

UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PÓS-GRADUAÇÃO EM ESTATÍSTICA

Avaliação de uma classe de coeficientes de correlação sob normalidade
bivariada

Cleber Martins Xavier

Orientador: Prof. Dr. Raydonal Ospina Martínez
Co-Orientador: Prof. Dr. Gustavo Henrique Esteves

Área de concentração: Estatística Aplicada

Recife, 19 fevereiro de 2014

UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PÓS-GRADUAÇÃO EM ESTATÍSTICA

Avaliação de uma classe de coeficientes de correlação sob normalidade
bivariada

Cleber Martins Xavier

Orientador: Prof. Dr. Raydonal Ospina Martínez
Co-Orientador: Prof. Dr. Gustavo Henrique Esteves

Área de concentração: Estatística Aplicada

Dissertação submetida como requerimento parcial para obtenção do grau de
Mestre em Estatística pela Universidade Federal de Pernambuco

Recife, 19 fevereiro de 2014

Catálogo na fonte
Bibliotecária Jane Souto Maior, CRB4-571

Xavier, Cleber Martins

**Avaliação de uma classe de coeficientes de correlação
sob normalidade bivariada / Cleber Martins Xavier. - Recife: O
Autor, 2014.**

98 f.: il., fig., tab.

Orientador: Raydonal Ospina Martínez.

**Dissertação (mestrado) - Universidade Federal de Pernambuco.
CCEN, Estatística, 2014.**

Inclui referências e apêndices.

1. Estatística aplicada. I. Martínez, Raydonal Ospina (orientador).

II. Título.

310

CDD (23. ed.)

MEI2014 – 025

Universidade Federal de Pernambuco
Pós-Graduação em Estatística

19 de fevereiro de 2014

Nós recomendamos que a Dissertação de Mestrado de autoria de

Cleber Martins Xavier

Intitulada:

“Avaliação de uma Classe de Coeficientes de Correlação Sob Normalidade Bivariada”

Seja aceita como cumprimento parcial dos requerimentos para o grau de Mestre em Estatística.

Coordenador da Pós-Graduação em Estatística

Banca Examinadora:

Raydonal Ospina Martínez	Orientador/UFPE
--------------------------	-----------------

Juvêncio Santos Nobre	UFC
-----------------------	-----

Francisco Cribari Neto	UFPE
------------------------	------

Este documento será anexado à versão final da tese.

Aos meus pais, minha esposa e minha filha.

Agradecimentos

Gostaria de agradecer primeiramente a meus pais por todo apoio durante minha vida acadêmica e a minha avó “Dona Cila” que sempre me ajudou na realização dos meus sonhos.

Agradeço à minha esposa Edielli (“Jay”) pela “paciência” durante toda esta jornada, principalmente nos momentos dedicados a nossa filha Alice Rafaela nos períodos em que não estive presente.

Ao professor Gustavo Henrique Esteves por me ajudar na escolha do meu caminho e, assim, escolhendo a Universidade Federal de Pernambuco para conduzir meus estudos de pós-graduação.

Ao meu orientador o professor Raydonal Ospina por me tornar seu orientando durante todo o período deste trabalho. Por sua paciência e entusiasmo, sempre me incentivando, por todos os ensinamentos e troca de experiências (acadêmicas e cotidianas) transmitidas durante este ano que, com certeza, levarei por toda minha vida.

Aos meus companheiros de mestrado, em especial Jéssica Rivas e Daniele Trindade (as “baianas”) pelas longas horas de estudo e pelos momentos de alegria. Também queria agradecer a Pedro Rafael pela ótima convivência que tivemos principalmente durante o primeiro ano de mestrado apesar de todas as dificuldades que passamos.

Aos Professores do Departamento de Estatística da UFPE, em especial aos professores Francisco Cribari, Gauss Cordeiro, Leandro Rêgo e Audrey Cysneiros por todo conhecimento que me passaram durante o curso.

Aos participantes da banca examinadora, pelas sugestões e correções para o aprimoramento deste trabalho.

A CAPES pelo apoio financeiro.

*A ausência da evidência
não significa evidência da ausência.*

*“The absence of evidence
is not the evidence of absence.”*

Carl Sagan

Resumo

O coeficiente de correlação é uma medida que permite mensurar a dependência entre duas variáveis. Na literatura estatística existem diversos coeficientes de correlação, como por exemplo o de Pearson (Pearson, 1931 e Pearson, 1932), Kendall (Kendall, 1938) e Spearman (Spearman, 1904) que são usados sob diferentes condições dos dados. Nesta dissertação discutimos sobre o coeficiente de correlação generalizado proposto por Chinchilli *et al.* (2005), o qual é obtido através de estatísticas U . Este coeficiente tem como casos especiais o coeficiente de correlação de Pearson e de Kendall. Neste trabalho foram desenvolvidos estudos de simulação a fim de avaliar o comportamento do coeficiente de correlação generalizado em pequenas e grandes amostras e sob diferentes cenários que incluem amostras normais bivariadas, amostras bivariadas contaminadas e perturbadas. Por fim, utilizando um banco de dados real de microarrays usamos o coeficiente de correlação generalizado na construção de uma rede de relevância.

Palavras chave: Coeficiente de correlação generalizado, Coeficiente de correlação de Spearman, Contaminação, Distribuição normal bivariada, Perturbação, Simulação.

Abstract

Correlation coefficient is a measure of dependence between two variables. In the literature there are several statistical correlation coefficients, such as the Pearson's (Pearson, 1931 and Pearson, 1932), Kendall's (Kendall, 1938) and Spearman's (Spearman, 1904) that are used under different conditions of the data. In this thesis we discuss the generalized correlation coefficient proposed by Chinchilli *et al.* (2005), which is obtained through U -statistics. This coefficient has as special cases the correlation coefficient of Pearson's and Kendall's. In this study we develop simulation studies in order to assess the behavior of the generalized correlation coefficient in small and large samples and under different scenarios that include bivariate normal samples, contaminated samples and disturbed bivariate samples. Finally, utilizing a real database we used the generalized correlation coefficient to build a relevance network.

Keywords: Bivariate normal distribution, Contamination, Disturbance, Generalized correlation coefficient, Simulation, Spearman's correlation coefficient.

Lista de Figuras

1.1	Gráfico com os possíveis tipos de correlação linear entre duas variáveis.	16
2.1	Gráfico para a forma explícita do coeficiente de correlação generalizado com diferentes valores de γ .	32
3.1	Gráfico da função densidade com as respectivas curvas de níveis da distribuição normal bivariada para os quatro casos com amostras sem perturbação e com $\rho = 0.3$.	36
3.2	Gráfico da função densidade com as respectivas curvas de níveis da distribuição normal bivariada para os cinco casos de misturas de normais com $\rho = 0.3$.	39
3.3	Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ para o primeiro caso das amostras normais bivariadas sem perturbação e considerando $\rho = 0$.	41
3.4	Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ para o primeiro caso das amostras normais bivariadas sem perturbação e considerando $\rho = 0.3$ e $\rho = 0.9$.	46
3.5	Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ para o segundo caso das amostras normais bivariadas sem perturbação e considerando $\rho = 0$.	47
3.6	Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ para o segundo caso das amostras normais bivariadas sem perturbação e considerando $\rho = 0.3$ e $\rho = 0.9$.	48
3.7	Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ para o terceiro caso das amostras normais bivariadas sem perturbação e considerando $\rho = 0$.	49

3.8	Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ para o terceiro caso das amostras normais bivariadas sem perturbação e considerando $\rho = 0.3$ e $\rho = 0.9$	50
3.9	Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ para o quarto caso das amostras normais bivariadas sem perturbação e considerando $\rho = 0$	51
3.10	Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ para o quarto caso das amostras normais bivariadas sem perturbação e considerando $\rho = 0.3$ e $\rho = 0.9$	54
3.11	Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ com 10% de perturbação da amostra e vários valores de γ considerando $\rho = 0.1$	55
3.12	Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ com 30% de perturbação da amostra e vários valores de γ considerando $\rho = 0.1$	56
3.13	Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ com 50% de perturbação da amostra e vários valores de γ considerando $\rho = 0.1$	57
3.14	Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ com 10% de perturbação da amostra e vários valores de γ considerando $\rho = 0.5$	59
3.15	Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ com 30% de perturbação da amostra e vários valores de γ considerando $\rho = 0.5$	60
3.16	Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ com 50% de perturbação da amostra e vários valores de γ considerando $\rho = 0.5$	61
3.17	Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ com 10% de perturbação da amostra e vários valores de γ considerando $\rho = 0.9$	63
3.18	Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ com 30% de perturbação da amostra e vários valores de γ considerando $\rho = 0.9$	64

3.19	Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ com 50% de perturbação da amostra e vários valores de γ considerando $\rho = 0.9$	65
3.20	Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ para o primeiro caso da mistura de duas normais bivariadas.	68
3.21	Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ para o segundo caso da mistura de duas normais bivariadas.	70
3.22	Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ para o terceiro caso da mistura de duas normais bivariadas.	71
3.23	Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ para o quarto caso da mistura de duas normais bivariadas.	73
3.24	Gráfico Boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ para o quinto caso da mistura de duas normais bivariadas.	75
4.1	Gráfico empírico para visualização da normalidade dos dados.	79
4.2	Gráfico (redes de relevância) usando o coeficiente de correlação generalizado e o de Spearman para diferentes valores de γ usando o pacote maigesPack.	80
4.3	Gráfico (redes de relevância) usando o coeficiente de correlação generalizado e o de Spearman para diferentes valores de γ usando o editor de gráficos yEd.	81

Lista de Tabelas

3.1	Raiz do erro quadrático médio das estimativas do coeficiente de correlação generalizado $\hat{\rho}_\gamma$, $\tilde{\rho}_\gamma$ e $\bar{\rho}_{S_\gamma}$ para o primeiro caso sem perturbação da amostra.	42
3.2	Raiz do erro quadrático médio das estimativas do coeficiente de correlação generalizado $\hat{\rho}_\gamma$, $\tilde{\rho}_\gamma$ e $\bar{\rho}_{S_\gamma}$ para o segundo caso sem perturbação da amostra.	44
3.3	Raiz do erro quadrático médio das estimativas do coeficiente de correlação generalizado $\hat{\rho}_\gamma$, $\tilde{\rho}_\gamma$ e $\bar{\rho}_{S_\gamma}$ para o terceiro caso sem perturbação da amostra.	45
3.4	Raiz do erro quadrático médio das estimativas do coeficiente de correlação generalizado $\hat{\rho}_\gamma$, $\tilde{\rho}_\gamma$ e $\bar{\rho}_{S_\gamma}$ para o quarto caso sem perturbação da amostra.	53
3.5	Raiz do erro quadrático médio das estimativas do coeficiente de correlação generalizado $\hat{\rho}_\gamma$, $\tilde{\rho}_\gamma$ e $\bar{\rho}_{S_\gamma}$ com perturbação da amostra aos níveis de 10%, 30% e 50% com $\rho = 0.1$	58
3.6	Raiz do erro quadrático médio das estimativas do coeficiente de correlação generalizado $\hat{\rho}_\gamma$, $\tilde{\rho}_\gamma$ e $\bar{\rho}_{S_\gamma}$ com perturbação da amostra aos níveis de 10%, 30% e 50% com $\rho = 0.5$	62
3.7	Raiz do erro quadrático médio das estimativas do coeficiente de correlação generalizado $\hat{\rho}_\gamma$, $\tilde{\rho}_\gamma$ e $\bar{\rho}_{S_\gamma}$ com perturbação da amostra aos níveis de 10%, 30% e 50% com $\rho = 0.9$	67
3.8	Raiz do erro quadrático médio das estimativas do coeficiente de correlação generalizado $\hat{\rho}_\gamma$, $\tilde{\rho}_\gamma$ e $\bar{\rho}_{S_\gamma}$ para o primeiro caso da mistura de duas distribuições normais bivariadas.	69

3.9	Raiz do erro quadrático médio das estimativas do coeficiente de correlação generalizado $\hat{\rho}_\gamma$, $\tilde{\rho}_\gamma$ e $\bar{\rho}_{S_\gamma}$ para o segundo caso da mistura de duas distribuições normais bivariadas.	69
3.10	Raiz do erro quadrático médio das estimativas do coeficiente de correlação generalizado $\hat{\rho}_\gamma$, $\tilde{\rho}_\gamma$ e $\bar{\rho}_{S_\gamma}$ para o terceiro caso da mistura de duas distribuições normais bivariadas.	72
3.11	Raiz do erro quadrático médio das estimativas do coeficiente de correlação generalizado $\hat{\rho}_\gamma$, $\tilde{\rho}_\gamma$ e $\bar{\rho}_{S_\gamma}$ para o quarto caso da mistura de duas distribuições normais bivariadas.	74
3.12	Raiz do erro quadrático médio das estimativas do coeficiente de correlação generalizado $\hat{\rho}_\gamma$, $\tilde{\rho}_\gamma$ e $\bar{\rho}_{S_\gamma}$ para o quinto caso da mistura de duas distribuições normais bivariadas.	76
4.1	Amostras de tecidos gastroesofágicos classificadas em normal, com inflamação e com metaplasia.	78

Sumário

1	Introdução	15
1.1	Introdução	15
1.2	Organização da Dissertação	17
1.3	Suporte Computacional	17
2	Coefficiente de Correlação	18
2.1	Introdução	18
2.2	Definições	18
2.2.1	Estimador de Máxima Verossimilhança do Coeficiente de Correlação Sob Normalidade Bivariada	21
2.3	Coefficiente de Correlação Generalizado	22
2.3.1	Estimação do Coeficiente de Correlação Generalizado Usando Estatísticas U	24
3	Estudo de Simulação	34
3.1	Introdução	34
3.2	Cenários de Simulação	34
3.2.1	Amostra Sem Contaminação	34
3.2.2	Amostra Perturbada	35
3.2.3	Mistura de Duas Distribuições Normais Bivariadas	37
3.3	Resultado das simulações	38
3.3.1	Análise dos estimadores sem perturbação da amostra	40
3.3.2	Análise dos estimadores com perturbação da amostra	55
3.3.3	Mistura de distribuições normais bivariadas	66
4	Aplicação	77

5	Considerações Finais	82
	Referências	84
A	Aspectos importantes	87
A.1	Distribuição Normal Bivariada	87
A.1.1	Função de Densidade e Distribuição Acumulada da Normal Bivariada	87
A.2	Distribuição t -Student Bivariada	88
A.3	Mistura de Distribuições	88
A.4	Método de Monte Carlo	89
B	Estatísticas U	92
B.1	Definições e Propriedades	92

Capítulo 1

Introdução

1.1 Introdução

Atualmente na Biologia existe um crescente interesse no estudo dos valores de expressão gênica interagentes. Uma forma de se estudar estes valores é através da avaliação das interações entre os níveis de expressão dos genes de interesse, o que nos dá um esboço dos perfis de interação produzidas por estes genes. Existem na literatura diversos métodos matemáticos e estatísticos que tentam extrair perfis de interação.

Uma abordagem adequada para avaliar os níveis de expressão gênica interagentes é a construção de redes de relevância (Butte e Kohane, 1999). Neste método utilizamos o coeficiente de correlação como forma de selecionar os genes que interagem entre si.

De modo geral, podem ser definidos diferentes coeficientes de correlação entre as variáveis de interesse. A forma mais conhecida na literatura é a correlação simples que envolve duas variáveis, X e Y . Dizemos que a relação entre duas variáveis é linear quando a relação pode ser interpretada por meio de uma reta do tipo $Y = \alpha + \beta X + e$ em que β mensura a força dessa relação e e representa o erro amostral. Neste caso o coeficiente de correlação linear consegue detectar esse comportamento.

Uma forma comumente utilizada para verificar o tipo de associação existente entre duas variáveis é através do gráfico de “dispersão”. Neste tipo de gráfico são apresentados os pares (X_i, Y_i) com $i = 1, 2, \dots, n$, em que n é o número total de observações. Na Figura 1.1 apresentamos algumas possíveis relações lineares entre duas variáveis com: (a) Correlação linear negativa, (b) Correlação linear positiva e (c) Correlação linear nula.

Em algumas situações não é possível ajustar uma reta para determinar a relação entre as variáveis. Nessa situação o coeficiente de correlação linear não é mais

adequado e assim novas medidas devem ser usadas para a detecção da dependência (linear ou não linear) entre as variáveis.

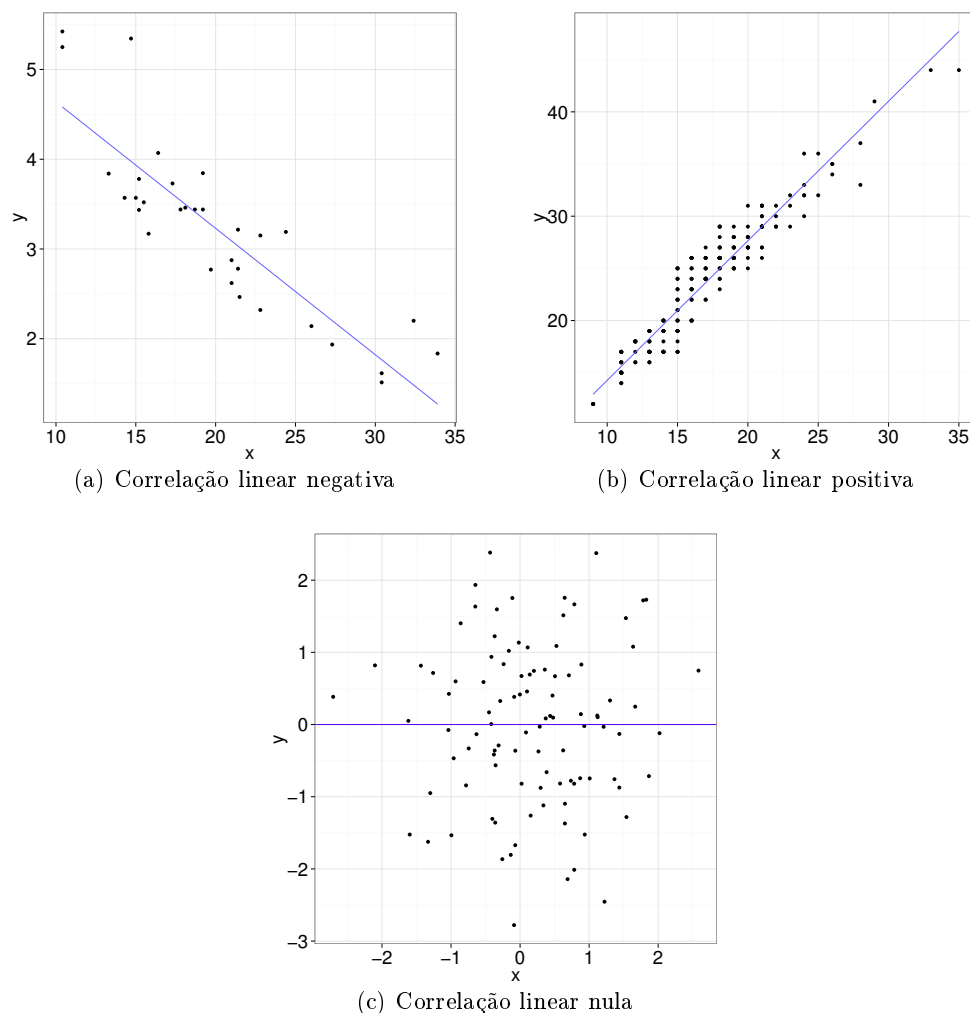


Figura 1.1: Gráfico com os possíveis tipos de correlação linear entre duas variáveis.

Nosso trabalho tem por objetivo introduzir uma nova abordagem para a construção das redes de relevância através da classe de coeficientes de correlação generalizado proposto por Chinchilli *et al.* (2005). Atualmente na literatura os pesquisadores utilizam o coeficiente de correlação de Pearson para a construção destas redes interagentes, porém uma vez que a normalidade dos dados é suspeitada, os pesquisadores recorrem ao coeficiente de correlação de Spearman. Assim apresentamos novas redes interagentes produzidas com a classe de coeficientes que com a ajuda de um especialista na área Biológica esta pode ser uma ótima alternativa na construção das redes de relevância.

1.2 Organização da Dissertação

Esta dissertação está organizada em 5 Capítulos e 2 Apêndices. No Capítulo 2, apresentamos os coeficientes de correlação de Pearson ρ_P , Kendall ρ_K e Spearman ρ_S bem como o coeficiente de correlação generalizado ρ_γ proposto por Chinchilli *et al.* (2005) o qual será objeto de nosso estudo. Aqui também apresentamos uma fórmula explícita para o coeficiente de correlação generalizado sob normalidade bivariada proposta por Chen *et al.* (2011). No Capítulo 3 analisamos o comportamento de vários estimadores para o coeficiente de correlação aqui proposto sob diferentes cenários de simulação. No Capítulo 4 utilizamos uma base de dados real com dados de expressão gênica e construímos redes de relevância através do coeficiente de correlação generalizado e do coeficiente de correlação de Spearman. Por fim, no Capítulo 5 apresentamos as conclusões obtidas neste trabalho. No Apêndice A apresentamos alguns aspectos preliminares em que introduzimos o conceito da função de densidade de probabilidade e distribuição acumulada da normal bivariada. Definimos também a mistura de distribuições, em especial a mistura de normais. Seguidamente, discutimos o método Monte Carlo o qual é importante para avaliar o comportamento de estimadores. Em seguida, no Apêndice B, apresentamos o conceito de estatística U mostrando algumas das suas propriedades, como por exemplo a sua distribuição assintótica.

1.3 Suporte Computacional

Todas as análises e programas de simulação foram realizadas com o **software** estatístico R na sua versão 3.0 que está disponível gratuitamente em <http://www.r-project.org>. Além disso vários gráficos no decorrer do trabalho foram confeccionados com este **software**. O seu amplo uso no meio acadêmico é uma grande vantagem pois contribui com uma imensa variedade de pacotes desenvolvidos em diversas áreas da Estatística.

Alguns gráficos também foram confeccionados com o **software** computacional Mathematica 8. Apesar de ser um **software** pago, possui um grande leque de funções gráficas e de cálculos algébricos de grande atrativo. Para maiores detalhes <http://www.wolfram.com/mathematica/>.

Para a execução das simulações utilizamos um **notebook** com processador Intel Core i3 primeira geração, 4 GB de memória RAM e sistema operacional Ubuntu 12.10 Quantal Quetzal.

Capítulo 2

Coeficiente de Correlação

2.1 Introdução

O coeficiente de correlação é uma medida que descreve a associação entre duas variáveis. Na literatura estatística, existem vários coeficientes de correlação e dentre estes o mais utilizado é o coeficiente de correlação de Pearson. Tal medida assume seus valores no intervalo fechado de -1 a 1 , em que -1 indica uma forte correlação negativa (linear) e 1 indica uma forte correlação positiva (linear). Rodgers e Nicewander (1988) apresentam uma série de propriedades bem como uma breve história sobre o desenvolvimento do coeficiente de correlação de Pearson. Nos casos em que a relação entre as variáveis seja não linear (quadrática, cúbica, exponencial, etc.) ou na presença de valores discrepantes, esta correlação pode não ser uma medida adequada para verificar a associação das variáveis. Porém, para este tipo de situação, o coeficiente de correlação de Spearman é mais adequado pois é mais robusto à presença de **outliers** (valores atípicos). Sua interpretação difere um pouco do coeficiente de correlação de Pearson sendo considerado como um índice de monotonicidade isto é, na medida que uma variável aumenta a associação com a outra variável também aumenta. Um fato importante deste coeficiente é que ele não é considerado como estimativa de um parâmetro populacional (Hollander e Wolfe, 2002). Por esta razão o coeficiente de correlação de Kendall torna-se uma ótima alternativa ao coeficiente de Spearman.

2.2 Definições

Sejam $(X_1, Y_1), \dots, (X_n, Y_n)$ uma amostra de variáveis aleatórias independentes e identicamente distribuídas de uma função de distribuição acumulada bivariada F_{XY} , e sejam F_X e F_Y as funções de distribuições marginais de X e Y , respectivamente.

Denotamos μ_X , μ_Y , σ_X^2 , σ_Y^2 e σ_{XY} como sendo a média de X , a média de Y , a variância de X , a variância de Y e a covariância entre X e Y , respectivamente. Seja também Σ a matriz de variância e covariância de (X, Y) representada por

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}.$$

O coeficiente de correlação de Pearson ρ_P é dado por

$$\rho_P = \frac{\sigma_{XY}}{\sqrt{\sigma_X\sigma_Y}} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

e pode ser escrito de forma equivalente como

$$\rho_P = \frac{E[(X_i - X_j)(Y_i - Y_j)]}{\sqrt{E[(X_i - X_j)^2]E[(Y_i - Y_j)^2]}},$$

em que (X_i, Y_i) e (X_j, Y_j) , $i \neq j = 1, 2, \dots, n$, são quaisquer dois vetores bivariados aleatórios independentes e identicamente distribuídos com função de distribuição bivariada F_{XY} .

De fato, note que

$$\begin{aligned} E[(X_i - X_j)(Y_i - Y_j)] &= E(X_i Y_i) - E(X_i Y_j) - E(X_j Y_i) + E(X_j Y_j) \\ &= 2E(X_i Y_i) - E(X_i)E(Y_j) - E(X_j)E(Y_i) \\ &= 2E(X_i Y_i) - 2E(X_i)E(Y_i) \\ &= 2\sigma_{XY}. \end{aligned}$$

De forma similar podemos calcular,

$$E(X_i - X_j)^2 = 2\sigma_{XX}, \quad E(Y_i - Y_j)^2 = 2\sigma_{YY},$$

logo

$$\rho_P = \frac{2\sigma_{XY}}{\sqrt{2\sigma_{XX}2\sigma_{YY}}} = \frac{\sigma_{XY}}{\sqrt{\sigma_{XX}\sigma_{YY}}}.$$

O coeficiente de correlação de Spearman (Spearman, 1904) é definido como o coeficiente de correlação de Pearson entre os postos das observações de X e Y , ou seja

$$\rho_S = \frac{E[\text{sign}(X_i - X_j)\text{sign}(Y_i - Y_j)]}{\sqrt{E[\text{sign}^2(X_i - X_j)]E[\text{sign}^2(Y_i - Y_j)]}},$$

em que $\text{sign}(\cdot)$ representa a função sinal, definida por -1 se $x < 0$, 0 se $x = 0$ e 1 se $x > 0$. Uma outra versão deste coeficiente pode ser definida como

$$\rho_S = 1 - \frac{6D^2}{n(n^2 - 1)},$$

em que $D^2 = \sum_{i=1}^n [\text{sign}(x_i) - \text{sign}(y_i)]^2$ é a soma dos postos dos valores observados de X e Y . É importante lembrar que para variáveis cuja mensuração é em nível ordinal e para amostras de tamanho pequeno o coeficiente de correlação de Spearman é uma ótima alternativa (Chok, 2010). Outra característica importante deste coeficiente é que a sua interpretação difere um pouco da interpretação do coeficiente de correlação de Pearson. Aqui a associação das variáveis não se reflete numa tendência linear mas um índice de monotonicidade, ou seja, para aumentos positivos da correlação, aumentos no valor de X correspondem a aumentos no valor de Y , e para coeficientes negativos ocorre o contrário.

O coeficiente de correlação de Kendall ρ_K (Kendall, 1938) pode ser definido como

$$\rho_K = \frac{C - D}{n(n-1)/2}$$

em que

$$\begin{aligned} C &= \sum_i \sum_j I(X_i > X_j)I(Y_i > Y_j) + I(X_i < X_j)I(Y_i < Y_j) \\ D &= \sum_i \sum_j I(X_i > X_j)I(Y_i < Y_j) + I(X_i < X_j)I(Y_i > Y_j) \end{aligned}$$

e $I(\cdot)$ representa a função indicadora definida por

$$I(t) = \begin{cases} 1, & t > 0 \\ 0, & t \leq 0. \end{cases}$$

Tanto o coeficiente de correlação de Kendall quanto o coeficiente de correlação de Spearman são utilizados para dados ordinais, com isto, ambas as variáveis devem ser medidas no mínimo em nível ordinal, de forma que seja possível atribuir postos a cada uma das variáveis.

Os coeficientes de correlação de Pearson, Kendall e Spearman são coeficientes de correlação usados frequentemente na literatura estatística. O coeficiente de correlação de Pearson é apropriado para amostras que seguem distribuição normal bivariada, porém se esta normalidade bivariada é suspeita os pesquisadores recorrem para os coeficientes de Kendall e Spearman que são na verdade dois coeficientes robustos. Neste contexto, estes dois estimadores são versões não paramétricas que apresentam maior robustez a **outliers**. Mesmo não sendo estimativa de um parâmetro populacional, o coeficiente de correlação de Spearman ainda atrai muitos pesquisadores por ser robusto a **outliers**. Como foi discutido anteriormente, o coeficiente de Kendall vem como uma alternativa ao coeficiente de correlação de Pearson quando suspeita-se da normalidade bivariada. Ainda segundo Raveh e Leshem (2011) um

fato importante sobre o coeficiente de correlação de Spearman é que em certos cenários de simulação este coeficiente tende a subestimar o verdadeiro valor do coeficiente de correlação.

2.2.1 Estimador de Máxima Verossimilhança do Coeficiente de Correlação Sob Normalidade Bivariada

Seja $(X_1, Y_1), \dots, (X_n, Y_n)$ uma amostra de uma distribuição normal bivariada e seja $\theta = (\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)^\top$ um vetor de parâmetros. Balakrishnan e Lai (2009) mostraram que as estimativas de máxima verossimilhança da distribuição normal bivariada (ver Apêndice A) são dadas por

$$\hat{\mu}_X = \frac{1}{n} \sum_{i=1}^n x_i, \quad \hat{\sigma}_X^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu}_X)^2,$$

$$\hat{\mu}_Y = \frac{1}{n} \sum_{i=1}^n y_i, \quad \hat{\sigma}_Y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\mu}_Y)^2,$$

$$\hat{\rho} = \frac{\sum_{i=1}^n (x_i - \hat{\mu}_X)(y_i - \hat{\mu}_Y)}{n \hat{\sigma}_X \hat{\sigma}_Y}.$$

Note que, no caso de normalidade bivariada o estimador de máxima verossimilhança de ρ coincide com o coeficiente de correlação amostral de Pearson r

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}. \quad (2.1)$$

A distribuição do coeficiente de correlação amostral pode ser escrita de várias formas, das quais destacamos as propostas por Fisher (1915) e Hotelling (1953). No livro de Johnson *et al.* (1995) existem outras formas desta distribuição em que os autores (veja Capítulo 32) mostram diferentes formas explícitas da distribuição do coeficiente de correlação e detalham como foram definidas.

Fisher (1915) definiu a densidade do coeficiente de correlação amostral (2.1) de duas formas. Na primeira forma ele expressou a densidade em termos de uma soma infinita, ou seja

$$\begin{aligned} f(r; \rho) &= \frac{2^{n-3} (1 - \rho^2)^{(1/2)(n-1)} (1 - r^2)^{(1/2)(n-4)}}{(n-3)! \pi} \\ &\quad \times \sum_{j=0}^{\infty} \frac{(2\rho r)^j}{j!} \left(\Gamma \left[\frac{1}{2}(n-1+j) \right] \right)^2, \end{aligned} \quad (2.2)$$

em que ρ é o coeficiente populacional e $\Gamma(\cdot)$ representa a função gama. A segunda forma proposta por Fisher (1915) foi expressa em termos das $(n - 2)$ derivadas de (2.2)

$$f(r; \rho) = \frac{2^{n-3}(1 - \rho^2)^{(1/2)(n-1)}(1 - r^2)^{(1/2)(n-4)}}{(n - 3)!\pi} \left[\frac{\partial^{n-2}}{\partial x^{n-2}} \left\{ \frac{\cos^{-1}(-x)}{\sqrt{1 - x^2}} \right\} \Big|_{x=r\rho} \right].$$

Para Hotelling (1953) a densidade do coeficiente de correlação r pode ser expressa da seguinte forma

$$\begin{aligned} f(r; \rho) &= \frac{n - 2}{\sqrt{2\pi}} \frac{\Gamma(n - 1)}{\Gamma(n - \frac{1}{2})} (1 - \rho^2)^{(1/2)(n-1)} (1 - r^2)^{(1/2)(n-4)} (1 - \rho r)^{-n+3/2} \times \\ &\times G\left(\frac{1}{2}, \frac{1}{2}; n - \frac{1}{2}, \frac{1 + \rho r}{2}\right), \end{aligned} \quad (2.3)$$

em que

$$G(a, b; c, x) = \sum_{j=0}^{\infty} \frac{\Gamma(a + j)}{\Gamma(a)} \frac{\Gamma(b + j)}{\Gamma(b)} \frac{\Gamma(c)}{\Gamma(c + j)} \frac{x^j}{j!}$$

é a função de distribuição acumulada da distribuição hipergeométrica.

2.3 Coeficiente de Correlação Generalizado

Segundo Chinchilli *et al.* (2005), a principal motivação para a definição do coeficiente de correlação generalizado foi o uso deste coeficiente dentro do contexto de um experimento **crossover** (cruzado). No experimento **crossover**, podemos testar um certo tipo de medicamento, cada indivíduo recebe mais de uma formulação de um mesma substância farmacológica em períodos diferentes. Por exemplo, o experimento **crossover** 2×2 é empregado quando dois medicamentos são administrados em dois períodos diferentes (Chellini, 2007). Os participantes deste tipo de experimento são alocados aleatoriamente para receber o medicamento a ser testado (T) e depois de um período recebem o medicamento de referência (R) ou vice-versa. Assim, é de grande interesse para os pesquisadores medir a correlação entre as repostas em diferentes períodos afim de avaliar a efetividade do tratamento. Porém, o cálculo das estimativas da correlação sem levar em consideração alguns efeitos, como os causados pelo tipo de período utilizado ou os efeitos da sequência utilizada no experimento, podem levar a uma análise estatística inadequada. Para resolver esse problema, Chinchilli *et al.* (2005) propuseram uma classe de coeficientes de correlação generalizados que contêm as estatísticas de Kendall e Pearson como casos particulares e também desenvolveram um método para estimar estes coeficientes

dentro do contexto de experimentos **crossover**. Uma discussão mais apropriada sobre este coeficiente de correlação pode ser vista em Chinchilli *et al.* (2005) e Chen *et al.* (2011).

