UNIVERSIDADE FEDERAL DO PARANÁ

ANÁLISE DE CORRELAÇÃO: ABORDAGEM TEÓRICA E DE CONSTRUÇÃO DOS COEFICIENTES COM APLICAÇÕES

CURITIBA

2004

SACHIKO ARAKI LIRA

ANÁLISE DE CORRELAÇÃO: ABORDAGEM TEÓRICA E DE CONSTRUÇÃO DOS COEFICIENTES COM APLICAÇÕES

Dissertação apresentada ao Curso de Pós-Graduação em Métodos Numéricos em Engenharia dos Setores de Ciências Exatas e de Tecnologia da Universidade Federal do Paraná, como requisito parcial à obtenção do Grau de "Mestre em Ciências".

Orientador: Prof. Dr. Anselmo Chaves Neto

CURITIBA

2004

TERMO DE APROVAÇÃO

SACHIKO ARAKI LIRA

ANÁLISE DE CORRELAÇÃO: ABORDAGEM TEÓRICA E DE CONSTRUÇÃO DOS COEFICIENTES COM APLICAÇÕES

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciências no Curso de Pós-Graduação em Métodos Numéricos em Engenharia da Universidade Federal do Paraná pela comissão formada pelos professores:

Orientador: Prof. Anselmo Chaves Neto, Dr.
Departamento de Estatística, UFPR

Prof. Jair Mendes Marques, Dr.
PPGMNE, UFPR

Prof. Sergio Aparecido Ignácio, Dr.
Pontifícia Universidade Católica do Paraná

Prof. Fernando Lang da Silveira, Dr.
Universidade Federal do Rio Grande do Sul

Curitiba, 16 de fevereiro de 2004

AGRADECIMENTOS

Ao orientador e amigo Prof. Anselmo Chaves Neto, pelos conhecimentos transmitidos desde o curso da graduação, pelo incentivo para fazer o Mestrado e pela orientação na realização deste trabalho.

Aos professores, colegas e amigos do Programa de Curso de Pós-Graduação em Métodos Numéricos em Engenharia.

Ao Prof. Fernando Lang da Silveira, da Universidade Federal do Rio Grande do Sul, que, mesmo sem me conhecer pessoalmente, gentilmente enviou-me seus trabalhos e sugeriu-me algumas leituras sobre diversas questões relacionadas ao tema.

Ao Instituto Paranaense de Desenvolvimento Econômico e Social (Ipardes), que me apoiou na decisão de fazer o Curso de Mestrado, e possibilitou a utilização do *software* SAS e dos microdados da Pesquisa Mensal de Emprego (PME).

À Ana Rita Barzick Nogueira e Estelita S. de Matias, que muito me ajudaram na editoração e revisão final do texto.

À Maria Luiza Pillati Lourenço, pela orientação quanto às normas para as referências citadas no trabalho.

À minha sobrinha Josiane, pela valiosa contribuição na localização de livros e trabalhos na biblioteca da UFRGS.

Ao meu esposo Herbert, pelo apoio irrestrito, pelo incentivo, carinho e compreensão em todos os momentos, não só durante o desenvolvimento deste trabalho, mas desde o momento em que decidi fazer o Curso de Mestrado.

Aos meus filhos Herbert Júnior e Bernard, pela compreensão nos momentos em que estive ausente.

A todas as pessoas que, direta ou indiretamente, estiveram presentes na realização deste trabalho.

SUMÁRIO

LISTA DE TABELAS viii
LISTA DE QUADROS ix
LISTA DE GRÁFICOS x
RESUMO xi
ABSTRACT xii
1 INTRODUÇÃO 1
1.1 PRELIMINARES 1
1.2 OBJETIVOS 2
1.3 JUSTIFICATIVA 3
1.4 RESUMO HISTÓRICO 3
1.5 APRESENTAÇÃO DOS CAPÍTULOS 4
2 REVISÃO DE LITERATURA 6
2.1 VARIÁVEL QUALITATIVA, QUANTITATIVA E ESCALAS 6
2.2 VARIÁVEL ALEATÓRIA 8
2.3 PARÂMETROS 8
2.4 DISTRIBUIÇÕES DE PROBABILIDADES 10
2.4.1 Distribuição Discreta 10
2.4.1.1 Distribuição de Bernoulli 10
2.4.2 Distribuições Contínuas 11
2.4.2.1 Distribuição normal univariada 12
2.4.2.2 Distribuição χ^2 (qui-quadrado) 14
2.4.2.3 Distribuição "t" de Student 16
2.4.2.4 Distribuição F de Snedecor 17
2.4.2.5 Distribuição normal multivariada 19
2.5 ESTIMADORES DOS PARÂMETROS 24
2.6 MÉTODOS DE ESTIMAÇÃO DOS PARÂMETROS 26
2.6.1 Método de Máxima Verossimilhança 26
2.6.2 Método dos Momentos 27
2.7 TESTES PARAMÉTRICOS E NÃO-PARAMÉTRICOS 28
2.7.1 Testes Paramétricos 28
2.7.2 Testes Não-Paramétricos 28
2.7.2.1 Testes de aderência 28
3 MEDIDAS DE CORRELAÇÃO 30

3.1 INTRODUÇÃO 30
3.2 MEDIDAS DE CORRELAÇÃO ENTRE DUAS VARIÁVEIS 33
3.2.1 Coeficiente de Correlação Linear de Pearson e a Distribuição Normal Bivariada 34
3.2.1.1 Estimadores de máxima verossimilhança 35
3.2.1.2 Suposições básicas para a utilização do Coeficiente de Correlação Linear de Pearson 39
3.2.1.3 Interpretação do Coeficiente de Correlação Linear de Pearson 41
3.2.1.4 Fatores que afetam o Coeficiente de Correlação Linear de Pearson 45
3.2.1.5 Distribuição Amostral do Coeficiente de Correlação Linear de Pearson 50
3.2.1.6 Teste de hipótese para $\rho = 0$ 62
3.2.1.7 Transformação Z de Fisher 66
3.2.1.8 Teste de hipótese para $\rho \neq 0$ 69
3.2.1.9 Intervalo de confiança para ρ 69
3.2.1.10 Confiabilidade 70
3.2.1.10.1 Confiabilidade de instrumentos de medida 70
3.2.1.10.1.1 Correção de atenuação do coeficiente de correlação 76
3.2.1.10.1.2 Aplicação da correção de atenuação 78
3.2.1.10.1.3 Aplicação da correção para restrição em variabilidade 79
3.2.1.10.2 Confiabilidade em Sistemas de Engenharia 80
3.2.1.10.2.1 Confiabilidade estrutural 81
3.2.1.10.2.2 Confiabilidade de sistemas 82
3.2.1.11 Teste de normalidade (Gaussianidade) 84
3.2.2 Coeficiente de Correlação Bisserial 86
3.2.2.1 Introdução 86
3.2.2.2 Estimador do Coeficiente de Correlação Bisserial e do erro padrão 87
3.2.2.3 Suposições básicas para a utilização do Coeficiente de Correlação Bisserial 88
3.2.2.4 Aplicação do Coeficiente de Correlação Bisserial 89
3.2.3 Coeficiente de Correlação Ponto Bisserial 91
3.2.3.1 Introdução 91
3.2.3.2 Estimador do Coeficiente de Correlação Ponto Bisserial e do erro padrão 91
3.2.3.3 Suposições básicas para a utilização do Coeficiente de Correlação Ponto Bisserial 94
3.2.3.4 Coeficiente de Correlação Ponto Bisserial e teste de médias 94
3.2.3.5 Aplicação do Coeficiente de Correlação Ponto Bisserial 95
3.2.4 Coeficiente de Correlação Tetracórico 96

3.2.4.1 Introdução 96
3.2.4.2 Estimador do Coeficiente de Correlação Tetracórico e do erro padrão 97
3.2.4.3 Suposições básicas para a utilização do Coeficiente de Correlação Tetracórico 100
3.2.4.4 Aplicação do Coeficiente de Correlação Tetracórico 100
3.2.5 Coeficiente de Correlação de Spearman 101
3.2.5.1 Introdução 101
3.2.5.2 Estimador do Coeficiente de Correlação de Spearman e significância 102
3.2.5.3 Suposições para a utilização do Coeficiente de Correlação de Spearman 104
3.2.5.4 Aplicação do Coeficiente de Correlação de Spearman 104
3.2.6 Coeficiente de Correlação por Postos de Kendall 106
3.2.6.1 Introdução 106
3.2.6.2 Estimador do Coeficiente de Correlação por Postos de Kendall e significância 106
3.2.6.3 Aplicação do Coeficiente de Correlação por Postos de Kendall 107
3.2.7 Coeficiente de Correlação Phi 108
3.2.7.1 Introdução 108
3.2.7.2 Estimador do Coeficiente de Correlação Phi e significância 108
3.2.7.3 O Coeficiente de Correlação Phi e a Análise de Agrupamento 111
3.2.7.4 Aplicação do Coeficiente de Correlação Phi 111
3.2.8 Coeficiente de Contingência 112
3.2.8.1 Introdução 112
3.2.8.2 Estimador do Coeficiente de Contingência e significância 113
3.2.8.3 Aplicação do Coeficiente de Contingência 114
3.2.9 Coeficiente de Correlação Eta 115
3.2.9.1 Introdução 115
3.2.9.2 Estimador do Coeficiente de Correlação Eta e significância 116
3.2.9.3 O Coeficiente de Correlação Eta e a Análise de Variância 117
3.2.9.4 Aplicação do Coeficiente de Correlação Eta 117
3.2.10 Resumo dos Coeficientes de Correlação entre Duas Variáveis 118
3.3 MEDIDAS DE CORRELAÇÃO ENTRE DIVERSAS VARIÁVEIS 119
3.3.1 Matriz de Correlações 119
3.3.1.1 Análise de Componentes Principais 119
3.3.1.1.1 Introdução 119
3.3.1.1.2 Aplicação da Análise de Componentes Principais 121
3.3.1.2 Análise Fatorial 122
3.3.1.2.1 Introdução 122

3.3.1.2.2 Aplicação da Análise Fatorial 126
3.3.2 Coeficiente de Correlação Múltipla e Parcial 128
3.3.2.1 Introdução 128
3.3.2.2 Suposições para a utilização do Coeficiente de Correlação Múltipla 129
3.3.2.3 Estimador do Coeficiente de Correlação Múltipla 130
3.3.2.4 Aplicação do Coeficiente de Correlação Múltipla 136
3.3.3 Análise de Correlação Canônica 138
3.3.3.1 Introdução 138
3.3.3.2 Aplicação da Análise de Correlação Canônica 140
4 RESULTADOS E DISCUSSÃO 143
4.1 INTRODUÇÃO 143
4.2 COEFICIENTES DE CORRELAÇÃO LINEAR DE PEARSON, BISSERIAL E TETRACÓRICO 143
4.2.1 Cálculo dos Coeficientes de Correlação 146
4.2.2 Comparação dos Erros Padrão 149
4.2.3 Comparação dos Coeficientes de Correlação Estimados 150
4.3 AVALIAÇÃO DOS RESULTADOS 151
CONCLUSÕES E RECOMENDAÇÕES 153
REFERÊNCIAS 155
APÊNDICE 1 - DISTRIBUIÇÕES AMOSTRAIS DO COEFICIENTE DE CORRELAÇÃO LINEAR DE PEARSON $(\hat{\rho})$ 158
APÊNDICE 2 - DISTRIBUIÇÕES AMOSTRAIS DE Z 162
APÊNDICE 3 - TESTE DE NORMALIDADE 165
APÊNDICE 4 - APLICAÇÃO DO COEFICIENTE DE CORRELAÇÃO PONTO BISSERIAL 172
APÊNDICE 5 - CÁLCULO DOS COEFICIENTES DE CORRELAÇÃO DE SPEARMAN E POR POSTOS DE KENDALL 177
APÊNDICE 6 - PROGRAMAS UTILIZADOS 179
ANEXO 1 - CO-RELATIONS AND THEIR MEASUREMENT, CHIEFLY FROM ANTHROPOMETRIC DATA 186
ANEXO 2 - VALORES CRÍTICOS DO COEFICIENTE DE CORRELAÇÃO 195

LISTA DE TABELAS

1 COEFICIENTES DE CONFIABILIDADE E DE CORRELAÇÃO ENTRE OS ESCORES DAS PROVAS DO CONCURSO VESTIBULAR DA UFRGS E DA PUCRS - 1999 79

2 COEFICIENTE DE CORRELAÇÃO ENTRE OS ESCORES DA PROVA DE REDAÇÃO E OUTRAS PROVAS DO CONCURSO VESTIBULAR DA UFRGS E DA PUCRS - 1999 80

3 POPULAÇÃO MIGRANTE TOTAL E ECONOMICAMENTE ATIVA NAS ATIVIDADES URBANAS, SEGUNDO MICRORREGIÕES DO PARANÁ - 1970 105

4 SITUAÇÃO OCUPACIONAL DA POPULAÇÃO ECONOMICAMENTE ATIVA SEGUNDO GÊNERO, NA RMC - AGOSTO 2003 112

5 COEFICIENTE DE CORRELAÇÃO ENTRE VARIÁVEIS DAS EQUAÇÕES DE INFILTRAÇÃO E PORCENTAGEM DE ARGILA E SILTE, EM JOÃO PESSOA 137

6 COEFICIENTES DE REGRESSÃO E CORRELAÇÃO MÚLTIPLA 138

LISTA DE QUADROS

1 VALORES DE V_1 E V_2 SEGUNDO TAMANHO DA AMOSTRA 67

2 ESTATÍSTICAS DA VARIÁVEL ALEATÓRIA X, SEGUNDO A ORDEM CRESCENTE 85

3 COEFICIENTES DE CORRELAÇÃO DE PEARSON E BISSERIAL ENTRE A PONTUAÇÃO TOTAL E RESPOSTA DE CADA ITEM, NO TESTE DE INTERPRETAÇÃO DE TEXTO DA 3.ª SÉRIE, DAS ESCOLAS MUNICIPAIS DE ANDIRÁ 90

4 MATRIZ DE CORRELAÇÃO TETRACÓRICA SEGUNDO ITENS DO TESTE ALÉRGICO 101

5 RESUMO DOS COEFICIENTES DE CORRELAÇÃO ENTRE DUAS VARIÁVEIS 118

6 MATRIZ DE CORRELAÇÃO ENTRE AS BANDAS LANDSAT-TM EM MACURURÉ - OUTUBRO 1987 121

7 AUTOVALORES E AUTOVETORES SEGUNDO COMPONENTES PRINCIPAIS 122

8 NÚMERO DE FAXINAIS, SEGUNDO MUNICÍPIOS DA REGIÃO CENTRO-SUL DO PARANÁ - AGOSTO 1997-JULHO 1998 126

9 RANQUEAMENTO DOS FAXINAIS DA REGIÃO CENTRO-SUL DO PARANÁ - AGOSTO 1997-JULHO 1998 127

10 CORRELAÇÕES CANÔNICAS ENTRE AS VARIÁVEIS DO GRUPO 1 E GRUPO 2 142

11 CORRELAÇÕES CANÔNICAS ENTRE AS VARIÁVEIS DO GRUPO 2 E GRUPO 3 142

12 PARÂMETROS UTILIZADOS NO PROCESSO DE SIMULAÇÃO PARA A OBTENÇÃO DAS AMOSTRAS NORMAIS BIVARIADAS 144

13 MÉDIA, DESVIO PADRÃO E MEDIANA DAS VARIÁVEIS ALEATÓRIAS X E Y, SEGUNDO O TAMANHO DA AMOSTRA 144

14 DESVIOS PADRÃO DAS VARIÁVEIS X E Y, RAZÃO F E VALOR-P, SEGUNDO O TAMANHO DA AMOSTRA 145

15 COEFICIENTE DE CORRELAÇÃO LINEAR DE PEARSON ($\hat{\rho}$) E ERRO PADRÃO, SEGUNDO O TAMANHO DA AMOSTRA 147

16 COEFICIENTE DE CORRELAÇÃO BISSERIAL ($\hat{\rho}_b$) E ERRO PADRÃO, SEGUNDO O TAMANHO DA AMOSTRA 148

17 COEFICIENTE DE CORRELAÇÃO TETRACÓRICO ($\hat{\rho}_t$) E ERRO PADRÃO, SEGUNDO O TAMANHO DA AMOSTRA 149

18 ERROS PADRÃO DOS COEFICIENTES DE CORRELAÇÃO LINEAR DE PEARSON, BISSERIAL E TETRACÓRICO, SEGUNDO O TAMANHO DA AMOSTRA 150

19 COEFICIENTES DE CORRELAÇÃO LINEAR DE PEARSON, BISSERIAL E TETRACÓRICO E ERROS RELATIVOS PERCENTUAIS, BISSERIAL E TETRACÓRICO, SEGUNDO O TAMANHO DA AMOSTRA 151

LISTA DE GRÁFICOS

1 CORRELAÇÃO LINEAR POSITIVA PERFEITA ENTRE AS VARIÁVEIS X E Y 31

2 CORRELAÇÃO LINEAR NULA ENTRE AS VARIÁVEIS X E Y 31

3 CORRELAÇÃO LINEAR NEGATIVA PERFEITA ENTRE AS VARIÁVEIS X E Y 31

4 CORRELAÇÃO NÃO-LINEAR ENTRE AS VARIÁVEIS X E Y 32

5 DISTRIBUIÇÃO AMOSTRAL DO COEFICIENTE DE CORRELAÇÃO DE PEARSON PARA $\rho = 0{,}80$ 58

6 DISTRIBUIÇÃO AMOSTRAL DO COEFICIENTE DE CORRELAÇÃO DE PEARSON PARA $\rho = -0{,}80$ 58

7 DISTRIBUIÇÃO AMOSTRAL DO COEFICIENTE DE CORRELAÇÃO DE PEARSON PARA $\rho = 0$ 60

8 DISTRIBUIÇÃO AMOSTRAL DE Z PARA $\rho = 0{,}80$ 68

9 DISTRIBUIÇÃO AMOSTRAL DE Z PARA $\rho = 0$ 68

RESUMO

A Análise de Correlação é uma ferramenta importante para as diferentes áreas do conhecimento, não somente como resultado final, mas como uma das etapas para a utilização de outras técnicas de análise. Dentre as principais técnicas que utilizam o Coeficiente de Correlação estão a Análise de Confiabilidade, a Análise da Estrutura de Variância-Covariância e o Teste de Normalidade ou Gaussianidade. É importante, desse modo, conhecer teoricamente os diferentes métodos e as suposições básicas requeridas para a sua utilização de forma adequada. Este trabalho apresenta os métodos de Análise de Correlação, envolvendo variáveis medidas em nível intervalar, nominal e ordinal e a Análise de Correlação Canônica. Os Coeficientes de Correlação Simples abordados no trabalho foram: Coeficiente Linear de Pearson, Coeficiente de Correlação Bisserial, Coeficiente de Correlação Ponto Bisserial, Coeficiente de Correlação Tetracórico, Coeficiente de Correlação Eta, Coeficiente de Correlação de Spearman, Coeficiente de Correlação por Postos de Kendall, Coeficiente de Correlação Phi e Coeficiente de Contingência. O presente trabalho discutiu alguns estudos realizados em diferentes áreas de pesquisa, os quais mostram as aplicações dos diferentes coeficientes de correlação.

Palavras-chave: Coeficiente de Correlação; Medida de Associação; Análise da Estrutura de Variância-Covariância.

ABSTRACT

Different research areas consider Correlation Analysis to be an important tool not only as a final result, but also as one of the steps of other analysis techniques. Among the main techniques making use of a Correlation Coefficient we can mention Reliability Analysis, Variance-covariance Structure Analysis and Normality or Gaussian Test. Thus, theoretically it is important to know different methods and the basic assumptions required to using such methods adequately. The present work shows Correlation Analysis methods involving variables measured at interval, nominal and ordinal levels, and Canonical Correlation Analysis. This work addresses the following Simple Correlation Coefficients: Pearson Linear Correlation Coefficient, Biserial Correlation Coefficient, Point Biserial Correlation Coefficient, Tetrachoric Correlation Coefficient, Eta Correlation Coefficient, Spearman Correlation Coefficient, Kendall Rank Correlation Coefficient, Phi Correlation Coefficient and Contingency Coefficient. The present work discusses some studies, carried out in different research areas, showing different uses of different correlation coefficients.

Key words: Correlation Coefficient; Association Measure; Variance-covariance Structure Analysis.

1 INTRODUÇÃO

1.1 PRELIMINARES

A Análise de Correlação e a Análise de Regressão são métodos estatísticos amplamente utilizados para estudar o grau de relacionamento entre variáveis.

A Análise de Correlação fornece um número, indicando como duas variáveis variam conjuntamente. Mede a intensidade e a direção da relação linear ou não-linear entre duas variáveis. É um indicador que atende à necessidade de se estabelecer a existência ou não de uma relação entre essas variáveis sem que, para isso, seja preciso o ajuste de uma função matemática. Não existe a distinção entre a variável explicativa e a variável resposta, ou seja, o grau de variação conjunta entre X e Y é igual ao grau de variação entre Y e X.

Já a análise de regressão, além de medir a associação entre uma variável resposta Y e um conjunto de variáveis independentes (X_1, X_2,..., X_p), também estima os parâmetros do comportamento sistemático entre as mesmas. Necessita a especificação da forma funcional que relaciona a variável resposta às outras covariáveis.

Quando o objetivo é estudar a relação entre as variáveis, nem sempre é necessário um detalhamento como o da Análise de Regressão, mas apenas determinar o grau de relacionamento entre as variáveis analisadas. Conforme descreve SIEGEL (1975, p. 220): “O estabelecimento da existência de uma correlação entre duas variáveis pode constituir o objetivo precípuo de uma pesquisa (...). Mas também representar apenas um passo, ou estágio, de uma pesquisa com outros objetivos, como, por exemplo, quando empregamos medidas de correlação para comprovar a confiabilidade de nossas observações".

Dado um conjunto de variáveis, pode haver somente uma relação numérica, sem relação causal. Diz-se, neste caso, que a correlação entre as variáveis envolvidas é espúria, devido apenas à coincidência.

Para o desenvolvimento teórico da Análise de Correlação, são feitas determinadas suposições sobre as variáveis envolvidas na análise. Na Análise de Regressão, as suposições são com relação aos erros do modelo ajustado. Entretanto, na prática, nem sempre é possível atender a tais suposições.

Quando as suposições não forem atendidas para a Análise de Correlação, são possíveis os seguintes procedimentos:

- utilizar os métodos não-paramétricos;
- adequar os dados às suposições através de uma transformação das variáveis envolvidas na análise.

Foram abordadas, no presente trabalho, a Análise de Correlação Simples Linear e Não-linear, Linear Múltipla, Análise de Componentes Principais, Análise Fatorial e Correlação Canônica.

A Análise de Correlação é amplamente utilizada em Análise de Confiabilidade, Análise da Estrutura de Variância-Covariância e Teste de Normalidade (Gaussianidade).

1.2 OBJETIVOS

Os objetivos deste trabalho foram:

a) Apresentar a teoria da Análise de Correlação;
b) Discutir os principais métodos e as suposições básicas de cada método;
c) Comparar, mediante simulação, o Coeficiente de Correlação Linear de Pearson com os Coeficientes de Correlação Bisserial e Tetracórico;
d) Apresentar as principais utilidades da Análise de Correlação com aplicações.

Considerando que se trata de um assunto bastante amplo, o objetivo não foi o de esgotar, mas de esclarecer algumas questões teóricas, de forma a contribuir na utilização adequada dos métodos discutidos na literatura que aborda o tema. Assim, procurou-se fazer um detalhamento teórico das técnicas.

1.3 JUSTIFICATIVA

A Análise de Correlação é uma ferramenta importante para as diferentes áreas do conhecimento, não somente como resultado final, mas como uma das etapas para a utilização de outras técnicas de análise.

A importância de conhecer teoricamente e em conjunto os diferentes métodos e as suposições básicas requeridas por parte de cada um deles é fundamental, para que não se utilize medida de correlação inadequada.

É comum o uso do Coeficiente de Correlação Linear de Pearson, por ser o mais conhecido, mas em muitas situações isto se dá sem que se tenha a clareza de que este coeficiente mede a relação linear entre duas variáveis.

Já alguns métodos de uso mais restrito, tais como o Coeficiente de Correlação Bisserial, Ponto Bisserial e o Tetracórico, são pouco abordados nas literaturas clássicas de Estatística.

Ao apresentar os diferentes métodos de Análise de Correlação e as suposições básicas para a sua utilização, pretendeu-se contribuir para o uso adequado de cada um deles, ilustrando com algumas aplicações, através de trabalhos já realizados em diferentes áreas do conhecimento.

1.4 RESUMO HISTÓRICO

A teoria da análise de correlação teve início na segunda metade do século XIX. Francis Galton (1822-1911) foi quem usou pela primeira vez os termos correlação e regressão. Publicou em 1869 o livro *Hereditary Genius*, sobre a teoria da regressão (SCHULTZ e SCHULTZ, 1992).

Galton adotou o termo regressão quando observou que filhos de homens altos não são, em média, tão altos quanto os pais, mas os filhos de homens baixos são, em média, mais altos do que os pais. Deve-se a Galton a forma gráfica de representar as propriedades básicas do coeficiente de correlação. O termo "co-relação" foi proposto por Galton, pela primeira vez, em 1888 (SCHULTZ e SCHULTZ, 1992).

A correlação foi observada analisando-se medidas antropométricas e definida da seguinte forma[1]: "*Two organs are said to be co-related or correlated, when variations in the one are generally accompanied by variations in the other, in the same direction, while the closeness of the relation differs in different pairs of organs*". (GALTON, 1889, p. 238).

Seu aluno, Karl Pearson, desenvolveu a fórmula matemática que usamos hoje e que tem seu nome em homenagem. O símbolo do coeficiente de correlação amostral *r* vem da primeira letra da palavra regressão, em reconhecimento a Galton (SCHULTZ e SCHULTZ, 1992).

No anexo 1, encontra-se o artigo sobre co-relação escrito pelo autor, na íntegra.[2]

1.5 APRESENTAÇÃO DOS CAPÍTULOS

No segundo capítulo, apresenta-se uma rápida revisão de literatura sobre alguns conceitos, distribuições de probabilidades discreta e contínua, estimadores de máxima verossimilhança e de momentos, testes paramétricos e não-paramétricos, importantes para o desenvolvimento do terceiro capítulo.

---

[1]Dois órgãos são ditos correlacionados quando a variação de um deles é geralmente acompanhada pela variação do outro, e na mesma direção, enquanto a proximidade da relação difere em diferentes pares de órgãos.

[2]O artigo foi obtido no endereço eletrônico: <http://www.mugu.com/galton>.

O terceiro capítulo trata da questão central deste trabalho, sendo apresentados, além da Teoria Estatística da Correlação, os diferentes Métodos de Correlação para variáveis medidas em nível intervalar, ordinal e nominal, e suas suposições básicas e a Análise de Correlação Canônica. Discutem-se, ainda, as principais utilidades dos diferentes Métodos de Análise de Correlação com suas aplicações, através de trabalhos realizados em diversas áreas do conhecimento.

No quarto capítulo são feitas comparações entre o Coeficiente de Correlação Linear de Pearson e os Coeficientes de Correlação Tetracórico e Bisserial, a partir de diferentes tamanhos de amostras, geradas por meio do processo de simulação.

Finalmente, faz-se recomendações para a utilização dos diferentes Métodos de Análise de Correlação envolvendo duas variáveis e a possibilidade da utilização do Coeficiente de Correlação Linear de Pearson mesmo em situações que não envolvam variáveis medidas em nível intervalar.

2 REVISÃO DE LITERATURA

2.1 VARIÁVEL QUALITATIVA, QUANTITATIVA E ESCALAS

Toda pesquisa envolve construções teóricas que o pesquisador deseja comprovar. Para isso faz-se necessária a definição de variáveis, através das quais pode-se aferir as questões de interesse. Assim, é possível entender que a variável é uma primeira forma de operacionalizar a construção teórica. E pode-se afirmar que a variável é uma característica que pode ser medida. Uma variável pode se apresentar das seguintes formas, quanto aos valores assumidos:

1.º Escala nominal: é aquela que permite o agrupamento da unidade de observação (unidade da pesquisa) de acordo com uma classificação qualitativa em categorias definidas, ou seja, consiste simplesmente em nomear ou rotular, não sendo possível estabelecer graduação ou ordenamento. Ao se trabalhar com essa escala, cada unidade de observação deve ser classificada em uma e somente uma categoria, isto é, deve ser mutuamente excludente. Citando um exemplo bastante comum, considerando que X seja a variável produção diária de peças de automóveis de uma determinada indústria, é possível classificar as peças em perfeitas e defeituosas. Neste caso, a variável X assume as categorias “perfeita” e “defeituosa”, sendo denominada dicotômica. Quando assume mais de duas categorias é denominada politômica.

2.º Escala ordinal: permite o agrupamento da unidade de observação de acordo com uma ordem de classificação. A escala ordinal fornece informações sobre a ordenação das categorias, mas não indica a grandeza das diferenças entre os valores. Considerando a produção diária das máquinas de uma fábrica de peças de equipamentos eletrônicos, é possível classificá-las em: primeira em produção, segunda em produção, terceira em produção, e assim por diante.

3.º Escala intervalar: ocorre quando as unidades de observação, além de estarem numa ordem de classificação, possibilitam quantificar as diferenças entre elas. Quando o zero está incluído como uma medida, é chamada escala de razão. Como exemplo, seja a variável X o número de peças de automóveis defeituosas produzidas diariamente numa certa indústria, essa variável pode assumir valores: 0, 1, 2, 3, ..., 1.000.

Sempre que possível, é preferível utilizar a medida de escala de razão, pois a partir desta pode-se transformar em escala intervalar, ordinal ou nominal, não ocorrendo o inverso.

De acordo com o nível de mensuração, a variável pode ser classificada em qualitativa ou quantitativa. Variável qualitativa é aquela cujo nível de mensuração é nominal ou ordinal, enquanto a quantitativa é aquela em que o nível de mensuração é intervalar ou de razão.

A variável quantitativa pode ser discreta ou contínua, sendo a primeira resultante de contagem, assumindo somente valores inteiros, e a última de medições, assumindo qualquer valor no campo dos números reais.

Outra diferença entre os dois tipos de variáveis está na interpretação de seus resultados. A variável discreta assume exatamente o valor a ela atribuído. Por exemplo, quando se diz que uma máquina produziu 100 peças durante o dia, isto significa dizer que a máquina produziu exatamente 100 peças no dia.

Já a interpretação de um valor de uma variável contínua é a de ser um valor aproximado, por não existirem instrumentos de medida capazes de medir com precisão absoluta, e mesmo porque pode não haver interesse em se determinar um valor contínuo com tanta precisão, considerando todas as suas casas decimais. Portanto, se a variável de interesse for o diâmetro externo de uma peça, e este for de 10,76 mm, o valor exato pode ser um valor entre 10,775 mm e 10,777 mm.

2.2 VARIÁVEL ALEATÓRIA

Variável aleatória é aquela cujo valor numérico não é conhecido antes da sua observação. Esta tem uma distribuição de probabilidades associada, o que permite calcular a probabilidade de ocorrência de certos valores.

A função p(x), que associa as probabilidades aos valores da variável, é chamada de função de probabilidade (f.p.), no caso da variável aleatória discreta, e de função densidade de probabilidade (f.d.p.), para variável aleatória contínua.

Existem distribuições teóricas de probabilidades para variáveis discretas e contínuas, que serão descritas adiante.

2.3 PARÂMETROS

O parâmetro é uma medida que descreve de forma reduzida uma característica, representada pela variável, da população ou universo. O parâmetro normalmente é desconhecido, e deseja-se estimá-lo através de dados amostrais.

População ou universo é composto pelos distintos elementos (unidades populacionais) que apresentam pelo menos uma característica em comum, aos quais os resultados do estudo deverão ser inferidos.

É importante distinguir a população-alvo da população amostrada, que é aquela da qual é selecionada a amostra para o estudo. A população-alvo ou população-objetivo é aquela da qual se desejam informações, e que deve coincidir com a amostrada, porém algumas vezes, por razões de operacionalidade ou comodidade, a população amostrada é mais restrita que a população-objetivo. Neste caso, deve-se ter claro que os resultados fornecidos pela amostra são válidos para a população amostrada (COCHRAN, 1965).

A esperança matemática E(X) de uma variável aleatória X, que é a média da distribuição, é definida, em CHAVES NETO (2003), por:

$$E(X) = \sum_{i=1}^{\infty} x_i P_X(X = x_i) \qquad (2.1)$$

para variável aleatória discreta, e por

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx \tag{2.2}$$

para variável aleatória contínua.

A variância da variável aleatória, representada por $V(X)$ ou σ^2, é definida por:

$$V(X) = \sigma^2 = E\big(X - E(X)\big)^2 = E(X^2) - \big[E(X)\big]^2 \tag{2.3}$$

onde:

$$E(X^2) = \sum_{i=1}^{\infty} x_i^2 P_X(X = x_i) \tag{2.4}$$

para variável aleatória discreta, e

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx \tag{2.5}$$

para variável aleatória contínua.

Segundo MOOD, GRAYBILL e BOES (1974), se X é uma variável aleatória, o *r*-ésimo momento[3] de X, representado por m'_r, é definido como $m'_r = E(X^r)$, se a esperança existe. Observe-se que se $r = 1$, tem-se $m'_1 = E(X) = \mu_x$, a média aritmética.

Se X é uma variável aleatória, o *r*-ésimo momento centrado em "a" é definido como $E[(X-a)^r]$. Se $a = \mu_x$, o *r*-ésimo momento centrado em μ_x será $m_r = E[(X-\mu_x)^r]$. Fazendo $r = 2$, obtém-se a variância de X, como se pode verificar:

$$m_2 = E[(X-\mu_x)^2] \tag{2.6}$$

Uma função que representa todos os momentos é chamada função geradora de momentos (f.g.m.). A f.g.m., representada por $m_x(t)$ ou $m(t)$, é dada por:

---

[3]O método de estimação de parâmetros, denominado Método dos Momentos, foi uma das contribuições de Karl Pearson.

$$m(t) = E\left[e^{tx}\right] = \sum_{x=0}^{\infty} e^{tx} p(x) \tag{2.7}$$

se a variável aleatória é discreta, e por

$$m(t) = E\left[e^{tx}\right] = \int_{-\infty}^{\infty} e^{tx} f(x)\,dx \tag{2.8}$$

se a variável aleatória é contínua.

Conforme apresentado em MOOD, GRAYBILL e BOES (1974), se a função geradora de momentos existe, então $m(t)$ é continuamente diferenciável em alguma vizinhança da origem. Calculando-se a diferencial da função geradora de momentos r vezes em relação a t, e fazendo t=0, tem-se:

$$\left.\frac{\partial^r m(t)}{\partial t^r}\right|_{t=0} = E\left[X^r\right] = m'_r \tag{2.9}$$

Se $r = 1$, tem-se $E(X) = m'_1(0)$, e para $r = 2$, $E(X^2) = m''_2(0)$.

Portanto, uma vez conhecida a f.g.m. da distribuição da variável aleatória, a derivada primeira da f.g.m. em relação a t, no ponto t=0, fornece a $E(X)$, ou seja, a média da distribuição, e a derivada segunda a $E(X^2)$.

2.4 DISTRIBUIÇÕES DE PROBABILIDADES

2.4.1 Distribuição Discreta

Dentre as distribuições de probabilidades discreta cita-se a de Bernoulli, importante para o desenvolvimento do estimador do Coeficiente de Correlação Ponto Bisserial, a ser tratada na seção 3.2.3.

2.4.1.1 Distribuição de Bernoulli

Uma variável aleatória X tem distribuição de Bernoulli, segundo CHAVES NETO (2003), se assume somente um de dois valores, 1 ou 0. A probabilidade de assumir o valor 1 é θ e a de assumir 0 é $(1-\theta)$, ou seja:

$$P_x(X=1)=\theta \quad \text{e} \quad P_x(X=0)=1-\theta \tag{2.10}$$

A função de probabilidade (f.p.) de X é dada por:

$$P_x(X=x)=\theta^x(1-\theta)^{1-x} \quad , \; x=0,\,1 \qquad , \; 0<\theta<1 \tag{2.11}$$

Resultado 2.1: Os parâmetros da distribuição de Bernoulli são: $E(X)=\theta$ e $V(X)=\theta(1-\theta)$.

Prova:

A esperança matemática de uma variável aleatória discreta é definida por:

$$E(X)=\sum_{i=1}^{\infty} x_i P_x(X=x_i)$$

logo, $$E(X)=0\times\left[\theta^0(1-\theta)^1\right]+1\times\left[\theta^1(1-\theta)^0\right]=\theta \tag{2.12}$$

A variância de uma variável aleatória é definida por: $V(X)=E(X^2)-\left[E(X)\right]^2$

onde: $E(X^2)=\sum_{i=1}^{\infty} x_i^2 P_X(X=x_i)$

logo, $E(X^2)=0^2\times\left[\theta^0(1-\theta)^1\right]+1^2\times\left[\theta^1(1-\theta)^0\right]=\theta$

portanto, $$V(X)=\theta-\left[\theta\right]^2=\theta(1-\theta). \tag{2.13}$$

Uma das aplicações da Distribuição de Bernoulli está na análise de discriminação de um item, onde a resposta ao item é "certo" ou "errado".

2.4.2 Distribuições Contínuas

Dentre as distribuições contínuas, uma das mais importantes é a distribuição normal ou distribuição de Gauss.

Adolph Quetelet, estatístico belga, foi o primeiro a aplicar a curva normal de probabilidade em 1870[4]. Quetelet demonstrou que medidas antropométricas de amostras aleatórias de pessoas formavam uma curva normal. Ele utilizou o termo *"l'homme moyen*" (o homem médio) para exprimir a descoberta de que a maioria dos indivíduos se concentra em torno da média (centro da distribuição), e à medida que se afasta encontra-se um número cada vez menor (SCHULTZ e SCHULTZ, 1992).

A distribuição de muitas estatísticas de testes é normal (Gaussiana) ou segue alguma forma que é derivada da distribuição normal, tais como t, χ^2 (qui-quadrado) e F.

2.4.2.1 Distribuição normal univariada

Uma variável aleatória X tem distribuição normal ou distribuição Gaussiana, segundo CHAVES NETO (2003), quando a sua função densidade de probabilidade (f.d.p.) é dada por:

$$f_x(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}, \; -\infty < \mu < \infty \; , \; \sigma > 0 \, , \; -\infty < x < \infty \qquad (2.14)$$

Resultado 2.2: Os parâmetros da distribuição normal univariada são: $E(X) = \mu$ e $V(X) = \sigma^2$.

Prova:

A esperança matemática de uma variável aleatória contínua é definida por:

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

$$E(X) = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2} dx$$

Fazendo $z = \frac{x-\mu}{\sigma}$, tem-se que $dz = \frac{dx}{\sigma}$

---

[4]Esta informação foi obtida no *site*:
http://stat-www.berkeley.edu/users/nrabbee/stat2/lecture5.pdf

$$E(X) = \int_{-\infty}^{\infty} (z\sigma + \mu) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

$$E(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (z\sigma + \mu) e^{-\frac{1}{2}z^2} dz$$

$$E(X) = \frac{1}{\sqrt{2\pi}} \left[0 + \mu \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} dz \right] = \mu \times \underbrace{\frac{1}{\sqrt{2\pi}} \left[\int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} dz \right]}_{=1} = \mu \qquad (2.15)$$

A variância é obtida através de: $V(X) = E(X^2) - [E(X)]^2$

onde: $E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx$

logo, $E(X^2) = \int_{-\infty}^{\infty} x^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$

Fazendo $z = \frac{X - \mu}{\sigma}$, tem-se que $dz = \frac{dx}{\sigma}$

então: $E(X^2) = \int_{-\infty}^{\infty} (z\sigma + \mu)^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$

$$E(X^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (z^2\sigma^2 + 2z\sigma\mu + \mu^2) e^{-\frac{1}{2}z^2} dz$$

$$E(X^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2\sigma^2 e^{-\frac{1}{2}z^2} dz + 2\mu\sigma \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{-\frac{1}{2}z^2} dz}_{=0} + \mu^2 \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} dz}_{=1}$$

Para calcular $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2\sigma^2 e^{-\frac{1}{2}z^2} dz$, faz-se integração por partes.

Fazendo: $z e^{-\frac{1}{2}z^2} = dv$ e $z = u$

$v = -e^{-\frac{1}{2}z^2}$ e $dz = du$

Obtém-se:

$$\sigma^2 \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\frac{1}{2}z^2} dz = \sigma^2 \left\{ \left[-\frac{z^2 e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}} \right]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} dz \right\}$$

$$E(X^2) = \sigma^2(0+1) + \mu^2 = \sigma^2 + \mu^2$$

$$V(X) = \sigma^2 + \mu^2 - \mu^2 = \sigma^2 \qquad (2.16)$$

Quando se tem média=0 e variância=1, a distribuição é chamada normal padrão e representada pela variável aleatória contínua Z. Então,

$$Z = \left(\frac{X - \mu}{\sigma} \right) \sim N(0,1) \qquad \text{e} \qquad (2.17)$$

$$f_z(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}, \qquad z \in R \qquad (2.18)$$

A Distribuição Normal tem grandes aplicações na inferência estatística, como testes de hipóteses e intervalos de confiança.

2.4.2.2 Distribuição χ^2 (qui-quadrado)

Uma variável aleatória X tem distribuição χ^2, segundo CHAVES NETO (2003), se sua função densidade de probabilidade (f.d.p.) é dada por:

$$f_x(x) = \frac{1}{\Gamma\left(\frac{\nu}{2}\right)} \left(\frac{1}{2} \right)^{\frac{\nu}{2}} x^{\left(\frac{\nu}{2}\right)-1} e^{-\frac{x}{2}}, \quad x > 0, \ \nu > 0 \qquad (2.19)$$

Resultado 2.3: Os parâmetros da distribuição χ^2 são: $E(X) = \nu$ e $V(X) = 2\nu$

Prova: Tem-se que:

$$E(X) = \int_0^{\infty} xf(x)dx = \int_0^{\infty} x \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} x^{\left(\frac{\nu}{2}\right)-1} e^{-\frac{x}{2}} dx$$

$$E(X) = \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} \int_0^{\infty} xx^{\left(\frac{\nu}{2}\right)-1} e^{-\frac{x}{2}} dx = \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} \int_0^{\infty} x^{\left(\frac{\nu}{2}\right)} e^{-\frac{x}{2}} dx$$

A função gama generalizada é definida por: $\int_0^{\infty} x^m e^{-ax^n} dx = \frac{\Gamma\left(\frac{m+1}{n}\right)}{n a^{\frac{m+1}{n}}}$ (2.20)

Assim, tem-se que:

$$E(X) = \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} \left[\frac{\Gamma\left(\frac{\nu}{2}+1 \middle/ 1\right)}{1 \times \left(\frac{1}{2}\right)^{\frac{\nu}{2}+1}} \right] = \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} \frac{\frac{\nu}{2}\Gamma\left(\frac{\nu}{2}\right)}{\left(\frac{1}{2}\right)^{\frac{\nu}{2}+1}}$$

$$E(X) = \frac{\frac{\nu}{2}}{2^{\frac{\nu}{2}}.2^{-\frac{\nu}{2}-1}} = \nu \qquad (2.21)$$

A variância da variável X é obtida por: $V(X) = E(X^2) - [E(X)]^2$

onde: $E(X^2) = \int_0^{\infty} x^2 f(x)dx = \int_0^{\infty} x^2 \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} x^{\left(\frac{\nu}{2}\right)-1} e^{-\frac{x}{2}} dx$, $\quad x>0$

$$E(X^2) = \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} \underbrace{\int_0^{\infty} x^{\left(\frac{\nu}{2}\right)+1} e^{-\frac{x}{2}} dx}_{\text{Gama generalizada}}$$

$$E(X^2) = \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} \left[\frac{\Gamma\left(\frac{\nu}{2}+1+1 \middle/ 1\right)}{1 \times \left(\frac{1}{2}\right)^{\frac{\nu}{2}+1+1}} \right] = \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} \frac{\left(\frac{\nu}{2}+1\right)\frac{\nu}{2}\Gamma\left(\frac{\nu}{2}\right)}{\left(\frac{1}{2}\right)^{\frac{\nu}{2}+2}} = \nu^2 + 2\nu$$

Portanto, $V(X) = \nu^2 + 2\nu - \nu^2 = 2\nu$ (2.22)

Dentre as aplicações da Distribuição Qui-quadrado cita-se a construção de intervalos de confiança para variâncias e testes de hipóteses.

2.4.2.3 Distribuição "t" de Student

Uma variável aleatória X tem distribuição "t" com ν graus de liberdade se sua função densidade de probabilidade (f.d.p.) é dada por:

$$f_X(x) = \frac{\Gamma\left[\frac{(\nu+1)}{2}\right]}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)\left(1+\frac{x^2}{\nu}\right)^{\frac{1}{2}(\nu+1)}}\ ,\quad x\in R,\quad \nu>0 \qquad (2.23)$$

Resultado 2.4: Os parâmetros da distribuição "t" são: $E(T)=0$ e $V(T)=\frac{\nu}{\nu-2}$, $\nu>2$

Prova:

A distribuição "t" é dada por $T=\frac{Z}{\sqrt{\frac{U}{V}}}$ onde $Z\sim N(0,1)$ e $U\sim\chi^2_\nu$ (2.24)

Tem-se que $f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}$, $z\in R$ e

$$f_x(x)=\frac{1}{\Gamma\left(\frac{\nu}{2}\right)}\left(\frac{1}{2}\right)^{\frac{\nu}{2}}x^{\left(\frac{\nu}{2}\right)-1}e^{-\frac{x}{2}}\ ,\quad x>0\ ,\ \nu>0$$

Então $E(T)=E\left[\frac{Z}{\sqrt{U/V}}\right]=\sqrt{V}\,E\left[\frac{Z}{\sqrt{U}}\right]=\sqrt{V}\,E[Z]\,E\left[\frac{1}{\sqrt{U}}\right]$

mas, $E\left[\frac{1}{\sqrt{U}}\right]=\int_0^\infty\frac{1}{\sqrt{u}}f(u)\,du=\frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)}\int_0^\infty u^{-\frac{1}{2}}u^{\left(\frac{\nu}{2}\right)-1}e^{-\frac{u}{2}}du=\frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)}\underbrace{\int_0^\infty u^{\frac{\nu-3}{2}}e^{-\frac{u}{2}}du}_{\text{gama generalizada}}$

$$E\left[\frac{1}{\sqrt{U}}\right]=\frac{\left(\frac{\nu-3}{2}\right)\Gamma\left(\frac{\nu-3}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{2}}\ ,\ \text{portanto}\ E(T)=\sqrt{V}\times0\times E\left[\frac{1}{\sqrt{U}}\right]=0 \qquad (2.25)$$

A variância é dada por: $V(T)=E(T^2)-[E(T)]^2$

onde: $E(T^2) = E\left[\frac{Z}{\sqrt{U/V}}\right]^2 = VE\left[Z^2\right]E\left[\frac{1}{U}\right]$ e,

$$E\left[Z^2\right] = \int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\frac{1}{2}z^2} dz = \frac{1}{\sqrt{2\pi}} \times 2 \times \underbrace{\int_0^{\infty} z^2 e^{-\frac{1}{2}z^2} dz}_{\text{gama generalizada}} = 1$$

mas, $E\left[\frac{1}{U}\right] = \int_0^{\infty} \frac{1}{u} f(u) du = \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} \int_0^{\infty} u^{-1} u^{\left(\frac{\nu}{2}\right)-1} e^{-\frac{u}{2}} du = \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} \underbrace{\int_0^{\infty} u^{\frac{\nu-4}{2}} e^{-\frac{u}{2}} du}_{\text{gama generalizada}}$

então, $E\left[\frac{1}{U}\right] = \frac{1}{\nu - 2}$, portanto $E(T^2) = \nu \times 1 \times \frac{1}{\nu - 2} = \frac{\nu}{\nu - 2}$ e $V(T) = \frac{\nu}{\nu - 2}$. (2.26)

Dentre as utilizações da Distribuição t, citam-se os testes de hipóteses e intervalos de confiança para amostras pequenas $(n < 30)$ e testes de hipóteses para coeficiente de correlação amostral.

