_0$). Também é proposto o uso de um método para estudo da distribuição exacta de $r_w$ sob $H_0$. Trata-se de um método já usado para estudo da distribuição de $r_s$.

O terceiro capítulo foi destinado à estimação não paramétrica da variância de $r_w$. Inicialmente são descritas as distribuições bivariadas donde são provenientes os conjuntos de dados simulados. Também descrevemos os métodos de estimação não paramétrica usados: método EBQP e os algoritmos de jackknife e bootstrap. Posteriormente são apresentados os resultados e as conclusões obtidos em simulações efectuadas usando o software Gauss 5.0 (Aptech Systems).

No quarto capítulo, descrevemos os métodos empregues na construção dos intervalos de confiança. Nestes métodos são usados os resultados do capítulo anterior, nomeadamente, as variâncias de $r_w$ estimadas pelos diferentes métodos.

As diversas conclusões tiradas ao longo do documento, são reunidas num capítulo final que resume o que se sabe a respeito do coeficiente de correlação $\rho_w$ e o que há para estudar.

# Capítulo 1

# Coeficiente de correlação de Spearman $\rho_S$

O coeficiente de correlação $\rho_s$, introduzido por Spearman (1904, 1906), tornou-se numa das estatísticas não paramétricas mais usadas. Tem sido frequente a sua utilização em estudos ligados à Medicina, Epidemiologia, Biologia, Psicologia e Ciências Sociais. A seguir é definido o seu estimador $r_s$ e é elaborado um resumo acerca do que se conhece sobre a variância de $r_s$.

## 1.1 Definição da estatística $r_s$

Consideremos uma amostra de dados $(x_i, y_i)$ bivariada de dimensão $n$. Se substituirmos os valores de $x_i$ pela ordem $R_i$ que lhe está associada tendo em conta os dados marginais relativos à variável $X$, e procedermos de igual modo, substituindo cada $y_i$ pela respectiva ordem $Q_i$, então $r_s$ é definido por:

$$r_s = \frac{\sum_{i=1}^{n} \left(R_i - \bar{R}\right)\left(Q_i - \bar{Q}\right)}{\sqrt{\sum_{i=1}^{n} \left(R_i - \bar{R}\right)^2} \sqrt{\sum_{i=1}^{n} \left(Q_i - \bar{Q}\right)^2}} \tag{1.1}$$

Considerando $D = \sum_{i=1}^{n}(R_i - Q_i)^2$, no caso de não existirem ordens (ranks) marginais repetidas obtém-se a partir de (1.1)

$$r_s = 1 - \frac{6D}{n^3 - n} \tag{1.2}$$

No caso em que existem ordens marginais repetidas, a equação é ligeiramente mais complexa. Se $f_k$ é o número de empates no k-ésimo grupo de empates do vector formado pelos $R_i$s, e se $g_m$ é o número de empates no m-ésimo grupo de empates do vector formado pelos $Q_i$s então:

$$r_s = \frac{1 - \frac{6}{n^3-n}\left[D + \frac{1}{12}\sum_k (f_k^3 - f_k) + \frac{1}{12}\sum_m (g_m^3 - g_m)\right]}{\sqrt{1 - \frac{\sum_k (f_k^3 - f_k)}{n^3 - n}}\sqrt{1 - \frac{\sum_m (g_m^3 - g_m)}{n^3 - n}}} \tag{1.3}$$

Neste caso, para calcular $D$, as ordens das observações com valores empatados são dadas pela média das ordens que são atribuidas préviamente. Por exemplo, consideremos $n = 5$ e os valores da variável $X$: 3.1, 4.6, 1.0, 2.3, 3.1. As ordens $R_i$s associadas a esta amostra são: 3.5, 5, 1, 2, 3.5, respectivamente. A ordem das duas observações de valor 3.1 resulta da atribuição prévia da ordem 3 e da ordem 4 a estas observações e posterior cálculo da média: $\frac{3+4}{2} = 3.5$.

Note-se que no caso em que não existem empates a equação (1.3) reduz-se a (1.2).

A seguir apresentamos algumas situações em que é conhecida a variância de $r_s$.

## 1.2 Variância de $r_s$

Seja $r_s$ o estimador de $\rho_s$ e $n$ a dimensão da amostra. Em determinadas condições especiais são conhecidas fórmulas para a estimação da variância de $r_s$ em amostras finitas.

Se $n = 2$ , $F(x, y)$ a função distribuição bivariada, então:

$$var(r_s) = 1 - [E(r_s)]^2 . \tag{1.4}$$

No caso de independência dos dados, por Pearson (1907), temos:

$$var(r_s) = \frac{1}{n-1}. \tag{1.5}$$

Assim, no caso de independência, a variância assimptótica de $r_s$ é dada por:

$$\sigma_a^2(r_s) = \lim_{n\to\infty} var(\sqrt{n} r_s) = \lim_{n\to\infty} \frac{n}{n-1} = 1 \tag{1.6}$$

Neste caso, é de referir que, a partir de Randles & Wolf (1979), podemos concluir a normalidade assimptótica da distribuição de $r_s$.

Considerando agora o caso de não independência, para a distribuição normal bivariada $(BVN)$ com correlação $\rho$ , Moran (1948) ao escrever o estimador do coeficiente de correlação de Spearman $r_s$ da seguinte forma:

$$r_s = \frac{12\left(S - \frac{1}{4}n(n-1)^2\right)}{n\left(n^2-1\right)}, \tag{1.7}$$

em que

$$S = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{m=1}^{n} H(x_i - x_j)H(y_i - y_m), \tag{1.8}$$

com $H(t)$ tal que:

$$H(t) = \begin{cases} 0 & \text{se } t \leq 0 \\ 1 & \text{se } t > 0 \end{cases}, \tag{1.9}$$

determina $E(r_s)$ através do cálculo de $E(S)$.

Ainda para a distribuição normal bivariada $(BVN)$, usando o mesmo método, Kendall (1949) determina uma aproximação para $E(S^2)$, que por sua vez, permitiu obter uma aproximação polinomial para a variância assimptótica $\sigma_a^2(r_s)$. A aproximação é um polinómio de grau 8 cujos monómios são potências de base $\rho$ e expoentes pares.

Posteriormente, Fieller et al.(1957) obtém uma melhor aproximação polinomial que a de Kendall para $\sigma_a^2(r_s)$. A aproximação é um polinómio de grau 12 cujos monómios são potências de base $\rho$ e expoentes pares.

Para outras distribuições bivariadas, Borkowf (2002) usa o método de Kendall (1949) para determinar $\sigma_a^2(r_s)$. A variância assimptótica de $r_s$ é escrita como função de cinco esperanças de funções distribuição conjuntas e funções distribuição marginais.

Infelizmente, este método é de difícil aplicação pois envolve numerosos cálculos de valores esperados. Por outro lado, este método não pode ser usado quando a distribuição dos dados é desconhecida. Justifica-se assim o estudo e aplicação de alguns métodos de estimação não paramétrica da variância de $r_s$ que se encontram em Borkowf (2000). No capítulo 3 deste trabalho apresentamos um estudo idêntico para um outro coeficiente de correlação de ordens (rankings), o coficiente $\rho_w$.

No capítulo que se segue apresentamos motivação para o estudo do coeficiente $\rho_w$ e alguns resultados relacionados com a distribuição do seu estimador $r_w$.

# Capítulo 2

# Coeficiente $\rho_w$

Iniciamos este capítulo com um exemplo, na Meta-Aprendizagem. Para recomendação de algoritmos com vista a resolver determinados problemas, tem-se em conta os desempenhos desses algoritmos na resolução de problemas do género dos que pretendemos resolver (tempo de resolução, potência do CPU). Esses desempenhos são avaliados previamente. Por vezes os algoritmos recomendados falham por falta de memória ou "bugs"de software, o que obriga a que seja recomendado um conjunto de alternativas. Por isso, a recomendação de algoritmos consiste em ordenar os algoritmos de acordo com os desempenhos esperados, formando sequências de algoritmos que constituem, cada uma, um método para a resolução do problema.

Para avaliar o melhor método, a estratégia passa por considerar os vectores das ordens como um todo. Os coeficientes de correlação entre o vector das ordens dos algoritmos que compõe cada um dos métodos a propor e o verdadeiro vector das ordens permite concluir qual dos vectores está em maior conformidade com o verdadeiro vector das ordens. Contudo, o uso do coeficiente de correlação de Spearman não é o mais adequado pois assume que todos as ordens são igualmente importantes. Isso não é verdade na recomendação de algoritmos pois algoritmos cujas ordens são mais baixas têm maior probabilidade de serem experimentados. Portanto, os erros nas ordens mais baixas têm maior impacto do que os erros nas ordens mais elevadas. Daí resulta a necessidade da

utilização do coeficiente de correlação $\rho_w$ (Costa § Soares, 2002) para atribuir maior penalização aos erros em ordens mais baixas.

Posteriormente apresentamos resultados conhecidos sobre a distribuição do estimador $r_w$.

## 2.1 Exemplo de aplicação do coeficiente $\rho_w$

Suponhamos que temos um verdadeiro vector de ordens ($r(X_1)$, $r(X_2)$, ..., $r(X_{10})$) de 10 algoritmos ($A_1$ a $A_{10}$) para resolver um determinado problema. Este vector de ordens (ranking) é obtido por ordenação dos desempenhos de cada algoritmo na resolução do problema (tabela 1). Queremos saber qual dos vectores fornecidos por dois métodos distintos ($M_1$ e $M_2$) mais se aproxima do verdadeiro vector das ordens.

Tabela 1: O verdadeiro vector das ordens e os vectores a recomendar

| | | $M_1$ | $M_2$ | | |
|---|---|---|---|---|---|
| Algoritmos | $r(X_i)$ | $r(Y_i)$ | $r(Z_i)$ | $(r(X_i)-r(Y_i))^2$ | $(r(X_i)-r(Z_i))^2$ |
| $A_1$ | 1 | 2 | 3 | 1 | 4 |
| $A_2$ | 2 | 1 | 5 | 1 | 9 |
| $A_3$ | 3 | 4 | 2 | 1 | 1 |
| $A_4$ | 4 | 6 | 1 | 4 | 9 |
| $A_5$ | 5 | 5 | 6 | 0 | 1 |
| $A_6$ | 6 | 3 | 4 | 9 | 4 |
| $A_7$ | 7 | 8 | 7 | 1 | 0 |
| $A_8$ | 8 | 9 | 8 | 1 | 0 |
| $A_9$ | 9 | 10 | 9 | 1 | 0 |
| $A_{10}$ | 10 | 7 | 10 | 9 | 0 |

Usando o estimador do coeficiente de correlação de Spearman $r_s$, calculado a partir da seguinte expressão:

$$r_s = 1 - \frac{6\sum_{i=1}^{n}(r(X_i) - r(Y_i))^2}{n^3 - n}, \tag{2.1}$$

obtém-se os resultados $r_s = 0,8303$ para $M_1$ e $r_s = 0,8303$ para $M_2$, o que significa que os vectores $(r(Y_1),\ r(Y_2),\ ...,\ r(Y_{10}))$ e $(r(Z_1),\ r(Z_2),\ ...,\ r(Z_{10}))$ são igualmente boas aproximações para o verdadeiro vector de ordens $(r(X_1),\ r(X_2),\ ...,\ r(X_{10}))$.

Contudo, pela observação directa dos vectores de ordens fornecidos por $M_1$ e $M_2$, verifica-se que nas ordens mais baixas o vector associado a $M_1$ aproxima-se melhor ao verdadeiro vector de ordens. Como algoritmos com ordens mais baixas têm maior probabilidade de serem experimentados, parece-nos que a melhor sugestão é o método $M_1$. Assim temos uma situação em que o coeficiente de correlação de Spearman não permite sugerir o melhor método. Para a resolução deste problema justifica-se a utilização de um coeficiente de correlação que penalize os erros nas ordens mais baixas.

Costa & Soares (2002) propuseram um novo coeficiente de correlação , designado por $r_w$ , em que a distância entre duas ordens (ranks) é:

$$WD_i^2 = (r(X_i) - r(Y_i))^2((n - r(X_i) + 1) + (n - r(Y_i) + 1)) \tag{2.2}$$

que é diferente da distância entre duas ordens considerada no cálculo do coeficiente de correlação de Spearman:

$$D_i^2 = (r(X_i) - r(Y_i))^2 \tag{2.3}$$

Observe-se que o termo $((n - r(X_i) + 1) + (n - r(Y_i) + 1))$ em (2.2) representa a importância das ordens $r(X_i)$ e $r(Y_i)$.

Na tabela 2 são calculadas as novas distâncias entre a ordem $r(X_i)$ e as ordens $r(Y_i)$ e $r(Z_i)$.