Seja F_{XY} a função de distribuição acumulada bivariada de $\mathbf{X} = (X, Y)$. Chinchilli *et al.* (2005) propuseram um coeficiente de correlação similar aos coeficientes ρ_P e ρ_K da forma

$$\rho_\gamma = \frac{E[g_\gamma(X_i - X_j)g_\gamma(Y_i - Y_j)]}{\sqrt{E[g_\gamma^2(X_i - X_j)]E[g_\gamma^2(Y_i - Y_j)]}} = \frac{\theta_1}{\sqrt{\theta_2\theta_3}}, \quad (2.4)$$

em que $g_\gamma(z)$ é a função a valor real

$$g_\gamma(z) = \text{sign}(z)|z|^\gamma, \quad (2.5)$$

para $\gamma \in [0, 1]$ e $\theta_1, \theta_2, \theta_3$ são parâmetros populacionais. Aqui (X_i, Y_i) e (X_j, Y_j) com $i \neq j = 1, 2, \dots, n$ são dois vetores aleatórios bivariados independentes seguindo a mesma função de distribuição bivariada F_{XY} . Note que para $\gamma = 1$ na Equação (2.4), o coeficiente de correlação generalizado é o coeficiente de correlação de Pearson ρ_P . Quando $\gamma = 0$ obtemos o coeficiente de correlação de Kendall ρ_K . Com isto temos que os coeficientes de Kendall e de Pearson são os limites inferior e superior da classe de coeficientes de correlação generalizados pois temos que $\gamma \in [0, 1]$.

Resultado 2.3.1. *O coeficiente de correlação populacional de Pearson é equivalente ao coeficiente de correlação generalizado quando $\gamma = 1$. Se tomamos $\gamma = 0$ temos o coeficiente de correlação de Kendall.*

Demonstração. Se $\gamma = 1$, então temos que $g_\gamma(z) = z$ e

$$\rho_{\gamma=1} = \frac{E[(X_i - X_j)(Y_i - Y_j)]}{\sqrt{E[(X_i - X_j)^2]E[(Y_i - Y_j)^2]}}.$$

Vimos na Seção 2.2 que o coeficiente de correlação populacional de Pearson é dado por

$$\rho_P = \frac{\sigma_{XY}}{\sqrt{\sigma_{XX}\sigma_{YY}}} = \frac{E[(X_i - X_j)(Y_i - Y_j)]}{\sqrt{E[(X_i - X_j)^2]E[(Y_i - Y_j)^2]}},$$

em que $\sigma_{XY}, \sigma_{XX}, \sigma_{YY}$ são a covariância e variância das variáveis aleatórias X e Y , respectivamente. Portanto, ρ_P e $\rho_{\gamma=1}$ são equivalentes.

Por outro lado, o coeficiente de correlação de Kendall é definido como

$$\begin{aligned} \rho_K &= \text{Prob}(\text{pares concordantes}) - \text{Prob}(\text{pares discordantes}) \\ &= [P(X_i > X_j, Y_i > Y_j) + P(X_i < X_j, Y_i < Y_j)] \\ &\quad - [P(X_i > X_j, Y_i < Y_j) + P(X_i < X_j, Y_i > Y_j)], \end{aligned}$$

em que (X_i, Y_i) e (X_j, Y_j) são dois vetores bivariados. Agora, se $\gamma = 0$ temos que $g_\gamma(z) = \text{sign}(z)$ e, desta forma,

$$\rho_{\gamma=0} = \frac{E[\text{sign}(X_i - X_j)\text{sign}(Y_i - Y_j)]}{\sqrt{E[\text{sign}^2(X_i - X_j)]E[\text{sign}^2(Y_i - Y_j)]}}.$$

Podemos agora calcular

$$\begin{aligned} E[\text{sign}(X_i - X_j)\text{sign}(Y_i - Y_j)] &= (1) \times P[\text{sign}(X_i - X_j)\text{sign}(Y_i - Y_j) = 1] \\ &\quad + (-1) \times P[\text{sign}(X_i - X_j)\text{sign}(Y_i - Y_j) = -1] \\ &\quad + 0 \times P[\text{sign}(X_i - X_j)\text{sign}(Y_i - Y_j) = 0] \\ &= P[\text{sign}(X_i - X_j)\text{sign}(Y_i - Y_j) = 1] \\ &\quad - P[\text{sign}(X_i - X_j)\text{sign}(Y_i - Y_j) = -1] \\ &= [P(X_i > X_j, Y_i > Y_j) + P(X_i < X_j, Y_i < Y_j)] \\ &\quad - [P(X_i > X_j, Y_i < Y_j) + P(X_i < X_j, Y_i > Y_j)]. \end{aligned}$$

Já para

$$\begin{aligned} E[\text{sign}^2(X_i - X_j)] &= (1)^2 \times P(X_i > X_j) + (-1)^2 \times P(X_i < X_j) \\ &\quad + 0 \times P(X_i = X_j) \\ &= P(X_i > X_j) + P(X_i < X_j) \\ &= 1 \end{aligned}$$

pois consideramos o caso das variáveis aleatórias serem contínuas. De modo similar temos que $E[\text{sign}^2(Y_i - Y_j)] = 1$, então

$$\begin{aligned} \rho_{\gamma=0} &= \frac{E[\text{sign}(X_i - X_j)\text{sign}(Y_i - Y_j)]}{\sqrt{E[\text{sign}^2(X_i - X_j)]E[\text{sign}^2(Y_i - Y_j)]}} \\ &= E[\text{sign}(X_i - X_j)\text{sign}(Y_i - Y_j)] \\ &= [P(X_i > X_j, Y_i > Y_j) + P(X_i < X_j, Y_i < Y_j)] \\ &\quad - [P(X_i > X_j, Y_i < Y_j) + P(X_i < X_j, Y_i > Y_j)] \\ &= \rho_K. \end{aligned}$$

Deste modo temos que $\rho_K = \rho_{\gamma=0}$.

2.3.1 Estimação do Coeficiente de Correlação Generalizado Usando Estatísticas U

Nesta seção apresentamos a estimação do coeficiente de correlação generalizado proposto por Chinchilli *et al.* (2005). Supomos que uma amostra bivariada de tamanho

n , (X_i, Y_i) , $i = 1, 2, \dots, n$ é selecionada independentemente de F_{XY} , ou seja,

$$(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n) \quad \text{i.i.d.} \quad \sim F_{XY}.$$

Definimos três funções núcleos diferentes de grau 2,

$$h_1\{(X_1, Y_1), (X_2, Y_2)\} = g_\gamma(X_1 - X_2)g_\gamma(Y_1 - Y_2),$$

$$h_2\{(X_1, Y_1), (X_2, Y_2)\} = g_\gamma^2(X_1 - X_2),$$

$$h_3\{(X_1, Y_1), (X_2, Y_2)\} = g_\gamma^2(Y_1 - Y_2).$$

As três estatísticas U para estes núcleos são, respectivamente

$$\begin{aligned} U_{\gamma, XY} &= \frac{1}{\binom{n}{2}} \sum_{i < j} h_1\{(X_i, Y_i), (X_j, Y_j)\} \\ &= \frac{2}{n(n-1)} \sum_{i < j} g_\gamma(X_i - X_j)g_\gamma(Y_i - Y_j), \end{aligned} \quad (2.6)$$

$$\begin{aligned} U_{\gamma, XX} &= \frac{1}{\binom{n}{2}} \sum_{i < j} h_2\{(X_i, Y_i), (X_j, Y_j)\} \\ &= \frac{2}{n(n-1)} \sum_{i < j} g_\gamma^2(X_i - X_j), \end{aligned} \quad (2.7)$$

$$\begin{aligned} U_{\gamma, YY} &= \frac{1}{\binom{n}{2}} \sum_{i < j} h_3\{(X_i, Y_i), (X_j, Y_j)\} \\ &= \frac{2}{n(n-1)} \sum_{i < j} g_\gamma^2(Y_i - Y_j), \end{aligned} \quad (2.8)$$

para uma amostra bivariada $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ de F_{XY} . Portanto, uma estimativa de ρ_γ baseada nas 3 estatísticas U é da forma

$$\tilde{\rho}_\gamma = \frac{U_{\gamma, XY}}{\sqrt{U_{\gamma, XX}U_{\gamma, YY}}}, \quad (2.9)$$

em que as estatísticas U são estimadores não-viesados tal que

$$E(U_{\gamma, XY}) = E(g_\gamma(X_i - X_j)g_\gamma(Y_i - Y_j)) = \theta_1$$

$$E(U_{\gamma, XX}) = E(g_\gamma^2(X_i - X_j)) = \theta_2$$

$$E(U_{\gamma, YY}) = E(g_\gamma^2(Y_i - Y_j)) = \theta_3.$$

Agora vamos determinar a distribuição assintótica de $\tilde{\rho}_\gamma$ e para isto é necessário derivar a distribuição assintótica conjunta das três estatísticas U definidas em (2.6),

(2.7) e (2.8). Primeiramente definimos os seguintes parâmetros:

$$\psi_{\gamma,XY1} = E[g_{\gamma}(X_i - X_j)g_{\gamma}(Y_i - Y_j)], \quad (2.10)$$

$$\psi_{\gamma,XY2} = E[g_{\gamma}(X_i - X_j)g_{\gamma}(Y_i - Y_j)g_{\gamma}(X_j - X_k)g_{\gamma}(Y_j - Y_k)], \quad (2.11)$$

$$\psi_{\gamma,XY3} = E[g_{\gamma}^2(X_i - X_j)g_{\gamma}^2(Y_i - Y_j)], \quad (2.12)$$

$$\psi_{\gamma,XY4} = E[g_{\gamma}(X_i - X_j)g_{\gamma}(Y_i - Y_j)g_{\gamma}^2(X_j - X_k)], \quad (2.13)$$

$$\psi_{\gamma,XY5} = E[g_{\gamma}^3(X_i - X_j)g_{\gamma}(Y_i - Y_j)], \quad (2.14)$$

$$\psi_{\gamma,XY6} = E[g_{\gamma}(X_i - X_j)g_{\gamma}(Y_i - Y_j)g_{\gamma}^2(Y_j - Y_k)], \quad (2.15)$$

$$\psi_{\gamma,XY7} = E[g_{\gamma}(X_i - X_j)g_{\gamma}^3(Y_i - Y_j)], \quad (2.16)$$

$$\psi_{\gamma,XY8} = E[g_{\gamma}^2(X_i - X_j)g_{\gamma}^2(Y_j - Y_k)], \quad (2.17)$$

$$\psi_{\gamma,XX1} = E[g_{\gamma}^2(X_i - X_j)], \quad (2.18)$$

$$\psi_{\gamma,XX2} = E[g_{\gamma}^2(X_i - X_j)g_{\gamma}^2(X_j - X_k)], \quad (2.19)$$

$$\psi_{\gamma,XX3} = E[g_{\gamma}^4(X_i - X_j)], \quad (2.20)$$

$$\psi_{\gamma,YY1} = E[g_{\gamma}^2(Y_i - Y_j)], \quad (2.21)$$

$$\psi_{\gamma,YY2} = E[g_{\gamma}^2(Y_i - Y_j)g_{\gamma}^2(Y_j - Y_k)], \quad (2.22)$$

$$\psi_{\gamma,YY3} = E[g_{\gamma}^4(Y_i - Y_j)]. \quad (2.23)$$

que podem ser estimados pelas correspondentes estatísticas U dadas por

$$\hat{\psi}_{\gamma,XY1} = \frac{2}{n(n-1)} \sum_{i < j} g_{\gamma}(X_i - X_j)g_{\gamma}(Y_i - Y_j) = U_{\gamma,XY}, \quad (2.24)$$

$$\begin{aligned} \hat{\psi}_{\gamma,XY2} &= \frac{6}{n(n-1)(n-2)} \sum_{i < j < k} g_{\gamma}(X_i - X_j)g_{\gamma}(Y_i - Y_j) \\ &\quad \times g_{\gamma}(X_j - X_k)g_{\gamma}(Y_j - Y_k), \end{aligned} \quad (2.25)$$

$$\hat{\psi}_{\gamma,XY3} = \frac{2}{n(n-1)} \sum_{i < j} g_{\gamma}^2(X_i - X_j)g_{\gamma}^2(Y_i - Y_j), \quad (2.26)$$

$$\begin{aligned} \hat{\psi}_{\gamma,XY4} &= \frac{6}{n(n-1)(n-2)} \sum_{i < j < k} g_{\gamma}(X_i - X_j) \\ &\quad \times g_{\gamma}(Y_i - Y_j)g_{\gamma}^2(X_j - X_k), \end{aligned} \quad (2.27)$$

$$\hat{\psi}_{\gamma,XY5} = \frac{2}{n(n-1)} \sum_{i < j} g_{\gamma}^3(X_i - X_j)g_{\gamma}(Y_i - Y_j), \quad (2.28)$$

$$\begin{aligned} \hat{\psi}_{\gamma,XY6} &= \frac{6}{n(n-1)(n-2)} \sum_{i < j < k} g_{\gamma}(X_i - X_j) \\ &\quad \times g_{\gamma}(Y_i - Y_j)g_{\gamma}^2(Y_j - Y_k), \end{aligned} \quad (2.29)$$

$$\hat{\psi}_{\gamma,XY7} = \frac{2}{n(n-1)} \sum_{i < j} g_{\gamma}(X_i - X_j) g_{\gamma}^3(Y_i - Y_j), \quad (2.30)$$

$$\hat{\psi}_{\gamma,XY8} = \frac{6}{n(n-1)(n-2)} \sum_{i < j < k} g_{\gamma}^2(X_i - X_j) g_{\gamma}^2(Y_j - Y_k), \quad (2.31)$$

$$\hat{\psi}_{\gamma,XX1} = \frac{2}{n(n-1)} \sum_{i < j} g_{\gamma}^2(X_i - X_j) = U_{\gamma,XX}, \quad (2.32)$$

$$\hat{\psi}_{\gamma,XX2} = \frac{6}{n(n-1)(n-2)} \sum_{i < j < k} [g_{\gamma}^2(X_i - X_j) g_{\gamma}^2(X_j - X_k)], \quad (2.33)$$

$$\hat{\psi}_{\gamma,XX3} = \frac{2}{n(n-1)} \sum_{i < j} g_{\gamma}^4(X_i - X_j), \quad (2.34)$$

$$\hat{\psi}_{\gamma,YY1} = \frac{2}{n(n-1)} \sum_{i < j} g_{\gamma}^2(Y_i - Y_j) = U_{\gamma,YY}, \quad (2.35)$$

$$\hat{\psi}_{\gamma,YY2} = \frac{6}{n(n-1)(n-2)} \sum_{i < j} g_{\gamma}^2(Y_i - Y_j) g_{\gamma}^2(Y_j - Y_k), \quad (2.36)$$

$$\hat{\psi}_{\gamma,YY3} = \frac{2}{n(n-1)} \sum_{i < j} g_{\gamma}^4(Y_i - Y_j). \quad (2.37)$$

Logo, temos o seguinte resultado.

Resultado 2.3.2. (*Chinchilli et al. (2005)*) *Dados os parâmetros das equações (2.10) à (2.23), então as três estatísticas U , $U_{\gamma,XX}$, $U_{\gamma,YY}$ e $U_{\gamma,XY}$, definidas em (2.6), (2.7) e (2.8), possuem a distribuição assintótica conjunta*

$$\left(\begin{bmatrix} U_{\gamma,XX} \\ U_{\gamma,YY} \\ U_{\gamma,XY} \end{bmatrix} - \begin{bmatrix} \psi_{\gamma,XX1} \\ \psi_{\gamma,YY1} \\ \psi_{\gamma,XY1} \end{bmatrix} \right) \xrightarrow{D} \mathcal{N}_3 \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \omega_{n,\gamma,XX,XX} & \omega_{n,\gamma,XX,YY} & \omega_{n,\gamma,XX,XY} \\ \omega_{n,\gamma,XX,YY} & \omega_{n,\gamma,YY,YY} & \omega_{n,\gamma,XY,YY} \\ \omega_{n,\gamma,XX,XY} & \omega_{n,\gamma,XY,YY} & \omega_{n,\gamma,XY,XY} \end{bmatrix} \right) \quad (2.38)$$

em que

$$\begin{aligned} \omega_{n,\gamma,XX,XX} &= \frac{2}{n(n-1)} [2(n-2)\psi_{\gamma,XX2} + \psi_{\gamma,XX3} - (2n-3)\psi_{\gamma,XX1}^2], \\ \omega_{n,\gamma,YY,YY} &= \frac{2}{n(n-1)} [2(n-2)\psi_{\gamma,YY2} + \psi_{\gamma,YY3} - (2n-3)\psi_{\gamma,YY1}^2], \\ \omega_{n,\gamma,XX,XY} &= \frac{2}{n(n-1)} [2(n-2)\psi_{\gamma,XY4} + \psi_{\gamma,XY5} - (2n-3)\psi_{\gamma,XY1}\psi_{\gamma,XX1}], \\ \omega_{n,\gamma,XX,YY} &= \frac{2}{n(n-1)} [2(n-2)\psi_{\gamma,XX8} + \psi_{\gamma,XX3} - (2n-3)\psi_{\gamma,XX1}\psi_{\gamma,YY1}], \\ \omega_{n,\gamma,XY,YY} &= \frac{2}{n(n-1)} [2(n-2)\psi_{\gamma,XX6} + \psi_{\gamma,XX7} - (2n-3)\psi_{\gamma,XY1}\psi_{\gamma,YY1}], \\ \omega_{n,\gamma,XY,XY} &= \frac{2}{n(n-1)} [2(n-2)\psi_{\gamma,XY2} + \psi_{\gamma,XY3} - (2n-3)\psi_{\gamma,XY1}^2]. \end{aligned}$$

Demonstração. Primeiramente calculamos

$$\begin{aligned} \sigma_{1,XY}^2 &= \text{Cov} \{h_1[(X_i, Y_i), (X_j, Y_j)], h_1[(X_j, Y_j), (X_k, Y_k)]\} \\ &= \text{Cov} \{g_{\gamma}(X_i - X_j)g_{\gamma}(Y_i - Y_j), g_{\gamma}(X_j - X_k)g_{\gamma}(Y_j - Y_k)\} \end{aligned}$$

$$\begin{aligned}
 &= E \{g_\gamma(X_i - X_j)g_\gamma(Y_i - Y_j)g_\gamma(X_j - X_k)g_\gamma(Y_j - Y_k)\} \\
 &- E \{g_\gamma(X_i - X_j)g_\gamma(Y_i - Y_j)\} E \{g_\gamma(X_j - X_k)g_\gamma(Y_j - Y_k)\} \\
 &= \psi_{\gamma,XY2} - (E \{g_\gamma(X_i - X_j)g_\gamma(Y_i - Y_j)\})^2 \\
 &= \psi_{\gamma,XY2} - \psi_{\gamma,XY1}^2
 \end{aligned}$$

e

$$\begin{aligned}
 \sigma_{2,XY}^2 &= \text{Cov} \{h_1[(X_i, Y_i), (X_j, Y_j)], h_1[(X_i, Y_i), (X_j, Y_j)]\} \\
 &= \text{Var} \{h_1[(X_i, Y_i), (X_j, Y_j)]\} \\
 &= E \{h_1[(X_i, Y_i), (X_j, Y_j)]\}^2 - (E \{h_1[(X_i, Y_i), (X_j, Y_j)]\})^2 \\
 &= E \{g_\gamma^2(X_i - X_j)g_\gamma^2(Y_i - Y_j)\} - (E \{g_\gamma(X_i - X_j)g_\gamma(Y_i - Y_j)\})^2 \\
 &= \psi_{\gamma,XY3} - \psi_{\gamma,XY1}^2.
 \end{aligned}$$

Usando o Teorema B.-19 no Apêndice B em que é definida a variância de uma estatística U , temos a variância de $U_{\gamma,XY}$ como sendo

$$\begin{aligned}
 \omega_{n,\gamma,XY,XY} &= \text{Var}(U_{\gamma,XY}) = \binom{n}{2}^{-1} \sum_{c=1}^2 \binom{2}{c} \binom{n-2}{2-c} \sigma_{c,XY}^2 \\
 &= \frac{2}{n(n-1)} [2(n-2)\sigma_{1,XY}^2 + \sigma_{2,XY}^2] \\
 &= \frac{2}{n(n-1)} [2(n-2)(\psi_{\gamma,XY2} - \psi_{\gamma,XY1}^2) + (\psi_{\gamma,XY3} - \psi_{\gamma,XY1}^2)] \\
 &= \frac{2}{n(n-1)} [2(n-2)\psi_{\gamma,XY2} + \psi_{\gamma,XY3} - (2n-3)\psi_{\gamma,XY1}^2].
 \end{aligned}$$

De maneira similar, as variâncias de $U_{\gamma,XX}$ e $U_{\gamma,YY}$ são definidas como

$$\begin{aligned}
 \omega_{n,\gamma,XX,XX} &= \frac{2}{n(n-1)} [2(n-2)\psi_{\gamma,XX2} + \psi_{\gamma,XX3} - (2n-3)\psi_{\gamma,XX1}^2] \\
 \omega_{n,\gamma,YY,YY} &= \frac{2}{n(n-1)} [2(n-2)\psi_{\gamma,YY2} + \psi_{\gamma,YY3} - (2n-3)\psi_{\gamma,YY1}^2]
 \end{aligned}$$

Definimos também

$$\begin{aligned}
 \sigma_{1,1,XYXX} &= \text{Cov} \{h_1[(X_i, Y_i), (X_j, Y_j)], h_2[(X_j, Y_j), (X_k, Y_k)]\} \\
 &= \text{Cov} \{g_\gamma(X_i - X_j)g_\gamma(Y_i - Y_j), g_\gamma^2(X_j - X_k)\} \\
 &= E \{g_\gamma(X_i - X_j)g_\gamma(Y_i - Y_j)g_\gamma^2(X_j - X_k)\} \\
 &- E \{(X_i - X_j)g_\gamma(Y_i - Y_j)\} E \{g_\gamma^2(X_j - X_k)\} \\
 &= \psi_{\gamma,XY4} - E \{(X_i - X_j)g_\gamma(Y_i - Y_j)\} E \{g_\gamma^2(X_j - X_k)\} \\
 &= \psi_{\gamma,XY4} - \psi_{\gamma,XY1}\psi_{\gamma,XX1}
 \end{aligned}$$

$$\begin{aligned}
 \sigma_{2,2,XYXX} &= \text{Cov} \{h_1[(X_i, Y_i), (X_j, Y_j)], h_2[(X_i, Y_i), (X_j, Y_j)]\} \\
 &= \text{Cov} \{g_\gamma(X_i - X_j)g_\gamma(Y_i - Y_j), g_\gamma^2(X_i - X_j)\} \\
 &= E \{g_\gamma(X_i - X_j)g_\gamma(Y_i - Y_j)g_\gamma^2(X_i - X_j)\} \\
 &\quad - E \{(X_i - X_j)g_\gamma(Y_i - Y_j)\} E \{g_\gamma^2(X_i - X_j)\} \\
 &= \psi_{\gamma,XY5} - \psi_{\gamma,XY1}\psi_{\gamma,XX1}.
 \end{aligned}$$

Agora, calculamos a covariância entre $U_{\gamma,XY}$ e $U_{\gamma,XX}$ como sendo

$$\begin{aligned}
 \omega_{n,\gamma,XY} &= \text{Cov}(U_{\gamma,XY}, U_{\gamma,XX}) = \binom{n}{2}^{-1} \sum_{c=1}^2 \binom{2}{c} \binom{n-2}{2-c} \sigma_{c,c,XYXX}^2 \\
 &= \frac{2}{n(n-1)} [2(n-2)\sigma_{1,1,XYXX}^2 + \sigma_{2,2,XYXX}^2] \\
 &= \frac{2}{n(n-1)} [2(n-2)(\psi_{\gamma,XY4} - \psi_{\gamma,XY1}\psi_{\gamma,XX1}) \\
 &\quad + (\psi_{\gamma,XY5} - \psi_{\gamma,XY1}\psi_{\gamma,XX1})] \\
 &= \frac{2}{n(n-1)} [2(n-2)\psi_{\gamma,XY4} + \psi_{\gamma,XY5} - (2n-3)\psi_{\gamma,XY1}\psi_{\gamma,XX1}].
 \end{aligned}$$

De maneira análoga temos as covariâncias entre $U_{\gamma,XY}$ e $U_{\gamma,YY}$, e a covariância entre $U_{\gamma,XX}$ e $U_{\gamma,YY}$, ou seja

$$\begin{aligned}
 \omega_{n,\gamma,XY,YY} &= \text{Cov}(U_{\gamma,XY}, U_{\gamma,YY}) \\
 &= \frac{2}{n(n-1)} [2(n-2)\psi_{\gamma,XX6} + \psi_{\gamma,XX7} - (2n-3)\psi_{\gamma,XY1}\psi_{\gamma,YY1}],
 \end{aligned}$$

$$\begin{aligned}
 \omega_{n,\gamma,XX,YY} &= \text{Cov}(U_{\gamma,XX}, U_{\gamma,YY}) \\
 &= \frac{2}{n(n-1)} [2(n-2)\psi_{\gamma,XX8} + \psi_{\gamma,XX3} - (2n-3)\psi_{\gamma,XX1}\psi_{\gamma,YY1}],
 \end{aligned}$$

respectivamente.

De acordo com os resultados anteriores, se todos os parâmetros ψ existem para a distribuição acumulada bivariada F_{XY} , então baseado no comportamento assintótico das estatísticas U temos que

$$\left(\begin{bmatrix} U_{\gamma,XX} \\ U_{\gamma,YY} \\ U_{\gamma,XY} \end{bmatrix} - \begin{bmatrix} \psi_{\gamma,XX1} \\ \psi_{\gamma,YY1} \\ \psi_{\gamma,XY1} \end{bmatrix} \right) \xrightarrow{\mathcal{D}} \mathcal{N}_3(\mathbf{0}, \Omega_{n,\gamma}),$$

em que

$$\Omega_{n,\gamma} = \begin{bmatrix} \omega_{n,\gamma,XX,XX} & \omega_{n,\gamma,XX,YY} & \omega_{n,\gamma,XX,XY} \\ \omega_{n,\gamma,XX,YY} & \omega_{n,\gamma,YY,YY} & \omega_{n,\gamma,XY,YY} \\ \omega_{n,\gamma,XY,YY} & \omega_{n,\gamma,XX,XY} & \omega_{n,\gamma,XY,XY} \end{bmatrix}.$$

Usando o método delta multivariado (Oehlert, 1992) e as equações (2.5) e (2.38), chegamos a

$$\tilde{\rho}_\gamma \sim N(\rho_\gamma, \sigma_{n,\gamma}^2 = D' \Omega_{n,\gamma} D) \quad (2.39)$$

em que

$$D' = \left[\frac{\partial \rho_\gamma}{\partial \psi_{\gamma,XY1}} \frac{\partial \rho_\gamma}{\partial \psi_{\gamma,XX1}} \frac{\partial \rho_\gamma}{\partial \psi_{\gamma,YY1}} \right] = \left[\frac{\rho_\gamma}{\psi_{\gamma,XY1}} \frac{-\rho_\gamma}{\psi_{\gamma,XX1}} \frac{-\rho_\gamma}{\psi_{\gamma,YY1}} \right], \quad (2.40)$$

no qual temos que a distribuição assintótica do coeficiente de correlação generalizado $\tilde{\rho}_\gamma$ é uma normal com média ρ_γ e variância $\sigma_{n,\gamma}^2$ em que

$$\sigma_{n,\gamma}^2 = \frac{2\rho_\gamma^2}{n(n-1)} [2(n-2)\lambda_1 + \lambda_2], \quad (2.41)$$

sendo

$$\begin{aligned} \lambda_1 &= \frac{\psi_{\gamma,XY2}}{\psi_{\gamma,XY1}\psi_{\gamma,XY1}} + \frac{\psi_{\gamma,XX2}}{4\psi_{\gamma,XX1}\psi_{\gamma,XX1}} + \frac{\psi_{\gamma,YY2}}{4\psi_{\gamma,YY1}\psi_{\gamma,YY1}} - \frac{\psi_{\gamma,XY4}}{\psi_{\gamma,XY1}\psi_{\gamma,XX1}} \\ &\quad - \frac{\psi_{\gamma,XY6}}{\psi_{\gamma,XY1}\psi_{\gamma,YY1}} + \frac{\psi_{\gamma,XY8}}{2\psi_{\gamma,XX1}\psi_{\gamma,YY1}}, \\ \lambda_2 &= \frac{\psi_{\gamma,XY3}}{\psi_{\gamma,XY1}\psi_{\gamma,XY1}} + \frac{\psi_{\gamma,XX3}}{4\psi_{\gamma,XX1}\psi_{\gamma,XX1}} + \frac{\psi_{\gamma,YY3}}{4\psi_{\gamma,YY1}\psi_{\gamma,YY1}} - \frac{\psi_{\gamma,XY5}}{\psi_{\gamma,XY1}\psi_{\gamma,XX1}} \\ &\quad - \frac{\psi_{\gamma,XY7}}{\psi_{\gamma,XY1}\psi_{\gamma,YY1}} + \frac{\psi_{\gamma,XY3}}{2\psi_{\gamma,XX1}\psi_{\gamma,YY1}}. \end{aligned}$$

A transformação Z de Fisher é normalmente aplicada a estatísticas de correlação amostral a fim de melhorar a taxa de convergência para normalidade. Neste caso, temos que

$$Z_\gamma = \frac{1}{2} \log \left(\frac{1 + \tilde{\rho}_\gamma}{1 - \tilde{\rho}_\gamma} \right) \sim \mathcal{N}(\varsigma_\gamma, \sigma_{n,Z_\gamma}^2) \quad (2.42)$$

em que

$$\varsigma_\gamma = \frac{1}{2} \log \left(\frac{1 + \rho_\gamma}{1 - \rho_\gamma} \right), \quad \sigma_{n,Z_\gamma}^2 = \frac{1}{(1 - \rho_\gamma^2)^2} \sigma_{n,\gamma}^2.$$

Usando as equações (2.24) a (2.37), podemos estimar λ_1 e λ_2 e, assim, estimamos a variância assintótica por

$$\hat{\sigma}_{n,\gamma}^2 = \frac{2\tilde{\rho}_\gamma^2}{n(n-1)} [2(n-2)\hat{\lambda}_1 + \hat{\lambda}_2].$$

Definida a forma de estimar a variância assintótica, podemos então definir os limites de um intervalo de confiança assintótico para ρ_γ como

$$\tilde{\rho}_\gamma \pm z_{1-\frac{\alpha}{2}} \sqrt{\hat{\sigma}_{n,\gamma}^2},$$

em que $z_{1-\frac{\alpha}{2}}$ é o quantil $1 - \frac{\alpha}{2}$ da distribuição normal padrão.

A fim de estudar algumas propriedades do coeficiente de correlação generalizado, Chen *et al.* (2011) definiram uma fórmula explícita para este coeficiente sob normalidade bivariada.

Resultado 2.3.3. *Dadas as equações (2.4) e (2.5) e sob as condições definidas na Seção 2.2 para uma amostra da distribuição normal bivariada, temos que*

$$\rho_\gamma = 2\pi^{-1/2} \frac{[\Gamma(\frac{1}{2}\gamma + 1)]^2}{\Gamma(\gamma + \frac{1}{2})} \rho(1 - \rho^2)^{\gamma + \frac{1}{2}} {}_2F_1(\frac{1}{2}\gamma + 1, \frac{1}{2}\gamma + 1; \frac{3}{2}; \rho^2) \quad (2.43)$$

$$= 2\pi^{-1/2} \frac{[\Gamma(\frac{1}{2}\gamma + 1)]^2}{\Gamma(\gamma + \frac{1}{2})} \rho {}_2F_1(\frac{1}{2}(1 - \gamma), \frac{1}{2}(1 - \gamma); \frac{3}{2}; \rho^2), \quad (2.44)$$

em que $\Gamma(\cdot)$ é a função gama e ${}_2F_1$ é a função hipergeométrica definida por

$${}_2F_1(a, b; c; x) = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k}{(c)_k k!} x^k$$

para $|x| < 1$ e os coeficientes definidos por

$$(a)_k = a(a+1)(a+2) \dots (a+k-1) = \frac{\Gamma(a+k)}{\Gamma(a)}, \quad (b)_k = \frac{\Gamma(b+k)}{\Gamma(b)} \quad e \quad (c)_k = \frac{\Gamma(c+k)}{\Gamma(c)}.$$

Demonstração. Omitida. Para mais detalhes, veja Chen (2006).

As expressões (2.43) e (2.44) permitem relacionar de forma explícita o coeficiente de correlação ρ (parâmetro da distribuição normal bivariada) e o coeficiente de correlação generalizado. Usando as propriedades de monotonicidade de ρ_γ (ver Chen *et al.*, 2011) se $\rho \geq 0$ então ρ_γ é uma função estritamente crescente em γ .

Pelo princípio da invariância e pelo Resultado (2.3.3) podemos estimar o coeficiente de correlação generalizado ρ_γ por máxima verossimilhança. De fato, se $\hat{\rho}$ é o estimador de máxima verossimilhança de ρ então para γ fixado

$$\hat{\rho}_\gamma = 2\pi^{-\frac{1}{2}} \frac{[\Gamma(\frac{1}{2}\gamma)]^2}{\Gamma(\gamma + \frac{1}{2})} \hat{\rho} {}_2F_1(\frac{1}{2}(1 - \gamma), \frac{1}{2}(1 - \gamma); \frac{3}{2}; \hat{\rho}^2)$$

é o estimador de máxima verossimilhança de ρ_γ .

Este resultado é verificado uma vez que a função (2.44) é estritamente crescente para $\rho \in [-1, 1]$ e γ fixado como vemos na Figura 2.1.