2.4.2.4 Distribuição F de Snedecor

A variável aleatória X tem distribuição F de Snedecor com ν_1 e ν_2 graus de liberdade se sua função densidade de probabilidade (f.d.p.) é dada por:

$$f_x(x) = \frac{\Gamma\left[\frac{1}{2}(\nu_1 + \nu_2)\right]\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \frac{x^{\frac{\nu_1}{2}-1}}{\left[1 + \frac{\nu_1}{\nu_2} x\right]^{\left(\frac{\nu_1+\nu_2}{2}\right)}}, \quad x \in R, \quad \nu_1, \nu_2 > 0 \qquad (2.27)$$

Resultado 2.5: Os parâmetros da distribuição F de Snedecor são:

$$E(X) = \frac{\nu_2}{\nu_2 - 2}, \quad \nu_2 > 2 \quad \text{e} \quad V(X) = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}, \quad \nu_2 > 4$$

Prova:

Seja $X = \frac{U/\nu_1}{V/\nu_2} \sim F_{\nu_1,\nu_2}$ então $E(X) = E\left[\frac{U/\nu_1}{V/\nu_2}\right] = \frac{\nu_2}{\nu_1} E\left[\frac{U}{V}\right]$ (2.28)

$$E(X) = \frac{\nu_2}{\nu_1} E[U] E\left[\frac{1}{V}\right]$$

$$E(U) = \nu_1$$

$$E\left(\frac{1}{V}\right) = \frac{1}{\Gamma\left(\frac{\nu_2}{2}\right) 2^{\frac{\nu_2}{2}}} \int_0^\infty \frac{1}{v} v^{\frac{\nu_2 - 2}{2}} e^{-\frac{v}{2}} dv$$

$$E\left(\frac{1}{V}\right) = \frac{1}{\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{1}{2}\right)^{\frac{\nu_2}{2}} \underbrace{\int_0^\infty v^{\frac{\nu_2 - 2}{2}} e^{-\frac{v}{2}} dv}_{\text{gama generalizada}}$$

$$E\left(\frac{1}{V}\right) = \frac{\Gamma\left[\frac{\nu_2 - 2}{2}\right]}{\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{1}{2}\right)^{\frac{\nu_2}{2}} \left(\frac{1}{2}\right)^{\frac{-(\nu_2 - 2)}{2}} = \frac{1}{\nu_2 - 2}$$

Então, tem-se que $E(X) = \frac{\nu_2}{\nu_1} \nu_1 \frac{1}{\nu_2 - 2} = \frac{\nu_2}{\nu_2 - 2}$ (2.29)

$$V(X) = E(X^2) - [E(X)]^2$$

$$E(X^2) = E\left[\left(\frac{U/\nu_1}{V/\nu_2}\right)^2\right] = \left(\frac{\nu_2}{\nu_1}\right)^2 E\left[\frac{U^2}{V^2}\right] = \left(\frac{\nu_2}{\nu_1}\right)^2 E[U^2] E\left[\frac{1}{V^2}\right]$$

$$E(U^2) = \int_0^\infty u^2 f(u) du = \int_0^\infty u^2 \frac{1}{\Gamma\left(\frac{\nu_1}{2}\right) 2^{\frac{\nu_1}{2}}} u^{\frac{\nu_1 - 1}{2}} e^{-\frac{u}{2}} du$$

$$E(U^2) = \frac{1}{\Gamma\left(\frac{\nu_1}{2}\right) 2^{\frac{\nu_1}{2}}} \underbrace{\int_0^\infty u^{\frac{\nu_1 + 1}{2}} e^{-\frac{u}{2}} du}_{\text{Gama generalizada}}$$

$$E(U^2) = \frac{1}{\Gamma\left(\frac{\nu_1}{2}\right) 2^{\frac{\nu_1}{2}}} \frac{\left(\frac{\nu_1}{2} + 1\right) \Gamma\left(\frac{\nu_1}{2} + 1\right)}{\left(\frac{1}{2}\right)^{\frac{\nu_1}{2} + 2}} = \nu_1(\nu_1 + 2)$$

$$E\left(\frac{1}{V^2}\right)=\frac{1}{\Gamma\left(\frac{\nu_2}{2}\right)2^{\frac{\nu_2}{2}}}\int_0^{\infty}\frac{1}{v^2}v^{\frac{\nu_2}{2}-1}e^{-\frac{v}{2}}dv$$

$$E\left(\frac{1}{V^2}\right)=\frac{1}{\Gamma\left(\frac{\nu_2}{2}\right)2^{\frac{\nu_2}{2}}}\int_0^{\infty}v^{\frac{\nu_2}{2}-3}e^{-\frac{v}{2}}dv=\frac{1}{\Gamma\left(\frac{\nu_2}{2}\right)2^{\frac{\nu_2}{2}}}\frac{\Gamma\left(\frac{\nu_2}{2}-2\right)}{\left(\frac{1}{2}\right)^{\frac{\nu_2}{2}-2}}=\frac{1}{(\nu_2-4)(\nu_2-2)}$$

$$E(X^2)=\frac{\nu_2^2}{\nu_1^2}\nu_1(\nu_1+2)\frac{1}{(\nu_2-4)(\nu_2-2)}$$

$$V(X)=\frac{\nu_2^2}{\nu_1^2}\nu_1(\nu_1+2)\frac{1}{(\nu_2-4)(\nu_2-2)}-\frac{\nu_2^2}{(\nu_2-2)^2}=\frac{2\nu_2^2(\nu_2+\nu_1-2)}{\nu_1(\nu_2-4)(\nu_2-2)^2} \quad (2.30)$$

Dentre as aplicações da Distribuição F é possível citar a análise de variância (ANOVA) e análise de regressão.

2.4.2.5 Distribuição normal multivariada

A função densidade de probabilidade da distribuição normal multivariada é uma generalização da normal univariada para $p \geq 2$ dimensões (JOHNSON e WICHERN, 1988).

Relembrando a função densidade de probabilidade da distribuição normal univariada, apresentada na seção 2.4.2.1, que é:

$$f_X(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},\quad -\infty<\mu<\infty,\ \sigma>0,\ -\infty<x<\infty$$

esta notação poderá ser estendida para o caso multivariado. O termo $\left(\frac{x-\mu}{\sigma}\right)^2=(x-\mu)(\sigma^2)^{-1}(x-\mu)$ pode ser generalizado para o vetor $\underline{X}$ de dimensão $p\times 1$ de observações de várias variáveis como $(\underline{X}-\underline{\mu})'\Sigma^{-1}(\underline{X}-\underline{\mu})$. O vetor $\underline{\mu}$ de dimensão $p\times 1$ representa o valor esperado do vetor aleatório $\underline{X}$ e a matriz Σ de dimensão $p\times p$ é sua matriz de variância-covariância. Assume-se que a matriz simétrica Σ é definida positiva e, então, a expressão $(\underline{X}-\underline{\mu})'\Sigma^{-1}(\underline{X}-\underline{\mu})$ é o quadrado da distância generalizada de $\underline{X}$ até $\underline{\mu}$.

A função densidade da distribuição normal multivariada é obtida substituindo a distância univariada pela distância generalizada multivariada. Quando isto é feito, a constante $(2\pi)^{-1/2}\left(\sigma^2\right)^{-1/2}$ deve ser substituída para uma constante que represente o volume sob a superfície da função densidade multivariada. Isto pode ser feito, conforme descrito em JOHNSON e WICHERN (1988), quando esta constante for $(2\pi)^{-p/2}|\Sigma|^{-1/2}$, onde p é a dimensão do vetor aleatório $\underline{X} = [X_1, X_2, ..., X_p]'$. A função densidade de probabilidade será dada por:

$$f_x(\underline{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \; e^{-\frac{1}{2}(\underline{x}-\underline{\mu})' \Sigma^{-1}(\underline{x}-\underline{\mu})}, \quad -\infty < x_i < \infty, \; i = 1,2,...,p \tag{2.31}$$

$\underline{\mu} \in R^p$, Σ definida não negativa.

Representa-se esta função densidade por $N_p(\underline{\mu}, \Sigma)$, onde Σ é a matriz de variância-covariância, ou seja, $V(\underline{X}) = E[(\underline{X}-\underline{\mu})(\underline{X}-\underline{\mu})']$ e $E(\underline{X}) = \underline{\mu}$

Os estimadores de máxima verossimilhança de $\underline{\mu}$ e Σ são apresentados a seguir, conforme demonstrados em JOHNSON e WICHERN (1988, p.140):

$$\hat{\underline{\mu}} = \overline{\underline{X}} \quad \text{e} \quad \hat{\Sigma} = \frac{1}{n}\sum_{j=1}^{n}(\underline{X}_j - \overline{\underline{X}})(\underline{X}_j - \overline{\underline{X}})' = \frac{(n-1)S}{n} \tag{2.32}$$

$$\text{onde } S = \frac{1}{n-1}\sum_{j=1}^{n}(\underline{X}_j - \overline{\underline{X}})(\underline{X}_j - \overline{\underline{X}})' \tag{2.33}$$

A distribuição normal bivariada é um caso particular da multivariada para p = 2.

Se as variáveis aleatórias X e Y, normalmente distribuídas, têm distribuição normal bivariada, então sua função densidade de probabilidade (f.d.p.) é dada por:

$$f_{X,Y}(X,Y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)}\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right) + \left(\frac{Y-\mu_Y}{\sigma_Y}\right)^2\right]\right\} \tag{2.34}$$

$X \in R$, $Y \in R$, $\mu_X \in R$, $\mu_Y \in R$, $\sigma_X \in R^+$, $\sigma_Y \in R^+$ e $-1 \le \rho \le 1$

A função geradora de momentos desta distribuição, conforme apresentada em MOOD, GRAYBILL e BOES (1974), é:

$$m(t_1,t_2)=e^{t_1\mu_X+t_2\mu_Y+\frac{1}{2}\left(t_1^2\sigma_X^2+2\rho t_1t_2\sigma_X\sigma_Y+t_2^2\sigma_Y^2\right)} \tag{2.35}$$

Tem-se, assim, os seguintes resultados:

Resultado 2.6: As médias (parâmetros) das variáveis aleatórias X e Y, com distribuição normal bivariada, são μ_X e μ_Y, respectivamente.

Prova:

Calculando-se a derivada primeira da função geradora de momentos em relação a t_1, no ponto t_1 e t_2 iguais a zero, tem-se:

$$E(X)=\frac{\partial m(t_1,t_2)}{\partial t_1}\Bigg|_{t_1,t_2=0}$$

$$E(X)=e^{t_1\mu_X+t_2\mu_Y+\frac{1}{2}\left(t_1^2\sigma_X^2+2\rho t_1t_2\sigma_X\sigma_Y+t_2^2\sigma_Y^2\right)}\times\left(\mu_X+t_1\sigma_X^2+\rho t_2\sigma_X\sigma_Y\right)\Bigg|_{t_1,t_2=0}$$

$$E(X)=\mu_X \tag{2.36}$$

Da mesma forma, calculando-se a derivada primeira da função geradora de momentos em relação a t_2, no ponto t_1 e t_2 iguais a zero, tem-se:

$$E(Y)=\frac{\partial m(t_1,t_2)}{\partial t_2}\Bigg|_{t_1,t_2=0}$$

$$E(Y)=e^{t_1\mu_X+t_2\mu_Y+\frac{1}{2}\left(t_1^2\sigma_X^2+2\rho t_1t_2\sigma_X\sigma_Y+t_2^2\sigma_Y^2\right)}\times\left(\mu_Y+t_2\sigma_Y^2+\rho t_1\sigma_X\sigma_Y\right)\Bigg|_{t_1,t_2=0}$$

$$E(Y)=\mu_Y \tag{2.37}$$

Resultado 2.7: As variâncias (parâmetros) das variáveis aleatórias X e Y, com distribuição normal bivariada, são σ_X^2 e σ_Y^2 , respectivamente.

Prova:

Calculando-se a derivada segunda da função geradora de momentos em relação t_1, no ponto t_1 e t_2 iguais a zero, tem-se:

$$E(X^2) = \frac{\partial^2 m(t_1, t_2)}{\partial t_1^2}\bigg|_{t_1, t_2 = 0}$$

$$E(X^2) = \sigma_X^2 + \mu_X^2$$

Tem-se que $V(X) = E(X^2) - [E(X)]^2$, logo

$$V(X) = \sigma_X^2 \tag{2.38}$$

Da mesma forma, obtém-se:

$$E(Y^2) = \sigma_Y^2 + \mu_Y^2 \quad \text{e} \quad V(Y) = \sigma_Y^2 \tag{2.39}$$

Resultado 2.8: O coeficiente de correlação (parâmetro) entre as variáveis aleatórias X e Y, com distribuição normal bivariada, é igual a ρ, definida por: $\rho = \rho_{x,y} = \dfrac{COV(X,Y)}{\sigma_X \sigma_Y}$

Prova:

A covariância de X e Y é dada por:

$$E[(X - \mu_X)(Y - \mu_Y)] = E[XY - X\mu_X - Y\mu_Y + \mu_X\mu_Y] = E[XY] - \mu_X\mu_Y \tag{2.40}$$

Para se obter momentos conjunto, diferencia-se $m(t_1, t_2)$, r vezes em relação a t_1 e *s* vezes em relação a t_2 e faz-se t_1 e t_2 iguais a zero.

Então, tem-se que: $E(XY) = \dfrac{\partial^2 m(t_1, t_2)}{\partial t_1 \partial t_2}\bigg|_{t_1, t_2 = 0}$

$$E(XY) = \rho\sigma_X\sigma_Y + \mu_X\mu_Y$$

$$E(XY) - \mu_X \mu_Y = \rho \sigma_X \sigma_Y$$

$$E[(X - \mu_X)(Y - \mu_Y)] = \rho \sigma_X \sigma_Y$$

onde ρ é o coeficiente de correlação entre X e Y e pode-se escrever:

$$\rho = \rho_{x,y} = \frac{COV(X,Y)}{\sigma_X \sigma_Y} \tag{2.41}$$

Resultado 2.9: O Coeficiente de Correlação populacional ρ varia entre –1 e +1, ou seja, $-1 \leq \rho \leq 1$.

Prova: A correlação entre duas variáveis X e Y é definida por:

$$\rho = \rho_{x,y} = \frac{COV(X,Y)}{\sigma_X \sigma_Y}$$

onde: σ_X é o desvio padrão de X;

σ_Y é o desvio padrão de Y;

$COV(X,Y)$ é a covariância entre X e Y.

A variância de qualquer valor é sempre positiva, por definição. Assim:

$$V\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) \geq 0 \tag{2.42}$$

Usando a propriedade da variância, tem-se:

$$V\left(\frac{X}{\sigma_X}\right) + V\left(\frac{Y}{\sigma_Y}\right) + 2COV\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) \geq 0$$

$$\frac{1}{\sigma_X^2} V(X) + \frac{1}{\sigma_Y^2} V(Y) + \frac{2}{\sigma_X \sigma_Y} COV(X,Y) \geq 0$$

$$1 + 1 + \frac{2}{\sigma_X \sigma_Y} COV(X,Y) \geq 0$$

$$1 + \frac{COV(X,Y)}{\sigma_X \sigma_Y} \geq 0$$

$$\rho_{XY} = \frac{COV(X,Y)}{\sigma_X \sigma_Y} \geq -1$$

De forma análoga:

$$V\left(\frac{X}{\sigma_X}-\frac{Y}{\sigma_Y}\right)\geq 0 \qquad (2.43)$$

$$V\left(\frac{X}{\sigma_X}\right)+V\left(\frac{Y}{\sigma_Y}\right)-2COV\left(\frac{X}{\sigma_X},\frac{Y}{\sigma_Y}\right)\geq 0$$

$$\frac{1}{\sigma_X^2}V(X)+\frac{1}{\sigma_Y^2}V(Y)-\frac{2}{\sigma_X\sigma_Y}COV(X,Y)\geq 0$$

$$1+1-\frac{2}{\sigma_X\sigma_Y}COV(X,Y)\geq 0$$

$$1-\frac{COV(X,Y)}{\sigma_X\sigma_Y}\geq 0$$

$$\rho_{x,y}=\frac{COV(X,Y)}{\sigma_X\sigma_Y}\leq 1$$

Portanto:

$$-1\leq \rho_{x,y}\leq 1 \qquad (2.44)$$

2.5 ESTIMADORES DOS PARÂMETROS

No caso das distribuições de probabilidades teóricas descritas na seção anterior (2.4), os parâmetros poderão ser estimados através de estimador ou estatística.

Estimador ou estatística é uma função dos valores da amostra, ou seja, é uma variável aleatória, pois depende dos elementos selecionados para compor a amostra.

Deve-se sempre levar em conta as qualidades de um estimador. Um bom estimador deve possuir as seguintes propriedades:

1.º Ser não-viciado, ou seja, $E(T)=\theta$ onde T = estimador

θ = parâmetro

2.º Ser eficiente (mínima variância)

Tendo dois estimadores T_1 e T_2, a serem utilizados para estimar o mesmo parâmetro θ, T_1 será dito mais eficiente que T_2 se para um mesmo tamanho de

amostra $E\left[(T_1-\theta)^2\right] < E\left[(T_2-\theta)^2\right]$, sendo T_1 e T_2 estimadores não-viciados de θ. Esta condição indica que a variância de T_1 é menor que a variância de T_2.

3.º Ser consistente

Um estimador é dito consistente se $\lim_{n\to\infty} P(|T-\theta| \geq \varepsilon) = 0$, $\forall \varepsilon > 0$.

Se o estimador for não-viciado, a condição de consistência equivale a dizer que sua variância tende a zero quando n tende a crescer infinitamente, ou seja, $\lim_{n\to\infty} V(T) = 0$ e $\lim_{n\to\infty} E(T) = \theta$, onde θ é o parâmetro.

Isto significa dizer que, à medida que se aumenta o tamanho da amostra (n), a diferença entre a estimativa e o parâmetro diminui, chegando a coincidir quando n = N (tamanho da população).

4.º Ser suficiente

O estimador ou estatística é suficiente para estimar um parâmetro θ quando é uma função dos valores da amostra, e resume todas as informações que a mesma tem sobre o parâmetro. Portanto, um estimador suficiente é aquele que depende somente dos dados amostrais.

Uma forma simples de obter-se estatísticas suficientes é usar propriedades das distribuições da família exponencial uniparamétrica ou k-paramétrica, conforme definições apresentadas em CHAVES NETO (2002a).

Definição 1: Uma variável aleatória em *R* possui distribuição da família exponencial uniparamétrica se a sua função de probabilidade (f.p.) ou função densidade de probabilidade (f.d.p.) é da forma $f(x/\theta) = \{\exp[c(\theta)T(x) + d(\theta) + S(x)]\} I_A(x)$, onde $\theta \in \Theta$, intervalo aberto de R e o conjunto $A = \{x / f(x/\theta) > 0\}$ é independente de θ, com I sendo a função indicadora.

Definição 2: A família de distribuição $\{P_\theta; \theta \in \Theta\}$ é dita família exponencial com k parâmetros ou k-paramétrica se existem as funções de valor real $c_1, c_2, ..., c_k$ e $d(\theta)$, e, ainda, $T_1, T_2, ..., T_k$, funções de variável real, e também S, definidas em R^n, e um conjunto $A \subset R^n$, tal que a f.d.p. (ou f.p.) P_θ pode ser escrita na forma:

$$p(\underline{x},\underline{\theta}) = \left\{ \exp\left[\sum_{i=1}^{k} c_i(\underline{\theta}) T_i(\underline{x}) + d(\underline{\theta}) + S(\underline{x}) \right] \right\} I_A(\underline{x})$$

Pelo Teorema da Fatorização o vetor $\underline{T}(\underline{x}) = \left[T_1(\underline{x}),...,T_k(\underline{x}) \right]'$ é suficiente para $\underline{\theta}' = \left(\theta_1, \theta_2,..., \theta_k \right)$.

Teorema da Fatorização ou de Neyman-Fisher: Seja uma amostra aleatória $\left[X_1, X_2,..., X_n \right]$ de uma distribuição $f(x;\theta)$, $\theta \in \Theta$. A estatística $T(\underline{x})$ é suficiente para θ se e somente se existe função $g(t,\theta)$, definida para todo t e para todo $\theta \in \Theta$, e $h(\underline{x})$ definida em R^n tal que: $P(\underline{x},\theta) = g(T(\underline{x}),\theta)\, h(\underline{x})$.

Cita-se, ainda, o Teorema da Família Exponencial para Estatísticas Suficientes e Completas:

Seja $\left\{ P_\theta / \theta \in \Theta \right\}$ uma família exponencial k-paramétrica dada por $p(\underline{x},\underline{\theta}) = \left\{ \exp\left[\sum_{i=1}^{k} c_i(\underline{\theta}) T_i(\underline{x}) + d(\underline{\theta}) + S(\underline{x}) \right] \right\} I_A(\underline{x})$. Suponha que a variação de $\underline{C} = \left[C_1(\theta), C_2(\theta),..., C_k(\theta) \right]$ tenha um interior não-vazio. Então $\underline{T}(\underline{x}) = \left[T_1(\underline{x}),..., T_k(\underline{x}) \right]$ é uma estatística suficiente e completa.

2.6 MÉTODOS DE ESTIMAÇÃO DOS PARÂMETROS

Diferentes métodos foram desenvolvidos para a estimação dos parâmetros. Citam-se os métodos de máxima verossimilhança e o dos momentos.

2.6.1 Método de Máxima Verossimilhança

Tem-se que X é a variável aleatória, e θ o parâmetro. A função de verossimilhança L é a função onde θ passa a ser a variável e X uma informação dada, de forma que $L(\theta,\underline{x}) = p(\theta,\underline{x})$.

No método da máxima verossimilhança, procura-se achar o valor $u(x)$ do parâmetro θ que maximiza $L(\theta,\underline{x})$ para cada valor de X. Sendo possível isso, $u(x)$ é o estimador de máxima verossimilhança de θ.

Sendo a função logaritmo natural (ln) uma função estritamente crescente, o valor máximo de $p(\theta, \underline{x})$ irá ocorrer no mesmo ponto do valor máximo de $\ln[L(\theta, \underline{x})]$.

Existindo o estimador de máxima verossimilhança $(\hat{\theta})$, deve-se verificar:

$$\frac{\partial \ln[p(\theta, \underline{x})]}{\partial \theta} = 0 \text{ em } \theta = \hat{\theta}$$

Deve-se citar um teorema importante para a obtenção do estimador de máxima verossimilhança, apresentado em CHAVES NETO (2002a):

Teorema da Família Exponencial para Estimador de Máxima Verossimilhança

Seja $p(\underline{x}, \underline{\theta}) = \left\{ \exp\left[\sum_{i=1}^{k} c_i(\theta) T_i(\underline{x}) + d(\theta) + S(\underline{x}) \right] \right\} I_A(\underline{x})$, $x \in A$, $\theta \in \Theta$ e seja C que denota o interior da variação de $\underline{C}(\theta)$, $\{c_1(\theta), c_2(\theta), c_3(\theta), ..., c_k(\theta)\}$. Se as equações: $E_\theta[T_i(\underline{x})] = T_i(\underline{x})$ para $i = 1,2,3,...,k$ têm solução $\underline{\hat{\theta}}' = (\hat{\theta}_1(\underline{x}), \hat{\theta}_2(\underline{x}), ..., \hat{\theta}_k(\underline{x}))$ para as quais $\{c_1(\underline{\hat{\theta}}(\underline{x})), c_2(\underline{\hat{\theta}}(\underline{x})), ..., c_k(\underline{\hat{\theta}}(\underline{x}))\} \in C$, então $\hat{\theta}$ é o único estimador de máxima verossimilhança de θ.

2.6.2 Método dos Momentos

É um método para obter estimadores de parâmetros, baseado na combinação do momento amostral com a correspondente distribuição de momentos. Seja $m'_j = E(X^j)$, que representa o j-ésimo momento de X no ponto 0.

Seja M'_j o j-ésimo momento amostral dado por:

$$M'_j = \frac{1}{n} \sum_{i=1}^{n} x_i^j, \quad j = 1, 2, 3, ..., k$$

Formando as equações:

$$M'_j = m'_j = f(\theta_1, \theta_2, ..., \theta_k), \quad j = 1, 2, 3, ..., k$$

Admitindo-se que tem solução única, $\hat{\theta}_j(X_1, X_2, ..., X_k)$, $j = 1, 2, 3, ..., k$. Estes k estimadores, solução do sistema de equações, são os estimadores dos parâmetros pelo método dos momentos.

2.7 TESTES PARAMÉTRICOS E NÃO-PARAMÉTRICOS

2.7.1 Testes Paramétricos

Quando é possível conhecer a distribuição de probabilidades teórica da variável em estudo, pode-se estimar os parâmetros e realizar testes de hipóteses para os mesmos de forma otimizada. Estes testes são conhecidos como testes paramétricos.

Os testes paramétricos incluem o requisito de que a variável em análise tenha distribuição de probabilidade conhecida. Também supõem que a variável tenha sido medida no mínimo em nível intervalar, e para alguns casos há a necessidade de as variáveis envolvidas terem as variâncias homogêneas (homocedasticidade).

2.7.2 Testes Não-Paramétricos

Um teste é não-paramétrico quando não há suposições formuladas sobre a natureza ou a forma das distribuições populacionais. Estes testes são chamados também de testes livres de distribuição. Dentre os testes não-paramétricos citam-se os de aderência.

2.7.2.1 Testes de aderência

A hipótese a ser testada refere-se à forma da distribuição da população. Admite-se, por hipótese, que a distribuição da variável em estudo siga o comportamento de uma distribuição teórica de probabilidade, na população.

Dentre os testes de aderência mais comuns cita-se o Qui-quadrado e de Kolmogorov-Smirnov.

No método de Kolmogorov-Smirnov a estatística do teste é a maior diferença observada entre a função de distribuição acumulada da distribuição teórica e a da variável em estudo.

O teste consiste na verificação do valor $d = \max\left|F(x) - G(x)\right|$ e da comparação com um valor crítico tabelado em função do nível de significância (α) e o tamanho da amostra (n). O teste é unilateral, rejeitando-se a hipótese H_0 de que a variável em estudo segue a distribuição de probabilidade ajustada na população, se *d* for maior que o valor crítico.

No método qui-quadrado calcula-se a estatística através da expressão:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_{oi} - f_{ei})^2}{f_{ei}} \tag{2.45}$$

onde: χ^2 é o qui-quadrado calculado;

f_{oi} é a freqüência observada de uma determinada classe ou valor da variável;

f_{ei} é a freqüência esperada, segundo modelo testado, dessa classe ou valor da variável;

$n = \sum_{i=1}^{k} f_{oi} = \sum_{i=1}^{k} f_{ei}$ é o número de observações da amostra;

k é o número de classes ou valores distintos observados na amostra.

O teste também é unilateral e rejeita-se H_0 quando o valor de χ^2 calculado for superior ao valor crítico.

3 MEDIDAS DE CORRELAÇÃO

3.1 INTRODUÇÃO

Em estudos que envolvem duas ou mais variáveis, é comum o interesse em conhecer o relacionamento entre elas, além das estatísticas descritivas normalmente calculadas.

A medida que mostra o grau de relacionamento entre duas variáveis, como se viu no Capítulo 2, é chamada de coeficiente de correlação. É também conhecida como medida de associação, de interdependência, de intercorrelação ou de relação entre as variáveis.

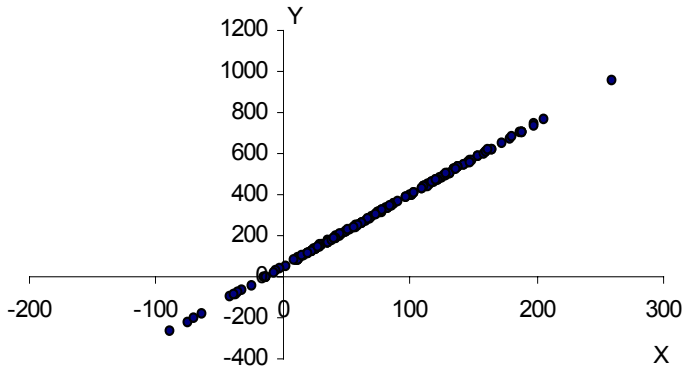
Diferentes formas de correlação podem existir entre as variáveis. O caso mais simples e mais conhecido é a correlação simples, envolvendo duas variáveis, X e Y. A relação entre duas variáveis será linear quando o valor de uma pode ser obtido aproximadamente por meio da equação da reta. Assim, é possível ajustar uma reta da forma $Y = \alpha + \beta X$ aos dados. Neste caso, a correlação é linear simples.

Entretanto, quando não for possível o ajuste da equação anterior, não significa que não existe correlação entre elas. Poderá haver correlação não-linear entre as mesmas.

Uma forma simples de verificar o tipo de correlação existente entre duas variáveis é através do gráfico chamado “diagrama de dispersão”. Trata-se de um gráfico onde são representados os pares (X_i, Y_i), $i = 1, 2, ..., n$, onde $n =$ número total de observações. Os gráficos 1, 2, 3 e 4 representam o “diagrama de dispersão” entre as variáveis X e Y.

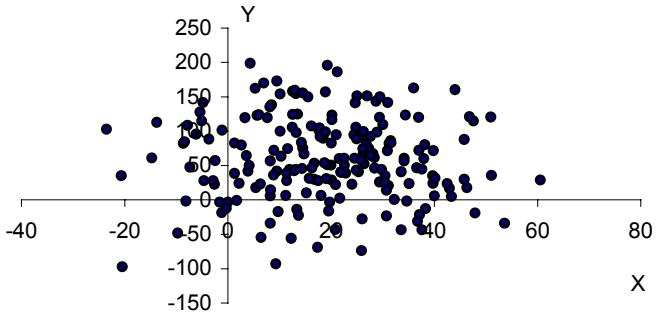
O método que permite estudar as relações ou associações é conhecido como Análise de Correlação. Esta análise mostra o grau de relacionamento entre as variáveis, fornecendo um número, indicando como as variáveis variam conjuntamente. Não há a necessidade de definir as relações de causa e efeito, ou seja, qual é a variável dependente e a independente. Os diagramas de dispersão a seguir mostram os tipos de correlações entre duas variáveis.

GRÁFICO 1 - CORRELAÇÃO LINEAR POSITIVA PERFEITA ENTRE AS VARIÁVEIS X E Y



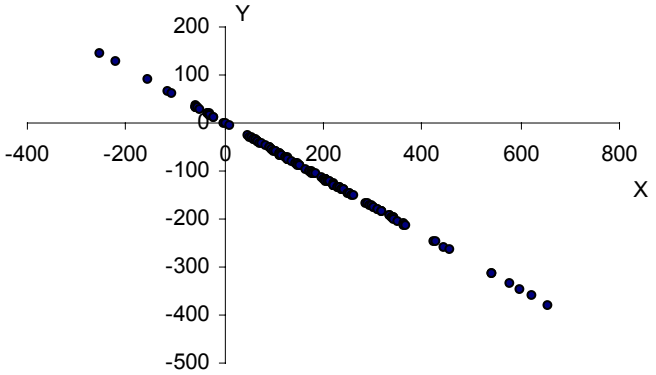
FONTE: A autora

GRÁFICO 2 - CORRELAÇÃO LINEAR NULA ENTRE AS VARIÁVEIS X E Y



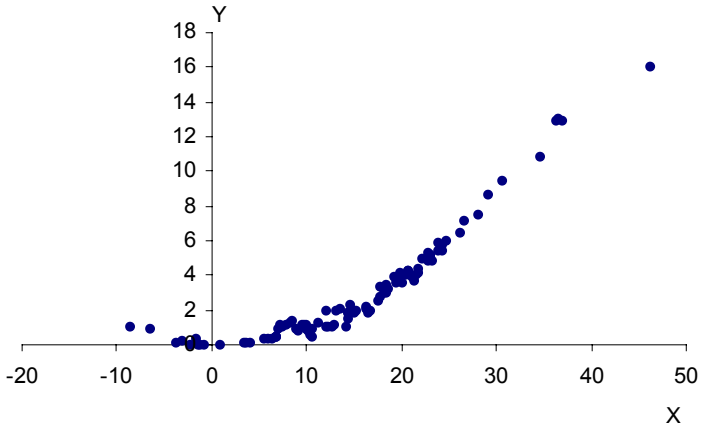
FONTE: A autora

GRÁFICO 3 - CORRELAÇÃO LINEAR NEGATIVA PERFEITA ENTRE AS VARIÁVEIS X E Y



FONTE: A autora

GRÁFICO 4 - CORRELAÇÃO NÃO-LINEAR ENTRE AS VARIÁVEIS X E Y



FONTE: A autora

Quando a análise envolve grande número de variáveis e há interesse em conhecer a correlação duas a duas, é comum a construção de uma matriz de correlações. Esta é uma matriz formada pelas correlações entre as variáveis X_i e X_j, $i \neq j$, fora da diagonal e 1 na diagonal, indicando a correlação das variáveis X_i e X_j, sendo $i = j$.

Pode ocorrer, ainda, situação onde se tem dois conjuntos de variáveis, um composto por uma variável (Y) e o outro com p variáveis (X_1, X_2, ..., X_p), e se deseja analisar a correlação entre a variável Y e a variável X_i, $i = 1, 2, ..., p$. Neste caso a correlação é chamada de múltipla e calculada por $R = \sqrt{\frac{SQ_{Regr}}{SQ_{Total}}}$, detalhada na seção 3.3.2. Evidentemente, o relacionamento entre Y e X_1, X_2, ..., X_p pode ser expresso pelo hiperplano $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$, admitindo relação linear entre Y e X_1, X_2, ..., X_p.

Ainda, se o interesse é analisar a correlação entre dois conjuntos de variáveis, X_i, $i = 1, 2, ..., p$ e Y_j, $j = 1, 2, ..., q$ sendo $p \leq q$, é possível utilizar a técnica de Análise Multivariada, conhecida como Análise de Correlação Canônica.

É possível, resumidamente, reunir os métodos de Análise de Correlação, os quais foram tratados neste trabalho em: Análise de Correlação Simples Linear e Não-linear, Análise de Correlação Linear Múltipla e Análise de Correlação Canônica.

Antes de aplicar qualquer método estatístico paramétrico é necessário verificar se as suposições (tais como Gaussianidade, homocedasticidade, independência) do modelo estão sendo razoavelmente satisfeitas, através de uma análise exploratória dos dados. Para SIQUEIRA (1983), a falha de uma das suposições altera o nível de significância do teste estatístico. O pesquisador pode pensar estar testando, por exemplo, a um nível de significância de 5%, e na realidade estar testando a um nível maior. Além disso, é possível causar perda de precisão das estimativas obtidas.

3.2 MEDIDAS DE CORRELAÇÃO ENTRE DUAS VARIÁVEIS

Para McNEMAR (1969), as situações mais freqüentes, na prática, para as quais as medidas de correlação simples são necessárias, podem ser agrupadas como se segue:

a) ocorrem medida contínua para uma variável e duas categorias para a outra variável;
b) ambas as variáveis são dicotomizadas;
c) ocorrem três ou mais categorias para uma variável e duas ou mais para a segunda;
d) ocorrem três ou mais categorias para uma variável e uma medida contínua para outra;
e) quando os dados são postos *(ranks)*;
f) as duas variáveis são contínuas.

Segundo DOWNIE e HEATH (1959), existem situações em que o relacionamento entre as duas variáveis não é linear, ou uma delas não é contínua, ou o número de pares das medidas é muito pequeno. Então, para cada uma dessas situações há necessidade de uma medida adequada de associação entre as variáveis.

3.2.1 Coeficiente de Correlação Linear de Pearson e a Distribuição Normal Bivariada

O método usualmente conhecido para medir a correlação entre duas variáveis é o Coeficiente de Correlação Linear de Pearson, também conhecido como Coeficiente de Correlação do Momento Produto. Este foi o primeiro método de correlação, estudado por Francis Galton e seu aluno Karl Pearson, em 1897[5] (SCHULTZ e SCHULTZ, 1992).

Este coeficiente de correlação é utilizado na Análise de Componentes Principais, Análise Fatorial, Análise de Confiabilidade, entre outras, que serão apresentadas neste trabalho.

O coeficiente de correlação populacional (parâmetro) ρ e sua estimativa amostral $\hat{\rho}$ estão intimamente relacionados com a distribuição normal bivariada, definida na seção 2.4.2.5.

Considerando a população normal bivariada, onde X é uma variável normalmente distribuída, com média μ_x e desvio padrão σ_x, e Y variável também normalmente distribuída com média μ_y e desvio padrão σ_y, a expressão matemática da distribuição (função densidade de probabilidade) é dada pela expressão abaixo, conforme já apresentada na seção 2.4.2.5 do Capítulo 2.

$$f_{X,Y}(X,Y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)+\left(\frac{Y-\mu_Y}{\sigma_Y}\right)^2\right]\right\} \quad (3.1)$$

onde a variação dos parâmetros é:

$$\mu_X \in R,\ \mu_Y \in R,\ \sigma_X^2 \in R^+,\ \sigma_Y^2 \in R^+ \text{ e } -1 \le \rho \le +1$$

Essa função contém os parâmetros obtidos no Capítulo 2: μ_X, μ_Y, σ_X^2, σ_Y^2 e ρ, onde ρ é o coeficiente de correlação para a população normal bivariada, e varia entre –1 e +1. O coeficiente de correlação ρ é definido como:

---

[5]Esta informação foi obtida no *site*: www.ime.br/~abe/cronologiajaneiro02.doc

$$\rho_{X,Y} = \rho = \frac{COV(X,Y)}{\sigma_X \sigma_Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \tag{3.2}$$

A covariância é uma medida que expressa a variação conjunta de duas variáveis, cuja expressão é dada por:

$$COV(X,Y) = E\left[(X - \mu_X)(Y - \mu_Y)\right] \tag{3.3}$$

Ela depende da escala das medidas, o que impossibilita a idéia de como as duas variáveis estão relacionadas. Quando se padroniza as variáveis tem-se o coeficiente de correlação, conforme expressão (3.2) acima, ou seja,

$$\rho = \frac{COV(X,Y)}{\sigma_X \sigma_Y} = \frac{E\left[(X - \mu_X)(Y - \mu_Y)\right]}{\sigma_X \sigma_Y} = E\left[\left(\frac{X - \mu_X}{\sigma_X}\right)\left(\frac{Y - \mu_Y}{\sigma_Y}\right)\right] = COV\left(Z_X, Z_Y\right) \tag{3.4}$$

e, é claro, a noção de associação entre as varáveis é percebida mais facilmente.

3.2.1.1 Estimadores de máxima verossimilhança

Os estimadores de máxima verossimilhança dos parâmetros μ_X, μ_Y, σ_X^2, σ_Y^2 e ρ são obtidos pelo resultado a seguir.

Resultado 3.1: Sejam n pares de observações $\left[(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\right]$ do vetor aleatório $\left[\underline{X}, \underline{Y}\right]$ que se distribui conforme a distribuição normal bivariada, ou seja, $\left[\underline{X}, \underline{Y}\right] \sim N_2\left(\underline{\mu}, \Sigma\right)$, com $\underline{\mu}' = \left[\mu_X, \mu_Y\right]$ e $\Sigma = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_Y\sigma_X & \sigma_Y^2 \end{bmatrix}$ e f.d.p. igual a

$$f_{X,Y}(X,Y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right) + \left(\frac{Y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}$$

Então, os estimadores de máxima verossimilhança dos parâmetros são:

$$\hat{\mu}_X = \overline{X},\ \hat{\mu}_Y = \overline{Y},\ \hat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2,\ \hat{\sigma}_Y^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 \quad \text{e}$$

$$\hat{\rho} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}}$$

Prova: A f.d.p. na forma da função distribuição de probabilidade conjunta é dada por:

$$f_{X,Y}(\underline{X},\underline{Y}) = \left[\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\right]^n \exp\left\{\frac{-1}{2(1-\rho^2)}\left[\sum_{i=1}^{n}\left(\frac{X_i-\mu_X}{\sigma_X}\right)^2 - 2\rho\sum_{i=1}^{n}\left(\frac{X_i-\mu_X}{\sigma_x}\right)\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right) + \sum_{i=1}^{n}\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)^2\right]\right\}$$

Passando para a forma da família exponencial:

$$f_{X,Y}(\underline{X},\underline{Y}) = \left\{\exp\left[-n\ln\left(2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}\right) - \frac{1}{2(1-\rho^2)}\left[\sum_{i=1}^{n}\left(\frac{X_i-\mu_X}{\sigma_x}\right)^2 - 2\rho\sum_{i=1}^{n}\left(\frac{X_i-\mu_X}{\sigma_x}\right)\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right) + \sum_{i=1}^{n}\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)^2\right]\right]\right\}$$

$$f_{X,Y}(\underline{X},\underline{Y}) = \left\{\exp\left[-n\ln\left(2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}\right) - \frac{1}{2(1-\rho^2)\sigma_X^2}\sum_{i=1}^{n}x_i^2 + \frac{\mu_X}{(1-\rho^2)\sigma_X^2}\sum_{i=1}^{n}x_i - \frac{n\mu_X^2}{2(1-\rho^2)\sigma_X^2}\right.\right.$$

$$- \frac{1}{2(1-\rho^2)\sigma_Y^2}\sum_{i=1}^{n}Y_i^2 + \frac{\mu_Y}{(1-\rho^2)\sigma_Y^2}\sum_{i=1}^{n}Y_i - \frac{n\mu_Y^2}{2(1-\rho^2)\sigma_Y^2} + \frac{\rho}{(1-\rho^2)\sigma_X\sigma_Y}\sum_{i=1}^{n}X_iY_i$$

$$\left.\left. - \frac{\rho\mu_Y}{(1-\rho^2)\sigma_X\sigma_Y}\sum_{i=1}^{n}X_i - \frac{\rho\mu_X}{(1-\rho^2)\sigma_X\sigma_Y}\sum_{i=1}^{n}Y_i + \frac{n\rho\mu_X\mu_Y}{(1-\rho^2)\sigma_X\sigma_Y}\right]\right\}$$

Pelo teorema da família exponencial k-paramétrica (definição 2 da seção 2.5) para estatísticas suficientes, tem-se que:

$$c_1(\theta) = \frac{\mu_X}{(1-\rho^2)\sigma_X^2} - \frac{\rho\mu_Y}{(1-\rho^2)\sigma_X\sigma_Y} \quad \text{e} \quad T_1(\underline{X}) = \sum_{i=1}^{n}X_i$$

$$c_2(\theta) = \frac{\mu_Y}{(1-\rho^2)\sigma_Y^2} - \frac{\rho\mu_X}{(1-\rho^2)\sigma_X\sigma_Y} \quad \text{e} \quad T_2(\underline{Y}) = \sum_{i=1}^{n}Y_i$$

$$c_3(\theta) = \frac{-1}{2(1-\rho^2)\sigma_X^2} \quad \text{e} \quad T_3(\underline{X}) = \sum_{i=1}^{n}X_i^2$$

$$c_4(\theta) = \frac{-1}{2(1-\rho^2)\sigma_Y^2} \quad \text{e} \quad T_4(\underline{Y}) = \sum_{i=1}^{n}Y_i^2$$

$$c_5(\theta) = \frac{\rho}{(1-\rho^2)\sigma_X\sigma_Y} \quad \text{e} \quad T_5(\underline{X},\underline{Y}) = \sum_{i=1}^{n}X_iY_i$$

Aplicando o Teorema da Família Exponencial para Estimador de Máxima Verossimilhança (seção 2.6.1) para a obtenção dos estimadores:

$$E\left[T_i(\underline{X})\right] = T_i(\underline{X})$$

Estimador de Máxima Verossimilhança (EMV) para μ_X

$$T_1(\underline{X}) = \sum_{i=1}^{n} X_i$$

$$E\left[T_1(\underline{X})\right] = \sum_{i=1}^{n} X_i$$

$$n\mu_X = \sum_{i=1}^{n} X_i$$

$$\hat{\mu}_X = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X} \qquad (3.5)$$

Estimador de Máxima Verossimilhança (EMV) para μ_Y

$$T_2(\underline{Y}) = \sum_{i=1}^{n} Y_i$$

$$E\left[T_2(\underline{Y})\right] = \sum_{i=1}^{n} Y_i$$

$$n\mu_Y = \sum_{i=1}^{n} Y_i$$

$$\hat{\mu}_Y = \frac{\sum_{i=1}^{n} Y_i}{n} = \overline{Y} \qquad (3.6)$$

Estimador de Máxima Verossimilhança (EMV) para σ_X^2

$$T_3(\underline{X}) = \sum_{i=1}^{n} X_i^2$$

$$E\left[T_3(\underline{X})\right] = \sum_{i=1}^{n} X_i^2$$

$$\sum_{i=1}^{n}\left[V(X_i)+E^2(X_i)\right]=\sum_{i=1}^{n}X_i^2$$

$$n\sigma_X^2+n\mu_X^2=\sum_{i=1}^{n}X_i^2$$

$$\hat{\sigma}_X^2=\frac{1}{n}\left[\sum_{i=1}^{n}X_i^2-n\mu_X^2\right]=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2 \qquad (3.7)$$

Estimador de Máxima Verossimilhança (EMV) para σ_Y^2

$$T_4(\underline{Y})=\sum_{i=1}^{n}Y_i^2$$

$$E\left[T_4(\underline{Y})\right]=\sum_{i=1}^{n}Y_i^2$$

$$\sum_{i=1}^{n}\left[V(Y_i)+E^2(Y_i)\right]=\sum_{i=1}^{n}Y_i^2$$

$$n\sigma_Y^2+n\mu_Y^2=\sum_{i=1}^{n}Y_i^2$$

$$\hat{\sigma}_Y^2=\frac{1}{n}\left[\sum_{i=1}^{n}Y_i^2-n\mu_Y^2\right]=\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2 \qquad (3.8)$$

Estimador de Máxima Verossimilhança (EMV) para ρ

$$T_5(\underline{X},\underline{Y})=\sum_{i=1}^{n}X_iY_i$$

$$E\left[T_5(\underline{X},\underline{Y})\right]=T_5(\underline{X},\underline{Y})$$

$$E\left[\sum_{i=1}^{n}X_iY_i\right]=\sum_{i=1}^{n}X_iY_i$$

$$\sum_{i=1}^{n}\left[E(X_i)E(Y_i)+\mathrm{cov}(X_i,Y_i)\right]=\sum_{i=1}^{n}X_iY_i$$

$$\sum_{i=1}^{n}\left[E(X_i)E(Y_i)+\rho\sigma_X\sigma_Y\right]=\sum_{i=1}^{n}X_iY_i$$

$$n\mu_X\mu_Y+\rho n\sigma_X\sigma_Y=\sum_{i=1}^{n}X_iY_i$$

$$\hat{\rho}_{X,Y} = \frac{\sum_{i=1}^{n} x_i Y_i - n\overline{X}\,\overline{Y}}{n\hat{\sigma}_X\hat{\sigma}_Y} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{n\hat{\sigma}_X\hat{\sigma}_Y} \tag{3.9}$$

$$\hat{\rho}_{X,Y} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{n\sqrt{\sum_{i=1}^{n}\frac{\left(x_i - \overline{X}\right)^2}{n}}\sqrt{\sum_{i=1}^{n}\frac{\left(Y_i - \overline{Y}\right)^2}{n}}} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}} \tag{3.10}$$

Então, pelo Teorema da Família Exponencial para Estimador de Máxima Verossimilhança, $\hat{\rho}$ é o único estimador de máxima verossimilhança de ρ.

Fazendo $x_i = X_i - \overline{X}$ e $y_i = Y_i - \overline{Y}$ a expressão acima poderá ser escrita da seguinte forma:

$$\hat{\rho}_{X,Y} = \frac{\sum_{i=1}^{n} x_i y_i}{n\hat{\sigma}_X\hat{\sigma}_Y} = \frac{\sum_{i=1}^{n} x_i y_i}{n\sqrt{\sum_{i=1}^{n}\frac{x_i^2}{n}}\sqrt{\sum_{i=1}^{n}\frac{y_i^2}{n}}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sum_{i=1}^{n} y_i^2} \tag{3.11}$$

Este coeficiente de correlação é também chamado de "coeficiente de correlação do momento produto", porque é calculado multiplicando-se os escores Z de duas variáveis (produto de duas variáveis) e então calcula-se a média (momento) do produto de um grupo de n observações (CHEN e POPOVICH, 2002).

$$\hat{\rho}_{X,Y} = \frac{\sum_{i=1}^{n} Z_{x_i} Z_{Y_i}}{n} \tag{3.12}$$

onde:

$$Z_{X_i} = \frac{X_i - \overline{X}}{\hat{\sigma}_X} \quad \text{e} \quad Z_{Y_i} = \frac{Y_i - \overline{Y}}{\hat{\sigma}_Y}$$

3.2.1.2 Suposições básicas para a utilização do Coeficiente de Correlação Linear de Pearson

A suposição básica para a utilização deste coeficiente é de que o relacionamento entre as duas variáveis seja linear, ou seja, é adequado para medir o relacionamento linear.

A segunda hipótese é de que as variáveis envolvidas sejam aleatórias e que sejam medidas no mínimo em escala intervalar.

Uma terceira hipótese é de que as duas variáveis tenham uma distribuição normal bivariada conjunta, o que equivale a dizer que para cada X dado, a variável Y é normalmente distribuída. Esta hipótese é necessária para fazer inferências estatísticas (teste de hipótese e intervalo de confiança), sendo dispensável quando se tratar de estudos amostrais.

Esta última hipótese é imprescindível para amostras pequenas, segundo BUNCHAFT e KELLNER (1999), e diminui a importância à medida que aumenta o tamanho da amostra, o que é justificado pelo Teorema Central do Limite para distribuições multivariadas apresentado em JOHNSON e WICHERN (1988, p.145).

Segundo SNEDECOR e COCHRAN (1980), na prática muitas vezes a distribuição bivariada de interesse está longe de ser normal. Assim, é possível fazer uma transformação de variáveis de forma que se aproxime da distribuição normal bivariada conjunta. Assim, torna-se possível estimar ρ na nova escala. Um dos objetivos das transformações, segundo SIQUEIRA (1983), é a correção da não-normalidade e também a homogeneização da variância das variáveis envolvidas na análise.

As transformações são lineares quando envolvem apenas uma mudança de origem e/ou de escala, podendo-se citar, como exemplo, a padronização de uma variável (Z). Este tipo de transformação não afeta as características essenciais de uma análise estatística (SIQUEIRA, 1983). A transformação linear não afeta a heterogeneidade das variâncias, e se a variável Y não é normal, uma transformação linear de Y não será normal. Entretanto, as transformações mais importantes são as não-lineares, em que um certo incremento na escala original normalmente não corresponde ao mesmo incremento na nova escala, que é o fator responsável pelo efeito da correção dos desvios das suposições.

Uma característica importante na transformação é que esta mantenha a relação de ordem, ou seja, que a ordenação das observações seja preservada. Uma

transformação $g(Y)$ é chamada monotônica estritamente crescente se para $\forall y' > y''$ implica necessariamente que $g(y') > g(y'')$.

As transformações não-lineares usuais são: logarítmica (qualquer base, embora as mais utilizadas sejam a base 10 e a natural), raiz quadrada, recíproca $\left(z = \frac{1}{y}\right)$ e angular $\left(\text{arc sen}\sqrt{y}\right)$.

3.2.1.3 Interpretação do Coeficiente de Correlação Linear de Pearson

Na prática, o coeficiente $(\hat{\rho})$ é interpretado como um indicador que descreve a interdependência entre as variáveis X e Y, com a forma $\hat{Y} = \hat{\alpha} + \hat{\beta}X$, onde $\hat{\alpha}$ e $\hat{\beta}$ são constantes.

A interpretação do coeficiente quando $\left|\hat{\rho}\right| = 1$ é de que existe correlação linear perfeita entre as variáveis X e Y. A correlação é linear perfeita positiva quando $\hat{\rho} = 1$ e linear perfeita negativa quando $\hat{\rho} = -1$. Quando se tem $\hat{\rho} = 0$, não existe correlação linear entre as variáveis X e Y.