Tabela 2: Novas distâncias entre as ordens.

| | | $M_1$ | | | $M_2$ | | |
|---|---|---|---|---|---|---|---|
| Algoritmos | $r(X_i)$ | $r(Y_i)$ | $D_i^2$ | $WD_i^2$ | $r(Z_i)$ | $D_i^2$ | $WD_i^2$ |
| $A_1$ | 1 | 2 | 1 | 19 | 3 | 4 | 72 |
| $A_2$ | 2 | 1 | 1 | 19 | 5 | 9 | 135 |
| $A_3$ | 3 | 4 | 1 | 15 | 2 | 1 | 17 |
| $A_4$ | 4 | 6 | 4 | 48 | 1 | 9 | 153 |
| $A_5$ | 5 | 5 | 0 | 0 | 6 | 1 | 11 |
| $A_6$ | 6 | 3 | 9 | 117 | 4 | 4 | 48 |
| $A_7$ | 7 | 8 | 1 | 7 | 7 | 0 | 0 |
| $A_8$ | 8 | 9 | 1 | 5 | 8 | 0 | 0 |
| $A_9$ | 9 | 10 | 1 | 3 | 9 | 0 | 0 |
| $A_{10}$ | 10 | 7 | 9 | 45 | 10 | 0 | 0 |

Com base nos dados registados na tabela 2 e usando o estimador do coeficiente $\rho_w$, cálculado a partir da seguinte expressão (Costa & Soares, 2002):

$$r_w = 1 - \frac{6\sum_{i=1}^{n}(r(X_i) - r(Y_i))^2((n - r(X_i) + 1) + (n - r(Y_i) + 1))}{n^4 + n^3 - n^2 - n}, \tag{2.4}$$

obtemos os resultados $r_w = 0,846832$ para o método $M_1$ e $r_w = 0,75978$ para o

método $M_2$. Isto significa que o vector de ordens (ranking) associado a $M_1$ é melhor que o vector de ordens associado a $M_2$, o que está em conformidade com a observação directa dos vectores.

O teorema seguinte garante-nos que $r_w$ só toma valores compreendidos entre -1 e 1, tal como acontece com $r_s$.

**Teorema 1**

(Costa & Soares,2002): A soma $\sum_{i=1}^{n}(r(X_i)-r(Y_i))^2((n-r(X_i)+1)+(n-r(Y_i)+1))$ é máxima quando $r(Y_i)=n-r(X_i)+1$ e assume o valor $\frac{n^4+n^3-n^2-n}{3}$.

A demonstração deste teorema (Apêndice A) encontra-se em Costa & Soares ( 2002)

Na secção seguinte apresentamos alguns resultados conhecidos respeitantes à distribuição de $r_w$.

## 2.2 Distribuição de $r_w$

Nesta secção vamos estudar a distribuição de $r_w$. Este estudo reporta-se sobretudo à hipótese de independência entre os dois vectores de ordens (hipótese $H_0$). No final da secção, tecemos algumas considerações sobre o comportamento da variância de $r_w$ para o caso da não independência dos dois vectores de ordens.

Começamos por designar $R_i = r(X_i)$ , $Q_i = r(Y_i)$ , $R = (R_1, ..., R_n)$ o primeiro vector de ordens e $Q = (Q_1, ..., Q_n)$ o segundo vector de ordens. Para determinar o valor esperado e a variância de $r_w$ é necessário recorrer a alguns resultados sobre estatísticas de ordens, apresentados em Randles & Wolf (1979) que passo a enunciar:

Se $S = \sum_{i=1}^{n} c(i)a(R_i^*)$ é uma estatística de ordens (ranks) em que as constantes $a(1), ..., a(n)$ são scores , $c(1), ..., c(n)$ constantes de regressão e $R^* = (R_1^*, ...., R_n^*)$ é uma ordenação aleatória, então:

$$
\begin{aligned}
&i) \quad E(S) = \bar{c}n\bar{a} \\
&ii) \quad var(S) = \tfrac{1}{n-1}\left[\sum_{i=1}^{n}(c(i)-\bar{c})^2\right]\left[\sum_{j=1}^{n}(a(j)-\bar{a})^2\right] \\
&iii) \quad cov(S,S') = \tfrac{1}{n-1}\left[\sum_{i=1}^{n}(c(i)-\bar{c})(c'(i)-\bar{c}')\right]\left[\sum_{j=1}^{n}(a(j)-\bar{a})(a'(j)-\bar{a}')\right],
\end{aligned}
\tag{2.5}
$$

em que $S' = \sum_{i=1}^{n} c'(i)\ a'(R_i^*)$

Com estes resultados é possível provar o seguinte teorema:

**Teorema 2**

(Costa & Soares, 2002): Se $R = (R_1, ..., R_n)$ e $Q = (Q_1, ..., Q_n)$ são vectores de ordens independentes, então

$i)$ $E(r_w) = 0$

$ii)$ $var(r_w) = \dfrac{31n^2 + 60n + 26}{30(n^3 + n^2 - n - 1)}$

A demonstração deste teorema (Apêndice B) encontra-se em Costa & Soares ( 2002).

Considerando

$$\begin{aligned} S_{1n} &= \sum_{i=1}^{n} iR_i^* \\ S_{2n} &= \sum_{i=1}^{n} iR_i^{*2} \\ S_{3n} &= \sum_{i=1}^{n} i^2 R_i^* \end{aligned} \tag{2.6}$$

podemos escrever:

$$\begin{aligned} r_w &= \frac{1}{n^4+n^3-n^2-n}[(n^4+n^3-n^2-n) - 24\frac{(n+1)^2 n(2n+1)}{6} \\ &+12\frac{n^2(n+1)^2}{4} + 24(n+1)(S_{1n} - E(S_{1n})) \\ &+24\frac{(n+1)^3 n}{4} - 6(S_{2n} - E(S_{2n})) - 6(S_{3n} - E(S_{3n})) - n(n+1)^2(2n+1)] \end{aligned} \tag{2.7}$$

o que nos permite concluir (Costa & Soares, 2002):

$$\frac{r_w}{\sqrt{var(r_w)}} = a_n \frac{S_{1n} - E(S_{1n})}{\sqrt{var(S_{1n})}} + b_n \frac{S_{2n} - E(S_{2n})}{\sqrt{var(S_{2n})}} + c_n \frac{S_{3n} - E(S_{3n})}{\sqrt{var(S_{3n})}} \tag{2.8}$$

em que:

$$\begin{aligned} a_n &\longrightarrow 2\sqrt{\frac{30}{31}} \\ b_n &\longrightarrow -\frac{1}{90}\sqrt{\frac{30}{31}} \\ c_n &\longrightarrow -\frac{1}{90}\sqrt{\frac{30}{31}} \end{aligned}$$

quando $n \longrightarrow \infty$.

Em Randles & Wolfe (1979) é provado que as distribuições das estatísticas:

$$\frac{S_{1n} - E(S_{1n})}{\sqrt{var(S_{1n})}},$$

$$\frac{S_{2n} - E(S_{2n})}{\sqrt{var(S_{2n})}} \text{ e}$$

$$\frac{S_{3n} - E(S_{3n})}{\sqrt{var(S_{3n})}}$$

se aproximam da distribuição $N(0,1)$ quando $n \to \infty$.

Em Billingsley (1978, pág.288), encontram-se resultados que nos permitem garantir que:

$$T_1 = a_n \frac{S_{1n} - E(S_{1n})}{\sqrt{var(S_{1n})}} \to N(0, \frac{120}{31}),$$

$$T_2 = b_n \frac{S_{2n} - E(S_{2n})}{\sqrt{var(S_{2n})}} \to N(0, \frac{1}{8370}) \text{ e}$$

$$T_3 = c_n \frac{S_{3n} - E(S_{3n})}{\sqrt{var(S_{3n})}} \to N(0, \frac{1}{8370})$$

Portanto, $\frac{r_w}{\sqrt{var(r_w)}}$ escreve-se como combinação linear de estatísticas que são assimptoticamente normais. Contudo, as estatísticas $T_1$, $T_2$, $T_3$ são dependentes entre si e, no caso de dependência, torna-se mais difícil concluir que a estatística $\frac{r_w}{\sqrt{var(r_w)}}$ segue uma distribuição $N(0,1)$.

Existe uma conjectura, baseada em simulações, que aponta nesse sentido. Em Costa & Soares (2002) encontra-se estimada a distribuição de $r_w$ com base em amostras aleatórias de um milhão de permutações. Após a comparação entre os valores estimados para os quantis mais importantes e os da distribuição $N(0,1)$ verifica-se que existem pequenas diferenças. As diferenças tendem a ser menores à medida que se aumenta à dimensão da amostra. No entanto, a demonstração teórica da normalidade assimptótica de $r_w$ não está ainda concluída (ver Costa & Soares,2002).

Fizemos algumas tentativas para estender o estudo do comportamento da variância assimptótica de $r_w$ para além da hipótese de independência dos dois vectores de ordens (ranks). Para isso, usamos o método desenvolvido por Kendall (1949) para determinar uma aproximação polinomial para $\sigma_a^2(r_s)$, no sentido de determinar uma aproximação para $\sigma_a^2(r_w)$ em amostras provenientes da distribuição normal bivariada. Contudo os

cálculos tornam-se extremamente fastidiosos, devido ao facto, de a variância de $r_w$ ser igual à variância da combinação linear de estatísticas do tipo $S$ (1.8). Refira-se que estas estatísticas são dependentes entre si, o que nos cria maiores dificuldades. O mesmo sucede quando se desenvolve o método de Borkowf (2002), no sentido de determinar uma aproximação para $\sigma_a^2(r_w)$, para outras distribuições bivariadas.

Na secção 2.3 propomos um método para estudar a distribuição de $r_w$ sob a hipótese de independência entre os dois vectores de ordens.

## 2.3 Método para determinar a distribuição exacta de $r_w$ sob $H_0$

Existem métodos para determinar a distribuição exacta de $r_w$ sob $H_0$ baseados na contagem de permutações. Estes métodos caracterizam-se por serem exaustivos, o que os torna pouco razoáveis no que confere ao tempo necessário para obtenção de resultados.

Seguindo o racíocinio descrito em Wiel et al. (2001) para estudar a distribuição de $r_s$, apresentamos um método que nos permite determinar a distribuição de $r_w$ sob $H_0$ numa amostra de dimensão $n$. Neste método, a função geradora de probabilidades da estatística em estudo representa-se através do permamente ('determinante sem sinal') de uma matriz cujas entradas são monómios.

O permanente de uma matriz quadrada $A_{n\times n}$ com entradas $a_{ij}$ define-se:

$$per(A) = \sum_{\sigma \in S_n} \prod_{i=1}^{n} a_{\sigma(i),i} \tag{2.9}$$

em que $S_n$ denota o grupo de permutações em {1,2,...,n}.

Tendo em consideração (2.4), para estudar a distribuição exacta de $r_w$ sob a hipótese

de independência ($H_0$) é necessário estudar a distribuição de:

$$D_{wn} = \sum_{i=1}^{n} (\tau(i) - \sigma(i))^2 (2n + 2 - \sigma(i) - \tau(i)) \tag{2.10}$$

com $\sigma, \tau \in S_n$. Sob a hipótese de independência dos vectores de ordens (rankings) a distribuição de $D_{wn}$ coincide com a distribuição de:

$$D^*_{wn} = \sum_{i=1}^{n} (i - \sigma(_i))^2 (2n + 2 - \sigma(i) - i) \tag{2.11}$$

Ao estudar a distribuição de $D^*_{wn}$ basta ter em conta que (M.A. Wiel et al., 2001):

$$\sum_{k=0}^{\infty} P(D^*_{wn} = k).a^k = \frac{1}{n!} per(A), \tag{2.12}$$

com

$$A_{ij} = a^{(i-j)^2.(2n+2-i-j)} \quad \forall a > 0 \tag{2.13}$$

Exemplificando, para $n = 4$ temos:

$$A = \begin{pmatrix} a^0 & a^7 & a^{24} & a^{45} \\ a^7 & a^0 & a^5 & a^{16} \\ a^{24} & a^5 & a^0 & a^3 \\ a^{45} & a^{16} & a^3 & a^0 \end{pmatrix} \tag{2.14}$$

Através da 'Package Algebra' do Maple obtemos:

$per(A) = 1 + a^6 + a^{10} + a^{14} + a^{20} + 2.a^{24} + a^{32} + 2.a^{36} + a^{48} + 2.a^{50} + 2.a^{60} + 2.a^{68} + 2.a^{72} + a^{80} + 3.a^{90} + a^{100}$.