Note que nessas condições, podemos construir um intervalo de confiança assintótico para ρ_γ baseado no estimador de máxima verossimilhança de ρ .

De fato, sob normalidade bivariada da amostra, temos que assintoticamente

$$\sqrt{n}(\hat{\rho} - \rho) \rightarrow \mathcal{N}(0, (1 - \rho^2)^2).$$

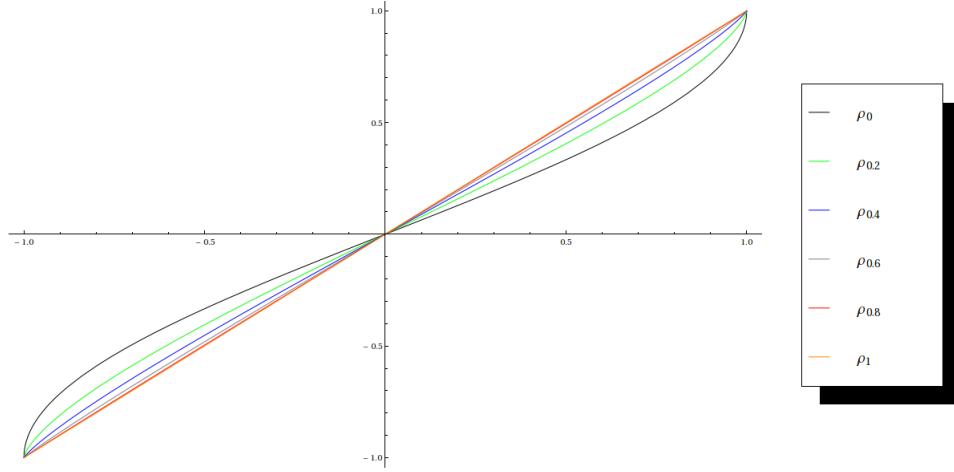


Figura 2.1: Gráfico para a forma explícita do coeficiente de correlação generalizado com diferentes valores de γ .

Assim, e considerando o método delta para γ fixado, temos que

$$\sqrt{n}(\hat{\rho}_\gamma - \rho_\gamma) \rightarrow \mathcal{N}(0, (\rho'_\gamma)^2(1 - \rho^2)^2).$$

Na prática, ρ não é conhecido e por isso usamos o seu estimador de máxima verossimilhança. Assim

$$\hat{\rho}_\gamma \pm z_{1-\frac{\alpha}{2}} \sqrt{\rho'_\gamma(\hat{\rho})(1 - \hat{\rho}^2)^2}$$

são os limites do intervalo de confiança assintótico para ρ_γ .

Adicionalmente podemos considerar uma versão robusta do estimador do coeficiente de correlação generalizado ρ_γ , a qual é baseada no estimador de Spearman, isto é, definimos o estimador $\bar{\rho}_{S_\gamma} = \hat{\rho}_\gamma(\rho_S)$ da expressão (2.44) da mesma forma que construímos o estimador do coeficiente de correlação generalizado por máxima verossimilhança.

Note também que, usando a distribuição assintótica de $\hat{\rho}_\gamma$ e $\bar{\rho}_{S_\gamma}$ podemos conduzir testes assintóticos para o coeficiente de correlação generalizado ρ_γ . De fato, considere

$$\begin{aligned} H_0 : \quad & \rho_\gamma = \rho_\gamma^* \\ H_1 : \quad & \rho_\gamma \neq \rho_\gamma^*. \end{aligned} \tag{2.45}$$

Assim, rejeitamos a hipótese nula se

$$|\hat{\rho}_\gamma - \rho_\gamma^*| > z_{1-\frac{\alpha}{2}} \sqrt{\rho'_\gamma(\hat{\rho})(1 - \hat{\rho}^2)^2}$$

(usando máxima verossimilhança)

ou

$$|\bar{\rho}_\gamma - \rho_\gamma^*| > z_{1-\frac{\alpha}{2}} \sqrt{\rho'_\gamma(\bar{\rho}_S)(1 - \bar{\rho}_S^2)^2}$$

(usando correlação de Spearman)

ou

$$|\tilde{\rho}_\gamma - \rho_\gamma^*| > z_{1-\frac{\alpha}{2}} \sqrt{\hat{\sigma}_{n,\gamma}^2}$$

(usando estatísticas U).

Afim de tornar viável a comparação dos coeficientes analisamos todos em função da Equação (2.43). Note que

$$\tilde{\rho}_\gamma \xrightarrow{p} \rho_\gamma, \quad \hat{\rho} \xrightarrow{p} \rho, \quad \text{e} \quad \rho_S \xrightarrow{p} \rho$$

em que $\xrightarrow{p}$ representa convergência em probabilidade.

Utilizando a inversa da Equação (2.43) no intervalo $(-1, 1)$ podemos comparar os estimadores em função de ρ , uma vez que estamos analisando os estimadores em função de ρ_γ .

Capítulo 3

Estudo de Simulação

3.1 Introdução

Neste capítulo apresentamos os cenários e os resultados das simulações em que analisamos o comportamento do estimador do coeficiente de correlação generalizado baseado em máxima verossimilhança (CCGMV) $\hat{\rho}_\gamma$, do coeficiente de correlação generalizado baseado em estatísticas U (CCGU) $\tilde{\rho}_\gamma$ proposto por Chinchilli *et al.* (2005) e do coeficiente de correlação generalizado baseado no coeficiente de Spearman (CCGS) $\bar{\rho}_{S_\gamma}$.

3.2 Cenários de Simulação

3.2.1 Amostra Sem Contaminação

Neste cenário abordamos as simulações baseadas no uso da distribuição normal bivariada sem contaminação. Contudo, definimos 4 casos distintos em que estudamos o comportamento dos estimadores definidos no Capítulo 2.

Caso 1

Uma distribuição normal bivariada

$$\mathcal{N}_2(0, 0; 1, 1; \rho)$$

com ρ igual a 0, 0.3 e 0.9 representando exemplos de variáveis aleatórias com nenhuma, moderada e alta correlação, respectivamente.

Caso 2

Uma distribuição normal bivariada

$$\mathcal{N}_2(-\sqrt{0.1}/2, \sqrt{0.1}/2; 1, 1; \rho)$$

com ρ igual a 0, 0.3 e 0.9 representando exemplos de variáveis aleatórias com nenhuma, moderada e alta correlação, respectivamente. Neste caso temos uma pequena alteração na média da distribuição normal bivariada.

Caso 3

Uma distribuição normal bivariada

$$\mathcal{N}_2(-0.5, 0.5; 1, 1; \rho)$$

com ρ igual a 0, 0.3 e 0.9 representando exemplos de variáveis aleatórias com nenhuma, moderada e alta correlação, respectivamente. Neste caso temos uma maior alteração na média da distribuição normal bivariada.

Caso 4

Uma distribuição normal bivariada

$$\mathcal{N}_2(\sqrt{0.25}/2, \sqrt{0.25}/2; (4/3)^2, (3/2)^2; \rho)$$

com ρ igual a 0, 0.3 e 0.9 representando exemplos de variáveis aleatórias com nenhuma, moderada e alta correlação, respectivamente.

Na Figura 3.1 apresentamos o gráfico da função de densidade e suas curvas de níveis para todos os casos sem contaminação da amostra com $\rho = 0.3$.

Se ρ for próximo de 1 representa distribuições com alta correlação entre as variáveis.

3.2.2 Amostra Perturbada

Nessa abordagem consideramos a distribuição amostrada perturbada, em que utilizamos valores da distribuição t -Student bivariada para perturbar a amostra. Para isso, utilizamos o seguinte algoritmo:

Entrada: n (tamanho da amostra), $\varepsilon \geq 0$ (nível de perturbação).

Para $i = 1, \dots, n$ faça:

(1) Gere

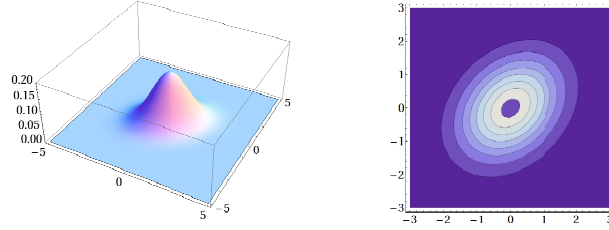
$$X_{1i} \sim \mathcal{N}_2(E(\underline{X}), \underline{\Sigma})$$

em que $E(\underline{X}), \underline{\Sigma}$ são os mesmos que definimos no Apêndice B;

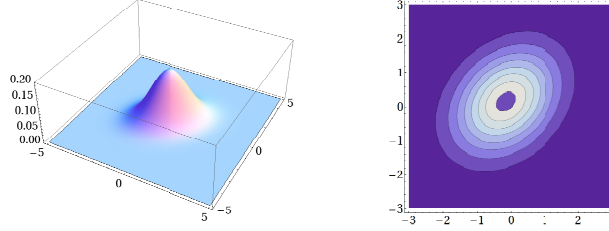
Gere

$$X_{2i} \sim t_2(\underline{\Sigma}, 3)$$

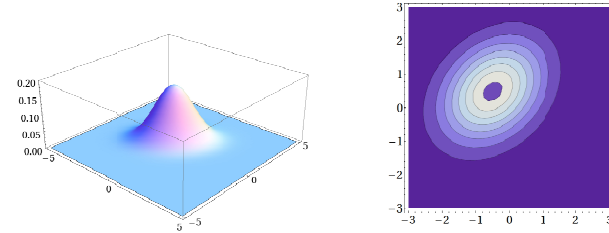
Caso 1



Caso 2



Caso 3



Caso 4

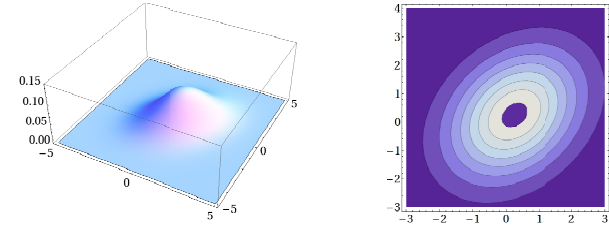


Figura 3.1: Gráfico da função densidade com as respectivas curvas de níveis da distribuição normal bivariada para os quatro casos com amostras sem perturbação e com $\rho = 0.3$.

em que Σ é a mesma matriz de variância e covariância definida para gerar números aleatórios normais bivariados e 3 denota os graus de liberdade da distribuição *t*-Student bivariada como definimos no Apêndice A;

Gere

$$U_i \sim \mathcal{U}(0, 1)$$

em que $\mathcal{U}$ representa a distribuição Uniforme no intervalo (0,1).

(2) Calcule

$$\underline{X}_i = \begin{cases} \underline{X}_{1i} & \text{se } U_i > \varepsilon, \\ \underline{X}_{2i} & \text{se } U_i \leq \varepsilon. \end{cases}$$

(3) Retorne $(\underline{X}_1, \dots, \underline{X}_n)$ (amostra contaminada).

Para geração dos números aleatórios normais bivariados e da distribuição t -Student bivariada utilizamos **software** estatístico **R** através do pacote **mvtnorm**. Este pacote fornece duas funções que geram estes números aleatórios: **rmvnorm** (geram números normais bivariados) e **rmvt** (geram números t -Student bivariados). Para maiores informações veja <http://cran.r-project.org/web/packages/mvtnorm/mvtnorm.pdf>.

No estudo de simulação perturbamos a amostra aos níveis de 10%, 30% e 50%.

3.2.3 Mistura de Duas Distribuições Normais Bivariadas

Considerando a Definição A.3 assumimos quatro casos para o estudo de simulação.

Caso 1

Neste primeiro caso avaliamos o comportamento dos estimadores onde a distribuição da mistura é da forma

$$Q_{\underline{X}} = 0.9 \times \mathcal{N}_2(0, 0; 1, 1; 0.9) + 0.1 \times \mathcal{N}_2(0, 0; 1, 1; 0.5)$$

em que a distribuição de referência é contaminada com uma distribuição com correlação moderada.

Caso 2

Neste cenário consideramos para a análise dos estimadores a mistura de distribuições

$$Q_{\underline{X}} = 0.9 \times \mathcal{N}_2(0, 0; 1, 1; 0.9) + 0.1 \times \mathcal{N}_2(0, 0; 1, 1; 0)$$

em que a distribuição de referência é contaminada com uma distribuição sem correlação.

Caso 3

Agora, temos o seguinte cenário para a distribuição da mistura

$$Q_{\underline{X}} = 0.9 \times \mathcal{N}_2(0, 0; 1, 1; 0.9) + 0.1 \times \mathcal{N}_2(0, 0; 1, 1; -0.9)$$

em que a distribuição de referência é contaminada com uma distribuição com alta correlação negativa.

Caso 4

Neste cenário temos a seguinte distribuição para a mistura

$$Q_{\underline{X}} = 0.6 \times \mathcal{N}_2(0, 0; 1, 1; 0.9) + 0.4 \times \mathcal{N}_2(0, 0; 1, 1; 0)$$

em que a distribuição de referência é contaminada com uma distribuição sem correlação com um nível de 40% de contaminação.

Caso 5

No último caso a distribuição da amostra é dada por

$$Q_{\underline{X}} = 0.2 \times \mathcal{N}_2(0, 0; 1, 1; 0.9) + 0.8 \times \mathcal{N}_2(0, 0; 1, 1; 0)$$

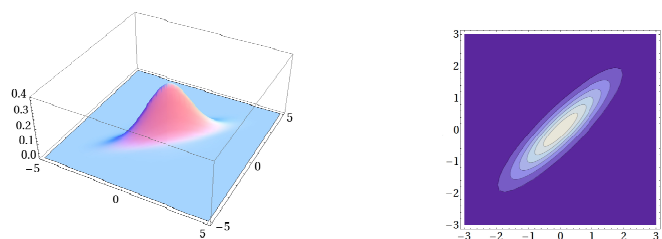
em que a distribuição de referência é contaminada com uma distribuição sem correlação com um nível de 80% de contaminação.

Na Figura 3.2 apresentamos a função de densidade e as curvas de nível para os 5 casos da mistura de normais com $\rho = 0.3$.

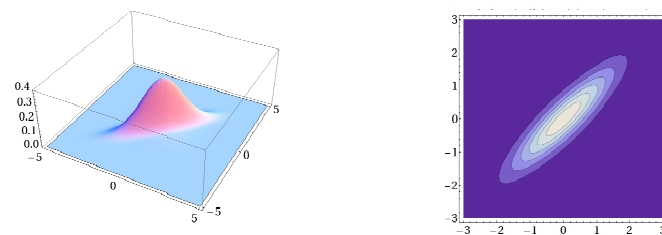
3.3 Resultado das simulações

Nesta seção apresentamos o estudo de simulações realizadas para os cenários definidos na Seção 3.2. Para as análises estatísticas fixamos em 5000 o número de réplicas de Monte Carlo. Nos casos em que temos a amostra sem contaminação e sem perturbação definimos os valores de γ em 0, 0.5 e 1. Utilizamos estes mesmos valores de γ para os casos em que temos a distribuição amostrada perturbada com valores da distribuição t -Student bivariada. Para os casos em que temos misturas de normais bivaridas definimos os valores de γ em 0, 0.35, 0.65 e 1. Os resultados de simulação são apresentados através de figuras onde se encontram gráficos boxplot de cada estimador sob os diferentes cenários de simulação. Nos gráficos boxplot observamos o comportamento distribucional e a tendência de convergência em todos os casos definidos. Apresentamos também tabelas com a raiz quadrada do erro quadrático médio para as estimativas dos três estimadores. Todos os `scripts` utilizados nestas simulações estão disponíveis para download em www.de.ufpe.br/~raydonal/materialsuplementar/Cleber.html.

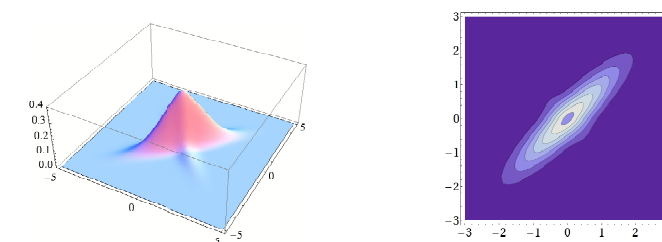
Caso 1



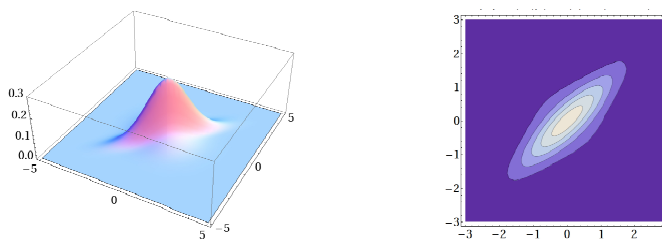
Caso 2



Caso 3



Caso 4



Caso 5

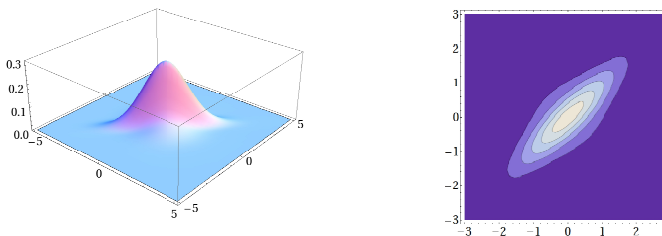


Figura 3.2: Gráfico da função densidade com as respectivas curvas de níveis da distribuição normal bivariada para os cinco casos de misturas de normais com $\rho = 0.3$.

3.3.1 Análise dos estimadores sem perturbação da amostra

Caso 1

Neste primeiro caso podemos observar pela Figura 3.3 que quando temos $\rho = 0$, $\gamma = 0$ e amostras de tamanho igual a 10 o coeficiente de correlação generalizado baseado em estatísticas U (CCGU) tende a subestimar o verdadeiro valor ρ_γ . Quando aumentamos o tamanho da amostra todos os estimadores tendem a diminuir suas respectivas variabilidades além de convergirem para o verdadeiro valor ρ_γ . Este comportamento permanece para os valores de γ , 0.5 e 1.

Na Figura 3.4 notamos que quando temos $\rho = 0.3$ (correlação moderada), $\gamma = 0$ e amostras de tamanho 10 o coeficiente de correlação generalizado (CCGU) superestima o verdadeiro valor ρ_γ . Para o coeficiente de correlação baseado em máxima verossimilhança (CCGMV) este comportamento é ainda mais evidente. Porém, a medida que aumentamos o tamanho da amostra ambos convergem para o verdadeiro valor ρ_γ . Ao mudar o valor de γ para 0.5 o comportamento destes estimadores permanecem o mesmo. Quanto temos $\gamma = 1$ o coeficiente de correlação generalizado (CCGU) tende superestimar o verdadeiro valor ρ_γ de uma maneira mais leve que o coeficiente baseado em máxima verossimilhança (CCGMV). O coeficiente de correlação generalizado baseado no coeficiente de Spearman (CCGS) tende a subestimar o verdadeiro valor ρ_γ para todos os valores de γ e todos os tamanhos da amostra.

Para $\rho = 0.9$ (forte correlação), na Figura 3.4, observamos que para todos os valores de γ o CCGU tende a superestimar o verdadeiro valor ρ_γ em amostras de tamanho 10. Além disso, quando aumentamos o valor de γ o seu comportamento distribucional torna-se ainda mais assimétrico. Observando o CCGMV podemos notar que suas estimativas tendem a superestimar menos do que o coeficiente de correlação generalizado além de possuir um comportamento menos variável. Quando aumentamos o tamanho da amostra ambos tendem a convergir para o verdadeiro valor ρ_γ mas CCGMV possui um comportamento menos variável. Para todos os valores de γ e nos diferentes tamanhos da amostra notamos também que o CCGS tende a subestimar o verdadeiro valor ρ_γ .

Na Tabela 3.1 apresentamos a raiz quadrada do erro quadrático médio para as estimativas dos três estimadores e observamos que na maioria dos casos o coeficiente CCGMV é o que apresenta os menores valores, isto é ele é melhor. Além disso podemos observar que, à medida que aumentamos o tamanho da amostra, os três estimadores diminuem drasticamente os seus valores da raiz do erro quadrático

médio como podemos perceber a partir das amostras de tamanho 50.

$$\rho = 0$$

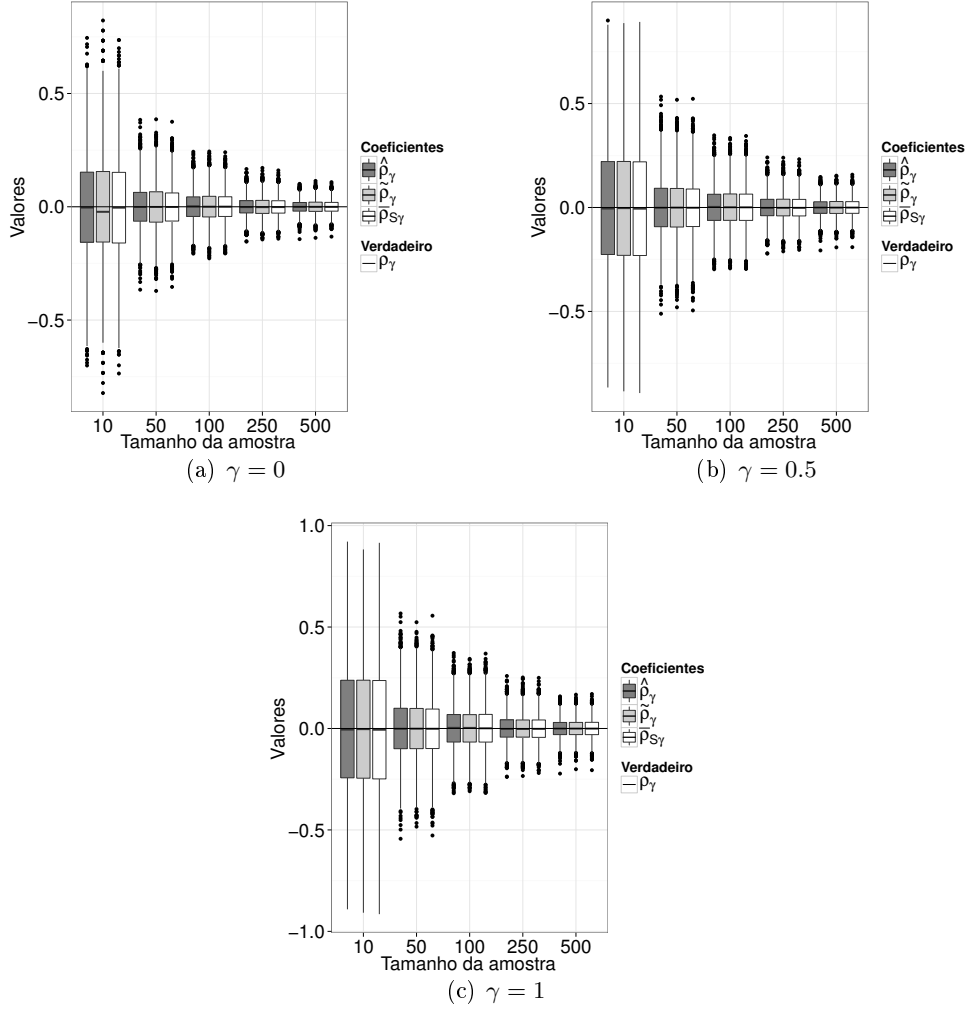


Figura 3.3: Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ para o primeiro caso das amostras normais bivariadas sem perturbação e considerando $\rho = 0$.

Caso 2

Pelos gráficos da Figura 3.5 observamos um comportamento bastante semelhante ao primeiro caso. O CCGU subestima o verdadeiro valor de ρ_γ quando temos $\gamma = 0$ e amostras de tamanho 10. Os estimadores CCGMV, CCGU e CCGS apresentam um comportamento bastante parecido tanto com relação a convergência quanto a variabilidade. Os três estimadores tendem a convergir para o verdadeiro valor ρ_γ a partir da amostra de tamanho 50, além de apresentarem um comportamento menos variável a medida que aumentamos o tamanho da amostra, o que era de se esperar.

Tabela 3.1: Raiz do erro quadrático médio das estimativas do coeficiente de correlação generalizado $\hat{\rho}_\gamma$, $\tilde{\rho}_\gamma$ e $\bar{\rho}_{S_\gamma}$ para o primeiro caso sem perturbação da amostra.

$\rho = 0$						
γ	Método	Tamanho da Amostra				
		10	50	100	250	500
0	$\hat{\rho}_\gamma$	0.2287	0.0953	0.0651	0.0411	0.0286
	$\tilde{\rho}_\gamma$	0.2480	0.0981	0.0676	0.0428	0.0298
	$\bar{\rho}_{S_\gamma}$	0.2246	0.0925	0.0642	0.0407	0.0284
0.5	$\hat{\rho}_\gamma$	0.3182	0.1376	0.0944	0.0597	0.0416
	$\tilde{\rho}_\gamma$	0.3154	0.1352	0.0941	0.0600	0.0418
	$\bar{\rho}_{S_\gamma}$	0.3129	0.1336	0.0931	0.0592	0.0414
1	$\hat{\rho}_\gamma$	0.3385	0.1480	0.1017	0.0644	0.0448
	$\tilde{\rho}_\gamma$	0.3315	0.1439	0.1003	0.0640	0.0446
	$\bar{\rho}_{S_\gamma}$	0.3330	0.1437	0.1003	0.0639	0.0446
$\rho = 0.3$						
0	$\hat{\rho}_\gamma$	0.2146	0.0859	0.0587	0.0375	0.0260
	$\tilde{\rho}_\gamma$	0.2402	0.0944	0.0649	0.0411	0.0286
	$\bar{\rho}_{S_\gamma}$	0.2187	0.0905	0.0628	0.0406	0.0289
0.5	$\hat{\rho}_\gamma$	0.2937	0.1200	0.0822	0.0526	0.0365
	$\tilde{\rho}_\gamma$	0.2969	0.1253	0.0869	0.0554	0.0386
	$\bar{\rho}_{S_\gamma}$	0.2993	0.1271	0.0884	0.0572	0.0407
1	$\hat{\rho}_\gamma$	0.3111	0.1278	0.0876	0.0561	0.0389
	$\tilde{\rho}_\gamma$	0.3097	0.1315	0.0912	0.0582	0.0406
	$\bar{\rho}_{S_\gamma}$	0.3171	0.1355	0.0943	0.0611	0.0434
$\rho = 0.9$						
0	$\hat{\rho}_\gamma$	0.0651	0.0295	0.0206	0.0133	0.0092
	$\tilde{\rho}_\gamma$	0.1374	0.0468	0.0314	0.0196	0.0135
	$\bar{\rho}_{S_\gamma}$	0.1413	0.0541	0.0379	0.0257	0.0198
0.5	$\hat{\rho}_\gamma$	0.0500	0.0227	0.0158	0.0102	0.0071
	$\tilde{\rho}_\gamma$	0.0943	0.0331	0.0225	0.0143	0.0099
	$\bar{\rho}_{S_\gamma}$	0.1352	0.0455	0.0310	0.0207	0.0158
1	$\hat{\rho}_\gamma$	0.0448	0.0202	0.0141	0.0091	0.0063
	$\tilde{\rho}_\gamma$	0.0846	0.0287	0.0193	0.0122	0.0085
	$\bar{\rho}_{S_\gamma}$	0.1296	0.0416	0.0281	0.0187	0.0142

Na Figura 3.6 temos que para $\rho = 0.3$ os estimadores CCGMV e CCGU tendem a superestimar o verdadeiro valor ρ_γ em amostras de tamanho 10. O CCGMV é o que apresenta a maior superestimação mas com uma variabilidade menor do que comparado ao CCGU. À medida que aumentamos o tamanho da amostra estes dois estimadores apresentam comportamento distribucional semelhante e tendem a convergir para o verdadeiro valor ρ_γ . Novamente, para todos os valores de γ e todos os tamanhos da amostra, CCGS tende a subestimar o verdadeiro valor ρ_γ .

Para $\rho = 0.9$ a tendência de superestimação tanto do CCGMV quanto do CCGU continuam para os valores de γ e quando temos amostras de tamanho 10. Porém o comportamento distribucional do CCGU torna-se mais assimétrico mas com maior variabilidade em comparação ao CCGMV. A medida que aumentamos o tamanho da amostra esta tendência assimétrica do CCGU atenua-se, porém ainda possui uma maior variabilidade do que o CCGMV. O CCGS permanece com a tendência de subestimar o verdadeiro valor ρ_γ mesmo, por exemplo, para amostras de tamanho 500 (grandes amostras).

Na Tabela 3.2 apresentamos a raiz quadrada do erro quadrático médio e observamos que mesmo fazendo uma pequena alteração no vetor de médias da distribuição normal bivariada, não há influência desta nos valores do erro quadrático médio. Assim, permanecem idênticos ao primeiro caso estudado e com os mesmos comportamentos.

Caso 3

Neste terceiro caso podemos observar nos gráficos da Figura 3.7 um comportamento semelhante aos dois casos anteriores para $\rho = 0$. Quando $\gamma = 0$ e o tamanho da amostra igual a 10 temos que o CCGU subestima o verdadeiro valor ρ_γ . Ao aumentarmos o tamanho da amostra os três estimadores apresentam um comportamento semelhante, porém o CCGMV apresenta menor variabilidade quando temos o tamanho da amostra igual a 500. Para os demais valores de γ e tamanhos da amostra os três estimadores apresentam uma variabilidade bastante semelhante.

Nos gráficos da Figura 3.8 observamos uma superestimação do verdadeiro valor ρ_γ tanto do CCGU bem como do CCGMV quando temos amostras de tamanho 10 e $\rho = 0.3$. Ao aumentarmos o valor de γ para 1 o CCGU tende a superestimar um pouco o verdadeiro valor ρ_γ . Quando aumentamos o tamanho da amostra percebemos que o comportamento do CCGU e do CCGMV são bastante semelhantes levando em consideração a variabilidade destes estimadores.

Analisando o comportamento dos estimadores quando $\rho = 0.9$ percebemos que o CCGU tende a superestimar o verdadeiro valor ρ_γ quando temos novamente amostras de tamanho 10. Neste mesmo cenário ao aumentarmos o valor de γ o comportamento distribucional deste estimador tende a se tornar mais assimétrico. Porém quando aumentamos o tamanho da amostra esta tendência diminui e percebemos também que o estimador CCGMV é o que possui a menor variabilidade entre os estimadores. O CCGS continua a subestimar o verdadeiro valor ρ_γ em todos os

Tabela 3.2: Raiz do erro quadrático médio das estimativas do coeficiente de correlação generalizado $\hat{\rho}_\gamma$, $\tilde{\rho}_\gamma$ e $\bar{\rho}_{S_\gamma}$ para o segundo caso sem perturbação da amostra.

$\rho = 0$						
γ	Método	Tamanho da Amostra				
		10	50	100	250	500
0	$\hat{\rho}_\gamma$	0.2287	0.0953	0.0651	0.0411	0.0286
	$\tilde{\rho}_\gamma$	0.2480	0.0981	0.0676	0.0428	0.0298
	$\bar{\rho}_{S_\gamma}$	0.2246	0.0925	0.0642	0.0407	0.0284
0.5	$\hat{\rho}_\gamma$	0.3182	0.1376	0.0944	0.0597	0.0416
	$\tilde{\rho}_\gamma$	0.3154	0.1352	0.0941	0.0600	0.0418
	$\bar{\rho}_{S_\gamma}$	0.3129	0.1336	0.0931	0.0592	0.0414
1	$\hat{\rho}_\gamma$	0.3385	0.1480	0.1017	0.0644	0.0448
	$\tilde{\rho}_\gamma$	0.3315	0.1439	0.1003	0.0640	0.0446
	$\bar{\rho}_{S_\gamma}$	0.3330	0.1437	0.1003	0.0639	0.0446
$\rho = 0.3$						
0	$\hat{\rho}_\gamma$	0.2146	0.0859	0.0587	0.0375	0.0260
	$\tilde{\rho}_\gamma$	0.2402	0.0944	0.0649	0.0411	0.0286
	$\bar{\rho}_{S_\gamma}$	0.2187	0.0905	0.0628	0.0406	0.0289
0.5	$\hat{\rho}_\gamma$	0.2937	0.1200	0.0822	0.0526	0.0365
	$\tilde{\rho}_\gamma$	0.2969	0.1253	0.0869	0.0554	0.0386
	$\bar{\rho}_{S_\gamma}$	0.2993	0.1271	0.0884	0.0572	0.0407
1	$\hat{\rho}_\gamma$	0.3111	0.1278	0.0876	0.0561	0.0389
	$\tilde{\rho}_\gamma$	0.3097	0.1315	0.0912	0.0582	0.0406
	$\bar{\rho}_{S_\gamma}$	0.3171	0.1355	0.0943	0.0611	0.0434
$\rho = 0.9$						
0	$\hat{\rho}_\gamma$	0.0651	0.0295	0.0206	0.0133	0.0092
	$\tilde{\rho}_\gamma$	0.1374	0.0468	0.0314	0.0196	0.0135
	$\bar{\rho}_{S_\gamma}$	0.1413	0.0541	0.0379	0.0257	0.0198
0.5	$\hat{\rho}_\gamma$	0.0500	0.0227	0.0158	0.0102	0.0071
	$\tilde{\rho}_\gamma$	0.0943	0.0331	0.0225	0.0143	0.0099
	$\bar{\rho}_{S_\gamma}$	0.1352	0.0455	0.0310	0.0207	0.0158
1	$\hat{\rho}_\gamma$	0.0448	0.0202	0.0141	0.0091	0.0063
	$\tilde{\rho}_\gamma$	0.0846	0.0287	0.0193	0.0122	0.0085
	$\bar{\rho}_{S_\gamma}$	0.1296	0.0416	0.0281	0.0187	0.0142

valores de γ e tamanhos da amostra para ρ igual a 0.3 e 0.9.