Entretanto, na prática ocorrem diferentes valores de $(\hat{\rho})$. A interpretação do valor de $\hat{\rho}$ depende muito dos objetivos de sua utilização e as razões pelas quais este é calculado. Segundo CALLEGARI-JACQUES (2003, p. 90), o coeficiente de correlação pode ser avaliado qualitativamente da seguinte forma:

se $0{,}00 < \left|\hat{\rho}\right| < 0{,}30$, existe fraca correlação linear;

se $0{,}30 \leq \left|\hat{\rho}\right| < 0{,}60$, existe moderada correlação linear;

se $0{,}60 \leq \left|\hat{\rho}\right| < 0{,}90$, existe forte correlação linear;

se $0{,}90 \leq \left|\hat{\rho}\right| < 1{,}00$, existe correlação linear muito forte.

Resultado 3.2: A relação existente entre o coeficiente da correlação estimado $(\hat{\rho})$ e o coeficiente angular estimado $(\hat{\beta})$ pode ser expressa conforme apresentada a seguir:

$$\hat{\rho}_{Y,X} = \hat{\beta}_{Y,X}\frac{S_X}{S_Y} \tag{3.13}$$

onde: $x_i = X_i - \overline{X}$ e $y_i = Y_i - \overline{Y}$

$$S_x = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n-1}} \quad \text{e} \quad S_Y = \sqrt{\frac{\sum_{i=1}^{n} y_i^2}{n-1}}$$

Prova:

Resolvendo as equações normais[6] da reta pelo método dos mínimos quadrados,

$$\text{tem-se que } \hat{\beta}_{Y,X} = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \qquad (3.14)$$

A expressão (3.11) poderá ser escrita como $\hat{\rho}_{Y,X} = \frac{\sum_{i=1}^{n} x_i y_i}{(n\text{-}1)\, S_x S_Y}$, utilizando-se os denominadores $(n-1)$, com $S_x = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n-1}}$ e $S_Y = \sqrt{\frac{\sum_{i=1}^{n} y_i^2}{n-1}}$.

Esta expressão poderá ser escrita como se segue:

$$\hat{\rho}_{Y,X} = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{(n\text{-}1) S_x S_Y} \qquad (3.15)$$

---

[6]Equações normais da reta: $\sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i$

$$\sum_{i=1}^{n} X_i Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2$$

A expressão (3.13) poderá ser escrita da seguinte forma:

$$\hat{\beta}_{X,Y} = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} \tag{3.16}$$

Substituindo a expressão (3.15) em (3.16) e dividindo por (n - 1), tem-se:

$$\hat{\beta}_{Y,X} = \frac{\hat{\rho}_{Y,X}(n-1)S_X S_Y}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\hat{\rho}_{Y,X} S_X S_Y}{S_X^2 S_Y} = \hat{\rho}_{Y,X}\frac{S_Y}{S_X}$$

Portanto: $\hat{\rho}_{Y,X} = \hat{\beta}_{Y,X}\frac{S_X}{S_Y}$

O coeficiente de correlação de X e Y é o mesmo que Y e X. O valor absoluto do coeficiente $\hat{\rho}$ não é afetado por qualquer transformação linear de X ou Y. Para ANDERBERG (1973), o coeficiente de correlação $\hat{\rho}$ é invariante frente às transformações lineares e quase-invariante em relação às transformações monotônicas.

Outra forma de interpretar o Coeficiente de Correlação é em termos de $\hat{\rho}^2$, denominado Coeficiente de Determinação ou de Explicação. Quando multiplicado por 100, o $\hat{\rho}^2 = \hat{R}^2$ fornece a percentagem da variação em Y (variável dependente), que pode ser explicada pela variação em X (variável independente), ou seja, o quanto de variação é comum às duas variáveis.

Resultado 3.3: A variação total da variável resposta Y é definida como $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ e pode ser decomposta em variação não-explicada mais a variação explicada pelo modelo $Y = f(X) + \varepsilon$, e pode ser escrita sob a forma:

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 \tag{3.17}$$

Prova:

Fazendo: $Y_i - \overline{Y} = \hat{Y}_i - \overline{Y} + Y_i - \hat{Y}_i$

e elevando ambos os membros ao quadrado, tem-se:

$$\left(Y_i - \overline{Y}\right)^2 = \left[\left(\hat{Y}_i - \overline{Y}\right) + \left(Y_i - \hat{Y}_i\right)\right]^2$$

Desenvolvendo o binômio e fazendo o somatório, obtém-se:

$$\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 + 2\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)\left(Y_i - \hat{Y}_i\right) + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

Tem-se que mostrar que $2\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)\left(Y_i - \hat{Y}_i\right) = 0$

Sabe-se que $\left(Y_i - \hat{Y}_i\right) = \hat{\varepsilon}_i$

Então

$$2\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)\hat{\varepsilon}_i = 2\sum_{i=1}^{n}\left(\hat{Y}_i\,\hat{\varepsilon}_i\right) - 2\overline{Y}\sum_{i=1}^{n}\hat{\varepsilon}_i$$

Mas $\sum_{i=1}^{n}\hat{\varepsilon}_i = 0$ e

$$2\sum_{i=1}^{n}\hat{\varepsilon}_i\hat{Y}_i = 2\sum_{i=1}^{n}\hat{\varepsilon}_i\left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)$$

$$2\sum_{i=1}^{n}\hat{\varepsilon}_i\hat{Y}_i = 2\hat{\beta}_0\sum_{i=1}^{n}\hat{\varepsilon}_i + 2\hat{\beta}_1\sum_{i=1}^{n}X_i\,\hat{\varepsilon}_i$$

$$2\sum_{i=1}^{n}\hat{\varepsilon}_i\hat{Y}_i = 2\hat{\beta}_1\sum_{i=1}^{n}X_i\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)$$

$2\sum_{i=1}^{n}\hat{\varepsilon}_i\hat{Y}_i = 2\hat{\beta}_1\sum_{i=1}^{n}X_i\,\hat{\varepsilon}_i = 0$, pois

$$\sum_{i=1}^{n}X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = \sum_{i=1}^{n}\left[X_iY_i - \hat{\beta}_0X_i - \hat{\beta}_1X_i^2\right] = \sum_{i=1}^{n}X_iY_i - \hat{\beta}_0\sum_{i=1}^{n}X_i - \hat{\beta}_1\sum_{i=1}^{n}X_i^2 = 0$$

Pois é a 2.ª equação normal do sistema de equações do método dos mínimos quadrados (ver nota de rodapé referente ao resultado 3.2).

Logo: $\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$

Os desvios $\left(\hat{Y}_i - \overline{Y}\right)$ têm um padrão definido, enquanto $\left(Y_i - \hat{Y}\right)$ comportam-se de forma imprevisível ou casual. O coeficiente entre a variação explicada (VE) pelo modelo e a variação total (VT) é chamado de coeficiente de determinação $\left(\hat{\rho}^2\right)$, como apresentado a seguir:

$$\hat{R}^2 = \hat{\rho}^2_{X,Y} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = \frac{VE}{VT} = \frac{SQ_{Regr}}{SQ_{Total}} \qquad (3.18)$$

Este coeficiente indica a proporção da variação total de Y explicada pelo ajuste do modelo.

O valor de $\hat{R} = \hat{\rho}_{X,Y} = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}}$ é o coeficiente de correção múltipla, quando se tem o modelo $Y = f(\underline{x}) + \varepsilon = f\left(X_1, X_2, ..., X_p\right) + \varepsilon$.

3.2.1.4 Fatores que afetam o Coeficiente de Correlação Linear de Pearson

Dentre os fatores que afetam a intensidade do coeficiente de correlação $\hat{\rho}$, bem como a precisão ao estimar a correlação populacional ρ, podem ser citados o tamanho da amostra, principalmente quando é pequena, os *outliers* (valores discrepantes), a restrição da amplitude de uma das variáveis ou de ambas e erros de medidas.

Os *outliers* afetam o coeficiente de correlação, principalmente quando a amostra é pequena. Quando a amostra é grande, eles têm menor efeito sobre o coeficiente de correlação. Estes podem ser detectados na análise exploratória de dados, através de teste e gráficos próprios.

A restrição da amplitude de uma das variáveis ou de ambas pode ocorrer quando o pesquisador seleciona amostra homogênea para o estudo. Este fator é conhecido como “restrição da variabilidade”. Quando a amostra é homogênea, o

valor do coeficiente de correlação tende a diminuir, pois um dos fatores que influenciam na intensidade da correlação é a variabilidade da amostra. Quanto maior a variabilidade das variáveis envolvidas na análise, maior a correlação entre elas.

McNEMAR (1969), SILVEIRA e PINENT (2001) e CHEN e POPOVICH (2002) expõem o coeficiente de correlação sem a restrição da variabilidade, isto é, o coeficiente de correlação para o grupo total, com base no coeficiente de correlação do grupo restrito. Um exemplo de aplicação pode ser encontrado na seção 3.2.1.10.1.3.

Resultado 3.4: O estimador do coeficiente de correlação sem a restrição da variabilidade é expresso por: $$\hat{\rho}_{(X,Y)_T} = \frac{\hat{\rho}_{X,Y}\dfrac{S_{X_T}}{S_X}}{\sqrt{1-\hat{\rho}^2_{X,Y}+\hat{\rho}^2_{X,Y}\left(\dfrac{S_{X_T}}{S_X}\right)^2}} \quad (3.19)$$

ou $$\hat{\rho}_{(X,Y)_T} = \frac{\hat{\rho}_{X,Y}S_{X_T}}{S_X\sqrt{1-\hat{\rho}^2_{X,Y}\left[1-\left(\dfrac{S_{X_T}}{S_X}\right)^2\right]}} \quad (3.20)$$

onde:

$\hat{\rho}_{(X,Y)_T}$ é o coeficiente de correlação entre as variáveis X e Y estimado para o grupo total;

$\hat{\rho}_{X,Y}$ é o coeficiente de correlação entre as variáveis X e Y do grupo restrito;

S_X é o desvio padrão da variável X do grupo restrito;

S_{X_T} é o desvio padrão da variável X do grupo total.

Prova:

Deve-se considerar duas suposições básicas, que são a linearidade da regressão de Y em X e a homocedasticidade da distribuição normal bivariada. Com base na suposição de linearidade é possível igualar a declividade da linha de regressão do grupo restrito à declividade da linha de regressão do grupo total, considerando as duas regressões paralelas.

Sabe-se, do resultado 3.2, que $\hat{\rho} = \hat{\beta}\frac{S_X}{S_Y}$, portanto $\hat{\beta} = \hat{\rho}\frac{S_Y}{S_X}$ e se as duas regressões são paralelas é possível a seguinte igualdade:

$$\hat{\rho}_{X,Y}\frac{S_Y}{S_X} = \hat{\rho}_{(X,Y)_T}\frac{S_{Y_T}}{S_{X_T}} \tag{3.21}$$

A suposição de homocedasticidade implica a igualdade dos erros padrão da estimativa (S) das duas regressões. O erro padrão da estimativa, que será discutido no resultado 3.9, pode ser obtido através de:

$$S = S_Y\sqrt{1-\hat{\rho}^2}$$

Igualando-se os erros padrão, tem-se:

$$S_Y\sqrt{1-\hat{\rho}^2_{X,Y}} = S_{Y_T}\sqrt{1-\hat{\rho}^2_{(X,Y)_T}} \tag{3.22}$$

onde:

$\hat{\rho}_{(X,Y)_T}$ é o coeficiente de correlação entre as variáveis X e Y estimado para o grupo total;

$\hat{\rho}_{X,Y}$ é o coeficiente de correlação entre as variáveis X e Y do grupo restrito;

S_X é o desvio padrão da variável X do grupo restrito;

S_{X_T} é o desvio padrão da variável X do grupo total;

S_Y é o desvio padrão da variável Y do grupo restrito;

S_{Y_T} é o desvio padrão da variável Y do grupo total.

De (3.21) tem-se:

$$S_{Y_T} = \frac{\hat{\rho}_{X,Y}S_Y S_{X_T}}{\hat{\rho}_{(X,Y)_T}S_X} \tag{3.23}$$

De (3.22) segue-se que:

$$S_Y^2(1-\hat{\rho}^2_{X,Y}) = S_{Y_T}^2(1-\hat{\rho}^2_{(X,Y)_T}) \tag{3.24}$$

Substituindo o valor de S_{Y_T} (3.23) em (3.24) tem-se:

$$S_Y^2(1-\hat{\rho}_{X,Y}^2)=\left[\frac{\hat{\rho}_{X,Y}S_Y S_{X_T}}{\hat{\rho}_{(X,Y)_T}S_X}\right]^2(1-\hat{\rho}_{(X,Y)_T}^2)$$

Dividindo ambos os membros por S_Y^2 tem-se:

$$(1-\hat{\rho}_{X,Y}^2)=\frac{\hat{\rho}_{X,Y}^2 S_{X_T}^2}{\hat{\rho}_{(X,Y)_T}^2 S_X^2}(1-\hat{\rho}_{(X,Y)_T}^2)$$

$$(1-\hat{\rho}_{X,Y}^2)=\frac{\hat{\rho}_{X,Y}^2 S_{X_T}^2}{\hat{\rho}_{(X,Y)_T}^2 S_X^2}-\frac{\hat{\rho}_{X,Y}^2 S_{X_T}^2 \hat{\rho}_{(X,Y)_T}^2}{\hat{\rho}_{(X,Y)_T}^2 S_X^2}$$

$$1-\hat{\rho}_{X,Y}^2+\frac{\hat{\rho}_{X,Y}^2 S_{X_T}^2}{S_X^2}=\frac{\hat{\rho}_{X,Y}^2 S_{X_T}^2}{\hat{\rho}_{(X,Y)_T}^2 S_X^2}$$

$$\hat{\rho}_{(X,Y)_T}^2=\frac{\hat{\rho}_{X,Y}^2 S_{X_T}^2}{S_X^2-\hat{\rho}_{X,Y}^2 S_X^2+\hat{\rho}_{X,Y}^2 S_{X_T}^2}$$

Dividindo por S_X^2 obtém-se:

$$\hat{\rho}_{(X,Y)_T}^2=\frac{\hat{\rho}_{X,Y}^2\frac{S_{X_T}^2}{S_X^2}}{1-\hat{\rho}_{X,Y}^2+\hat{\rho}_{X,Y}^2\frac{S_{X_T}^2}{S_X^2}}$$

Portanto: $\hat{\rho}_{(X,Y)_T}=\dfrac{\hat{\rho}_{X,Y}\dfrac{S_{X_T}}{S_X}}{\sqrt{1-\hat{\rho}_{X,Y}^2+\hat{\rho}_{XY}^2\dfrac{S_{X_T}^2}{S_X^2}}}$

Os erros de medidas, devido a uma medição incorreta das variáveis, por diversas razões, também têm efeito sobre a correlação. BROWNLEE (1960) apresenta os efeitos de erros de medidas no coeficiente de correlação.

Resultado 3.5: O coeficiente de correlação entre as variáveis observadas X' e Y' ($\hat{\rho}_{X',Y'}$), com erros de medidas, é menor do que o coeficiente de correlação entre as variáveis verdadeiras X e Y ($\hat{\rho}_{X,Y}$).

Prova:

Representando-se por X e Y as verdadeiras variáveis e por X' e Y' as variáveis observadas, com erros de medidas, tem-se:

$$X' = X + u$$

$$Y' = Y + v$$

onde u e v são os erros de medidas, normalmente distribuídas com média zero e variância σ_u^2 e σ_v^2. Assumindo que os erros são independentes entre si e de X e Y, tem-se: $Cov[u,v] = Cov[X,u] = Cov[Y,v] = Cov[X,v] = Cov[Y,u] = 0$

Supondo, por conveniência, que X e Y têm média zero, então $E[X'] = E[Y'] = 0$ e

$$V[X'] = V[X] + V[u]$$
$$V[Y'] = V[Y] + V[v]$$
$$Cov[X',Y'] = E[X'Y'] - E[X']\,E[Y'] = E[XY] + E[uv] + E[Xv] + E[Yu]$$
$$Cov[X',Y'] = E[XY] - E[X]\,E[Y] = Cov[X,Y]$$

A correlação entre as variáveis observadas X' e Y' será:

$$\hat{\rho}_{X',Y'} = \frac{Cov[X',Y']}{\sqrt{V[X']V[Y']}} = \frac{Cov[X,Y]}{\sqrt{(V[X]+V[u])(V[Y]+V[v])}}$$

$$\hat{\rho}_{X',Y'} = \frac{\hat{\rho}_{X,Y}}{\sqrt{\left(1+\frac{V[u]}{V[X]}\right)\left(1+\frac{V[v]}{V[Y]}\right)}} \quad \text{ou} \tag{3.25}$$

$$\hat{\rho}_{Y',X'} = \frac{\hat{\rho}_{Y,X}}{\sqrt{\left(1+\frac{V[u]}{V[X]}\right)\left(1+\frac{V[v]}{V[Y]}\right)}} \tag{3.26}$$

É evidente que o coeficiente de regressão $\hat{\beta}_{Y',X'}$ é também afetado, pois existe relação entre $\hat{\beta}$ e $\hat{\rho}$, como apresentado a seguir:

$$\hat{\beta}_{Y,X} = \hat{\rho}_{Y,X}\frac{S_Y}{S_X} = \hat{\rho}_{Y,X}\sqrt{\frac{V[Y]}{V[X]}} \tag{3.27}$$

Substituindo (3.26) em (3.27) tem-se:

$$\hat{\beta}_{Y',X'} = \hat{\rho}_{Y',X'}\sqrt{\frac{V[Y']}{V[X']}} = \hat{\rho}_{Y,X}\frac{\sqrt{\frac{V[Y']}{V[X']}}}{\sqrt{\left(1+\frac{V[u]}{V[X]}\right)\left(1+\frac{V[v]}{V[Y]}\right)}}$$

$$\hat{\beta}_{Y',X'} = \frac{\hat{\beta}_{YX}}{\left(1+\frac{V[u]}{V[X]}\right)} \tag{3.28}$$

Se X é medido com erro, então o coeficiente de regressão das variáveis observadas é um estimador viesado do coeficiente de regressão das verdadeiras variáveis X e Y. No entanto, se X é medido sem erro, então $\hat{\beta}_{Y',X'} = \hat{\beta}_{Y,X}$.

A presença de erro em Y não causa viés no coeficiente de regressão, como se pode observar na expressão (3.28).

3.2.1.5 Distribuição Amostral do Coeficiente de Correlação Linear de Pearson

Como qualquer outra estatística, é esperado que $\hat{\rho}$ difira do seu parâmetro ρ. A distribuição de $\hat{\rho}$ não é simétrica; esta depende do tamanho de ρ e do tamanho da amostra (GUILFORD, 1950).

Fisher[7], em 1915, citado por ANDERSON (1958, p. 69), foi quem desenvolveu a distribuição de $\hat{\rho}$. Hotelling[8], em 1953, citado por ANDERSON (1958, p. 69), fez um estudo exaustivo e recomendou a forma apresentada a seguir.

---

[7]FISHER, R. A. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. **Biometrika**, v. 10, p. 507-521, 1915.

Resultado 3.6: O coeficiente de correlação $\hat{\rho}$, estimado a partir da amostra de tamanho n, proveniente de distribuição normal bivariada com $\rho \neq 0$, é distribuído com função densidade de probabilidade dada por:

$$f(\hat{\rho}) = \frac{(n-2)\Gamma(n-1)(1-\rho^2)^{\frac{n-1}{2}}(1-\hat{\rho}^2)^{\frac{n-4}{2}}}{\sqrt{2\pi}\Gamma\left(n-\frac{1}{2}\right)(1-\rho\hat{\rho})^{n-\frac{3}{2}}} \times \left[1+\frac{1}{4}\frac{(\rho\hat{\rho}+1)}{2n-1}+\frac{9}{16}\frac{(\rho\hat{\rho}+1)^2}{2(2n-1)(2n+1)}+...\right] \quad (3.29)$$

Prova:

A função densidade de probabilidade da distribuição normal bivariada, conforme apresentada na seção 2.4.2.5, é:

$$f_{X,Y}(X,Y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2-2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)+\left(\frac{Y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}$$

onde: $\mu_X \in R$, $\mu_Y \in R$, $\sigma_X^2 \in R^+$, $\sigma_Y^2 \in R^+$ e $-1 \leq \rho \leq +1$

Fazendo: $t = \frac{x-\mu_X}{\sigma_X}$

e $u = \frac{Y-\mu_Y}{\sigma_Y}$

então, tem-se que:

$$f(t,u) = \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[t^2-2\rho tu+u^2\right]\right\}$$

$$f(t,u) = \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[(u-\rho t)^2+\left(1-\rho^2\right)\times t^2\right]\right\}$$

$$f(t,u) = \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left\{-\frac{(u-\rho t)^2}{2(1-\rho^2)}-\frac{\left(1-\rho^2\right)\times t^2}{2(1-\rho^2)}\right\}$$

---

[8]HOTELLING, Harold. New light on the correlation coefficient and its transforms. **J. Roy. Stat. Soc.** , B, v. 15, p. 193-225, 1953.

Fazendo $v = \frac{u - \rho t}{\sqrt{(1-\rho^2)}}$, para $v = u$ tem-se $\rho = 0$

e então $f(t,u) = \frac{1}{2\pi}\exp\left\{-\frac{v^2}{2} - \frac{t^2}{2}\right\} = \frac{1}{2\pi} e^{-\frac{v^2}{2}} e^{-\frac{t^2}{2}}$

Assim, t e v são variáveis normais padrão e portanto $\sum_{i=1}^{n} v_i^2 \sim \chi_n^2$.

Fazendo uma transformação ortogonal de v_i para um novo conjunto de variáveis $\xi_1, \xi_2, ..., \xi_N$, onde se escolhe

$$\xi_1 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} v_i = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{u_i - \rho t_i}{\sqrt{(1-\rho^2)}} = \frac{\sqrt{n}}{\sqrt{(1-\rho^2)}}\sum_{i=1}^{n}\frac{u_i - \rho t_i}{n} = \frac{\sqrt{n}}{\sqrt{(1-\rho^2)}}\left(\bar{u} - \rho\bar{t}\right)$$

Então, tem-se

$$\sum_{i=1}^{n}\xi_i^2 = \sum_{i=1}^{n} v_i^2 = \sum_{i=1}^{n}\left[\frac{u_i - \rho t_i}{\sqrt{(1-\rho^2)}}\right]^2 = \frac{1}{(1-\rho^2)}\left[\sum_{i=1}^{n} u_i^2 - 2\rho\sum_{i=1}^{n} u_i t_i + \rho^2\sum_{i=1}^{n} t_i^2\right]$$

$$\sum_{i=1}^{n}\xi_i^2 = \frac{1}{(1-\rho^2)}\left[\sum_{i=1}^{n}\left(u_i - \bar{u}\right)^2 - 2\rho\sum_{i=1}^{n}\left(u_i - \bar{u}\right)\left(t_i - \bar{t}\right) + \rho^2\sum_{i=1}^{n}\left(t_i - \bar{t}\right)^2 + n\bar{u}^2 - 2\rho n\bar{u}\bar{t} + \rho^2 n\bar{t}^2\right]$$

$$\sum_{i=1}^{n}\xi_i^2 = \frac{1}{(1-\rho^2)}\left[S_2^2 - 2\rho\hat{\rho}S_2S_1 + \rho^2 S_1^2\right] + \xi_1^2$$

onde $S_1^2 = \sum_{i=1}^{n}\left(t_i - \bar{t}\right)^2$ e $S_2^2 = \sum_{i=1}^{n}\left(u_i - \bar{u}\right)^2$

portanto,

$$\sum_{i=2}^{n}\xi_i^2 = \frac{1}{(1-\rho^2)}\left[S_2^2 - 2\rho\hat{\rho}S_2S_1 + \rho^2 S_1^2\right] \text{ com distribuição } \chi_{n-1}^2 \qquad (3.30)$$

Escolhe-se agora $\xi_2 = \frac{1}{S_1}\sum_{i=1}^{n}\left(t_i - \bar{t}\right)v_i$, que é ortogonal a ξ_1.

Substituindo o valor de v_i em ξ_2 tem-se:

$$\xi_2 = \frac{1}{S_1}\sum_{i=1}^{n}\left(t_i - \bar{t}\right)\left(\frac{u_i - \rho t_i}{\sqrt{(1-\rho^2)}}\right) = \frac{1}{S_1\sqrt{(1-\rho^2)}}\sum_{i=1}^{n}\left(t_i - \bar{t}\right)\left(u_i - \rho t_i\right)$$

$$\xi_2 = \frac{1}{S_1\sqrt{(1-\rho^2)}}\sum_{i=1}^{n}\left(t_i - \bar{t}\right)\left[\left(u_i - \bar{u}\right) - \rho\left(t_i - \bar{t}\right)\right] = \frac{1}{\sqrt{(1-\rho^2)}}\left[\sum_{i=1}^{n}\frac{\left(t_i - \bar{t}\right)\left(u_i - \bar{u}\right)}{S_1} - \rho\sum_{i=1}^{n}\frac{\left(t_i - \bar{t}\right)^2}{S_1}\right]$$

$$\xi_2 = \frac{1}{\sqrt{(1-\rho^2)}}\left[\frac{\hat{\rho}S_1S_2}{S_1} - \rho\frac{S_1^2}{S_1}\right] = \frac{1}{\sqrt{(1-\rho^2)}}\left[\hat{\rho}S_2 - \rho S_1\right]$$

Tem-se, então, que:

$$\xi_2^2 = \frac{1}{1-\rho^2}\left[\hat{\rho}^2S_2^2 - 2\rho\hat{\rho}S_1S_2 + \rho^2S_1^2\right] \qquad (3.31)$$

De (3.30) e (3.31) tem-se que: $\sum_{i=3}^{n}\xi_i^2 = \sum_{i=2}^{n}\xi_i^2 - \xi_2^2$

$$\sum_{i=3}^{n}\xi_i^2 = \frac{1}{(1-\rho^2)}\left[S_2^2 - 2\rho\hat{\rho}S_2S_1 + \rho^2S_1^2\right] - \left[\frac{\left[\hat{\rho}S_2 - \rho S_1\right]}{\sqrt{(1-\rho^2)}}\right]^2 = \frac{S_2^2\left(1-\hat{\rho}^2\right)}{(1-\rho^2)} \sim \chi^2_{n-2}$$

Além disso, $S_1^2 = \sum_{i=1}^{n}\left(t_i - \bar{t}\right)^2 \sim \chi^2_{n-1}$

Tem-se três variáveis independentes:

$$a = \xi_2 = \frac{1}{\sqrt{(1-\rho^2)}}\left[\hat{\rho}S_2 - \rho S_1\right] \sim N(0,1)$$

$$b = \frac{1}{2}\sum_{i=3}^{n}\xi_i^2 = \frac{S_2^2\left(1-\hat{\rho}^2\right)}{2(1-\rho^2)} \sim \chi^2_{n-2} \qquad (3.32)$$

$$c = \frac{1}{2}\sum_{i=1}^{n}\left(t_i - \bar{t}\right)^2 = \frac{S_1^2}{2} \sim \chi^2_{n-1}$$

$f(a) = \frac{1}{\sqrt{2\pi}}e^{-\frac{a^2}{2}}$ f.d.p da distribuição normal padrão apresentada na seção 2.4.2.1.

$f(b) = \frac{1}{\Gamma\left(\frac{n-2}{2}\right)}\left(\frac{1}{2}\right)^{\frac{n-2}{2}} b^{\frac{n-4}{2}} e^{-\frac{b}{2}}$ f.d.p da distribuição Qui-quadrado (χ^2) apresentada

na seção 2.4.2.2.

$f(c) = \frac{1}{\Gamma\left(\frac{n-1}{2}\right)}\left(\frac{1}{2}\right)^{\frac{n-1}{2}} c^{\frac{n-3}{2}} e^{-\frac{c}{2}}$ f.d.p da distribuição Qui-quadrado (χ^2) apresentada

na seção 2.4.2.2.

a, b e c são independentes, portanto:

$$f(a,b,c) = f(a)f(b)f(c) = \frac{1}{\sqrt{2\pi}} e^{-\frac{a^2}{2}} \frac{1}{\Gamma\left(\frac{n-2}{2}\right)}\left(\frac{1}{2}\right)^{\frac{n-2}{2}} b^{\frac{n-4}{2}} e^{-\frac{b}{2}} \frac{1}{\Gamma\left(\frac{n-1}{2}\right)}\left(\frac{1}{2}\right)^{\frac{n-1}{2}} c^{\frac{n-3}{2}} e^{-\frac{c}{2}}$$

$$f(a,b,c) = \frac{1}{\sqrt{2\pi}\Gamma\left(\frac{n-2}{2}\right)\Gamma\left(\frac{n-1}{2}\right)} b^{\frac{n-4}{2}} c^{\frac{n-3}{2}} e^{-\left[\frac{a^2}{2}+b+c\right]}$$

Mas tem-se que: $\frac{a^2}{2} + b + c = \frac{1}{2(1-\rho^2)}\left[S_1^2 + S_2^2 - 2\rho\hat{\rho}S_1S_2\right]$

$$b^{\frac{n-4}{2}} = \left[\frac{S_2^2(1-\hat{\rho}^2)}{2(1-\rho^2)}\right]^{\frac{n-4}{2}} = \frac{S_2^{n-4}(1-\hat{\rho}^2)^{\frac{n-4}{2}}}{2^{\frac{n-4}{2}}(1-\rho^2)^{\frac{n-4}{2}}}$$

$$c^{\frac{n-3}{2}} = \left[\frac{S_1^2}{2}\right]^{\frac{n-3}{2}} = \frac{S_1^{n-3}}{2^{\frac{n-3}{2}}}, \quad \text{portanto}$$

$$f(a,b,c) = \frac{1}{\sqrt{2\pi}\Gamma\left(\frac{n-2}{2}\right)\Gamma\left(\frac{n-1}{2}\right)} \frac{S_2^{n-4}(1-\hat{\rho}^2)^{\frac{n-4}{2}}}{2^{\frac{n-4}{2}}(1-\rho^2)^{\frac{n-4}{2}}} \frac{S_1^{n-3}}{2^{\frac{n-3}{2}}} e^{-\frac{1}{2(1-\rho^2)}\left[S_1^2+S_2^2-2\rho\hat{\rho}S_1S_2\right]}$$

$$f(a,b,c) = \frac{(1-\hat{\rho}^2)^{\frac{n-4}{2}} S_2^{n-4} S_1^{n-3} e^{-\frac{1}{2(1-\rho^2)}\left[S_1^2+S_2^2-2\rho\hat{\rho}S_1S_2\right]}}{\sqrt{2\pi}(1-\rho^2)^{\frac{n-4}{2}} 2^{n-\frac{7}{2}}\Gamma\left(\frac{n-2}{2}\right)\Gamma\left(\frac{n-1}{2}\right)}$$

Trocando-se as variáveis a, b, c por $\hat{\rho}, S_1, S_2$. De (3.32), o Jacobiano de transformação é:

$$J = \left(\frac{a,b,c}{\hat{\rho}, S_1, S_2} \right) = (1-\rho^2)^{-3/2} \begin{vmatrix} S_2 & -\rho & \hat{\rho} \\ -\hat{\rho}S_2^2 & 0 & (1-\hat{\rho}^2)S_2 \\ 0 & S_1 & 0 \end{vmatrix} = -(1-\rho^2)^{-3/2} S_1 S_2^2$$

então,

$$f(\hat{\rho}, S_1, S_2) = \frac{(1-\hat{\rho}^2)^{\frac{n-4}{2}} S_1 S_2^2 S_2^{n-4} S_1^{n-3} e^{-\frac{1}{2(1-\rho^2)}\left[S_1^2 + S_2^2 - 2\rho\hat{\rho}S_1S_2\right]}}{\sqrt{2\pi}(1-\rho^2)^{\frac{n-4}{2}} (1-\rho^2)^{\frac{3}{2}} 2^{n-\frac{7}{2}} \Gamma\left(\frac{n-2}{2}\right)\Gamma\left(\frac{n-1}{2}\right)}$$

A função densidade de $\hat{\rho}$ é obtida integrando em relação a S_1 e S_2 no intervalo de zero a ∞.

$$f(\hat{\rho}) = \int_0^\infty \int_0^\infty \frac{(1-\hat{\rho}^2)^{\frac{n-4}{2}} S_2^{n-2} S_1^{n-2} e^{-\frac{1}{2(1-\rho^2)}\left[S_1^2 + S_2^2 - 2\rho\hat{\rho}S_1S_2\right]}}{\sqrt{2\pi}(1-\rho^2)^{\frac{n-1}{2}} 2^{n-\frac{7}{2}} \Gamma\left(\frac{n-2}{2}\right)\Gamma\left(\frac{n-1}{2}\right)} dS_1 dS_2$$

Devido à dificuldade para calcular a integral, FISHER[9], citado por KENNY e KEEPING (1951, p. 219), utilizou a seguinte transformação:

$$S_1 = \alpha^{1/2} e^{\beta/2}$$

$$S_2 = \alpha^{1/2} e^{-\beta/2}$$

$$J = \begin{vmatrix} e^{\beta/2}\frac{1}{2}\alpha^{-1/2} & \alpha^{1/2}e^{\beta/2}\frac{1}{2} \\ e^{-\beta/2}\frac{1}{2}\alpha^{-1/2} & \alpha^{1/2}e^{-\beta/2}\left(-\frac{1}{2}\right) \end{vmatrix}$$

[9]FISHER, R. A. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. **Biometrika**, v. 10, p. 507-521, 1915.

O Jacobiano de transformação é igual a -1/2, portanto:

$$f(S_1,S_2/\alpha,\beta)=\frac{(1-\hat{\rho}^2)^{\frac{n-4}{2}}\left(\alpha^{1/2}e^{\beta/2}\right)^{n-2}\left(\alpha^{1/2}e^{-\beta/2}\right)^{n-2}e^{-\frac{1}{2(1-\rho^2)}\left[\left(\alpha^{1/2}e^{\beta/2}\right)^2+\left(\alpha^{1/2}e^{-\beta/2}\right)^2-2\rho\hat{\rho}\left(\alpha^{1/2}e^{\beta/2}\right)\left(\alpha^{1/2}e^{-\beta/2}\right)\right]}}{2\sqrt{2\pi}(1-\rho^2)^{\frac{n-1}{2}}2^{n-\frac{7}{2}}\Gamma\left(\frac{n-2}{2}\right)\Gamma\left(\frac{n-1}{2}\right)}$$

$$f(S_1,S_2/\alpha,\beta)=\frac{(1-\hat{\rho}^2)^{\frac{n-4}{2}}\alpha^{n-2}e^{-\frac{1}{2(1-\rho^2)}\left[\alpha(e^{\beta}+e^{-\beta}-2\rho\hat{\rho}\right]}}{2\sqrt{2\pi}(1-\rho^2)^{\frac{n-1}{2}}2^{n-\frac{7}{2}}\Gamma\left(\frac{n-2}{2}\right)\Gamma\left(\frac{n-1}{2}\right)}$$

Tem-se que $\frac{1}{2}\left(e^{z}+e^{-z}\right)=\cosh(z)$, e pela Fórmula de Duplicação de Legendre $2^{n-3}\Gamma\left(\frac{n-2}{2}\right)\Gamma\left(\frac{n-1}{2}\right)=\sqrt{\pi}\,\Gamma(n-2)$, então

$$f(S_1,S_2/\alpha,\beta)=\frac{(1-\hat{\rho}^2)^{\frac{n-4}{2}}\alpha^{n-2}e^{\frac{-\alpha(\cosh\beta-\rho\hat{\rho})}{(1-\rho^2)}}}{2\pi(1-\rho^2)^{\frac{n-1}{2}}\Gamma(n-2)}$$

e $$f(\hat{\rho})=\int_{-\infty}^{\infty}\int_{0}^{\infty}\frac{(1-\hat{\rho}^2)^{\frac{n-4}{2}}\alpha^{n-2}e^{\frac{-\alpha(\cosh\beta-\rho\hat{\rho})}{(1-\rho^2)}}}{2\pi(1-\rho^2)^{\frac{n-1}{2}}\Gamma(n-2)}d\alpha d\beta$$

$$f(\hat{\rho})=\int_{-\infty}^{\infty}\frac{(1-\hat{\rho}^2)^{\frac{n-4}{2}}}{2\pi(1-\rho^2)^{\frac{n-1}{2}}\Gamma(n-2)}\int_{0}^{\infty}\alpha^{n-2}e^{\frac{-\alpha(\cosh\beta-\rho\hat{\rho})}{(1-\rho^2)}}d\alpha d\beta$$

$$f(\hat{\rho})=\frac{(1-\hat{\rho}^2)^{\frac{n-4}{2}}}{2\pi(1-\rho^2)^{\frac{n-1}{2}}\Gamma(n-2)}2\int_{0}^{\infty}\frac{\Gamma(n-1)(1-\rho^2)^{n-1}}{(\cosh\beta-\rho\hat{\rho})^{n-1}}d\beta$$

$$f(\hat{\rho})=\frac{(1-\hat{\rho}^2)^{\frac{n-4}{2}}\Gamma(n-1)(1-\rho^2)^{n-1}}{2\pi(1-\rho^2)^{\frac{n-1}{2}}\Gamma(n-2)}2\int_{0}^{\infty}\frac{1}{(\cosh\beta-\rho\hat{\rho})^{n-1}}d\beta$$

$$f(\hat{\rho})=\frac{(n-2)(1-\hat{\rho}^2)^{\frac{n-4}{2}}(1-\rho^2)^{\frac{n-1}{2}}}{\pi}\int_{0}^{\infty}\frac{d\beta}{(\cosh\beta-\rho\hat{\rho})^{n-1}} \qquad (3.33)$$

A integral pode ser expressa como uma função hipergeométrica, apresentada em KENNEY e KEEPING (1951, p. 219):

$$\int_0^{\infty} \frac{d\beta}{(\cosh\beta - \rho\hat{\rho})^{n-1}} = \left(\frac{\pi}{2}\right)^{1/2} \frac{\Gamma(n-1)}{\Gamma(n-\frac{1}{2})}(1-\rho\hat{\rho})^{-(n-3/2)} F\left(\frac{1}{2};\frac{1}{2};\frac{2n-1}{2};\frac{\rho\hat{\rho}+1}{2}\right) \qquad (3.34)$$

A função hipergeométrica tem a seguinte solução:

$$F(a;b;c;z) = 1 + \frac{ab}{1!c}z + \frac{a(a+1)b(b+1)}{2!c(c+1)}z^2 + ... \qquad (3.35)$$

e $f(\hat{\rho})$ pode ser escrita na forma de série convergente, como segue:

$$f(\hat{\rho}) = \frac{(n-2)\Gamma(n-1)(1-\rho^2)^{\frac{n-1}{2}}(1-\hat{\rho}^2)^{\frac{n-4}{2}}}{\sqrt{2\pi}\,\Gamma\left(n-\frac{1}{2}\right)(1-\rho\hat{\rho})^{n-\frac{3}{2}}} \times \left[1 + \frac{1}{4}\frac{(\rho\hat{\rho}+1)}{2n-1} + \frac{9}{16}\frac{(\rho\hat{\rho}+1)^2}{2(2n-1)(2n+1)} + ...\right]$$

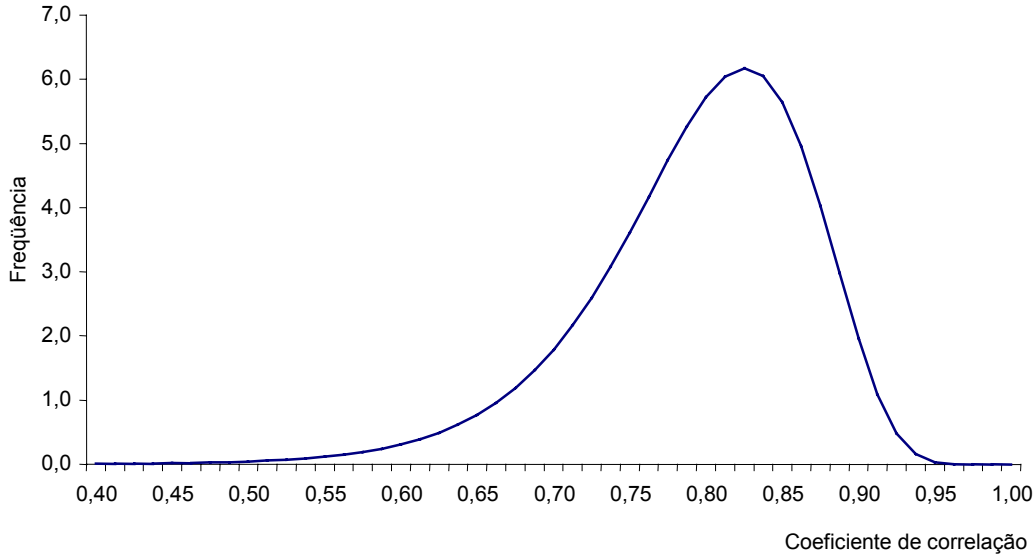
Para valores altos e positivos de ρ, a distribuição é assimétrica negativa, e no caso de serem altos e negativos, a distribuição é assimétrica positiva, como mostram os gráficos a seguir.

Para um mesmo valor de ρ, quanto menor o tamanho da amostra maior é a assimetria da distribuição. À medida que aumenta o tamanho da amostra, tende para uma distribuição simétrica.

Os gráficos 5 e 6 apresentam a distribuição amostral de $\hat{\rho}$ para amostra de tamanho $n = 29$ e $\rho = 0{,}80$ e $\rho = -0{,}80$, respectivamente. A escolha do tamanho da amostra foi arbitrária. Os cálculos para a obtenção dos valores de $f(\hat{\rho})$ encontram-se no Apêndice 1.

(I) $n = 29$ e $\rho = 0{,}80$

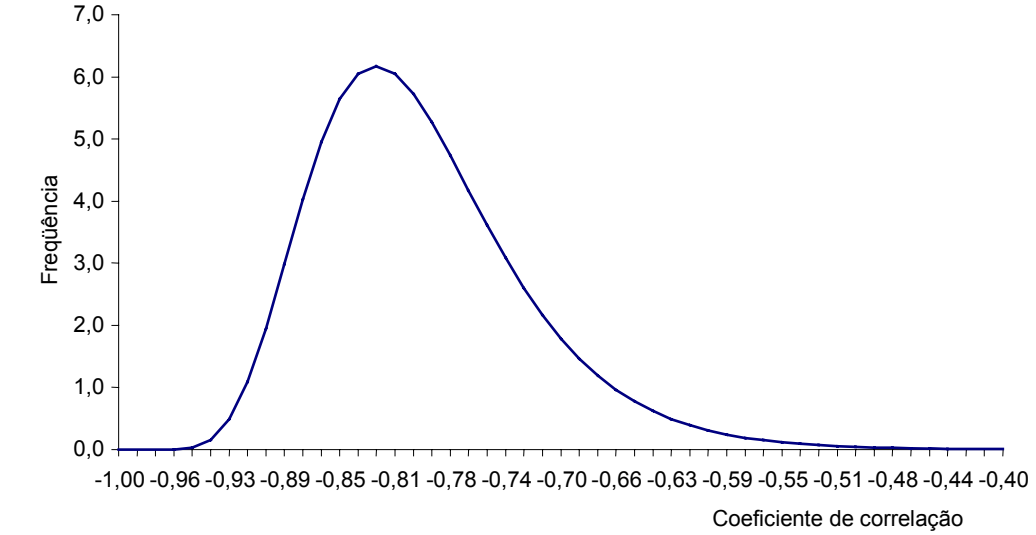
GRÁFICO 5 - DISTRIBUIÇÃO AMOSTRAL DO COEFICIENTE DE CORRELAÇÃO DE PEARSON PARA $\rho = 0{,}80$



FONTE: A autora

(II) $n = 29$ e $\rho = -0{,}80$

GRÁFICO 6 - DISTRIBUIÇÃO AMOSTRAL DO COEFICIENTE DE CORRELAÇÃO DE PEARSON PARA $\rho = -0{,}80$



FONTE: A autora

Resultado 3.7: O coeficiente de correlação $\hat{\rho}$, estimado a partir da amostra de tamanho n, proveniente de distribuição normal bivariada com $\rho = 0$, é distribuído com função densidade de probabilidade dada por:

$$f(\hat{\rho}) = \frac{\Gamma\left[\frac{1}{2}(n-1)\right](1-\hat{\rho}^2)^{\frac{n-4}{2}}}{\Gamma\left[\frac{1}{2}(n-2)\right]\sqrt{\pi}} \tag{3.36}$$

Prova:

Tem-se, da expressão (3.33), que:

$$f(\hat{\rho}) = \frac{(n-2)(1-\hat{\rho}^2)^{\frac{n-4}{2}}(1-\rho^2)^{\frac{n-1}{2}}}{\pi}\int_0^{\infty}\frac{d\beta}{(\cosh\beta - \rho\hat{\rho})^{n-1}}$$

Mas se $\rho = 0$, tem-se que: $f(\hat{\rho}) = \frac{(n-2)(1-\hat{\rho}^2)^{\frac{n-4}{2}}}{\pi}\int_0^{\infty}\frac{d\beta}{(\cosh\beta)^{n-1}}$

e $\int_0^{\infty}\frac{d\beta}{(\cosh\beta)^{n-1}} = \left(\frac{\pi}{2}\right)^{1/2}\frac{\Gamma(n-1)}{\Gamma(n-\frac{1}{2})}F\left(\frac{1}{2};\frac{1}{2};\frac{2n-1}{2};\frac{1}{2}\right) = \left(\frac{\pi}{2}\right)^{1/2}\frac{\Gamma(n-1)}{\Gamma(n-\frac{1}{2})}\frac{\Gamma(n-\frac{1}{2})2^{(3/2)-n}\sqrt{\pi}}{\left[\Gamma\left(\frac{n}{2}\right)\right]^2}$

Pois, $F\left(\frac{1}{2};\frac{1}{2};\frac{2n-1}{2};\frac{1}{2}\right) = \frac{\Gamma(n-\frac{1}{2})2^{(3/2)-n}\sqrt{\pi}}{\left[\Gamma\left(\frac{n}{2}\right)\right]^2}$

então, $\int_0^{\infty}\frac{d\beta}{(\cosh\beta)^{n-1}} = \left(\frac{\pi}{2}\right)^{1/2}\frac{\Gamma(n-1)}{\Gamma(n-\frac{1}{2})}\frac{\Gamma(n-\frac{1}{2})2^{(3/2)-n}\sqrt{\pi}}{\left[\Gamma\left(\frac{n}{2}\right)\right]^2}$

logo, $f(\hat{\rho}) = \frac{(n-2)(1-\hat{\rho}^2)^{\frac{n-4}{2}}}{\pi}\frac{\sqrt{\pi}}{\sqrt{2}}\frac{\Gamma(n-1)2^{(3/2)-n}\sqrt{\pi}}{\left[\Gamma\left(\frac{n}{2}\right)\right]^2}$

$$f(\hat{\rho}) = \frac{(n-2)(1-\hat{\rho}^2)^{(n-4)/2}2^{(3/2)-n}}{\sqrt{2}\sqrt{\pi}}\frac{\sqrt{\pi}\Gamma(n-1)}{\left[\Gamma\left(\frac{n}{2}\right)\right]^2}$$

Pela fórmula de duplicação de Legendre tem-se que:

$$\sqrt{\pi}\Gamma(n-1) = 2^{n-2}\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{n-1}{2}\right)$$

Assim,

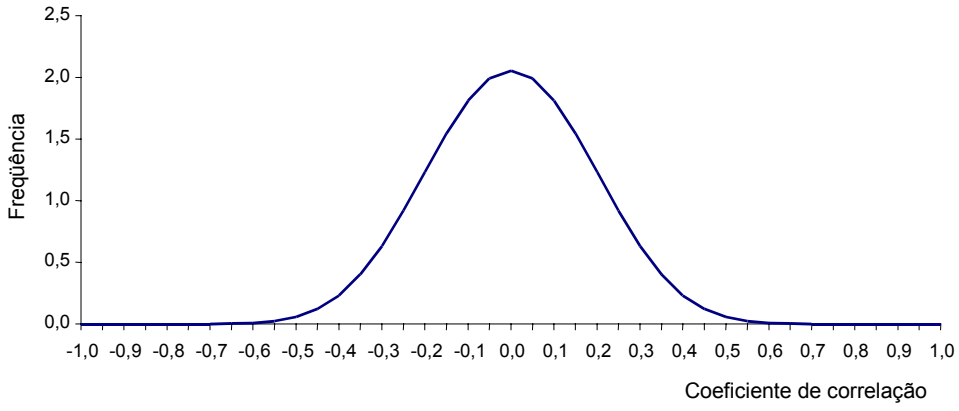
$$f(\hat{\rho}) = \frac{(n-2)(1-\hat{\rho}^2)^{(n-4)/2}2^{3/2-n}2^{-1/2}2^{n-2}}{\sqrt{\pi}} \frac{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{n-1}{2}\right)}{\left[\Gamma\left(\frac{n}{2}\right)\right]^2}$$

$$f(\hat{\rho}) = \frac{(n-2)(1-\hat{\rho}^2)^{(n-4)/2}2^{-1}}{\sqrt{\pi}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} = \frac{(n-2)(1-\hat{\rho}^2)^{(n-4)/2}}{2\sqrt{\pi}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\frac{(n-2)}{2}\Gamma\left(\frac{n-2}{2}\right)}$$

$$f(\hat{\rho}) = \frac{\Gamma\left[\frac{1}{2}(n-1)\right](1-\hat{\rho}^2)^{(n-4)/2}}{\Gamma\left[\frac{1}{2}(n-2)\right]\sqrt{\pi}}$$

Segundo BRYANT (1960), quando $\rho = 0$ a distribuição é simétrica, embora não exatamente Gaussiana. O gráfico 7 mostra a distribuição amostral de $\hat{\rho}$ para amostra de tamanho $n = 29$. Manteve-se o mesmo tamanho de amostra dos gráficos 5 e 6, para possibilitar a comparação entre os mesmos. Os cálculos de $f(\hat{\rho})$ encontram-se no Apêndice 1.