Podemos então concluir que:

$P(D^*_{wn} = 0) = \frac{1}{4!}$, $P(D^*_{wn} = 1) = ... = P(D^*_{wn} = 5) = 0$,

$P(D^*_{wn} = 6) = \frac{1}{4!}$, $P(D^*_{wn} = 24) = \frac{2}{4!}$,..., $P(D^*_{wn} = 90) = \frac{3}{4!}$ e $P(D^*_{wn} = 100) = \frac{1}{4!}$

donde:

$P(r_w = 1) = P(D^*_{wn} = 0) = \frac{1}{4!}$, $P(r_w = 0.88) = P(D^*_{wn} = 6) = \frac{1}{4!}$, $P(r_w = 0.52) = P(D^*_{wn} = 24) = \frac{2}{4!}$,..., $P(r_w = -0.8) = P(D^*_{wn} = 90) = \frac{3}{4!}$ e $P(r_w = -1) = P(D^*_{wn} = 100) = \frac{1}{4!}$.

Para $n > 10$ não nos foi possível calcular o permanente de matrizes, com entradas monomiais, através da 'Package Algebra' do Maple. Em Wiel et al. (2001) é apresentado um algoritmo para cálculo de permanentes de matrizes de maiores dimensões. Devido às dificuldades enunciadas na secção anterior, o capítulo que se segue foi reservado para a estimação não paramétrica da variância de $r_w$.

# Capítulo 3

# Variância de $r_w$ e $r_s$

Neste capítulo abordamos alguns métodos de estimação não paramétrica e aplicamos esses métodos para estimar a variância de $r_w$. Este trabalho é análogo ao realizado em Borkowf (2000) para estimar a variância de $r_s$, e do qual apresentamos as principais conclusões extraídas na secção 3.5. Na secção 3.4 apresentamos os resultados obtidos nas simulações que efectuamos. Em ambos os estudos, os conjuntos de dados simulados são provenientes de diversas distribuições bivariadas que definimos na secção que se segue.

## 3.1 Algumas distribuições bivariadas de interesse

Para estudar o comportamento da variância de $r_w$ consideramos algumas distribuições bivariadas. São usadas amostras provenientes das distribuições: normal bivariada ($BVN$) de médias 0, variâncias 1, correlação $\rho$, com função densidade de probabilidade

$$f(x, y, \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{(x^2+y^2-2\rho xy)}{2(1-\rho^2)}}; \tag{3.1}$$

qui-quadrado bivariada usual ($BCS$); normal contaminada bivariada ($BCN$) com parâmetros $\rho, \tau, \varepsilon$; e "Three-Squares"($TS$).

Recordamos ao leitor que, se $(X, Y)$ segue a distribuição normal bivariada usual $(BVN(\rho))$ então $(X^2, Y^2)$ segue a distribuição qui-quadrado bivariada usual $(BCS)$ com correlação $\rho^2$ (Kotz et al., 2000), $(e^X, e^Y)$ segue a distribuição lognormal bivariada $(BLN)$ com correlação $\rho$ e $(|X|, |Y|)$ segue a distribuição semi-normal bivariada $(BHN)$ com correlação $\rho$. Se $(X, Y)$ segue a distribuição $BVN(\rho)$ usual e $W$ é tal que: $W = \tau$ ou $W = 1$, com $P(W = \tau) = \varepsilon$, então $(WX, WY)$ segue a distribuição $BCN(\rho, \tau, \varepsilon)$ (no estudo foram considerados $\rho = 0.75$ ; $\tau = 2, 4$ e $\varepsilon = 0.5$ ). A função densidade de probabilidade desta distribuição é:

$$g(x, y, \rho) = (1 - \varepsilon) f(x, y, \rho) + \frac{\varepsilon}{\tau^2} f(\frac{x}{\tau}, \frac{y}{\tau}, \rho); \tag{3.2}$$

em que $f(x, y, \rho)$ é a função densidade de probabilidade da distribuição $BVN(\rho)$ usual definida em (3.1). Note-se que esta distribuição corresponde a ter uma mistura de duas distribuições normais e permite-nos obter algumas conclusões acerca do comportamento de $r_w$ em distribuições de caudas pesadas.

Foi ainda considerada a distribuição $TS$ com densidade 3 em cada um dos três quadrados ($\left[0, \frac{1}{3}\right] \times \left[0, \frac{1}{3}\right]$ , $\left]\frac{1}{3}, \frac{2}{3}\right] \times \left]\frac{2}{3}, 1\right]$ e $\left]\frac{2}{3}, 1\right] \times \left]\frac{1}{3}, \frac{2}{3}\right]$). Nesta distribuição, as variáveis marginais $X$ e $Y$ seguem cada uma a distribuição uniforme em $[0, 1]$ e são dependentes com $\rho = \frac{4}{9}$. A distribuição $TS$ (Borkowf et al., 1997) é representativa de situações em que os dados podem ser agrupados (3 grupos) e observa-se frequentemente em estudos da Psicologia Experimental e em estudos epidemiológicos. Esta distribuição foi definida com o objectivo de demonstrar potenciais comportamentos estranhos de certas estatísticas calculadas a partir de dois vectores de ordens (ranks). Estes comportamentos estranhos reflectem-se no facto de as ordens (ranks) baixas, médias, e elevadas associadas às observações de $X$ corresponderem sempre a ordens baixas, elevadas, e médias associadas às observações de $Y$, respectivamente.

Em consequência da distribuição de $r_w$ ser invariante por transformações monótonas das distribuições originais marginais dos dados, o estudo do comportamento de $r_w$ em

distribuições como $BLN(\rho)$ e $BHN(\rho)$ reduz-se ao estudo do comportamento deste coeficiente em amostras provenientes das distribuições $BVN(\rho)$ e $BCS(\rho^2)$, respectivamente. Pelo facto de $\sigma^2(r_w) = \sigma^2(-r_w)$ nós estudamos o comportamento de $r_w$ nas distribuições $BVN$ e $BCS$ para correlações não negativas ($\rho \geq 0$).

A seguir é apresentado o método EBQP ("empirical bivariate quantile-partitioned") (Borkowf et al., 1997) adaptado para a estimação da variância de $r_w$ em amostras finitas.

## 3.2 Método EBQP para estimação da variância de $r_w$

Suponhamos que temos um conjunto de dados bivariados $(X_i, Y_i)$ $(i = 1, ..., n)$ independentes e identicamente distribuidos cuja função de distribuição é $F(x, y)$. Sejam $G(x)$, $H(y)$ funções de distribuição marginais; $G(x|y)$, $H(y|x)$ funções de distribuição condicionadas e $\hat{F}(x, y)$, $\hat{G}(x)$ e $\hat{H}(y)$ funções de distribuição empíricas. Note-se que:

$$\hat{F}(x, y) = \frac{1}{n} \sum_{i=1}^{n} I\{X_i \leq x, Y_i \leq y\}, \tag{3.3}$$

em que $I\{.\}$ é a função indicatriz.

Consideremos $R_i$ e $S_i$ ordens (ranks) marginais das observações $(X_i, Y_i)$. Sejam $\mu_1, ..., \mu_r$ e $\nu_1, ..., \nu_c$ ordens marginais distintas dos dados $(X_i, Y_i)$, respectivamente. Note-se que quando não há dados marginais com valores repetidos temos: $R_i = n\hat{G}(X_i)$; $S_i = n\hat{H}(Y_i)$; $\mu_k = k$; $\nu_l = l$; $r = c = n$. Para o caso em que existem dados marginais com valores repetidos podemos ter $r < n$, $c < n$ e, neste caso, $\mu_k$ e $\nu_l$ não são necessariamente inteiros.

Definem-se as proporcões marginais cumulativas como:

$$\gamma_k = \frac{1}{n} \sum_{i=1}^{n} I\{R_i \leq \mu_k\}$$

e

$$\eta_l = \frac{1}{n}\sum_{i=1}^{n} I\{S_i \leq \nu_l\}.$$

Denote-se: os quantis populacionais de $X$ e $Y$ correspondentes às proporções marginais cumulativas por $\xi_k = G^{-1}(\gamma_k)$ e $\psi_l = H^{-1}(\eta_l)$; as proporções cumulativas por $\phi_{kl} = F(\xi_k, \psi_l)$ e as proporções condicionais por $\gamma_{k|l} = G(\xi_k|\psi_l)$ e $\eta_{l|k} = H(\psi_l|\xi_k)$.

A seguir são definidas as células da tabela EBQP $\{\pi_{kl}\}$ $(k = 1, ..., r\ ; l = 1, ..., c)$ por:

$$\pi_{kl} = \phi_{kl} - \phi_{(k-1)l} - \phi_{k(l-1)} + \phi_{(k-1)(l-1)}.$$

Portanto, $\pi_{kl}$ designa a probabilidade:

$$P\{(\xi_{k-1} < X \leq \xi_k) \cap (\psi_{l-1} < Y \leq \psi_l)\}.$$

As proporções cumulativas empíricas $\left\{\hat{\phi}_{kl}\right\}$ e as células da tabela EBQP $\left\{\hat{\pi}_{kl}\right\}$ podem ser calculadas a partir das ordens das observações por:

$$\hat{\phi}_{kl} = \frac{1}{n}\sum_{i=1}^{n} I\{R_i \leq n\gamma_k\}\, I\{S_i \leq n\eta_l\} \tag{3.4}$$

e

$$\hat{\pi}_{kl} = \hat{\phi}_{kl} - \hat{\phi}_{(k-1)l} - \hat{\phi}_{k(l-1)} + \hat{\phi}_{(k-1)(l-1)}. \tag{3.5}$$

Refira-se que as células $\left\{\hat{\pi}_{kl}\right\}$formam uma tabela EBQP esparsa de dimensão $r \times c$. Podemos agora estimar as proporções condicionadas por (Borkowf, 2000):

$$\hat{\gamma}_{k|l} = \frac{1}{2}\left[\frac{1}{\hat{\pi}_{+l}}\sum_{i=1}^{k}\hat{\pi}_{il} + \frac{1}{\hat{\pi}_{+(l+1)}}\sum_{i=1}^{k}\hat{\pi}_{i(l+1)}\right]$$

e

$$\hat{\eta}_{l|k} = \frac{1}{2}\left[\frac{1}{\hat{\pi}_{k+}}\sum_{i=1}^{l}\hat{\pi}_{ki} + \frac{1}{\hat{\pi}_{(k+1)+}}\sum_{i=1}^{l}\hat{\pi}_{(k+1)i}\right]$$

com

$$\hat{\pi}_{+l} = \sum_{i=1}^{r}\hat{\pi}_{il}$$

e

$$\hat{\pi}_{k+} = \sum_{j=1}^{c}\hat{\pi}_{kj}.$$

Em Borkowf & al. (1997) encontra-se a prova de que:

$$E(\hat{\phi}_{kl}) \longrightarrow \phi_{kl}$$

e

$$Cov(\sqrt{n}\hat{\phi}_{kl}, \sqrt{n}\hat{\phi}_{ij}) \longrightarrow \lambda_{kl}^{T}\underset{klij}{\Omega}\lambda_{ij} \tag{3.6}$$

em que $n \longrightarrow \infty$, $\lambda_{kl}^{T} \equiv (1, -\eta_{l|k}, -\gamma_{k|l})$ , $\lambda_{ij}^{T} \equiv (1, -\eta_{j|i}, -\gamma_{i|j})$ e

$$\underset{klij}{\Omega} \equiv \begin{pmatrix} (\phi_{mn} - \phi_{kl}\phi_{ij}) & (\phi_{ml} - \phi_{kl}\gamma_i) & (\phi_{kn} - \phi_{kl}\eta_j) \\ (\phi_{mj} - \gamma_k\phi_{ij}) & (\gamma_m - \gamma_k\gamma_i) & (\phi_{kj} - \gamma_k\eta_j) \\ (\phi_{in} - \eta_l\phi_{ij}) & (\phi_{il} - \eta_l\gamma_i) & (\eta_n - \eta_l\eta_j) \end{pmatrix}, \tag{3.7}$$

com $m = \min\{i, k\}$ e $n = \min\{j, l\}$ . Todas estas definições e fórmulas são necessárias para as estimações que se seguem.

Estamos agora em condições de estimar $var(r_w)$ tendo em consideração que :

$$r_w = 1 - 6\frac{D^*}{n^3 + n^2 - n - 1}, \tag{3.8}$$

com

$$D^* = \frac{1}{n}\sum_{i=1}^{n}(R_i - S_i)^2(2n+2-R_i-S_i) \tag{3.9}$$
$$= \sum_{k=1}^{r}\sum_{l=1}^{c}(\mu_k - \nu_l)^2(2n+2-\mu_k-\nu_l)\hat{\pi}_{kl}$$

Esta última igualdade e as duas fórmulas que se seguem foram obtidas da mesma forma como foi obtida a fórmula para estimação da variância de $r_s$ pelo método EBQP (Borkowf, 2000), ainda que, com as necessárias adaptações para $r_w$.