Na Tabela 3.3 observamos o mesmo comportamento dos dois casos anteriores. O coeficiente de correlação generalizado por máxima verossimilhança é o que apresenta os menores valores da raiz do erro quadrático médio. Podemos observar também que ao aumentarmos o tamanho da amostra a variabilidade dos três estimadores diminui principalmente se compararmos os valores da raiz do erro quadrático médio quando temos a amostra de tamanho 10 e 500.

Tabela 3.3: Raiz do erro quadrático médio das estimativas do coeficiente de correlação generalizado $\hat{\rho}_\gamma$, $\tilde{\rho}_\gamma$ e $\bar{\rho}_{S_\gamma}$ para o terceiro caso sem perturbação da amostra.

$\rho = 0$						
γ	Método	Tamanho da Amostra				
		10	50	100	250	500
0	$\hat{\rho}_\gamma$	0.2287	0.0953	0.0651	0.0411	0.0286
	$\tilde{\rho}_\gamma$	0.2480	0.0981	0.0676	0.0428	0.0298
	$\bar{\rho}_{S_\gamma}$	0.2246	0.0925	0.0642	0.0407	0.0284
0.5	$\hat{\rho}_\gamma$	0.3182	0.1376	0.0944	0.0597	0.0416
	$\tilde{\rho}_\gamma$	0.3154	0.1352	0.0941	0.0600	0.0418
	$\bar{\rho}_{S_\gamma}$	0.3129	0.1336	0.0931	0.0592	0.0414
1	$\hat{\rho}_\gamma$	0.3385	0.1480	0.1017	0.0644	0.0448
	$\tilde{\rho}_\gamma$	0.3315	0.1439	0.1003	0.0640	0.0446
	$\bar{\rho}_{S_\gamma}$	0.3330	0.1437	0.1003	0.0639	0.0446
$\rho = 0.3$						
0	$\hat{\rho}_\gamma$	0.2146	0.0859	0.0587	0.0375	0.0260
	$\tilde{\rho}_\gamma$	0.2402	0.0944	0.0649	0.0411	0.0286
	$\bar{\rho}_{S_\gamma}$	0.2187	0.0905	0.0628	0.0406	0.0289
0.5	$\hat{\rho}_\gamma$	0.2937	0.1200	0.0822	0.0526	0.0365
	$\tilde{\rho}_\gamma$	0.2969	0.1253	0.0869	0.0554	0.0386
	$\bar{\rho}_{S_\gamma}$	0.2993	0.1271	0.0884	0.0572	0.0407
1	$\hat{\rho}_\gamma$	0.3111	0.1278	0.0876	0.0561	0.0389
	$\tilde{\rho}_\gamma$	0.3097	0.1315	0.0912	0.0582	0.0406
	$\bar{\rho}_{S_\gamma}$	0.3171	0.1355	0.0943	0.0611	0.0434
$\rho = 0.9$						
0	$\hat{\rho}_\gamma$	0.0651	0.0295	0.0206	0.0133	0.0092
	$\tilde{\rho}_\gamma$	0.1374	0.0468	0.0314	0.0196	0.0135
	$\bar{\rho}_{S_\gamma}$	0.1413	0.0541	0.0379	0.0257	0.0198
0.5	$\hat{\rho}_\gamma$	0.0500	0.0227	0.0158	0.0102	0.0071
	$\tilde{\rho}_\gamma$	0.0943	0.0331	0.0225	0.0143	0.0099
	$\bar{\rho}_{S_\gamma}$	0.1352	0.0455	0.0310	0.0207	0.0158
1	$\hat{\rho}_\gamma$	0.0448	0.0202	0.0141	0.0091	0.0063
	$\tilde{\rho}_\gamma$	0.0846	0.0287	0.0193	0.0122	0.0085
	$\bar{\rho}_{S_\gamma}$	0.1296	0.0416	0.0281	0.0187	0.0142

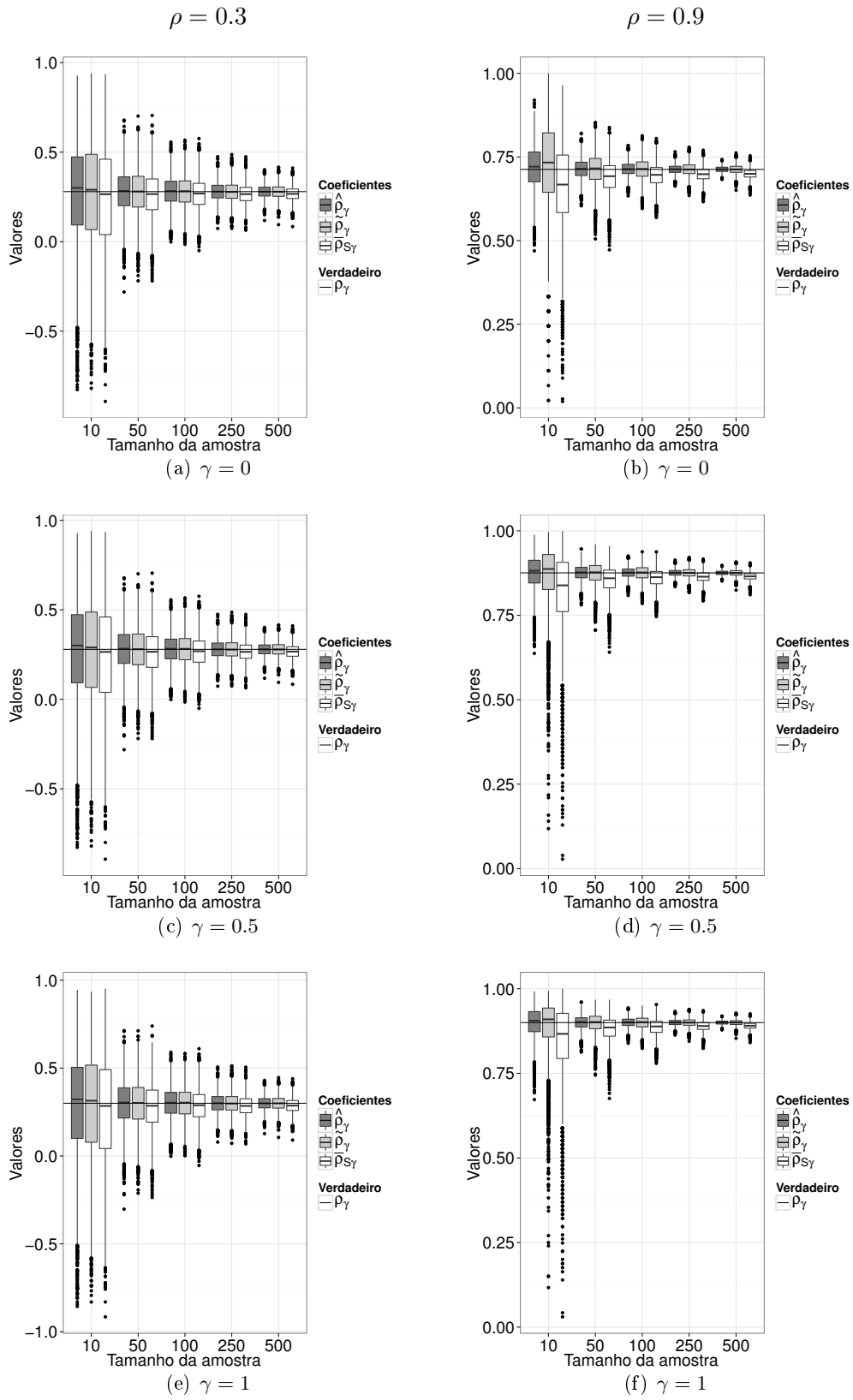


Figura 3.4: Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ para o primeiro caso das amostras normais bivariadas sem perturbação e considerando $\rho = 0.3$ e $\rho = 0.9$.

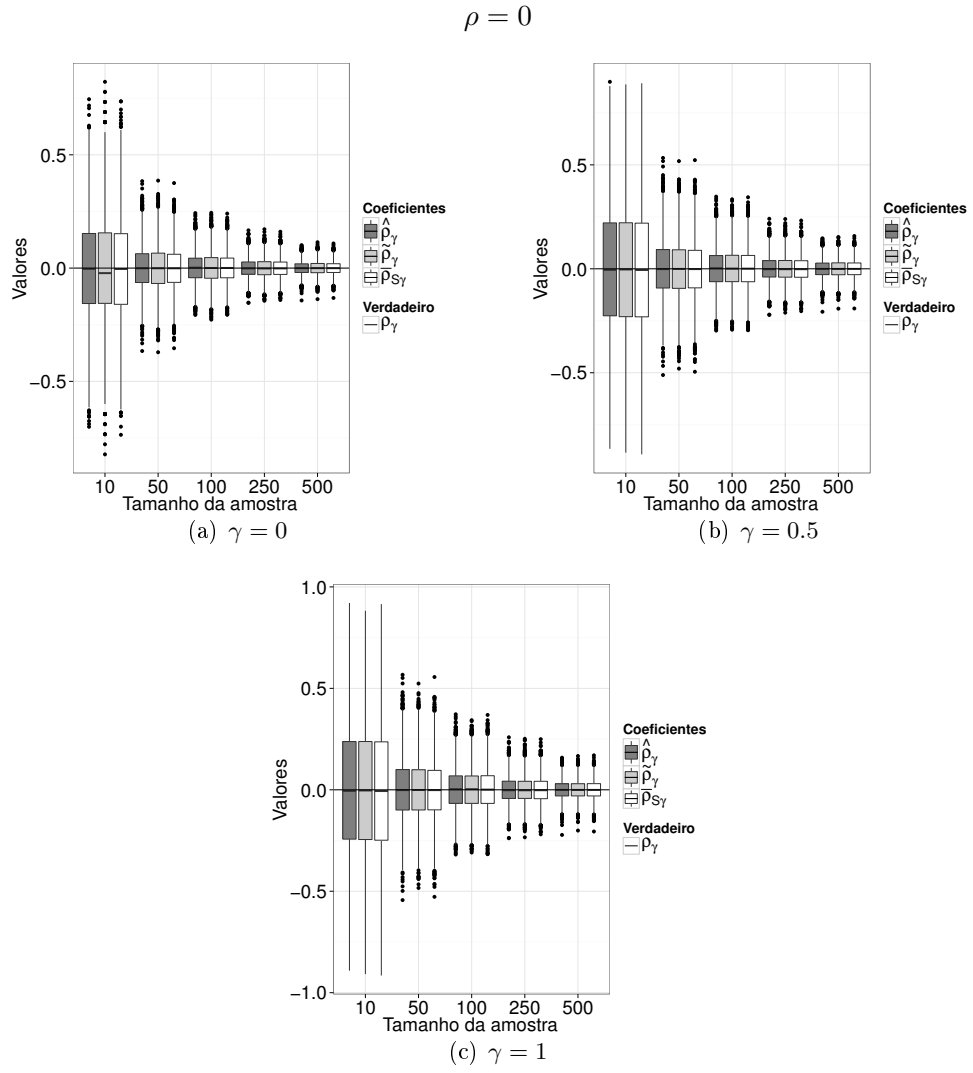


Figura 3.5: Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ para o segundo caso das amostras normais bivariadas sem perturbação e considerando $\rho = 0$.

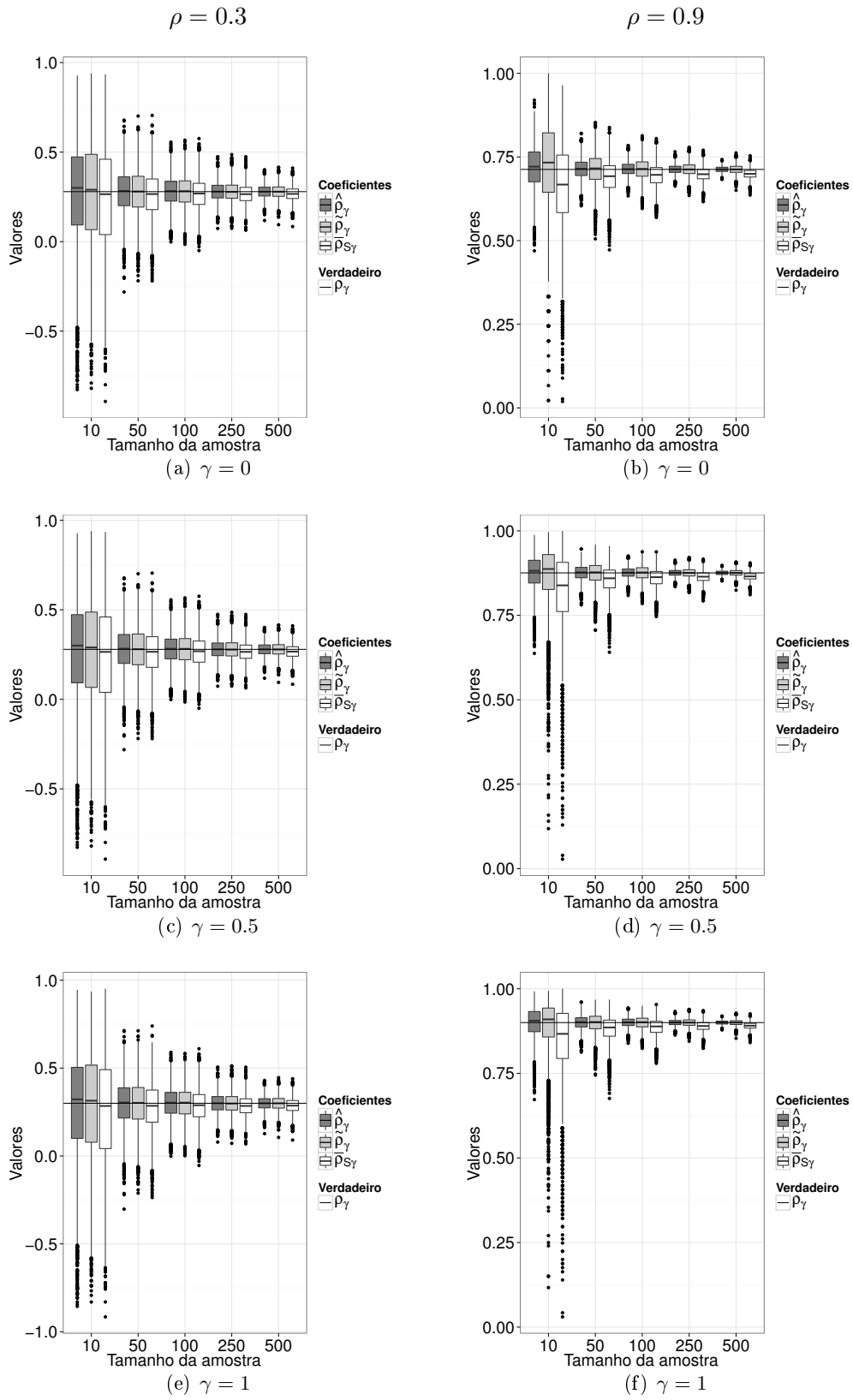


Figura 3.6: Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ para o segundo caso das amostras normais bivariadas sem perturbação e considerando $\rho = 0.3$ e $\rho = 0.9$.

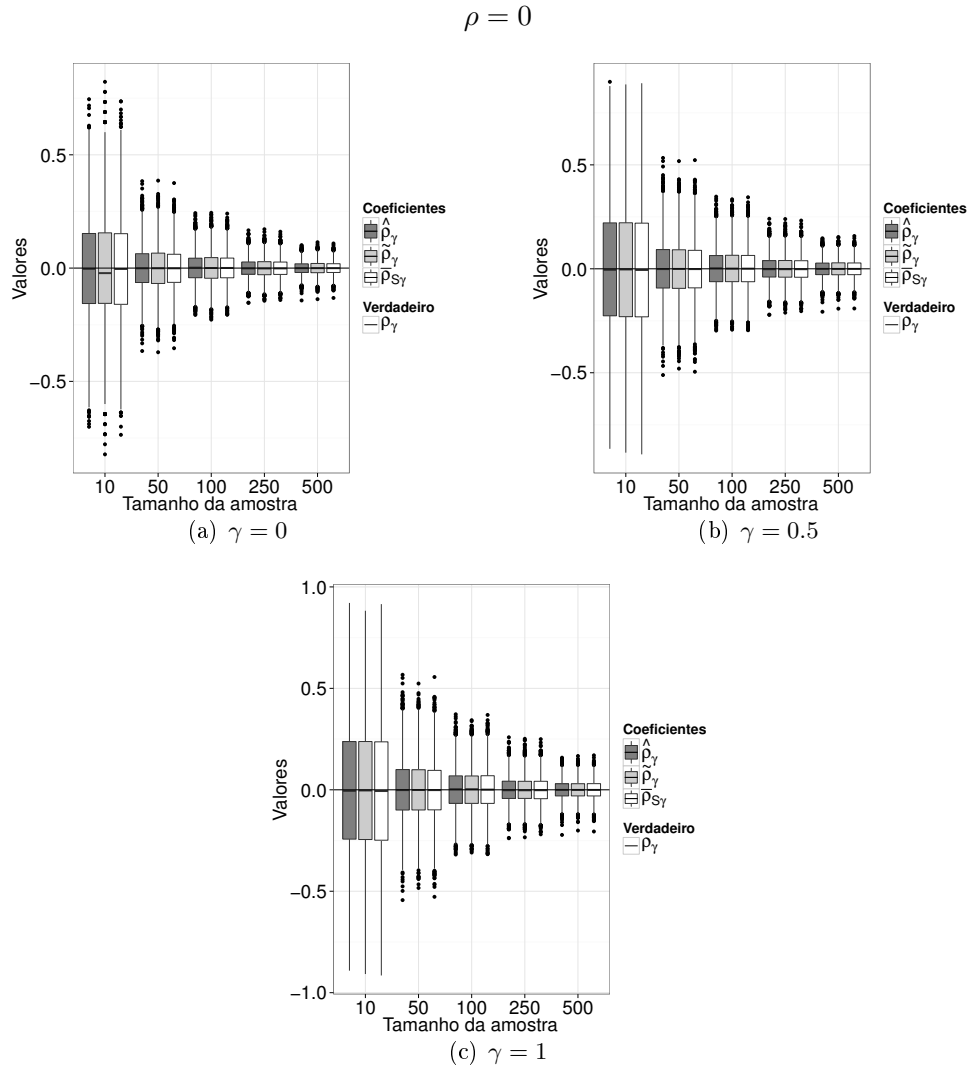


Figura 3.7: Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ para o terceiro caso das amostras normais bivariadas sem perturbação e considerando $\rho = 0$.

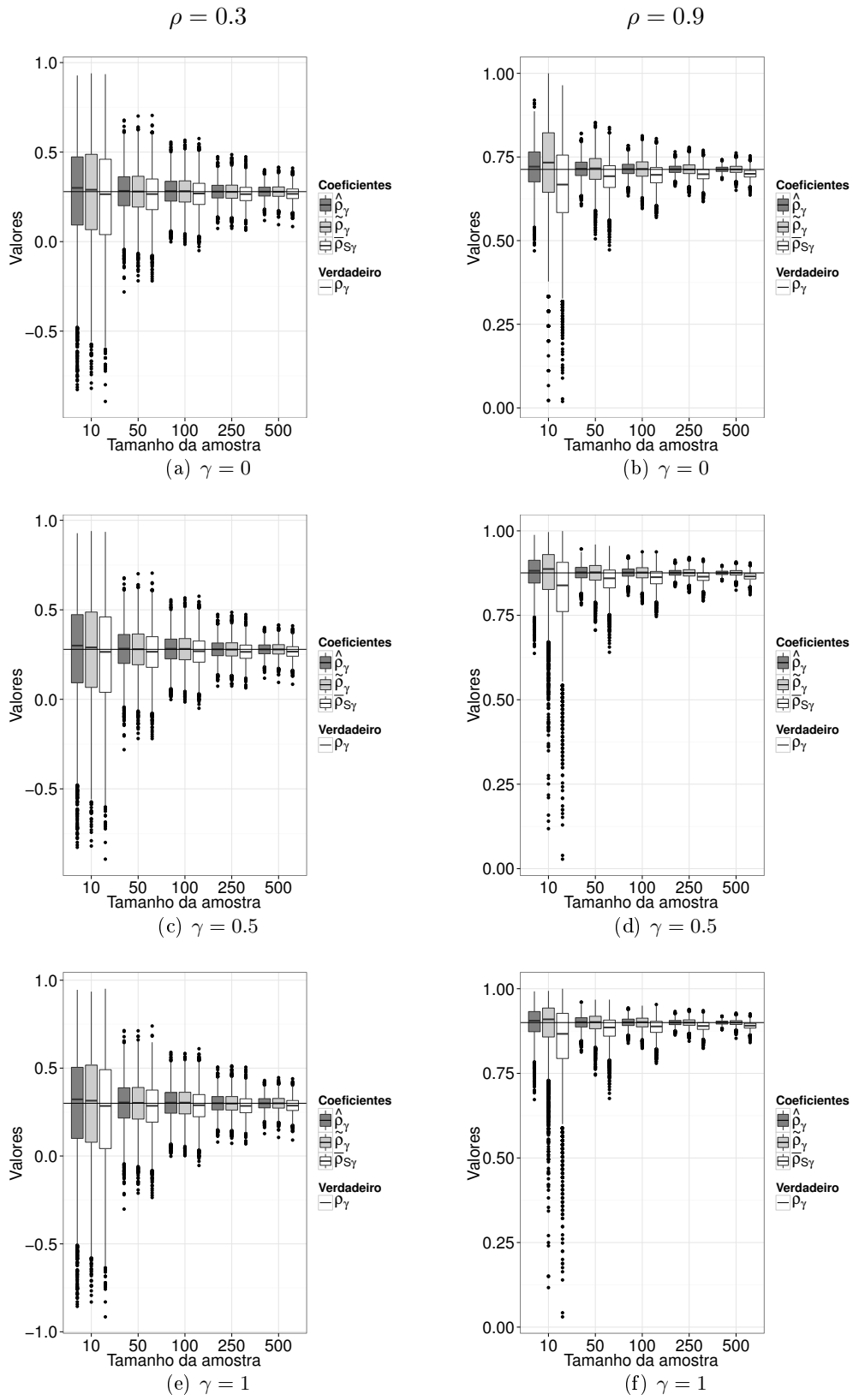


Figura 3.8: Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ para o terceiro caso das amostras normais bivariadas sem perturbação e considerando $\rho = 0.3$ e $\rho = 0.9$.

Caso 4

Para este caso observamos nos gráficos da Figura 3.9 que o CCGMV tende a subestimar o verdadeiro valor ρ_γ em todos os valores de γ e todos os tamanhos da amostra quando temos $\rho = 0$. Vemos também que o CCGU possui este mesmo comportamento quando temos $\gamma = 0$ e amostras de tamanho 10. Porém a medida que aumentamos o tamanho da amostra este estimador tende a convergir ao verdadeiro valor ρ_γ . Analisando o CCGS notamos a convergência ao verdadeiro valor ρ_γ em todo os valores de γ e tamanhos da amostra.

$$\rho = 0$$

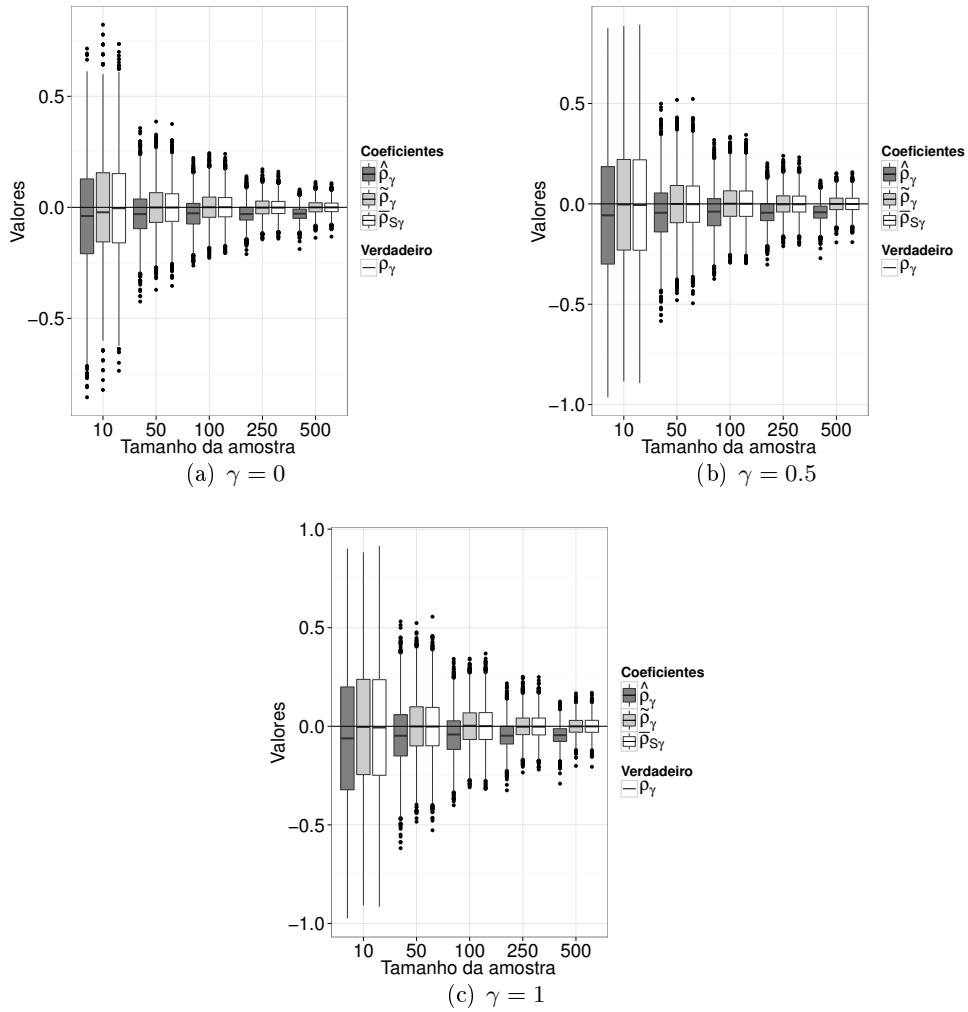


Figura 3.9: Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ para o quarto caso das amostras normais bivariadas sem perturbação e considerando $\rho = 0$.

Analisando a Figura 3.10 notamos que quando temos $\rho = 0.3$ o CCGMV tende a subestimar o verdadeiro valor ρ_γ para todos os valores de γ . Ao aumentarmos o

tamanho da amostra observamos que este estimador tende a divergir do verdadeiro valor ρ_γ . Obsevando o CCGU notamos que quando temos amostras de tamanho 10 este estimador tende a superestimar o verdadeiro valor ρ_γ para todos os valores de γ . Entretanto, ao aumentarmos o tamanho da amostra esta tendência de superestimação desaparece e sua variabilidade tende a diminuir. O CCGS continua a ter um comportamento de subestimação do verdadeiro valor ρ_γ para todos os valores de γ , porém com menos intensidade em comparação ao CCGMV, principalmente quando temos amostras de tamanho maior que 50.

Agora para $\rho = 0.9$ notamos que o CCGMV continua a subestimar o verdadeiro valor ρ_γ porém com menos variabilidade se comparado ao caso de $\rho = 0.3$. Este comportamento permanece para todos os valores de γ e ao aumentarmos o tamanho da amostra observamos uma tendência de divergência do verdadeiro valor ρ_γ . O CCGU tende a superestimar o verdadeiro valor ρ_γ quando temos amostras de tamanho 10 para todos os valores de γ . Outra característica deste estimador é a presença de um comportamento distribucional assimétrico para este tamanho de amostra e que torna-se mais evidente a medida que aumentamos o valor de γ . Porém, à medida que aumentamos o tamanho da amostra tanto a tendência de superestimação quanto o comportamento assimétrico são atenuados para todos os valores de γ . O CCGS tende a subestimar o verdadeiro valor ρ_γ para todos os valores de γ porém com menor intensidade se comparado ao CCGMV quando temos amostras de tamanho maior que 100. No caso em que temos amostras de tamanho 10, observamos que CCGS tende a subestimar o verdadeiro valor ρ_γ mais do que CCGMV. Além disso, possui um comportamento distribucional bastante assimétrico e com mais variabilidade.

A Tabela 3.4 apresenta os valores da raiz quadrada do erro quadrático médio das estimativas dos três estimadores. Nela podemos observar que quando temos $\rho = 0$ o CCGS é o que apresenta os menores valores seguido do CCGU para todos os tamanhos de amostra e valores de γ . Quando temos $\rho = 0.3$, notamos que CCGU é o que possui os menores valores da raiz do erro quadrático médio principalmente quando temos amostras de tamanho 500. Para amostras de tamanho 10 e considerando γ igual a zero CCGS é o que apresenta o menor valor da raiz quadrada do erro quadrático médio. Porém ao aumentarmos o valor de γ CCGU é o que apresenta o menor valor da raiz quadrada do erro quadrático médio.

Tabela 3.4: Raiz do erro quadrático médio das estimativas do coeficiente de correlação generalizado $\hat{\rho}_\gamma$, $\tilde{\rho}_\gamma$ e $\bar{\rho}_{S_\gamma}$ para o quarto caso sem perturbação da amostra.

$\rho = 0$						
γ	Método	Tamanho da Amostra				
		10	50	100	250	500
0	$\hat{\rho}_\gamma$	0.2489	0.1067	0.0754	0.0523	0.0417
	$\tilde{\rho}_\gamma$	0.2480	0.0981	0.0676	0.0428	0.0298
	$\bar{\rho}_{S_\gamma}$	0.2246	0.0925	0.0642	0.0407	0.0284
0.5	$\hat{\rho}_\gamma$	0.3433	0.1537	0.1092	0.0760	0.0606
	$\tilde{\rho}_\gamma$	0.3154	0.1352	0.0941	0.0600	0.0418
	$\bar{\rho}_{S_\gamma}$	0.3129	0.1336	0.0931	0.0592	0.0414
1	$\hat{\rho}_\gamma$	0.3645	0.1652	0.1176	0.0820	0.0653
	$\tilde{\rho}_\gamma$	0.3315	0.1439	0.1003	0.0640	0.0446
	$\bar{\rho}_{S_\gamma}$	0.3330	0.1437	0.1003	0.0639	0.0446
$\rho = 0.3$						
0	$\hat{\rho}_\gamma$	0.2361	0.0931	0.0640	0.0432	0.0328
	$\tilde{\rho}_\gamma$	0.2404	0.0944	0.0649	0.0411	0.0286
	$\bar{\rho}_{S_\gamma}$	0.2187	0.0905	0.0628	0.0406	0.0288
0.5	$\hat{\rho}_\gamma$	0.3252	0.1315	0.0905	0.0612	0.0465
	$\tilde{\rho}_\gamma$	0.2969	0.1253	0.0869	0.0554	0.0386
	$\bar{\rho}_{S_\gamma}$	0.2993	0.1271	0.0884	0.0572	0.0407
1	$\hat{\rho}_\gamma$	0.3450	0.1405	0.0967	0.0654	0.0497
	$\tilde{\rho}_\gamma$	0.3097	0.1315	0.0912	0.0582	0.0406
	$\bar{\rho}_{S_\gamma}$	0.3171	0.1356	0.0943	0.0611	0.0434
$\rho = 0.9$						
0	$\hat{\rho}_\gamma$	0.0631	0.0343	0.0281	0.0243	0.0228
	$\tilde{\rho}_\gamma$	0.1380	0.0470	0.0314	0.0196	0.0135
	$\bar{\rho}_{S_\gamma}$	0.1418	0.0542	0.0378	0.0258	0.0199
0.5	$\hat{\rho}_\gamma$	0.0525	0.0280	0.0228	0.0195	0.0181
	$\tilde{\rho}_\gamma$	0.0940	0.0331	0.0225	0.0142	0.0099
	$\bar{\rho}_{S_\gamma}$	0.1357	0.0456	0.0310	0.0208	0.0158
1	$\hat{\rho}_\gamma$	0.0480	0.0254	0.0205	0.0175	0.0163
	$\tilde{\rho}_\gamma$	0.0839	0.0286	0.0193	0.0122	0.0085
	$\bar{\rho}_{S_\gamma}$	0.1301	0.0417	0.0281	0.0187	0.0142

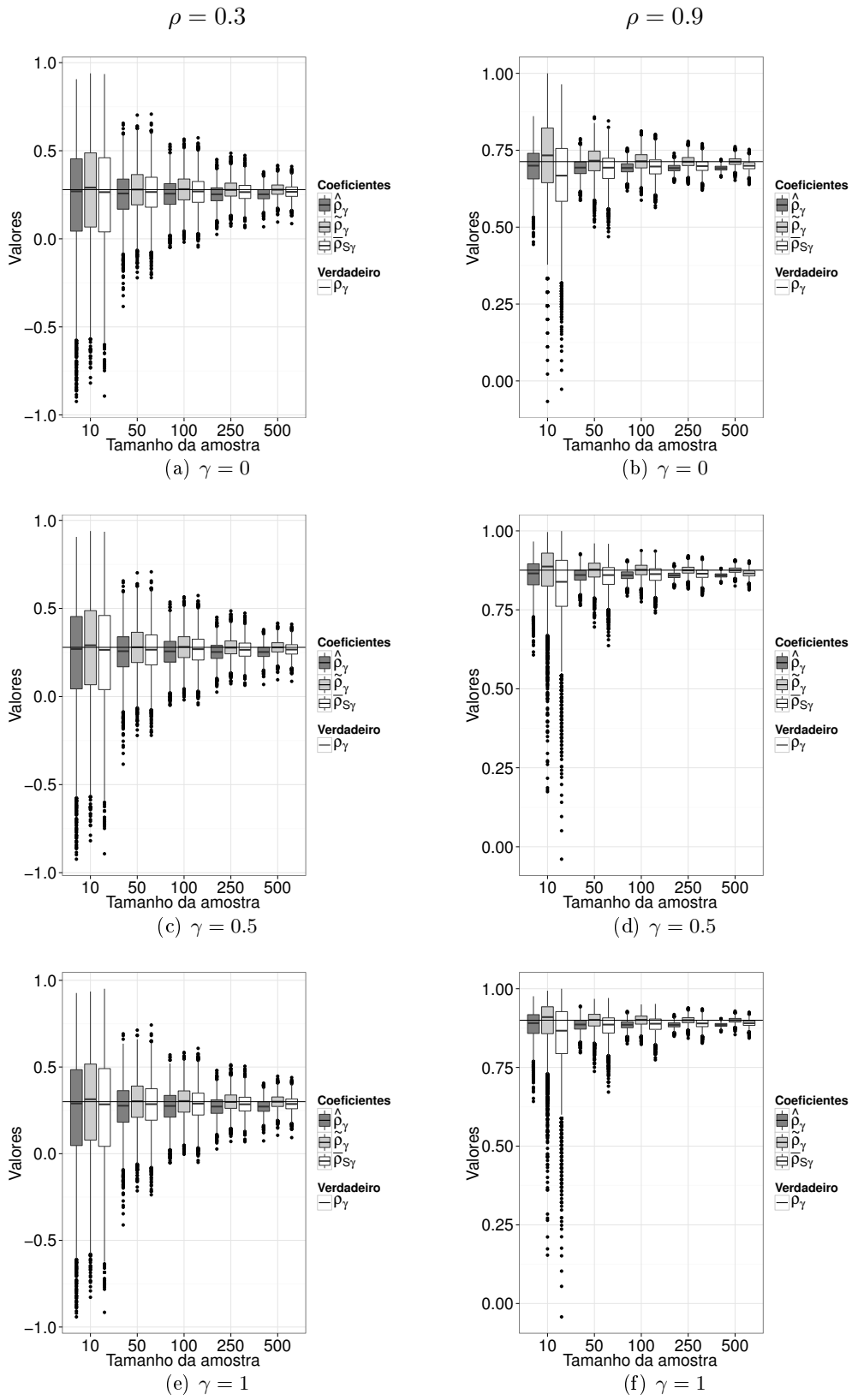


Figura 3.10: Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ para o quarto caso das amostras normais bivariadas sem perturbação e considerando $\rho = 0.3$ e $\rho = 0.9$.