GRÁFICO 7 - DISTRIBUIÇÃO AMOSTRAL DO COEFICIENTE DE CORRELAÇÃO DE PEARSON PARA $\rho = 0$



FONTE: A autora

Resultado 3.8: Se $\rho = 0$, a distribuição amostral de $\hat{\rho}$ será simétrica com

$$E(\hat{\rho}) = 0, \quad \hat{\sigma}^2_{\hat{\rho}} = \frac{1}{n-1} \quad \text{e} \quad \hat{\sigma}_{\hat{\rho}} = \frac{1}{\sqrt{n-1}} \tag{3.37}$$

Prova:

A curva de freqüências de Pearson tipo II, citado por ELDERTON (1953, p. 51a), apresentada a seguir, é simétrica tendo a média como origem, que coincide com a moda e portanto $E(y) = 0$.

$$f(y) = y_0 \left[1 - \left(\frac{y}{a} \right)^2 \right]^m$$

A função densidade de $\hat{\rho}$ é uma curva de freqüências de Pearson do tipo II, como se pode observar na comparação de ambas. Como já apresentada no resultado 3.7, a f.d.p. de $\hat{\rho}$ quando $\rho = 0$ é:

$$f(\hat{\rho}) = \frac{\Gamma\left[\frac{1}{2}(n-1)\right](1-\hat{\rho}^2)^{(n-4)/2}}{\Gamma\left[\frac{1}{2}(n-2)\right]\sqrt{\pi}}$$

Fazendo $y_0 = \dfrac{\Gamma\left[\frac{1}{2}(n-1)\right]}{\Gamma\left[\frac{1}{2}(n-2)\right]\sqrt{\pi}}$, $\hat{\rho}^2 = \left(\dfrac{y}{a} \right)^2$ e $m = (n-4)/2$, as duas funções são equivalentes. Portanto, a f.d.p. de $\hat{\rho}$ é simétrica com $E(\hat{\rho}) = 0$.

O estimador da variância é obtido através de $V(\hat{\rho}) = E(\hat{\rho}^2) - \left[E(\hat{\rho})\right]^2$ e $E(\hat{\rho}^2) = \int_{-1}^{1} \hat{\rho}^2 f(\hat{\rho}) d\hat{\rho}$, apresentados na seção 2.3.

Fazendo $\hat{\rho}^2 = x$, então $\hat{\rho} = x^{1/2}$ e $d\hat{\rho} = \frac{1}{2} x^{-1/2} dx$ e tem-se $-1 \le \hat{\rho} \le 1$ então $0 \le x \le 1$.

Portanto: $E(\hat{\rho}^2) = 2E(X) = 2\int_0^1 x \frac{\Gamma\left[\frac{1}{2}(n-1)\right](1-x)^{(n-4)/2}}{\Gamma\left[\frac{1}{2}(n-2)\right]\sqrt{\pi}} \frac{1}{2} x^{-1/2} dx$

$$E(\hat{\rho}^2) = 2E(X) = \frac{\Gamma\left[\frac{1}{2}(n-1)\right]}{\Gamma\left[\frac{1}{2}(n-2)\right]\sqrt{\pi}} \underbrace{\int_0^1 x^{1/2}(1-x)^{(n-4)/2} dx}_{\text{Função Beta}}$$

A função beta é definida por:

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1} dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \quad , \quad a > 0, \ b > 0 \tag{3.38}$$

Tem-se que $a = \frac{3}{2}$ e $b = \frac{n-2}{2}$, portanto:

$$E(\hat{\rho}^2) = 2E(X) = 2\frac{\Gamma\left[\frac{1}{2}(n-1)\right]}{\Gamma\left[\frac{1}{2}(n-2)\right]\sqrt{\pi}} \frac{\Gamma\left(\frac{3}{2}\right)\Gamma\left(\frac{n-2}{2}\right)}{\Gamma\left(\frac{3}{2}+\frac{n-2}{2}\right)}$$

$$E(\hat{\rho}^2) = 2E(X) = \frac{\Gamma\left[\frac{1}{2}(n-1)\right]}{\Gamma\left[\frac{1}{2}(n-2)\right]\sqrt{\pi}} \frac{\sqrt{\pi}}{2}\Gamma\left(\frac{1}{2}(n-2)\right) = \frac{\frac{1}{2}\Gamma\left[\frac{1}{2}(n-1)\right]}{\left(\frac{n-1}{2}\right)\Gamma\left[\frac{1}{2}(n-1)\right]} = \frac{1}{(n-1)}$$

e $\hat{\sigma}^2_{\hat{\rho}} = E(\hat{\rho}^2) - \left[E(\hat{\rho})\right]^2 = \frac{1}{n-1} - 0 = \frac{1}{n-1}$

e $\hat{\sigma}_{\hat{\rho}} = \frac{1}{\sqrt{n-1}}$

3.2.1.6 Teste de hipótese para $\rho = 0$

A forma simétrica da distribuição quando $\rho = 0$ torna possível testar a hipótese $H_0 : \rho = 0$ contra a hipótese $H_1 : \rho \neq 0$, através da distribuição t de Student.

Resultado 3.9: A estatística para testar a hipótese $H_0 : \rho = 0$ contra $H_1 : \rho \neq 0$, tem distribuição t com n - 2 graus de liberdade, ou seja:

$$t = \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}} \sim t_{n-2} \qquad (3.39)$$

Prova:

Testar a hipótese $H_0 : \rho = 0$ equivale a testar a hipótese de que $H_0 : \beta = 0$, devido à relação entre os dois coeficientes, como já apresentado no resultado 3.2.

$$\hat{\rho} = \hat{\beta}\frac{S_X}{S_Y} \quad \text{e portanto} \quad \hat{\beta} = \hat{\rho}\frac{S_Y}{S_X} \quad \text{onde} \quad S_X = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n-1}} \quad \text{e} \quad x_i = X_i - \overline{X}$$

$$S_Y = \sqrt{\frac{\sum_{i=1}^{n} y_i^2}{n-1}} \quad \text{e} \quad y_i = Y_i - \overline{Y}$$

Das equações normais da reta pelo método dos mínimos quadrados obtém-se (expressão 3.14): $\hat{\beta} = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$ onde $x_i = X_i - \overline{X}$ e $y_i = Y_i - \overline{Y}$

A expressão acima pode ser reescrita como:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i \left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i Y_i - \overline{Y} \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} \qquad (3.40)$$

Sabe-se que $\sum_{i=1}^{n} x_i = 0$ e fazendo $w_i = \dfrac{x_i}{\sum_{i=1}^{n} x_i^2}$, tem-se:

$$\sum_{i=1}^{n} w_i = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} = 0$$

$$\sum_{i=1}^{n} w_i^2 = \frac{\sum_{i=1}^{n} x_i^2}{\left(\sum_{i=1}^{n} x_i^2\right)^2} = \frac{1}{\sum_{i=1}^{n} x_i^2}$$

$$\sum_{i=1}^{n} w_i X_i = \sum_{i=1}^{n} w_i \left(x_i + \overline{X}\right) = \sum_{i=1}^{n} w_i x_i + \overline{X} \sum_{i=1}^{n} w_i = \sum_{i=1}^{n} w_i x_i = \frac{\sum_{i=1}^{n} x_i x_i}{\sum_{i=1}^{n} x_i^2} = 1$$

A expressão (3.40) poderá ser reescrita da seguinte forma:

$\hat{\beta} = \alpha \sum_{i=1}^{n} w_i \, Y_i = \sum_{i=1}^{n} w_i \left(\alpha + \beta X_i + \varepsilon_i\right)$, pois tem-se do modelo de regressão linear simples que $Y = \alpha + \beta X + \varepsilon$

e, portanto, $\hat{\beta} = \alpha \sum_{i=1}^{n} w_i + \beta \sum_{i=1}^{n} w_i X_i + \sum_{i=1}^{n} w_i \varepsilon_i = \beta + \sum_{i=1}^{n} w_i \varepsilon_i$

A esperança e a variância de $\hat{\beta}$ serão:

$$E(\hat{\beta}) = E(\beta + \sum_{i=1}^{n} w_i \varepsilon_i) = E(\beta) + \sum_{i=1}^{n} w_i E(\varepsilon_i) \qquad (3.41)$$

Porém, tem-se no modelo de regressão linear simples as seguintes suposições sobre os erros:

$$E(\varepsilon_i) = 0 \qquad (3.42)$$

$$V(\varepsilon_i) = \sigma^2 \qquad (3.43)$$

Assim, substituindo (3.42) em (3.41) tem-se:

$$E(\hat{\beta}) = E(\beta) = \beta$$

e $V(\hat{\beta}) = V(\beta + \sum_{i=1}^{n} w_i \varepsilon_i) = V(\beta) + \sum_{i=1}^{n} w_i^2 \, V(\varepsilon_i) = \sum_{i=1}^{n} w_i^2 \, V(\varepsilon_i)$ (3.44)

Substituindo (3.43) e o valor de $\sum_{i=1}^{n} w_i^2 = \frac{1}{\sum_{i=1}^{n} x_i^2}$ em (3.44) tem-se:

$V(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$, portanto

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}\right)$$

Mas $S^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left[Y_i - \left(\hat{\alpha} + \hat{\beta}X_i\right)\right]^2 = \frac{\sum_{i=1}^{n} y_i^2 - \hat{\beta}^2 \sum_{i=1}^{n} x_i^2}{n-2}$ é o estimador não-viesado de σ^2 (WONNACOTT e WONNACOTT, 1978, p. 50),

e $\hat{\beta}^2 = \hat{\rho}^2\left[\frac{S_Y}{S_X}\right]^2 = \hat{\rho}^2 \frac{\sum_{i=1}^{n} y_i^2}{\sum_{i=1}^{n} x_i^2}$ então

$$S^2 = \frac{\frac{\hat{\beta}^2 \sum_{i=n}^{n} x_i^2}{\hat{\rho}^2} - \hat{\beta}^2 \sum_{i=1}^{n} x_i^2}{n-2} = \frac{-\hat{\beta}^2 \sum_{i=1}^{n} x_i^2 \left(\frac{1}{\hat{\rho}^2} - 1\right)}{n-2}$$

Tem-se que $U = \frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$ (NETER, et. al.,1996, p. 75) e $t = \frac{Z}{\sqrt{\frac{U}{n-2}}} \sim t_{n-2}$ (JAMES, 1981, p. 85)

Então, $U = \frac{(n-2)S^2}{\sigma^2} = \frac{(n-2)\frac{\hat{\beta}^2 \sum_{i=1}^{n} x_i^2 \left(\frac{1}{\hat{\rho}^2} - 1\right)}{n-2}}{\sigma^2} = \frac{\hat{\beta}^2 \sum_{i=1}^{n} x_i^2 \left(\frac{1}{\hat{\rho}^2} - 1\right)}{\sigma^2} \sim \chi^2_{n-2}$

Fazendo $Z = \frac{\hat{\beta} - \beta}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}}} \sim N(0,1)$, então

$$t = \frac{Z}{\sqrt{\frac{U}{n-2}}} = \frac{(\hat{\beta}-\beta)\sqrt{n-2}}{\sqrt{\frac{\sigma^2\hat{\beta}^2\sum_{i=1}^{n}x_i^2\left(\frac{1}{\hat{\rho}^2}-1\right)}{\sum_{i=1}^{n}x_i^2\sigma^2}}} = \frac{(\hat{\beta}-\beta)\sqrt{n-2}}{\sqrt{\hat{\beta}^2\left(\frac{1}{\hat{\rho}^2}-1\right)}}$$

$$t = \frac{(\hat{\beta}-\beta)\sqrt{n-2}}{\sqrt{\hat{\beta}^2\left(\frac{1-\hat{\rho}^2}{\hat{\rho}^2}\right)}} = \frac{(\hat{\beta}-\beta)\sqrt{n-2}}{\frac{\hat{\beta}}{\hat{\rho}}\sqrt{1-\hat{\rho}^2}}, \text{ mas se } \beta = 0 \text{ então}$$

$$t = \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}} \sim t_{n-2}$$

3.2.1.7 Transformação Z de Fisher

Devido às divergências entre a distribuição amostral de $\hat{\rho}$ e a distribuição normal e as limitações para interpretação, Ronald A. Fisher desenvolveu uma estatística em que qualquer valor de $\hat{\rho}$ pode ser transformado. Esta estatística é chamada de Z, que não é a distribuição normal padronizada (GUILFORD, 1950).

$$Z = \frac{1}{2}\ln\left[\frac{1+\hat{\rho}}{1-\hat{\rho}}\right] = \tanh^{-1}\hat{\rho} \tag{3.45}$$

onde ln é o logaritmo natural.

A média e a variância da distribuição amostral de Z é apresentada a seguir e se encontra em KENNEY e KEEPING (1951, p. 222):

$$E(Z) = \frac{1}{2}\ln\left[\frac{1+\rho}{1-\rho}\right] + \frac{\rho}{2n-1} \tag{3.46}$$

$$V(Z) = \frac{1}{n-1} + \frac{4-\rho^2}{2(n-1)^2} \tag{3.47}$$

Fazendo $k = \frac{4-\rho^2}{2}$, a variância (expressão 3.47) pode ser escrita na forma a seguir:

$$V_1 = \frac{1}{n-1}\left[1+\frac{k}{n-1}\right] \tag{3.48}$$

A expressão (3.48) se aproxima de $V_2 = \frac{1}{n-1-k}$ quando k = 2 e à medida que o tamanho da amostra (n) aumenta, como se pode observar no quadro 1:

QUADRO 1 - VALORES DE V_1 E V_2 SEGUNDO TAMANHO DA AMOSTRA

TAMANHO DA AMOSTRA (n)	$V_1 = \frac{1}{n-1}\left[1+\frac{2}{n-1}\right]$	$V_2 = \frac{1}{n-1-2}$
20	0,05817	0,05882
30	0,03686	0,03704
50	0,02124	0,02128
100	0,01031	0,01031
200	0,00508	0,00508

FONTE: A autora

Para valores de n moderado, verificando-se a igualdade das expressões apresentada no quadro, é possível utilizar os estimadores para variância e erro padrão apresentados a seguir:

$$\hat{\sigma}_Z^2 = \frac{1}{n-3} \qquad \text{e} \qquad \hat{\sigma}_Z = \frac{1}{\sqrt{n-3}} \tag{3.49}$$

Em 1938, DAVID[10], citado por ANDERSON (1958, p. 75), fez algumas comparações entre as probabilidades tabeladas e calculadas, assumindo Z como sendo distribuição Gaussiana. Segundo a autora, para $n > 25$ é possível tratar Z como normalmente distribuída com média $E(Z) = \frac{1}{2}\ln\left[\frac{1+\rho}{1-\rho}\right] + \frac{\rho}{2n-1}$ e variância $\hat{\sigma}_Z^2 = \frac{1}{n-3}$.

A função densidade de probabilidade da distribuição normal ou Gaussiana já foi apresentada na seção 2.4.2.1.

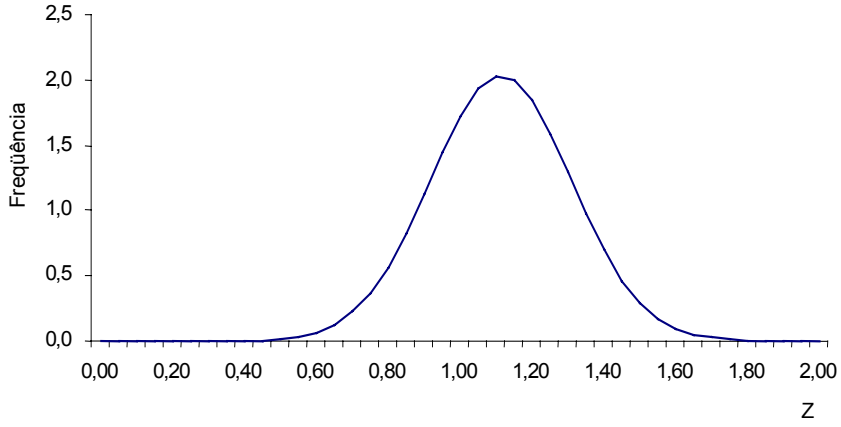
[10]DAVID, F. N. Tables of the ordinates and Probability Integral of the Distribution of the Correlation Coefficient in Small Samples. **Biometrika**, 1938.

Para $n > 25$, como discutido anteriormente, a distribuição de Z terá a seguinte f.d.p.: $f(Z) = \frac{1}{\hat{\sigma}_Z \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{Z-E(Z)}{\hat{\sigma}_Z}\right)^2}$

Os gráficos 8 e 9 mostram a distribuição amostral de Z para as situações apresentadas nos gráficos 5 ($n = 29$ e $\rho = 0{,}80$) e 7 ($n = 29$ e $\rho = 0$), mostrando as distribuições amostrais de $\hat{\rho}$. Os cálculos para a obtenção dos f(Z) encontram-se no Apêndice 2.

(I) Para $n = 29$ e $\rho = 0{,}80$

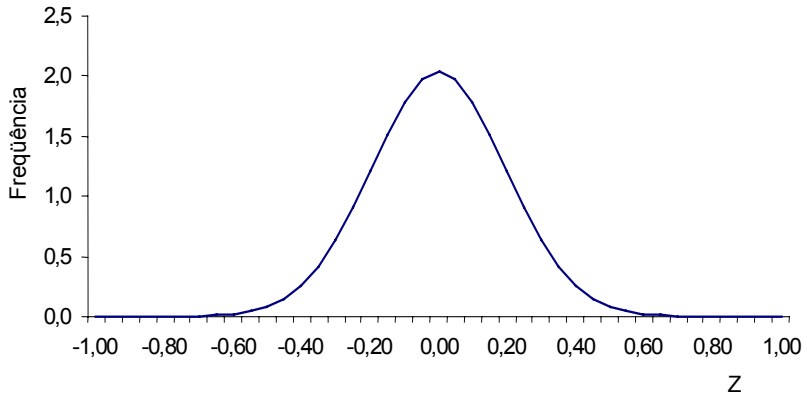
GRÁFICO 8 - DISTRIBUIÇÃO AMOSTRAL DE Z PARA $\rho = 0{,}80$



FONTE: A autora

(II) Para $n = 29$ e $\rho = 0$

GRÁFICO 9 - DISTRIBUIÇÃO AMOSTRAL DE Z PARA $\rho = 0$



FONTE: A autora

3.2.1.8 Teste de hipótese para $\rho \neq 0$

A transformação abordada anteriormente é útil, também, quando se deseja testar a hipótese $H_0 : \rho = \rho_0$ contra $H_1 : \rho \neq \rho_0$.

Neste caso, pode-se usar o teste Z, calculado através de $Z = \frac{Z_{\hat{\rho}} - Z_{\rho}}{\hat{\sigma}_z}$, que é aproximadamente normal (BRYANT, 1960); os valores de Z_{ρ} e $Z_{\hat{\rho}}$ podem ser obtidos através das expressões a seguir:

$$Z_{\rho} = \frac{1}{2} \ln \left[\frac{1+\rho}{1-\rho} \right] \quad \text{e} \quad Z_{\hat{\rho}} = \frac{1}{2} \ln \left[\frac{1+\hat{\rho}}{1-\hat{\rho}} \right] \tag{3.50}$$

onde ρ é o parâmetro populacional que se está testando e $\hat{\rho}$ é a estimativa amostral.

Ainda, a significância da diferença de coeficientes de correlação de duas amostras pode ser testada por:

$H_0 : \rho_1 - \rho_2 = 0$

$H_1 : \rho_1 - \rho_2 \neq 0$

A estatística do teste é definida por:

$$Z = \frac{Z_1 - Z_2}{\sqrt{\hat{\sigma}_{Z_1}^2 + \hat{\sigma}_{Z_2}^2}} \quad \text{onde} \quad \hat{\sigma}_{Z_1} = \frac{1}{\sqrt{n_1 - 3}} \quad \text{e} \quad \hat{\sigma}_{Z_2} = \frac{1}{\sqrt{n_2 - 3}} \tag{3.51}$$

Os valores de Z_1 e Z_2 podem ser obtidos substituindo-se os valores de ρ_1 e ρ_2 na expressões a seguir:

$$Z_1 = \frac{1}{2} \ln \left[\frac{1+\rho_1}{1-\rho_1} \right] \quad \text{e} \quad Z_2 = \frac{1}{2} \ln \left[\frac{1+\rho_2}{1-\rho_2} \right]$$

3.2.1.9 Intervalo de confiança para ρ

Os limites de confiança, de nível $1-\alpha$ para o parâmetro ρ, apresentados em BRYANT (1960), podem ser obtidos através de:

$$P[Z_{\hat{\rho}} - Z_{\alpha/2}\hat{\sigma}_z < Z_\rho < Z_{\hat{\rho}} + Z_{\alpha/2}\hat{\sigma}_z] = 1-\alpha \quad (3.52)$$

onde $Z_{\hat{\rho}}$ é o valor de Z correspondente ao valor do coeficiente de correlação amostral, e $Z_{\alpha/2}$ é o valor da área sob a distribuição normal padrão para um nível de significância de $\alpha/2$.

A partir dos limites de confiança obtidos para Z_ρ, obtém-se os limites para ρ, fazendo:

$$Z_{\hat{\rho}_1} = Z_{\hat{\rho}} - Z_{\alpha/2}\hat{\sigma}_z \qquad \text{e} \qquad Z_{\hat{\rho}_2} = Z_{\hat{\rho}} + Z_{\alpha/2}\hat{\sigma}_z$$

Então, o intervalo de confiança para ρ será obtido a partir da expressão $P[\hat{\rho}_1 < \rho < \hat{\rho}_2] = 1-\alpha$, onde ρ_1 e ρ_2 serão obtidos a partir de:

$$\hat{\rho}_1 = \frac{e^{2Z_{\hat{\rho}_1}} - 1}{e^{2Z_{\hat{\rho}_1}} + 1} \qquad \text{e} \qquad \hat{\rho}_2 = \frac{e^{2Z_{\hat{\rho}_2}} - 1}{e^{2Z_{\hat{\rho}_2}} + 1} \quad (3.53)$$

3.2.1.10 Confiabilidade

3.2.1.10.1 Confiabilidade de instrumentos de medida

A Análise de Correlação é bastante útil em instrumentos de avaliação, particularmente os de educação (testes), quando se está estudando a confiabilidade do instrumento.

Entende-se por confiabilidade em educação a consistência dos escores obtidos pelos examinandos (alunos) em determinado teste.

Um instrumento é confiável quando um aluno obtém grau X no teste, hoje, e dias após obtém um grau muito próximo daquele. Esta consistência expressa a confiabilidade do teste. Para medir a confiabilidade utiliza-se a Análise de Correlação.

Resultado 3.10:

O coeficiente de confiabilidade é estimado pelo coeficiente de correlação.

Prova:

Considerando que cada medida possa ser avaliada em dois momentos distintos, tem-se, então, duas observações para cada elemento ou indivíduo. Supondo que ambas são referentes a uma mesma característica e ambas sujeitas a erro, então é possível escrever, conforme apresentado em FERGUSON (1981):

$$X_{1i} = X_i + e_{1i} \tag{3.54}$$

$$X_{2i} = X_i + e_{2i} \tag{3.55}$$

onde: X_{1i} é a primeira medida obtida para indivíduo i;

X_{2i} é a segunda medida obtida para indivíduo i;

X_i é a medida verdadeira do indivíduo i;

e_{1i} é o erro da primeira medida do indivíduo i;

e_{2i} é o erro da segunda medida do indivíduo i.

Assim, é possível escrever os modelos:

$$(X_{1i} - \mu) = (X_i - \mu) + e_{1i}$$

$$(X_{2i} - \mu) = (X_i - \mu) + e_{2i}$$

e fazendo o produto das duas equações tem-se:

$$(X_{1i} - \mu)(X_{2i} - \mu) = \{[(X_i - \mu) + e_{1i}] \times [(X_i - \mu) + e_{2i}]\}$$

$$(X_{1i} - \mu)(X_{2i} - \mu) = \left[(X_i - \mu)^2 + (X_i - \mu) \times e_{2i} + e_{1i} \times (X_i - \mu) + e_{1i} \times e_{2i}\right]$$

e fazendo o somatório e dividindo por $N\sigma_1\sigma_2$, obtém-se:

$$\frac{\sum_{i=1}^{N}(X_{1i} - \mu)(X_{2i} - \mu)}{N\sigma_1\sigma_2} = \frac{\sum_{i=1}^{N}\left[(X_i - \mu)^2 + (X_i - \mu) \times e_{2i} + e_{1i} \times (X_i - \mu) + e_{1i} \times e_{2i}\right]}{N\sigma_1\sigma_2}$$

$$\frac{\sum_{i=1}^{N}(X_{1i} - \mu)(X_{2i} - \mu)}{N\sigma_1\sigma_2} = \frac{\sum_{i=1}^{N}(X_i - \mu)^2 + \sum_{i=1}^{N}(X_i - \mu) \times e_{2i} + \sum_{i=1}^{N} e_{1i} \times (X_i - \mu) + \sum_{i=1}^{N} e_{1i} \times e_{2i}}{N\sigma_1\sigma_2}$$

E, ainda, supondo que os erros sejam aleatórios e não correlacionados com a verdadeira medida, os três últimos termos da expressão acima são iguais a zero e $\sigma_1 = \sigma_2 = \sigma$. Assim, obtém-se:

$$\frac{\sum_{i=1}^{N}(X_{1i}-\mu)(X_{2i}-\mu)}{N\sigma_1\sigma_2} = \frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N\sigma_1\sigma_2} = \frac{N\sigma_i^2}{N\sigma^2} = \frac{\sigma_i^2}{\sigma^2}, \text{ logo}$$

$$\rho_{X_1X_2} = \frac{\sigma_i^2}{\sigma^2} \tag{3.56}$$

onde: $\rho_{X_1X_2}$ é o coeficiente de confiabilidade;

σ_i^2 é a variância verdadeira;

σ^2 é a variância observada.

Como se pode observar, $\rho_{X_1X_2}$ é o coeficiente de correlação entre as duas medidas, que representa o coeficiente de confiabilidade. Quando as medidas referem-se às amostras, o coeficiente de confiabilidade será obtido a partir de:

$$\hat{\rho}_{X_1X_2} = \frac{S_i^2}{S^2} \tag{3.57}$$

onde: $\hat{\rho}_{X_1X_2}$ é o coeficiente de confiabilidade amostral;

S_i^2 é a variância amostral verdadeira;

S^2 é a variância amostral observada.

É possível ainda considerar um teste constituído por n itens, aplicado a uma amostra de N indivíduos. Seja $P_1, P_2, ..., P_n$ o número total de escores obtidos em cada um dos itens, pelos N indivíduos. A proporção média de acertos do item i é $p_i = \frac{P_i}{N}$, e a variância $S_i^2 = p_i(1-p_i) = p_iq_i$.

Representando-se por $X_1, X_2, ..., X_N$ o total de acertos (escores) de N indivíduos, tem-se:

$\overline{X} = \frac{\sum_{j=1}^{N} X_j}{N}$, a média de escores do teste

$S_X^2 = \frac{\sum_{j=1}^{N} (X_j - \overline{X})^2}{N-1}$, a variância de escores do teste

Em testes constituídos por diferentes itens, cada item está correlacionado com os outros itens. Assim, é possível obter a variância total S_X^2 através de

$S_X^2 = \sum_{i=1}^{n} S_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \hat{\phi}_{ij} S_i S_j$ onde $\hat{\phi}_{ij}$ é o Coeficiente de Correlação Phi, que será apresentado na seção 3.2.7

$S_X^2 - \sum_{i=1}^{n} S_i^2 = 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \hat{\phi}_{ij} S_i S_j$, mas $S_i^2 = p_i(1-p_i) = p_i q_i$

e $S_X^2 - \sum_{i=1}^{n} p_i q_i = 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \hat{\phi}_{ij} S_i S_j$

Segundo GUILFORD (1950), a verdadeira variância está na covariância (expressão à direita da igualdade da equação acima). Assim, a verdadeira variância poderá ser definida como se segue:

$S_V^2 = S_X^2 - \sum_{i=1}^{n} p_i q_i$

dividindo-se a expressão acima por S_X^2

$\hat{\rho}_{XX} = \frac{S_V^2}{S_X^2} = \frac{S_X^2 - \sum_{i=1}^{n} p_i q_i}{S_X^2}$, que é o método de consistência interna, descrito no item d, a seguir.

A partir dos modelos apresentados foram desenvolvidos diferentes métodos para estimar a confiabilidade:

a) Método do Teste-Reteste

Neste método, o mesmo instrumento de medida é aplicado em duas ocasiões distintas para a mesma amostra. Calcula-se, então, o Coeficiente de Correlação Linear de Pearson para o conjunto de medidas. O tempo decorrido entre a aplicação dos testes é importante, pois quanto maior o tempo transcorrido entre os dois testes menor é a correlação. O teste é freqüentemente utilizado para calcular a confiabilidade de testes escritos, sendo conhecido como coeficiente de estabilidade.

b) Método da Forma Paralela

É também conhecido como forma equivalente. Neste método, administra-se um teste da forma “A” para um grupo de pessoas, e imediatamente após administra-se um teste da forma “B”, com o mesmo conteúdo. As duas formas são feitas com os mesmos tipos de itens. O Coeficiente de Correlação Linear de Pearson é calculado para o conjunto de escores dos dois testes.

c) Método *Split-Half*

Sua vantagem é que necessita somente de um conjunto de dados. Neste método, normalmente considera-se o número de acertos das questões pares e o número de acertos das questões ímpares. Ou, ainda, as duas primeiras questões para o primeiro escore, as duas seguintes para o segundo escore, e assim alternadamente. Não é aconselhável fazer a divisão dos itens exatamente ao meio, pois é comum as primeiras questões serem mais fáceis do que as últimas. O Coeficiente de Correlação Linear de Pearson é calculado para o conjunto de escores.

d) Método de Consistência Interna

Este método era inicialmente utilizado para escores dicotômicos, como, por exemplo, 1 para “certo” e zero para “errado”. Conforme citado por FERGUSON (1981, p. 438), KUDER e RICHARDSON desenvolveram um método para obter o coeficiente de confiabilidade usando estatística de teste de itens. Uma estimativa da confiabilidade é dada por:

$$\hat{\rho}_{XX} = \frac{n}{n-1}\frac{S_X^2 - \sum_{i=1}^{n} p_i q_i}{S_X^2} = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} p_i q_i}{S_X^2}\right] \quad (3.58)$$

onde: $\hat{\rho}_{XX}$ é o coeficiente de confiabilidade de KUDER e RICHARDSON;

n é o número de itens;

S_X^2 é a variância de escores do teste obtida por: $S_x^2 = \frac{\sum_{j=1}^{N}\left(escj - \overline{esc}\right)^2}{N-1}$;

N é o total de examinados (participantes do teste);

escj é o total de escores do teste para cada examinando;

$\overline{esc}$ é a média dos escores do teste;

$\sum_{i=1}^{n} p_i q_i$ é a soma do produto de proporções de acertos e erros em cada item i.

Lee Cronbach generalizou a expressão de KUDER e RICHARDSON para o caso em que os itens não são todos dicotômicos (CRONBACH, 1951). Esta expressão recebeu o nome de “alfa de Cronbach”, apresentada a seguir:

$$\alpha = \frac{n}{n-1}\frac{S^2 - \sum_{i=1}^{n} S_i^2}{S^2} = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} S_i^2}{S^2}\right] \quad (3.59)$$

onde: α é o coeficiente alfa de CRONBACH;

n é o número de itens;

S^2 é a variância dos escores do teste obtida por: $S^2 = \frac{\sum_{j=1}^{N}\left(escj - \overline{esc}\right)^2}{N-1}$;

N é o total de examinados (participantes do teste);

escj é o total de escores do teste para cada examinando;

$\overline{esc}$ é a média dos escores do teste;

S_i^2 é a variância dos escores no item i obtida por: $S_i^2 = \frac{\sum_{j=1}^{N}\left(escij - \overline{esc_i}\right)^2}{N-1}$;

escij é o escore do examinando j no item i;

$\overline{esc_i}$ é a média dos escores do item i.

3.2.1.10.1.1 Correção de atenuação do coeficiente de correlação

Uma importante utilização do coeficiente de confiabilidade, apresentada por GUILFOD (1950) e FERGUSON (1981), é para solucionar o problema de erros de medida.

É importante considerar a possibilidade de erros de medida das variáveis envolvidas. Tais erros, já descritos, têm influência direta no coeficiente de correlação. Os erros normalmente tendem a diminuir o coeficiente de correlação entre as duas variáveis.

Resultado 3.11: O estimador do coeficiente de correlação corrigido ou desatenuado é conforme a expressão a seguir:

$$\hat{\rho}_{X,Y} = \frac{\hat{\rho}_{X',Y'}}{\sqrt{\hat{\rho}_{X',X'}\hat{\rho}_{Y',Y'}}} \tag{3.60}$$

onde: $\hat{\rho}_{X,Y}$ é o coeficiente de correlação corrigido ou desatenuado;

$\hat{\rho}_{X',Y'}$ é o coeficiente de correlação entre as variáveis X' e Y' (observadas);

$\hat{\rho}_{X',X'}$ é o coeficiente de confiabilidade da variável X' (observada);

$\hat{\rho}_{Y',Y'}$ é o coeficiente de confiabilidade da variável Y' (observada).

Prova:

Sejam as variáveis observadas:

$$X' = X + u$$

$$Y' = Y + v$$

onde: X' e Y' são as variáveis observadas;

X e Y são as variáveis verdadeiras (sem erros de medidas);

u e v são os erros de medidas das variáveis X e Y, respectivamente.

O coeficiente de correlação entre as variáveis observadas X' e Y' é conforme a expressão (3.11) do resultado 3.1:

$$\hat{\rho}_{X',Y'} = \frac{\sum_{i=1}^{n} (X'_i - \overline{X'})\,(Y'_i - \overline{Y'})}{n\sqrt{\sum_{i=1}^{n} \frac{(X'_i - \overline{X'})^2}{n} \sum_{i=1}^{n} \frac{(Y'_i - \overline{Y'})^2}{n}}} = \frac{\sum_{i=1}^{n} x'_i y'_i}{n\hat{\sigma}_{X'}\hat{\sigma}_{Y'}}$$

As variáveis observadas podem ser substituídas pelas verdadeiras, mais os erros de medidas.

Utilizaram-se as seguintes notações para cada uma das variáveis: $x_i = X_i - \overline{X}$, $y_i = Y_i - \overline{Y}$, $x'_i = X'_i - \overline{X'}$ e $y'_i = Y'_i - \overline{Y'}$.

Reescrevendo a expressão anterior tem-se:

$$\hat{\rho}_{X',Y'} = \frac{\sum_{i=1}^{n} (x_i + u_i)(y_i + v_i)}{n\hat{\sigma}_{X'}\hat{\sigma}_{Y'}} = \frac{\sum_{i=1}^{n} (x_i y_i + x_i v_i + y_i u_i + u_i v_i)}{n\hat{\sigma}_{X'}\hat{\sigma}_{Y'}}$$

$$\hat{\rho}_{X',Y'} = \frac{\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} x_i v_i + \sum_{i=1}^{n} y_i u_i + \sum_{i=1}^{n} u_i v_i}{n\hat{\sigma}_{X'}\hat{\sigma}_{Y'}}$$

Supondo que os erros sejam independentes entre si e de x e y, tem-se

$\hat{\rho}_{X',Y'} = \frac{\sum_{i=1}^{n} x_i y_i}{n\hat{\sigma}_{X'}\hat{\sigma}_{Y'}}$, mas $\hat{\rho}_{X,Y} = \frac{\sum_{i=1}^{n} x_i y_i}{n\hat{\sigma}_X\hat{\sigma}_Y}$ e portanto $\sum_{i=1}^{n} x_i y_i = \hat{\rho}_{X,Y} n\hat{\sigma}_X\hat{\sigma}_Y$

então $\hat{\rho}_{X',Y'} = \frac{\hat{\rho}_{X,Y} n\hat{\sigma}_X\hat{\sigma}_Y}{n\hat{\sigma}_{X'}\hat{\sigma}_{Y'}} = \hat{\rho}_{X,Y} \frac{\hat{\sigma}_X}{\hat{\sigma}_{X'}} \frac{\hat{\sigma}_Y}{\hat{\sigma}_{Y'}}$

Tem-se, do resultado 3.10 (expressão 3.57), que o coeficiente de confiabilidade é medido pela razão entre a variância verdadeira e a variância observada da variável X. Assim, tem-se $\frac{\hat{\sigma}_X}{\hat{\sigma}_{X'}} = \sqrt{\hat{\rho}_{X',X'}}$ e $\frac{\hat{\sigma}_Y}{\hat{\sigma}_{Y'}} = \sqrt{\hat{\rho}_{Y',Y'}}$ e é possível escrever a expressão como segue:

$\hat{\rho}_{X',Y'} = \hat{\rho}_{X,Y} \sqrt{\hat{\rho}_{X',X'}} \sqrt{\hat{\rho}_{Y',Y'}}$, e portanto $\hat{\rho}_{X,Y} = \frac{\hat{\rho}_{X',Y'}}{\sqrt{\hat{\rho}_{X',X'}} \sqrt{\hat{\rho}_{Y',Y'}}}$

3.2.1.10.1.2 Aplicação da correção de atenuação

A aplicação descrita a seguir refere-se ao trabalho realizado por SILVEIRA e PINENT (2001), cujo objetivo foi estudar a validade e o poder decisório da redação em concursos de ingresso à universidade no Brasil.

O estudo foi elaborado com os dados dos participantes do Concurso Vestibular de 1999 da Universidade Federal do Rio Grande do Sul (CV-UFRGS) e da Pontifícia Universidade Católica do Rio Grande do Sul (CV-PUCRS). O total de candidatos foi de 35.787 e 10.547, respectivamente da UFRGS e da PUCRS. Destes, 6.516 candidatos participaram dos dois concursos, os quais foram objeto do estudo.

Os candidatos do CV-UFRGS receberam o total de escores entre zero e trinta para cada uma das nove provas a que se submeteram, mais o escore final da redação, entre zero e vinte. Com relação aos candidatos do CV-PUCRS, os candidatos receberam o total de escores em quatro pares de provas (Biologia e Química, Física e Matemática, História e Geografia, Língua Estrangeira e Literatura Brasileira), acrescidos das provas de Língua Portuguesa e de Redação. Para os pares de provas os escores variaram entre zero e cinqüenta, para Língua Portuguesa de zero a vinte e cinco, e para a prova de Redação entre zero e quatro.

A tabela 1 apresenta os resultados obtidos, pelos autores, de coeficientes de confiabilidade[11] das provas nos CV-UFRGS e CV-PUCRS, os coeficientes de

---

[11] Os autores utilizaram a seguinte expressão: $\rho_y = 1 - \dfrac{\sum_{i=1}^{n}\left[(1-\rho_i)S_i^2\right]}{S_y^2}$

onde: ρ_y = coeficiente de confiabilidade do escore obtido da soma de dois ou mais escores X_i

ρ_i = coeficiente de confiabilidade do escore X_i

y = escore total ou $y = \sum_{i=1}^{n} x_i$

S_i^2 = variância do escore X_i

S_y^2 = variância do escore y

Correlação Linear de Pearson entre os escores obtidos nos dois concursos para cada par de provas e o coeficiente de correlação desatenuado.

TABELA 1 - COEFICIENTES DE CONFIABILIDADE E DE CORRELAÇÃO ENTRE OS ESCORES DAS PROVAS DO CONCURSO VESTIBULAR DA UFRGS E DA PUCRS - 1999

PROVA	COEFICIENTE DE CONFIABILIDADE		COEFICIENTE DE CORRELAÇÃO DE PEARSON	COEFICIENTE DE CORRELAÇÃO DESATENUADO
	PUCRS	UFRGS		
Biologia e Química	0,84	0,79	0,80	0,98
Matemática e Física	0,84	0,84	0,78	0,93
História e Geografia	0,83	0,79	0,80	0,99
Língua Estrangeira e Língua Brasileira	0,81	O,84	0,79	0,96
Língua Portuguesa	0,66	0,69	0,52	0,92
Número total de acertos nas 9 provas	0,95	0,96	0,92	0,96

FONTE: SILVEIRA e PINENT (2001)

Conforme os autores, os coeficientes de confiabilidade das provas da PUCRS e UFRGS que versam sobre os mesmos conteúdos (por exemplo, biologia e química, matemática e física, etc.) são semelhantes, aproximadamente iguais aos coeficientes de correlação linear de Pearson.

Os coeficientes de correlação desatenuado ou corrigido são próximos de um, indicando que as provas do CV-UFRGS e do CV-PUCRS medem os mesmos conteúdos.

3.2.1.10.1.3 Aplicação da correção para restrição em variabilidade

No trabalho apresentado na seção 3.2.1.10.1.2, de SILVEIRA e PINENT (2001), pode-se encontrar a aplicação da Correção para a Restrição em Variabilidade (resultado 3.4), quando apresentam os coeficientes de correlação para os 6.516 candidatos das duas universidades, obtidos a partir de estatísticas de um grupo menor de candidatos.

No CV-UFRGS, as redações são avaliadas somente quando o candidato faz mais de 108 acertos (40%) do total de 270 questões de escolha múltipla. Do total de

6.516 candidatos comuns aos dois concursos, 4.184 tiveram a redação avaliada no CV-UFRGS.

A partir do grupo composto por 4.184 candidatos, foram estimados os coeficientes de correlação dos escores na redação do CV-UFRGS de todos os candidatos (6.516), com as demais provas dos dois concursos, incluindo a redação do CV-PUCRS (tabela 2).

TABELA 2 - COEFICIENTE DE CORRELAÇÃO ENTRE OS ESCORES DA PROVA DE REDAÇÃO E OUTRAS PROVAS DO CONCURSO VESTIBULAR DA UFRGS E DA PUCRS - 1999

PROVA	COEFICIENTE DE CORRELAÇÃO ENTRE OS ESCORES DA PROVA DE REDAÇÃO E OUTRAS PROVAS	
	CV-UFRGS	CV-PUCRS
Biologia e Química - PUCRS	0,29	0,39
Biologia e Química - UFGRS	0,28	0,38
Matemática e Física - PUCRS	0,24	0,36
Matemática e Física - UFRGS	0,20	0,34
História e Geografia - PUCRS	0,32	0,40
História e Geografia - UFRGS	0,29	0,38
Língua Estrangeira e Língua Brasileira - PUCRS	0,47	0,47
Língua Estrangeira e Língua Brasileira - UFRGS	0,49	0,46
Língua Portuguesa - PUCRS	0,49	0,44
Língua Portuguesa - UFRGS	0,55	0,46
Redação - PUCRS	0,41	
Redação - UFRGS		0,41

FONTE: SILVEIRA e PINENT (2001)

Os coeficientes de correlação apresentados na tabela 2 são relativamente baixos, podendo indicar que, segundo os autores, a questão de redação avalia aspectos independentes aos que são medidos em questão de múltipla escolha.

Ainda, os autores concluem que há fortes indícios de que a confiabilidade dos escores de redação é pequena, de forma que a correlação entre a redação e outra prova não poderá ser grande.

3.2.1.10.2 Confiabilidade em Sistemas de Engenharia

O objetivo da confiabilidade em sistemas de engenharia é avaliar a segurança de um sistema. Assim, tem-se a avaliação da probabilidade de não haver falha durante a sua vida útil, atendendo aos objetivos para os quais o sistema foi projetado.

3.2.1.10.2.1 Confiabilidade estrutural

A avaliação da probabilidade de falha tem como base a função de performance do sistema, conhecida como função de estado limite, ou função de falha ou margem de segurança, representada por $g(\underline{X})$, onde $\underline{X}$ é o vetor de variáveis aleatórias envolvidas na análise. A avaliação da probabilidade de falha é usualmente identificada como análise de confiabilidade estrutural.

Sendo $f_x(\underline{X})$ a função densidade de probabilidades conjunta das variáveis aleatórias $\underline{X}$, a probabilidade de falha pode ser obtida através de:

$P(\text{falha}) = \int_F f_x(\underline{X})dx$ onde F indica o domínio de falha $(g(\underline{X}) \leq 0)$.

A avaliação da expressão acima não é simples, pois envolve a avaliação de uma integral n-dimensional com domínio $(g(\underline{X}) \leq 0)$, onde n é o número de variáveis aleatórias de $\underline{X}$. Em função da dificuldade, métodos alternativos são utilizados. Citem-se dois métodos analíticos bastante utilizados:

(i) First Order Reliability Method (FORM): Quando se tem uma função de falha linear, a confiabilidade pode ser obtida através da distância da função até a origem.

Neste método, as variáveis aleatórias $\underline{X}$, com distribuições quaisquer, podendo ser dependentes ou não entre si, são transformadas em variáveis normais padrões $\underline{X}'$, estatisticamente independentes.

(ii) Second Order Reliability Method (SORM): A diferença deste método para o anterior está na aproximação feita para a superfície de falha. Neste método, faz-se uma aproximação por uma superfície não-linear (quadrática), em vez de linear.

Os métodos FORM e SORM assumem implicitamente (HALDAR e MAHADEVAN, 2000) que as variáveis $(X_1, X_2, ..., X_n)$ são não correlacionadas. Deve-se, inicialmente, portanto, obter a matriz de correlação dessas variáveis. Considerando a função de estado limite $g(X_1, X_2, ..., X_n)$, a matriz de covariância será representada como:

$$[C]=\begin{bmatrix} \sigma_{X_1}^2 & \mathrm{cov}(X_1,X_2) & ... & \mathrm{cov}(X_1,X_n) \\ \mathrm{cov}(X_2,X_1) & \sigma_{X_2}^2 & ... & \mathrm{cov}(X_2,X_n) \\ ... & ... & ... & ... \\ \mathrm{cov}(X_n,X_1) & \mathrm{cov}(X_n,X_2) & ... & \sigma_{X_n}^2 \end{bmatrix} \tag{3.61}$$

Definindo as variáveis padronizadas como: $X_i' = \dfrac{X_i - \mu_{X_i}}{\sigma_{X_i}}$ $(i = 1,2,...,n)$, então a matriz $[C']$ será:

$$[C']=\begin{bmatrix} 1 & \rho_{X_1,X_2} & ... & \rho_{X_1,X_n} \\ \rho_{X_2,X_1} & 1 & ... & \rho_{X_2,X_n} \\ ... & ... & ... & ... \\ \rho_{X_n,X_1} & \rho_{X_n,X_2} & ... & 1 \end{bmatrix} \tag{3.62}$$

onde ρ_{X_i,X_j} é o coeficiente de correlação entre X_i e X_j.

Os métodos FORM e SORM poderão ser utilizados se as variáveis $(X_1,X_2,...,X_n)$ forem transformadas para variáveis não-correlacionadas. Em grande parte dos problemas práticos, as variáveis correlacionadas podem ser transformadas em não-correlacionadas através de uma transformação ortogonal da forma: $Y = L^{-1}(X')'$ onde L é a matriz triangular inferior obtida pela fatoração de Cholesky da matriz $[C']$ (HALDAR e MAHADEVAN, 2000).

3.2.1.10.2.2 Confiabilidade de sistemas

Existem situações em que mais de uma função de performance ou estado limites é envolvida. Neste caso é possível calcular a probabilidade de falha para cada modo ou componente, usando o método FORM, e depois calcular a probabilidade do sistema como um todo, levando-se em conta a contribuição de cada um dos componentes.

Um sistema é chamado em série quando a falha de um de seus componentes leva a falhar o sistema. A probabilidade de falha de um sistema em série pode ser obtida através de (UFRJ. COPPE. PEC):

$$P_i = \Phi(-\beta_i) \tag{3.63}$$

$$P_{ij} = \Phi(-\beta_i,-\beta_j,\rho_{ij}) \tag{3.64}$$

onde: β_i,β_j são os índices de confiabilidade de cada um dos componentes;

ρ_{ij} é a correlação entre os dois componentes, ou seja, $\rho_{ij} = \alpha_i\alpha_j$, onde α_i e α_j são os vetores normais nos pontos de mínimo de cada um dos componentes;

$\Phi(\ )$ é a função cumulativa de probabilidade normal padrão;

$\Phi(,,\rho)$ é a função cumulativa bidimensional normal padrão dada por:

$$\Phi\left(-\beta_i,-\beta_j,\rho_{ij}\right) = \Phi\left(-\beta_i\right)\Phi\left(-\beta_j\right) + \int_0^{\rho_{ij}} \varphi\left(-\beta_i,-\beta_j,z\right)dz \tag{3.65}$$

e $\Phi(,,\rho)$ é a função densidade de probabilidade bidimensional padrão dada por:

$$\Phi\left(x,y,\rho\right) = \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2}\left(\frac{x^2+y^2-2\rho}{1-\rho^2}\right)\right] \tag{3.66}$$

Um sistema é chamado em paralelo quando a falha do sistema ocorre após a falha de todos os seus componentes ou modos.

A probabilidade de falha de um sistema em paralelo, utilizando o método FORM, para o caso de dois componentes, pode ser obtida através de:

$$P_{ij} = \Phi\left(-\beta_i,-\beta_j,\rho_{ij}\right)$$

onde: β_i,β_j são os índices de confiabilidade de cada um dos componentes;

ρ_{ij} é a correlação entre os dois componentes, ou seja, $\rho_{ij} = \alpha_i\alpha_j$, onde α_i e α_j são os vetores normais nos pontos de mínimo de cada um dos componentes;

$\Phi(\ )$ é a função cumulativa de probabilidade normal padrão;

$\Phi(,,\rho)$ é a função cumulativa bidimensional normal padrão dada por:

$$\Phi\left(-\beta_i,-\beta_j,\rho_{ij}\right) = \Phi\left(-\beta_i\right)\Phi\left(-\beta_j\right) + \int_0^{\rho_{ij}} \varphi\left(-\beta_i,-\beta_j,z\right)dz$$

e $\Phi(,,\rho)$ é a função densidade de probabilidade bidimensional padrão dada por:

$$\Phi(x,y,\rho)=\frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2}\left(\frac{x^2+y^2-2\rho}{1-\rho^2}\right)\right],$$ conforme já apresentado anteriormente.

3.2.1.11 Teste de normalidade (Gaussianidade)

FILLIBEN (1975) propõe o teste de normalidade de uma variável através do cálculo do coeficiente de correlação, utilizando a mediana da distribuição normal padronizada.

O autor apresenta algumas vantagens de se utilizar a mediana, em vez da média, pois segundo ele esta última medida apresenta algumas propriedades indesejáveis, tais como: a técnica de integração para o cálculo da média varia drasticamente de uma distribuição para outra; para algumas distribuições, as médias são difíceis de serem obtidas ou requerem grande tempo de cálculo e precisam ser aproximadas e, ainda, em algumas distribuições, a média pode não ser definida.