O valor esperado e a variância de $r_w$ calculam-se por:

$$E(r_w) = 1 - \frac{6\sum_{k=1}^{r}\sum_{l=1}^{c}(\mu_k - \nu_l)^2(2n+2-\mu_k-\nu_l)\pi_{kl}}{n^3+n^2-n-1} \tag{3.10}$$

e

$$var(\sqrt{n}r_w) = 36\frac{var(\sqrt{n}D^*)}{(n^3+n^2-n-1)^2}$$
$$= 36\frac{\sum_{k=1}^{r}\sum_{l=1}^{c}\sum_{i=1}^{r}\sum_{j=1}^{c}(\mu_k-\nu_l)^2(2n+2-\mu_k-\nu_l)(\mu_i-\nu_j)^2(2n+2-\mu_i-\nu_j)cov(\sqrt{n}\hat{\pi}_{kl},\sqrt{n}\hat{\pi}_{ij})}{(n^3+n^2-n-1)^2} \tag{3.11}$$

As covariâncias da expressão (3.11) são estimadas usando (3.6) e a relação (3.5).

Denota-se a variância de $r_w$ fornecida pelo método EBQP por $\hat{\sigma}^2_E(r_w)$. De acordo com Borkowf (2000), a estimação da variância de $r_s$ fornecida pelo método EBQP pode ser melhorada com uma correcção empírica. Segundo o mesmo autor, a correcção deve-se ao relacionamento entre a distribuição assimptótica normal para as tabelas EBQP e a

distribuição assimptótica hipergeométrica multivariada (MXH) (Borkowf et al., 1997) , de onde é obtida a correcção exacta baseada no caso de independência dos dados (Borkowf , 2000).

Para a estimação da variância de $r_w$ nós efectuamos a mesma correcção. Designamos por EBQP* o método EBQP com a correcção empírica e por $\hat{\sigma}^2_{E^*}(\sqrt{n}r_w)$ a variância estimada pelo método EBQP*, que é dada por $\hat{\sigma}^2_{E^*}(\sqrt{n}r_w) = \hat{\sigma}^2_E(\sqrt{n}r_w)\frac{n}{n-1}$.

## 3.3 Algoritmos de bootstrap e jackknife para estimação da variância de $r_w$

Vamos considerar dois outros métodos para estimação da variância de $r_w$ em amostras finitas: os algoritmos de bootstrap e jackknife. O algoritmo de bootstrap envolve processos de reamostragens com substituição do conjunto de dados $\{(X_i, Y_i)\}$ por um novo conjunto $\{(X_i^*, Y_i^*)\}$.

O processo repete-se $B$ vezes ($B = 250$ é um valor razoável). Em cada repetição é gerado um conjunto $\left\{\left(X^*_{i(b)}, Y^*_{i(b)}\right)\right\}$, onde é calculado $r^*_w(b)$. A seguir calcula-se a média e a variância de $r_w$ através das fórmulas (Efron & Tibshirani , 1993):

$$\bar{r}^*_w = \frac{1}{B}\sum_{b=1}^{B} r^*_w(b)$$

e

$$\hat{\sigma}^2_B(r_w) = \frac{1}{B-1}\sum_{b=1}^{B}\left[r^*_w(b) - \bar{r}^*_w\right]^2. \tag{3.12}$$

$\hat{\sigma}^2_B(r_w)$ é usado para estimar a variância de $r_w$.

O algoritmo de jackknife envolve a eliminação sucessiva de observações. Em cada repetição é criado um novo conjunto de dados com a $i$-ésima observação eliminada e calcula-se $r^{\#}_w(i)$ $(i = 1, ...n)$. A seguir determina-se a média e a variância de jacknife por

(Efron &Tibshirani, 1993):

$$\bar{r}_w^{\#} = \frac{1}{n}\sum_{i=1}^{n} r_w^{\#}(i)$$

e

$$\hat{\sigma}_J^2(r_w) = \frac{n-1}{n}\sum_{i=1}^{n}\left[r_w^{\#}(i) - \bar{r}_w^{\#}\right]^2. \tag{3.13}$$

Assim, $\hat{\sigma}_J^2(r_w)$ pode ser usado para estimar a variância de $r_w$. Na secção seguinte são apresentados os valores da variância de $\sqrt{n}r_w$, fornecidos pelos algoritmos de bootstrap e jackknife e pelos métodos EBQP e EBQP*.

## 3.4 Resultados da estimação da variância de $r_w$ obtidos em simulações

Nas simulações que efectuamos foi usada a linguagem de programação do software Gauss 5.0 (Aptech Systems). Estimamos a variância de $\sqrt{n}r_w$ em amostras finitas ($n = 5, 8, 10, 12, 15, 20$ e 25) provenientes das distribuições bivariadas apresentadas na secção 3.1 e correlação $\rho \simeq 0,\ 0.25,\ 0.5,\ 0.75,\ 0.9$. As variâncias de $\sqrt{n}r_w$ foram estimadas pelos métodos EBQP, EBQP*, bootstrap ( com $B = 250$) e jackknife.

Para cada distribuição bivariada combinada com cada uma das dimensões da amostra foram simulados 40 000 conjuntos de dados (10 000 para $n = 20, 25$), a que se seguiu a estimação pontual de $\rho_w$ e as estimações da variância de $\sqrt{n}r_w$ fornecidas por $\hat{\sigma}_E^2(\sqrt{n}r_w)$, $\hat{\sigma}_{E^*}^2(\sqrt{n}r_w)$, $\hat{\sigma}_J^2(\sqrt{n}r_w)$, e $\hat{\sigma}_B^2(\sqrt{n}r_w)$ com $B = 250$. Para a distribuição normal bivariada, no caso de independência dos dados, também calculamos o valor da variância correcta, $\sigma^2(\sqrt{n}r_w)$. As variâncias estimadas foram aproximadas a duas casas decimais. Na tabela 3 estão registados os resultados obtidos.

Tabela 3

Variâncias de $\sqrt{n}r_w$ estimadas em amostras provenientes de algumas distribuições bivariadas[1]

| | | Distribuição($\rho$) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | BVN | BVN | BVN | BVN | BVN | BCS | BCS | BCS | BCS | BCN | BCN | TS |
| | | 0.00 | 0.25 | 0.50 | 0.75 | 0.90 | 0.25 | 0.50 | 0.75 | 0.90 | (a) | (b) | 0.44 |
| n | Método | | | | | | | | | | | | |
| 5 | EBQP | 1.02 | 0.98 | 0.85 | 0.59 | 0.34 | 1.00 | 0.94 | 0.74 | 0.49 | 0.60 | 0.63 | 0.89 |
| | EBQP* | 1.27 | 1.22 | 1.06 | 0.74 | 0.42 | 1.26 | 1.17 | 0.93 | 0.61 | 0.75 | 0.79 | 1.11 |
| | BOOT | 1.08 | 1.03 | 0.90 | 0.66 | 0.42 | 1.05 | 0.97 | 0.79 | 0.55 | 0.66 | 0.68 | 0.97 |
| | JACK | 2.10 | 2.01 | 1.73 | 1.17 | 0.64 | 2.04 | 1.87 | 1.45 | 0.91 | 1.18 | 1.26 | 1.97 |
| | $\sigma^2(\sqrt{n}r_w)$ | 1.27 | | | | | | | | | | | |
| 8 | EBQP | 1.07 | 1.02 | 0.84 | 0.51 | 0.24 | 1.06 | 0.96 | 0.71 | 0.41 | 0.55 | 0.61 | 0.90 |
| | EBQP* | 1.23 | 1.16 | 0.96 | 0.58 | 0.28 | 1.21 | 1.10 | 0.81 | 0.46 | 0.63 | 0.69 | 1.03 |
| | BOOT | 1.12 | 1.06 | 0.90 | 0.60 | 0.33 | 1.09 | 1.00 | 0.76 | 0.48 | 0.63 | 0.67 | 1.01 |
| | JACK | 1.63 | 1.54 | 1.25 | 0.74 | 0.33 | 1.58 | 1.41 | 1.01 | 0.56 | 0.80 | 0.89 | 1.46 |
| | $\sigma^2(\sqrt{n}r_w)$ | 1.17 | | | | | | | | | | | |
| 10 | EBQP | 1.08 | 1.02 | 0.81 | 0.47 | 0.20 | 1.06 | 0.95 | 0.68 | 0.37 | 0.51 | 0.59 | 0.88 |
| | EBQP* | 1.20 | 1.13 | 0.90 | 0.52 | 0.23 | 1.17 | 1.06 | 0.75 | 0.41 | 0.57 | 0.66 | 0.98 |
| | BOOT | 1.12 | 1.07 | 0.88 | 0.56 | 0.28 | 1.09 | 0.99 | 0.74 | 0.44 | 0.59 | 0.65 | 1.00 |
| | JACK | 1.50 | 1.41 | 1.11 | 0.63 | 0.26 | 1.45 | 1.28 | 0.90 | 0.47 | 0.68 | 0.80 | 1.29 |
| | $\sigma^2(\sqrt{n}r_w)$ | 1.14 | | | | | | | | | | | |
| 12 | EBQP | 1.08 | 1.01 | 0.80 | 0.45 | 0.18 | 1.06 | 0.95 | 0.66 | 0.34 | 0.49 | 0.58 | 0.85 |
| | EBQP* | 1.18 | 1.11 | 0.87 | 0.49 | 0.20 | 1.15 | 1.03 | 0.73 | 0.37 | 0.54 | 0.64 | 0.93 |
| | BOOT | 1.12 | 1.06 | 0.87 | 0.53 | 0.25 | 1.09 | 0.99 | 0.72 | 0.40 | 0.56 | 0.64 | 0.97 |
| | JACK | 1.41 | 1.32 | 1.04 | 0.56 | 0.22 | 1.37 | 1.21 | 0.83 | 0.41 | 0.62 | 0.74 | 1.17 |
| | $\sigma^2(\sqrt{n}r_w)$ | 1.12 | | | | | | | | | | | |

Tabela 3: (continuação)

| | | Distribuição($\rho$) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | BVN | BVN | BVN | BVN | BVN | BCS | BCS | BCS | BCS | BCN | BCN | TS |
| | | 0.00 | 0.25 | 0.50 | 0.75 | 0.90 | 0.25 | 0.50 | 0.75 | 0.90 | (a) | (b) | 0.44 |
| n | Método | | | | | | | | | | | | |
| 15 | EBQP | 1.08 | 1.01 | 0.78 | 0.42 | 0.16 | 1.05 | 0.94 | 0.65 | 0.31 | 0.46 | 0.56 | 0.82 |
| | EBQP* | 1.15 | 1.08 | 0.84 | 0.45 | 0.17 | 1.13 | 1.01 | 0.69 | 0.34 | 0.50 | 0.60 | 0.88 |
| | BOOT | 1.12 | 1.05 | 0.84 | 0.49 | 0.21 | 1.09 | 0.98 | 0.70 | 0.37 | 0.53 | 0.61 | 0.94 |
| | JACK | 1.33 | 1.25 | 0.96 | 0.50 | 0.18 | 1.29 | 1.14 | 0.77 | 0.36 | 0.56 | 0.68 | 1.06 |
| | $\sigma^2(\sqrt{n}r_w)$ | 1.10 | | | | | | | | | | | |
| 20 | EBQP | 1.07 | 1.00 | 0.76 | 0.38 | 0.13 | 1.05 | 0.92 | 0.62 | 0.29 | 0.44 | 0.55 | 0.77 |
| | EBQP* | 1.13 | 1.05 | 0.80 | 0.40 | 0.14 | 1.10 | 0.97 | 0.65 | 0.30 | 0.46 | 0.58 | 0.81 |
| | BOOT | 1.10 | 1.03 | 0.81 | 0.44 | 0.17 | 1.08 | 0.95 | 0.66 | 0.33 | 0.49 | 0.59 | 0.88 |
| | JACK | 1.25 | 1.16 | 0.88 | 0.44 | 0.14 | 1.22 | 1.06 | 0.71 | 0.32 | 0.50 | 0.63 | 0.94 |
| | $\sigma^2(\sqrt{n}r_w)$ | 1.08 | | | | | | | | | | | |
| 25 | EBQP | 1.06 | 0.98 | 0.74 | 0.36 | 0.12 | 1.04 | 0.92 | 0.61 | 0.27 | 0.42 | 0.54 | 0.74 |
| | EBQP* | 1.11 | 1.02 | 0.77 | 0.38 | 0.12 | 1.08 | 0.96 | 0.63 | 0.28 | 0.43 | 0.56 | 0.77 |
| | BOOT | 1.09 | 1.02 | 0.78 | 0.41 | 0.15 | 1.07 | 0.95 | 0.64 | 0.31 | 0.46 | 0.57 | 0.84 |
| | JACK | 1.20 | 1.11 | 0.83 | 0.40 | 0.13 | 1.17 | 1.03 | 0.67 | 0.29 | 0.46 | 0.60 | 0.86 |
| | $\sigma^2(\sqrt{n}r_w)$ | 1.07 | | | | | | | | | | | |

**Observação 1** : *Distribuições: normal bivariada (BVN), qui-quadrado bivariada (BCS), normal contaminada bivariada (BCN), "three squares"(TS), (a) BCN (0.75,2,0.5), (b) BCN (0.75,4,0.5). Métodos: EBQP = variância EBQP, EBQP* = variância EBQP com correcção empírica, BOOT = variância bootstrap (B=250), JACK = variância jackknife. Os valores da tabela são as médias das variâncias calculadas para 40 000 conjuntos de dados (10 000 para n=20,25).* $\sigma^2(\sqrt{n}r_w)$ *designa a variância correcta.*

Os dados registados na tabela 3 mostram que a variância estimada pelo método EBQP, $\hat{\sigma}^2_E(\sqrt{n}r_w)$, é sempre inferior às variâncias estimadas pelos outros métodos.