3.3.2 Análise dos estimadores com perturbação da amostra

Neste cenário, temos a distribuição da amostra perturbada com valores da distribuição t -Student bivariada com 3 graus de liberdade como definimos na Seção 3.2.2. Consideramos ρ igual a 0.1 (correlação fraca), 0.5 (correlação moderada) e 0.9 (correlação forte). Já 10%, 30% e 50% são os níveis de perturbação da amostra e os valores de γ são 0, 0.5 e 1.

Nos gráficos da Figura 3.11 apresentamos os diagramas boxplot para uma amostra perturbada em 10% e com $\rho = 0.1$. Neste caso observamos um comportamento muito semelhante dos três estimadores para os três valores de γ . Porém notamos

10% de perturbação da amostra

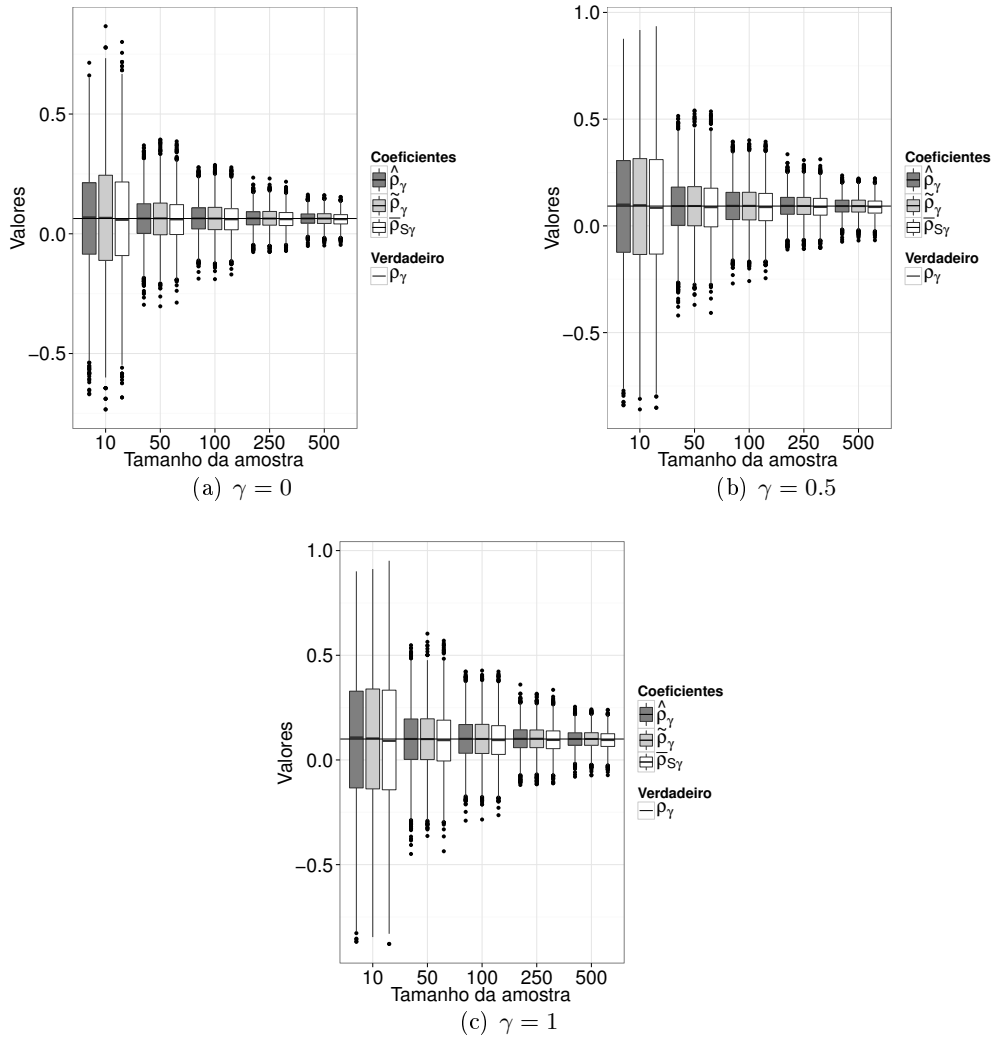


Figura 3.11: Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ com 10% de perturbação da amostra e vários valores de γ considerando $\rho = 0.1$.

30% de perturbação da amostra

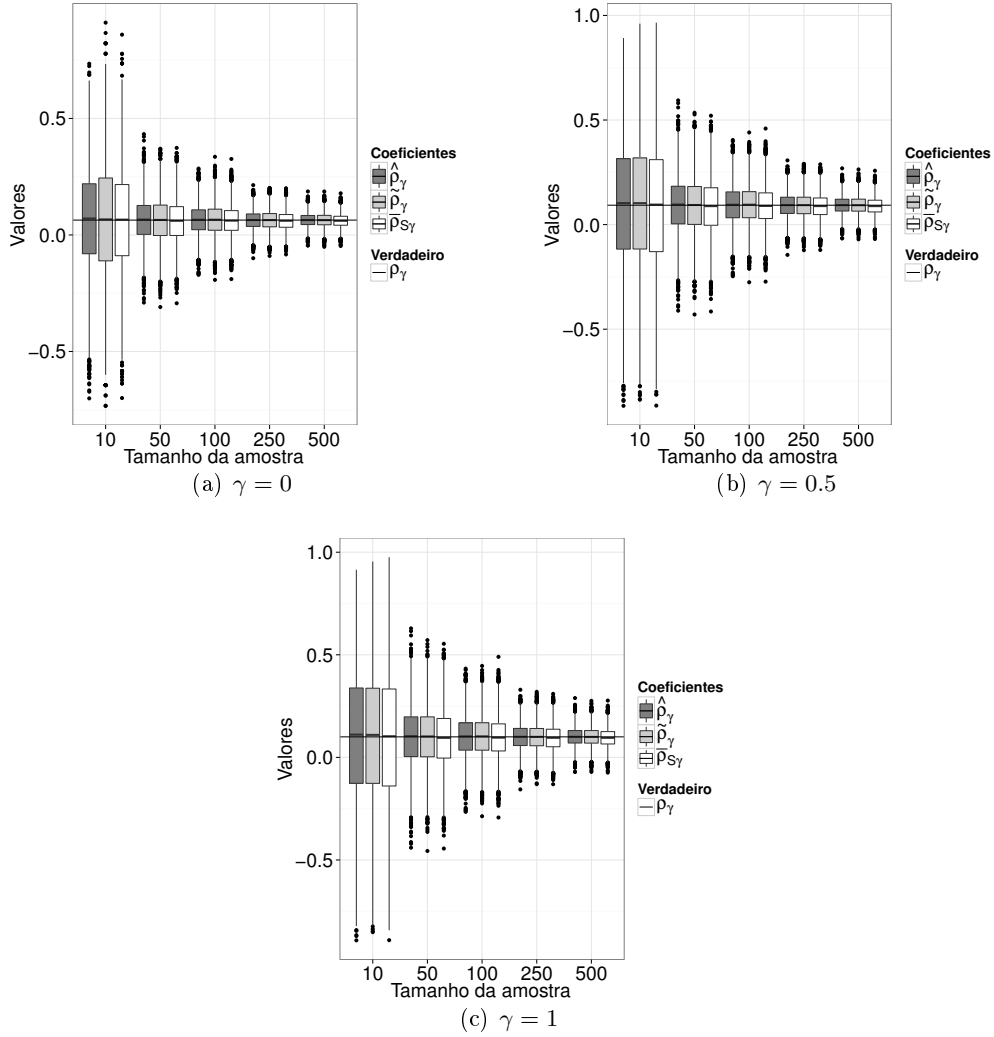


Figura 3.12: Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ com 30% de perturbação da amostra e vários valores de γ considerando $\rho = 0.1$.

que quando temos amostras de tamanho 10 e γ igual a 0 o CCGU apresenta uma maior variabilidade do que os demais estimadores. Para γ igual a 1 temos uma pequena subestimação do CCGS nas amostras de tamanho 10 mas a medida que aumentamos o tamanho da amostra essa tendência é atenuada.

Na Figura 3.12 apresentamos os gráficos boxplot com perturbação de 30% da amostra e o mesmo valor de ρ . Nesta situação, podemos observar que para $\gamma = 0.5$ e amostra de tamanho 10 há uma superestimação dos estimadores CCGMV e CCGU com ambos apresentando variabilidade bastante semelhante. Já para $\gamma = 0$ o CCGU apresenta maior variabilidade quando temos amostras de tamanho 10.

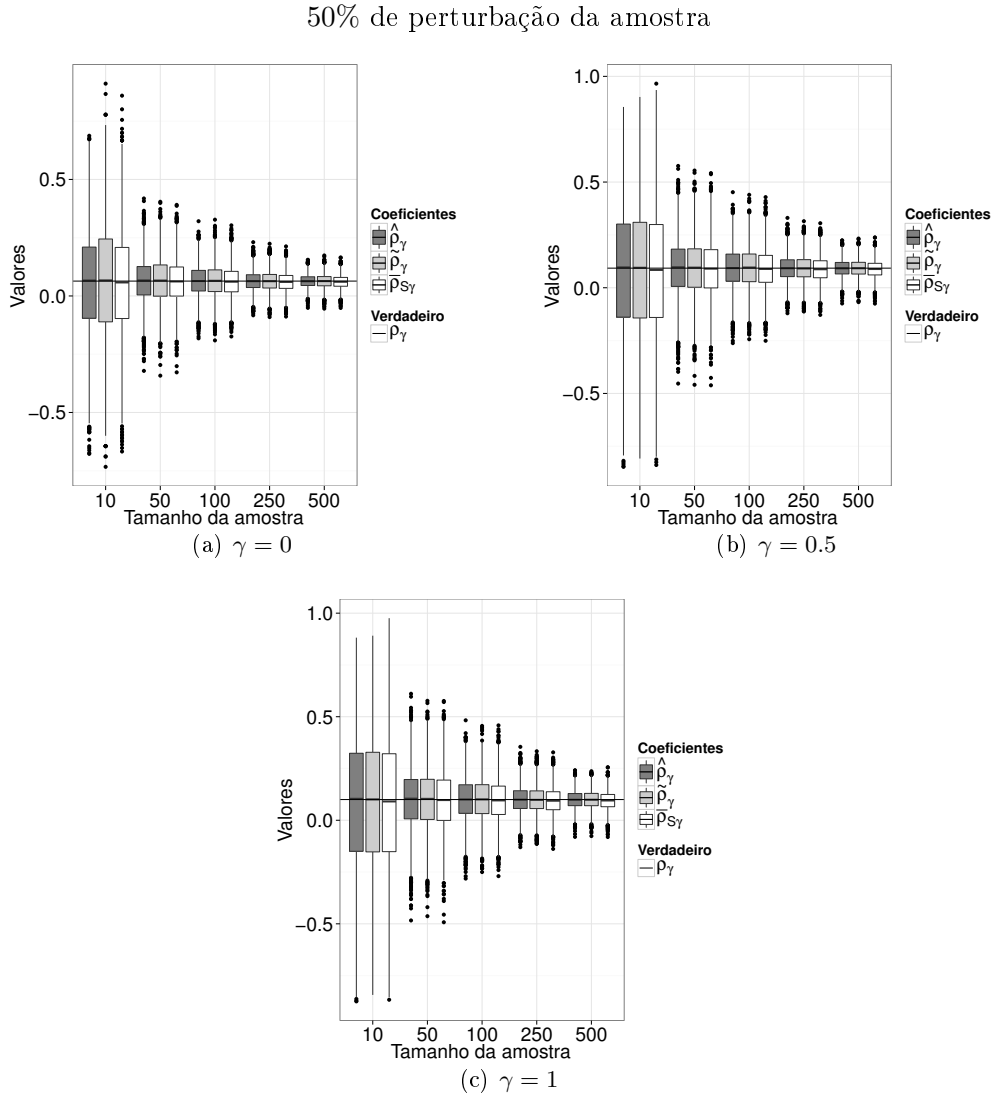


Figura 3.13: Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ com 50% de perturbação da amostra e vários valores de γ considerando $\rho = 0.1$.

Ainda para $\rho = 0.1$, temos na Figura 3.13 os gráficos boxplot para uma perturbação de 50% da amostra. Observamos, então, um comportamento bastante semelhante entre os três estimadores, principalmente nos casos que temos γ igual a 1.

A Tabela 3.5 apresenta os valores da raiz quadrada do erro quadrático médio para os três estimadores. Nela podemos notar que quando temos $\gamma = 1$ e amostras de tamanho 10 a medida que aumentamos o nível de contaminação há um pequeno aumento na variabilidade destes estimadores. Contudo, ao aumentarmos o tamanho da amostra os três estimadores diminuem consideravelmente os valores da raiz

quadrada do erro quadrático médio.

Tabela 3.5: Raiz do erro quadrático médio das estimativas do coeficiente de correlação generalizado $\hat{\rho}_\gamma$, $\tilde{\rho}_\gamma$ e $\bar{\rho}_{S_\gamma}$ com perturbação da amostra aos níveis de 10%, 30% e 50% com $\rho = 0.1$.

γ	Método	Tamanho da Amostra				
		10	50	100	250	500
10% de perturbação						
0	$\hat{\rho}_\gamma$	0.2238	0.0991	0.0768	0.0550	0.0456
	$\tilde{\rho}_\gamma$	0.2495	0.1115	0.0864	0.0619	0.0514
	$\bar{\rho}_{S_\gamma}$	0.3117	0.1427	0.1107	0.0795	0.0666
0.5	$\hat{\rho}_\gamma$	0.2685	0.1090	0.0771	0.0482	0.0341
	$\tilde{\rho}_\gamma$	0.3224	0.1417	0.1025	0.0649	0.0468
	$\bar{\rho}_{S_\gamma}$	0.3444	0.1694	0.1339	0.0959	0.0762
1	$\hat{\rho}_\gamma$	0.3128	0.1328	0.0943	0.0593	0.0421
	$\tilde{\rho}_\gamma$	0.3321	0.1424	0.1013	0.0637	0.0453
	$\bar{\rho}_{S_\gamma}$	0.3327	0.1428	0.1015	0.0639	0.0454
30% de perturbação						
0	$\hat{\rho}_\gamma$	0.2179	0.1085	0.0875	0.0663	0.0566
	$\tilde{\rho}_\gamma$	0.2431	0.1220	0.0984	0.0746	0.0638
	$\bar{\rho}_{S_\gamma}$	0.3041	0.1559	0.1261	0.0958	0.0819
0.5	$\hat{\rho}_\gamma$	0.2779	0.1146	0.0802	0.0498	0.0360
	$\tilde{\rho}_\gamma$	0.3423	0.1594	0.1158	0.0733	0.0537
	$\bar{\rho}_{S_\gamma}$	0.3745	0.2098	0.1703	0.1253	0.1051
1	$\hat{\rho}_\gamma$	0.3189	0.1371	0.0961	0.0603	0.0437
	$\tilde{\rho}_\gamma$	0.3383	0.1471	0.1032	0.0647	0.0469
	$\bar{\rho}_{S_\gamma}$	0.3389	0.1474	0.1035	0.0649	0.0471
50% de perturbação						
0	$\hat{\rho}_\gamma$	0.2188	0.1132	0.0898	0.0697	0.0584
	$\tilde{\rho}_\gamma$	0.2440	0.1272	0.1010	0.0785	0.0658
	$\bar{\rho}_{S_\gamma}$	0.3050	0.1626	0.1293	0.1008	0.0844
0.5	$\hat{\rho}_\gamma$	0.2835	0.1177	0.0823	0.0518	0.0364
	$\tilde{\rho}_\gamma$	0.3572	0.1709	0.1227	0.0788	0.0563
	$\bar{\rho}_{S_\gamma}$	0.3957	0.2349	0.1849	0.1395	0.1109
1	$\hat{\rho}_\gamma$	0.3233	0.1393	0.0976	0.0622	0.0437
	$\tilde{\rho}_\gamma$	0.3428	0.1493	0.1048	0.0668	0.0470
	$\bar{\rho}_{S_\gamma}$	0.3434	0.1497	0.1050	0.0670	0.0471

Para $\rho = 0.5$ e 10% de perturbação da amostra, podemos observar na Figura 3.14 que quando temos amostras de tamanho 10 o CCGMV apresenta uma superestimação do verdadeiro valor ρ_γ e um comportamento distribucional assimétrico. O CCGU apresenta uma superestimação do verdadeiro valor ρ_γ quando temos γ igual a 0.5 e 1 em amostras de tamanho 10. O CCGS tende a subestimar o verdadeiro valor ρ_γ em todos os valores de γ e em todos os tamanhos da amostra.

10% de perturbação da amostra

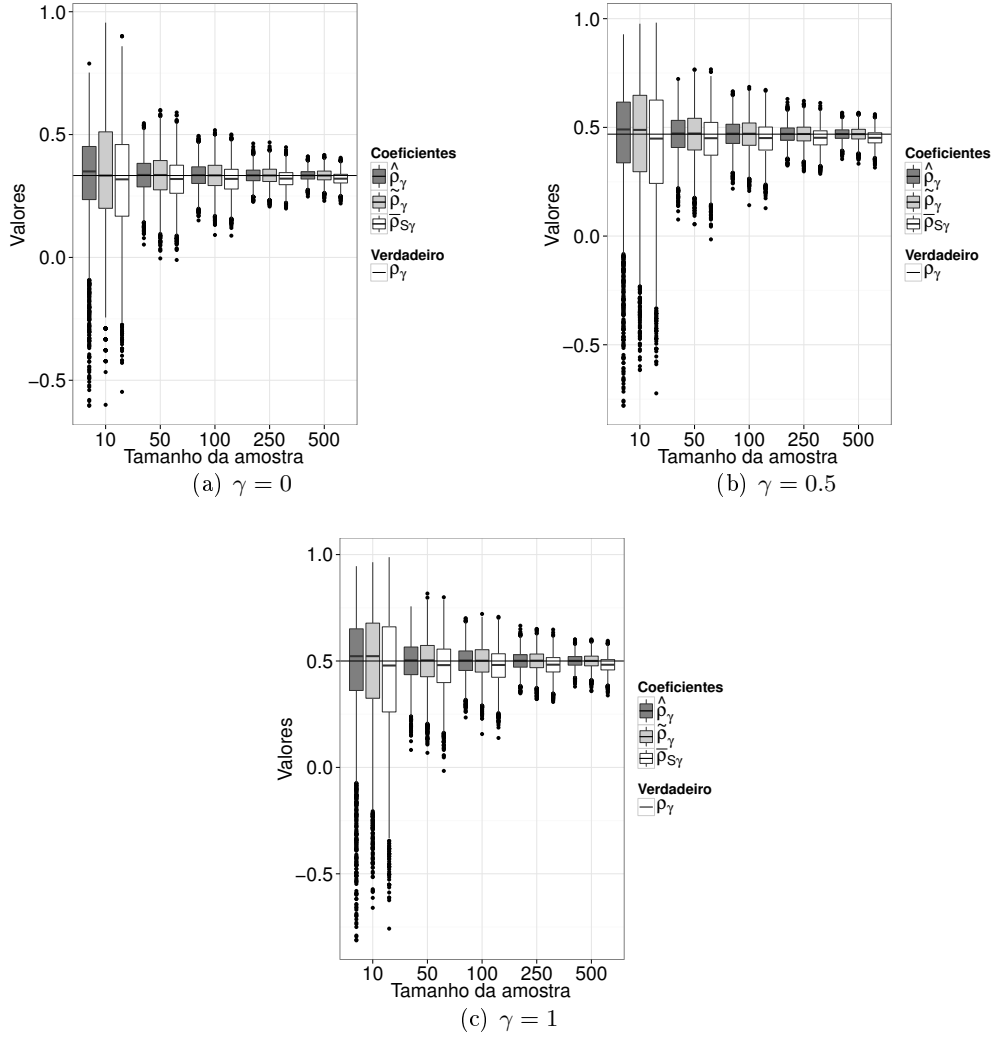


Figura 3.14: Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ com 10% de perturbação da amostra e vários valores de γ considerando $\rho = 0.5$.

Com 30% de perturbação da amostra podemos observar na Figura 3.15 que os três estimadores apresentam comportamento bastante semelhante ao caso em que temos 10% de perturbação da amostra (Figura 3.14). Para amostras de tamanho 10 e para γ igual a 0.5 e 1 tanto o CCGMV quanto o CCGU superestimam o verdadeiro valor ρ_γ . Quando temos $\gamma = 0$ apenas o CCGMV superestima o verdadeiro valor ρ_γ . O CCGS novamente subestima o verdadeiro valor ρ_γ para todos os valores de γ e em todos os tamanhos da amostra.

Analisando os estimadores com 50% de perturbação da amostra podemos ver nos gráficos da Figura 3.16 comportamentos parecidos com os níveis de perturbação

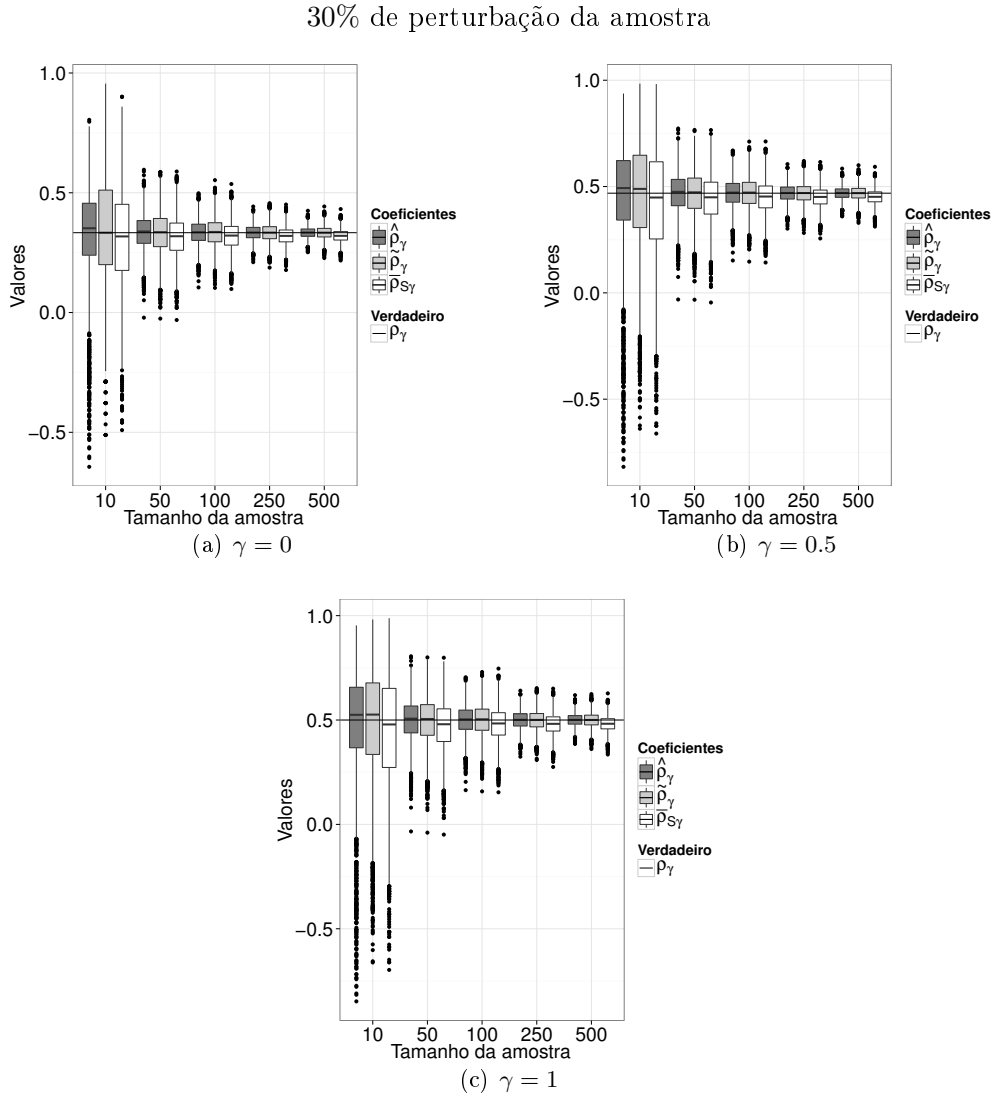


Figura 3.15: Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ com 30% de perturbação da amostra e vários valores de γ considerando $\rho = 0.5$.

10% e 30% analisados anteriormente. Assim, CCGMV e o CCGU superestimam o verdadeiro valor ρ_γ quando temos amostras de tamanho 10 e γ igual a 0.5 e 1. O CCGS continua subestimando o verdadeiro valor ρ_γ para todos os valores de γ e todos os tamanhos da amostra.

Na Tabela 3.6 apresentamos os valores da raiz quadrada do erro quadrático médio para $\rho = 0.5$ e com 10%, 30% e 50% de perturbação da amostra. Podemos ver que a medida que aumentamos o nível de perturbação da amostra, os valores da raiz do erro quadrático médio também aumentam, o que era de se esperar. Notamos também que quando temos amostras de tamanho 10 e $\gamma = 0$ os três estimadores

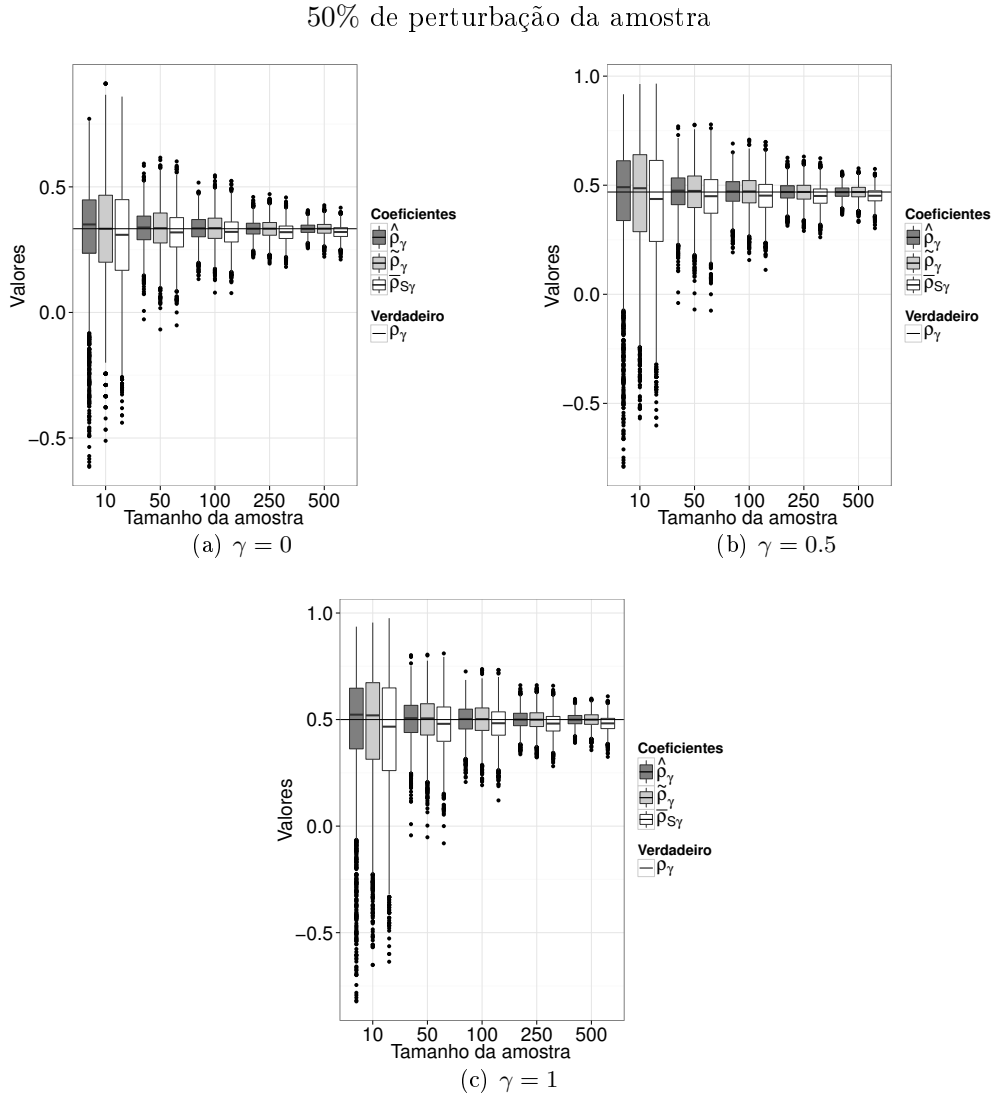


Figura 3.16: Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ com 50% de perturbação da amostra e vários valores de γ considerando $\rho = 0.5$.

apresentam os menores valores da raiz do erro quadrático médio em comparação aos outros valores de γ .

Por fim, analisamos os três estimadores quando temos $\rho = 0.9$ e os níveis 10%, 30% e 50% de perturbação da amostra. A Figura 3.17 mostra o comportamento dos três estimadores quando temos 10% de perturbação da amostra. Nela podemos observar que quando temos amostras de tamanho 10 o CCGU tende a superestimar o verdadeiro valor ρ_γ para todos os valores de γ . Além disso, este estimador apresenta um comportamento distribucional assimétrico quando temos γ igual a 0.5 e 1. Observamos também que o CCGMV apresenta uma superestimação do verdadeiro

Tabela 3.6: Raiz do erro quadrático médio das estimativas do coeficiente de correlação generalizado $\hat{\rho}_\gamma$, $\tilde{\rho}_\gamma$ e $\bar{\rho}_{S_\gamma}$ com perturbação da amostra aos níveis de 10%, 30% e 50% com $\rho = 0.5$.

γ	Método	Tamanho da Amostra				
		10	50	100	250	500
10% de perturbação						
0	$\hat{\rho}_\gamma$	0.1906	0.0896	0.0749	0.0576	0.0535
	$\tilde{\rho}_\gamma$	0.2098	0.0991	0.0829	0.0637	0.0592
	$\bar{\rho}_{S_\gamma}$	0.2534	0.1207	0.1014	0.0778	0.0724
0.5	$\hat{\rho}_\gamma$	0.2370	0.0946	0.0664	0.0414	0.0293
	$\tilde{\rho}_\gamma$	0.2693	0.1142	0.0822	0.0515	0.0373
	$\bar{\rho}_{S_\gamma}$	0.2834	0.1331	0.1060	0.0725	0.0601
1	$\hat{\rho}_\gamma$	0.2731	0.1150	0.0817	0.0531	0.0401
	$\tilde{\rho}_\gamma$	0.2861	0.1204	0.0854	0.0554	0.0419
	$\bar{\rho}_{S_\gamma}$	0.2865	0.1205	0.0855	0.0555	0.0419
30% de perturbação						
0	$\hat{\rho}_\gamma$	0.1964	0.1167	0.1045	0.0917	0.0881
	$\tilde{\rho}_\gamma$	0.2168	0.1295	0.1161	0.1017	0.0977
	$\bar{\rho}_{S_\gamma}$	0.2638	0.1596	0.1432	0.1250	0.1198
0.5	$\hat{\rho}_\gamma$	0.2479	0.0997	0.0693	0.0432	0.0311
	$\tilde{\rho}_\gamma$	0.2864	0.1294	0.0929	0.0588	0.0429
	$\bar{\rho}_{S_\gamma}$	0.3075	0.1680	0.1345	0.0970	0.0812
1	$\hat{\rho}_\gamma$	0.2832	0.1204	0.0847	0.0567	0.0436
	$\tilde{\rho}_\gamma$	0.2968	0.1260	0.0886	0.0593	0.0455
	$\bar{\rho}_{S_\gamma}$	0.2972	0.1262	0.0887	0.0594	0.0455
50% de perturbação						
0	$\hat{\rho}_\gamma$	0.2066	0.1347	0.1213	0.1117	0.1075
	$\tilde{\rho}_\gamma$	0.2285	0.1498	0.1348	0.1240	0.1192
	$\bar{\rho}_{S_\gamma}$	0.2795	0.1855	0.1667	0.1527	0.1464
0.5	$\hat{\rho}_\gamma$	0.2527	0.1022	0.0711	0.0450	0.0316
	$\tilde{\rho}_\gamma$	0.3020	0.1393	0.0988	0.0632	0.0452
	$\bar{\rho}_{S_\gamma}$	0.3308	0.1894	0.1477	0.1096	0.0873
1	$\hat{\rho}_\gamma$	0.2888	0.1231	0.0877	0.0605	0.0459
	$\tilde{\rho}_\gamma$	0.3025	0.1289	0.0918	0.0633	0.0480
	$\bar{\rho}_{S_\gamma}$	0.3030	0.1291	0.0919	0.0634	0.0481

valor ρ_γ quando temos $\gamma = 0$ e amostras de tamanho 10. O CCGS mantém o comportamento de subestimar o verdadeiro valor ρ_γ em todos os valores de γ e em todos os tamanhos da amostra. Outra característica deste estimador é o comportamento distribucional mais assimétrico quando temos amostras de tamanho 10.