A proposta apresentada para o cálculo do coeficiente de correlação é:

$$\hat{\rho}=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(M_i-\overline{M}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\sum_{i=1}^{n}\left(M_i-\overline{M}\right)^2}} \tag{3.67}$$

Os valores de M_i necessários para o cálculo da estatística $\hat{\rho}$ são os inversos da função densidade acumulada da distribuição $N(0,1)$ de m_i, ou seja, $M_i=\Phi^{-1}(m_i)$. Os valores de m_i podem ser obtidos através da expressão apresentada a seguir:

$$m_i=\begin{cases}1-m_n & \text{para } i=1\\ (i-0{,}3175)/(n+0{,}365) & \text{para } i=2,3,...,n-1\\ 0{,}5^{1/n} & \text{para } i=n\end{cases} \tag{3.68}$$

Os valores de m_i correspondem às áreas sob a curva normal e, os de M_i, aos respectivos valores de z (distribuição normal padrão).

Para um exemplo prático, considere-se uma amostra aleatória com $n = 200$ observações obtida através do processo de simulação. A variável aleatória é normalmente distribuída com média igual a 92,84155 e variância igual a $(57{,}98319)^2$. Os resultados da simulação são apresentados resumidamente a seguir. A amostra aleatória e as estatísticas calculadas encontram-se no Apêndice 3.

Os valores da variável são ordenados em ordem crescente e os m_i são obtidos conforme a expressão apresentada anteriormente.

Calculou-se inicialmente o $m_{200} = 0{,}5^{(1/200)} = 0{,}99654$, e, após, obteve-se o $m_1 = 1 - m_{200} = 1 - 0{,}99654 = 0{,}00346$. A partir de m_2, até m_{199}, basta substituir o valor de i em: $(i - 0{,}3175)/(n + 0{,}365)$. O m_2 será obtido por: $(2 - 0{,}3175)/(200 + 0{,}365) = 0{,}00840$; $m_3 = (3 - 0{,}3175)/(200 + 0{,}365) = 0{,}01339$ e assim até m_{199} (quadro 2).

Para $m_1 = 0{,}00346$, o valor de z correspondente é -2,70 (áreas sob a curva normal); para $m_2 = 0{,}00840$, o valor de z é -2,39 e assim até $m_{200} = 0{,}99654$, cujo valor de z correspondente é 2,70.

QUADRO 2 - ESTATÍSTICAS DA VARIÁVEL ALEATÓRIA X, SEGUNDO A ORDEM CRESCENTE

i	X_i	$(X_i - \overline{X})$	m_i	M_i
1	-66,08907	-158,93062	0,00346	-2,70067
2	-55,34452	-148,18607	0,00840	-2,39106
3	-54,64370	-147,48525	0,01339	-2,21471
4	-33,28091	-126,12246	0,01838	-2,08842
5	-29,60415	-122,44570	0,02337	-1,98865
6	-14,42701	-107,26856	0,02836	-1,90547
7	-10,38914	-103,23069	0,03335	-1,83369
8	-9,61244	-102,45399	0,03834	-1,77029
9	-8,57903	-101,42058	0,04333	-1,71329
10	-7,46465	-100,30620	0,04832	-1,66137
11	-5,66421	-98,50576	0,05332	-1,61348
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
196	200,60562	107,76407	0,97663	1,98865
197	211,89209	119,05054	0,98162	2,08842
198	212,51855	119,67700	0,98661	2,21471
199	222,03666	129,19511	0,99160	2,39106
200	247,78060	154,93905	0,99654	2,70067

FONTE: A autora

Foram obtidos os seguintes valores, necessários para o cálculo de $\hat{\rho}$:

$$\overline{X} = 92{,}84155 \;;\; \overline{M} = 0 \;;\; \sum_{i=1}^{n}(X_i - \overline{X})(M_i) = 11.424{,}30554 \;;$$

$$\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = 669.048{,}10709 \;;\; \sum_{i=1}^{n} M_1^2 = 195{,}55906$$

A expressão (3.67) pode ser apresentada de forma resumida, pois $\overline{M} = 0$.

$$\hat{\rho} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right) M_i}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \sum_{i=1}^{n} M_i^2}} \qquad (3.69)$$

Substituindo os valores na expressão acima, obtém-se o coeficiente de correlação $\hat{\rho} = 0{,}99876$. Este coeficiente é superior ao valor crítico (quadro A.2.1 do Anexo 2) igual a 0,98700, para nível de significância de 5%. Portanto, aceita-se a hipótese H_0 de que a variável aleatória X é normalmente distribuída.

3.2.2 Coeficiente de Correlação Bisserial

3.2.2.1 Introdução

O Coeficiente de Correlação Bisserial é uma estimativa do Coeficiente de Correlação Linear de Pearson entre uma variável contínua X e uma variável "latente" Y_L (contínua e normal), subjacente à variável dicotômica Y (LORD e NOVICK, 1967), (FERGUSON, 1976) e (WHERRY, 1984).

Uma aplicação possível deste coeficiente é na análise de itens (questões que geram escores dicotômicos do tipo certo ou errado) de uma prova; utiliza-se então a hipótese de que, subjacente à resposta de cada item, exista uma variável "latente", contínua e normal, que determina o resultado (certo ou errado, zero ou um) no item. O Coeficiente Bisserial estima o Coeficiente de Pearson entre o escore total na prova (X) e a variável "latente", subjacente ao item.

De acordo com GUILFORD (1950), o Coeficiente Bisserial é utilizado em situações em que ambas as variáveis correlacionadas são passíveis de ser medidas como contínuas, mas, por alguma razão, uma delas foi reduzida a duas categorias. Esta redução pode ser em conseqüência de ser a única forma de obtenção dos dados, como, por exemplo, a situação em que o aluno foi aprovado ou reprovado, conforme algum critério.

Quando uma das variáveis (Y) é medida como dicotômica, ou seja, reduzida a duas categorias por alguma razão, e a outra é contínua, o Coeficiente de Correlação Bisserial ($\hat{\rho}_b$) é utilizado, descrito em GUILFORD (1950), DOWNIE e HEATH (1959), McNEMAR (1969) e BUNCHAFT e KELLNER (1999).

3.2.2.2 Estimador do Coeficiente de Correlação Bisserial e do erro padrão

$$\hat{\rho}_b = \frac{\overline{X}_p - \overline{X}_t}{S_t} \times \frac{p}{y} \quad \text{ou} \tag{3.70}$$

$$\hat{\rho}_b = \frac{\overline{X}_p - \overline{X}_q}{S_t} \times \frac{p \times q}{y} \tag{3.71}$$

onde: $\hat{\rho}_b$ é o Coeficiente de Correlação Bisserial;

$\overline{X}_p$ é a média dos valores de X para o grupo superior (grupo cujos valores de X estão acima do ponto de dicotomização da variável Y);

$\overline{X}_q$ é a média dos valores de X para o grupo inferior (grupo cujos valores de X estão abaixo do ponto de dicotomização da variável Y);

$\overline{X}_t$ é a média total de X da amostra;

S_t é o desvio padrão total de X da amostra;

p é a proporção de casos do grupo superior (grupo cujos valores de X estão acima do ponto de dicotomização da variável Y);

q é a proporção de casos do grupo inferior (grupo cujos valores de X estão abaixo do ponto de dicotomização da variável Y);

y é a ordenada da distribuição normal no ponto de dicotomização (p) da variável Y. Inicialmente obtém-se o valor de z, correspondente à área menor ou igual a p e calcula-se y = f (z), dada por $f(z) = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}$.

Erro padrão do Coeficiente de Correlação Bisserial:

$$\hat{\sigma}_{\hat{\rho}b} = \frac{\frac{\sqrt{pq}}{y} - \hat{\rho}_b^2}{\sqrt{n}} \tag{3.72}$$

onde: $\hat{\sigma}_{\hat{\rho}b}$ é o erro padrão;

$\hat{\rho}_b^2$ é o quadrado do Coeficiente de Correlação Bisserial;

n é o número de observações da amostra.

Para testar a hipótese nula de que $\rho_b = 0$ o estimador do erro padrão será:

$$\hat{\sigma}_{\hat{\rho}b} = \frac{\frac{\sqrt{pq}}{y}}{\sqrt{n}} \tag{3.73}$$

3.2.2.3 Suposições básicas para a utilização do Coeficiente de Correlação Bisserial

As suposições básicas para a utilização da Correlação Bisserial são apresentadas em GUILFORD (1950), McNEMAR (1969) e BUNCHAFT e KELLNER (1999). A primeira é que a variável Y seja medida como dicotômica, porém existindo uma variável normal e contínua, subjacente a ela. Como segunda suposição, a variável X deve ser contínua.

Segundo GUILFORD (1950), a utilização das quantidades p, q, e y, na expressão (3.70) e (3.71), está diretamente associada à distribuição normal da variável subjacente à variável dicotômica. Não sendo normalmente distribuída, recairá numa estimativa não confiável da correlação.

Finalmente, a variável Y deve ser dicotomizada (ao ser medida) em um ponto mais próximo possível da mediana.

Quando $\hat{\rho}_b = 0{,}00$, o erro padrão de $\hat{\rho}_b$ é pelo menos 25% maior que de $\hat{\rho}$, para o mesmo tamanho de amostra. À medida que p se aproxima de 1,0 ou 0,0, a razão $\frac{\sqrt{p \times q}}{y}$ torna-se maior. Para $p = 0{,}94$, o valor da ordenada y é igual a 0,1200 e

esta razão é igual a 2,0. Para $p = 0{,}5$, o valor de y é $0{,}3989$, e a razão assume o menor valor, igual a 1,25. Esta é, segundo GUILFORD (1950), a razão pela qual se recomenda que a dicotomização de Y seja feita mais próxima da mediana.

GUILFORD ainda se refere à diferença entre as médias para o cálculo do Coeficiente de Correlação Bisserial, como pode ser visto em (3.71). A diferença não é muito estável, a não ser que as amostras sejam grandes. Segundo ele, mesmo que a amostra seja de 1.000 casos, se apenas 1% dos casos estiver em uma das categorias (0 ou 1), a média é baseada em 10 casos, o que não é favorável para realizar estimativas com base nessa média.

Comparando-se as características das duas correlações, a de Pearson e a Bisserial, sempre que possível é preferível utilizar a primeira, principalmente quando a amostra é pequena (GUILFORD, 1950).

3.2.2.4 Aplicação do Coeficiente de Correlação Bisserial

A aplicação descrita a seguir refere-se ao trabalho realizado por CHAVES NETO e TURIM (2003). O objetivo do estudo foi abordar as teorias da avaliação educacional, tanto a Teoria Clássica, quanto a Teoria de Resposta ao Item (TRI) nos seus vários aspectos.

Para CHAVES NETO e TURIM (2003), o instrumento de medida educacional é um dos aspectos mais importantes da avaliação escolar. E, para eles, os bons instrumentos de avaliação normalmente têm as seguintes propriedades: validade, confiabilidade, objetividade e praticabilidade.

Ainda, é desejável, segundo os autores, que os itens que compõem o instrumento tenham as características do grau de discriminação e de dificuldade, conhecidos *a priori*. Assim, é possível classificar os examinandos (alunos) em três grupos: bom, médio e fraco.

Foram aplicados testes avaliativos em 5 escolas da rede municipal, do período matutino, envolvendo as disciplinas de Língua Portuguesa e Matemática, do

município de Andirá. Participaram todos os alunos devidamente matriculados nas 3.ª e 4.ª séries do ensino fundamental regular, num total de aproximadamente 1.400 alunos.

O teste de Língua Portuguesa, aplicado nas 3ª. e 4ª. séries, compreendeu três partes:

- parte I: interpretação de textos;
- parte II: produção de textos;
- parte III: leitura de textos.

A discriminação de cada item foi estimada tanto pela Teoria de Resposta ao Item (TRI), quanto pela Teoria Clássica. Na análise utilizando a Teoria Clássica, a estimação da discriminação do item foi feita calculando-se o Coeficiente de Correlação Bisserial e o Coeficiente de Correlação de Pearson.

O quadro 3 apresenta os Coeficientes de Correlação de Pearson e Bisserial, calculados entre a pontuação total (X) e resposta de cada item (Y), no teste de interpretação de texto dos alunos da 3.ª série, totalizando 369 examinandos (alunos).

QUADRO 3 - COEFICIENTES DE CORRELAÇÃO DE PEARSON E BISSERIAL ENTRE A PONTUAÇÃO TOTAL E RESPOSTA DE CADA ITEM, NO TESTE DE INTERPRETAÇÃO DE TEXTO DA 3.ª SÉRIE, DAS ESCOLAS MUNICIPAIS DE ANDIRÁ

NÚMERO DO ITEM	TOTAL DE ALUNOS EXAMINANDOS	ACERTOS	COEFICIENTE DE CORRELAÇÃO	
			Pearson	Bisserial
01	369	311	0,356	0,539
02	369	292	0,325	0,460
03	369	208	0,471	0,593
04	369	237	0,492	0,631
05	369	150	0,476	0,602
06	369	202	0,469	0,589
07	369	126	0,382	0,494
08	369	272	0,451	0,609
09	369	233	0,540	0,691
10	369	268	0,495	0,663
11	369	296	0,433	0,620
12	369	294	0,551	0,785
13	369	221	0,505	0,640
14	369	187	0,423	0,530
15	369	314	0,207	0,317
16	369	226	0,366	0,466
17	369	261	0,433	0,573
18	369	261	0,523	0,692
19	369	268	0,511	0,684
20	369	306	0,451	0,669

FONTE: CHAVES NETO e TURIM (2003)

Quanto maior o coeficiente de correlação, maior é a discriminação do item. Observa-se, no quadro, que o item de maior discriminação é o 12, pois apresenta Coeficiente de Correlação Bisserial igual a 0,785.

3.2.3 Coeficiente de Correlação Ponto Bisserial

3.2.3.1 Introdução

Embora seja usada normalmente como medida de correlação entre escores e itens de testes, a Correlação Ponto Bisserial pode ser empregada em outras situações, onde a variável dicotômica pode ser, a título de exemplo, gênero masculino ou feminino, pessoas normais ou neuróticas, etc.

O Coeficiente de Correlação Ponto Bisserial ($\hat{\rho}_{pb}$) é derivado do Coeficiente de Correlação de Pearson. Este método é indicado quando uma das variáveis (Y) é dicotômica e a outra é contínua.

Conforme apresentado em FERGUSON (1981), a Correlação Ponto Bisserial fornece uma medida da relação entre uma variável contínua, como escores de testes, e outra variável com duas categorias ou dicotômicas, como aprovado ou reprovado.

Segundo GUILFORD (1950), DOWNIE e HEATH (1959) e FERGUSON (1981), a Correlação Ponto Bisserial é a Correlação do Momento Produto. Se se atribuir 1 para observações de uma categoria e zero para outra, e se calcular o Coeficiente de Correlação do Momento Produto, o resultado será o Coeficiente Ponto Bisserial. Ele é interpretado da mesma forma que $\hat{\rho}$.

3.2.3.2 Estimador do Coeficiente de Correlação Ponto Bisserial e do erro padrão

O estimador do Coeficiente de Correlação Ponto Bisserial foi obtido a partir do estimador do Coeficiente de Correlação Linear de Pearson, conforme apresentado em GUILFORD (1950).

Fazendo $x_i = X_i - \overline{X}$ e $y_i = Y_i - \overline{Y}$, o estimador do coeficiente linear de Pearson é (resultado 3.1):

$$\hat{\rho} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} = \frac{\sum_{i=1}^{n} x_i y_i}{n\sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n} \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n}}} = \frac{\sum_{i=1}^{n} x_i y_i}{n\hat{\sigma}_x \hat{\sigma}_y} \tag{3.74}$$

X é uma variável aleatória contínua e Y uma variável aleatória com distribuição de Bernoulli, tem-se, então, que, por conveniência:

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

$S_y = \sqrt{pq}$, onde p = θ e q = (1 - θ) da distribuição de Bernoulli (conforme resultado 2.1).

Desenvolvendo (3.74) tem-se:

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} \left[X_i Y_i - X_i \overline{Y} - \overline{X} Y_i + \overline{X}\overline{Y}\right]$$

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} X_i Y_i - n\overline{X}\overline{Y} \tag{3.75}$$

Substituindo (3.75) em (3.74) tem-se:

$$\hat{\rho} = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\overline{Y}}{nS_x\sqrt{pq}} \qquad \text{mas} \qquad \sum_{i=1}^{n} X_i Y_i = n_p \times \overline{X}_p$$

e $n\overline{X}\overline{Y} = n \times \overline{X} \times p = n_p \times \overline{X}$, então,

$$\hat{\rho} = \frac{n_p \times \overline{X}_p - n_p \times \overline{X}}{nS_x\sqrt{pq}}$$

Dividindo por n, tem-se:

$$\hat{\rho} = \frac{p \times \overline{X}_p - p \times \overline{X}}{S_x \sqrt{pq}} = \frac{\left(\overline{X}_p - \overline{X}\right) \times p}{S_x \sqrt{pq}}$$

Dividindo por $\sqrt{p}$, tem-se que

$$\hat{\rho}_{pb} = \frac{\left(\overline{X}_p - \overline{X}\right)}{S_x} \sqrt{\frac{p}{q}} \qquad \text{ou} \qquad (3.76)$$

$$\hat{\rho}_{pb} = \frac{\overline{X}_p - \overline{X}_q}{S_x} \sqrt{pq}$$

onde: $\hat{\rho}_{pb}$ é o Coeficiente de Correlação Ponto Bisserial;

$\overline{X}_p$ é a média dos valores de X para o grupo superior (grupo cuja variável Y assume valor 1);

$\overline{X}$ é a média total de X da amostra;

S_x é o desvio padrão total de X da amostra;

p é a proporção de casos do grupo superior (grupo cuja variável Y assume valor 1);

q é a proporção de casos do grupo inferior (grupo cuja variável Y assume valor 0).

Erro padrão do Coeficiente de Correlação Ponto Bisserial:

$$\hat{\sigma}_{\hat{\rho}_{pb}} = \sqrt{\frac{1 - \hat{\rho}_{pb}^2}{n - 2}} \qquad (3.77)$$

onde: $\hat{\sigma}_{\hat{\rho}_{pb}}$ é o erro padrão;

$\hat{\rho}_{pb}^2$ é o quadrado do Coeficiente de Correlação Ponto Bisserial;

n é o número de observações da amostra.

A relação existente entre os Coeficientes de Correlação Bisserial e Ponto Bisserial é apresentada em GUILFORD (1950):

$$\hat{\rho}_b = \hat{\rho}_{pb}\frac{\sqrt{pq}}{y} \quad \text{e} \quad \hat{\rho}_{pb} = \hat{\rho}_b\frac{y}{\sqrt{pq}}$$

3.2.3.3 Suposições básicas para a utilização do Coeficiente de Correlação Ponto Bisserial

Sendo o Coeficiente de Correlação Ponto Bisserial igual ao Coeficiente de Correlação do Momento Produto, a suposição é de relação linear.

O que difere este coeficiente do Coeficiente de Correlação Bisserial é que, neste, a variável Y é originalmente dicotômica, não necessitando ser contínua e nem normalmente distribuída (BUNCHAFT e KELLNER, 1999). Este método é mais utilizado do que o Coeficiente de Correlação Bisserial, pois não exige que a variável Y tenha distribuição normal na população. Havendo qualquer dúvida a respeito da distribuição da variável dicotômica, deve-se utilizar este coeficiente.

3.2.3.4 Coeficiente de Correlação Ponto Bisserial e teste de médias

O cálculo do Coeficiente de Correlação Ponto Bisserial pode ser comparado ao teste de hipóteses para diferença de duas médias (GUILFORD, 1950) e (CHEN e POPOVICH, 2002). A variável contínua (X) representa a característica de interesse para o estudo e a variável dicotômica (Y) representa os grupos. Quando é testada a hipótese de que $H_0 : \rho_{pb} = 0$, isto equivale a testar a hipótese de que $H_0 : \mu_1 - \mu_2 = 0$. Aceitando-se a hipótese $H_0 : \rho_{pb} = 0$, pode-se concluir que as médias dos grupos são iguais.

É possível testar as hipóteses $H_0 : \rho_{pb} = 0$ e $H_1 : \rho_{pb} \neq 0$ utilizando-se a estatística t, pois o Coeficiente de Correlação Ponto Bisserial é o Coeficiente de Correlação Linear de Pearson. A estatística do teste foi obtida no resultado 3.9, dada por:

$$t = \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}} \sim t_{n-2}$$

onde: t é a estatística do teste;

$\hat{\rho}_{pb} = \hat{\rho}$ é o coeficiente de correlação amostral;

n é o número de observações da amostra.

3.2.3.5 Aplicação do Coeficiente de Correlação Ponto Bisserial

Os dados utilizados foram obtidos a partir da Pesquisa Mensal de Emprego (PME) na Região Metropolitana de Curitiba (RMC). A PME é uma pesquisa domiciliar de periodicidade mensal que tem por objetivo acompanhar a situação do mercado de trabalho na RMC. São pesquisadas mensalmente cerca de 10.000 pessoas com 10 anos e mais de idade.

As variáveis da análise foram a renda recebida no trabalho principal pelas pessoas ocupadas na semana de referência, na condição de empregados com carteira de trabalho assinada no setor privado, no grupo de atividade relativa a intermediação financeira e atividades imobiliárias, aluguéis e serviços prestados às empresas, com 11 anos ou mais de estudo e que trabalharam entre 35 e 45 horas, na semana de referência e gênero. Tem-se uma situação em que uma variável é medida em nível intervalar e outra dicotômica. Os dados referentes a esta aplicação encontram-se no Apêndice 4.

Tendo em vista que a variável renda não é normalmente distribuída, fez-se uma transformação logarítmica na variável, pois, conforme descreve SIQUEIRA (1983), a transformação logarítmica reduz a variância, mesmo que a variável original seja bastante heterogênea, e muitas vezes esta transformação também resolve o problema da não-normalidade, pois deixa a nova variável mais próxima da normal.

Após a transformação, calculou-se o Coeficiente de Correlação Ponto Bisserial e o Coeficiente de Correlação Linear de Pearson entre as variáveis logaritmo natural da renda (ln renda) e gênero. O coeficiente estimado foi $\hat{\rho}_{pb} = \hat{\rho} = 0{,}21544$, significativo para $\alpha = 0{,}02$. Evidentemente que as estimativas são iguais, pois trata-se do mesmo coeficiente de correlação.

Calculou-se também o Coeficiente Linear de Pearson entre a variável original renda e gênero. O coeficiente estimado foi $\hat{\rho} = 0{,}18412$, significativo para $\alpha = 0{,}04$.

Cabe destacar que o objetivo foi mostrar que, embora a variável renda não seja normalmente distribuída e tenha sofrido uma transformação logarítmica, os resultados não sofreram grandes alterações, o que mostra a propriedade do Coeficiente de Correlação de Pearson ser quase-invariante frente às transformações monotônicas (ANDERBERG , 1973).

3.2.4 Coeficiente de Correlação Tetracórico

3.2.4.1 Introdução

O Coeficiente de Correlação Tetracórico é uma estimativa do Coeficiente de Correlação Linear de Pearson entre uma variável "latente" (X_L) e uma variável "latente" (Y_L) (ambas contínuas e normais), subjacentes às variáveis dicotômicas X e Y efetivamente observadas (LORD e NOVICK, 1967), (FERGUSON, 1976) e (WHERRY, 1984).

O Coeficiente de Correlação Tetracórico é utilizado na aplicação da Teoria de Resposta ao Item (TRI). Para determinar a dimensionalidade de uma medida, um dos índices utilizados é com base na Análise Fatorial a partir da matriz dos Coeficientes de Correlação Tetracórico. É possível encontrar um maior detalhamento sobre o assunto em NOJOSA (2001).

As literaturas iniciais sobre a análise de dados categóricos tratavam este coeficiente como índice de associação. O assunto causou intenso debate entre estatísticos, como Karl Pearson e G. Udny Yule, sobre como medir a associação. Karl Pearson pensou na tabela de classificação cruzada de uma distribuição contínua bivariada. O Coeficiente de Correlação Tetracórico é uma medida de associação para variáveis contínuas, porém transformadas em tabela 2x2 (AGRESTI, 1990).

Esse coeficiente é utilizado, segundo DOWNIE e HEATH (1959), McNEMAR (1969) e BUNCHAFT e KELLNER (1999), para se relacionar duas variáveis X e Y contínuas, mas dicotomizadas (ao serem medidas) pelo pesquisador, por alguma razão.

3.2.4.2 Estimador do Coeficiente de Correlação Tetracórico e do erro padrão

Apresenta-se, a seguir, a equação tetracórica. A demonstração para a obtenção desta equação, a partir da transformação da distribuição normal bivariada em variáveis dicotômicas, pode ser encontrada em ELDERTON (1953, p. 175).

$$\frac{ad-bc}{yy'n^2} = \hat{\rho}_t + \hat{\rho}_t^2 \frac{zz'}{2} + \hat{\rho}_t^3 \frac{\left(z^2-1\right)\left(z'^2-1\right)}{6} + \hat{\rho}_t^4 \frac{z\left(z^2-3\right)\left(z'^2-3\right)}{24} + \hat{\rho}_t^5 \frac{\left(z^4-6z^2+3\right)\left(z'^4-6z'^2+3\right)}{120} +$$

$$+ \hat{\rho}_t^6 \frac{z\left(z^4-10z^2+15\right)z'\left(z'^4-10z'^2+15\right)}{720} + \hat{\rho}_t^7 \frac{\left(z^6-15z^4+45z^2-15\right)\left(z'^6-15z'^4+45z'^2+15\right)}{5040} + \ldots \qquad (3.78)$$

Após a dicotomização das variáveis X e Y, obtém-se a tabela 2x2, como se segue:

		Variável X: 1	Variável X: 0	TOTAL
Variável Y	1	a	b	a+b
	0	c	d	c+d
TOTAL		a+c	b+d	n

$$p = \frac{(a+b)}{n} \text{ e } q = \frac{(c+d)}{n} = 1-p \qquad (3.79)$$

$$p' = \frac{(a+c)}{n} \text{ e } q' = \frac{(b+d)}{n} = 1-p' \qquad (3.80)$$

n = a + b + c + d (total de observações)

Assim, tem-se que:

a,b,c,d são as freqüências da tabela 2x2;

z é o valor correspondente à área menor ou igual a p. Por exemplo, se p = 0,50, então tem-se que z = 0 (tabela de áreas sob a curva normal);

z´ é o valor correspondente à área menor ou igual a p'. Se p' = 0,50, então tem-se que z´ = 0;

y é o valor da ordenada no ponto p e pode ser obtida fazendo-se $y = f(z) = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}$.

Para o exemplo citado, se z = 0, então $y = f(0) = \frac{e^{-0}}{\sqrt{2\pi}} = 0{,}39894$ (tabela de ordenadas da curva normal);

y´ é o valor da ordenada no ponto p' e pode ser obtida fazendo-se $y' = f(z') = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}$.

GUILFORD (1950) apresenta uma solução aproximada do cálculo do Coeficiente de Correlação Tetracórico, ignorando os termos de grau superior a 2, na expressão (3.78):

$$\frac{ad - bc}{yy'n^2} = \hat{\rho}_t + \hat{\rho}_t^2 \frac{zz'}{2} \qquad (3.81)$$

onde: $\hat{\rho}_t$ é o Coeficiente de Correlação Tetracórico;

a,b,c,d são as freqüências da tabela 2x2;

z é o valor correspondente à área menor ou igual a p;

z´ é o valor correspondente à área menor ou igual a p';

y é o valor da ordenada no ponto p;

y´ é o valor da ordenada no ponto p';

$n = (a + b + c + d)$ é o número de observações da amostra.

Chamando o primeiro termo da expressão (3.81) de **c**; o coeficiente de $\hat{\rho}_t$ de **b** ; e $\frac{zz'}{2}$ de **a**, tem-se uma equação do 2.º grau:

$$a\hat{\rho}_t^2 + b\hat{\rho}_t + c = 0 \qquad (3.82)$$

que poderá ser resolvida através de: $\hat{\rho}_t = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$

Uma outra expressão é apresentada em GUILFORD (1950), utilizando o cosseno[12]:

$$\hat{\rho}_t = \cos\left(\frac{180\sqrt{bc}}{\sqrt{ad} + \sqrt{bc}} \right) \tag{3.83}$$

Quando o produto bc é igual a ad, o ângulo é 90^o e o cosseno é igual a zero, conseqüentemente $\hat{\rho}_t = 0$.

Erro Padrão aproximado do Coeficiente de Correlação Tetracórico:

$$\hat{\sigma}_{\hat{\rho}_t} = \frac{\sqrt{p \times q \times p' \times q'}}{y' \times y \times \sqrt{n}} \sqrt{\left(1 - \hat{\rho}_t^2\right) \times \left[1 - \left(\frac{sen^{-1} \hat{\rho}_t}{90^o} \right)^2 \right]} \tag{3.84}$$

onde: $\hat{\sigma}_{\hat{\rho}_t}$ é o erro padrão;

$\hat{\rho}_t$ é o Coeficiente de Correlação Tetracórico;

$sen^{-1} \hat{\rho}_t$ é o arco seno de $\hat{\rho}_t$;

$n = (a + b + c + d)$ é o número de observações da amostra.

Para testar a hipótese de que $\rho_t = 0$, o que poderá ser feito através da estatística $t = \frac{\hat{\rho}_t}{\hat{\sigma}_{\hat{\rho}_t}}$, o erro padrão poderá ser calculado considerando apenas a primeira parte da expressão (3.84), como apresenta McNEMAR (1969):

$$\hat{\sigma}_{\hat{\rho}_t} = \frac{\sqrt{p \times q \times p' \times q'}}{y' \times y \times \sqrt{n}} \tag{3.85}$$

---

[12]Conforme demonstrado em WONNACOTT e WONNACOTT (1978), existe uma relação entre o Coeficiente de Correlação $\hat{\rho}$ e o $\cos\theta$, $\hat{\rho} = \cos\theta$ e $-1 \leq \cos\theta \leq +1$.

3.2.4.3 Suposições básicas para a utilização do Coeficiente de Correlação Tetracórico

As suposições básicas para a utilização do Coeficiente de Correlação Tetracórico são de que as variáveis X_L e Y_L (latentes) devem ser contínuas e normalmente distribuídas, relacionadas linearmente; ainda, X e Y devem ser dicotomizadas (ao serem medidas) o mais próximo possível à mediana.

O Coeficiente de Correlação Tetracórico ($\hat{\rho}_t$) é menos confiável que o de Pearson, sendo que sua variabilidade é cerca de 50% maior (GUILFORD, 1950), quando $\rho = 0$. Para obter a mesma confiabilidade[13] para o Coeficiente de Correlação Tetracórico que a obtida no Coeficiente de Correlação de Pearson, é necessário o dobro do tamanho da amostra. Recomenda-se que se utilizem amostras superiores a 300.

3.2.4.4 Aplicação do Coeficiente de Correlação Tetracórico

FACHEL (1986) apresenta exemplos de aplicação do Coeficiente de Correlação Tetracórico a partir de dados empíricos. Dentre eles, cita-se o que ela denomina de *Weinreich data*. Uma amostra foi composta de 802 pacientes, e estes foram submetidos a um teste alérgico, em que a resposta para cada um dos 5 itens (causas de alergia) é "nenhuma reação" ou "reação positiva". Os Coeficientes de Correlação Tetracórico foram obtidos para cada par de diferentes causas de alergia e a matriz de correlação tetracórica é apresentada no quadro 4. Os 5 tipos de itens do teste alérgico foram: 1) *onion couch*; 2) *fescue grass*; 3) *couch grass*; 4) *cock's foot grass*; 5) *rye grass*.

---

[13]A confiabilidade, aqui, é usada como sinônimo de erro padrão.

QUADRO 4 - MATRIZ DE CORRELAÇÃO TETRACÓRICA SEGUNDO ITENS DO TESTE ALÉRGICO

ITENS	ONION COUCH	FESCUE GRASS	COUCH GRASS	COCK´S FOOT GRASS
Fescue grass	0,90	1,00	0,89	0,87
Couch grass	0,88	0,89	1,00	0,88
Cock´s foot grass	0,91	0,87	0,88	1,00
Rye grass	0,81	0,87	0,85	0,81

FONTE: FACHEL (1986)

NOTA: Assumindo que as variáveis são realmente contínuas e normais.

O quadro acima indica que existe alta correlação entre os cinco itens do teste alérgico. Um paciente que apresenta “reação positiva” a um tipo de item também apresenta para os demais. A correlação é maior entre os itens *onion couch* e *cock's foot grass,* com $\hat{\rho}_t = 0{,}91$. Em seguida, entre os itens *onion couch* e *fescue grass,* com $\hat{\rho}_t = 0{,}90$. Os itens que apresentam correlações menores são *rye grass* com os itens *onion couch* $(\hat{\rho}_t = 0{,}81)$ e *cock's foot grass* $(\hat{\rho}_t = 0{,}81)$.

3.2.5 Coeficiente de Correlação de Spearman

3.2.5.1 Introdução

Este coeficiente é o mais antigo e também o mais conhecido para variáveis mensuradas em nível ordinal, chamado também de Coeficiente de Correlação por Postos de Spearman, designado “rho” e representado por $\hat{\rho}_s$. Quando as amostras são pequenas, este método deve ser usado, segundo GUILFORD (1950), em substituição ao Coeficiente de Correlação do Momento Produto. É conveniente para número de pares menor que 30 e quando os dados já estão ordenados.

Para as variáveis cuja mensuração é em nível ordinal, pode-se citar os Coeficientes de Correlação Ordinal de Spearman e Postos de Kendall.

É importante enfatizar, segundo BUNCHAFT e KELLNER (1999), que as correlações ordinais não podem ser interpretadas da mesma maneira que as correlações de Pearson. Inicialmente, não mostram necessariamente tendência linear, mas podem ser consideradas como índices 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 oposto. O quadrado do índice de correlação não pode ser interpretado como a proporção da variância comum às duas variáveis.

3.2.5.2 Estimador do Coeficiente de Correlação de Spearman e significância

Seu estimador foi derivado a partir do estimador do Coeficiente de Correlação Linear de Pearson, conforme apresentado em SIEGEL (1975).

$$\hat{\rho} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} \tag{3.86}$$

onde: $x_i = X_i - \overline{X}$

$y_i = Y_i - \overline{Y}$

Pode-se escrever: $\sum_{i=1}^{n} X_i = \frac{n(n+1)}{2}$ onde $n = \text{postos} = 1,\ 2,\ 3,...,\ n$

Os quadrados dos postos são: $1^2,\ 2^2,\ 3^2,...,\ n^2$

Então $\sum_{i=1}^{n} X_i^2 = \frac{n(n+1)(2n+1)}{6}$

Assim, $\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 = \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}$

$$\sum_{i=1}^{n} x_i^2 = \frac{n(n+1)(2n+1)}{6} - \frac{[n(n+1)/2]^2}{n}$$

$$\sum_{i=1}^{n} x_i^2 = \frac{(n^2+n)(2n+1)}{6} - \frac{n(n^2+2n+1)}{4}$$

$$\sum_{i=1}^{n} x_i^2 = \frac{n^3 - n}{12} \tag{3.87}$$

Da mesma forma, obtém-se que:

$$\sum_{i=1}^{n} y_i^2 = \frac{n^3 - n}{12} \tag{3.88}$$

Fazendo a diferença de postos:

$$d_i = x_i - y_i$$

elevando ao quadrado tem-se:

$$d_i^2 = (x_i - y_i)^2 = x_i^2 - 2x_i y_i + y_i^2$$

fazendo o somatório:

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2 - 2\sum_{i=1}^{n} x_i y_i \tag{3.89}$$

fazendo $\hat{\rho}_s = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}$, tem-se que $\sum_{i=1}^{n} x_i y_i = \hat{\rho}_s \sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}$ (3.90)

substituindo (3.87), (3.88) e (3.90) em (3.89) tem-se:

$$\sum_{i=1}^{n} d_i^2 = 2\left(\frac{n^3 - n}{12}\right) - 2\hat{\rho}_s \sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}$$

Assim, obtém-se:

$$\hat{\rho}_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{3.91}$$

onde: $\hat{\rho}_s$ é o Coeficiente de Correlação de Spearman;

d_i é a diferença entre as ordenações;

n é o número de pares de ordenações.

Quando a seleção dos elementos que compõem a amostra é feita de forma aleatória, a partir de uma população, é possível determinar se as variáveis em estudo são associadas, na população. Ou seja, é possível testar a hipótese de que as duas variáveis estão associadas na população.

Para amostras superiores a 10, segundo SIEGEL (1975), a significância de um valor obtido de $\hat{\rho}_s$ pode ser verificada através de t calculado pelo estimador apresentado a seguir.

$$t = \hat{\rho}_s \sqrt{\frac{n-2}{1-\hat{\rho}_s^2}} \sim t_{n-2} \qquad (3.92)$$

onde: t é a estatística do teste;

$\hat{\rho}_s$ é o Coeficiente de Correlação de Spearman;

n é o número de pares de ordenações.

Para n grande $(n \geq 10)$, a expressão acima tem distribuição t de Student com n-2 graus de liberdade.

3.2.5.3 Suposições para a utilização do Coeficiente de Correlação de Spearman

Segundo SIEGEL (1975), o Coeficiente de Correlação de Spearman é uma medida que exige que as duas variáveis se apresentem em escala de mensuração pelo menos ordinal, de forma que os elementos (indivíduos ou objetos) em estudo formem duas séries ordenadas.

3.2.5.4 Aplicação do Coeficiente de Correlação de Spearman

A aplicação apresentada a seguir refere-se ao trabalho de MENEZES, FAISSOL e FERREIRA (1978), que utilizaram o Coeficiente de Correlação de Spearman para analisar a correlação entre “população total migrante de destino urbano e origem rural” e “população economicamente ativa nas atividades urbanas”,

denominadas de X e Y, respectivamente. Tomaram como unidades observacionais as microrregiões homogêneas do Estado do Paraná. Os dados apresentados na tabela 3 são referentes ao Censo Demográfico de 1970.

TABELA 3 - POPULAÇÃO MIGRANTE TOTAL E ECONOMICAMENTE ATIVA NAS ATIVIDADES URBANAS, SEGUNDO MICRORREGIÕES DO PARANÁ - 1970

MICRORREGIÃO	POPULAÇÃO MIGRANTE TOTAL (X)	POPULAÇÃO ECONOMICAMENTE ATIVA NAS ATIVIDADES URBANAS (Y)
701	42 116	226 657
702	2 448	21 064
703	250	690
704	137	803
705	1 845	10 792
706	14 796	48 967
707	750	3 304
708	613	2 434
709	3 580	11 085
710	623	2 455
711	7 401	13 957
712	28 528	45 664
713	7 172	9 219
714	86 938	111 618
715	39 501	47 809
716	36 216	37 141
717	32 740	34 848
718	45 510	42 589
719	26 437	29 485
720	1 387	2 482
721	40 978	48 198
722	27 713	23 832
723	3 637	17 125
724	6 268	14 318

FONTE: MENEZES, FAISSOL e FERREIRA (1978)

NOTA: População migrante total de destino urbano e origem rural.

O Coeficiente de Correlação de Spearman obtido foi de 0,92, indicando que existe forte correlação entre a população migrante e economicamente ativa, considerando as microrregiões. Os cálculos encontram-se no Apêndice 5 do trabalho.

3.2.6 Coeficiente de Correlação por Postos de Kendall

3.2.6.1 Introdução

O Coeficiente de Correlação por Postos de Kendall (τ) é uma medida de correlação utilizada para dados ordinais, como no caso do Coeficiente de Correlação de Spearman. Ambas as variáveis devem ser medidas no mínimo em nível ordinal, de forma que seja possível atribuir postos a cada uma das variáveis.

3.2.6.2 Estimador do Coeficiente de Correlação por Postos de Kendall e significância

O estimador do Coeficiente de Correlação por Postos de Kendall é definido como apresentado a seguir:

$$\hat{\tau} = \frac{S}{\frac{1}{2}n(n-1)} \tag{3.93}$$

onde: $\hat{\tau}$ é o Coeficiente de Correlação por Postos de Kendall;

n é o número de elementos aos quais se atribuíram postos em X e Y;

S é a soma do número de postos da variável Y à direita que são superiores menos o número de postos à direita que são inferiores.

Para o cálculo do Coeficiente de Correlação por Postos de Kendall ordena-se inicialmente uma das variáveis em ordem crescente de postos e o S correspondente a cada elemento será obtido fazendo o número de elementos cujo posto é superior ao que se está calculando menos o número de elementos cujo posto é inferior ao mesmo.

Encontra-se, no Apêndice 5, o cálculo detalhado do exemplo de aplicação da seção 3.2.6.3.

Quando n é maior que 10, de acordo com SIEGEL (1975), $\hat{\tau}$ pode ser considerado distribuído normalmente com média $(\hat{\mu}_{\hat{\tau}})$ igual a zero e desvio padrão $\hat{\sigma}_{\hat{\tau}})$ dado por:

$$\hat{\sigma}_{\hat{\tau}} = \sqrt{\frac{2(2n+5)}{9n(n-1)}} \tag{3.94}$$

e pode-se obter $Z = \frac{\hat{\tau} - \hat{\mu}_{\hat{\tau}}}{\hat{\sigma}_{\hat{\tau}}}$, que tem distribuição normal com média zero e variância unitária. A significância de z pode ser obtida mediante a tabela da distribuição normal.

Ainda, o autor faz uma comparação entre Coeficiente de Correlação de Spearman e Coeficiente de Correlação por Postos de Kendall. Os valores numéricos não são iguais, quando calculados para os mesmos pares de postos, e não são comparáveis numericamente. Contudo, pelo fato de utilizarem a mesma quantidade de informação contida nos dados, ambos têm o mesmo poder de detectar a existência de associação na população, e rejeitarão a hipótese da nulidade para um mesmo nível de significância.

3.2.6.3 Aplicação do Coeficiente de Correlação por Postos de Kendall

MENEZES, FAISSOL e FERREIRA (1978) calcularam o Coeficiente de Correlação de Kendall para os dados apresentados na tabela 3. O Coeficiente de Correlação obtido foi de 0,79. Apesar de inferior ao obtido pelo método do Coeficiente de Correlação de Spearman, indica que há correlação entre as duas variáveis. Segundo SIEGEL (1975), tanto $\hat{\rho}_s$ como $\hat{\tau}$ apresentam o mesmo poder na rejeição da hipótese de que não há correlação entre as duas variáveis (H_0) e tem eficiência de 91% quando comparados ao $\hat{\rho}$. Os cálculos vêm apresentados no Apêndice 5.

3.2.7 Coeficiente de Correlação Phi

3.2.7.1 Introdução

O Coeficiente de Correlação Phi é utilizado na aplicação da Teoria de Resposta ao Item (TRI). Para determinar a dimensionalidade de uma medida, um dos índices utilizados é com base na Análise Fatorial a partir da matriz dos coeficientes de Correlação Phi. É possível encontrar um maior detalhamento sobre o assunto em NOJOSA (2001).

Este coeficiente é também utilizado na análise de confiabilidade, já apresentada na seção 3.2.1.10.1.

Em algumas situações, as variáveis são medidas em nível nominal ou por categorias discretas e expressas em forma de freqüências. Nesses casos, não é possível a utilização de nenhum dos métodos vistos anteriormente.

O Coeficiente de Correlação Phi deve ser utilizado quando ambas as variáveis correlacionadas são dicotomizadas (ao serem medidas) ou genuinamente dicotômicas. George Udny Yule publicou, em 1912, no *Journal of Royal Statistical Society*, um artigo sobre o Coeficiente de Correlação Phi. Yule acreditava que era possível definir um coeficiente sem assumir a distribuição contínua. Ele defendia que variáveis como “vacinado” e “não vacinado”, ou “morreu” e “sobreviveu", são inerentemente discretas e que mesmo o melhor coeficiente considerando distribuição normal poderia somente dizer como essas variáveis hipotéticas se correlacionariam entre si (AGRESTI, 1990).

3.2.7.2 Estimador do Coeficiente de Correlação Phi e significância

O Estimador do Coeficiente de Correlação Phi foi obtido a partir do estimador do Coeficiente Linear de Pearson, bastando fazer com que a variável X também seja dicotômica e distribuída conforme apresentada a seguir:

		Variável X		
		1	0	TOTAL
Variável Y	1	a	b	n_p
	0	c	d	n_q
TOTAL		$n_{p'}$	$n_{q'}$	n

Tem-se, da expressão (3.76), que:

$$\hat{\rho}_{pb} = \frac{\left(\overline{X}_p - \overline{X}\right)}{S_x}\sqrt{\frac{p}{q}} \tag{3.95}$$

$$\text{mas} \quad \overline{X}_p = \frac{a}{n_p} = \frac{a}{a+b} \quad \text{e} \quad \overline{X}_q = \frac{c}{n_q} = \frac{c}{c+d} \tag{3.96}$$

$$p = \frac{(a+b)}{n} \quad \text{e} \quad q = \frac{(c+d)}{n} \tag{3.97}$$

$$\overline{X} = p\overline{X}_p + q\overline{X}_q = \frac{(a+b)}{n}\frac{a}{(a+b)} + \frac{(c+d)}{n}\frac{c}{(c+d)} = \frac{(a+c)}{n} \tag{3.98}$$

$$S_x = \sqrt{\frac{n_{p'}}{n}\frac{n_{q'}}{n}} = \sqrt{\frac{(a+c)}{n}\frac{(b+d)}{n}} = \frac{1}{n}\sqrt{(a+c)(b+d)} \tag{3.99}$$

Então, substituindo as expressões (3.96), (3.97), (3.98) e (3.99) em (3.95), tem-se:

$$\hat{\phi} = \frac{\dfrac{a}{(a+b)} - \dfrac{(a+c)}{n}}{\dfrac{1}{n}\sqrt{(a+c)(b+d)}}\frac{\sqrt{(a+b)}}{\sqrt{(a+c)}} = \frac{\dfrac{na-(a+b)(a+c)}{n(a+b)}}{\dfrac{1}{n}\sqrt{(a+c)(b+d)}}\frac{\sqrt{(a+b)}}{\sqrt{(a+c)}}$$

$$\hat{\phi} = \frac{na-(a+b)(a+c)}{(a+b)\sqrt{(a+c)(b+d)}}\frac{\sqrt{(a+b)}}{\sqrt{(a+c)}}$$

$$\hat{\phi} = \frac{(ad-bc)}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \tag{3.100}$$

onde: $\hat{\phi}$ é o Coeficiente de Correlação Phi;

a,b,c,d são as freqüências da tabela 2x2;

$n = (a+b+c+d)$ é o número de observações da amostra.

O coeficiente Phi está relacionado com χ^2 para a tabela 2x2, dada pela expressão a seguir, como apresentada em FERGUSON (1981):

$$\hat{\phi} = \sqrt{\frac{\chi^2}{n}} \quad \text{ou} \quad \chi^2 = n\hat{\phi}^2 \tag{3.101}$$

Por essa razão, pode-se testar a significância de $\hat{\phi}$ calculando o valor de $\chi^2 = n\hat{\phi}^2$ e comparando com o valor de χ^2, com 1 grau de liberdade (FERGUSON, 1981).

Os valores de $\hat{\phi}$ variam entre -1 e +1. Entretanto, para BUNCHAFT e KELLNER (1999) é suficiente que **a** e **d** indiquem ou concordância ou discordância, o mesmo acontecendo com **b** e **c**.

Devido à crescente utilização do Coeficiente Phi, particularmente relacionado com intercorrelação em teste de item, tornou-se importante conhecer o valor máximo que esse coeficiente pode assumir. O valor máximo do Coeficiente de Correlação Phi pode ser calculado através de:

$$\hat{\phi}_{máx} = \sqrt{\left(\frac{p_j}{q_j}\right)\left(\frac{q_i}{p_i}\right)} \quad \text{onde } p_i \geq p_j \geq 0{,}5 \tag{3.102}$$

onde: $\hat{\phi}_{máx}$ é o valor máximo do Coeficiente de Correlação Phi;

p_i é a maior proporção marginal da tabela de contingência 2x2;

p_j é a maior proporção marginal na outra variável;

q_i e q_j são seus complementares.

Quando $p_i = p_j$ o valor máximo de $\hat{\phi}$ é igual a 1.

Quando obtiver um valor do Coeficiente de Correlação Phi negativo, este pode ser comparado com o valor de Phi mínimo, dado por:

$$\hat{\phi}_{min} = \sqrt{\left(\frac{q_i}{p_i}\right)\left(\frac{q_j}{p_j}\right)} \quad \text{onde } p_i \leq p_j \tag{3.103}$$

onde: $\hat{\phi}_{min}$ é o valor mínimo do Coeficiente de Correlação Phi;

p_i é a menor proporção marginal da tabela de contingência 2x2;

p_j é a menor proporção marginal na outra variável;

q_i e q_j são seus complementares.

Quando $p_i = p_j$ o valor mínimo de $\hat{\phi}$ é igual a -1.

3.2.7.3 O Coeficiente de Correlação Phi e a Análise de Agrupamento

A Análise de Agrupamento é uma técnica de estatística multivariada que permite agrupar unidades semelhantes com base nas distâncias ou similaridades.

Quando as unidades observacionais são agrupadas, a proximidade é normalmente indicada por algum tipo de distância. Entretanto, as variáveis são usualmente agrupadas com base nos coeficientes de correlação ou em outras medidas de avaliação.

Conforme apresentado em CHAVES NETO (2002b), quando as variáveis são binárias pode-se obter uma tabela de contingência. Para cada par de variáveis, existem n objetos categorizados, conforme se mostra a seguir:

		Variável k		
		1	0	TOTAL
Variável i	1	a	b	a+b
	0	c	d	c+d
TOTAL		a+c	b+d	n

Como uma medida de similaridade entre i e k, poderá ser tomado o coeficiente de correlação obtido através de:

$$\hat{\phi} = \frac{(ad - bc)}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

3.2.7.4 Aplicação do Coeficiente de Correlação Phi

Os dados utilizados para a aplicação a seguir foram obtidos a partir da Pesquisa Mensal de Emprego (PME) na Região Metropolitana de Curitiba (RMC).