Para a distribuição normal bivariada usual ($BVN(\rho)$), se $\rho \lesssim 0.5$ observa-se que a variância estimada pelo método EBQP*, $\hat{\sigma}^2_{E^*}(\sqrt{n}r_w)$, é quase sempre maior do que a variância estimada pelo algoritmo de bootstrap, $\hat{\sigma}^2_B(\sqrt{n}r_w)$. No entanto, se $\rho \gtrsim 0.5$ e $n > 10$, $\hat{\sigma}^2_{E^*}(\sqrt{n}r_w)$ é quase sempre menor do que $\hat{\sigma}^2_B(\sqrt{n}r_w)$.

Para a distribuição qui-quadrado bivariada usual ($BCS(\rho^2)$), se $\rho \lesssim 0.75$ e $n \geq 8$ verifica-se que $\hat{\sigma}^2_{E^*}(\sqrt{n}r_w)$ é quase sempre maior do que $\hat{\sigma}^2_B(\sqrt{n}r_w)$. Contudo, se $\rho \gtrsim 0.75$ e $n \gtrsim 10$ então $\hat{\sigma}^2_{E^*}(\sqrt{n}r_w)$ tende a ser menor do que $\hat{\sigma}^2_B(\sqrt{n}r_w)$.

Em amostras provenientes das distribuições bivariada contaminada normal ($BCN$ com $\tau = 2,4$ e $\varepsilon = 0.5$) e "three squares"($TS$) a variância estimada pelo algoritmo de bootstrap é maior do que $\hat{\sigma}^2_{E^*}(\sqrt{n}r_w)$ para $n \geq 10$.

De um modo geral, a variância estimada pelo algoritmo de jackknife $\sigma^2_J(\sqrt{n}r_w)$ é maior do que $\hat{\sigma}^2_{E^*}(\sqrt{n}r_w)$ e $\hat{\sigma}^2_B(\sqrt{n}r_w)$, com excepção de amostras com $n \gtrsim 8$ e correlações altas ($\rho \simeq 0.9$), nas quais, $\hat{\sigma}^2_B(\sqrt{n}r_w)$ é ligeiramente maior do que $\hat{\sigma}^2_J(\sqrt{n}r_w)$. Também verifica-se que, em amostras de pequenas dimensões ($n < 8$), o algoritmo de jackknife fornece valores para $\sigma^2(\sqrt{n}r_w)$ bastante superiores aos fornecidos pelos outros métodos.

De referir, que na distribuição $BVN$, para o caso de independência dos dados ($\rho = 0$), se $n \lesssim 12$ então o método EBQP* produz melhores resultados que os outros métodos. Para $12 \lesssim n \lesssim 25$ constata-se que as variâncias estimadas pelo algoritmo de bootstrap e

pelo método EBQP* são a que mais se aproximam de $\sigma^2(\sqrt{n}r_w)$ por valores superiores. Neste caso, podemos afirmar que o método EBQP* foi adaptado com sucesso para a estimação de $r_w$. Para as outras situações não nos foi possível comparar as variâncias estimadas pelos diferentes métodos com a variância correcta, $\sigma^2(\sqrt{n}r_w)$, por esta não ser conhecida.

A seguir enunciamos algumas das conclusões de um trabalho idêntico para estudo da variância do estimador do coeficiente de correlação de Spearman $r_s$.

## 3.5 Análise dos resultados obtidos na estimação da variância de $r_s$

Em Borkowf (2000) é possível encontrar um estudo do comportamento da variância do coeficiente de correlação de Spearman $\rho_s$, em amostras finitas. O estudo foi efectuado em amostras de dimensão $n = 5, 8, 10, 12, 15, 20, 25$, e correlação $\rho \simeq 0,\ 0.25,\ 0.5,\ 0.75,\ 0.9$. No estudo foram usadas amostras provenientes das distribuições apresentadas na secção 3.1.

As variâncias de $r_s$ foram estimadas pelos métodos EBQP, EBQP*, bootstrap (com $B = 250$) e jackknife nas mesmas condições em que foram estimadas as variâncias de $r_w$ na secção anterior. Este estudo é mais completo do que o apresentado para a variância de $r_w$ pois apresenta mais alguns resultados. Entre esses resultados destaca-se a média das variâncias simuladas baseadas em 10 repetições de 1 000 000 de conjuntos de dados simulados.

Observou-se que o método EBQP* estima quase sempre com maior precisão $\sigma^2(\sqrt{n}r_s)$

em amostras finitas do que os algoritmos de bootstrap e jackknife. Os valores fornecidos pelo método EBQP* são aproximadamente iguais ou ligeiramente superiores aos valores correctos $\sigma^2(\sqrt{n}r_s)$. Em amostras de pequenas dimensões, as variâncias estimadas a partir do método EBQP são inferiores às variâncias correctas $\sigma^2(\sqrt{n}r_s)$. Contudo, para correlações altas ou $n$ grande pode acontecer que o método EBQP estime melhor $\sigma^2(\sqrt{n}r_s)$ que o método EBQP*.

A estimação da variância obtida pelo algoritmo de bootstrap aproxima-se da variância correcta para $\rho$ pequeno e $n$ grande, mas fornece resultados bastante superiores à variância correcta quando $\rho$ é grande e $n$ pequeno. Isso é consequência do facto dos processos de reamostragens originarem muitas repetições nas ordens associadas aos dados marginais, o que faz diminuir o viés da estimação, aumentando assim, a variância estimada.

A variância de $r_s$ estimada pelo algoritmo de jackknife, tal como sucede ao estimar a variância de $r_w$, é quase sempre superior às variâncias estimadas pelo método EBQP* e pelo algoritmo de bootstrap (excepto em amostras com correlações altas).

Em suma, na generalidade das amostras desse estudo, verificou-se que o método EBQP* fornece melhores estimativas para a variância de $r_s$ do que os algoritmos de bootstrap e jackknife.

Não nos é possível afirmar que o método EBQP* fornece melhores estimativas para a variância de $r_w$. No entanto é de salientar o facto de, à semelhança do que acontece para a variância de $r_s$, os valores estimados pelo método EBQP* serem quase sempre inferiores aos valores estimados pelo algoritmo de jackknife e ligeiramente inferiores ou, por vezes, ligeiramente superiores aos valores estimados pelo algoritmo de bootstrap.

No capítulo 4 abordamos alguns métodos de construção de intervalos de confiança em que usamos os valores da variância de $r_w$ estimados neste capítulo.

# Capítulo 4

# Intervalos de confiança para $\rho_w$ e $\rho_s$

Numa primeira fase deste capítulo descrevemos alguns dos métodos que usamos para construir intervalos de confiança. Posteriormente, apresentamos intervalos de confiança a 90% para $\rho_w$. A finalizar o capítulo, analisamos os resultados obtidos num estudo similar para o coeficiente $\rho_s$ realizado em Borkowf (2000). Em particular, são analisadas as proporções do coeficiente de correlação de Spearman $\rho_s$ estar compreendido entre os limites inferior e superior de cada um dos intervalos de confiança. Estas proporções vamos designá-las por porporções de cobertura dos intervalos de confiança e podem ser interpretados como um indicador dos melhores métodos para construir intervalos de confiança para $\rho_s$. Refira-se que não nos foi possível realizar um estudo idêntico para os intervalos construídos para $\rho_w$ por desconhecimento dos valores esperados para $r_w$ nas distribuições estudadas.

## 4.1 Métodos para construção de intervalos de confiança para $\rho_w$

Neste capítulo são construídos intervalos de confiança a 90% para o coeficiente $\rho_w$ nas mesmas distribuições bivariadas para as quais estimamos a variância de $r_w$ na tabela 3. Um dos intervalos de confiança construídos é o assimptoticamente normal, simétrico, a

$(1-\alpha)100\%$ que é da forma $r_w \mp \Phi^{-1}(1-\frac{\alpha}{2})\hat{\sigma}(r_w)$, em que $\Phi$ é a função de distribuição de $N(0,1)$. Refira-se que ao construir estes intervalos estamos a assumir que $r_w$ segue uma distribuição que se aproxima da normal e embora haja indicações nesse sentido, a prova formal está por concluir (secção 2.2).

Na construção dos intervalos assimptoticamente normais usamos os valores de $\hat{\sigma}^2(r_w)$ fornecidos pelo método EBQP*e pelos algoritmos de bootstrap e jackknife, abordados no capítulo anterior.

Em amostras de pequena dimensão e valores de $r_w$ próximos de $\mp 1$,os intervalos de confiança podem ter limites inferior ou superior não pertencentes ao intervalo $[-1,1]$. Este problema pode ser resolvido usando as transformações de Fisher. Ao aplicar essas transformações considera-se:

$$\varepsilon_w = \operatorname{arctanh}(\rho_w)$$

e

$$\hat{\varepsilon_w} = \operatorname{arctanh}(r_w).$$

O intervalo de confiança é da forma $\hat{\varepsilon_w} \pm \sqrt{\frac{31}{30}}\Phi^{-1}(1-\frac{\alpha}{2})\hat{\sigma}(\hat{\varepsilon_w})$. Contudo, não foi possível aplicar este método devido ao desconhecimento do valor $\hat{\sigma}(\hat{\varepsilon_w})$. Sempre que nos surgiram limites inferiores ou superiores não pertencentes ao intervalo $[-1,1]$ adoptamos o procedimento usual para truncatura, ou seja, truncamos em 1 ou -1, consoante o caso, e alteramos o outro limite.

Em outro dos métodos aplicado na construção dos intervalos de confiança a $(1-\alpha)100\%$ é usado o conjunto das estimativas de bootstrap $\{r_w^*(b)\}$ $b = 1, ..., B$. Ordenam-se as estimativas $\{r_w^*(b)\}$, $b = 1, ..., B$ e depois cálculam-se os percentis de bootstrap correspondentes a $\frac{\alpha}{2}$ e $1-\frac{\alpha}{2}$ (é usada interpolação linear), que são os limites inferior e superior do intervalo, respectivamente. O intervalo é então: $(r_w^{*(\frac{\alpha}{2})}, r_w^{*(1-\frac{\alpha}{2})})$ em que $r_w^{*(\alpha)}$ designa o percentil de ordem $100.\alpha$ das réplicas de bootstrap $r_w^*(1), ..., r_w^*(B)$. Por exemplo, se $B = 2000$ e $\alpha = 0.1$, então o intervalo é $(r_w^{*(0.05)}, r_w^{*(0.95)})$, o qual contém os valores compreendidos entre o centésimo e o 1900-ésimo valores ordenados do conjunto

$\{r_w^*(b)\}$, $b = 1, ..., 2000$.

Neste estudo são construídos intervalos de confiança através dos percentis de bootstrap com viés corrigido $BC_a$ (Efron & Tibshirani, 1993). O intervalo é $(r_w^{*(\alpha_1)}, r_w^{*(\alpha_2)})$, com:

$$\begin{aligned} \alpha_1 &= \Phi\left\{ z_0 + \frac{\left[z_0 + \Phi^{-1}(\frac{\alpha}{2})\right]}{\left[1 - a(z_0 + \Phi^{-1}(\frac{\alpha}{2}))\right]} \right\} \\ \alpha_2 &= \Phi\left\{ z_0 + \frac{\left[z_0 + \Phi^{-1}(1 - \frac{\alpha}{2})\right]}{\left[1 - a(z_0 + \Phi^{-1}(1 - \frac{\alpha}{2}))\right]} \right\}, \end{aligned} \tag{4.1}$$

em que o valor para a correcção do viés $\hat{z}_0$ é obtido directamente a partir da proporção das réplicas de bootstrap menores do que o valor de $r_w$ estimado inicialmente,

$$\hat{z}_0 = \Phi^{-1}\left\{ \frac{1}{B} \sum_{b=1}^{B} \left[ I\left\{r_w^*(b) < r_w\right\} + \frac{1}{2} I\left\{r_w^*(b) = r_w\right\} \right] \right\}, \tag{4.2}$$

e $\hat{a}$, que representa a razão de variação do desvio-padrão de $r_w$ relativamente a $\rho_w$, pode ser estimado por:

$$\hat{a} = -\frac{\left[\sum_{k=1}^{n} (r_w^{\#}(k) - \bar{r}_w^{\#})^3\right]}{6\left[\sum_{k=1}^{n} (r_w^{\#}(k) - \bar{r}_w^{\#})^2\right]^{\frac{3}{2}}}. \tag{4.3}$$

Note-se que para estimar $\hat{z}_0$ recorre-se às amostras de bootstrap e para estimação de $\hat{a}$ recorre-se às amostras de jackknife.