Com 30% de perturbação da amostra, podemos observar na Figura 3.18 que o comportamento dos três estimadores é semelhante aos encontrados quando temos 10% de perturbação da amostra. O CCGU tende a superestimar o verdadeiro valor

10% de perturbação da amostra

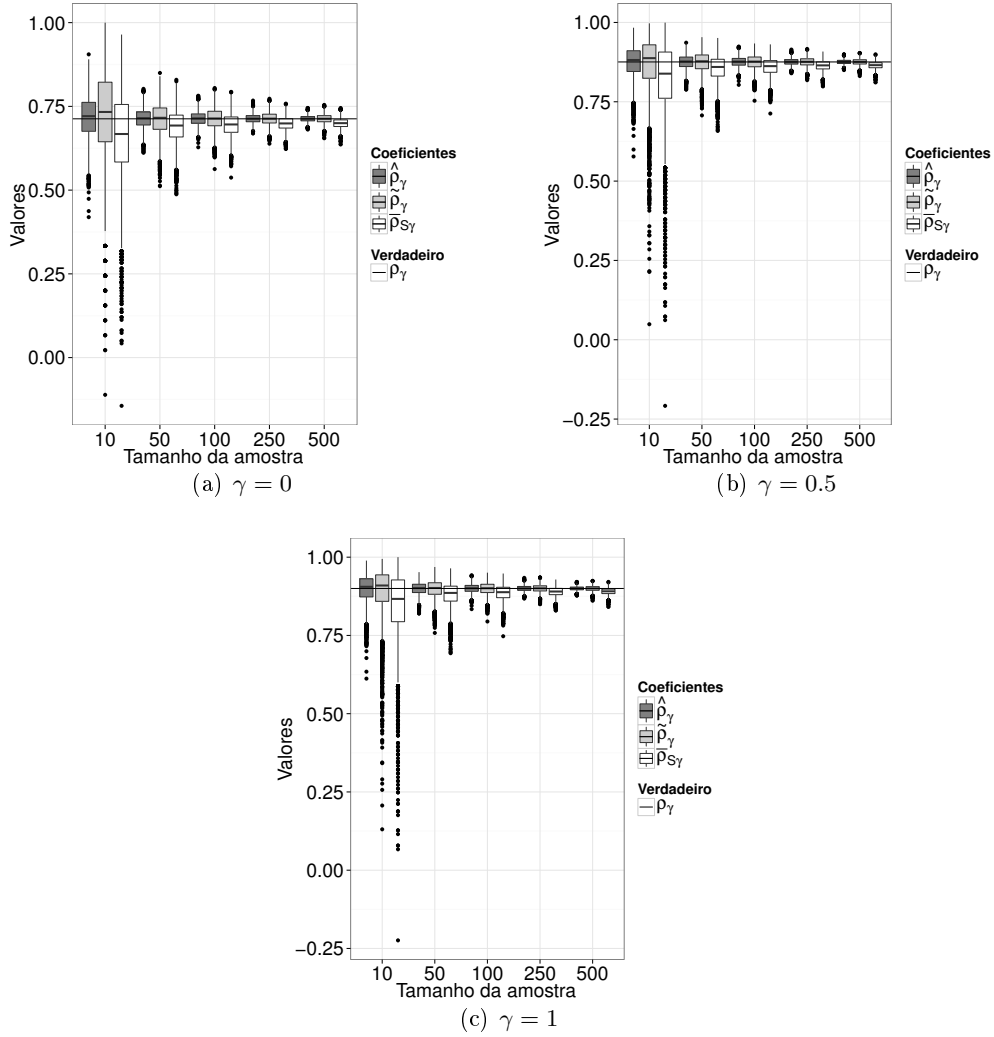


Figura 3.17: Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ com 10% de perturbação da amostra e vários valores de γ considerando $\rho = 0.9$.

ρ_γ quando temos amostras de tamanho 10 para todos os valores de γ . Além disso, observamos um comportamento distribucional assimétrico quando temos amostras de tamanho 10 e γ igual a 0.5 e 1.

O CCGMV tende a superestimar o verdadeiro valor ρ_γ quando temos γ igual a 0 e 0.5 e amostras de tamanho 10, porém superestima menos se comparado ao estimador do coeficiente de correlação generalizado nestas condições. Também podemos observar que o CCGMV é o que apresenta um comportamento distribucional menos assimétrico quando temos amostras de tamanho 10. O CCGS permanece subestimando o verdadeiro valor ρ_γ para todos os valores de γ e em todos os tamanhos da

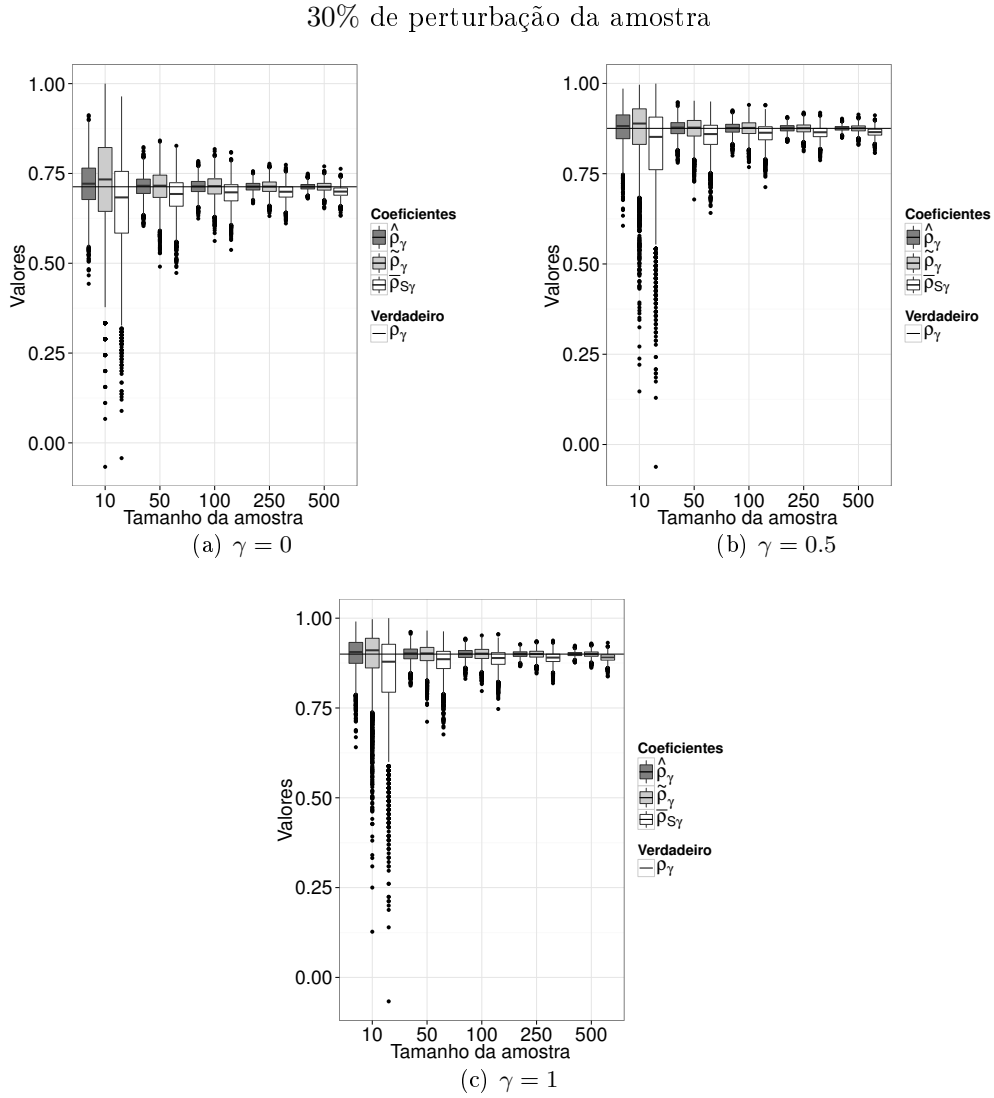


Figura 3.18: Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ com 30% de perturbação da amostra e vários valores de γ considerando $\rho = 0.9$.

amostra. Este estimador também apresenta um forte comportamento assimétrico quando temos amostras de tamanho 10.

Na Figura 3.19 apresentamos através dos gráficos boxplot os comportamentos dos três estimadores quando temos 50% de perturbação da amostra. Nela notamos um comportamento semelhante aos casos em que temos 10% e 30% de perturbação da amostra. O CCGU tende a superestimar o verdadeiro valor ρ_γ em todos os valores de γ quando temos amostras de tamanho 10. O comportamento distribucional assimétrico continua para os valores de γ igual a 0.5 e 1 para este mesmo tamanho da amostra. O CCGS também apresenta um comportamento assimétrico quando

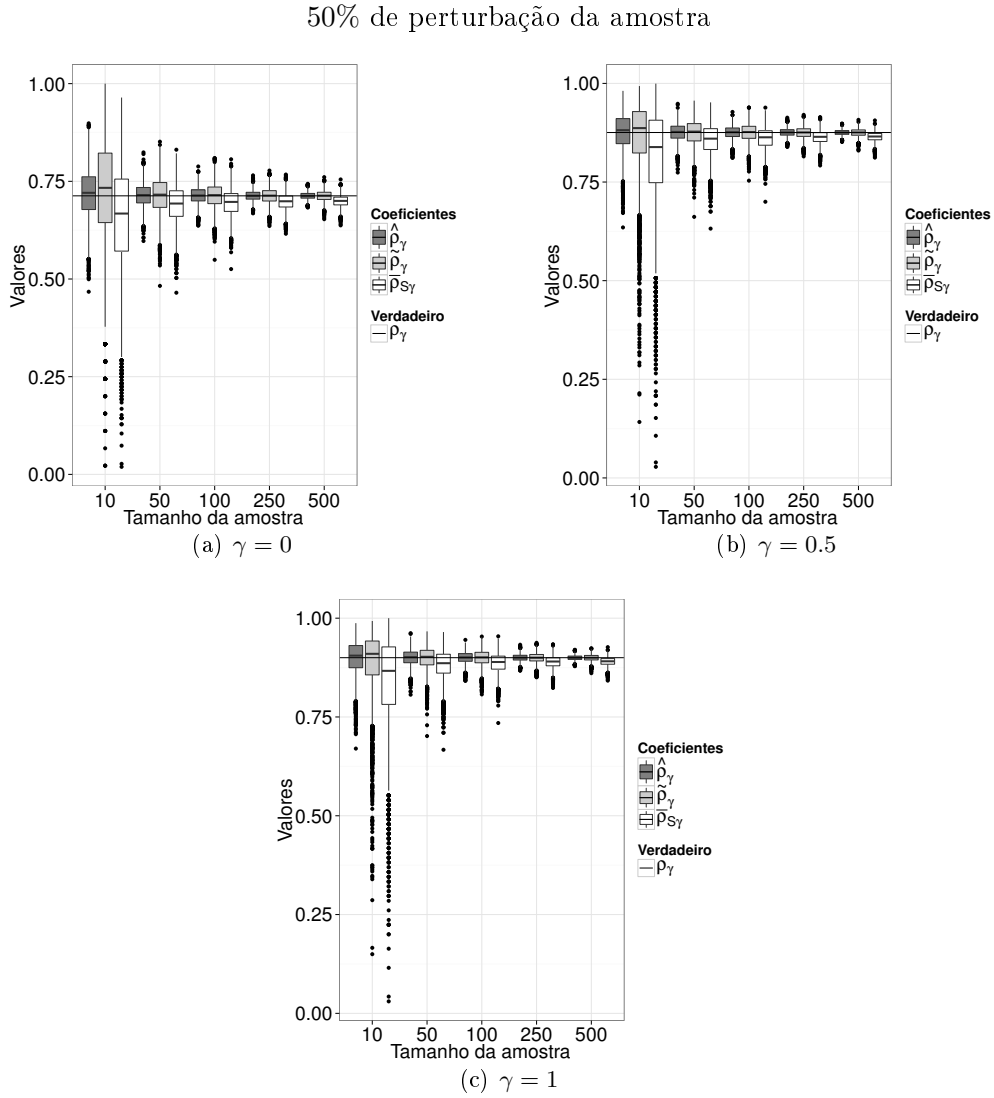


Figura 3.19: Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ com 50% de perturbação da amostra e vários valores de γ considerando $\rho = 0.9$.

temos amostras de tamanho 10 e γ igual a 0.5 e 1. Este estimador continua com a tendência de subestimar o verdadeiro valor ρ_γ em todos os tamanhos da amostra e em todos os valores de γ . O CCGMV tende a superestimar o verdadeiro valor de γ quando temos amostras de tamanho 10. Porém à medida que aumentamos o tamanho da amostra tanto o CCGMV quanto o CCGU tendem a diminuir este comportamento de superestimar o verdadeiro valor de ρ_γ em todos os valores de γ .

Na Tabela 3.7 temos os valores da raiz quadrada do erro quadrático médio das estimativas dos três estimadores. Podemos observar que a medida que aumentamos o nível de perturbação da amostra os valores da raiz quadrada do erro quadrático

médio tendem a aumentar. Para os níveis 30% e 50% de perturbação da amostra e $\gamma = 0$, quando aumentamos o tamanho da amostra há uma pequena diminuição dos valores da raiz quadrada do erro quadrático médio. Já para γ igual a 0.5 e 1 há uma maior diminuição destes valores quando aumentamos o tamanho da amostra. Observamos também que quando temos amostras de tamanho 10 e $\gamma = 1$ os valores da raiz quadrada do erro quadrático médio são os maiores em comparação aos outros valores de γ .

Como a distribuição normal bivariada e a distribuição t -Student bivariada são distribuições simétricas, mesmo aumentando o nível de perturbação da amostra o comportamento dos 3 estimadores são bastante parecidos em todos os níveis de perturbação. Para amostras de tamanho 10 os três estimadores ou apresentam tendências de superestimar ou subestimar o verdadeiro valor ρ_γ . Porém ao aumentarmos o tamanho da amostra tanto CCGMV quanto CCGU conseguem corrigir suas respectivas superestimações. Entretanto, CCGS tende a continuar subestimando o verdadeiro valor ρ_γ para todos os tamanhos da amostra em todos os valores de γ com todos os níveis de perturbação da amostra. Além disso, possui comportamento distribucional bastante assimétrico quando temos amostras de tamanho 10 em todos os valores de γ .

3.3.3 Mistura de distribuições normais bivariadas

A fim de analisar o comportamento dos estimadores CCGMV, CCGU e CCGS utilizamos a mistura de duas normais bivariadas definida na Seção 3.2.3. Apresentamos 5 casos dentro deste tipo de cenário e analisamos o comportamento destes estimadores.

Caso 1: $Q_{\underline{X}} = 0.9 \times \mathcal{N}_2(0, 0; 1, 1; 0.9) + 0.1 \times \mathcal{N}_2(0, 0; 1, 1; 0.5)$

Nos gráficos da Figura 3.20 podemos observar que para todos os valores γ e para amostras de tamanho 10, o CCGS possui melhores estimativas se comparado com os demais estimadores. Porém a medida que aumentamos o tamanho da amostra tende a divergir do verdadeiro valor ρ_γ . Outra característica deste estimador, quando temos amostras de tamanho 10, é seu comportamento distribucional bastante assimétrico que torna-se mais evidente a medida que aumentamos o valor de γ . Tanto o CCGMV quanto o CCGU tendem a superestimar o verdadeiro valor ρ_γ para todos os tamanhos da amostra e para todos os valores de γ . Aumentando o tamanho da amostra este comportamento permanece para os dois estimadores mesmo

Tabela 3.7: Raiz do erro quadrático médio das estimativas do coeficiente de correlação generalizado $\hat{\rho}_\gamma$, $\tilde{\rho}_\gamma$ e $\bar{\rho}_{S_\gamma}$ com perturbação da amostra aos níveis de 10%, 30% e 50% com $\rho = 0.9$.

γ	Método	Tamanho da Amostra				
		10	50	100	250	500
10% de perturbação						
0	$\hat{\rho}_\gamma$	0.0843	0.0518	0.0474	0.0370	0.0361
	$\tilde{\rho}_\gamma$	0.0837	0.0517	0.0477	0.0367	0.0359
	$\bar{\rho}_{S_\gamma}$	0.0763	0.0471	0.0446	0.0325	0.0324
0.5	$\hat{\rho}_\gamma$	0.1286	0.0456	0.0311	0.0192	0.0135
	$\tilde{\rho}_\gamma$	0.0997	0.0369	0.0269	0.0160	0.0116
	$\bar{\rho}_{S_\gamma}$	0.0933	0.0371	0.0337	0.0189	0.0159
1	$\hat{\rho}_\gamma$	0.1408	0.0490	0.0340	0.0225	0.0180
	$\tilde{\rho}_\gamma$	0.1358	0.0451	0.0310	0.0204	0.0162
	$\bar{\rho}_{S_\gamma}$	0.1356	0.0449	0.0309	0.0203	0.0161
30% de perturbação						
0	$\hat{\rho}_\gamma$	0.1097	0.0868	0.0818	0.0738	0.0747
	$\tilde{\rho}_\gamma$	0.1113	0.0880	0.0823	0.0735	0.0744
	$\bar{\rho}_{S_\gamma}$	0.1076	0.0840	0.0767	0.0663	0.0670
0.5	$\hat{\rho}_\gamma$	0.1335	0.0486	0.0329	0.0206	0.0147
	$\tilde{\rho}_\gamma$	0.1096	0.0443	0.0303	0.0191	0.0137
	$\bar{\rho}_{S_\gamma}$	0.1076	0.0538	0.0404	0.0266	0.0218
1	$\hat{\rho}_\gamma$	0.1450	0.0541	0.0377	0.0268	0.0216
	$\tilde{\rho}_\gamma$	0.1401	0.0500	0.0345	0.0243	0.0195
	$\bar{\rho}_{S_\gamma}$	0.1399	0.0498	0.0344	0.0242	0.0195
50% de perturbação						
0	$\hat{\rho}_\gamma$	0.1313	0.1139	0.1072	0.1051	0.1035
	$\tilde{\rho}_\gamma$	0.1345	0.1158	0.1080	0.1053	0.1034
	$\bar{\rho}_{S_\gamma}$	0.1337	0.1110	0.1010	0.0966	0.0938
0.5	$\hat{\rho}_\gamma$	0.1378	0.0506	0.0344	0.0217	0.0153
	$\tilde{\rho}_\gamma$	0.1221	0.0476	0.0330	0.0208	0.0149
	$\bar{\rho}_{S_\gamma}$	0.1265	0.0599	0.0450	0.0326	0.0280
1	$\hat{\rho}_\gamma$	0.1508	0.0583	0.0417	0.0304	0.0245
	$\tilde{\rho}_\gamma$	0.1461	0.0539	0.0382	0.0276	0.0222
	$\bar{\rho}_{S_\gamma}$	0.1459	0.0537	0.0381	0.0275	0.0221

com a diminuição de suas respectivas variabilidades.

Na Tabela 3.8 apresentamos a média quadrática do CCGMV, CCGU e CCGS. Observamos que para amostras de tamanho 10 o CCGMV é o que apresenta os menores valores. Ao aumentarmos o tamanho da amostra percebemos que os três estimadores tendem a diminuir os valores de suas respectivas médias quadráticas.

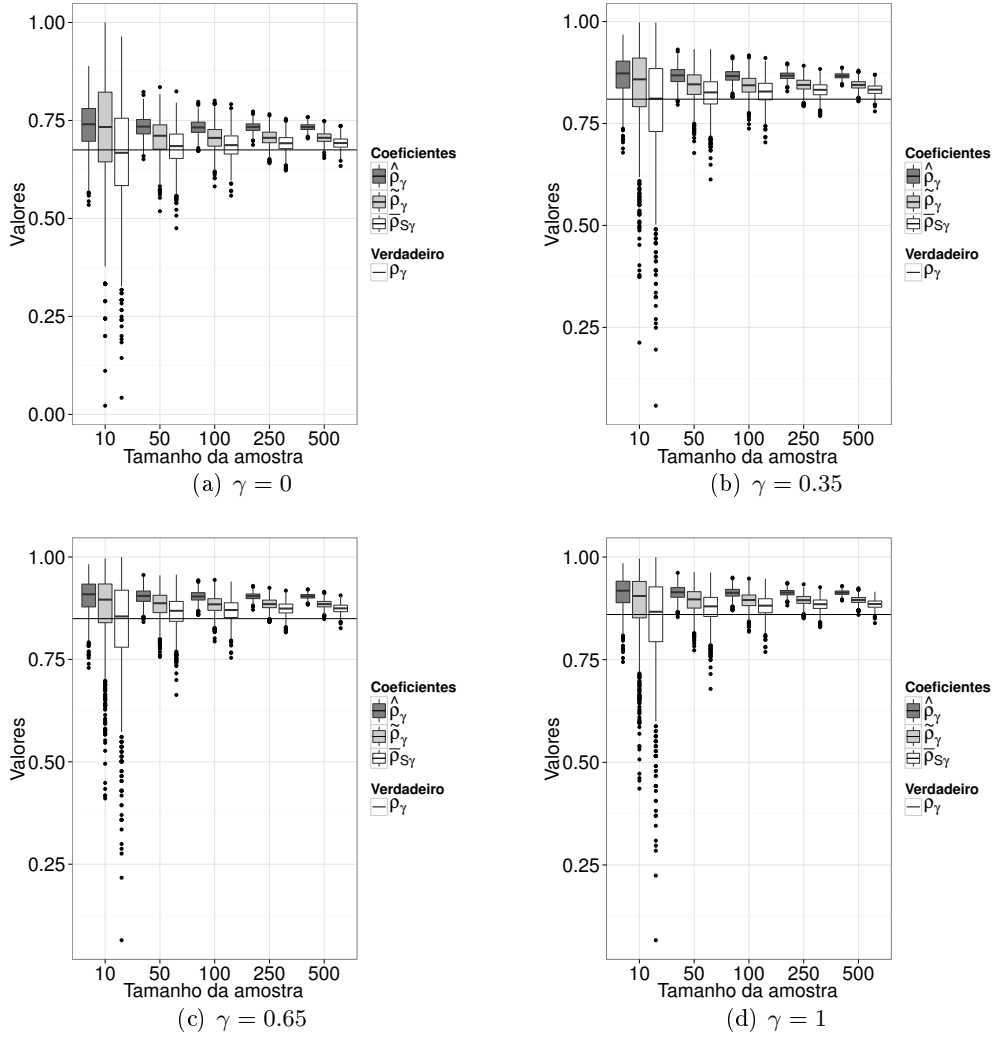


Figura 3.20: Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ para o primeiro caso da mistura de duas normais bivariadas.

Caso 2: $Q_{\underline{X}} = 0.9 \times \mathcal{N}(0, 0; 1, 1; 0.9) + 0.1 \times \mathcal{N}(0, 0; 1, 1; 0)$

Neste caso podemos observar na Figura 3.21 que todos os estimadores tendem a superestimar o verdadeiro valor ρ_γ , porém o CCGS possui, em média, melhores estimativas quando temos amostras de tamanho 10 e $\gamma = 0$. Aumentando o tamanho da amostra este estimador tende a divergir do verdadeiro valor ρ_γ para todos os valores de γ .

Observando o CCGMV notamos que possui uma maior superestimação do verdadeiro valor ρ_γ para todos os tamanhos da amostra e todos os valores de γ . Para o CCGU observamos que também tende a superestimar o verdadeiro valor ρ_γ porém com menor intensidade do que o CCGMV. Este comportamento permanece para

Tabela 3.8: Raiz do erro quadrático médio das estimativas do coeficiente de correlação generalizado $\hat{\rho}_\gamma$, $\tilde{\rho}_\gamma$ e $\bar{\rho}_{S_\gamma}$ para o primeiro caso da mistura de duas distribuições normais bivariadas.

γ	Método	Tamanho da Amostra				
		10	50	100	250	500
0	$\hat{\rho}_\gamma$	0.0879	0.0651	0.0610	0.0598	0.0589
	$\tilde{\rho}_\gamma$	0.1438	0.0561	0.0439	0.0371	0.0342
	$\bar{\rho}_{S_\gamma}$	0.1337	0.0495	0.0362	0.0272	0.0235
0.35	$\hat{\rho}_\gamma$	0.0760	0.0615	0.0588	0.0583	0.0577
	$\tilde{\rho}_\gamma$	0.1073	0.0499	0.0422	0.0386	0.0368
	$\bar{\rho}_{S_\gamma}$	0.1306	0.0456	0.0348	0.0291	0.0269
0.65	$\hat{\rho}_\gamma$	0.0683	0.0575	0.0556	0.0553	0.0549
	$\tilde{\rho}_\gamma$	0.0922	0.0458	0.0402	0.0379	0.0366
	$\bar{\rho}_{S_\gamma}$	0.1249	0.0423	0.0332	0.0292	0.0277
1	$\hat{\rho}_\gamma$	0.0657	0.0560	0.0542	0.0540	0.0536
	$\tilde{\rho}_\gamma$	0.0873	0.0443	0.0393	0.0375	0.0364
	$\bar{\rho}_{S_\gamma}$	0.1224	0.0409	0.0325	0.0291	0.0278

todos os valores de γ e para todos os tamanhos da amostra.

Na Tabela 3.9 observamos que para amostras de tamanho 10 CCGMV é o que apresenta os menores valores para todos os valores de γ . Quando aumentamos o tamanho da amostra percebemos uma pequena diminuição da média quadrática dos três estimadores, principalmente a partir das amostras de tamanho 50.

Tabela 3.9: Raiz do erro quadrático médio das estimativas do coeficiente de correlação generalizado $\hat{\rho}_\gamma$, $\tilde{\rho}_\gamma$ e $\bar{\rho}_{S_\gamma}$ para o segundo caso da mistura de duas distribuições normais bivariadas.

γ	Método	Tamanho da Amostra				
		10	50	100	250	500
0	$\hat{\rho}_\gamma$	0.1097	0.0892	0.0864	0.0849	0.0840
	$\tilde{\rho}_\gamma$	0.1528	0.0735	0.0649	0.0594	0.0574
	$\bar{\rho}_{S_\gamma}$	0.1362	0.0596	0.0517	0.0464	0.0446
0.35	$\hat{\rho}_\gamma$	0.1097	0.0982	0.0966	0.0959	0.0953
	$\tilde{\rho}_\gamma$	0.1254	0.0803	0.0761	0.0736	0.0726
	$\bar{\rho}_{S_\gamma}$	0.1339	0.0672	0.0634	0.0617	0.0611
0.65	$\hat{\rho}_\gamma$	0.1075	0.0997	0.0986	0.0982	0.0978
	$\tilde{\rho}_\gamma$	0.1147	0.0825	0.0799	0.0785	0.0779
	$\bar{\rho}_{S_\gamma}$	0.1290	0.0701	0.0682	0.0678	0.0676
1	$\hat{\rho}_\gamma$	0.1066	0.0999	0.0990	0.0986	0.0983
	$\tilde{\rho}_\gamma$	0.1114	0.0832	0.0810	0.0800	0.0794
	$\bar{\rho}_{S_\gamma}$	0.1269	0.0711	0.0697	0.0697	0.0696

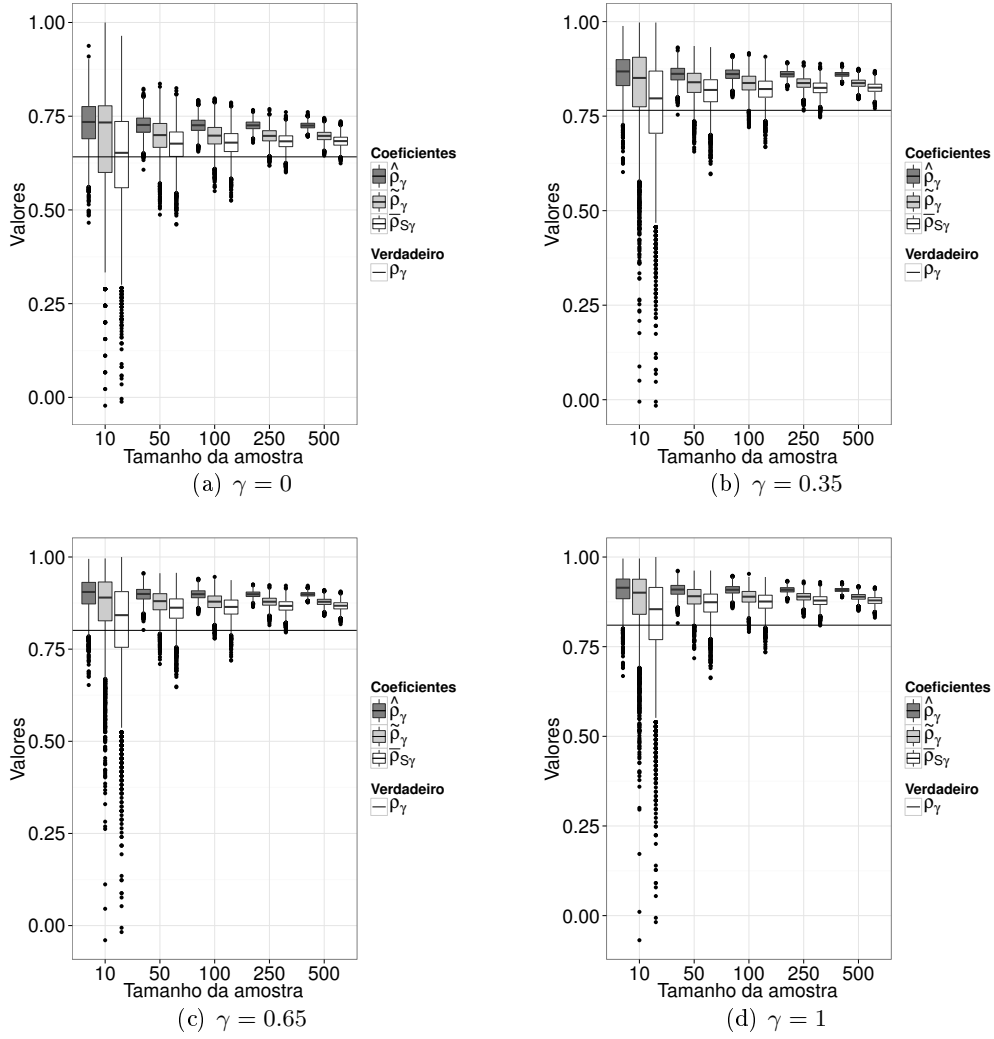


Figura 3.21: Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ para o segundo caso da mistura de duas normais bivariadas.

Caso 3: $Q_{\underline{X}} = 0.9 \times \mathcal{N}(0, 0; 1, 1; 0.9) + 0.1 \times \mathcal{N}(0, 0; 1, 1; -0.9)$

Pelos gráficos boxplot da Figura 3.22 observamos que os três estimadores tendem a superestimar o verdadeiro valor ρ_γ para todos os valores de γ e para todos os tamanhos da amostra. Notamos também um comportamento distribucional bastante assimétrico do CCGU e CCGS quando temos amostras de tamanho 10. Tendência esta que se torna mais evidente quando aumentamos o valor de γ .

Na Tabela 3.10 apresentamos a média quadrática dos três estimadores e notamos que mesmo aumentando o tamanho da amostra os valores da médias quadráticas sofrem pouca alteração. Observamos também que o CCGS é o que apresenta os menores valores da média quadrática em todos os valores de γ e para todos os

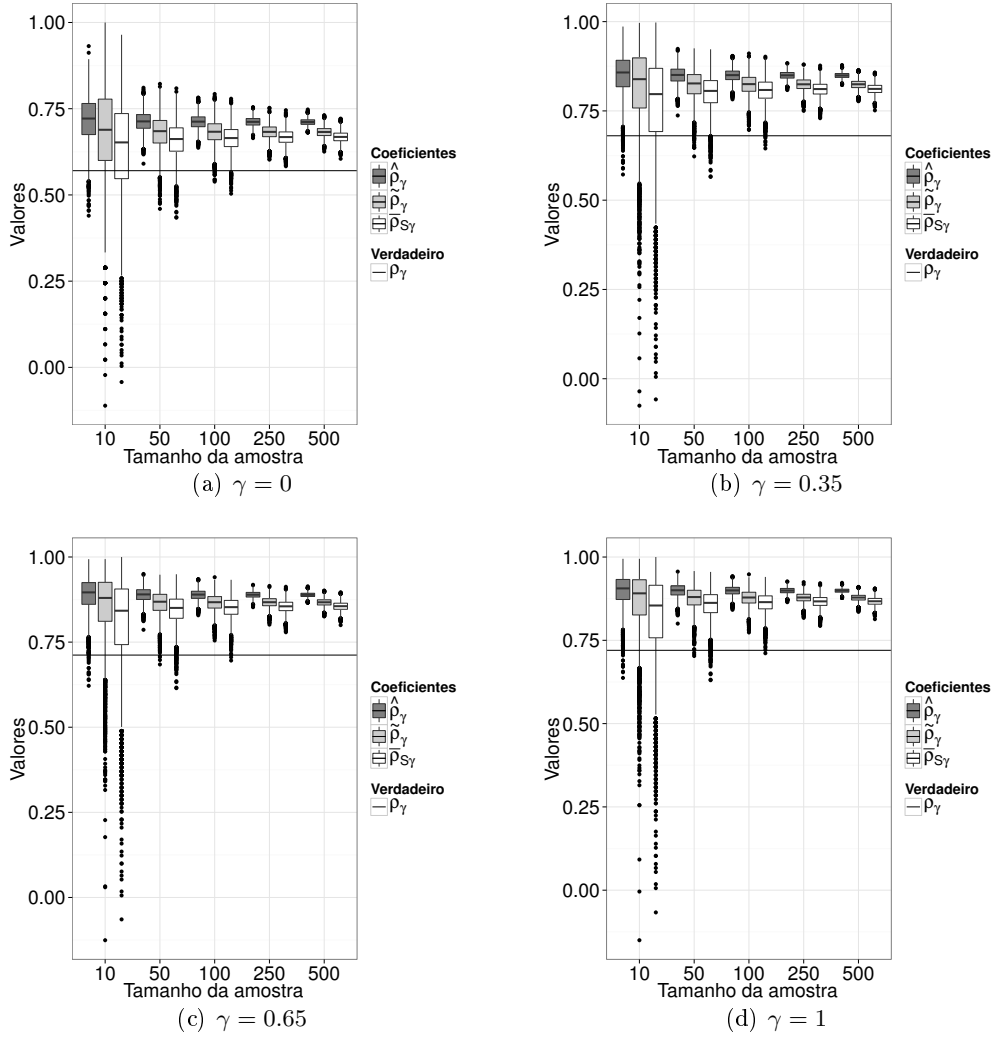


Figura 3.22: Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ para o terceiro caso da mistura de duas normais bivariadas.

tamanhos da amostra. Analisando o CCGMV e fixando o tamanho da amostra podemos notar que ao aumentarmos o valor de γ sua média quadrática tende a aumentar, ou seja, sua variabilidade torna-se maior a medida que aumentamos o valor de γ .