Dentre os diversos indicadores disponibilizados pela PME, foram escolhidas duas variáveis dicotômicas, uma delas indicando a situação ocupacional das pessoas (pessoas com trabalho e pessoas desempregadas, ou seja, aquelas pessoas sem trabalho, e que efetivamente procuraram trabalho no período de referência da pesquisa) e a outra que caracteriza o gênero (masculino e feminino). A seguir, apresenta-se a tabela 2X2.

TABELA 4 - SITUAÇÃO OCUPACIONAL DA POPULAÇÃO ECONOMICAMENTE ATIVA SEGUNDO GÊNERO, NA RMC - AGOSTO 2003

SITUAÇÃO OCUPACIONAL	GÊNERO		TOTAL
	Homem	Mulher	
Ocupados	2 896	2 157	5 053
Desempregados	221	251	472
TOTAL	3 117	2 408	5 525

FONTE: PME - IPARDES/IBGE
NOTA: A tabulação dos dados foi feita pela autora.

O Coeficiente de Correlação Phi obtido foi $\hat{\phi} = 0{,}05913$ com significância $< 0{,}005$, indicando que existe correlação, embora muito pequena, ou seja, existe uma fraca tendência no sentido de que a incidência de desemprego entre as mulheres seja maior do que entre homens.

É evidente que ao calcular o Coeficiente de Correlação Linear de Pearson para as variáveis dicotômicas, obtém-se o mesmo valor, pois trata-se do mesmo coeficiente.

3.2.8 Coeficiente de Contingência

3.2.8.1 Introdução

Quando se pretende relacionar dados em nível nominal, dispostos em tabelas politômicas, utiliza-se o coeficiente de contingência C. Este não exige nenhuma suposição quanto à forma da distribuição populacional dos escores, sendo necessário, apenas, que a variável seja medida em nível nominal.

Este coeficiente não pode ser comparado a qualquer outro coeficiente de correlação, podendo-se comparar vários coeficientes de contingência quando estes forem provenientes de tabelas de mesmas dimensões.

Outra limitação de C é que os dados devem satisfazer aos requisitos para o cálculo de χ^2. Conforme descrito em SIEGEL (1975), a prova χ^2 somente pode ser utilizada adequadamente se menos de 20% das células apresentam freqüência esperada (f_e) menor que 5, sendo que nenhuma célula deve ter freqüência esperada menor que 1.

3.2.8.2 Estimador do Coeficiente de Contingência e significância

O estimador do Coeficiente de Contingência é conforme apresentado a seguir:

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}} \tag{3.104}$$

onde: C é o Coeficiente de Contingência;

χ^2 é o qui-quadrado calculado para os dados;

n é o número de elementos da amostra.

O χ^2 é calculado através de:

$$\chi^2 = \sum_{i=1}^{n} \frac{(f_{oi} - f_{ei})^2}{f_{ei}} \tag{3.105}$$

onde: f_o é a freqüência observada;

f_e é a freqüência esperada.

O valor máximo do Coeficiente de Contingência nunca atinge a unidade, mesmo que as variáveis sejam perfeitamente correlacionadas, embora seja nulo quando não há correlação.

O valor de $C_{máx}$ poderá ser calculado se o número de linhas (ℓ) for igual ao número de colunas (c), através de:

$$C_{máx} = \sqrt{\frac{\ell - 1}{\ell}} \qquad (3.106)$$

Para verificar se o valor observado de C indica existência de associação entre duas variáveis na população, utiliza-se o valor de χ^2 observado com gl=(ℓ-1)(c-1). Se χ^2 calculado para a amostra for significativo, a um certo nível de significância, pode-se concluir que a associação entre as duas variáveis é diferente de zero.

3.2.8.3 Aplicação do Coeficiente de Contingência

O trabalho discutido a seguir, intitulado *Cumplimento del régimen terapêutico y su relación con las características biológicas y sociales del indivíduo con insuficiência renal crônica terminal en hemodiálisis*, foi desenvolvido por TOBO et al. (1995).

O estudo foi realizado com amostra de 68 pessoas com insuficiência renal crônica terminal, que se submeteram a hemodiálise em 3 unidades da cidade de Cali, Colômbia, no período de setembro a outubro de 1994. O objetivo foi determinar a relação entre as características biológicas e psicológicas com o cumprimento do regime terapêutico.

A seleção da amostra foi aleatória e o tamanho determinado por meio de uma prova estatística para estudos descritivos, com 94% de confiança e 6% de margem de erro.

O estudo consistiu no cálculo de estatísticas descritivas, teste χ^2 (Qui-quadrado), Coeficiente de Correlação Phi e Coeficiente de Contigência (C).

As informações foram obtidas mediante um questionário, com três enfoques: dados de identificação, características sociais e biológicas e cumprimento do regime terapêutico.

As características sociais e biológicas contempladas foram: idade, sexo, enfermidade associada, limitação física, escolaridade, estado civil, tempo de hemodiálise, opinião sobre a doença e tratamento, condição socioeconômica e apoio familiar.

Alguns dos resultados alcançados foram: (i) a escolaridade relaciona-se significativamente com os níveis séricos de nitrogênio uréico, potássio, cálcio, fósforo e albumina, sendo o coeficiente de contingência resultante igual a $C = 0{,}32$, indicando uma correlação moderada; (ii) a associação entre o tempo de hemodiálise e o cumprimento da terapia dialítica apresentou Coeficiente de Contingência igual a $C = 0{,}35$, indicando que a correlação entre estas variáveis é moderada, ou seja, quanto menor o tempo de hemodiálise, maior o cumprimento desta terapia; (iii) a associação entre as variáveis conhecimento da doença e do tratamento e volume total de sangue teve um resultado estatisticamente significativo, apresentando coeficiente igual a $\phi = 0{,}31$.

3.2.9 Coeficiente de Correlação Eta

3.2.9.1 Introdução

O coeficiente de correlação a ser calculado quando se tem uma variável quantitativa Y e outra variável categórica ou nominal X, conforme descrito em SILVEIRA (1999), é o Coeficiente de Correlação Eta. Este resulta sempre em um valor no intervalo fechado 0 e 1.

Conforme descreve FERGUSON (1981) e CHEN e POPOVICH (2002), a Correlação Eta tem sido apresentada como a medida apropriada para descrever a relação não-linear entre duas variáveis. Se uma das variáveis - digamos, a independente - é uma variável nominal, e a outra variável é intervalar ou de razão, a idéia de linearidade ou não-linearidade praticamente não tem sentido.

Para DOWNIE e HEATH (1959), o coeficiente correto quando a relação entre dois conjuntos de dados é curvilínea é o Coeficiente Eta. Os valores de Eta e $\hat{\rho}$ devem ser idênticos, quando a relação é linear. Se a relação é curvilínea, Eta é maior que $\hat{\rho}$, e a diferença entre os dois indica o grau de distância da linearidade.

Conforme apresenta CHEN e POPOVICH (2002), o Coeficiente Eta é também um caso especial de $\hat{\rho}$. Se os valores de Y (variável nominal) forem substituídos pela média de X, correspondente a cada categoria, o resultado será equivalente ao $|\hat{\rho}|$.

3.2.9.2 Estimador do Coeficiente de Correlação Eta e significância

O estimador do Coeficiente de Correlação Eta é a raiz quadrada da expressão a seguir:

$$\eta^2_{y,x} = \frac{\text{soma de quadrados entre grupos}}{\text{soma de quadrados total}} \tag{3.107}$$

O erro padrão do quadrado do Coeficiente Eta é dado por:

$$\hat{\sigma}_{\hat{\eta}^2_{y,x}} = \frac{1-\eta^2_{y,x}}{n-k} \tag{3.108}$$

onde: $\hat{\sigma}_{\hat{\eta}^2_{y,x}}$ é o erro padrão do quadrado do Coeficiente Eta;

$\eta^2_{y,x}$ é o quadrado Coeficiente Eta;

n é o número de observações da amostra;

k é o número de categorias da variável nominal.

Na Análise da Variância (ANOVA) a um critério de classificação ou experimento de um fator são envolvidas duas variáveis, sendo que a variável independente é normalmente do tipo nominal e a dependente é medida em nível intervalar ou de razão.

Na ANOVA, a soma de quadrados total é dividida em soma de quadrados entre grupos e soma de quadrados dentro dos grupos. A soma de quadrados entre grupos é a parte da variação atribuída à variável independente, e dentro dos grupos a outros fatores.

A Correlação Eta ao quadrado é a razão entre a soma de quadrados entre grupos e a soma de quadrados total, equivalente ao $\hat{\rho}^2$ do modelo de regressão linear simples[14].

---

[14] $\hat{\rho}^2$ = variação explicada/variação total.

Para testar a significância do Coeficiente de Correlação Eta ($H_0 : \eta = 0$ e $H_1 : \eta \neq 0$), usa-se a razão F (que é exatamente a razão F da ANOVA), dada por:

$$F = \frac{\eta^2_{y,x} /(k-1)}{(1-\eta^2_{y,x})/(n-k)} \qquad (3.109)$$

onde: F é a estatística do teste;

k é o número de categorias da variável nominal;

n é o número total de observações.

3.2.9.3 O Coeficiente de Correlação Eta e a Análise de Variância

A Análise de Variância é utilizada para testar a hipótese de diferença entre duas ou mais médias. A hipótese a ser testada será $H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_k$.

É possível, através do Coeficiente de Correlação Eta, testar a hipótese de diferença entre duas ou mais médias. Neste caso, a variável nominal ou ordinal assume duas ou mais categorias. Da mesma forma que no caso anterior, a hipótese a ser testada será de $H_0 : \eta = 0$.

Rejeitando-se a hipótese de que a correlação é igual a zero, está-se aceitando a hipótese de que pelo menos uma das médias é diferente. Para identificar qual média difere das demais, pode-se utilizar, entre outros, os testes de Diferença Mínima Significante (DMS), Duncan e Scheffé, abordados em SNEDECOR e COCHRAN (1980).

3.2.9.4 Aplicação do Coeficiente de Correlação Eta

A aplicação apresentada a seguir refere-se ao trabalho realizado por SILVEIRA (1999), em que se utilizou o Coeficiente de Correlação Eta para estudar a relação entre o desempenho na prova de Biologia do Concurso Vestibular de 1999

da Universidade Federal do Rio Grande do Sul, e o tipo de ensino médio cursado pelos candidatos.

O desempenho na prova de Biologia, de um total de 35.463 candidatos, foi correlacionado com o tipo de ensino médio que cursaram, a saber: não-profissionalizante, profissionalizante, magistério, militar e supletivo. A prova de Biologia era composta de 30 itens de múltipla escolha, com resposta única.

O Coeficiente de Correlação Eta obtido foi $\eta = 0{,}27$, o que indica a existência de alguma relação entre as variáveis. Foi possível também observar, através dos resultados, que os candidatos que cursaram o ensino médio em escolas militares apresentaram, em média, melhor desempenho, com média em torno de 15 acertos. Por outro lado, os que cursaram o supletivo tiveram o pior desempenho, com média em torno de 9 acertos.

3.2.10 Resumo dos Coeficientes de Correlação entre Duas Variáveis

Apresenta-se, no quadro 5, o resumo dos diferentes métodos para obtenção do coeficiente de correlação entre duas variáveis.

QUADRO 5 - RESUMO DOS COEFICIENTES DE CORRELAÇÃO ENTRE DUAS VARIÁVEIS

COEFICIENTE	SÍMBOLO	INTERVALO DE VARIAÇÃO	VARIÁVEIS	
			X	Y
Pearson	ρ	$-1 \le \rho \le 1$	Contínua	Contínua
Ponto Bisserial	ρ_{pb}	$-1 \le \rho_{pb} \le 1$	Contínua	Dicotômica
Bisserial	ρ_{b}	$-1 \le \rho_{b} \le 1$	Contínua	Contínua, mas dicotomizada
Tetracórico	ρ_{t}	$-1 \le \rho_{t} \le 1$	Contínua, mas dicotomizada	Contínua, mas dicotomizada
Phi	ϕ	$-1 \le \phi \le 1$	Dicotômica	Dicotômica
Spearman	ρ_{s}	$-1 \le \rho_{s} \le 1$	Dados em *ranks* ou passíveis de serem transformados	Dados em *ranks* ou passíveis de serem transformados
Kendall	τ	$-1 \le \tau \le 1$	Dados em *ranks*	Dados em *ranks*
Contingência	C	$0 \le C < 1$	Nominal	Nominal
Eta	η	$0 \le \eta \le 1$	Contínua	Contínua ou discreta; pode assumir valores nominais ou outros tipos de valores

FONTE: DOWNIE e HEATH (1959)

3.3 MEDIDAS DE CORRELAÇÃO ENTRE DIVERSAS VARIÁVEIS

3.3.1 Matriz de Correlações

Quando se tem $p > 2$ variáveis, e o interesse é conhecer as correlações existentes entre as variáveis, duas a duas, ou seja, X_i com X_j, $i \neq j$. A partir de coeficientes simples obtém-se a matriz de correlações, representada da seguinte forma:

$$\rho = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \ldots & \rho_{1p} \\ \rho_{21} & 1 & \rho_{23} & \ldots & \rho_{2p} \\ \rho_{31} & \rho_{32} & 1 & \ldots & \rho_{3p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \rho_{n1} & \rho_{n2} & \rho_{n3} & \ldots & 1 \end{bmatrix} \quad \text{matriz de correlação populacional} \qquad (3.110)$$

A matriz ρ é um parâmetro populacional estimado por:

$$\hat{\rho} = \begin{bmatrix} 1 & \hat{\rho}_{12} & \hat{\rho}_{13} & \ldots & \hat{\rho}_{1p} \\ \hat{\rho}_{21} & 1 & \hat{\rho}_{23} & \ldots & \hat{\rho}_{2p} \\ \hat{\rho}_{31} & \hat{\rho}_{32} & 1 & \ldots & \hat{\rho}_{3p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \hat{\rho}_{n1} & \hat{\rho}_{n2} & \hat{\rho}_{n3} & \ldots & 1 \end{bmatrix} \quad \text{matriz de correlação amostral} \qquad (3.111)$$

Uma das principais aplicações da matriz de correlação está na análise da estrutura de variância-covariância de um vetor aleatório $\underline{X}$.

3.3.1.1 Análise de Componentes Principais

3.3.1.1.1 Introdução

Uma das importantes aplicações no estudo da Análise de Covariância e Correlação está a Análise de Componentes Principais. Como se sabe, a matriz Σ (covariância) ou ρ (correlação) resume a estrutura de associação entre as p variáveis de um vetor aleatório $\underline{X}$.

A partir de Σ ou de ρ inicia-se o procedimento da Análise de Componentes Principais, conforme descrita a seguir.

Seja o vetor aleatório $\underline{X}' = [X_1, X_2, ..., X_p]$, que tem a matriz de covariância Σ, com autovalores $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. Considerando-se as seguintes combinações lineares:

$$
\begin{aligned}
Y_1 &= \underline{e}'_1 \underline{X} = e_{11}X_1 + e_{21}X_2 + ... + e_{p1}X_p \\
Y_2 &= \underline{e}'_2 \underline{X} = e_{12}X_1 + e_{22}X_2 + ... + e_{p2}X_p \\
&... \quad ... \quad ... \quad + \quad ... \quad + .. . + ... \\
Y_p &= \underline{e}'_p \underline{X} = e_{1p}X_1 + e_{2p}X_2 + ... + e_{pp}X_p
\end{aligned}
\qquad (3.112)
$$

com $i = 1, 2, ..., p$

As componentes principais são as combinações lineares $Y_1, Y_2, ..., Y_p$, não correlacionadas, cujas variâncias são tão grandes quanto possível.

A primeira componente principal é a combinação linear $\underline{\ell}'_1 \underline{X}$, que maximiza $Var(\underline{\ell}'_1 \underline{X})$, sujeito a $\underline{\ell}'_1 \underline{\ell}_1 = 1$. A segunda componente é a combinação linear $\underline{\ell}'_2 \underline{X}$, que maximiza $Var(\underline{\ell}'_2 \underline{X})$, sujeito a $\underline{\ell}'_2 \underline{\ell}_2 = 1$ e $COV(\underline{\ell}'_1 \underline{X}, \underline{\ell}'_2 \underline{X}) = 0$, e assim até a i-ésima componente principal.

Então, conforme descrito em JOHNSTON e WICHERN (1988), tem-se:

$$Var(Y_i) = \underline{e}'_i \Sigma \underline{e}_i = \lambda_i \qquad i = 1, 2, ..., p \qquad (3.113)$$

$$COV(Y_i, Y_k) = \underline{e}'_i \Sigma \underline{e}_k = 0 \qquad i \neq k = 1, 2, ..., p \qquad (3.114)$$

Os pares $(\lambda_1, \underline{e}_1), (\lambda_2, \underline{e}_2), ..., (\lambda_p, \underline{e}_p)$, com $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$, são os pares de autovalores e autovetores de Σ. É possível calcular os coeficientes de correlação entre as componentes Y_i e as variáveis X_k, através de:

$$\rho(Y_i, X_k) = \frac{COV(Y_i, X_k)}{\sqrt{Var(Y_i) Var(X_k)}} = \frac{\lambda_i e_{ki}}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}} = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \qquad i, k = 1, 2, ..., p \qquad (3.115)$$

As componentes principais podem ser obtidas, ainda, a partir da matriz de correlação ρ, obtendo-se os autovalores e autovetores e as componentes, exatamente

da mesma forma como foi descrita acima, apenas substituindo Σ por ρ e , quando se trata de amostra, substituindo por S e $\hat{\rho}$.

Os escores obtidos para cada unidade observacional e para cada cada uma das componentes principais podem ser utilizados como dados de entrada, ou seja, variáveis independentes, para a análise de regressão múltipla.

3.3.1.1.2 Aplicação da Análise de Componentes Principais

A aplicação apresentada a seguir refere-se ao trabalho desenvolvido por ALMEIDA FILHO (2001), cujo objetivo foi estudar a possibilidade de a microexsudação de hidrocarbonetos ter criado alterações mineralógicas e/ou associações rocha-solo-vegetação, na região localizada no Estado da Bahia, município de Macururé, região de influência da cidade de Paulo Afonso. Para viabilizar o estudo, foram analisadas imagens do Landsat Thematic Mapper (TM).

Foram utilizadas imagens multiespaciais do dia 7 de outubro de 1987. A seleção de conjunto de imagens mais antigas (sem cobertura de nuvens), conforme descreve o autor, visou minimizar possíveis influências de atividade antrópica na cobertura vegetal.

Segundo o autor, uma característica marcante das imagens multiespaciais é que as informações referentes a distintas bandas são muitas vezes redundantes, fazendo com que as correlações entre elas sejam elevadas. O quadro 6 apresenta os coeficientes de correlações entre as bandas.

QUADRO 6 - MATRIZ DE CORRELAÇÃO ENTRE AS BANDAS LANDSAT-TM EM MACURURÉ - OUTUBRO 1987

BANDAS	TM1	TM2	TM3	TM4	TM5	TM7
TM1	1,000	0,902	0,840	0,715	0,689	0,728
TM2	0,902	1,000	0,946	0,851	0,764	0,812
TM3	0,840	0,946	1,000	0,841	0,863	0,898
TM4	0,715	0,851	0,841	1,000	0,711	0,715
TM5	0,689	0,764	0,863	0,711	1,000	0,959
TM7	0,728	0,812	0,898	0,715	0,959	1,000

FONTE: ALMEIDA FILHO (2001)

Utilizou-se a técnica de Análise de Componentes Principais para evitar as correlações entre as bandas e separar as informações que são específicas de cada banda espectral. Os autovalores e autovetores obtidos a partir da matriz de covariância encontram-se no quadro 7.

QUADRO 7 - AUTOVALORES E AUTOVETORES SEGUNDO COMPONENTES PRINCIPAIS

COMPONENTE PRINCIPAL	AUTOVALORES		AUTOVETORES					
	Abs.	%	TM1	TM2	TM3	TM4	TM5	TM7
1	2 683	68	0,083	0,187	0,332	0,302	0,640	0,595
2	581	13	0,116	0,304	0,300	0,771	-0,383	-0,255
3	491	11	0,199	0,344	0,430	-0,454	-0,522	0,423
4	223	5	0,249	0,311	0,412	-0,324	0,409	-0,631
5	89	2	0,610	0,458	-0,642	0,019	0,042	0,065
6	44	1	0,711	0,699	0,199	0,063	0,045	0,019

FONTE: ALMEIDA FILHO (2001)

A análise se concentrou nas componentes 3, 4 e 5, que embora tenham totalizado apenas 18% da variação dos dados, segundo o autor, as informações espectrais contidas nessas componentes, isentas das contribuições de albedo (componente 1) e da cobertura vegetal (componente 2), estão relacionadas ao comportamento espectral de feições do terreno.

Conforme analisa o autor, a componente 3 pode ser entendida como expressando respostas de solo, enquanto a componente 4 sugere influência de resposta espectral de argilas, podendo também estar sendo influenciada pelo material barbonático.

A componente 5 é denominada pelas bandas do visível, podendo-se inferir, como afirma o autor, a contribuição de material limonítico.

3.3.1.2 Análise Fatorial

3.3.1.2.1 Introdução

Uma aplicação importante da Análise de Covariância e Correlação está na técnica conhecida como Análise Fatorial. Esta técnica parte da matriz de covariância Σ ou de correlação ρ, que resume a estrutura de relacionamento entre as variáveis.

Então, da matriz de dados X de ordem n x p, onde n é o número de observações e p o número de variáveis, obtém-se a matriz de covariância Σ ou de correlação ρ de ordem pxp. A partir daí inicia-se a técnica de Análise Fatorial, descrita a seguir.

O objetivo principal da análise fatorial é descrever a estrutura de covariância dos relacionamentos do conjunto com p variáveis através de variáveis não observáveis chamadas fatores.

Supondo que as variáveis possam ser agrupadas por suas correlações, tem-se grupos cujas variáveis são altamente correlacionadas entre si, porém com correlações relativamente baixas com as variáveis de outros grupos, podendo-se, assim, dizer que cada grupo de variáveis representa um fator.

Seja $\underline{X}$ o vetor aleatório com *p* componentes, então $\underline{X} \sim \bullet(\underline{\mu}, \Sigma)$. O modelo fatorial postula, segundo CHAVES NETO (2002), que $\underline{X}$ é linearmente dependente de algumas variáveis aleatórias não observáveis $F_1, F_2, ..., F_m$, chamadas fatores comuns e *p* fontes de variação aditivas, $\varepsilon_1, \varepsilon_2, ..., \varepsilon_p$, chamadas erros, ou fatores específicos. Assim, tem-se o modelo na forma matricial, como segue:

$$\underline{X} - \underline{\mu} = L\underline{F} + \underline{\varepsilon} \tag{3.116}$$

onde L é a matriz de carregamentos dos fatores, e o elemento da i-ésima linha e j-ésima coluna, coeficiente ℓ_{ij}, é chamado de carregamento do j-ésimo fator na i-ésima variável.

Os desvios $X_1 - \mu_1, X_2 - \mu_2, ..., X_p - \mu_p$ são expressos em termos de $p + m$ variáveis aleatórias: $F_1, F_2, ..., F_m, \varepsilon_1, \varepsilon_2, ..., \varepsilon_p$, que não são observáveis. A diferença entre este modelo e o de regressão múltipla está, justamente, no fato de que as variáveis independentes (F_i) $i = 1, 2, ..., n$ não são observáveis.

Assumem-se algumas suposições:

$$E(\underline{F}) = \underline{0} \tag{3.117}$$

$$COV(\underline{F}) = E(\underline{F}\underline{F}') = I \quad \text{(matriz identidade)} \tag{3.118}$$

$$E(\underline{\varepsilon}) = \underline{0} \tag{3.119}$$

$$COV(\underline{\varepsilon}) = E(\underline{\varepsilon}\underline{\varepsilon}') = \psi \quad \text{(matriz das variâncias específicas)} \tag{3.120}$$

$$COV(\underline{\varepsilon},\underline{F}) = E(\underline{\varepsilon}'\underline{F}) = 0 \quad \text{com} \quad m = p \tag{3.121}$$

Então, o modelo $\underline{X} - \underline{\mu} = L\underline{F} + \underline{\varepsilon}$ é chamado modelo fatorial ortogonal e pode ser escrito como: $\underline{X} = \underline{\mu} + L\underline{F} + \underline{\varepsilon}$.

A parte da variância da i-ésima variável aleatória X, devida à contribuição dos m fatores comuns, é chamada de comunalidade, e a parte devida ao fator específico é chamada de variância específica. Assim, tem-se:

$$V(X_i) = V(\mu_i + \ell_{i1}F_1 + \ell_{i2}F_2 + ... + \ell_{im}F_m + \varepsilon_i) \tag{3.122}$$

$$V(X_i) = \ell_{i1}^2 + \ell_{i2}^2 + ... + \ell_{im}^2 + \psi_i \tag{3.123}$$

Fazendo $h_i^2 = \ell_{i1}^2 + \ell_{i2}^2 + ... + \ell_{im}^2$, tem-se $V(Xi) = h_i^2 + \psi_i$, $i = 1,2,...,p$, onde h_i^2 é a comunalidade e ψ_i a variância específica.

O modelo fatorial ortogonal procura representar de forma adequada o conjunto de dados, através de um número menor de fatores. A matriz de covariância S é um estimador da matriz populacional Σ desconhecida e é, geralmente, usada, pois não se conhece o parâmetro Σ, ou, ainda, o estimador de ρ, $\hat{\rho}$, é usado quando se faz a análise a partir da matriz de correlação.

Conforme apresentado em JOHNSTON e WICHERN (1988), se os elementos fora diagonal de S (matriz de covariância amostral) são baixos, ou na matriz de correlação amostral $\hat{\rho}$ são praticamente nulos, as variáveis não são relacionadas e a análise fatorial não é útil. Contudo, se S é significativamente diferente de uma matriz diagonal, então é possível utilizar o modelo fatorial. Para tal deve-se estimar os carregamentos ℓ_{ij} e as variâncias específicas ψ_i. A estimação poderá ser feita pelo método das Componentes Principais, que é o preferido, ou pelo método da Máxima Verossimilhança.

A matriz de carregamentos estimados $\hat{\ell}_{ij}$ é dada por:

$$\hat{L} = \left[\sqrt{\hat{\lambda}_1}\,\underline{\hat{e}}_1, \sqrt{\hat{\lambda}_2}\,\underline{\hat{e}}_2, ..., \sqrt{\hat{\lambda}_m}\,\underline{\hat{e}}_m\right] \tag{3.124}$$

onde $\hat{\lambda}_1, \hat{\lambda}_2, ..., \hat{\lambda}_m$ são os autovalores de S e $\underline{\hat{e}}_1, \underline{\hat{e}}_2, \ldots, \underline{\hat{e}}_m$, os autovetores e m o número de fatores, quando se faz a análise a partir de S ou, então, tem-se estimativas equivalentes quando a análise é a partir de $\hat{\rho}$.

As variâncias específicas são estimadas por:

$$\hat{\psi} = \begin{bmatrix} \hat{\psi}_1 & 0 & ... & 0 \\ 0 & \hat{\psi}_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \hat{\psi}_p \end{bmatrix} \text{ com } \hat{\psi}_i = S_{ii} - \sum_{j=1}^{m} \hat{\ell}_{ij}^2 \text{ ou } \hat{\psi}_i = 1 - \sum_{j=1}^{m} \hat{\ell}_{ij}^2 \qquad (3.125)$$

Na Análise Fatorial, a interpretação dos fatores será facilitada pela rotação dos mesmos. A rotação poderá ser ortogonal (com independência dos fatores extraídos) ou oblíquos (os fatores são correlacionados). Na rotação ortogonal, os métodos mais utilizados são o quartimax e o varimax. O primeiro método procura maximizar a carga fatorial de uma variável com um fator e minimizar com os outros fatores. Já o segundo método busca, inversamente, simplificar as colunas da matriz de cargas fatoriais, isto é, procura definir mais claramente quais variáveis estão associadas com um determinado fator e quais não estão. Mais detalhes poderão ser obtidos em FACHEL (1976) e JOHNSON e WICHERN (1988).

Em muitas aplicações, os valores estimados dos fatores comuns, denominados escores fatoriais, são importantes e se necessita obter.

Os escores fatoriais são estimativas dos valores para os vetores fatoriais aleatórios não observáveis $\underline{F}_j$, $j = 1, 2, ..., m$. Uma técnica bastante utilizada na estimação é o método dos mínimos quadrados ponderados, desenvolvido por Bartlett, embora existam outras, tais como a da Regressão e a da Regressão para Fatores Correlacionados.

O estimador dos escores fatorais pelo método de Mínimos Quadrados Ponderados para o j-ésimo fator é dado por:

$$\hat{\underline{f}}_j = \left[\hat{L}'\hat{\psi}^{-1}\hat{L}\right]^{-1}\hat{L}'\hat{\psi}^{-1}\left(\underline{x}_j - \overline{\underline{x}}\right) \qquad (3.126)$$

A Análise Fatorial pode ser utilizada a partir da matriz de covariância Σ, ou de correlação ρ, como já se citou. Se o estudo for realizado a partir dos dados amostrais, tem-se a matriz de variância e de correlação amostral S e $\hat{\rho}$, respectivamente.

Em FACHEL (1986) são apresentadas as diferentes matrizes de correlação que são utilizadas como dados de entrada (*input*), na análise fatorial, estando, dentre elas, as matrizes formadas pelos coeficientes tetracórico e Phi. Faz-se uma comparação entre diferentes métodos de Análise Fatorial e a Análise Fatorial de Bartholomew, para dados categóricos.

3.3.1.2.2 Aplicação da Análise Fatorial

A aplicação apresentada a seguir refere-se ao trabalho realizado por FURTADO (1999), cujo objetivo foi fazer um ranqueamento (hierarquização) de áreas especialmente protegidas, chamadas de faxinais do Estado do Paraná, considerando as variáveis avaliadas pelo Instituto Ambiental do Paraná (IAP).

Os faxinais são entendidos, de acordo com o Decreto Estadual n^{o}. 3.446/97, conforme descreve o autor, como um sistema tradicional, característico da região Centro-Sul do Paraná, que tem como característica marcante o uso coletivo da terra para a produção animal e a conservação ambiental.

As informações levantadas neste trabalho referem-se ao ano agrícola de agosto de 1997 a julho de 1998. Os questionários foram aplicados às famílias e lideranças locais. O universo de famílias foi de 1.947 e foram aplicados os questionários em uma amostra de 316 famílias, tendo sido adotada uma precisão da estimativa de 5,5%, considerando-se um nível de confiança de 95%.

Os dados foram coletados através da aplicação de questionários às famílias selecionadas para compor a amostra, que se distribuem em 20 faxinais pertencentes a 4 municípios. Os faxinais estão localizados conforme mostra o quadro 8, apresentado a seguir:

QUADRO 8 - NÚMERO DE FAXINAIS, SEGUNDO MUNICÍPIOS DA REGIÃO CENTRO-SUL DO PARANÁ - AGOSTO 1997-JULHO 1998

MUNICÍPIO	NÚMERO DE FAXINAIS
Prudentópolis	14
Rebouças	3
Irati	2
Boa Ventura de São Roque	1

FONTE: FURTADO (1999)

Criou-se uma matriz composta de 20 linhas e 80 colunas, em que 20 é o número de faxinais e 80 o número de variáveis. As descrições das variáveis poderão ser encontradas detalhadamente em FURTADO (1999) e FURTADO e CHAVES NETO (2003).

Estimou-se a matriz de correlação das variáveis e, em seguida, os pares de autovalores e autovetores dessa matriz. Considerando os autovalores superiores a 1, foram escolhidos 17 fatores, que correspondem a um grau de explicação de 97,764%.

A matriz de carregamentos fatoriais foi obtida a partir de autovalores e autovetores associados. Utilizou-se o método varimax normal para a obtenção da matriz de carregamentos fatoriais rotacionados e, após, foram determinadas as comunalidades e as variâncias específicas de cada variável.

Os escores fatoriais foram estimados pelo método de mínimos quadrados ponderados. Os escores de cada Faxinal foram obtidos ponderando-os pela importância de cada fator, ou seja, pelo autovalor.

O quadro a seguir apresenta o ranqueamento dos faxinais estudados. Os escores brutos foram obtidos através de média aritmética dos 17 escores fatoriais, ponderada pelos autovalores da matriz de correlação. Na seqüência, os escores foram colocados na escala entre 0 e 2.

QUADRO 9 - RANQUEAMENTO DOS FAXINAIS DA REGIÃO CENTRO-SUL DO PARANÁ - AGOSTO 1997-JULHO 1998

FAXINAL	ESCORES BRUTOS	ESCORES PADRONIZADOS
1º São Pedro	406,0385	1,8000
2º Ivaí - Anta Gorda	255,5732	1,4586
3º Ponte Nova	163,4198	1,2495
4º Linha Brasília	150,4757	1,2201
5º Patos Velhos	116,3825	1,1427
6º Papanduva de Baixo	92,0459	1,0875
7º Queimadas	71,1442	1,0401
8º Cachoeira do Palmital	62,9660	1,0215
9º Rio dos Couros	2,5779	0,8845
10º Rio do Meio	-9,2271	0,8577
11º Tijuco Preto	-11,4835	0,8526
12º Paraná - Anta Gorda	-19,0331	0,8355
13º Guanabara	-71,2905	0,7169
14º Salto	-101,9150	0,6474
15º Taboãozinho	-105,5850	0,6391
16º Dos Mellos	-113,3440	0,6215
17º Marmeleiro de Baixo	-123,3310	0,5988
18º Dos Krieger	-202,2800	0,4197
19º Marmeleiro de Cima	-264,0430	0,2795
20º Rio Bonito	-299,0920	0,2000

FONTE: FURTADO (1999)

3.3.2 Coeficiente de Correlação Múltipla e Parcial

3.3.2.1 Introdução

O Coeficiente de Correlação Múltipla indica o grau de relacionamento entre as variáveis independentes representado pelo vetor $\underline{X}$, onde $\underline{X} = [X_1, X_2, X_3, ..., X_p]$ e a variável dependente (Y).

Os princípios gerais do método para a Correlação Múltipla constituem apenas uma extensão direta dos conceitos e raciocínios apresentados para o Coeficiente Linear de Pearson.

A Correlação Múltipla não é simplesmente a soma de correlações da variável dependente com as independentes tomadas separadamente (GUILFORD, 1950). Uma das razões é que as variáveis independentes são normalmente intercorrelacionadas, conhecidas também como multicolineares. Quando as intercorrelações forem iguais a zero, então o quadrado do coeficiente de correlação múltipla será a soma dos quadrados dos coeficientes de cada variável independente com a dependente.

A Correlação Múltipla aumenta quando aumenta o tamanho da correlação entre as variáveis dependentes e independentes e quando o tamanho das intercorrelações entre as variáveis independentes diminui (GUILFORD, 1950).

Da mesma forma que a análise de correlação simples e a regressão simples estão ligadas, a correlação e regressão múltipla também estão.

A análise de regressão múltipla é tratada através do modelo linear geral:

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \tag{3.127}$$

onde: $\underline{Y}$ é o vetor das observações (respostas) de dimensão n;

X a matriz de dados de ordem $n \times p$;

$\underline{\beta}$ vetor dos parâmetros de dimensão p;

$\underline{\varepsilon}$ vetor dos erros de dimensão n.

É comum que algumas ou todas as variáveis explicativas (independentes) estejam correlacionadas umas com as outras, o que dificulta isolar suas influências separadamente e obter uma estimativa razoavelmente precisa de seus efeitos relativos.

Uma das formas de resolver a multicolinearidade é através da utilização de componentes principais (NETER et al., 1996), uma vez que as componentes principais são combinações lineares independentes. Mais uma vez necessita-se da matriz de correlação, agora das variáveis explicativas.

A análise de componentes principais procura, segundo CHAVES NETO (2002b), explicar a estrutura de variância-covariância da matriz de dados a partir de combinações lineares não correlacionadas das *p* variáveis originais. Freqüentemente, a maior parte da variabilidade do conjunto de variáveis pode ser explicada por um número menor, *k*, de componentes principais. As *k* componentes principais contêm quase a mesma quantidade de informações que as p variáveis originais. É possível, assim, utilizar as componentes principais em substituição das variáveis originais.

3.3.2.2 Suposições para a utilização do Coeficiente de Correlação Múltipla

A primeira suposição para a utilização da Correlação Múltipla é que as variáveis sejam aleatórias. Como segunda suposição, deve-se considerar que as relações entre as variáveis sejam lineares e, finalmente, as variâncias sejam iguais (homocedasticidade) e as distribuições condicionais todas normais.

Uma vez que existe relação entre a análise de correlação múltipla e regressão múltipla, é possível, através da segunda, obter-se o coeficiente de correlação múltipla. Neste caso, as suposições sobre o erro do modelo de regressão múltipla devem ser consideradas. As suposições usuais sobre a componente ε são as seguintes:

(i) $E(\varepsilon_i) = 0 \quad , i = 1, 2, ..., n$ (3.128)

(ii) $VAR(\varepsilon_i) = \sigma^2 \quad , i = 1, 2, ..., n$ (3.129)

(iii) $COV(\varepsilon_i, \varepsilon_j) = 0 \quad , i, j = 1, 2, ..., n, i \neq j$ (3.130)

Conforme descrito em SIQUEIRA (1983), para fazer inferências estatísticas (teste de hipóteses e estimação por intervalos) é necessário atender à suposição de que:

(iv) $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, 2, ..., n$ (3.131)

Quando as suposições não são atendidas, é possível fazer alguma transformação nas variáveis, conforme já apresentado na seção 3.2.1.2.

3.3.2.3 Estimador do Coeficiente de Correlação Múltipla

Seja Y a variável dependente e X_1 e X_2 as independentes. O modelo de regressão linear poderá ser escrito sob a forma:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i, \quad i = 1,2,3,...,n \tag{3.132}$$

A estimativa do modelo poderá ser escrita na forma:

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \hat{\varepsilon}_i \tag{3.133}$$

onde tem-se que $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$, que representa o erro.

A soma de quadrados dos erros é representada pela expressão a seguir:

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i}\right)^2 \tag{3.134}$$

Derivando-se parcialmente a expressão acima em relação a b_0 e igualando-se a zero, tem-se:

$$2\sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i}\right) = 0$$

$$\sum_{i=1}^{n} Y_i - nb_0 - b_1 \sum_{i=1}^{n} X_{1i} - b_2 \sum_{i=1}^{n} X_{2i} = 0$$

$$\sum_{i=1}^{n} Y_i = nb_0 + b_1 \sum_{i=1}^{n} X_{1i} + b_2 \sum_{i=1}^{n} X_{2i}$$

Dividindo por n tem-se:

$$\overline{Y} = b_0 + b_1\overline{X}_1 + b_2\overline{X}_2 \tag{3.135}$$

Subtraindo (3.135) de (3.133) tem-se:

$$\hat{y}_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$$

Escrevendo a soma de quadrados dos erros na forma:

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}\right)^2$$

Derivando-se parcialmente em relação a $\hat{\beta}_1$ e $\hat{\beta}_2$ tem-se as duas equações normais:

$$\frac{\partial \sum_{i=1}^{n} \varepsilon_i^2}{\partial \hat{\beta}_1} = 2\left[-\sum_{i=1}^{n} x_{1i} y_i + \hat{\beta}_1 \sum_{i=1}^{n} x_{1i}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{1i} x_{2i}\right]$$

$$\frac{\partial \sum_{i=1}^{n} \varepsilon_i^2}{\partial \hat{\beta}_2} = 2\left[-\sum_{i=1}^{n} x_{2i} y_i + \hat{\beta}_1 \sum_{i=1}^{n} x_{1i} x_{2i} + \hat{\beta}_2 \sum_{i=1}^{n} x_{2i}^2\right]$$

Logo:

$$\sum_{i=1}^{n} x_{1i} y_i = \hat{\beta}_1 \sum_{i=1}^{n} x_{1i}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{1i} x_{2i}$$

$$\sum_{i=1}^{n} x_{2i} y_i = \hat{\beta}_1 \sum_{i=1}^{n} x_{1i} x_{2i} + \hat{\beta}_2 \sum_{i=1}^{n} x_{2i}^2$$

Resolvendo as equações tem-se:

$$\hat{\beta}_1 = \frac{\left(\sum_{i=1}^{n} y_i x_{1i}\right)\left(\sum_{i=1}^{n} x_{2i}^2\right) - \left(\sum_{i=1}^{n} y_i x_{2i}\right)\left(\sum_{i=1}^{n} x_{1i} x_{2i}\right)}{\left(\sum_{i=1}^{n} x_{1i}^2\right)\left(\sum_{i=1}^{n} x_{2i}^2\right) - \left(\sum_{i=1}^{n} x_{1i} x_{2i}\right)^2}$$

$$\hat{\beta}_1 = \frac{\hat{\rho}_{Y,X_1} \hat{\rho}_{Y,X_2} \hat{\rho}_{X_1,X_2}}{1 - \left(\hat{\rho}_{X_1,X_2}\right)^2} \frac{S_Y}{S_{X_1}} \tag{3.136}$$

$$\hat{\beta}_2 = \frac{\left(\sum_{i=1}^{n} y_i x_{2i}\right)\left(\sum_{i=1}^{n} x_{1i}^2\right) - \left(\sum_{i=1}^{n} y_i x_{1i}\right)\left(\sum_{i=1}^{n} x_{1i} x_{2i}\right)}{\left(\sum_{i=1}^{n} x_{1i}^2\right)\left(\sum_{i=1}^{n} x_{2i}^2\right) - \left(\sum_{i=1}^{n} x_{1i} x_{2i}\right)^2}$$

$$\hat{\beta}_2 = \frac{\hat{\rho}_{Y,X_2}\hat{\rho}_{Y,X_1}\hat{\rho}_{X_1,X_2}}{1-\left(\hat{\rho}_{X_1,X_2}\right)^2}\frac{S_Y}{S_{X_2}} \qquad (3.137)$$

A variância do erro é dada por:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

Mas $\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i\left(y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}\right)$

Então tem-se que $(n-1)S^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i\left(y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}\right)$

Logo $(n-1)S^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i y_i - \hat{\beta}_1\sum_{i=1}^{n}\hat{\varepsilon}_i x_{1i} - \hat{\beta}_2\sum_{i=1}^{n}\hat{\varepsilon}_i x_{2i}$,

mas $\sum_{i=1}^{n}\hat{\varepsilon}_i x_{1i} = \sum_{i=1}^{n}\hat{\varepsilon}_i x_{2i} = 0$, então

$$\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i y_i$$

$$\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \sum_{i=1}^{n} y_i\left(y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}\right)$$

$$\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \sum_{i=1}^{n} y_i^2 - \hat{\beta}_1\sum_{i=1}^{n} y_i x_{1i} - \hat{\beta}_2\sum_{i=1}^{n} y_i x_{2i}$$

Tem-se que: $\hat{\rho}^2 = 1 - \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{\sum_{i=1}^{n} y_i^2} = 1 - \frac{\sum_{i=1}^{n} y_i^2 - \hat{\beta}_1\sum_{i=1}^{n} y_i x_{1i} - \hat{\beta}_2\sum_{i=1}^{n} y_i x_{2i}}{\sum_{i=1}^{n} y_i^2}$

e, $\hat{\rho}^2 = \frac{\hat{\beta}_1\sum_{i=1}^{n} y_i x_{1i} + \hat{\beta}_2\sum_{i=1}^{n} y_i x_{2i}}{\sum_{i=1}^{n} y_i^2}$ (3.138)

Substituindo (3.136) e (3.137) em (3.138) tem-se:

$$\hat{\rho}^2_{Y,X_1,X_2} = \frac{\hat{\rho}^2_{Y,X_1} + \hat{\rho}^2_{Y,X_2} - 2\hat{\rho}_{Y,X_1}\hat{\rho}_{Y,X_2}\hat{\rho}_{X_1,X_2}}{1-\left(\hat{\rho}_{X_1,X_2}\right)^2} \tag{3.139}$$

Portanto, o estimador do Coeficiente de Correlação Múltipla entre três variáveis é obtido através de:

$$\hat{\rho}_{Y,X_1,X_2} = \sqrt{\frac{\hat{\rho}^2_{X_1,Y} + \hat{\rho}^2_{X_2,Y} - 2\hat{\rho}_{X_1,Y}\hat{\rho}_{X_2,Y}\hat{\rho}_{X_1,X_2}}{1-\hat{\rho}^2_{X_1,X_2}}} \tag{3.140}$$

Ou ainda, através da raiz quadrada do coeficiente de determinação ou explicação, dada por:

$$\hat{\rho} = \sqrt{\frac{SQ_{Regr}}{SQ_{Total}}} \tag{3.141}$$

Quando se tratar de amostras pequenas, deve-se fazer a seguinte correção (BUNCHAFT e KELLNER, 1999):

$$\hat{\rho}^2_c = 1-\left(1-\hat{\rho}^2\right)\frac{(n-1)}{(n-m)}$$

logo:

$$\hat{\rho}_c = \sqrt{1-\left(1-\hat{\rho}^2\right)\frac{(n-1)}{(n-m)}} \tag{3.142}$$

onde: $\hat{\rho}_c$ é o coeficiente de correlação corrigido;

$\hat{\rho}$ é o coeficiente de correlação;

n é o tamanho da amostra (número de observações da amostra);

m é o número de variáveis correlacionadas.

Quanto maior a amostra e menor o número de variáveis, menor será a diferença entre os coeficientes. Uma amostra pequena, bem como um número grande de variáveis, levam ao aumento do coeficiente de correlação.

A significância do Coeficiente de Correlação Múltipla é calculada através da razão F:

$$F = \frac{\hat{\rho}^2 / k}{(1-\hat{\rho}^2)/(n-k-1)} \tag{3.143}$$

onde: $\hat{\rho}^2$ é o coeficiente de determinação;

n é o tamanho da amostra (número de observações da amostra);

k é o número de variáveis independentes.

Já o Coeficiente de Correlação Parcial é usado quando se deseja conhecer a correlação entre duas variáveis quaisquer, quando os efeitos das outras variáveis forem controlados, ou seja, desconsiderados. Para representar a correlação amostral entre as variáveis X_1 e X_2, controlando X_3, utiliza-se a correlação parcial com notação $\hat{\rho}_{12,3}$. Esta notação pode se estender a qualquer número de variáveis controladas, acrescentando-se, à direita da vírgula, as outras variáveis.

Para calcular o coeficiente $\hat{\rho}_{12,3}$, elimina-se a influência linear de X_3 de X_1 e de X_2. Sejam as regressões lineares entre X_1 e X_3 e X_2 e X_3 dadas por:

$$X_{1i} = a_{13} + b_{13}X_{3i} + u_i \tag{3.144}$$

$$X_{2i} = a_{23} + b_{23}X_{3i} + u_i \tag{3.145}$$

Escrevendo-as nas formas de desvios, tem-se:

$$x_{1i} = \hat{\beta}_{13}x_{3i} + u_i$$

$$x_{2i} = \hat{\beta}_{23}x_{3i} + u_i$$

Os resíduos não explicados de X_1 e X_2 das regressões são dados por:

$$u_i = X_{1i} - a_{13} - b_{13}X_{3i} = x_{1i} - \hat{\beta}_{13}x_{3i}$$

$$v_i = X_{2i} - a_{23} - b_{23}X_{3i} = x_{2i} - \hat{\beta}_{23}x_{3i}$$

O coeficiente de correlação entre X_1 e X_2, com X_3 fixo, é obtido calculando-se o coeficiente de correlação simples entre u_i e v_i, ou seja:

$$\hat{\rho}_{12,3} = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2}\sqrt{\sum_{i=1}^{n} v_i^2}} \tag{3.146}$$

u_i e v_i são resíduos (erros) das regressões de mínimos quadrados, portanto têm médias iguais a zero. Assim, é possível escrever:

$$\hat{\rho}_{12,3} = \frac{\sum_{i=1}^{n} \left(x_{1i} - \hat{\beta}_{13} x_{3i}\right)\left(x_{2i} - \hat{\beta}_{23} x_{3i}\right)}{\sqrt{\sum_{i=1}^{n} u_i^2}\sqrt{\sum_{i=1}^{n} v_i^2}} \tag{3.147}$$

Tem-se da expressão (3.18) na seção 3.2.1.3, que:

$$\hat{\rho}^2_{X,Y} = \frac{\sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2} = 1 - \frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} y_i^2} \tag{3.148}$$

Portanto: $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} y_i^2 \left(1 - \hat{\rho}^2_{X,Y}\right)$

Da mesma forma tem-se que:

$$\sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} x_{1i}^2 \left(1 - \hat{\rho}^2_{1,3}\right) \quad \text{e} \quad \sum_{i=1}^{n} v_i^2 = \sum_{i=1}^{n} x_{2i}^2 \left(1 - \hat{\rho}^2_{2,3}\right)$$

Tem-se ainda que $\hat{\beta}_{13} = \hat{\rho}_{13} \frac{S_1}{S_3}$ e $\hat{\beta}_{23} = \hat{\rho}_{23} \frac{S_2}{S_3}$, logo:

$$\hat{\rho}_{12,3} = \frac{\sum_{i=1}^{n} x_{1i} x_{2i} - \hat{\rho}_{13} \frac{S_1}{S_3} \sum_{i=1}^{n} x_{2i} x_{3i} - \hat{\rho}_{23} \frac{S_2}{S_3} \sum_{i=1}^{n} x_{1i} x_{3i} + \hat{\rho}_{13} \hat{\rho}_{23} \frac{S_1}{S_3} \frac{S_2}{S_3} \sum_{i=1}^{n} x_{3i}^2}{\sqrt{\sum_{i=1}^{n} x_{1i}^2}\sqrt{\sum_{i=1}^{n} x_{2i}^2}\sqrt{1 - \hat{\rho}^2_{13}}\sqrt{1 - \hat{\rho}^2_{23}}}$$

$$\hat{\rho}_{12,3} = \frac{nS_1S_2\hat{\rho}_{12} - nS_1S_2\hat{\rho}_{13}\hat{\rho}_{23}}{nS_1S_2\sqrt{1-\hat{\rho}_{13}^2}\sqrt{1-\hat{\rho}_{23}^2}}$$

Desse modo, o Coeficiente de Correlação Parcial pode ser obtido através de:

$$\hat{\rho}_{12,3} = \frac{\hat{\rho}_{12} - \hat{\rho}_{13}\hat{\rho}_{23}}{\sqrt{\left(1-\hat{\rho}_{13}^2\right)\left(1-\hat{\rho}_{23}^2\right)}} \qquad (3.149)$$

A expressão acima representa o Coeficiente de Correlação Parcial de primeira ordem entre as variáveis X_1 e X_2, com X_3 fixa.