## 4.2 Intervalos de confiança para $\rho_w$

Nesta secção apresentamos os intervalos de confiança a 90% para $\rho_w$ que se encontram na tabela 4. Estes intervalos foram construídos a partir dos resultados otidos nas simulações para estimar a variância de $r_w$ (capítulo 3). Os extremos dos intervalos foram aproximados a duas casas decimais.

Tabela 4

Intervalos de confiança a 90% para $\rho_w$ construídos por diversos métodos[2]

| | | Distribuição($\rho$) | | | | |
|---|---|---|---|---|---|---|
| | | BVN | BVN | BVN | BVN | BVN |
| | | 0.00 | 0.25 | 0.50 | 0.75 | 0.90 |
| n | Método | | | | | |
| 5 | EBQP*-N | [-0.77,0.77] | [-0.55,0.95] | [-0.33,1.00] | [-0.20,1.00] | [0.34,1.00] |
| | BOOT-N | [-0.75,0.74] | [-0.52,0.92] | [-0.29,1.00] | [-0.20,1.00] | [0.29,1.00] |
| | JACK-N | [-0.99,0.98] | [-0.87,1.00] | [-0.62,1.00] | [-0.17,1.00] | [0.24,1.00] |
| | BOOT-P | [-0.57,0.86] | [-0.44,0.93] | [-0.27,0.97] | [0.00,0.99] | [0.30,1.00] |
| | BOOT_B | [-0.64,0.62] | [-0.55,0.77] | [-0.43,0.88] | [-0.24,0.96] | [0.01,0.99] |
| 8 | EBQP*-N | [-0.62,0.63] | [-0.39,0.82] | [-0.10,0.97] | [0.24,1.00] | [0.52,1.00] |
| | BOOT-N | [-0.60,0.61] | [-0.37,0.80] | [-0.10,0.97] | [0.20,1.00] | [0.45,1.00] |
| | JACK-N | [-0.71,0.72] | [-0.48,0.91] | [-0.19,1.00] | [0.16,1.00] | [0.49,1.00] |
| | BOOT-P | [-0.53,0.67] | [-0.37,0.80] | [-0.16,0.89] | [0.16,0.97] | [0.45,0.99] |
| | BOOT_B | [-0.59,0.58] | [-0.44,0.73] | [-0.24,0.85] | [0.07,0.95] | [0.34,0.99] |
| 10 | EBQP*-N | [-0.55,0.56] | [-0.32,0.76] | [-0.02,0.91] | [0.31,1.00] | [0.60,1.00] |
| | BOOT-N | [-0.54,0.55] | [-0.31,0.75] | [-0.03,0.92] | [0.29,1.00] | [0.53,1.00] |
| | JACK-N | [-0.62,0.62] | [-0.38,0.82] | [-0.07,0.96] | [0.28,1.00] | [0.58,1.00] |
| | BOOT-P | [-0.50,0.59] | [-0.32,0.73] | [-0.09,0.85] | [0.23,0.94] | [0.52,0.98] |
| | BOOT_B | [-0.54,0.53] | [-0.37,0.69] | [-0.14,0.84] | [0.19,0.93] | [0.47,0.98] |
| 12 | EBQP*-N | [-0.51,0.51] | [-0.27,0.71] | [0.02,0.87] | [0.39,0.99] | [0.65,1.00] |
| | BOOT-N | [-0.50,0.50] | [-0.26,0.70] | [0.02,0.88] | [0.35,1.00] | [0.58,1.00] |
| | JACK-N | [-0.56,0.55] | [-0.31,0.75] | [-0.01,0.91] | [0.36,1.00] | [0.63,1.00] |
| | BOOT-P | [-0.47,0.53] | [-0.28,0.68] | [-0.04,0.81] | [0.29,0.92] | [0.57,0.98] |
| | BOOT_B | [-0.50,0.49] | [-0.31,0.66] | [-0.07,0.79] | [0.26,0.91] | [0.55,0.97] |

Tabela 4 (continuação)

| | | Distribuição($\rho$) | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | BCS | BCS | BCS | BCS | BCN | BCN | TS |
| | | 0.25 | 0.50 | 0.75 | 0.90 | (a) | (b) | 0.44 |
| n | Método | | | | | | | |
| 5 | EBQP*-N | [-0.64,0.90] | [-0.45,1.00] | [-0.20,1.00] | [0.14,1.00] | [0.01,1.00] | [-0.01,1.00] | [-0.32,1.00] |
| | BOOT-N | [-0.60,0.86] | [-0.40,0.98] | [-0.16,1.00] | [0.13,1.00] | [0.01,1.00] | [-0.01,1.00] | [-0.34,1.00] |
| | JACK-N | [-0.92,1.00] | [-0.77,1.00] | [-0.40,1.00] | [0.01,1.00] | [-0.17,1.00] | [-0.20,1.00] | [-0.65,1.00] |
| | BOOT-P | [-0.48,0.90] | [-0.36,0.95] | [-0.14,0.98] | [0.14,1.00] | [0.01,0.99] | [0.02,0.99] | [-0.39,0.92] |
| | BOOT_B | [-0.58,0.72] | [-0.50,0.83] | [-0.34,0.93] | [0.12,0.98] | [-0.22,0.96] | [-0.21,0.96] | [-0.49,0.80] |
| 8 | EBQP*-N | [-0.48,0.75] | [-0.28,0.88] | [0.04,1.00] | [0.34,1.00] | [0.21,1.00] | [0.19,1.00] | [-0.07,0.98] |
| | BOOT-N | [-0.46,0.73] | [-0.26,0.87] | [0.04,1.00] | [0.29,1.00] | [0.18,1.00] | [0.16,1.00] | [-0.11,1.00] |
| | JACK-N | [-0.57,0.84] | [-0.36,0.96] | [-0.04,1.00] | [0.29,1.00] | [0.13,1.00] | [0.11,1.00] | [-0.20,1.00] |
| | BOOT-P | [-0.43,0.76] | [-0.28,0.84] | [-0.01,0.93] | [0.28,0.98] | [0.14,0.96] | [0.13,0.97] | [-0.24,0.85] |
| | BOOT_B | [-0.50,0.68] | [-0.36,0.78] | [-0.11,0.90] | [0.17,0.96] | [0.04,0.94] | [0.03,0.95] | [-0.31,0.79] |
| 10 | EBQP*-N | [-0.41,0.69] | [-0.21,0.83] | [0.13,0.97] | [0.42,1.00] | [0.30,1.00] | [0.26,1.00] | [0.02,0.94] |
| | BOOT-N | [-0.40,0.67] | [-0.20,0.82] | [0.12,0.97] | [0.39,1.00] | [0.27,1.00] | [0.24,1.00] | [-0.01,0.98] |
| | JACK-N | [-0.47,0.75] | [-0.26,0.88] | [0.09,1.00] | [0.40,1.00] | [0.25,1.00] | [0.21,1.00] | [-0.04,1.00] |
| | BOOT-P | [-0.39,0.68] | [-0.22,0.79] | [0.06,0.90] | [0.35,0.96] | [0.22,0.94] | [0.20,0.95] | [-0.14,0.81] |
| | BOOT_B | [-0.44,0.63] | [-0.28,0.75] | [-0.01,0.87] | [0.30,0.95] | [0.17,0.93]] | [0.13,0.93] | [-0.19,0.79] |
| 12 | EBQP*-N | [-0.36,0.64] | [-0.16,0.78] | [0.17,0.93] | [0.49,1.00] | [0.37,1.00] | [0.32,1.00] | [0.09,0.91] |
| | BOOT-N | [-0.35,0.63] | [-0.15,0.77] | [0.16,0.94] | [0.45,1.00] | [0.34,1.00] | [0.30,1.00] | [0.05,0.94] |
| | JACK-N | [-0.41,0.68] | [-0.20,0.82] | [0.14,0.96] | [0.48,1.00] | [0.33,1.00] | [0.28,1.00] | [0.04,0.95] |
| | BOOT-P | [-0.35,0.63] | [-0.18,0.74] | [0.10,0.87] | [0.41,0.95] | [0.27,0.92] | [0.24,0.93] | [-0.07,0.79] |
| | BOOT_B | [-0.39,0.60] | [-0.22,0.71] | [0.06,0.85] | [0.38,0.94] | [0.24,0.91] | [0.20,0.92] | [-0.10,0.78] |

Tabela 4 (continuação)

| | | Distribuição($\rho$) | | | | |
|---|---|---|---|---|---|---|
| | | BVN | BVN | BVN | BVN | BVN |
| | | 0.00 | 0.25 | 0.50 | 0.75 | 0.90 |
| n | Método | | | | | |
| 15 | EBQP*-N | [-0.45,0.45] | [-0.30,0.74] | [0.08,0.83] | [0.43,0.96] | [0.69,1.00] |
| | BOOT-N | [-0.45,0.44] | [-0.29,0.74] | [0.07,0.83] | [0.41,0.98] | [0.65,1.00] |
| | JACK-N | [-0.49,0.48] | [-0.33,0.78] | [0.05,0.85] | [0.42,0.97] | [0.69,1.00] |
| | BOOT-P | [-0.43,0.46] | [-0.31,0.70] | [0.02,0.78] | [0.35,0.90] | [0.62,0.96] |
| | BOOT_B | [-0.45,0.44] | [-0.33,0.68] | [0.00,0.77] | [0.34,0.90] | [0.63,0.97] |
| 20 | EBQP*-N | [-0.39,0.39] | [-0.14,0.60] | [0.14,0.78] | [0.49,0.92] | [0.74,0.99] |
| | BOOT-N | [-0.38,0.38] | [-0.14,0.60] | [0.14,0.79] | [0.47,0.94] | [0.71,1.00] |
| | JACK-N | [-0.41,0.41] | [-0.16,0.62] | [0.13,0.80] | [0.48,0.93] | [0.74,0.99] |
| | BOOT-P | [-0.37,0.39] | [-0.16,0.58] | [0.10,0.74] | [0.42,0.88] | [0.68,0.95] |
| | BOOT_B | [-0.39,0.38] | [-0.17,0.57] | [0.09,0.73] | [0.42,0.88] | [0.69,0.96] |
| 25 | EBQP*-N | [-0.34,0.34] | [-0.10,0.56] | [0.18,0.75] | [0.52,0.90] | [0.77,0.97] |
| | BOOT-N | [-0.34,0.34] | [-0.10,0.56] | [0.18,0.75] | [0.51,0.91] | [0.75,0.99] |
| | JACK-N | [-0.36,0.36] | [-0.11,0.57] | [0.17,0.76] | [0.51,0.91] | [0.76,0.98] |
| | BOOT-P | [-0.33,0.35] | [-0.11,0.54] | [0.15,0.71] | [0.47,0.86] | [0.72,0.95] |
| | BOOT_B | [-0.34,0.34] | [-0.12,0.54] | [0.14,0.71] | [0.47,0.86] | [0.73,0.95] |