Caso 4: $Q_{\underline{X}} = 0.6 \times \mathcal{N}(0, 0; 1, 1; 0.9) + 0.4 \times \mathcal{N}(0, 0; 1, 1; 0)$

Pelos gráficos boxplot da Figura 3.23 notamos que quando temos $\gamma = 0$ o CCGS tende a subestimar o verdadeiro valor ρ_γ em todos os tamanho da amostra. Além disso, apresenta um comportamento distribucional com maior variabilidade quando temos amostras de tamanho 10. Ainda neste cenário observamos também que CCGU tende a superestimar o verdadeiro valor ρ_γ em amostras de tamanho 10, porém

Tabela 3.10: Raiz do erro quadrático médio das estimativas do coeficiente de correlação generalizado $\hat{\rho}_\gamma$, $\tilde{\rho}_\gamma$ e $\bar{\rho}_{S_\gamma}$ para o terceiro caso da mistura de duas distribuições normais bivariadas.

γ	Método	Tamanho da Amostra				
		10	50	100	250	500
0	$\hat{\rho}_\gamma$	0.1621	0.1454	0.1433	0.1422	0.1415
	$\tilde{\rho}_\gamma$	0.1853	0.1225	0.1173	0.1142	0.1129
	$\bar{\rho}_{S_\gamma}$	0.1524	0.1027	0.1007	0.0996	0.0991
0.35	$\hat{\rho}_\gamma$	0.1798	0.1710	0.1700	0.1695	0.1690
	$\tilde{\rho}_\gamma$	0.1770	0.1482	0.1460	0.1448	0.1442
	$\bar{\rho}_{S_\gamma}$	0.1623	0.1301	0.1308	0.1317	0.1317
0.65	$\hat{\rho}_\gamma$	0.1839	0.1783	0.1776	0.1773	0.1770
	$\tilde{\rho}_\gamma$	0.1749	0.1571	0.1559	0.1554	0.1550
	$\bar{\rho}_{S_\gamma}$	0.1638	0.1405	0.1420	0.1435	0.1438
1	$\hat{\rho}_\gamma$	0.1848	0.1801	0.1795	0.1793	0.1790
	$\tilde{\rho}_\gamma$	0.1745	0.1597	0.1588	0.1585	0.1582
	$\bar{\rho}_{S_\gamma}$	0.1640	0.1436	0.1454	0.1471	0.1474

quando aumentamos o tamanho da amostra este estimador tende a convergir ao verdadeiro valor ρ_γ . Contudo, ainda com $\gamma = 0$, notamos que CCGMV tende a superestimar o verdadeiro valor ρ_γ em todos os tamanhos da amostra. Observamos também um comportamento bastante assimétrico deste estimador quando temos amostras de tamanho 10. Ao aumentarmos o valor de γ percebemos que todos os estimadores tendem a superestimar o verdadeiro valor ρ_γ .

Na Tabela 3.11 observamos que o CCGMV apresenta os maiores valores das médias quadráticas para todos os valores de γ quando temos amostras de tamanho maior que 10. Para $\gamma = 0$ e amostras de tamanho maior que 10, o CCGU apresenta os menores valores da média quadrática. Notamos também que os três estimadores tendem a aumentar suas respectivas médias quadráticas à medida que aumentamos o valor de γ .

Caso 5: $Q_{\tilde{X}} = 0.2 \times \mathcal{N}(0, 0; 1, 1; 0.9) + 0.8 \times \mathcal{N}(0, 0; 1, 1; 0)$

Nos gráficos da Figura 3.24 podemos observar um comportamento bastante semelhante com o quarto caso da mistura de normais. Quando temos $\gamma = 0$ e amostras de tamanho 10, o CCGU tende a superestimar o verdadeiro valor ρ_γ mas à medida que aumentamos o tamanho da amostra este estimador tende a convergir para o verdadeiro valor ρ_γ . Analisando o CCGS, com $\gamma = 0$, tende a subestimar o verdadeiro valor ρ_γ para todos os tamanhos da amostra. Neste mesmo cenário, o CCGMV tende a superestimar o verdadeiro valor ρ_γ em todos os tamanhos da amostra.

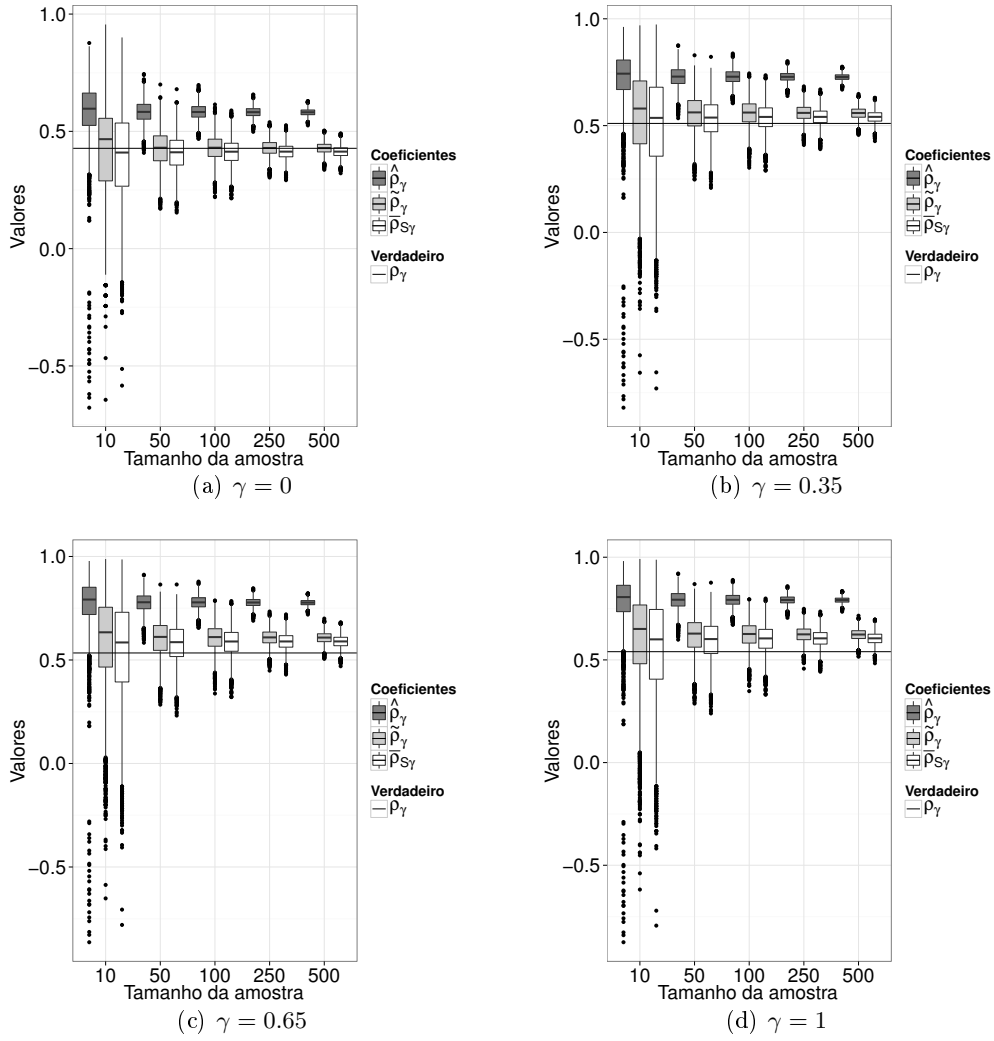


Figura 3.23: Gráficos boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ para o quarto caso da mistura de duas normais bivariadas.

Quando aumentamos o valor de γ todos os estimadores tendem a superestimar o verdadeiro valor ρ_γ , porém CCGS apresenta a menor tendência de superestimação. Para amostras de tamanho 10 o CCGMV apresenta um comportamento distribucional bastante assimétrico, enquanto que o CCGU e o CCGS apresentam uma maior variabilidade. Esta variabilidade tende a diminuir à medida que aumentamos os tamanhos da amostra.

Observando a média quadrática dos três estimadores na Tabela 3.12 temos que os maiores valores ocorrem em amostras de tamanho 10. A medida que aumentamos o tamanho da amostra estes valores tendem a diminuir. O CCGMV é o que apresenta os maiores valores da média quadrática quando temos amostras de tamanho menor que 250 para todos os valores de γ . Para amostras de tamanho 250 e 500 o CCGS

Tabela 3.11: Raiz do erro quadrático médio das estimativas do coeficiente de correlação generalizado $\hat{\rho}_\gamma$, $\tilde{\rho}_\gamma$ e $\bar{\rho}_{S_\gamma}$ para o quarto caso da mistura de duas distribuições normais bivariadas.

γ	Método	Tamanho da Amostra				
		10	50	100	250	500
0	$\hat{\rho}_\gamma$	0.2017	0.1618	0.1582	0.1553	0.1540
	$\tilde{\rho}_\gamma$	0.2068	0.0774	0.0554	0.0336	0.0241
	$\bar{\rho}_{S_\gamma}$	0.1931	0.0787	0.0574	0.0365	0.0282
0.35	$\hat{\rho}_\gamma$	0.2552	0.2227	0.2205	0.2186	0.2176
	$\tilde{\rho}_\gamma$	0.2237	0.0983	0.0786	0.0619	0.0553
	$\bar{\rho}_{S_\gamma}$	0.2294	0.0939	0.0713	0.0503	0.0417
0.65	$\hat{\rho}_\gamma$	0.2749	0.2469	0.2455	0.2440	0.2432
	$\tilde{\rho}_\gamma$	0.2309	0.1127	0.0963	0.0836	0.0786
	$\bar{\rho}_{S_\gamma}$	0.2415	0.1054	0.0857	0.0690	0.0626
1	$\hat{\rho}_\gamma$	0.2810	0.2546	0.2534	0.2521	0.2514
	$\tilde{\rho}_\gamma$	0.2334	0.1184	0.1031	0.0919	0.0873
	$\bar{\rho}_{S_\gamma}$	0.2450	0.1101	0.0918	0.0767	0.0709

apresenta os maiores valores da média quadrática.

Neste estudo notamos que o CCGU é uma ótima alternativa em diversos casos. Quando temos amostras sem perturbação o CCGU possui um comportamento bastante semelhante ao CCGMV apesar de apresentar uma maior variabilidade. No caso em que temos perturbação da amostra, estamos avaliando o comportamento do CCGU na presença de **outliers**, ou seja, sua robustez, e percebemos que para amostras pequenas e $\gamma = 0$ este estimador tende a convergir para o verdadeiro valor ρ_γ principalmente quando temos $\rho = 0.1$ e $\rho = 0.5$. Nestes cenários observamos que o CCGS tende a subestimar o verdadeiro valor ρ_γ em todos os tamanhos da amostra. Por fim, no caso em que temos contaminação da amostra, o CCGMV não é necessariamente único (multimodalidade da função de log-verossimilhança) e por isso pode apresentar resultados não tão satisfatórios, já que podemos estar analisando um máximo local e não global.

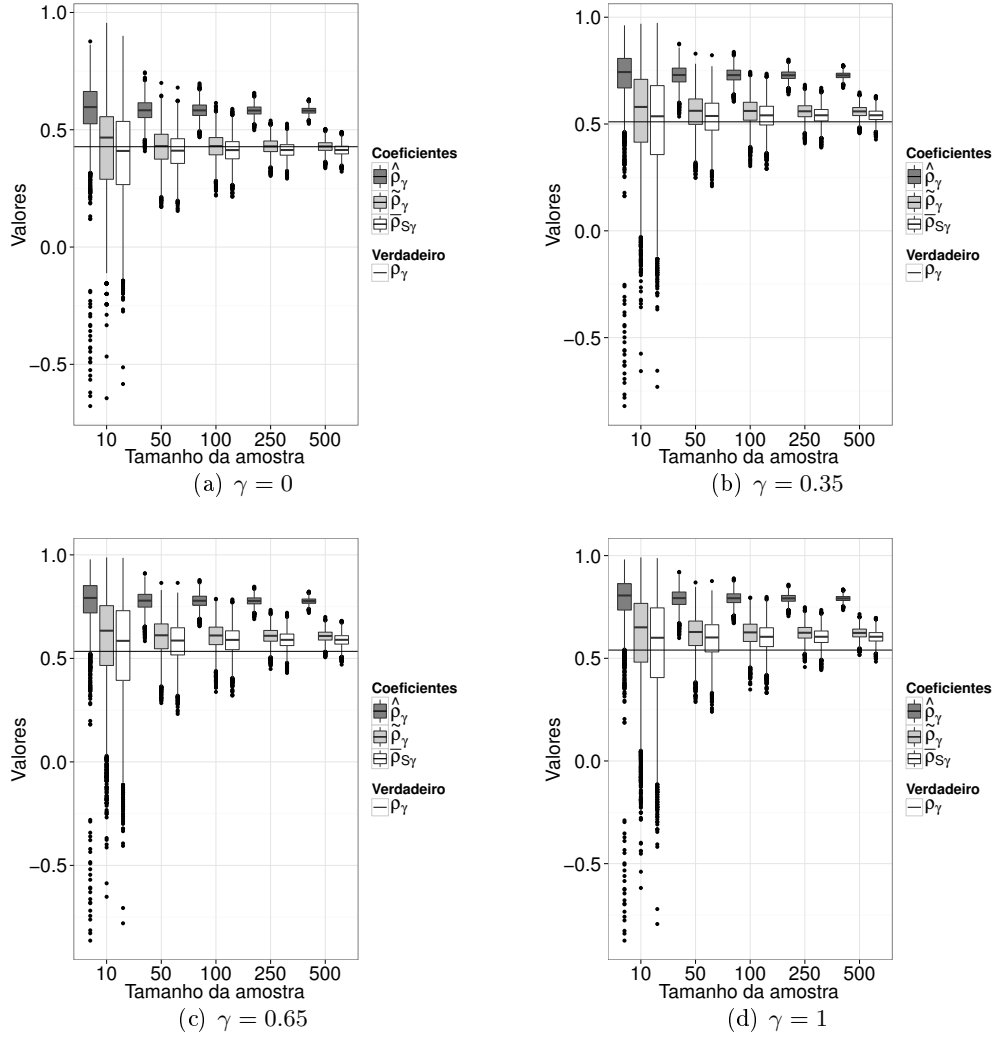


Figura 3.24: Gráfico Boxplot para a comparação dos estimadores CCGMV, CCGU e CCGS de ρ_γ para o quinto caso da mistura de duas normais bivariadas.

Tabela 3.12: Raiz do erro quadrático médio das estimativas do coeficiente de correlação generalizado $\hat{\rho}_\gamma$, $\tilde{\rho}_\gamma$ e $\bar{\rho}_{S_\gamma}$ para o quinto caso da mistura de duas distribuições normais bivariadas.

γ	Método	Tamanho da Amostra				
		10	50	100	250	500
0	$\hat{\rho}_\gamma$	0.3258	0.1807	0.1404	0.1086	0.0955
	$\tilde{\rho}_\gamma$	0.2725	0.1457	0.1278	0.1167	0.1126
	$\bar{\rho}_{S_\gamma}$	0.2518	0.1434	0.1275	0.1174	0.1137
0.35	$\hat{\rho}_\gamma$	0.4190	0.2331	0.1774	0.1308	0.1103
	$\tilde{\rho}_\gamma$	0.3277	0.1773	0.1520	0.1359	0.1299
	$\bar{\rho}_{S_\gamma}$	0.3235	0.1771	0.1532	0.1375	0.1316
0.65	$\hat{\rho}_\gamma$	0.4539	0.2535	0.1914	0.1383	0.1142
	$\tilde{\rho}_\gamma$	0.3492	0.1880	0.1589	0.1405	0.1337
	$\bar{\rho}_{S_\gamma}$	0.3506	0.1890	0.1613	0.1428	0.1357
1	$\hat{\rho}_\gamma$	0.4648	0.2602	0.1959	0.1405	0.1151
	$\tilde{\rho}_\gamma$	0.3558	0.1914	0.1604	0.1414	0.1344
	$\bar{\rho}_{S_\gamma}$	0.3592	0.1926	0.1636	0.1440	0.1365

Capítulo 4

Aplicação

Neste capítulo apresentamos uma aplicação do uso do coeficiente de correlação generalizado e do coeficiente de correlação de Spearman em uma base de dados real. A base de dados para análise é resultado do projeto de pesquisa da FAPESP nº 06/03227-2 intitulado “Expressão gênica em tumores do estômago e do esôfago: da biologia ao diagnóstico”. Os dados foram obtidos através da parceria do Profº Dr. Gustavo Henrique Esteves, da Universidade Estadual da Paraíba, e do Profº Phd Luiz Fernando Lima Reis, Diretor de pesquisa do Hospital Sírrio-Libânes e coordenador do projeto.

Neste projeto foram coletados dados de biópsias de cerca de mil pacientes com problemas (ou suspeitas de problemas) em tecidos do esôfago e estômago associados ao câncer. Um dos focos do projeto era avaliar o perfil de expressão gênica entre os diferentes tipos de tecidos com o uso de experimentos de microarray. Neste caso, foram usados chips da plataforma Agilent (para mais detalhes sobre esta lâmina acesse http://www.genomics.agilent.com/article.jsp?pageId=1516&_requestid=2409114).

Após uma seleção dessa coleta foram utilizadas 146 amostras de tecidos gastroesofágicos a fim de identificar assinaturas moleculares capazes de classificar precisamente as amostras em mucosa normal, mucosa com inflamação e mucosa com metaplasia (processo inflamatório crônico que leva a substituição da mucosa normal por um tecido colunar do tipo intestinal). Por questões de segurança estes dados não foram publicados e também não foram disponibilizados ao público. A Tabela 4.1 apresenta a classificação das amostras coletadas do estômago, esôfago e do TEG (Transição esôfago-gástrica) em mucosa normal, mucosa com inflamação e mucosa com metaplasia.

Nosso intuito é colaborar com os pesquisadores da área na avaliação da interação

Tabela 4.1: Amostras de tecidos gastroesofágicos classificadas em normal, com inflamação e com metaplasia.

	Estômago	Esôfago	TEG	Total
Normal	57	12	4	73
Inflamação	16	0	0	16
Metaplasia	42	8	7	57
Total	115	20	11	146

gênica entre duas variáveis (moléculas). Uma abordagem para avaliar tais interações é através da construção de redes de expressão gênica interagentes e conhecidas como redes de relevância, introduzidas por Butte e Kohane (1999). Este método consiste em calcular o quadrado do coeficiente de correlação linear de Pearson, r^2 , entre todos os pares de genes para cada tipo biológico, definindo um grafo completamente interligado. Em seguida, é definido um ponto de corte $r^{2'}$ afim de dividir o grafo em pequenos sub-grafos em que $r^2 > r^{2'}$. Estes sub-grafos são denominados por redes de relevância.

Para a construção das redes de relevância neste estudo utilizamos a amostra de 57 tecidos do estômago classificados como normal (ver Tabela 4.1). Normalmente utilizamos o coeficiente de correlação de Pearson para construção destas redes, porém nesta aplicação usamos o coeficiente de correlação generalizado e o coeficiente de correlação de Spearman. Empiricamente, verificamos que é possível assumir que os dados de expressão gênica são normais. Por exemplo, através de uma amostra aleatória construímos um gráfico em que observamos o histograma de frequências relativas, o gráfico da densidade estimada comparada com a curva da distribuição normal, como podemos ver na Figura 4.1 a suposição de normalidade parece ser razoável.

Na construção das redes de relevância, através do coeficiente de correlação generalizado, definimos os valores de γ em 1, 0.86, 0.71, 0.57, 0.43, 0.29, 0.14 e 0. Definimos os valores de γ a fim de observar o comportamento da rede quando usamos estes valores. Em seguida, por analogia ao trabalho de Butte e Kohane (1999), definimos o ponto de corte $r^{2'} = 0.5$ e construímos os respectivos sub-grafos para o tipo de tecido selecionado. A Figura 4.2 mostra as redes de relevância construídas através do coeficiente de correlação generalizado e do coeficiente de correlação de Spearman. Podemos observar que a medida que diminuimos o tamanho de γ os grafos tendem a diminuir a quantidade de interações entre os genes. Somente o especialista da área poderá decidir se tais redes e interações fazem sentido biológico. Observando o

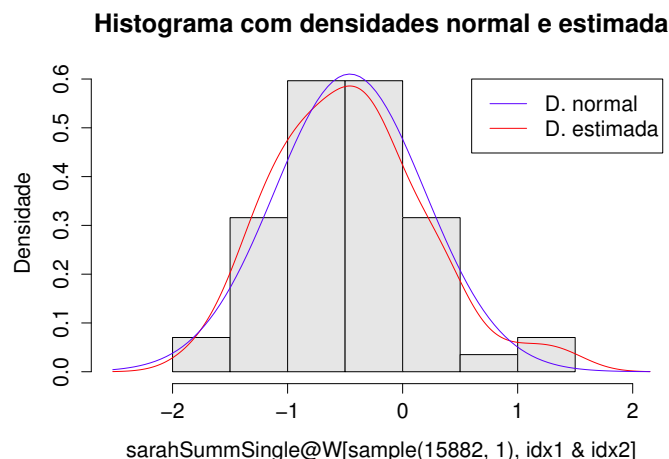


Figura 4.1: Gráfico empírico para visualização da normalidade dos dados.

coeficiente de correlação de Spearman notamos características bastante semelhantes com o coeficiente de correlação generalizado quando temos $\gamma = 0.71$.

Na Figura 4.3 apresentamos as mesmas informações da Figura 4.2, porém este tipo de gráfico é a versão mais utilizada entre os pesquisadores das áreas biológicas.

Usualmente utilizamos o coeficiente de correlação de Pearson, coeficiente de correlação generalizado quando $\gamma = 1$, na construção das redes de relevância mas nem sempre conseguimos apresentar interações moleculares significativas no aspecto biológico. Contudo, através do coeficiente de correlação generalizado podemos variar o valor de γ mostrando diversos cenários a um especialista tentando entender as interações moleculares em cada sub-grafo formado.

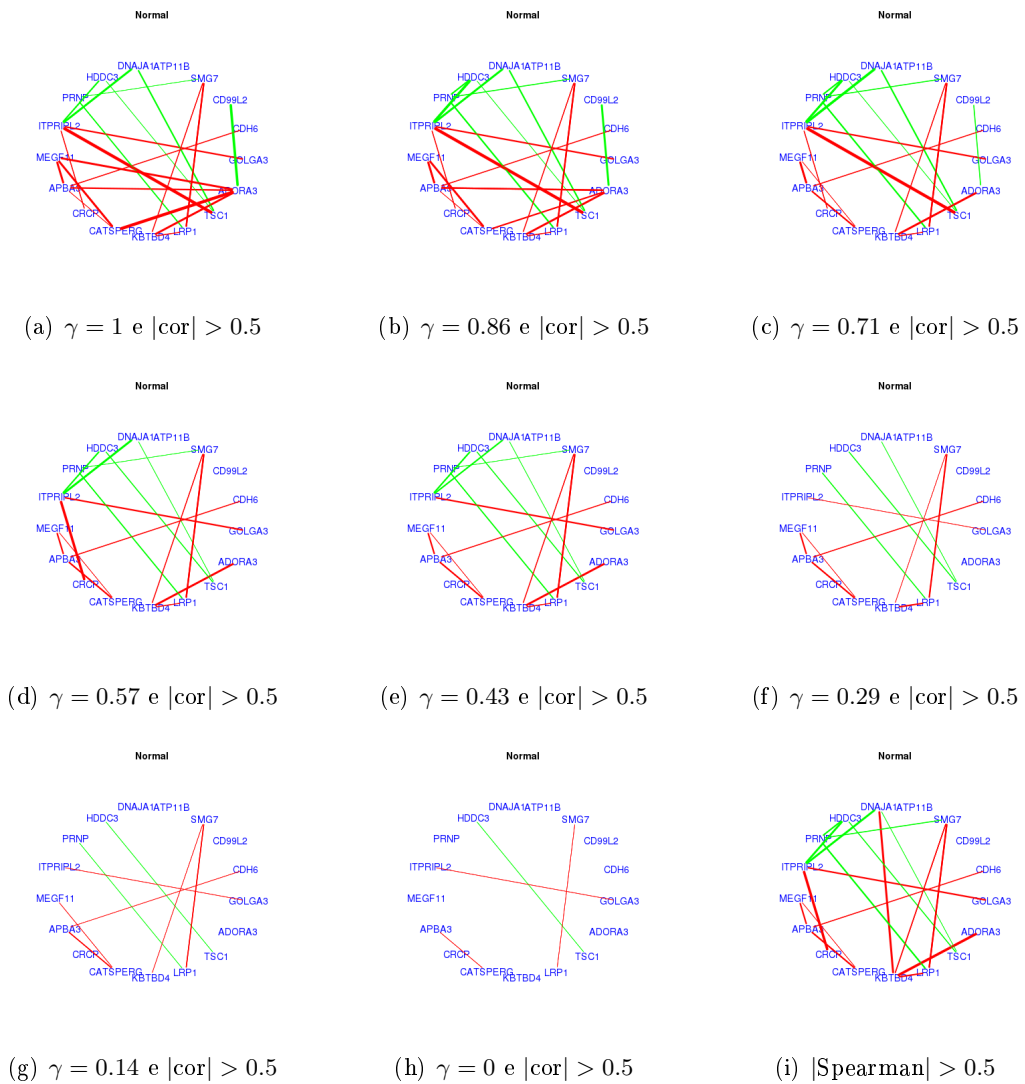


Figura 4.2: Gráfico (redes de relevância) usando o coeficiente de correlação generalizado e o de Spearman para diferentes valores de γ usando o pacote maigesPack.

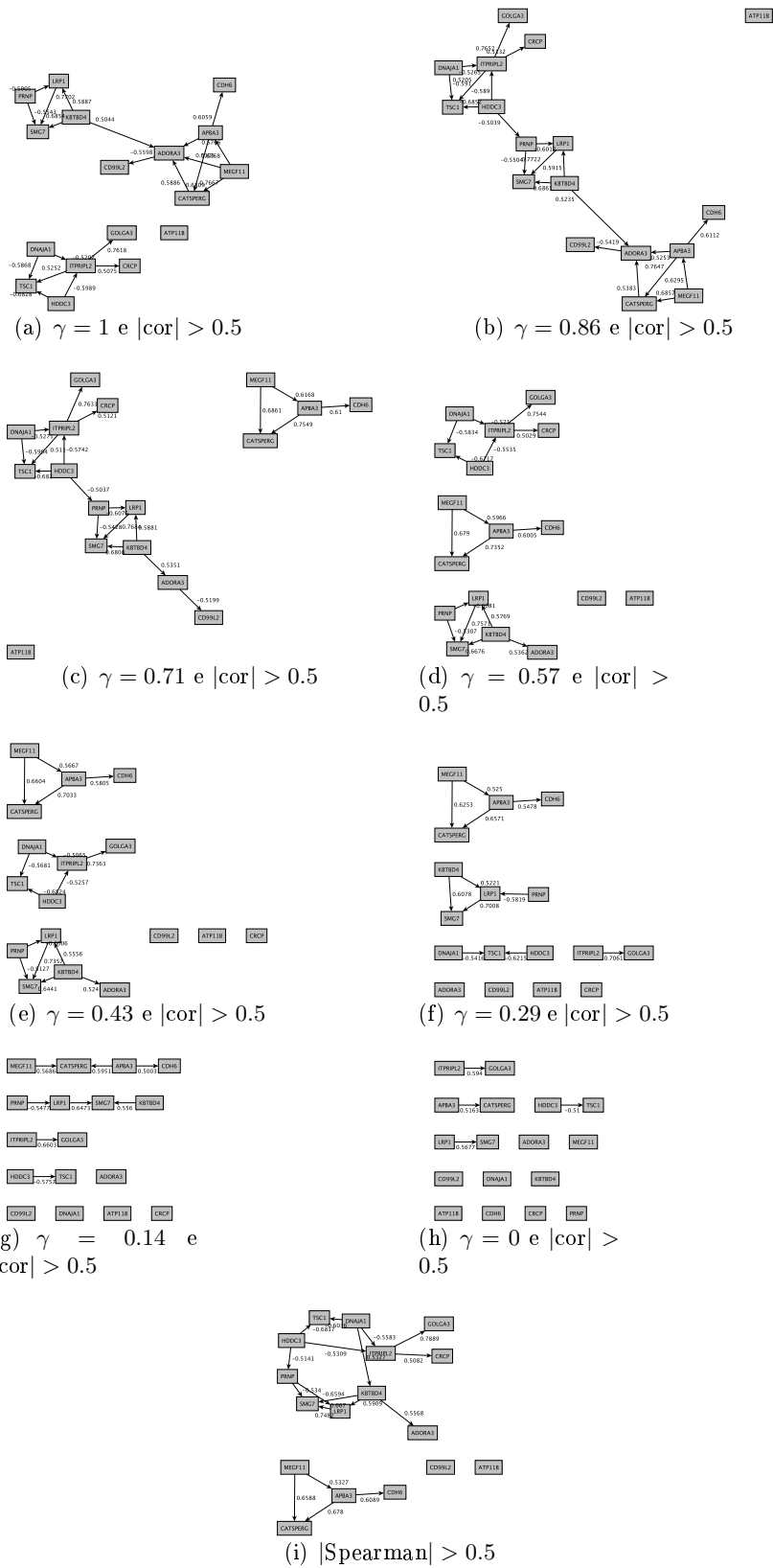


Figura 4.3: Gráfico (redes de relevância) usando o coeficiente de correlação generalizado e o de Spearman para diferentes valores de γ usando o editor de gráficos yEd.

Capítulo 5

Considerações Finais

Esta dissertação tem por objetivo estudar o comportamento do estimador do coeficiente de correlação generalizado proposto por Chinchilli *et al.* (2005) em diversos cenários comparando-o com o estimador do coeficiente de correlação por máxima verossimilhança e do coeficiente de correlação de Spearman que propomos. Definimos os seguintes cenários para avaliação: amostra sem contaminação, amostras com perturbação nos níveis 10%, 30% e 50% e amostras contaminadas pela mistura de duas distribuições normais bivariadas.

Para o cenário em que temos amostras sem contaminação definimos 4 casos. Nos 4 casos observamos que quando temos $\rho = 0$ os três estimadores apresentam características semelhantes tanto na convergência para o verdadeiro valor como na variabilidade das estimativas. Quando temos ρ igual a 0.3 e 0.9 e amostras de tamanho 10 observamos que o estimador do coeficiente de correlação generalizado por máxima verossimilhança e estimador do coeficiente de correlação generalizado tendem a superestimar o verdadeiro valor, porém a medida que aumentamos o tamanho da amostra esta tendência desaparece. Para todos os casos e todos os tamanhos da amostra o estimador do coeficiente de correlação por Spearman tende a subestimar o verdadeiro valor ρ_γ .

No cenário em que abordamos a perturbação da amostra, quando temos $\rho = 0.1$ observamos que o comportamento dos três estimadores é praticamente o mesmo independente do nível de perturbação. Para $\rho = 0.5$ e amostras de tamanho 10, o estimador do coeficiente de correlação generalizado tende a superestimar o verdadeiro valor ρ_γ quando γ igual a 0.5 e 1. Este comportamento é bastante semelhante em todos os níveis de perturbação. O estimador do coeficiente de correlação generalizado por Spearman tende a subestimar o verdadeiro valor ρ_γ em todos os níveis de perturbação e em todos os tamanhos da amostra.

Por último consideramos 5 casos de mistura de normais e notamos que o estimador do coeficiente de correlação generalizado por Spearman tende a superestimar menos do que o estimador do coeficiente de correlação generalizado e o estimador do coeficiente de correlação por máxima verossimilhança na maioria dos casos. Quando temos a mistura de uma distribuição normal bivariada com $\rho = 0.9$ e uma distribuição normal bivariada com $\rho = 0$ e tomamos $\gamma = 0$ notamos que o estimador do coeficiente de correlação generalizado tende a convergir ao verdadeiro valor ρ_γ para amostras de tamanho maior que 10 independente do peso da distribuição usada como distribuição de contaminação.