3.3.2.4 Aplicação do Coeficiente de Correlação Múltipla

A aplicação apresentada refere-se ao trabalho de LIMA e SILANS (1999), que estudaram a variabilidade espacial da infiltração e dos parâmetros hidrodinâmicos do solo das equações de Philip e de Green e Ampt, apresentadas a seguir.

A Equação de Philip é dada pela expressão: $I = St^{1/2} + At$

onde: I é a lâmina de água infiltrada no solo (cm);

S é a absortividade $(cm/min^{1/2})$;

t é o tempo (min);

A é a constante da equação de Philip.

E a Equação de Green e Ampt, dada por: $I = K_o t - (h_f - h_o)\Delta\theta \ln\left[1 - \frac{I}{\Delta\theta(h_f - h_o)}\right]$

onde: I é a lâmina de água infiltrada no solo (cm);

K_o é condutividade hidráulica à saturação (cm/min);

h_f é a sucção na frente de umedecimento do solo (cm);

h_o é a carga hidráulica acima do plano representada pela superfície do solo (cm);

$\Delta\theta$ é a diferença entre a umidade volumétrica da frente de umidificação e a umidade volumétrica inicial do solo (cm^3/cm^3).

A parcela selecionada para o estudo situa-se na Fazenda Experimental da EMEPA, em João Pessoa. Demarcou-se uma área de 5.000 m^2, e foi traçada uma malha retangular com espaçamento de 15 m, com um total de 32 nós. Em cada nó efetuaram-se testes de infiltração com duração de 90 minutos, com infiltrômetro duplo-anel de carga constante. Utilizando-se o método gravimétrico-padrão, determinaram as umidades volumétricas de amostras coletadas antes e depois do teste de infiltração. O peso específico aparente do solo seco foi determinado a partir de amostras nos 20 primeiros centímetros do solo. Também a análise granulométrica foi efetuada em ponto de medição, e obtida a porcentagem da fração de argila + silte.

Foram calculados os coeficientes de correlação simples entre a porcentagem de argila + silte e as demais variáveis das equações de infiltração de Philip e Green e Ampt, conforme apresentada na tabela 5.

TABELA 5 - COEFICIENTE DE CORRELAÇÃO ENTRE VARIÁVEIS DAS EQUAÇÕES DE INFILTRAÇÃO E PORCENTAGEM DE ARGILA E SILTE, EM JOÃO PESSOA

VARIÁVEL DEPENDENTE	VARIÁVEIS DAS EQUAÇÕES DE INFILTRAÇÃO	COEFICIENTE DE CORRELAÇÃO
Porcentagem (argila + silte)	Teor da umidade do solo, após o teste de infiltração	-0,16
	Absortividade	-0,32
	Constante da equação de Philip	0,09
	Condutividade hidráulica na superfície em regime permanente de infiltração	0,02
	Variação do teor de umidade volumétrica	-0,09

FONTE: LIMA E SILANS (1999)
NOTAS: Equações de Philip e Green e Ampt.
A área selecionada para o estudo situa-se na fazenda experimental da Empresa de Estudos e Pesquisas Agropecuárias do Estado da Paraíba, em João Pessoa.

Segundo os autores, não existem correlações significativas entre o fator textual (porcentagem de argila + silte) e as propriedades hidrodinâmicas do solo (variáveis das equações de infiltração).

Utilizando a Correlação Múltipla, os autores procuraram detectar a existência de uma possível direção privilegiada das propriedades físicas e hidrodinâmicas do solo. Foi utilizado um sistema de referência ortogonal, onde o eixo das ordenadas é orientado na direção longitudinal da malha. Procuraram estabelecer a relação linear da seguinte forma: $W = aX + bY + c$.

Na tabela 6 apresentam-se os coeficientes de regressão e correlação múltipla, obtidos pelos autores.

TABELA 6 - COEFICIENTES DE REGRESSÃO E CORRELAÇÃO MÚLTIPLA

VARIÁVEL DEPENDENTE	COEFICIENTES DE REGRESSÃO			COEFICIENTE DE CORRELAÇÃO MÚLTIPLA
	a	b	c	
Porcentagem de argila + silte	-0,01	0,04	1,35	0,82
Teor da umidade do solo, após o teste de infiltração	0,00	0,00	0,24	0,35
Absortividade	0,09	-0,02	3,94	0,43
Constante da equação de Philip	0,09	0,02	1,50	0,59
Condutividade hidráulica na superfície em regime permanente de infiltração	0,09	0,01	2,17	0,57
Variação do teor de umidade volumétrica	0,00	0,00	0,17	0,30

FONTE: LIMA E SILANS (1999)

Os resultados da tabela acima indicam que a porcentagem de argila + silte é fortemente correlacionada com a direção ($\hat{\rho}_{w,x,y} = 0{,}82$). É possível observar, ainda, que 67,0% (coeficiente de explicação) da variância da porcentagem de argila + silte na parcela é explicada pela posição do ponto de amostragem (direção). Aproximadamente 35% das variâncias da constante da Equação de Philip (A) e condutividade hidráulica na superfície (K_0) são explicadas pela posição do ponto de amostragem, mas sem a indicação de uma direção previlegiada, verificada através dos coeficientes de correlação múltipla iguais a $\hat{\rho}_{w,x,y} = 0{,}59$ e $\hat{\rho}_{w,x,y} = 0{,}57$, respectivamente.

3.3.3 Análise de Correlação Canônica

3.3.3.1 Introdução

A análise de correlação canônica é uma técnica para a identificação e quantificação da associação entre dois grupos de variáveis. Conforme descrito em CHAVES NETO (2002b), o objetivo dessa técnica é determinar as combinações lineares $U = \underline{c}'_1 X$ e $V = \underline{c}'_2 Y$ tais que tenham a maior correlação possível. A Análise de Correlação Canônica pode ser entendida como uma extensão da Análise de

Regressão Múltipla. Na Análise de Regressão Múltipla, as variáveis formam o conjunto das covariáveis $\underline{X}$ (variáveis independentes) com p variáveis e a variável resposta Y (variável dependente). No problema de Análise de Regressão, a solução está em achar a combinação linear $\beta'\underline{X}$ que é altamente correlacionada com Y e na análise de correlação canônica o conjunto $\underline{Y}$ contém $p \geq 1$ variáveis, devendo-se achar os vetores $\underline{c}_1$ e $\underline{c}_2$ para os quais a correlação entre $U = \underline{c}'_1\underline{X}$ e $V = \underline{c}'_2\underline{Y}$ é máxima.

Tem-se interesse em medir a associação entre os dois grupos de variáveis. O primeiro grupo de p variáveis é representado pelo vetor aleatório $\underline{X}$ $(p \times 1)$ e o segundo de q variáveis $\underline{Y}$ $(q \times 1)$, sendo $p \leq q$.

Tem-se para os vetores aleatórios:

$$E(\underline{X}) = \mu_1;\ \ COV(\underline{X}) = \Sigma_{11};\ E(\underline{Y}) = \mu_2;\ \ COV(\underline{Y}) = \Sigma_{22};\ \ COV(\underline{X},\underline{Y}) = \Sigma_{12} = \Sigma_{21}$$

Sejam as combinações lineares:

$$U = \underline{c}_1{}'\underline{X} \quad \text{e} \quad V = \underline{c}_2{}'\underline{Y}$$

$$\text{Então} \quad Corr(U,V) = \frac{COV(U,V)}{\sqrt{V(U)V(V)}} = \frac{E\left[\left(U-\overline{U}\right)\left(V-\overline{V}\right)\right]}{\sqrt{V(U)V(V)}} \tag{3.150}$$

$$\text{Onde: } E\left[\left(U-\overline{U}\right)\left(V-\overline{V}\right)\right] = E\left[\left(U-\underline{c}_1{}'\mu_1\right)\left(V-\underline{c}_2{}'\mu_2\right)\right] = \underline{c}_1{}'\Sigma_{12}\underline{c}_2$$

$$V(U) = V(\underline{c}_1{}'\underline{X}) = \underline{c}_1{}'COV(\underline{X}) = \underline{c}_1{}'\Sigma_{11}\underline{c}_1$$

$$V(V) = V(\underline{c}_2{}'Y) = \underline{c}_2{}'COV(\underline{Y}) = \underline{c}_2{}'\Sigma_{22}\underline{c}_2$$

$$\text{Portanto,} \quad Corr(U,V) = \frac{\underline{c}_1{}'\Sigma_{12}\underline{c}_2}{\sqrt{\underline{c}_1{}'\Sigma_{11}\underline{c}_1 \times \underline{c}_2{}'\Sigma_{22}\underline{c}_2}} \tag{3.151}$$

O primeiro par de variáveis canônicas são as combinações lineares U_1, V_1, com variância unitária que maximiza a correlação (3.151).

O segundo par de variáveis canônicas são as combinações lineares, com variância unitária, que maximiza a correlação (3.151) entre todas as escolhas que

não são correlacionadas com o primeiro par de variáveis canônicas. E assim, até a k-ésima variável canônica.

Sendo os vetores $\underline{X}$ e $\underline{Y}$ de dimensão p e q com matrizes de covariâncias Σ_1 e Σ_2, respectivamente, e covariância cruzada Σ_{12}, com combinações lineares $U = \underline{c}_1' \underline{X}$ e $V = \underline{c}_2' \underline{Y}$. A correlação máxima Corr(U,.V) é alcançada em $Corr(U,V) = \rho_1^*$ com $\underline{c}'_1 = \underline{e}'_1 \Sigma_1^{-1/2}$ e $\underline{c}'_2 = \underline{f}'_1 \Sigma_2^{-1/2}$, onde $\underline{e}_1$ é o autovetor correspondente ao maior autovalor $\rho_1^{*^2}$ de $\Sigma_1^{-1/2}\Sigma_{12}\Sigma_2^{-1}\Sigma_{21}\Sigma_1^{1/2}$ com p autovalores $\rho_1^{*^2} \geq \rho_2^{*^2} \geq ... \geq \rho_p^{*^2}$ e p autovetores $\underline{e}_k$, $k = 1,2,3,...,p$. Já $\underline{f}_1$ é o autovetor correspondente ao maior autovalor de $\Sigma_2^{-1/2}\Sigma_{21}\Sigma_1^{-1}\Sigma_{12}\Sigma_2^{1/2}$ que tem q autovetores $\underline{f}_k$ correspondentes aos autovalores $\rho_1^{*^2} \geq \rho_2^{*^2} \geq ... \geq \rho_q^{*^2}$.

As variáveis canônicas U_k com $k = 1,2,3,...,p$ são da forma $U_k = \underline{e}'_k \Sigma_{11}^{-1/2} \underline{X}$, e as variáveis canônicas V_k com $k = 1,2,3,...,p$, da forma $V_k = \underline{f}'_k \Sigma_{22}^{-1/2} \underline{Y}$. Sendo $a'_k = \underline{e}'_k \Sigma_{11}^{-1/2}$ tem-se que $U_k = \underline{a}'_k \underline{X}$ e sendo $b'_k = \underline{f}'_k \Sigma_{22}^{-1/2}$ tem-se $V_k = \underline{b}'_k \underline{Y}$ e são formados os pares de variáveis canônicas U_1 e V_1, U_2 e V_2, ..., U_p e V_p, sendo que a máxima correlação canônica é obtida para o primeiro par.

Assim, a correlação entre U_1 e V_1 é dada por: $Corr(U_1, V_1) = \rho_1^* = \sqrt{\rho_1^{*^2}}$. Da mesma forma até k-ésimas variáveis canônicas, quando se tem:

$$Corr(U_k, V_k) = \rho_k^* = \sqrt{\rho_k^{*^2}} \qquad (3.152)$$

As matrizes de covariâncias Σ podem ser substituídas pelas matrizes de correlação ρ. Em se tratando de estudos a partir de dados amostrais, a matriz de covariância e de correlação serão R e $\hat{\rho}$, respectivamente. As correlações canônicas serão obtidas da mesma forma, a partir da matriz de covariância ou de correlação.

3.3.3.2 Aplicação da Análise de Correlação Canônica

A aplicação apresentada a seguir refere-se ao trabalho de FEY NETO (1999), que utilizou a Análise de Correlação Canônica, com o objetivo de estimar o grau de

associação entre o grupo de variáveis que representam as características da qualidade do papel, e o que representa as características da matéria-prima (madeira) e as características do processo (pasta). O objetivo da pesquisa era identificar o grupo de variáveis mais fortemente relacionado com a qualidade (madeira ou pasta).

O trabalho foi realizado com dados levantados em uma indústria de fabricação de papel, PISA - Papel de Imprensa S.A., no período de 23 de julho de 1998 a 31 de março de 1999.

São duas as etapas fundamentais na elaboração do papel. A primeira etapa consiste no recebimento da matéria-prima, em que esta é picada e transformada em cavaco. Na segunda etapa, o cavaco produzido na etapa anterior é transformado em pasta, e por sucessivas operações obtém-se o papel. O problema está na identificação das etapas que têm maior influência na qualidade do papel.

Foram definidas as variáveis que caracterizam cada uma das etapas e a qualidade do papel, conforme descritas a seguir.

Grupo 1 - Variáveis que caracterizam a madeira

Totalizam um conjunto de 15 variáveis: densidade básica; umidade; resina; espessura da fibra 4 mm; espessura da fibra 6 mm; espessura da fibra 8 mm; espessura da fibra 18 mm; comprimento dos cavacos finos; comprimento dos cavacos palitos; comprimento do cavaco > 45 mm; comprimento do cavaco < 45 mm; largura da fibra; diâmetro do lúmem; comprimento da fibra e espessura da fibra.

Grupo 2 - Variáveis que caracterizam a qualidade

São 5 as variáveis deste grupo: alvura, tração, rasgo, densidade e csf.

Grupo 3 - Variáveis que caracterizam a elaboração da pasta

Este grupo é composto por 27 variáveis: produção; cs-04; csf-04; gapte-04; gapde-04; pressão Te-04; pressão De-04; diluição-04; pressão de operação do disco -04; potência-04; cee-04; durabilidade disco externo-04; durabilidade disco interno estator-04; durabilidade disco interno rotor-04; cs-05; csf-05; gapte-05;

gapde-05; pressão Te-05; pressão De-05; diluição-05; pressão de operação do disco-05; potência-05; cee-05; durabilidade disco externo-05; durabilidade disco interno; estator-05 e durabilidade disco interno rotor-05.

Foi inicialmente obtida a matriz de correlação, a partir das correlações simples para cada par de variáveis do grupo 1 (características da madeira), em seguida para o grupo 2 (características da qualidade do papel), e finalmente entre as variáveis dos grupos 1 e 2.

A partir da matriz de correlação foram obtidas as seguintes correlações canônicas entre variáveis do grupo 1 (características da madeira) e grupo 2 (características da qualidade do papel):

QUADRO 10 - CORRELAÇÕES CANÔNICAS ENTRE AS VARIÁVEIS DO GRUPO 1 E GRUPO 2

$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$	$\hat{\rho}_4$	$\hat{\rho}_5$
0,734628	0,48529	0,44777	0,3558	0,30075

FONTE: FEY NETO (1999)

Da mesma forma, foi obtida a matriz de correlação, a partir das correlações simples para cada par de variáveis do grupo 3 (características da elaboração da pasta) e a matriz de correlação entre cada par de variáveis dos grupos 2 e 3.

As correlações canônicas entre as variáveis dos grupos 2 e 3 vêm apresentadas a seguir:

QUADRO 11 - CORRELAÇÕES CANÔNICAS ENTRE AS VARIÁVEIS DO GRUPO 2 E GRUPO 3

$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$	$\hat{\rho}_4$	$\hat{\rho}_5$
0,979863	0,909272	0,707575	0,621093	0,515732

FONTE: FEY NETO (1999)

As correlações canônicas entre os grupos de variáveis que representam as características do processo de elaboração da pasta (grupo 3) e qualidade do papel (grupo 2) são superiores às correlações entre os grupos de variáveis referentes à madeira (grupo 1) e qualidade do papel (grupo 2). O que significa que a qualidade do papel depende mais fortemente das variáveis do processo de produção (elaboração de pasta) do que das variáveis que caracterizam a matéria-prima (papel).

4 RESULTADOS E DISCUSSÃO

4.1 INTRODUÇÃO

O objetivo deste capítulo foi fazer a comparação entre os coeficientes de correlação estimados pelo método de Correlação Linear de Pearson e os métodos de Correlação Bisserial e Tetracórico, utilizando amostras de diferentes tamanhos e mediana como ponto de dicotomização. As amostras foram obtidas pelo processo de simulação.

Utilizou-se o programa disponibilizado pelo Statistical Analysis Software (SAS), para obter as amostras com distribuições normais bivariadas. Os programas encontram-se no Apêndice 6.

Para o cálculo do Coeficiente de Correlação Linear de Pearson utilizou-se a *Procedure Correlation* (PROC CORR). O Coeficiente de Correlação Bisserial foi calculado através do programa desenvolvido também no SAS, e o Coeficiente de Correlação Tetracórico foi obtido através da opção PLCORR, disponível na *Procedure Frequency* (PROC FREQ). Este procedimento adota o método iterativo para o cálculo do Coeficiente de Correlação Tetracórico, através do algoritmo de Newton-Raphson, permitindo definir o número máximo de iterações e o critério de convergência.

4.2 COEFICIENTES DE CORRELAÇÃO LINEAR DE PEARSON, BISSERIAL E TETRACÓRICO

Para avaliar os métodos de Correlação de Pearson, Bisserial e Tetracórico, utilizaram-se amostras com distribuições normais bivariadas, de diferentes tamanhos e parâmetros, obtidas pelo processo de simulação. O quadro 12 apresenta os tamanhos de amostra e parâmetros adotados. No quadro 13 estão apresentadas as médias, desvios padrão e as medianas das variáveis X e Y.

QUADRO 12 - PARÂMETROS UTILIZADOS NO PROCESSO DE SIMULAÇÃO PARA A OBTENÇÃO DAS AMOSTRAS NORMAIS BIVARIADAS

NÚMERO DA AMOSTRA	TAMANHO DA AMOSTRA	SEMENTE	PARÂMETROS				
			μ_X	σ_X^2	μ_Y	σ_Y^2	ρ
1	100	123	10	4	20	4	0,90
2	100	123	5	2	20	2	-0,80
3	100	123	5	3	20	3	0,40
4	200	123	40	5	50	5	0,85
5	200	123	15	5	20	5	-0,70
6	200	123	30	8	20	8	0,30
7	300	123	20	7	25	7	0,80
8	300	123	15	5	25	5	-0,90
9	300	123	20	10	35	10	0,25
10	500	123	5	2	20	2	0,80
11	500	123	80	25	70	25	-0,75
12	500	123	60	35	50	35	0,35
13	1 000	123	80	30	75	30	0,80
14	1 000	123	60	25	45	25	-0,85
15	1 500	123	30	20	45	20	0,70
16	2 000	123	45	25	30	25	0,90
17	2 500	123	35	15	70	15	0,80
18	3 000	123	15	9	25	9	0,75
19	4 000	123	65	30	55	30	0,85
20	5 000	123	10	6	14	6	0,70
21	10 000	123	90	30	60	30	0,90

FONTE: A autora

QUADRO 13 - MÉDIA, DESVIO PADRÃO E MEDIANA DAS VARIÁVEIS ALEATÓRIAS X E Y, SEGUNDO O TAMANHO DA AMOSTRA

AMOSTRA	TAMANHO DA AMOSTRA	VARIÁVEL X			VARIÁVEL Y		
		Média	Desvio Padrão	Mediana	Média	Desvio Padrão	Mediana
1	100	9,9202	1,7577	9,8322	19,8159	1,8250	19,7586
2	100	4,9436	1,2429	4,8814	19,9359	1,2252	19,9187
3	100	4,9309	1,5222	4,8547	19,7680	1,6310	19,6763
4	200	39,8386	2,1135	39,7528	49,7350	2,1916	49,6677
5	200	14,8386	2,1135	14,7528	19,9398	2,1338	19,9827
6	200	29,7958	2,6734	29,6873	19,6461	2,8529	19,5381
7	300	19,9017	2,9462	19,8717	24,8569	2,6341	24,8305
8	300	14,9169	2,1096	14,8916	25,0352	2,0558	25,0609
9	300	19,8825	2,9835	19,8467	34,8463	3,1624	34,9224
10	500	4,9445	1,3392	4,93319	19,9088	1,3715	19,8967
11	500	79,8038	4,7347	79,7638	69,9646	4,7565	69,9133
12	500	59,7679	5,6022	59,7205	49,6130	5,8221	49,5430
13	1 000	79,8360	5,2155	79,7644	74,6442	5,3334	74,5860
14	1 000	59,8503	4,7611	59,7849	44,9471	4,7972	45,0081
15	1 500	29,9193	4,3006	29,8879	44,7371	4,4488	44,7477
16	2 000	44,8611	4,9314	44,8338	29,6962	4,9960	29,6922
17	2 500	34,8761	3,8223	34,8867	39,7716	3,8559	39,7339
18	3 000	14,9222	2,9479	14,9211	24,8472	2,9957	24,7976
19	4 000	64,8446	5,4385	64,8399	54,7883	5,4910	54,7473
20	5 000	9,95013	2,4390	9,92922	13,9267	2,4621	13,9141
21	10 000	89,9673	5,4623	89,9417	59,9408	5,5036	59,9822

FONTE: A autora

Verificou-se, inicialmente, a homogeneidade das variâncias das amostras através de testes de hipóteses.

A hipótese $H_0 : \sigma_X^2 = \sigma_Y^2$ contra $H_1 : \sigma_X^2 \neq \sigma_Y^2$ foi testada pela razão F definida como:

$$F = \frac{S_1^2}{S_2^2}$$ onde: F é a estatística do teste; (4.1)

S_1^2 é a variância da primeira amostra;

S_2^2 é a variância da segunda amostra.

O quadro 14 apresenta os desvios padrão da variável X e Y, a razão F e o valor-p.

QUADRO 14 - DESVIOS PADRÃO DAS VARIÁVEIS X E Y, RAZÃO F E VALOR-P, SEGUNDO O TAMANHO DA AMOSTRA

AMOSTRA	TAMANHO DA AMOSTRA	S_X	S_y	F	VALOR-P
1	100	1,7577	1,8250	0,9276	0,7093
2	100	1,2429	1,2252	1,0291	0,8867
3	100	1,5222	1,6310	0,8711	0,4935
4	200	2,1135	2,1916	0,9308	0,6138
5	200	2,1135	2,1338	0,9811	0,8929
6	200	2,6734	2,8529	0,8781	0,3599
7	300	2,9462	2,6341	0,8980	0,3527
8	300	2,1096	2,0558	1,0531	0,6551
9	300	2,9835	3,1624	0,8900	0,3143
10	500	1,3392	1,3715	0,9534	0,5940
11	500	4,7347	4,7565	0,9909	0,9183
12	500	5,6022	5,8221	0,9259	0,3900
13	1 000	5,2155	5,3334	0,9563	0,4798
14	1 000	4,7611	4,7972	0,9850	0,8110
15	1 500	4,3006	4,4488	0,9345	0,1897
16	2 000	4,9314	4,9960	0,9743	0,5606
17	2 500	3,8223	3,8559	0,9826	0,6617
18	3 000	2,9479	2,9957	0,9684	0,3791
19	4 000	5,4385	5,4991	0,9778	0,4769
20	5 000	2,4390	2,4621	0,9813	0,5055
21	10 000	5,4623	5,5036	0,9851	0,4520

FONTE: A autora

Os valores-p referentes aos testes de hipóteses para verificar a homogeneidade das variâncias, apresentados no quadro acima, são todos superiores a 0,05; portanto, aceita-se a hipótese H_0 de que as variâncias são iguais para todos os tamanhos de amostra.

Tem-se, então, as variáveis X e Y com distribuição normal bivariada e variâncias homogêneas.

4.2.1 Cálculo dos Coeficientes de Correlação

Os quadros 15, 16 e 17 apresentam os coeficientes de correlação obtidos pelos métodos de Pearson, Bisserial e Tetracórico, calculados para os diferentes tamanhos de amostras. Utilizou-se a mediana como ponto de dicotomização da variável Y, para o método de Coeficiente de Correlação Bisserial e das variáveis X e Y para o método de Coeficiente de Correlação Tetracórico.

Os estimadores utilizados para os cálculos dos coeficientes de correlação e os erros padrão já foram apresentados no Capítulo 3 e transcritos a seguir.

A distribuição amostral do Coeficiente Linear de Pearson $(\hat{\rho})$ discutida na seção 3.2.1.5, é assimétrica quando o parâmetro populacional (ρ) é diferente de zero, dificultando a sua interpretação. Para resolver este problema, Ronald A. Fisher desenvolveu a estatística Z, discutida na seção 3.2.1.7. Contudo, não existem estatísticas equivalentes a Z, para os métodos de Correlação Bisserial e Tetracórico.

Portanto, para os cálculos dos erros padrão, considerou-se o coeficiente de correlação populacional (parâmetro) iguais a zero, para os três métodos.

Calculou-se também o erro relativo percentual, com o objetivo de avaliar os erros entre o Coeficiente de Correlação Linear de Pearson e o Coeficiente de Correlação Bisserial e entre o Coeficiente de Correlação Linear de Pearson e o Coeficiente de Correlação Tetracórico, para diferentes tamanhos de amostras.

a) Estimadores do Método de Coeficiente de Correlação Linear de Pearson

I) Coeficiente de Correlação

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \qquad (4.2)$$

II) Erro padrão

$$\hat{\sigma}_{\hat{\rho}} = \frac{1}{\sqrt{n-1}} \tag{4.3}$$

QUADRO 15 - COEFICIENTE DE CORRELAÇÃO LINEAR DE PEARSON ($\hat{\rho}$) E ERRO PADRÃO, SEGUNDO O TAMANHO DA AMOSTRA

NÚMERO DA AMOSTRA	TAMANHO DA AMOSTRA	$\hat{\rho}$	$\hat{\sigma}_{\hat{\rho}}$
1	100	0,89704	0,10050
2	100	-0,76775	0,10050
3	100	0,43492	0,10050
4	200	0,84261	0,07089
5	200	-0,66200	0,07089
6	200	0,32073	0,07089
7	300	0,80669	0,05783
8	300	-0,88534	0,05783
9	300	0,31393	0,05783
10	500	0,79475	0,04477
11	500	-0,73125	0,04477
12	500	0,35775	0,04477
13	1 000	0,78949	0,03164
14	1 000	-0,83720	0,03164
15	1 500	0,68755	0,02583
16	2 000	0,89782	0,02237
17	2 500	0,79524	0,02000
18	3 000	0,74349	0,01826
19	4 000	0,84814	0,01581
20	5 000	0,70072	0,01414
21	10 000	0,90049	0,01000

FONTE: A autora

b) Estimadores do Método de Coeficiente de Correlação Bisserial

I) Coeficiente de correlação

$$\hat{\rho}_b = \frac{\overline{X}_p - \overline{X}_t}{S_t} \times \frac{p}{y} \tag{4.4}$$

II) Erro padrão

$$\hat{\sigma}_{\hat{\rho}_b} = \frac{\frac{\sqrt{pq}}{y}}{\sqrt{n}} \tag{4.5}$$

QUADRO 16 - COEFICIENTE DE CORRELAÇÃO BISSERIAL ($\hat{\rho}_b$) E ERRO PADRÃO, SEGUNDO O TAMANHO DA AMOSTRA

NÚMERO DA AMOSTRA	TAMANHO DA AMOSTRA	$\hat{\rho}_b$	$\hat{\sigma}_{\hat{\rho}_b}$
1	100	0,94610	0,12533
2	100	-0,78635	0,12534
3	100	0,45559	0,12534
4	200	0,89507	0,08862
5	200	-0,69485	0,08862
6	200	0,37090	0,08863
7	300	0,85043	0,07236
8	300	-0,88661	0,07236
9	300	0,27665	0,07236
10	500	0,83054	0,05605
11	500	-0,71946	0,05605
12	500	0,29871	0,05605
13	1 000	0,81672	0,03963
14	1 000	-0,82284	0,03963
15	1 500	0,68462	0,03236
16	2 000	0,90806	0,02803
17	2 500	0,80552	0,02507
18	3 000	0,74258	0,02288
19	4 000	0,85946	0,01982
20	5 000	0,70337	0,01773
21	10 000	0,90574	0,01253

FONTE: A autora

c) Estimadores do Método de Coeficiente de Correlação Tetracórico

I) Coeficiente de correlação

$$\frac{ad-bc}{yy'n^2} = \hat{\rho}_t + \hat{\rho}_t^2 \frac{zz'}{2} + \hat{\rho}_t^3 \frac{(z^2-1)(z'^2-1)}{6} + ... \quad (4.6)$$

II) Erro padrão

$$\hat{\sigma}_{\hat{\rho}_t} = \frac{\sqrt{p \times q \times p' \times q'}}{y \times y' \times \sqrt{n}} \quad (4.7)$$

QUADRO 17 - COEFICIENTE DE CORRELAÇÃO TETRACÓRICO ($\hat{\rho}_t$) E ERRO PADRÃO, SEGUNDO O TAMANHO DA AMOSTRA

NÚMERO DA AMOSTRA	TAMANHO DA AMOSTRA	$\hat{\rho}_t$	$\hat{\sigma}_{\hat{\rho}_t}$
1	100	0,95110	0,15705
2	100	-0,68450	0,15714
3	100	0,58780	0,15712
4	200	0,89100	0,11103
5	200	-0,63740	0,11114
6	200	0,36810	0,11112
7	300	0,85540	0,09079
8	300	-0,85540	0,09079
9	300	0,30900	0,09068
10	500	0,83750	0,07013
11	500	-0,68450	0,07020
12	500	0,33280	0,07027
13	1 000	0,84090	0,04956
14	1 000	-0,79780	0,04962
15	1 500	0,69970	0,04055
16	2 000	0,91400	0,03521
17	2 500	0,80750	0,03131
18	3 000	0,74590	0,02863
19	4 000	0,85830	0,02481
20	5 000	0,70260	0,02217
21	10 000	0,90850	0,01565

FONTE: A autora

4.2.2 Comparação dos Erros Padrão

O quadro 18 apresenta a comparação dos erros padrão estimados pelos três métodos. É interessante observar que a razão entre os erros padrão dos Coeficientes de Correlação Bisserial e de Pearson é aproximadamente de 1,25, ou seja, o primeiro é 25% superior, confirmando o que foi observado por GUILFORD (1950) e apresentado na seção 3.2.2.3. Em relação à razão entre os erros padrão dos Coeficientes de Correlação Tetracórico e de Pearson, esta é de aproximadamente 1,56, ou seja, o erro padrão do Coeficiente de Correlação Tetracórico é cerca de 56% superior ao de Pearson, também observado por GUILFORD (1950), discutido na seção 3.2.4.3.

QUADRO 18 - ERROS PADRÃO DOS COEFICIENTES DE CORRELAÇÃO LINEAR DE PEARSON, BISSERIAL E TETRACÓRICO, SEGUNDO O TAMANHO DA AMOSTRA

NÚMERO DA AMOSTRA	TAMANHO DA AMOSTRA	$\hat{\sigma}_{\hat{\rho}}$	$\hat{\sigma}_{\hat{\rho}_b}$	$\hat{\sigma}_{\hat{\rho}_t}$	$\hat{\sigma}_{\hat{\rho}_b} / \hat{\sigma}_{\hat{\rho}}$	$\hat{\sigma}_{\hat{\rho}_t} / \hat{\sigma}_{\hat{\rho}}$
1	100	0,10050	0,12533	0,15705	1,24702	1,56261
2	100	0,10050	0,12534	0,15714	1,24706	1,56352
3	100	0,10050	0,12534	0,15712	1,24708	1,56334
4	200	0,07089	0,08862	0,11103	1,25014	1,56632
5	200	0,07089	0,08862	0,11114	1,25014	1,56777
6	200	0,07089	0,08863	0,11112	1,25024	1,56756
7	300	0,05783	0,07236	0,09079	1,25115	1,56989
8	300	0,05783	0,07236	0,09079	1,25129	1,56989
9	300	0,05783	0,07236	0,09068	1,25120	1,56800
10	500	0,04477	0,05605	0,07013	1,25203	1,56663
11	500	0,04477	0,05605	0,07020	1,25204	1,56826
12	500	0,04477	0,05605	0,07027	1,25207	1,56970
13	1 000	0,03164	0,03963	0,04956	1,25269	1,56652
14	1 000	0,03164	0,03963	0,04962	1,25261	1,56843
15	1 500	0,02583	0,03236	0,04055	1,25295	1,57003
16	2 000	0,02237	0,02803	0,03521	1,25314	1,57407
17	2 500	0,02000	0,02507	0,03131	1,25311	1,56543
18	3 000	0,01826	0,02288	0,02863	1,25285	1,56796
19	4 000	0,01581	0,01982	0,02481	1,25333	1,56919
20	5 000	0,01414	0,01773	0,02217	1,25333	1,56762
21	10 000	0,01000	0,01253	0,01565	1,25330	1,56471

FONTE: A autora

4.2.3 Comparação dos Coeficientes de Correlação Estimados

As comparações entre os Coeficientes de Correlação Linear de Pearson e os Coeficientes de Correlação Bisserial e Tetracórico foram feitas através do cálculo do erro relativo percentual. Este erro indica relativamente o quanto o Coeficiente de Correlação Bisserial e o Tetracórico diferem do Coeficiente de Correlação Linear de Pearson.

O erro relativo percentual foi obtido pela expressão:

$$\text{erp}(\hat{\rho}_b) = \left| \frac{(\hat{\rho}_b - \hat{\rho})}{\hat{\rho}} \right| \times 100 \quad \text{e} \quad \text{erp}(\hat{\rho}_t) = \left| \frac{(\hat{\rho}_t - \hat{\rho})}{\hat{\rho}} \right| \times 100 \tag{4.8}$$

onde:

$\text{erp}(\hat{\rho}_b)$ é o erro relativo percentual do Coeficiente de Correlação Bisserial em relação ao Coeficiente de Correlação Linear de Pearson

$erp(\hat{\rho}_t)$ é o erro relativo percentual do Coeficiente de Correlação Tetracórico em relação ao Coeficiente de Correlação Linear de Pearson

$\hat{\rho}$ é o Coeficiente de Correlação Linear de Pearson estimado

$\hat{\rho}_b$ é o Coeficiente de Correlação Bisserial estimado

$\hat{\rho}_t$ é o Coeficiente de Correlação Tetracórico estimado

O quadro 19 apresenta os erros relativos percentuais entre o Coeficiente de Correlação Bisserial e o de Pearson e entre os do Coeficiente de Correlação Tetracórico e de Pearson.

QUADRO 19 - COEFICIENTES DE CORRELAÇÃO LINEAR DE PEARSON, BISSERIAL E TETRACÓRICO E ERROS RELATIVOS PERCENTUAIS, BISSERIAL E TETRACÓRIcO, SEGUNDO O TAMANHO DA AMOSTRA

NÚMERO DA AMOSTRA	TAMANHO DA AMOSTRA	$\hat{\rho}$	$\hat{\rho}_b$	$\hat{\rho}_t$	$erp(\hat{\rho}_b)$	$erp(\hat{\rho}_t)$
1	100	0,89704	0,94610	0,95110	5,46910	6,02649
2	100	-0,76775	-0,78635	-0,68450	2,42266	10,84337
3	100	0,43492	0,45559	0,58780	4,75260	35,15129
4	200	0,84261	0,89507	0,89100	6,22589	5,74287
5	200	-0,66200	-0,69485	-0,63740	4,96224	3,71601
6	200	0,32073	0,37090	0,36810	15,64244	14,76943
7	300	0,80669	0,85043	0,85540	5,42216	6,03826
8	300	-0,88534	-0,88661	-0,85540	0,14345	3,38175
9	300	0,31393	0,27665	0,30900	11,87526	1,57041
10	500	0,79475	0,83054	0,83750	4,50330	5,37905
11	500	-0,73125	-0,71946	-0,68450	1,61231	6,39316
12	500	0,35775	0,29871	0,33280	16,50314	6,97414
13	1 000	0,78949	0,81672	0,84090	3,44906	6,51180
14	1 000	-0,83720	-0,82284	-0,79780	1,71524	4,70616
15	1 500	0,68755	0,68462	0,69970	0,42615	1,76714
16	2 000	0,89782	0,90806	0,91400	1,14054	1,80214
17	2 500	0,79524	0,80552	0,80750	1,29269	1,54167
18	3 000	0,74349	0,74258	0,74590	0,12240	0,32415
19	4 000	0,84814	0,85946	0,85830	1,33469	1,19792
20	5 000	0,70072	0,70337	0,70260	0,37818	0,26830
21	10 000	0,90049	0,90574	0,90850	0,58302	0,88952

FONTE: A autora

4.3 AVALIAÇÃO DOS RESULTADOS

A análise do quadro 19 mostra que tanto o Coeficiente Correlação Bisserial quanto o Coeficiente de Correlação Tetracórico diferem do Coeficiente Linear de

Pearson para todos os tamanhos de amostra, sendo o erro relativo percentual maior para amostras de tamanho menor.

Para todos os tamanhos de amostra os Coeficientes de Correlação Bisserial e o Tetracórico fornecem estimativas maiores do que o Coeficiente de Correlação Linear de Pearson. Embora os erros relativos diminuam à medida que se aumenta o tamanho da amostra, devemos considerar que os erros padrão dos Coeficientes de Correlação Bisserial são aproximadamente 25% superiores aos do Coeficiente de Correlação Linear de Pearson e os do Coeficiente de Correlação Tetracórico, em torno de 56% superiores.

É importante destacar que estas são as situações ideais, em que se tem distribuições normais bivariadas com variâncias homogêneas, o que na prática dificilmente ocorre, e, ainda, utilizando as medianas como pontos de dicotomização.

Para a utilização dos Coeficientes de Correlação Bisserial e Tetracórico é necessário que se atenda à suposição da existência de variáveis subjacentes (latentes) às variáveis medidas como dicotômicas, normalmente distribuídas, caso contrário não é possível a sua utilização.

Dentre os três métodos discutidos, é preferível, sempre que possível, utilizar o Coeficiente de Correlação Linear de Pearson.

CONCLUSÕES E RECOMENDAÇÕES

O Coeficiente de Correlação Linear de Pearson, conhecido também como Coeficiente de Correlação do Momento Produto, é, sem dúvida, o mais importante e o mais utilizado, como as aplicações apresentadas no Capítulo 3.

As Técnicas de Análise Multivariada, como a Análise Fatorial, Análise de Componentes Principais e Análise Canônica, utilizam a matriz de correlações, constituída a partir de Coeficientes Linear de Pearson, para cada par de variáveis envolvidas na análise.

As Análises de Confiabilidade em Sistemas de Engenharia e de Instrumentos de Medidas também fazem uso do Coeficiente de Correlação Linear de Pearson.

Comprovou-se que é possível a utilização do Coeficiente Linear de Pearson em situações que envolvem duas variáveis dicotômicas, uma variável dicotômica e outra medida em nível intervalar e duas variáveis medidas em nível ordinal. Os Coeficientes de Correlação Ponto Bisserial, Correlação Phi e Correlação de Spearman fornecem a mesma estimativa do Coeficiente de Correlação Linear de Pearson, pois os seus estimadores são derivados deste último.

Já no caso dos Coeficientes de Correlação Bisserial e Tetracórico, só são possíveis as suas utilizações se existirem variáveis subjacentes (latentes) às variáveis medidas como dicotômicas, normalmente distribuídas. Observou-se que os erros padrão destes coeficientes são superiores aos do Coeficiente de Correlação Linear de Pearson. Além disso, as estimativas dos coeficientes de correlação também são maiores se comparadas às do Coeficiente de Correlação Linear de Pearson.

Quando se tratar de amostras pequenas (normalmente consideradas para $n < 30$), deve-se verificar a suposição da normalidade das variáveis envolvidas na análise. É possível verificar a normalidade das variáveis utilizando o método apresentado na seção 3.2.1.11 (Teste de Normalidade).

Em situações que não atendem à suposição da normalidade das variáveis é possível fazer alguma transformação, como as apresentadas na seção 3.2.1.2.

Ainda, é possível utilizar o recurso dos *ranks* (atribuindo uma ordem aos dados) e utilizar o Coeficiente de Correlação Linear de Pearson.

Apresenta-se a seguir, de forma resumida, as situações em que se pode utilizar cada um dos métodos de coeficientes de correlação envolvendo duas variáveis, discutidos neste trabalho.

a) Coeficiente de Correlação Linear de Pearson: este método pode ser utilizado em situações que envolvem variáveis medidas em nível intervalar e ordinal e variáveis dicotômicas.
b) Coeficiente de Correlação Bisserial: pode ser empregado quando se tem uma variável medida em nível intervalar e outra dicotômica ou dicotomizada (ao serem medidas), porém a suposição da existência de uma variável normalmente distribuída, subjacente à variável dicotômica, deve ser atendida.
c) Coeficiente de Correlação Ponto Bisserial: trata-se do Coeficiente de Correlação Linear de Pearson, quando calculado para uma variável dicotômica e outra medida em nível intervalar.
d) Coeficiente de Correlação Tetracórico: este método deve ser utilizado quando se tem duas variáveis dicotômicas ou dicotomizadas (ao serem medidas), porém a suposição da existência de variáveis normalmente distribuídas, subjacentes às variáveis dicotômicas, deve ser atendida;
e) Coeficiente de Correlação de Spearman: é o Coeficiente de Correlação Linear de Pearson quando se tem duas variáveis medidas em nível ordinal.
f) Coeficiente de Correlação por Postos de Kendall: as duas variáveis envolvidas na análise são medidas em nível ordinal.
g) Coeficiente de Correlação Phi: é o Coeficiente de Correlação Linear de Pearson quando se tem duas variáveis dicotômicas.
h) Coeficiente de Contingência: as duas variáveis são medidas em nível nominal.
i) Coeficiente de Correlação Eta: uma variável é medida em nível intervalar e a outra em nível nominal.

REFERÊNCIAS

AGRESTI, Alan. **Categorical data analysis**. New York: J. Wiley & Sons, 1990. 557p.

ALMEIDA FILHO, Raimundo. Processamento digital de imagens Landsat-TM na detecção de áreas de microexsudação de hidrocarbonetos, região da Serra do Tona, Bahia. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 10., 2001, Foz do Iguaçu. **Anais**. São José dos Campos: INPE, p. 235-242, 2001.

ANDERBERG, Michael R. **Cluster analysis for applications**. New York: Academic Press, 1973. 359p.

ANDERSON, T. W. **An introduction to multivariate statistical analysis**. New York: J. Wiley & Sons, 1958. 375p.

BROWNLEE, K. A. **Statistical theory and methodology in science and engineering**. New York: J. Wiley & Sons, 1960. 570p.

BRYANT, Edward C. **Statistical analysis**. New York: McGraw-Hill Book, 1960. 303p.

BUNCHAFT, Guenia; KELLNER, Sheilah R.O. **Estatística sem mistérios**. 2.ed. Petrópolis: Vozes, 1999. v.2, 303p.

CALLEGARI-JACQUES, Sidia M. **Bioestatística: princípios e aplicações**. Porto Alegre: Artemed, 2003. 255p.

CHAVES NETO, Anselmo. **Probabilidade e estatística matemática II**. Curitiba: UFPR, 1.º semestre de 2002a. Notas de aula.

CHAVES NETO, Anselmo. **Análise multivariada aplicada à pesquisa**. Curitiba: UFPR, 2.º semestre de 2002b. Notas de aula.

CHAVES NETO, Anselmo. **Probabilidade e estatística matemática I**. Curitiba: UFPR, 1.º semestre de 2003. Notas de aula.

CHAVES NETO, Anselmo; TURIM, Maria Elisa. Análise de itens pela teoria clássica da avaliação e TRI em dados reais do ensino fundamental. In:SEMINÁRIO IASI DE ESTATÍSTICA APLICADA, 9., **Anais**. Rio de Janeiro, 2003.

CHEN, Peter Y.; POPOVICH, Paula M. **Correlation:** parametric and nonparametric measures. London: Sage, 2002. 95p.

COCHRAN, William G. **Técnicas de amostragem**. Rio de Janeiro: Fundo de Cultura, 1965. 555p.

CRONBACH, Lee J. Coefficient alpha and the internal structure of testes. **Psychometrika**, v. 16, n. 3, p. 297-333, Sept. 1951.

DOWNIE, N. M.; HEATH, R. W. **Basic statistical methods**. New York: Harper & Brothers, 1959. 289p.

ELDERTON, William P. **Frequency curves and correlation**. 4.ed. Washington: Harren Press, 1953. 272p.

FACHEL, Jandyra M. G. **Análise fatorial**. São Paulo, 1976. 81p. Dissertação (Mestrado) - IME, USP.

FACHEL, Jandyra M. G. **The C-type distribution as an underlying model for categorical data and its use in factor analysis**. London, 1986. 235p. Tese (Doutorado).

FERGUSON, G. A. **Statistical analysis in psycology and education**. Tokyo: McGraw-Hill Kogagusha, 1976.

FERGUSON, George A. **Statistical analysis in psychology and education**. 5.ed. New York: McGraw-Hill book, 1981. 549p.

FEY NETO, Emílio Rudolfo. **Análise de correlação canônica aplicada em sistema de produção contínuo**. Curitiba, 1999. 150p. Dissertação (Mestrado) - Departamento de Informática, Curso de Informática Aplicada, PUC-PR.

FILLIBEN, James J. The Probability plot correlation coefficient test for normality. **Technometrics**, v. 17, n. 1, p. 111-117, Feb. 1975.

FURTADO, Emerson Marcos. **Automação do ranqueamento qualitativo de áreas especialmente protegidas do Estado do Paraná através da análise fatorial**.Curitiba, 1999. 220 p. Dissertação (Mestrado) - Setor de Ciências Exatas, UFPR.

FURTADO, Emerson Marcos; CHAVES NETO, Anselmo et al. Ranqueamento de faxinais do Estado do Paraná. **Revista de Ciências Exatas e Naturais**, v.5, n.1, jan.-jun. 2003.

GALTON, Francis. Correlations and their measurement, chiefly from antropometric data. **Nature**, p. 238, 3 Jan. 1889.

GUILFORD, J. P. **Fundamental statistics in psychology and education**. 4.ed. New York: McGraw-hill Book, 1950. 605p.

HALDAR, A.; MAHADEVAN, S. **Probability, reliability and statistical methods in engineering design**. New York: J. Willey & Sons, 2000. 320p.

JAMES, Barry R. **Probabilidade**: um curso em nível intermediário. Rio de Janeiro: Instituto de Matemática Pura e Aplicada, 1981. 304p.

JOHNSON, Richard A.; WICHERN, Dean W. **Applied multivariate statistical analysis**. 2. ed. New Jesery: Prentice Hall International, 1988. 607p.

KENNEY, J. F.; KEEPING, E. S. **Mathematics of statistics**. 2.ed. Princeton, Van Nostrand, 1951. 429p.

LIMA, Cícero A. G.; SILANS, Alain P. de. Variabilidade espacial da infiltração de água no solo. **Pesquisa Agropecuária Brasileira**, Brasília, v. 34, n. 12, p. 2311-2320, dez. 1999.

LORD, F. e NOVICK, M. R. **Statistical theories of mental test scores**. Reading: Addison-Wesley, 1967. 568p.

McNEMAR, Quinn. **Psychological statistics**. 4. ed. New York: J. Wiley & Sons, 1969. 529p.

MENEZES, Antônio C. F.; FAISSOL, Speridião; FERREIRA, Marilourdes L. Análise da matriz geográfica: estruturas e inter-relaçãoes. In: IBGE. **Tendências atuais da geografia urbano/regional**: teorização e quantificação. Rio de Janeiro, 1978. p. 67-109.

MOOD, Alexander M.; GRAYBILL, Franklin A.; BOES, Duane C. **Introduction to the theory of statistics**. 3. ed. Singapore: McGraw-Hill Book, 1974. 564p.

NETER, John et al. **Applied linear statistical models**. New York: McGraw-Hill, 1996. 1408p.

NOJOSA, Ronald T. **Modelos multidimensionais para a teoria da resposta ao item**. Recife, 2001. 66p. Dissertação (Mestrado), UFPE.

NUNNALLY, Jum C. **Introduccion a la medicion psicologica**. Buenos Aires: McGraw-Hill, 1970. 619 p.

SCHULTZ, Duane P.; SCHULTZ, Sydney Ellen. **História da psicologia moderna**. 16. ed. São Paulo: Cultrix, 1992. 439 p.

SIEGEL, Sidney. **Estatística não-paramétrica**: para as ciências do comportamento. São Paulo: McGraw-Hill do Brasil, 1975. 350 p.

SILVEIRA, Fernando L. Um exemplo de análise multivariada aplicada à pesquisa quantitativa em ensino de ciências: explicando o desempenho dos candidatos ao concurso vestibular de 1999 da Universidade Federal do Rio Grande do Sul. **Investigações em Ensino de Ciências**, Porto Alegre, v. 4, n. 2, p. 161-180, 1999.

SILVEIRA, Fernando L.; PINENT, Carlos E. C. A questão de redação no concurso vestibular à universidade: validade e poder decisório. **Estudos em Avaliação Educacional**, São Paulo, v. 24, p. 147-162, 2001.

SIQUEIRA, Arminda Lucia. **Uso de transformação em análise de variância e análise de regressão**. São Paulo, 1983. 154p. Dissertação (Mestrado), USP/IME.

SNEDECOR, George W.; COCHRAN, William G. **Statistical methods**. 7.ed. Ames: Iowa Sate University, 1980. 507p.

TOBO, Natividad et al. Cumplimiento del régimen terapéutico y su relación con las características biológicas y sociales del individuo con insuficiencia renal crónica terminal en hemodiálisis. **Colombia Médica**, Colombia, v. 26, p. 141-145, 1995.

UFRJ.COPPE.PEC. **COC796-Confiabilidade estrutural**. Métodos analíticos para análise de confiabilidade. Disponível em: http://www.ufrj/coppe/Coc796.doc Acesso em: 2º semestre de 2003.