Tabela 4 (continuação)

| | | Distribuição($\rho$) | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | BCS | BCS | BCS | BCS | BCN | BCN | TS |
| | | 0.25 | 0.50 | 0.75 | 0.90 | (a) | (b) | 0.44 |
| n | Método | | | | | | | |
| 15 | EBQP*-N | [-0.31,0.58] | [-0.11,0.73] | [0.22,0.89] | [0.54,0.99] | [0.42,0.97] | [0.37,0.97] | [0.16,0.87] |
| | BOOT-N | [-0.30,0.58] | [-0.10,0.73] | [0.21,0.90] | [0.51,1.00] | [0.40,0.98] | [0.36,0.98] | [0.13,0.90] |
| | JACK-N | [-0.34,0.61] | [-0.13,0.76] | [0.20,0.91] | [0.53,0.99] | [0.40,0.98] | [0.35,0.99] | [0.12,0.90] |
| | BOOT-P | [-0.30,0.57] | [-0.13,0.70] | [0.16,0.83] | [0.46,0.93] | [0.34,0.90] | [0.30,0.91] | [0.01,0.77] |
| | BOOT_B | [-0.33,0.55] | [-0.16,0.68] | [0.13,0.82] | [0.45,0.93] | [0.32,0.90] | [0.28,0.90] | [0.01,0.77] |
| 20 | EBQP*-N | [-0.24,0.52] | [-0.04,0.68] | [0.28,0.85] | [0.58,0.96] | [0.46,0.93] | [0.42,0.94] | [0.23,0.83] |
| | BOOT-N | [-0.24,0.52] | [-0.04,0.67] | [0.27,0.85] | [0.57,0.97] | [0.45,0.94] | [0.41,0.95] | [0.21,0.86] |
| | JACK-N | [-0.26,0.54] | [-0.05,0.69] | [0.26,0.86] | [0.58,0.96] | [0.46,0.94] | [0.40,0.95] | [0.21,0.85] |
| | BOOT-P | [-0.24,0.51] | [-0.06,0.65] | [0.23,0.80] | [0.52,0.91] | [0.40,0.88] | [0.36,0.89] | [0.12,0.74] |
| | BOOT_B | [-0.26,0.50] | [-0.07,0.64] | [0.21,0.80] | [0.52,0.91] | [0.40,0.88] | [0.35,0.88] | [0.12,0.75] |
| 25 | EBQP*-N | [-0.20,0.48] | [0.00,0.64] | [0.31,0.82] | [0.61,0.94] | [0.50,0.91] | [0.45,0.92] | [0.28,0.81] |
| | BOOT-N | [-0.20,0.48] | [0.00,0.64] | [0.31,0.82] | [0.60,0.95] | [0.49,0.92] | [0.45,0.92] | [0.26,0.83] |
| | JACK-N | [-0.21,0.49] | [-0.01,0.65] | [0.30,0.82] | [0.61,0.94] | [0.49,0.92] | [0.45,0.92] | [0.27,0.82] |
| | BOOT-P | [-0.20,0.47] | [-0.02,0.61] | [0.27,0.78] | [0.56,0.90] | [0.45,0.87] | [0.41,0.87] | [0.18,0.73] |
| | BOOT_B | [-0.21,0.47] | [-0.03,0.61] | [0.26,0.78] | [0.56,0.90] | [0.45,0.87] | [0.40,0.87] | [0.19,0.73] |

**Observação 2** : *Distribuições: normal bivariada (BVN), qui-quadrado bivariada (BCS), normal contaminada bivariada (BCN), "three squares"(TS), (a) BCN (0.75,2,0.5), (b) BCN (0.75,4,0.5). Métodos: EBQP*-N = I.C. assimptoticamente normal construído com a variância EBQP* com correcção empírica, BOOT -N = I.C. a. normal construído com a variância de*

*bootstrap (B=250), JACK-N =I.C. a. normal construído com a variância de jacknife, BOOT-P = I.C. construído com o percentil de bootstrap, BOOT-B = I.C. construído com o percentil* $BC_a$.

Os resultados registados na tabela 4 mostram que os intervalos assimptoticamente normais construídos com a variância estimada pelo método EBQP*, $\hat{\sigma}^2_{E^*}(r_w)$, e com a variância estimada pelo algoritmo de bootstrap, $\hat{\sigma}^2_B(r_w)$, têm, de um modo geral, amplitudes bastante aproximadas.

Em amostras de pequena dimensão ($n \lesssim 10$), os intervalos assimptoticamente normais construídos com a variância estimada pelo algoritmo de jackknife, $\hat{\sigma}^2_J(r_w)$, têm amplitude bastante superior aos intervalos construídos pelos outros métodos. Esta diferença de amplitudes já era esperada como consequência de $\hat{\sigma}^2_J(r_w)$ ser bastante superior à variância estimada pelos outros métodos em amostras de pequena dimensão, como é possível observar na tabela 3.

Para amostras provenientes das distribuições $BVN$ e $BCS$ com correlação baixa ($\rho \lesssim 0.25$), os intervalos construídos com percentis de bootstrap com viés corrigido $BC_a$ têm menor amplitude que os intervalos construídos com percentis simples de bootstrap. Contudo, essa tendência inverte-se quando $\rho \gtrsim 0.25$.

Em amostras provenientes das distribuições normal contaminada bivariada ($BCN$ com $\tau = 2,4$ e $\varepsilon = 0.5$) e "three squares"($TS$ com $n \gtrsim 8$) os intervalos construídos com percentis de bootstrap com viés corrigido $BC_a$ têm maior amplitude que os intervalos construídos com percentis simples de bootstrap.

De um modo geral, em amostras de dimensão $n \gtrsim 10$ verifica-se que as amplitudes dos intervalos assimptoticamente normais tendem a ser menores que as amplitudes dos intervalos construídos com os percentis de bootstrap

Em amostras com dimensão $n = 25$ já é possível observar que os intervalos construídos pelos diferentes métodos apresentam diferenças muito ligeiras de amplitudes, isto obviamente, considerando as aproximações a duas casas decimais.

A análise a respeito dos intervalos construídos para $\rho_w$ fica-se apenas pela observação das suas amplitudes. Recordo que não nos foi possível estudar as proporções de $\rho_w$ estar compreendido entre os limites inferior e superior do intervalo de confiança construído.

Em Borkowf (2000) foram calculadas as proporções de cobertura para os intervalos de confiança construídos para $\rho_s$. A seguir apresentamos algumas conclusões desse estudo que, na globalidade, legitimam a aplicação dos diversos métodos usados neste trabalho para construção de intervalos de confiança para $\rho_s$. De algum modo, essas conclusões convidam a aplicar estes processos de construção de intervalos de confiança para outras estatísticas não paramétricas como, por exemplo, o coeficiente $\rho_w$.

## 4.3 Intervalos de confiança para $\rho_s$

Em Borkowf (2000), encontram-se registados resultados de simulações realizadas para o estudo das proporções do coeficiente de correlação de Spearman $\rho_s$ estar compreendido entre os limites inferior e superior de cada um dos intervalos de confiança a 90% e 95% construídos. Estas proporções designam-se de porporções de cobertura e foram calculadas com base em 40 000 conjuntos de dados simulados (10 000 para $n = 20,\ 25$). As amostras são provenientes das distribuições apresentadas na secção 3.1 e de dimensões $n = 5, 8, 10, 12, 15, 20, 25$.

Os intervalos de confiança foram construídos pelos diversos métodos descritos acima (foram usadas as transformações de Fisher) combinados com o uso das variâncias estimadas pelos métodos EBQP*, bootstrap e jackknife.

Nas tabelas apresentadas em Borkowf (2000) é possível observar que em geral, em amostras de dimensão ($n \leq 8$), nenhum dos métodos produz bons resultados devido ao facto do viés da estimação de $r_s$ tender para zero e as distribuições serem discretas.

Em amostras de dimensão ($10 \leq n \leq 25$), os intervalos de confiança assimptoticamente normais construídos com a variância estimada pelos métodos EBQP*, bootstrap ($\rho$ pequeno), e jackknife formam uma cobertura sub-nominal (proporção de cobertura inferior ao nível de confiança do intervalo), especialmente nos intervalos a 95%. Além disso, intervalos de confiança assimptoticamente normais construídos com a variância estimada pelo método de bootstrap para $\rho$ grande formam uma cobertura supra-nominal (proporção de cobertura superior ao nível de confiança do intervalo), devido ao facto de, nesse caso, a variância de bootstrap ser maior que as variâncias estimadas pelos outros métodos.

Em amostras de dimensão ($10 \leq n \leq 25$) com distribuição $BVN$ e $BCS$, os intervalos de confiança construídos com a transformação de Fisher e variância estimada pelo algoritmo de jackknife formam uma cobertura ligeiramente supra-nominal para qualquer valor de $\rho$. Os intervalos construídos com a transformação de Fisher e variância de bootstrap tendem a formar uma cobertura supra-nominal com o crescer da correlação $\rho$, enquanto os intervalos construídos com a variância estimada pelo método EBQP* e com transformação de Fisher formam uma cobertura nominal (proporção de cobertura muito próxima do nível de confiança do intervalo) para todos os valores de $\rho$.

Intervalos construídos com o percentil simples de bootstrap tendem a formar uma cobertura nominal para $\rho$ pequeno e uma cobertura supra-nominal para $\rho$ grande. Intervalos construídos com o percentil de bootstrap $BC_a$ formam uma cobertura ligeiramente supra-nominal para qualquer valor de $\rho$.

Para distribuições suaves ( $BVN$ e $BCS$ ), os intervalos de confiança construídos com as transformações de Fisher combinadas com variância estimada pelo método EBQP* ou pelo algoritmo de jackknife e os intervalos construídos pelo percentil de bootstrap $BC_a$ são os que têm maior proporção de cobertura.

Para a distribuição $TS$ ou distribuições com caudas pesadas ($BCN$), os intervalos construídos com o percentil simples de bootstrap e os intervalos construídos com o uso da transformação de Fisher e com a variância de bootstrap apresentam maiores proporções

de cobertura.

De referir que os intervalos construídos com o uso da transformação de Fisher e com variância estimada pelo método EBQP* tendem a ter menor amplitude relativamente aos intervalos com proporções de cobertura comparáveis, especialmente em valores médios de $\rho$.

# Capítulo 5

# Conclusões gerais

Em determinados problemas, o uso do coeficiente de correlação de ordens $\rho_w$ pode ser mais adequado do que o uso do coeficiente de correlação de Spearman $\rho_s$.

Como o leitor se deve ter apercebido, pelo levantamento de resultados relacionados com a distribuição dos estimadores destes coeficientes efectuado nos primeiro e segundo capítulos, existem muitas dificuldades ao estudar o comportamento das suas variâncias. Essas dificuldades são maiores, sobretudo, quando não existe independência dos dados e quando a distribuição dos dados é desconhecida. Assim, justifica-se o recurso a alguns métodos de estimação não paramétrica com vista a determinar as variâncias de $r_s$ e $r_w$, e a construir intervalos de confiança para $\rho_s$ e $\rho_w$.

No que diz respeito à estimação não paramétrica da variância de $r_s$, Borkowf (2000) apresenta um estudo, baseado em simulações, em que usa o método EBQP e os algoritmos de bootstrap e jackknife. Nesse estudo também foram determinadas as proporções de cobertura dos intervalos de confiança para $\rho_s$ construídos com as variâncias estimadas pelos diferentes métodos. Os resultados desse estudo mostram que o método EBQP pode ser adaptado com sucesso na estimação da variância amostral para $r_s$ e sugerem a aplicação deste método para estimação da variância de outras estatísticas não paramétricas calculadas a partir das ordens (ranks) de amostras bivariadas.

Tendo em consideração as conclusões do referido estudo e as dificuldades enunciadas

no capítulo 2 para determinar a variância de $r_w$, desenvolvemos um estudo idêntico em que usamos os mesmos métodos de estimação não paramétrica com as necessárias adaptações.

Ao contrário do estudo desenvolvido por Borkowf (2000), no nosso estudo não nos foi possível observar quais os métodos que fornecem os melhores resultados (excepto no caso de independência dos dados) e quais os intervalos de confiança com maior proporção de cobertura. Contudo, no nosso estudo podemos observar que a variância estimada para $r_w$ pelo método EBQP* é quase sempre inferior à variância estimada pelo algoritmo de jackknife, tal como sucede na estimação da variância de $r_s$. Ao contrário do que se verifica na estimação da variância de $r_s$, a variância estimada para $r_w$ pelo método EBQP* nem sempre toma valores inferiores à variância estimada pelo algoritmo de bootstrap.

Tendo em conta os resultados obtidos nas simulações para estimação da variância de $r_w$ no caso de independência dos dados, sugerimos o método EBQP* e o algoritmo de bootstrap como os melhores métodos para estimação da variância de $r_w$.

Na construção de intervalos de confiança para $\rho_w$ podemos ser tentados a concluir que, dos intervalos assimptoticamente normais, o intervalo assimptoticamente normal construído com a variância estimada pelo método EBQP* poderá ser o que apresenta maior proporção de cobertura.

Como trabalhos futuros pensamos: concluir a prova formal da normalidade assimptótica da distribuição do coeficiente $r_w$; determinar uma aproximação para a variância assimptótica de $r_w$ em diversas distribuições bivariadas; testar as proporções de cobertura dos intervalos de confiança a 90% construídos para o coeficiente $\rho_w$; estender a aplicação de $\rho_w$ a amostras multivariadas e propor outros coeficientes de correlação de ordens que se adaptem melhor na resolução de determinados problemas.

# Bibliografia


[1] Aptech Systems, Inc., 1999. The GAUSS System, Versão 5.0. Aptech Systems, Maple Valley, Washington.