Na construção das redes de relevância notamos que o coeficiente de correlação generalizado pode ser uma ótima alternativa. Quando alteramos o valor de γ podemos construir diversas redes interagentes e com a ajuda de um especialista esta pode ser uma ótima alternativa na construção de redes de relevância que possam conduzir a resultados biológicos mais confiáveis.

Referências

- [1] BALAKRISHNAN, N. e LAI, C.-D. (2009). *Continuous Bivariate Distributions*. Springer, second edition.
- [2] BUTTE, A. J. e KOHANE, I.S. (1999). *Unsupervised Knowledge discovery in medical databases using relevance networks*. Proc. AMIA Symp., 711-715.
- [3] CHELLINI, P. R. (2007) *Boas Práticas Estatísticas em Estudos de Bioequivalência com Delineamento Crossover 2×2* . Dissertação de Mestrado, Universidade Federal de Minas Gerais.
- [4] CHEN, V. Y.-J.; CHINCHILLI, V. M. e RICHARDS, D. S. P. (2011). *Robustness and monotonicity properties of generalized correlation coefficients*. Journal of Statistical Planning and Inference, **141**, 924-936.
- [5] CHEN, Y.-J. (2006). *Robustness properties of generalized correlation coefficients, with applications to cross-over designs*. Tese de Doutorado, The Pennsylvania State University.
- [6] CHINCHILLI, V. M.; PHILIPS, B. R.; MAUGER, D. T. e SZEFLER, S. J. (2005). *A general class of correlation coefficients for the 2×2 crossover design*. Biometrical Journal, **47**, 644-653.
- [7] CHOK, N. S. (2010). *Pearson's versus Spearman's and Kendall's Correlation Coefficients For Continuous Data*. Dissertação de Mestrado, University of Pittsburgh.
- [8] FISHER, R. A. (1915). *Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population*. Biometrika, **10**, 507-521.

-
- [9] FOSDICK, B. K. e RAFTERY, A. E. (2012). *Estimating the correlation in bivariate normal data with known variances and small sample sizes*. The American Statistician, **66**, 34-41.
- [10] HOEFFDING, W. (1948). *A class of U-statistics with asymptotically normal distribution*. The Annals of Mathematical Statistics, **19**, 293-325.
- [11] HOLLANDER, M. e WOLFE, D. A. (2002). *Nonparametric Statistical Methods, Second Edition*. Wiley, New York.
- [12] HOTELLING, H. (1953). *New light on the correlation coefficient and its transform*. Journal of the Royal Statistical Society. Series B, **15**, 193-232.
- [13] JOHNSON, N. L.; KOTZ, S. e BALAKRISHNAN, N. (1995). *Continuous Univariate Distributions 2 edition*. John Wiley & Sons.
- [14] KENDALL, M. G. (1938). *A New Measure of Rank Correlation*. Biometrika, **1**, 81-93.
- [15] LEE, A. J. (1990). *U-Statistics: Theory and Practice*. Marcel Dekker. New York.
- [16] MCNEIL, A. J., FREY, R., e EMBRECHTS, P. (2005). *Quantitative Risk Management: Concepts, Techniques, Tools*. Princeton University Press.
- [17] OEHLERT, G. W. (1992). *A Note on the Delta Method*. The American Statistician, **46**, 27-29.
- [18] PEARSON, E. S. (1931). *The test of significance for the correlation coefficient*. Journal of the American Statistical Association, **26**, 128-134.
- [19] PEARSON, E. S. (1932). *The test of significance for the correlation coefficient: Some further results*. Journal of the American Statistical Association, **27**, 424-426.
- [20] RAVEH, A. e LESHEM, G. (2011). *Bias in estimation of classical statistical coefficient*. Journal of Statistics and Management Systems, **14:1**, 141-150.
- [21] ROBERT, C. P. e CASELLA, G. (2010) *Introducing Monte Carlo Methods with R*. Springer.
- [22] RODGERS, J. L. e NICEWANDER, W. A. (1988). *Thirteen Ways to Look at the Correlation Coefficient*. The American Statistician, **42**, 59-66.

-
- [23] SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*.
Jonh Wiley & Sons.
- [24] SPEARMAN, C. (1904). *The Proof and Measurement of Association between
Two Things*. The American Journal of Psychology, **15**, 72-101.

Apêndice A

Aspectos importantes

Neste Apêndice apresentamos alguns aspectos importantes para esta dissertação. Primeiramente apresentamos a função densidade da distribuição Normal bivariada a qual utilizamos para avaliar o comportamento dos estimadores propostos. Em seguida, definimos a distribuição *t*-Student Bivariada a qual utilizamos para estudar o comportamento dos estimadores na presença de valores atípicos (**outliers**). Apresentamos também a definição da mistura de distribuições e definimos a mistura de duas normais bivariadas. Por fim discutimos de maneira muito sucinta o método Monte Carlo importantíssimo para a realização das simulações.

A.1 Distribuição Normal Bivariada

Segundo Balakrishnan e Lai (2009) as origens de dados com comportamento distribucional normal bivariado datam da primeira metade do século XIX nos trabalhos de Laplace, Plana, Gauss e Bravais. Francis Galton foi o primeiro a reconhecer a necessidade de uma medida de correlação para dados bivariados a fim de analisar a dependência entre a altura dos pais e de seus filhos. Ele também observou que as distribuições marginais dos dados eram distribuições normais.

A.1.1 Função de Densidade e Distribuição Acumulada da Normal Bivariada

Seja $\mathbf{X} = (X, Y)$ um vetor aleatório seguindo uma distribuição normal acumulada conjunta F_{XY} com $E(\mathbf{X}) = (E(X), E(Y)) = (\mu_X, \mu_Y)$ e matriz de variância-covariância

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \quad (\text{A.1})$$

em que $\sigma_X^2 = \text{Var}(X)$, $\sigma_Y^2 = \text{Var}(Y)$ e

$$\rho = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sqrt{\sigma_X^2 \sigma_Y^2}}$$

é o coeficiente de correlação entre X e Y . A função densidade de probabilidade de $\mathbf{X}$ é definida por

$$\begin{aligned} f(x, y) = & \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_X}{\sigma_X} \right)^2 \right. \right. \\ & \left. \left. - 2\rho \left(\frac{x-\mu_X}{\sigma_X} \right) \left(\frac{y-\mu_Y}{\sigma_Y} \right) + \left(\frac{y-\mu_Y}{\sigma_Y} \right)^2 \right] \right\}. \end{aligned} \quad (\text{A.1})$$

Denotamos, $\mathbf{X} \sim \mathcal{N}_2 \left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \right)$ ou $\mathcal{N}_2(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ a distribuição Normal bivariada de $\mathbf{X}$.

Para Balakrishnan e Lai (2009) não existe uma expressão em termos de funções elementares para a função de distribuição acumulada da normal bivariada. No entanto, para $\underline{x} = (x, y)$ e supondo que $\mathbb{E}(\underline{X}) = (0, 0)$ podemos escrevê-la como sendo

$$F_{\underline{X}}(\underline{x}) = F(x, y) = \frac{ab\sqrt{1-\rho^2}}{2\pi} \int_{-\infty}^x \int_{-\infty}^y e^{-\frac{1}{2}(a^2u^2 - 2ab\rho uv + b^2v^2)} dudv, \quad (\text{A.2})$$

em que $1/a = \sigma_X\sqrt{1-\rho^2}$, $1/b = \sigma_Y\sqrt{1-\rho^2}$, que em algumas situações facilitam o cálculo.

A.2 Distribuição t -Student Bivariada

Seja $\mathbf{X} = (X, Y)$ um vetor aleatório seguindo uma distribuição t -Student bivariada, definimos sua densidade por

$$f(\mathbf{x}) = \frac{\Gamma((\nu+2)/2)}{\Gamma(\nu/2)(\nu\pi)^{|\Sigma|^{1/2}}} \left(1 + \frac{(\mathbf{x} - \mathbb{E}[\mathbf{x}])^\top \Sigma^{-1} (\mathbf{x} - \mathbb{E}[\mathbf{x}])}{\nu} \right)^{-\frac{(\nu+2)}{2}} \quad (\text{A.3})$$

em que $\mathbb{E}(\mathbf{X}) = (\mu_X, \mu_Y)$, $|\Sigma|$ representa o determinante da matriz positiva Σ definida em (A.1), $\Gamma(\cdot)$ é a função gama e ν são os graus de liberdade da distribuição t -Student (McNeil *et al*, 2005). Para que $E(\mathbf{X}) < \infty$ consideramos $\nu > 1$.

A.3 Mistura de Distribuições

Definição A.3.1. Dado um conjunto finito de funções de distribuição acumuladas $F_1(x), F_2(x), \dots, F_m(x)$ e pesos $\varepsilon_1, \dots, \varepsilon_m$ em que $\varepsilon_i \geq 0$ e $\sum_{i=1}^m \varepsilon_i = 1$. A distri-

buição acumulada da mistura finita é uma combinação linear convexa da forma

$$F(x) = \sum_{i=1}^m \varepsilon_i F_i(x). \quad (\text{A.4})$$

Para $\mathbf{X}_i = (X_i, Y_i)$ a mistura de m distribuições normais bivariadas avaliada no ponto $\mathbf{x} = (x, y)$ é

$$Q(\mathbf{x}) = \varepsilon_1 F_{\mathbf{X}_1}(\mathbf{x}) + \cdots + \varepsilon_m F_{\mathbf{X}_m}(\mathbf{x}), \quad (\text{A.5})$$

em que $F_{\mathbf{X}_i}(\mathbf{x})$ são funções de distribuição normal bivariada com vetor de médias $E(\mathbf{X}_i) = (\mu_{X_i}, \mu_{Y_i})$, coeficiente de correlação ρ_i e matriz de variância e covariância dada por $\Sigma_i = \begin{pmatrix} \sigma_{X_i}^2 & \rho_i \sigma_{X_i} \sigma_{Y_i} \\ \rho_i \sigma_{X_i} \sigma_{Y_i} & \sigma_{Y_i}^2 \end{pmatrix}$.

No caso de duas distribuições, a mistura é definida como

$$Q(\mathbf{x}) = (1 - \varepsilon) F_{\mathbf{X}_1}(\mathbf{x}) + \varepsilon F_{\mathbf{X}_2}(\mathbf{x}) \quad (\text{A.6})$$

tomando como base $F_{\mathbf{X}_1}(\mathbf{x})$ temos que o segundo termo desta expressão representa a “contaminação” de $F_{\mathbf{X}_1}(\mathbf{x})$ pela fração ε de $F_{\mathbf{X}_2}(\mathbf{x})$. $Q(\mathbf{x})$ é também chamada de distribuição contaminada.

A.4 Método de Monte Carlo

Dois dos principais problemas que surgem na inferência estatística envolvem problemas de integração e de otimização. Por exemplo, nem sempre é possível calcular analiticamente os estimadores associados aos parâmetros (máxima verossimilhança, método dos momentos, etc.) ou avaliar as suas propriedades distribucionais como a consistência e eficiência que envolvem cálculos de integrais e, assim, muitas vezes devemos considerar soluções numéricas. Uma abordagem geral é o uso de simulação para calcular estas quantidades de interesse. Através destas simulações podemos utilizar resultados assintóticos e probabilísticos, como por exemplo a Lei dos Grandes Números e o Teorema do Limite Central, a fim de avaliar, por exemplo, a convergência dos estimadores.

O problema genérico e abordado no método Monte Carlo é como avaliar a esperança matemática (integral)

$$E_f[h(X)] = \int_{\Delta} h(x) f(x) dx, \quad (\text{A.7})$$

em que h é uma função mensurável e Δ determina a região em que a variável aleatória X assume seus valores o que geralmente coincide com o suporte da densidade

f (Robert e Casella, 2010). O princípio de Monte Carlo para a aproximação de (A.7) consiste em gerar uma amostra $(X_1, \dots, X_n)$ da densidade f e propor como aproximação da integral a média empírica

$$\bar{h}_n = \frac{1}{n} \sum_{j=1}^n h(x_j),$$

em que x_j é o valor observado de X_j e desde que $\bar{h}_n$ convirja quase certamente (ou seja, para quase todas as sequências geradas) para $E_f[h(X)]$.

Seja uma amostra $X_1, X_2, \dots, X_n$ obtida através de algum modelo estatístico indexado pelo parâmetro θ , isto é, $X_i \sim F_\theta(x)$. O nosso objetivo é obter informações de θ através da amostra. Consideramos $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ uma estimativa de θ e suponhamos que θ representa o parâmetro de localização da distribuição, isto é

$$\theta = E(X) = \int x dF_\theta(x),$$

a aproximação desta integral pode ser feita da seguinte forma

- (1) Para $i = 1, 2, \dots, N$ geram-se n amostras de tamanho N

$$(X_1^{(i)}, X_2^{(i)}, \dots, X_n^{(i)}).$$

- (2) Calcula-se

$$\hat{\theta}^{(i)} = \hat{\theta}(X_1^{(i)}, X_2^{(i)}, \dots, X_n^{(i)})$$

para $i = 1, 2, \dots, N$;

- (3) O valor aproximado de θ pelo método de Monte Carlo é dado por

$$\theta \approx \hat{\theta} = \frac{1}{N} \sum_{i=1}^N \hat{\theta}^{(i)}.$$

Note que $\hat{\theta}$ é também uma função dos dados, logo $\hat{\theta}$ é uma variável aleatória. Podemos calcular o viés do estimador de θ (erro de aproximação) por

$$\text{viés}(\hat{\theta}) \approx \widehat{\text{viés}}(\hat{\theta}) = \frac{1}{N} \sum_{i=1}^N (\hat{\theta}^{(i)} - \theta).$$

Uma medida da qualidade de $\hat{\theta}$ é o Erro Quadrático Médio (EQM), o qual é definido por

$$\begin{aligned} \text{EQM}(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] = E[(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2] \\ &= E\{[\hat{\theta} - E(\hat{\theta})]^2\} + 2E\{[\hat{\theta} - E(\hat{\theta})][E(\hat{\theta}) - \theta]\} + E\{[E(\hat{\theta}) - \theta]^2\} \\ &= E\{[\hat{\theta} - E(\hat{\theta})]^2\} + E\{[E(\hat{\theta}) - \theta]^2\} = \text{Var}(\hat{\theta}) + [\text{viés}(\hat{\theta})]^2, \end{aligned}$$

uma vez que $E(\hat{\theta}) - \theta$ é uma constante e $E[\hat{\theta} - E(\hat{\theta})] = 0$. Na estimação do EQM pelo método de Monte Carlo temos que

$$\text{EQM}(\hat{\theta}) \approx \widehat{\text{EQM}}(\hat{\theta}) = \frac{1}{N} \sum_{i=1}^N (\hat{\theta}^{(i)} - \theta)^2.$$

Tomando a raiz quadrada de EQM produzimos o Desvio Quadrático Médio que é uma medida de precisão também conhecida como Média Quadrática. Suponhamos que $\hat{\theta}_1$ e $\hat{\theta}_2$ são dois estimadores de θ , então dizemos que $\hat{\theta}_1$ é melhor que $\hat{\theta}_2$ se

$$\text{EQM}(\hat{\theta}_1) \leq \text{EQM}(\hat{\theta}_2).$$

No Erro Quadrático Médio avaliamos a diferença entre um estimador e o verdadeiro valor da quantidade estimada. No caso de $\hat{\theta}_1$ e $\hat{\theta}_2$ serem estimadores não viesados de θ , isto é, $E(\hat{\theta}_1) = E(\hat{\theta}_2) = \theta$, então $\hat{\theta}_1$ é melhor que $\hat{\theta}_2$ se

$$\text{Var}(\hat{\theta}_1) \leq \text{Var}(\hat{\theta}_2).$$

Em nosso estudo utilizamos a raiz do erro quadrático médio a fim de comparar o desempenho dos estimadores propostos, ou seja, o estimador que possuir o menor valor da raiz do erro quadrático médio consideramos como o melhor estimador do parâmetro.

Apêndice B

Estatísticas U

B.1 Definições e Propriedades

Seja $\mathcal{F}$ uma família de medidas de probabilidade em um espaço mensurável arbitrário. Seja $\theta(F)$ uma função real mensurável (parâmetro) definida para $F \in \mathcal{F}$. Consideramos uma amostra $X_1, X_2, \dots, X_n$ de variáveis aleatórias independentes e identicamente distribuídas (i.i.d.) com função de distribuição F . Suponhamos que queremos estimar um parâmetro $\theta = \theta(F)$ baseado nesta amostra.

Assumimos que $\mathcal{F}$ consiste de todas as distribuições contínuas ou discretas ou subclasses destas. Note que nenhuma hipótese sobre o conhecimento de F é assumida, somente que ela é uma distribuição da família $\mathcal{F}$.

Definição B.1.1. $\theta(F)$ é um parâmetro estimável em $\mathcal{F}$ se para algum inteiro m existe um estimador não viesado de $\theta(F)$ baseado em m variáveis aleatórias i.i.d. distribuídas de acordo com F , para o qual existe uma função real mensurável $h(x_1, \dots, x_m)$ tal que

$$E_F(h(X_1, \dots, X_m)) = \theta(F) \quad \text{para todo } F \in \mathcal{F}. \quad (\text{B.1})$$

O menor inteiro m com esta propriedade é chamado grau de $\theta(F)$ e a função $h(x_1, \dots, x_m)$ que satisfaz (B.1) é chamada de núcleo de θ .

Sem perda de generalidade, suponhamos que $h(x_1, \dots, x_m)$ é simétrica em seus argumentos, ou seja,

$$h(x_1, \dots, x_m) = h(x_{\alpha_1}, \dots, x_{\alpha_m})$$

para qualquer permutação $\{\alpha_1, \dots, \alpha_m\}$ do conjunto $\{1, \dots, m\}$. No caso em que $h(x_1, \dots, x_m)$ não for simétrica, podemos construir a partir de $h(x_1, \dots, x_m)$ uma função que seja simétrica nos seus argumentos,

$$h^*(x_1, \dots, x_m) = \frac{1}{m!} \sum_{\zeta_m} h(x_{\alpha_1}, \dots, x_{\alpha_m}),$$

em que ζ_m indica a soma sobre todas as possíveis permutações no conjunto $\{1, \dots, m\}$. Portanto através desta quantidade, podemos definir um estimador de θ que é não viesado e simétrico chamado de estatística U (onde U refere-se a “unbiased”).

Definição B.1.2. Para uma função real mensurável $h(x_1, \dots, x_m)$, com $m \leq n$, e para uma amostra de variáveis aleatórias i.i.d. $X_1, \dots, X_n$ de uma distribuição F , uma estatística U com núcleo h é definida como

$$U_n = U_n(h) = \frac{(n-m)!}{n!} \sum_{\zeta_{m,n}} h(X_{i_1}, \dots, X_{i_m}), \quad (\text{B.0})$$

em que o somatório é sobre o conjunto $\zeta_{m,n}$ de todas as $\frac{n!}{(n-m)!}$ permutações do conjunto $\{i_1, i_2, \dots, i_m\}$ de tamanho m a partir de $\{1, 2, \dots, n\}$. Se o núcleo h é simétrico em seus argumentos, U_n possui a forma equivalente

$$U_n = U_n(h) = \frac{1}{\binom{n}{m}} \sum_{C_{m,n}} h(X_{i_1}, \dots, X_{i_m}),$$

em que o somatório é sobre o conjunto $C_{m,n}$ de todas as combinações de m inteiros $i_1 < i_2 < \dots < i_m$ a partir de $\{1, 2, \dots, n\}$.

Existem funcionais $\theta(F)$ que não podem ser expressos da forma (B.0). Por exemplo, suponha que desejamos estimar a probabilidade de que a distribuição atribua valores positivos. Logo

$$\Psi(x) = \begin{cases} 1, & x > 0 \\ 0, & \text{caso contrário.} \end{cases}$$

Então $E_F[\Psi(X_1)] = P_F(X_1 > 0) = p_F$. Note que $\Psi(\cdot)$ fornece um estimador não viesado de $P_F(X_1 > 0)$ para cada F mas o valor numérico da probabilidade de p_F não é a mesma para todo F . Assim, a correspondente estatística U é dada por

$$\begin{aligned} U_n(X_1, \dots, X_n) &= \frac{1}{n} [\text{o número de } X_i\text{'s que são positivos}] \\ &= \frac{1}{n} B, \end{aligned}$$

em que B é a estatística usada no teste sinal.

Uma estatística U pode ser representada como sendo o resultado da esperança condicional do núcleo com relação as estatísticas de ordem, isto é, para um núcleo $h(x_1, \dots, x_m)$ e uma amostra $X_1, \dots, X_n$ em que $m \leq n$, temos

$$U_n = E\{h(X_1, \dots, X_m) | X_{(1:n)}, \dots, X_{(n:n)}\},$$

em que $(X_{(1:n)}, \dots, X_{(n:n)}) = (X_{(1)}, \dots, X_{(n)})$ denotam as estatísticas de ordem (Serfling, 1980).

Teorema B.1.1. *Seja $S = S(X_1, \dots, X_n)$ um estimador não viesado de $\theta(F)$ baseado em uma amostra $X_1, \dots, X_n$ de uma distribuição F . Então, a estatística U correspondente é também não viesada, ou seja*

$$E(S) = \theta(F) \quad e \quad E(U_n) = \theta(F)$$

e adicionalmente

$$\text{Var}_F(U_n) \leq \text{Var}_F(S),$$

em que a igualdade ocorre se, e somente se, $P_F(U = S) = 1$.

Demonstração. O núcleo associado com S é

$$\frac{1}{m!} \sum_{\zeta_m} S(x_{i_1}, \dots, x_{i_m}),$$

que no caso ($m = n$) é a estatística U associada com ela mesma. Isto é, a estatística U associada com S pode ser expressa como

$$U = E\{S|X_{(1:n)}, \dots, X_{(n:n)}\}.$$

Visto que $E_F(U) = E_F(S)$, pela desigualdade de Jensen temos

$$\begin{aligned} E_F(U_n^2) &\leq E_F(S^2) \\ E_F\{E^2(S|X_{(1:n)}, \dots, X_{(n:n)})\} &\leq E_F\{E(S^2|X_{(1:n)}, \dots, X_{(n:n)})\} \end{aligned}$$

com a igualdade se e somente se $E(S|X_{(1:n)}, \dots, X_{(n:n)})$ é igual a S com probabilidade 1.

Uma vez que a estatística de ordem $X_{(1:n)}, \dots, X_{(n:n)}$ é suficiente para alguma família de distribuições $\mathcal{F}$ contendo F , temos que a estatística U resultante é uma função condicionada à estatística suficiente. Para que U_n seja um estimador não viesado de variância mínima de θ é necessário que $X_{(1:n)}, \dots, X_{(n:n)}$ seja suficiente completa.

Lema B.1.1. *Seja $h_c(x_1, \dots, x_c) = E\{h(X_1, \dots, X_m)|X_1 = x_1, \dots, X_c = x_c\}$, $c = 1, \dots, m$. Então*

$$(i) \quad h_c(x_1, \dots, x_c) = E\{h_d(X_1, \dots, X_d)|X_1 = x_1, \dots, X_c = x_c\} \text{ para } 1 \leq c < d \leq m$$

$$(ii) \quad E h_c(X_1, \dots, X_c) = E h(X_1, \dots, X_m)$$

Lema B.1.2. Para $F \in \mathcal{F}$ e $\{i_1, \dots, i_m\}$ e $\{j_1, \dots, j_m\}$ em $C_{m,n}$

$$\begin{aligned} \text{Cov}(h(X_{i_1}, \dots, X_{i_m}), h(X_{j_1}, \dots, X_{j_m})) &= \text{Cov}(h_c(X_1, \dots, X_c), h(X_1, \dots, X_m)) \\ &= \sigma_c^2 \end{aligned}$$

em que c é o número de inteiros em comum para $\{i_1, \dots, i_m\}$ e $\{j_1, \dots, j_m\}$.

Demonstração. Pela definição de variância temos que $\sigma_c^2 = \text{E}h_c^2(X_1, \dots, X_c) - [\text{E}h_c(X_1, \dots, X_c)]^2$ e utilizando o resultado do lema B.1.1 (ii) temos que

$$\sigma_c^2 = \text{E}h_c^2(X_1, \dots, X_c) - [\text{E}h(X_1, \dots, X_m)]^2.$$

Dado que $X_1, \dots, X_n$ são identicamente distribuídas e que $\{i_1, \dots, i_m\}$, $\{j_1, \dots, j_m\}$ possuem c elementos em comum, podemos escrever

$$\begin{aligned} \text{Cov}\{h(X_{i_1}, \dots, X_{i_m}), h(X_{j_1}, \dots, X_{j_m})\} &= \\ &= \text{E}\{h(X_1, \dots, X_m)h(X_1, \dots, X_c, X_{m+1}, \dots, X_{2m-c})\} \\ &\quad - \text{E}h(X_1, \dots, X_m)\text{E}h(X_1, \dots, X_c, X_{m+1}, \dots, X_{2m-c}) \\ &= \text{E}\{h(X_1, \dots, X_m)h(X_1, \dots, X_c, X_{m+1}, \dots, X_{2m-c})\} \\ &\quad - [\text{E}h(X_1, \dots, X_m)]^2. \end{aligned}$$

Agora basta provar que

$$\text{E}\{h(X_1, \dots, X_m)h(X_1, \dots, X_c, X_{m+1}, \dots, X_{2m-c})\} = \text{E}h_c^2(X_1, \dots, X_c).$$

Como $X_1, \dots, X_n$ são i.i.d. temos que

$$\begin{aligned} \text{E}\{h(X_1, \dots, X_m)h(X_1, \dots, X_c, X_{m+1}, \dots, X_{2m-c})\} &= \\ &= \int \cdots \int h(x_1, \dots, x_m)h(x_1, \dots, x_c, x_{m+1}, \dots, x_{2m-c}) \prod_{i=1}^{2m-c} dF(x_i) \\ &= \int \cdots \int \left\{ \int \cdots \int h(x_1, \dots, x_m) \prod_{i=c+1}^m dF(x_i) \right\} \\ &\quad \cdot \left\{ \int \cdots \int h(x_1, \dots, x_c, x_{m+1}, \dots, x_{2m-c}) \prod_{i=m+1}^{2m-c} dF(x_i) \right\} \prod_{i=1}^c dF(x_i) \\ &= \int \cdots \int h_c^2(x_1, \dots, x_c) \prod_{i=1}^c dF(x_i) \\ &= \text{E}h_c^2(X_1, \dots, X_c) \end{aligned}$$

e com isto se segue o resultado. Para mais detalhes veja Lee (1990).

Para um dado parâmetro estimável $\theta = \theta(F)$ com o correspondente núcleo simétrico $h(x_1, \dots, x_m)$ e sendo $\mathcal{F}$ a classe das distribuições cuja variância seja finita ($\text{Var}\{h(X_1, \dots, X_m)\} < \infty$). Definindo uma sequência de funções relativa a h e para $c = 0, 1, \dots, m$ temos

$$h_c(x_1, \dots, x_c) = E_F h(x_1, \dots, x_c, X_{c+1}, \dots, X_m),$$

em que $X_{c+1}, \dots, X_m$ são i.i.d. seguindo a F . Note que $h_0 = \theta$ e $h_m(x_1, \dots, x_m) = h(x_1, \dots, x_m)$ e são tal que possuem esperança θ , isto é,

$$E h_c(X_1, \dots, X_c) = E h(X_1, \dots, X_c, X_{c+1}, \dots, X_m) = \theta$$

contudo elas não podem ser chamadas de funções núcleo pois podem depender da distribuição F .

Assim, para $c = 0, 1, \dots, m$,

$$\sigma_c^2 = \text{Var}(h_c(X_1, \dots, X_c))$$

de modo que $\sigma_0^2 = 0$ e $\sigma_m^2 = \text{Var}(h(X_1, \dots, X_m))$.

Teorema B.1.2. Para $F \in \mathcal{F}$

$$\text{Var}(U_n) = \binom{n}{m}^{-1} \sum_{c=1}^m \binom{m}{c} \binom{n-m}{m-c} \sigma_c^2.$$

Se $\sigma_c^2 < \infty$, então $\text{Var}(U_n) \approx m^2 \sigma_1^2 / n$ para n suficientemente grande.

Demonstração. Seja

$$\begin{aligned} \text{Var}(U_n) &= \text{Var} \left\{ \binom{n}{m}^{-1} \sum_{C_{n,m}} h(X_{\alpha_1}, \dots, X_{\alpha_m}) \right\} = \\ &= \text{Cov} \left\{ \binom{n}{m}^{-1} \sum_{C_{n,m}} h(X_{\alpha_1}, \dots, X_{\alpha_m}), \binom{n}{m}^{-1} \sum_{C'_{n,m}} h(X_{\beta_1}, \dots, X_{\beta_m}) \right\} \\ &= \binom{n}{m}^{-2} \sum_{C_{n,m}} \sum_{C'_{n,m}} \text{Cov}\{h(X_{\alpha_1}, \dots, X_{\alpha_m}), h(X_{\beta_1}, \dots, X_{\beta_m})\} \end{aligned}$$

Precisamos agora continuar separando a soma dos termos com exatamente c elementos em comum. O número de pares de $\{h(X_{\alpha_1}, \dots, X_{\alpha_m}), h(X_{\beta_1}, \dots, X_{\beta_m})\}$ que possuem exatamente c elementos em comum é $\binom{n}{m} \binom{m}{c} \binom{n-m}{m-c}$, pois o primeiro membro do par pode ser escolhido de $\binom{n}{m}$ maneiras, o segundo membro tem que ter c elementos em comum com o primeiro membro, portanto existem $\binom{m}{c}$ maneiras de

escolher c . E por fim como não queremos ter mais do que c elementos em comum com o primeiro membro, os $m - c$ elementos restantes devem ser escolhidos de $\binom{n-m}{m-c}$ maneiras.

Então,

$$\begin{aligned} \text{Var}(U_n) &= \binom{n}{m}^{-2} \sum_{c=1}^m \binom{n}{m} \binom{m}{c} \binom{n-m}{m-c} \sigma_c^2 \\ &= \binom{n}{m}^{-1} \sum_{c=1}^m \binom{m}{c} \binom{n-m}{m-c} \sigma_c^2 \end{aligned} \quad (\text{B.-20})$$

Considerando a variância definida em (B.-19), temos que $\text{Var}(U_n) = O(n^{-1})$ ¹ se $\sigma_1^2 > 0$. No caso em que a estatística U é estacionária de ordem d , ou seja, $0 = \sigma_1^2 = \sigma_2^2 = \dots = \sigma_d < \sigma_{d+1}$, temos $\text{Var}(U_n) = O(n^{-(d+1)})$. Outro fato importante é que $\text{Var}(\sqrt{n}U_n)$ é uma função decrescente de n de tal forma que $\text{Var}(U_n) = \frac{m^2\sigma_1^2}{n} + O(n^{-2})$.

Nos casos em que temos $m \geq 2$, não podemos aplicar diretamente o Teorema do Limite Central clássico com o propósito de obter a distribuição assintótica de U_n , pois os termos de (B.0) não são todos independentes. Assim, uma alternativa seria a exploração da estrutura de tipo martingal (ou tipo martingal reverso) associado ao processo de construção da estatística U e utilizar o TLC para martingais (ou para martingais reversos). Entretanto, nesta construção nem sempre é fácil a verificação das condições de regularidade principalmente quando a estatística é estacionária.

A distribuição assintótica de U_n pode ser obtida, no caso em que $\sigma_1^2 > 0$, por meio de uma projeção de U_n em relação a $X_1, \dots, X_n$ resultando em uma soma de variáveis aleatórias independentes e identicamente distribuídas mais uma parcela desprezável em probabilidade (Hoeffding, 1948). Por este motivo é possível utilizar o TLC clássico juntamente com o Teorema de Slutsky de forma a obter a distribuição assintótica da estatística U_n , isto é,

$$\sqrt{n}(U_n - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, m^2\sigma_1^2). \quad (\text{B.-19})$$

Para o caso de estacionaridade maior que 1, veja Serfling (1980).

Uma forma de representar as estatísticas U é através da decomposição H , bastante útil para obter propriedades assintóticas de U_n . Nessa decomposição representamos a estatística U de grau m como a soma de m estatísticas U não-correlacionadas de graus $1, 2, \dots, m$. Seja U_n uma estatística U de grau m , então a

¹O termo $O(\cdot)$ ("de ordem no máximo igual a") é usado para comparar as ordens de magnitude de sequências de constantes.

correspondente decomposição H é dada por

$$U_n = \theta + \sum_{j=1}^m \binom{m}{j} H_n^{(j)}$$

em que

$$H_n^{(j)} = \binom{n}{j}^{-1} \sum_{(n,j)} h^{(j)}(X_{v_1}, \dots, X_{v_j}),$$

com $\sum_{(n,j)}$ indicando a soma estendida sob todas as $\binom{n}{j}$ combinações de n elementos $\{i_1, \dots, i_m\}$ de $\{1, \dots, n\}$ e com a função $h^{(j)}(x_{v_1}, \dots, x_{v_j})$ definida como

$$\begin{aligned} h^{(1)}(x_1) &= h_1(x_1) - \theta \\ h^{(c)}(x_1, \dots, x_c) &= h_c(x_1, \dots, x_c) - \sum_{j=1}^{c-1} \sum_{(c,j)} h^{(j)}(x_{i_1}, \dots, x_{i_j}) - \theta, \quad c = 2, \dots, m, \end{aligned}$$

com $\text{Cov}(H_n^{(j)}, H_n^{(j')}) = 0$, para todo $j \neq j'$. Para mais detalhes veja Lee (1990).

Também podemos definir uma extensão das estatísticas U , conhecida como estatísticas U generalizadas de grau $(m_1, \dots, m_t)$, definidas como

$$U_n = \prod_{i=1}^t \binom{n_i}{m_i}^{-1} \sum_{(n_1, m_1)} \cdots \sum_{(n_t, m_t)} h(X_{i_1}^1, \dots, X_{i_{m_1}}^1, \dots, X_{i_1}^t, \dots, X_{i_{m_t}}^t).$$

Para mais detalhes sobre as estatísticas U generalizadas veja Serfling (1980) e Lee (1990), por exemplo.