[2] Billingsley , 1978. *Probability and measure*, Probability and Mathematical Statistics, Wiley.

[3] Bishop, Y.M.M., Fienberg, S. E., Holland, P.W., 1975. *Discrete Multivariate Analysis.* MIT Press, Cambridge, MA.

[4] Borkowf, C.B., Gail, M.H., Raymond, J.C., Richard, D.G.,1997. Analyzing bivariate continuous data grouped into categories defined by empirical quantiles of marginal distributions. *Biometrics* 53, 1054-1069.

[5] Borkowf, C.B., Gail, M.H., 1997. On measures of agreement calculated from contigency tables with categories defined by the empirical quantiles of the marginal distributions. *American Journal of Epidemiology* 146, 520-526.

[6] Borkowf, C.B., 2000. A new nonparametric method for variance estimation and confidence interval construction for Spearman's rank correlation. *Computacional Statistics & Data Analysis* 34, 219-241.

[7] Borkowf, C.B., 2000. On multidimensional contingency tables with categories defined by the empirical quantiles of the marginal data. *Journal of Statistical Planning and Inference* 91, 33-51.

[8] Borkowf, C.B., 2002. Computing the nonnull asymptotic variance and the asymptotic relative efficiency of Spearman's rank correlation. *Computacional Statistics & Data Analysis* 39, 271-286.

[9] Costa, J.F.P., Soares, C.M.M.O, 2002. Weighted Rank correlation (Submetido em 2002).

[10] David, S.T., Kendall, M.G., Stuart, A., 1951. Some questions of distribution in the theory of rank correlation. *Biometrika* 38, 131-140.

[11] David, F.N., Mallows, C. L., 1961. The variance of Spearman's rho in normal samples. *Biometrika* 48, 19-28.

[12] Efron, B., Tibshirani, R., 1993. *An Introduction to the Bootstrap.* Chapman & Hall.

[13] Fieller, E.C., Hartey, H.O., Pearson, E.S., 1957. Tests for rank correlation coefficients. I. *Biometrika* 44, 470-481.

[14] Kendall, M.G., 1949. Rank and product-moment correlation. *Biometrika* 36, 177-193.

[15] Kendall, M.G., Kendall, F.H., Smith, B.B., 1939. The distribution of Spearman's coefficient of rank correlation in a universe in which all rankings occur an equal number of times. *Biometrika* 30, 251-273.

[16] Kotz, S., Johnson, N.L., Balakrishnan, N., 2000. *Continuous multivariate distributions: models and applications.* Wiley, New York.

[17] Lehmann, E.L., 1975. *Nonparametrics: Statistical Methods Based on Ranks.* Holden-Day.

[18] Meier, U., 1997. On the asymptotic normality of rank tests for independence. *Journal of Statistical Planning and Inference* 61, 279-296.

[19] Moran, P.A.P., 1948. Rank correlation and product moment correlation. *Biometrika* 35, 203-206.

[20] Randles and Wolf, 1979. *Introduction to the theory of Nonparametric Statistics,* Probability and Mathematical Statistics, Wiley.

[21] Spearman, C., 1904. The proof and measurement of association between two things. *Amer. J. Psychol.* 15, 72-101.

[22] *Spearman, C., 1906. 'Footrule' for measuring correlation. British J. Psychol.* 2, 89-108.

[23] Waterloo Maple, Inc., 2001. MAPLE Software, versão 7.0. Waterloo Maple, Ontario.

[24] Wiel, M.A., Bucchianico, A., 2001. Fast computation of the exact null distribution of Spearman's $\rho$ and Page's L statistics for samples with and without ties. *Journal of Statistical Planning and Inference* 92, 133-145.

# Apêndice A

**Teorema 1**

(Costa & Soares,2002): A soma $\sum_{i=1}^{n}(r(X_i)-r(Y_i))^2((n-r(X_i)+1)+(n-r(Y_i)+1))$ é máxima quando $r(Y_i)=n-r(X_i)+1$ e assume o valor $\frac{n^4+n^3-n^2-n}{3}$.

**Demonstração:**

Queremos provar que a permutação $(R_1,...,R_n)$ que maximiza a expressão

$$\sum_{i=1}^{n}(i-R_i)^2((n-i+1)+(n-R_i+1))=\sum_{i=1}^{n}(i-R_i)^2(2(n+1)-i-R_i) \qquad (A1)$$

é a permutação $(n,...,1)$, ou seja, $r(X_i)=R_i=n-i+1$. Note-se que estamos a assumir $r(Y_i)=i$, sem perda de generalidade.

Suponhamos que $(R_1,...,R_n)$ não é a permutação que maximiza a soma $(A1)$ e que $(R'_1,...,R'_n)$ é essa permutação. Existem dois inteiros $l$ e $m$, $l\leq n$ e $m\leq n$, tal que $R'_l\neq R_l$ e $R'_m\neq R_m$. Seja $l=\min_{i=1,...,n}\{i:R'_i\neq n-i+1\}$. Note-se que $l$ é o primeiro inteiro tal que $R'_l\neq n-l+1$. É óbvio que $R'_l<R_l=n-l+1$. Consideremos $m$ tal que $R'_m=R_l=n-l+1$. Temos a seguinte situação:

| | 1 | 2 | ... | $l-1$ | $l$ | ... | $m$ | ... | $n$ |
|---|---|---|---|---|---|---|---|---|---|
| $R$ | $n$ | $n-1$ | ... | $n-l+2$ | $n-l+1$ | ... | $n-m+1$ | ... | 1 |
| $R'$ | $n$ | $n-1$ | ... | $n-l+2$ | $R'_l$ | ... | $n-l+1$ | ... | $R'_n$ |

Vamos demonstrar que se trocarmos $R'_l$ com $R'_m$ na permutação $R'$ obtemos uma permutação que faz com que a soma $(A1)$ seja maior, o que é absurdo, visto que, por hipótese, $R'$ é a permutação que maximiza a soma $(A1)$.

Se trocarmos $R'_l$ com $R'_m$ então $R_l$ passa a ocupar a posição $R'_l$ e $R'_l$ passa a ocupar a posição $R'_m$. Tendo em atenção que $m-l>0, R'_l<R_l$ e $l,m,R_l,R'_l\leq n$; podemos concluir que:

$$(l-R_l)^2(n-l+1+n-R_l+1)+(m-R'_l)^2(n-m+1+n-R'_l+1)-$$
$$(l-R'_l)^2(n-l+1+n-R'_l+1)+(m-R_l)^2(n-m+1+n-R_l+1)$$

$= -(m-l)(R'_l - R_l)(4n+4-l-m-R_l-R'_l) > 0.$

Isto significa que a soma $(A1)$ é maior quando efectuamos as trocas de $R'_l$ com $R'_m$ na permutação $R'$, o que é absurdo, pois por hipótese $R'$ é a permutação que maximiza a soma $(A1)$. Como esta hipótese é falsa, a permutação $R = (n, ..., 1)$ é que maximiza a soma $(A1)$.

Para obter o valor máximo da soma, basta substituir em $(A1)$ $R_i$ por $n-i+1$.

# Apêndice B

**Teorema 2**

(Costa & Soares, 2002): Se $R = (R_1, ..., R_n)$ e $Q = (Q_1, ..., Q_n)$ são vectores de ordens independentes, então

*i)* $E(r_w) = 0$

*ii)* $var(r_w) = \dfrac{31n^2 + 60n + 26}{30(n^3 + n^2 - n - 1)}$

**Demonstração:**

Vamos começar por escrever $r_w$ como combinação linear de estatísticas lineares ordinais.

$$\sum_{i=1}^{n}(i - R_i^*)^2(2(n+1) - i - R_i^*) = \sum_{i=1}^{n}(i^2 - R_i^{*2} - 2iR_i^*)(2(n+1) - i - R_i^*)$$

$$= 2(n+1)\sum_{i=1}^{n} i^2 - \sum_{i=1}^{n} i^3 + 2(n+1)\sum_{i=1}^{n} R_i^{*2} - \sum_{i=1}^{n} R_i^{*3} + \sum_{i=1}^{n} i^2 R_i^* + \sum_{i=1}^{n} iR_i^{*2}$$

$$-4(n+1)\sum_{i=1}^{n} iR_i^* = 4(n+1)\sum_{i=1}^{n} i^2 - 2\sum_{i=1}^{n} i^3 + \sum_{i=1}^{n} i^2 R_i^* + \sum_{i=1}^{n} iR_i^{*2}$$

$$-4(n+1)\sum_{i=1}^{n} iR_i^*.$$

O termo $C = 4(n+1)\sum_{i=1}^{n} i^2 - 2\sum_{i=1}^{n} i^3$ é constante pois não depende de $R^*$.

Sejam (ver 2.6) ,

$S_{1n} = \sum_{i=1}^{n} iR_i^*$,

$S_{2n} = \sum_{i=1}^{n} iR_i^{*2}$ e

$S_{3n} = \sum_{i=1}^{n} i^2 R_i^*$.

A distribuição de $r_w$ sob $H_0$ é a mesma de

$$1 - \frac{6C}{n(n^3+n^2-n-1)} + \frac{24(n+1)}{n(n^3+n^2-n-1)} S_{1n}$$
$$- \frac{6}{n(n^3+n^2-n-1)} S_{2n} - \frac{6}{n(n^3+n^2-n-1)} S_{3n} \qquad \text{(B1)}$$

Por (2.5 $i$)) temos,

$$E(S_{1n}) = n \frac{\sum_{i=1}^{n} i}{n} \frac{\sum_{i=1}^{n} i}{n} = \frac{n(n+1)^2}{4},$$

$$E(S_{2n}) = n \frac{\sum_{i=1}^{n} i}{n} \frac{\sum_{i=1}^{n} i^2}{n} = \frac{n(n+1)}{2} \frac{n(n+1)(2n+1)}{6n} = \frac{n(n+1)^2(2n+1)}{12},$$

$$E(S_{3n}) = n \frac{\sum_{i=1}^{n} i^2}{n} \frac{\sum_{i=1}^{n} i}{n} = \frac{n(n+1)}{2} \frac{n(n+1)(2n+1)}{6n} = \frac{n(n+1)^2(2n+1)}{12}.$$

Assim,

$$E(r_w) = 1 - \frac{6C - 6(n+1)^3 n + n(n+1)^2(2n+1)}{n(n^3+n^2-n-1)} = 0.$$

Portanto, se dois vectores de ordens são independentes, o valor esperado de $r_w$ é 0.

A expressão para a variância de $r_w$ sob $H_0$ é obtida tendo em conta que:

$$\begin{aligned} var(r_w) &= \left( \frac{6}{n(n^3+n^2-n-1)} \right)^2 var(4(n+1)S_{1n} - S_{2n} - S_{3n}) \\ &= \left( \frac{6}{n(n^3+n^2-n-1)} \right)^2 (16(n+1)^2 var(S_{1n}) + var(S_{2n}) + var(S_{3n}) \\ &\quad -8(n+1)cov(S_{1n}, S_{2n}) - 8(n+1)cov(S_{1n}, S_{3n}) + 2cov(S_{2n}, S_{3n})) \qquad \text{(B2)} \end{aligned}$$

Por (2.5 *ii*)), considerando $c(i) = i$ e $a(i) = i$ :

$$var(S_{1n}) = \frac{1}{n-1}\left[\sum_{i=1}^{n} i^2 - n\frac{(n+1)^2}{4}\right]\left[\sum_{i=1}^{n} i^2 - n\frac{(n+1)^2}{4}\right]$$

$$= \frac{n^2(n+1)^2(n-1)}{144},$$

considerando $c(i) = i$ e $a(i) = i^2$ :

$$var(S_{2n}) = var(S_{3n}) = \frac{1}{n-1}\left[\sum_{i=1}^{n} i^2 - n\frac{(n+1)^2}{4}\right]\left[\sum_{i=1}^{n} i^4 - n\frac{(n+1)^2(2n+1)^2}{36}\right]$$

$$= \frac{n^2(n+1)^2}{2160}(16n^3 + 14n^2 - 19n - 11).$$

Por (2.5 *iii*)),

$$cov(S_{1n}, S_{2n}) = cov(S_{1n}, S_{3n}) = \frac{1}{144}(n^4 + 2n^3 - 2n - 1)n^2$$

e

$$cov(S_{2n}, S_{3n}) = \frac{1}{144}(n^5 + 3n^4 + 2n^3 - 2n^2 - 3n - 1)n^2.$$

Finalmente, substituindo estes resultados em $(B2)$, obtemos a variância de $r_w$ sob a hipótese de independência ente dois vectores de ordens:

$$var(r_w) = \frac{31n^2 + 60n + 26}{30(n^3 + n^2 - n - 1)}.$$