WANNACOTT, Ronald J.; WANNACOTT, Thomas H. **Econometria**. 2.ed. São Paulo: Livros Técnicos e Científicos, 1978. 424p.

WHERRY, R. J. **Contributions to correlational analysis**. Orlando: Academic Press, 1984. 463p.

APÊNDICE 1 - DISTRIBUIÇÕES AMOSTRAIS DO COEFICIENTE DE CORRELAÇÃO LINEAR DE PEARSON $(\hat{\rho})$

DISTRIBUIÇÕES AMOSTRAIS DO COEFICIENTE DE CORRELAÇÃO LINEAR DE PEARSON $(\hat{\rho})$

(I) PARA QUANDO $\rho \neq 0$

A função densidade de probabilidade de $\hat{\rho}$ para quando $\rho \neq 0$ é conforme apresentado no resultado 3.6:

$$f(\hat{\rho}) = \frac{(n-1)\Gamma(n-1)(1-\rho^2)^{(n-1)/2}(1-\hat{\rho}^2)^{(n-4)/2}}{\sqrt{2\pi}\,\Gamma\left(n-\frac{1}{2}\right)(1-\rho\hat{\rho})^{(n-3/2)}}\left[1+\frac{1}{4}\frac{(\rho\hat{\rho}+1)}{2n-1}+\frac{9}{16}\frac{(\rho\hat{\rho}+1)^2}{2(2n-1)(2n+1)}\right]$$

Considerando a amostra de tamanho $n = 29$ e $\rho = 0{,}80$ tem-se:

$$f(\hat{\rho}) = \frac{(27)\times\Gamma(28)(1-0{,}8^2)^{14}(1-\hat{\rho}^2)^{25/2}}{\sqrt{2\pi}\times\Gamma\left(\frac{57}{2}\right)(1-0{,}8\times\hat{\rho})^{55/2}}\left[1+\frac{1}{4}\frac{(0{,}8\hat{\rho}+1)}{57}+\frac{9}{16}\frac{(0{,}8\hat{\rho}+1)^2}{2\times57\times59}\right]$$

$$f(\hat{\rho}) = \frac{0{,}000001256(1-\hat{\rho}^2)^{25/2}}{(1-0{,}8\times\hat{\rho})^{55/2}}\left[1+\frac{(0{,}8\hat{\rho}+1)}{228}+\frac{9\times(0{,}8\hat{\rho}+1)^2}{2\times53\,808}\right]$$

Substituindo valores para $\hat{\rho}$, obtém-se os correspondentes para $f(\hat{\rho})$. Para a construção do gráfico 5, utilizou-se intervalo para $\hat{\rho}$ igual a 0,0125, iniciando em 0,20. A tabela a seguir apresenta alguns valores como exemplo.

TABELA A.1.1 - COEFICIENTE DE CORRELAÇÃO LINEAR DE PEARSON ($\hat{\rho}$) E RESPECTIVO VALOR DE $f(\hat{\rho})$

$\hat{\rho}$	$f(\hat{\rho})$
0,20	0,00009
0,30	0,00074
0,40	0,00577
0,50	0,04373
0,60	0,30834
0,70	1,78533
0,80	5,72350
0,90	1,94955
1,00	0,00000

FONTE: A autora

Considerando a amostra de tamanho n=29 e $\rho = -0,80$ tem-se:

$$f(\hat{\rho}) = \frac{(27)\times\Gamma(28)(1-(-0,8)^2)^{14}(1-\hat{\rho}^2)^{25/2}}{\sqrt{2\pi}\times\Gamma\left(\frac{57}{2}\right)(1-(-0,8)\times\hat{\rho})^{55/2}}\left[1+\frac{1}{4}\frac{(-0,8\hat{\rho}+1)}{57}+\frac{9}{16}\frac{(-0,8\hat{\rho}+1)^2}{2\times57\times59}\right]$$

$$f(\hat{\rho}) = \frac{0,000001256(1-\hat{\rho}^2)^{25/2}}{(1+0,8\times\hat{\rho})^{55/2}}\left[1+\frac{(-0,8\hat{\rho}+1)}{228}+\frac{9\times(-0,8\hat{\rho}+1)^2}{2\times53\,808}\right]$$

Substituindo valores para $\hat{\rho}$, obtém-se os correspondentes para $f(\hat{\rho})$. Alguns valores são apresentados na tabela a seguir. Para a construção do gráfico 6, o intervalo utilizado para $\hat{\rho}$ foi de 0,0125 e o valor inicial igual a -1,0.

TABELA A.1.2 - COEFICIENTE DE CORRELAÇÃO LINEAR DE PEARSON ($\hat{\rho}$)E RESPECTIVO VALOR DE $f(\hat{\rho})$

$\hat{\rho}$	$f(\hat{\rho})$
-1,00	0,00000
-0,90	1,94955
-0,80	5,72350
-0,70	1,78533
-0,60	0,30834
-0,50	0,04373
-0,40	0,00577
-0,30	0,00074
-0,20	0,00009

FONTE: A autora

(II) PARA QUANDO $\rho = 0$

A função densidade de probabilidade de $\hat{\rho}$ para quando $\rho = 0$ é conforme apresentada no Resultado 3.7 :

$$f(\hat{\rho}) = \frac{\Gamma\left[\frac{1}{2}(n-1)\right]}{\Gamma\left[\frac{1}{2}(n-2)\right]\sqrt{\pi}}(1-\hat{\rho}^2)^{(n-4)/2}$$

Considerando o tamanho da amostra $n = 29$ e substituindo na expressão acima tem-se:

$$f(\hat{\rho}) = \frac{\Gamma\left[\frac{1}{2}(28)\right]}{\Gamma\left[\frac{1}{2}(27)\right]\sqrt{\pi}}(1-\hat{\rho}^2)^{25/2} = \frac{\Gamma[14]}{\Gamma\left[\frac{1}{2}(27)\right]\sqrt{\pi}}(1-\hat{\rho}^2)^{25/2} = 2{,}0563864(1-\hat{\rho}^2)^{25/2}$$

Substituindo valores para $\hat{\rho}$, obtém-se os correspondentes para $f(\hat{\rho})$. Alguns valores são apresentados na tabela a seguir. Para a construção do gráfico 7, o intervalo utilizado para $\hat{\rho}$ foi de 0,05, iniciando em -1,0.

TABELA A.1.3 - COEFICIENTE DE CORRELAÇÃO LINEAR DE PEARSON ($\hat{\rho}$) E RESPECTIVO VALOR DE $f(\hat{\rho})$

$\hat{\rho}$	$f(\hat{\rho})$
-1,0	0,00000
-0,8	0,00001
-0,6	0,00776
-0,4	0,23231
-0,2	1,23300
-0,0	2,05386
0,2	1,23300
0,4	0,23231
0,6	0,00776
0,8	0,00001
1,0	0,00000

FONTE: A autora

APÊNDICE 2 - DISTRIBUIÇÕES AMOSTRAIS DE Z

DISTRIBUIÇÕES AMOSTRAIS DE Z

Conforme apresentado na seção 3.2.1.7, a função densidade de Z, para $n > 25$ é :

$$f(Z) = \frac{1}{\hat{\sigma}_Z \sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{Z-E(Z)}{\hat{\sigma}_Z}\right)^2}$$

com:

$$E(Z) = \frac{1}{2} \ln\left[\frac{1+\rho}{1-\rho}\right] + \frac{\rho}{2n-1} \quad \text{e} \quad \hat{\sigma}_Z = \frac{1}{\sqrt{n-3}}$$

Assim, para amostra de tamanho $n = 29$ e $\rho = 0$ tem-se:

$$E(Z) = \frac{1}{2} \ln\left[\frac{1}{1}\right] + \frac{0}{57} = 0$$

$$\hat{\sigma}_Z^2 = \frac{1}{29-3} \quad \text{e} \quad \hat{\sigma}_Z = 0{,}1961$$

Portanto, f(Z) será:

$$f(z) = \frac{1}{0{,}1961\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{Z}{0{,}1961}\right)^2} = 2{,}0344 e^{\frac{-1}{2}\left(\frac{Z}{0{,}1961}\right)^2}$$

A tabela a seguir mostra alguns valores de z e os correspondentes f(z). Para a construção do gráfico 8, utilizou-se intervalo para Z igual a 0,05, iniciando-se em -1,0.

TABELA A.2.1 - VARIÁVEL Z E RESPECTIVO VALOR DE f(Z)

Z	f(Z)
-1,0	0,00000
-0,8	0,00049
-0,6	0,01886
-0,4	0,25407
-0,2	1,20939
0,0	2,03340
0,2	1,20939
0,4	0,25407
0,6	0,01886
0,8	0,00049
1,0	0,00000

FONTE: A autora
NOTA: Z é a transformação de Fisher.

Para amostra de tamanho $n = 29$ e $\rho = 0,80$, tem-se:

$$E(Z) = \frac{1}{2}\ln\left[\frac{1+0,8}{1-0,8}\right] + \frac{0,8}{57} = 1,1126$$

$$\hat{\sigma}_Z^2 = \frac{1}{29-3} \qquad \text{e} \qquad \hat{\sigma}_Z = 0,1961$$

Portanto, $f(Z)$ será:

$$f(Z) = \frac{1}{0,1961\sqrt{\pi}} e^{\frac{-1}{2}\left(\frac{Z-1,1126}{0,1961}\right)^2} = 2,0344e^{\frac{-1}{2}\left(\frac{Z-1,1126}{0,1961}\right)^2}$$

Alguns valores de $f(Z)$, para cada valor de Z são apresentados na tabela a seguir. Para a construção do gráfico 9, utilizou-se o valor inicial para Z igual a 0,00, e o intervalo de 0,05.

TABELA A.2.2 - VARIÁVEL Z E RESPECTIVO VALOR DE $f(Z)$

Z	f(Z)
0,0	0,00000
0,2	0,00004
0,4	0,00276
0,6	0,06679
0,8	0,57101
1,0	1,72521
1,2	1,84205
1,4	0,69506
1,6	0,09268
1,8	0,00437
2,0	0,00007
2,2	0,00000

FONTE: A autora
NOTA: Z é a transformação de Fisher.

APÊNDICE 3 - TESTE DE NORMALIDADE

1 AMOSTRA ALEATÓRIA GERADA PELO PROCESSO DE SIMULAÇÃO

O quadro abaixo apresenta a amostra aleatória de 200 observações gerada através do processo de simulação.

QUADRO A.3.1 - VARIÁVEL ALEATÓRIA X GERADA PELO PROCESSO DE SIMULAÇÃO

ORDEM	VARIÁVEL X	ORDEM	VARIÁVEL X	ORDEM	VARIÁVEL X	ORDEM	VARIÁVEL X	ORDEM	VARIÁVEL X
1	72,18700	41	8,12555	81	212,51855	121	147,14338	161	164,82585
2	17,44974	42	41,55470	82	169,85026	122	97,03626	162	183,53943
3	102,67841	43	3,79202	83	162,01862	123	152,89828	163	102,36524
4	160,48252	44	68,67889	84	95,32578	124	80,69527	164	97,12284
5	156,10761	45	87,18117	85	89,10925	125	45,33027	165	192,85433
6	186,05545	46	90,57455	86	114,40728	126	149,25564	166	86,61525
7	-33,28091	47	70,92790	87	87,57117	127	56,41797	167	52,21369
8	75,88585	48	211,89209	88	29,93820	128	118,89907	168	139,81303
9	150,32126	49	-8,57903	89	-10,38914	129	119,56322	169	88,44523
10	28,14476	50	47,75729	90	135,38656	130	71,27952	170	147,19482
11	50,34857	51	-55,34452	91	113,87657	131	64,31710	171	113,34344
12	-5,66421	52	170,06952	92	123,60274	132	12,50440	172	170,38835
13	-14,42701	53	17,33324	93	100,95450	133	200,60562	173	82,21271
14	34,25275	54	52,39952	94	31,49187	134	57,27668	174	35,14380
15	45,68360	55	131,43197	95	158,33893	135	93,82323	175	2,83909
16	-29,60415	56	115,13586	96	71,57206	136	75,88139	176	-54,64370
17	57,19621	57	137,97809	97	41,96438	137	247,78060	177	94,77852
18	66,37334	58	122,36154	98	43,01682	138	159,11080	178	105,75475
19	96,55177	59	12,17640	99	64,97641	139	138,33079	179	88,00390
20	68,53239	60	26,51864	100	63,61176	140	163,60119	180	100,22796
21	70,68852	61	127,86369	101	59,66748	141	99,52077	181	103,46104
22	68,26653	62	107,06764	102	136,56805	142	119,34963	182	171,66572
23	164,18793	63	36,36909	103	117,73961	143	75,02128	183	93,77593
24	84,22407	64	75,50692	104	54,93497	144	20,61910	184	26,46274
25	78,57517	65	96,36600	105	162,07885	145	159,93437	185	11,16490
26	60,26039	66	120,23667	106	-9,61244	146	72,82231	186	99,72138
27	199,44387	67	96,76804	107	119,83202	147	15,87099	187	142,26056
28	137,47769	68	37,98617	108	82,48876	148	131,79823	188	4,50761
29	222,03666	69	194,55137	109	91,33751	149	132,91981	189	54,84214
30	52,32523	70	116,28309	110	52,37297	150	112,97667	190	118,31782
31	182,04153	71	93,04538	111	58,22274	151	111,61737	191	76,86138
32	-1,29427	72	135,47110	112	136,56193	152	52,70658	192	69,75953
33	119,18111	73	137,27243	113	133,97053	153	129,49351	193	79,84534
34	139,52272	74	53,35424	114	114,55931	154	118,41611	194	64,37847
35	196,45927	75	37,30127	115	127,71663	155	149,20216	195	44,05315
36	84,26704	76	151,22657	116	180,31494	156	-66,08907	196	125,10860
37	59,66748	77	21,73399	117	106,01921	157	114,97364	197	92,29407
38	126,80775	78	1,70204	118	121,98375	158	171,15971	198	159,75330
39	67,64586	79	128,83419	119	192,75883	159	-7,46465	199	103,68259
40	131,42624	80	124,75075	120	70,20881	160	91,54767	200	92,19966

FONTE: A autora

NOTA: Média = 92,84155 e Desvio Padrão=57,98319

2 TESTE DE NORMALIDADE

O método utilizado para testar a normalidade da variável aleatória X foi o proposto por FILLIBEN (1975). O quadro A.3.2 apresenta as estatísticas da variável aleatória X necessárias para a aplicação do método proposto.

QUADRO A.3.2 - ESTATÍSTICAS DA VARIÁVEL ALEATÓRIA X

continua

ORDEM	X_i ORDENADA	m_i	M_i	$(X - \bar{X})$	$(X - \bar{X}) M_i$	$(X - \bar{X})^2$	M_i^2
1	- 66,08907	0,00346	-2,70067	-158,93062	429,21915	25 258,94075	7,29362
2	- 55,34452	0,00840	-2,39106	-148,18607	354,32178	21 959,11020	5,71717
3	- 54,64370	0,01339	-2,21471	-147,48525	326,63705	21 751,89783	4,90494
4	- 33,28091	0,01838	-2,08842	-126,12246	263,39666	15 906,87395	4,36150
5	- 29,60415	0,02337	-1,98865	-122,44570	243,50163	14 992,94851	3,95473
6	- 14,42701	0,02836	-1,90547	-107,26856	204,39702	11 506,54314	3,63082
7	- 10,38914	0,03335	-1,83369	-103,23069	189,29308	10 656,57456	3,36242
8	- 9,61244	0,03834	-1,77029	-102,45399	181,37327	10 496,81928	3,13393
9	- 8,57903	0,04333	-1,71329	-101,42058	173,76286	10 286,13327	2,93536
10	- 7,46465	0,04832	-1,66137	-100,30620	166,64571	10 061,33299	2,76015
11	- 5,66421	0,05332	-1,61348	-98,50576	158,93707	9 703,38399	2,60332
12	- 1,29427	0,05831	-1,56912	-94,13582	147,71039	8 861,55188	2,46214
13	1,70204	0,06330	-1,52765	-91,13951	139,22927	8 306,40958	2,33371
14	2,83909	0,06829	-1,48865	-90,00246	133,98216	8 100,44211	2,21608
15	3,79202	0,07328	-1,45179	-89,04953	129,28121	7 929,81811	2,10769
16	4,50761	0,07827	-1,41681	-88,33394	125,15240	7 802,88428	2,00735
17	8,12555	0,08326	-1,38348	-84,71600	117,20289	7 176,80000	1,91402
18	11,16490	0,08825	-1,35161	-81,67665	110,39497	6 671,07453	1,82685
19	12,17640	0,09324	-1,32107	-80,66515	106,56430	6 506,86580	1,74523
20	12,50440	0,09823	-1,29171	-80,33715	103,77230	6 454,05705	1,66851
21	15,87099	0,10322	-1,26342	-76,97056	97,24614	5 924,46651	1,59623
22	17,33324	0,10822	-1,23605	-75,50831	93,33204	5 701,50430	1,52782
23	17,44974	0,11321	-1,20964	-75,39181	91,19694	5 683,92443	1,46323
24	20,61910	0,11820	-1,18404	-72,22245	85,51427	5 216,08173	1,40195
25	21,73399	0,12319	-1,15919	-71,10756	82,42717	5 056,28454	1,34372
26	26,46274	0,12818	-1,13504	-66,37881	75,34260	4 406,14591	1,28832
27	26,51864	0,13317	-1,11153	-66,32291	73,71990	4 398,72788	1,23550
28	28,14476	0,13816	-1,08863	-64,69679	70,43086	4 185,67414	1,18512
29	29,93820	0,14315	-1,06628	-62,90335	67,07258	3 956,83096	1,13695
30	31,49187	0,14814	-1,04445	-61,34968	64,07667	3 763,78276	1,09088
31	34,25275	0,15313	-1,02310	-58,58880	59,94220	3 432,64703	1,04673
32	35,14380	0,15812	-1,00222	-57,69775	57,82584	3 329,02991	1,00444
33	36,36909	0,16311	-0,98176	-56,47246	55,44240	3 189,13830	0,96385
34	37,30127	0,16811	-0,96166	-55,54028	53,41086	3 084,72227	0,92479
35	37,98617	0,17310	-0,94199	-54,85538	51,67322	3 009,11229	0,88735
36	41,55470	0,17809	-0,92267	-51,28685	47,32083	2 630,34059	0,85132
37	41,96438	0,18308	-0,90369	-50,87717	45,97719	2 588,48604	0,81666
38	43,01682	0,18807	-0,88503	-49,82473	44,09638	2 482,50334	0,78328
39	44,05315	0,19306	-0,86668	-48,78840	42,28393	2 380,30760	0,75113
40	45,33027	0,19805	-0,84861	-47,51128	40,31854	2 257,32136	0,72014
41	45,68360	0,20304	-0,83081	-47,15795	39,17929	2 223,87189	0,69025
42	47,75729	0,20803	-0,81328	-45,08426	36,66612	2 032,59015	0,66142
43	50,34857	0,21302	-0,79599	-42,49298	33,82398	1 805,65302	0,63360
44	52,21369	0,21801	-0,77893	-40,62786	31,64626	1 650,62270	0,60673

QUADRO A.3.2 - ESTATÍSTICAS DA VARIÁVEL ALEATÓRIA X

continua

ORDEM	X_i ORDENADA	m_i	M_i	$(x - \bar{x})$	$(x - \bar{x}) M_i$	$(x - \bar{x})^2$	M_i^2
45	52,32523	0,22301	-0,76207	-40,51632	30,87627	1 641,57187	0,58075
46	52,37297	0,22800	-0,74545	-40,46858	30,16730	1 637,70566	0,55570
47	52,39952	0,23299	-0,72904	-40,44203	29,48385	1 635,55748	0,53150
48	52,70658	0,23798	-0,71282	-40,13497	28,60901	1 610,81551	0,50811
49	53,35424	0,24297	-0,69678	-39,48731	27,51397	1 559,24735	0,48550
50	54,84214	0,24796	-0,68093	-37,99941	25,87494	1 443,95487	0,46367
51	54,93497	0,25295	-0,66524	-37,90658	25,21697	1 436,90852	0,44254
52	56,41797	0,25794	-0,64971	-36,42358	23,66476	1 326,67690	0,42212
53	57,19621	0,26293	-0,63434	-35,64534	22,61126	1 270,58999	0,40239
54	57,27668	0,26792	-0,61912	-35,56487	22,01892	1 264,85970	0,38331
55	58,22274	0,27291	-0,60404	-34,61881	20,91114	1 198,46174	0,36486
56	59,66748	0,27791	-0,58906	-33,17407	19,54152	1 100,51866	0,34699
57	59,66748	0,28290	-0,57425	-33,17407	19,05021	1 100,51866	0,32976
58	60,26039	0,28789	-0,55956	-32,58116	18,23111	1 061,53174	0,31311
59	63,61176	0,29288	-0,54499	-29,22979	15,92994	854,38040	0,29701
60	64,31710	0,29787	-0,53054	-28,52445	15,13336	813,64403	0,28147
61	64,37847	0,30286	-0,51619	-28,46308	14,69236	810,14670	0,26645
62	64,97641	0,30785	-0,50196	-27,86514	13,98718	776,46581	0,25196
63	66,37334	0,31284	-0,48782	-26,46821	12,91172	700,56594	0,23797
64	67,64586	0,31783	-0,47378	-25,19569	11,93721	634,82260	0,22447
65	68,26653	0,32282	-0,45983	-24,57502	11,30033	603,93142	0,21144
66	68,53239	0,32781	-0,44597	-24,30916	10,84115	590,93507	0,19889
67	68,67889	0,33281	-0,43217	-24,16266	10,44238	583,83395	0,18677
68	69,75953	0,33780	-0,41848	-23,08202	9,65936	532,77947	0,17513
69	70,20881	0,34279	-0,40486	-22,63274	9,16309	512,24075	0,16391
70	70,68852	0,34778	-0,39132	-22,15303	8,66892	490,75657	0,15313
71	70,92790	0,35277	-0,37785	-21,91365	8,28007	480,20789	0,14277
72	71,27952	0,35776	-0,36445	-21,56203	7,85828	464,92097	0,13282
73	71,57206	0,36275	-0,35112	-21,26949	7,46814	452,39104	0,12329
74	72,18700	0,36774	-0,33785	-20,65455	6,97814	426,61028	0,11414
75	72,82231	0,37273	-0,32463	-20,01924	6,49884	400,76982	0,10538
76	75,02128	0,37772	-0,31148	-17,82027	5,55066	317,56189	0,09702
77	75,50692	0,38271	-0,29837	-17,33463	5,17213	300,48926	0,08902
78	75,88139	0,38770	-0,28532	-16,96016	4,83907	287,64690	0,08141
79	75,88585	0,39270	-0,27229	-16,95570	4,61687	287,49563	0,07414
80	76,86138	0,39769	-0,25933	-15,98017	4,14414	255,36571	0,06725
81	78,57517	0,40268	-0,24642	-14,26638	3,51552	203,52949	0,06072
82	79,84534	0,40767	-0,23354	-12,99621	3,03513	168,90137	0,05454
83	80,69527	0,41266	-0,22071	-12,14628	2,68080	147,53202	0,04871
84	82,21271	0,41765	-0,20791	-10,62884	2,20984	112,97216	0,04323
85	82,48876	0,42264	-0,19515	-10,35279	2,02035	107,18018	0,03808
86	84,22407	0,42763	-0,18241	-8,61748	1,57191	74,26090	0,03327
87	84,26704	0,43262	-0,16971	-8,57451	1,45518	73,52216	0,02880
88	86,61525	0,43761	-0,15703	-6,22630	0,97772	38,76676	0,02466
89	87,18117	0,44260	-0,14438	-5,66038	0,81725	32,03986	0,02085
90	87,57117	0,44760	-0,13173	-5,27038	0,69427	27,77686	0,01735
91	88,00390	0,45259	-0,11912	-4,83765	0,57626	23,40282	0,01419
92	88,44523	0,45758	-0,10653	-4,39632	0,46834	19,32760	0,01135
93	89,10925	0,46257	-0,09396	-3,73230	0,35069	13,93003	0,00883
94	90,57455	0,46756	-0,08141	-2,26700	0,18456	5,13927	0,00663
95	91,33751	0,47255	-0,06886	-1,50404	0,10357	2,26212	0,00474
96	91,54767	0,47754	-0,05633	-1,29388	0,07288	1,67412	0,00317
97	92,19966	0,48253	-0,04381	-0,64189	0,02812	0,41202	0,00192

QUADRO A.3.2 - ESTATÍSTICAS DA VARIÁVEL ALEATÓRIA X

continua

ORDEM	X_i ORDENADA	m_i	M_i	$(x - \bar{x})$	$(x - \bar{x}) M_i$	$(x - \bar{x})^2$	M_i^2
98	92,29407	0,48752	-0,03129	-0,54748	0,01713	0,29973	0,00098
99	93,04538	0,49251	-0,01878	0,20383	-0,00383	0,04155	0,00035
100	93,77593	0,49750	-0,00627	0,93438	-0,00586	0,87307	0,00004
101	93,82323	0,50250	0,00627	0,98168	0,00616	0,96370	0,00004
102	94,77852	0,50749	0,01878	1,93697	0,03638	3,75187	0,00035
103	95,32578	0,51248	0,03129	2,48423	0,07773	6,17142	0,00098
104	96,36600	0,51747	0,04381	3,52445	0,15441	12,42177	0,00192
105	96,55177	0,52246	0,05633	3,71022	0,20900	13,76576	0,00317
106	96,76804	0,52745	0,06886	3,92649	0,27038	15,41735	0,00474
107	97,03626	0,53244	0,08141	4,19471	0,34149	17,59562	0,00663
108	97,12284	0,53743	0,09396	4,28129	0,40227	18,32948	0,00883
109	99,52077	0,54242	0,10653	6,67922	0,71154	44,61203	0,01135
110	99,72138	0,54741	0,11912	6,87983	0,81953	47,33211	0,01419
111	100,22796	0,55240	0,13173	7,38641	0,97301	54,55911	0,01735
112	100,95450	0,55740	0,14438	8,11295	1,17135	65,82002	0,02085
113	102,36524	0,56239	0,15703	9,52369	1,49551	90,70074	0,02466
114	102,67841	0,56738	0,16971	9,83686	1,66941	96,76389	0,02880
115	103,46104	0,57237	0,18241	10,61949	1,93710	112,77365	0,03327
116	103,68259	0,57736	0,19515	10,84104	2,11563	117,52823	0,03808
117	105,75475	0,58235	0,20791	12,91320	2,68478	166,75083	0,04323
118	106,01921	0,58734	0,22071	13,17766	2,90844	173,65082	0,04871
119	107,06764	0,59233	0,23354	14,22609	3,32236	202,38175	0,05454
120	111,61737	0,59732	0,24642	18,77582	4,62674	352,53156	0,06072
121	112,97667	0,60231	0,25933	20,13512	5,22164	405,42321	0,06725
122	113,34344	0,60730	0,27229	20,50189	5,58246	420,32765	0,07414
123	113,87657	0,61230	0,28532	21,03502	6,00171	442,47223	0,08141
124	114,40728	0,61729	0,29837	21,56573	6,43457	465,08088	0,08902
125	114,55931	0,62228	0,31148	21,71776	6,76465	471,66127	0,09702
126	114,97364	0,62727	0,32463	22,13209	7,18474	489,82958	0,10538
127	115,13586	0,63226	0,33785	22,29431	7,53213	497,03643	0,11414
128	116,28309	0,63725	0,35112	23,44154	8,23079	549,50598	0,12329
129	117,73961	0,64224	0,36445	24,89806	9,07410	619,91358	0,13282
130	118,31782	0,64723	0,37785	25,47627	9,62621	649,04053	0,14277
131	118,41611	0,65222	0,39132	25,57456	10,00784	654,05832	0,15313
132	118,89907	0,65721	0,40486	26,05752	10,54965	678,99455	0,16391
133	119,18111	0,66220	0,41848	26,33956	11,02258	693,77262	0,17513
134	119,34963	0,66719	0,43217	26,50808	11,45600	702,67851	0,18677
135	119,56322	0,67219	0,44597	26,72167	11,91706	714,04785	0,19889
136	119,83202	0,67718	0,45983	26,99047	12,41103	728,48568	0,21144
137	120,23667	0,68217	0,47378	27,39512	12,97926	750,49281	0,22447
138	121,98375	0,68716	0,48782	29,14220	14,21615	849,26805	0,23797
139	122,36154	0,69215	0,50196	29,51999	14,81786	871,43004	0,25196
140	123,60274	0,69714	0,51619	30,76119	15,87862	946,25105	0,26645
141	124,75075	0,70213	0,53054	31,90920	16,92911	1 018,19729	0,28147
142	125,10860	0,70712	0,54499	32,26705	17,58522	1 041,16276	0,29701
143	126,80775	0,71211	0,55956	33,96620	19,00613	1 153,70300	0,31311
144	127,71663	0,71710	0,57425	34,87508	20,02702	1 216,27147	0,32976
145	127,86369	0,72209	0,58906	35,02214	20,63014	1 226,55056	0,34699
146	128,83419	0,72709	0,60404	35,99264	21,74100	1 295,47041	0,36486
147	129,49351	0,73208	0,61912	36,65196	22,69196	1 343,36645	0,38331
148	131,42624	0,73707	0,63434	38,58469	24,47581	1 488,77860	0,40239
149	131,43197	0,74206	0,64971	38,59042	25,07258	1 489,22081	0,42212
150	131,79823	0,74705	0,66524	38,95668	25,91554	1 517,62322	0,44254

QUADRO A.3.2 - ESTATÍSTICAS DA VARIÁVEL ALEATÓRIA X

continua

ORDEM	X_i ORDENADA	m_i	M_i	$(x - \bar{x})$	$(x - \bar{x}) M_i$	$(x - \bar{x})^2$	M_i^2
151	132,91891	0,75204	0,68093	40,07736	27,28988	1 606,19509	0,46367
152	133,97053	0,75703	0,69678	41,12898	28,65785	1 691,59331	0,48550
153	135,38656	0,76202	0,71282	42,54501	30,32694	1 810,07820	0,50811
154	135,47110	0,76701	0,72904	42,62955	31,07865	1 817,27886	0,53150
155	136,56193	0,77200	0,74545	43,72038	32,59136	1 911,47196	0,55570
156	136,56805	0,77699	0,76207	43,72650	33,32266	1 912,00714	0,58075
157	137,27243	0,78199	0,77893	44,43088	34,60855	1 974,10344	0,60673
158	137,47769	0,78698	0,79599	44,63614	35,52992	1 992,38534	0,63360
159	137,97809	0,79197	0,81328	45,13654	36,70865	2 037,30759	0,66142
160	138,33079	0,79696	0,83081	45,48924	37,79292	2 069,27131	0,69025
161	139,52272	0,80195	0,84861	46,68117	39,61411	2 179,13199	0,72014
162	139,81303	0,80694	0,86668	46,97148	40,70925	2 206,32030	0,75113
163	142,26056	0,81193	0,88503	49,41901	43,73731	2 442,23893	0,78328
164	147,14338	0,81692	0,90369	54,30183	49,07202	2 948,68916	0,81666
165	147,19482	0,82191	0,92267	54,35327	50,15014	2 954,27838	0,85132
166	149,20216	0,82690	0,94199	56,36061	53,09113	3 176,51879	0,88735
167	149,25564	0,83189	0,96166	56,41409	54,25118	3 182,54998	0,92479
168	150,32126	0,83689	0,98176	57,47971	56,43128	3 303,91750	0,96385
169	151,22657	0,84188	1,00222	58,38502	58,51464	3 408,81101	1,00444
170	152,89828	0,84687	1,02310	60,05673	61,44404	3 606,81128	1,04673
171	156,10761	0,85186	1,04445	63,26606	66,07824	4 002,59484	1,09088
172	158,33893	0,85685	1,06628	65,49738	69,83855	4 289,90729	1,13695
173	159,11080	0,86184	1,08863	66,26925	72,14270	4 391,61401	1,18512
174	159,75330	0,86683	1,11153	66,91175	74,37442	4 477,18280	1,23550
175	159,93437	0,87182	1,13504	67,09282	76,15304	4 501,44701	1,28832
176	160,48252	0,87681	1,15919	67,64097	78,40874	4 575,30134	1,34372
177	162,01862	0,88180	1,18404	69,17707	81,90842	4 785,46755	1,40195
178	162,07885	0,88679	1,20964	69,23730	83,75221	4 793,80424	1,46323
179	163,60119	0,89178	1,23605	70,75964	87,46246	5 006,92720	1,52782
180	164,18793	0,89678	1,26342	71,34638	90,14045	5 090,30649	1,59623
181	164,82585	0,90177	1,29171	71,98430	92,98285	5 181,74000	1,66851
182	169,85026	0,90676	1,32107	77,00871	101,73390	5 930,34201	1,74523
183	170,06952	0,91175	1,35161	77,22797	104,38210	5 964,15994	1,82685
184	170,38835	0,91674	1,38348	77,54680	107,28445	6 013,50679	1,91402
185	171,15971	0,92173	1,41681	78,31816	110,96196	6 133,73479	2,00735
186	171,66572	0,92672	1,45179	78,82417	114,43615	6 213,25038	2,10769
187	180,31494	0,93171	1,48865	87,47339	130,21727	7 651,59463	2,21608
188	182,04153	0,93670	1,52765	89,19998	136,26636	7 956,63712	2,33371
189	183,53943	0,94169	1,56912	90,69788	142,31586	8 226,10613	2,46214
190	186,05545	0,94668	1,61348	93,21390	150,39877	8 688,83187	2,60332
191	192,75883	0,95168	1,66137	99,91728	165,99958	9 983,46361	2,76015
192	192,85433	0,95667	1,71329	100,01278	171,35090	10 002,55693	2,93536
193	194,55137	0,96166	1,77029	101,70982	180,05588	10 344,88827	3,13393
194	196,45927	0,96665	1,83369	103,61772	190,00278	10 736,63270	3,36242
195	199,44387	0,97164	1,90547	106,60232	203,12753	11 364,05545	3,63082
196	200,60562	0,97663	1,98865	107,76407	214,30503	11 613,09561	3,95473
197	211,89209	0,98162	2,08842	119,05054	248,62754	14 173,03199	4,36150
198	212,51855	0,98661	2,21471	119,67700	265,04986	14 322,58525	4,90494
199	222,03666	0,99160	2,39106	129,19511	308,91327	16 691,37744	5,71717
200	247,78060	0,99654	2,70067	154,93905	418,43925	24 006,11041	7,29362
TOTAL			0,00000		11 424,30554	669 048,10709	195,55906

FONTE: A autora

As hipóteses a serem testadas:

H_0 : A variável aleatória X é normalmente distribuída

H_1 : A variável aleatória X não é normalmente distribuída

O coeficiente de correlação é calculado através da seguinte expressão:

$$\hat{\rho} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \sum_{i=1}^{n} M_i^2}}$$, pois $\overline{M} = \sum_{i=1}^{n} M_i = 0$ (ver total da 4.ª coluna, do quadro acima)

Tem-se que: $\sum_{i=1}^{n}\left(X_i - \overline{X}\right)M_i = 11.424{,}30554$; $\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = 669.048{,}10709$;

$\sum_{i=1}^{n} M_i^2 = 195{,}55906$

Substituindo-se os valores na expressão acima, obtém-se o valor $\hat{\rho} = 0{,}99876$, superior ao valor crítico igual a $0{,}98700$ (Quadro A.2.1 do Anexo 2), para nível de significância de 5%. Portanto, aceita-se a hipótese H_0 de que a variável aleatória X é normalmente distribuída.

APÊNDICE 4 - APLICAÇÃO DO COEFICIENTE DE CORRELAÇÃO PONTO BISSERIAL

1 ARQUIVO DE DADOS

O arquivo de dados utilizado para os cálculos é apresentado no quadro a seguir.

QUADRO A.4.1 - RENDA DAS PESSOAS OCUPADAS, SEGUNDO GÊNERO, NA RMC - AGO 2003

continua

OBS.	RENDA (R$ 1,00)	GÊNERO	OBS.	RENDA (R$ 1,00)	GÊNERO
1	350	1	64	450	1
2	600	1	65	2 900	1
3	900	1	66	600	1
4	2 300	1	67	1 000	1
5	500	1	68	1 500	1
6	350	1	69	320	0
7	800	1	70	530	0
8	4 000	1	71	400	0
9	1 500	1	72	360	0
10	1 000	1	73	1 000	0
11	350	1	74	400	0
12	700	1	75	400	0
13	1 000	1	76	800	0
14	800	1	77	650	0
15	400	1	78	1 000	0
16	500	1	79	900	0
17	800	1	80	240	0
18	380	1	81	400	0
19	591	1	82	300	0
20	3 000	1	83	500	0
21	900	1	84	300	0
22	600	1	85	1 000	0
23	2 300	1	86	900	0
24	900	1	87	2 200	0
25	2 500	1	88	900	0
26	400	1	89	400	0
27	1 400	1	90	3 000	0
28	2 000	1	91	700	0
29	2 200	1	92	450	0
30	1 500	1	93	330	0
31	1 000	1	94	2 000	0
32	1 500	1	95	1 200	0
33	500	1	96	480	0
34	1 700	1	97	500	0
35	1 800	1	98	1 700	0
36	4 000	1	99	250	0
37	2 500	1	100	590	0
38	1 500	1	101	1 600	0
39	680	1	102	800	0
40	400	1	103	600	0
41	500	1	104	1 900	0
42	470	1	105	500	0
43	1 200	1	106	4 000	0
44	900	1	107	350	0
45	1 000	1	108	900	0
46	3 500	1	109	900	0
47	500	1	110	240	0

QUADRO A.4.1 - RENDA DAS PESSOAS OCUPADAS, SEGUNDO GÊNERO, NA RMC - AGO 2003

conclusão

OBS.	RENDA (R$ 1,00)	GÊNERO	OBS.	RENDA (R$ 1,00)	GÊNERO
48	350	1	111	1 500	0
49	1 200	1	112	2 000	0
50	2 400	1	113	1 200	0
51	800	1	114	2 100	0
52	600	1	115	300	0
53	3 000	1	116	800	0
54	520	1	117	1 500	0
55	800	1	118	740	0
56	400	1	119	900	0
57	600	1	120	800	0
58	1 200	1	121	600	0
59	350	1	122	340	0
60	1 300	1	123	280	0
61	1 000	1	124	860	0
62	1 500	1	125	600	0
63	810	1			

FONTE: PME-IPARDES/IBGE

NOTAS: Pessoas ocupadas na condição de empregados com carteira de trabalho assinada no setor privado, no grupo de atividades relativa a intermediação financeira e ativiades imobiliárias, aluguéis e serviços prestados às empresas, com 11 anos ou mais de estudo e que trabalharam entre 35 e 45 horas, na semana de referência.

A variável gênero assume os valores 0 e 1, sendo: 1= masculino; 0=feminino.

2 ESTATÍSTICAS DESCRITIVAS DA VARIÁVEL RENDA

TABELA A.4.1 - ESTATÍSTICAS DESCRITIVAS DA RENDA DAS PESSOAS OCUPADAS SEGUNDO GÊNERO E TOTAL NA RMC - AGOSTO 2003

ESTATÍSTICAS DESCRITIVAS	GÊNERO		TOTAL
	Homem	Mulher	
Tamanho da amostra	68,00	57,00	125,00
Mínimo (R$ 1,00)	350,00	240,00	240,00
Máximo (R$ 1,00)	4 000,00	4 000,00	4 000,00
Média (R$ 1,00)	1 212,51	901,93	1 070,89
Mediana (R$ 1,00)	900,00	700,00	800,00
Desvio Padrão (R$ 1,00)	910,19	729,73	843,55

FONTE: PME – IPARDES/IBGE

NOTAS: Pessoas ocupadas na condição de empregados com carteira de trabalho assinada no setor privado, no grupo de atividades relativa a intermediação financeira e ativiades imobiliárias, aluguéis e serviços prestados às empresas, com 11 anos ou mais de estudo e que trabalharam entre 35 e 45 horas, na semana de referência.

3 TESTE DE NORMALIDADE DA VARIÁVEL RENDA

H_0 : a variável renda provém de uma distribuição normal

H_1 : a variável renda não provém de uma distribuição normal

Estatísticas de Kolmogorov:

DN = 0,221489

Valor-p aproximado = 0,00000943433

Conclusão: Sendo o valor-p menor que 0,05, podemos rejeitar H_0 e concluir que a distribuição da variável em estudo não provém de uma distribuição normal.

4 TRANSFORMAÇÃO DA VARIÁVEL RENDA

Tendo em vista que a variável renda não é normalmente distribuída, fez-se uma transformação logarítmica (base e) na variável, e testou-se a hipótese da normalidade.

H_0 : a variável ln renda provém de uma distribuição normal

H_1 : a variável ln renda não provém de uma distribuição normal

Estatísticas de Kolmogorov:

DN = 0,086597

Valor-p aproximado = 0,307337

Conclusão: Sendo o valor-p maior que 0,05, pode-se aceitar H_0 e concluir que a distribuição da variável em estudo provém de uma distribuição normal.

5 CÁLCULO DOS COEFICIENTES DE CORRELAÇÃO

O Coeficiente Linear de Pearson foi obtido utilizando-se a *Procedure Correlation* (PROC CORR) disponível no Statistical Software Analysis (SAS) e o Coeficiente de Correlação Ponto Bissserial, utilizando-se o programa que se encontra no Apêndice 6.

QUADRO A.4.2 - COEFICIENTES DE CORRELAÇÃO PONTO BISSERIAL E LINEAR DE PEARSON ENTRE AS VARIÁVEIS EM ESTUDO

VARIÁVEIS	COEFICIENTE DE CORRELAÇÃO PONTO BISSERIAL		COEFICIENTE DE CORRELAÇÃO LINEAR DE PEARSON	
	$\hat{\rho}_{pb}$	Significância	$\hat{\rho}$	Significância
Renda e gênero	0, 18412	0,04	0,18412	0,04
Ln de renda e gênero	0,21544	0,02	0,21544	0,02

FONTE: PME - IPARDES/IBGE

NOTAS: Pessoas ocupadas na condição de empregados com carteira de trabalho assinada no setor privado, no grupo de atividades relativa a intermediação financeira e ativiades imobiliárias, aluguéis e serviços prestados às empresas, com 11 anos ou mais de estudo e que trabalharam entre 35 e 45 horas, na semana de referência.

Observa-se no quadro acima que as estimativas dos dois coeficientes de correlação são exatamente iguais, o que é correto, pois trata-se do mesmo coeficiente.

APÊNDICE 5 - CÁLCULO DOS COEFICIENTES DE CORRELAÇÃO DE SPEARMAN E POR POSTOS DE KENDALL

1 CÁLCULO DOS COEFICIENTES DE CORRELAÇÃO DE SPEARMAN E POR POSTOS DE KENDALL

TABELA A.5.1 - POPULAÇÃO MIGRANTE TOTAL E ECONOMICAMENTE ATIVA NAS ATIVIDADES URBANAS, POSTO DE X E Y, ELEMENTOS SUPERIORES E INFERIORES E S

MICRORREGIÕES	POPULAÇÃO MIGRANTE TOTAL (X)	POPULAÇÃO ECONOMICAMENTE ATIVA (Y)	POSTO DE X (1)	POSTO DE Y (2)	ELEMENT0S SUPERIORES (3)	ELEMENTOS INFERIORES (4)	S (5)
704	137	803	1	2	22	1	21
703	250	690	2	1	22	0	22
708	613	2 434	3	3	21	0	21
710	623	2 455	4	4	20	0	20
707	750	3 304	5	6	18	1	17
720	1 387	2 482	6	5	18	0	18
705	1 845	10 792	7	8	16	1	15
702	2 448	21 064	8	13	11	5	6
709	3 580	11 085	9	9	14	1	13
723	3 637	17 125	10	12	11	3	8
724	6 268	14 318	11	11	11	2	9
713	7 172	9 219	12	7	12	0	12
711	7 401	13 957	13	10	11	0	11
706	14 796	48 967	14	22	2	8	-6
719	26 437	29 485	15	15	8	1	7
722	27 713	23 832	16	14	8	0	8
712	28 528	45 664	17	19	4	3	1
717	32 740	34 848	18	16	6	0	6
716	36 216	37 141	19	17	5	0	5
715	39 501	47 809	20	20	3	1	2
721	40 978	48 198	21	21	2	1	1
701	42 116	226 657	22	24	0	2	-2
718	45 510	42 589	23	18	1	0	1
714	86 938	111 618	24	23	0	0	0
TOTAL							216

FONTE DOS DADOS: Menezes, Faissol e Ferreira (1978)

NOTAS: População migrante total de destino urbano e origem rural. As colunas (1), (2), (3), (4) e (5) foram elaboradas pela autora.

O Coeficiente de Correlação de Spearman calculado a partir dos postos foi $\rho_S = 0{,}922609$. Este coeficiente é o Coeficiente de Correlação Linear de Pearson.

Coeficiente de Correlação por Postos de Kendall obtido foi:

$$\tau = \frac{216}{0{,}5 \times 24 \times 23} = 0{,}782609$$

APÊNDICE 6 - PROGRAMAS UTILIZADOS

1 PROGRAMA PARA GERAR AMOSTRA NORMAL BIVARIADA

O Programa que deu origem a este, foi obtido no site:

www.sas.com_>service and support_>tecnical support_

```
data a500;
keep x y;
m1=5; m2=20; v1=2; v2=10; ro=0.80;
do i=1 to 500; /* tamanho da amostra */
x=m1+sqrt(v1)*rannor(123);
y=(m2+ro*(sqrt(v2)/sqrt(v1))*(x-m1))+ sqrt(v2*(1-ro**2))*rannor(123);
output;
end;
run;
```

2 PROGRAMAS PARA OS CÁLCULOS DOS COEFICIENTES DE CORRELAÇÃO

2.1 COEFICIENTE DE CORRELAÇÃO BISSERIAL

O Programa que deu origem a este, foi obtido no site:

www.sas.com_>service and support_>tecnical support_

```
*************************************************************
* ESTE PROGRAMA CALCULA O COEFICIENTE DE CORRELAÇÃO BISSERIAL  *
*************************************************************
data arq;
set dados;
if y>=116854 then dicoty=1 /* ponto de dicotomização*/;
else dicoty=0;
run;
```

```
* calcula a proporção da variável dicotômica,desvio padrão e n *;
proc means data=arq noprint;
var dicoty x;
output out=temp(keep=p stdx n) mean=p std=stdy stdx n=n;
run;

* ordena a variável dicotômica*;
proc sort data=arq;
by descending dicoty;
run;

*calcula a média da variável continua *;
proc means data=arq noprint;
by notsorted dicoty;
var x;
output out=out2 mean=m1;
run;

* organiza a média calculada acima *;
proc transpose data=out2 out=out3(rename=(col1=mx1 col2=mx0));
var m1;
run;

* calcula o coeficiente de correlação bisserial *;
data out4;
set out3(drop= _name_);
run;
```

```
*calcula o coeficiente bisserial *;
data out5;
merge  temp out4;
z=probit(1-p);
y=exp(-z*z/2)/sqrt(2*arcos(-1));
bis=p*(1-p)*(mx1-mx0)/stdx/y;
rbis=((((sqrt(p*(1-p))/y)-(bis*bis)))/sqrt(n));
run;

proc print data=out5;
title1 ' correlação bisserial';
var bis rbis p u mx1 mx0;
format bis rbis p u mx1 mx0 comma15.4;
run;
```

2.2 COEFICIENTE DE CORRELAÇÃO TETRACÓRICO

```
*********************************************************
* ESTE PROGRAMA CALCULA O COEFICIENTE DE CORRELAÇÃO *
* TETRACÓRICO  *
*********************************************************

* define o ponto de dicotomização*;
data arq;
set dados;
if y>=23.2831 then dicoty=1;
else dicoty=0;
if x>=6.98211 then dicotx=1;
else dicotx=0;
run;
```

```
proc freq  data=arq;
title1 'ponto de dicotomização: mediana';
tables dicotx*dicoty /  measures chisq plcorr converge=0.0001
maxiter=200;
run;
```

2.3 COEFICIENTE DE CORRELAÇÃO PONTO BISSERIAL

O Programa que deu origem a este, foi obtido no site:

www.sas.com_>service and support_>tecnical support_

```
*****************************************************************
* PROGRAMA PARA CALCULAR O COEFICIENTE DE CORRELAÇÃO PONTO *
*  BISSERIAL    *       *       *
*****************************************************************;

*define a variável dicotômica*;
data arq;
set dados;
dicoty=y;
x=x;
run;

* calcula a proporção da variável binária,
  desvio padrão da variável continua, e n *;

proc means data=arq noprint;
var dicoty x;
output out=temp(keep=p stdx n) mean=p std=stdy stdx n=n;
run;
```

```
* ordena a variável dicotômica *;
proc sort data=arq;
by descending dicoty;
run;

*calcula a média da variável continua *;
proc means data=arq noprint;
by notsorted dicoty;
var x;
output out=out2 mean=m1;
run;

* organiza o arquivo gerado acima*;
proc transpose data=out2 out=out3(rename=(col1=mx1 col2=mx0));
var m1;
run;

* calcula o coeficiente ponto bisserial *;
data out4;
set out3(drop= _name_);
run;
proc corr data=arq  noprint outp=temp1;
var dicoty x;
run;

* retira o coeficiente ponto bisserial da matriz *;
data temp2(keep=pbis);
set temp1(rename=(x=pbis));
if _TYPE_='CORR' and dicoty<>1 then output;
run;
```

```
*calculo do coeficiente de correlação ponto bisserial *;
data out5;
merge temp2 temp out4;
if pbis=1 then delete;
rpbis=sqrt(((1-(pbis*pbis)))/(n-2));
keep mx1 mx0 p pbis rpbis;
run;

proc print data=out5;
title1 'correlação ponto bisserial';
var pbis rpbis p mx1 mx0;
format  pbis rpbis p mx1 mx0 comma10.6;
run;
```

ANEXO 1 - CO-RELATIONS AND THEIR MEASUREMENT, CHIEFLY FROM ANTHROPOMETRIC DATA

CO-RELATIONS AND THEIR MEASUREMENT, CHIEFLY FROM ANTHROPOMETRIC DATA

By FRANCIS GALTON, F.R.S.
Received December 5, 1888.
[Proceedings of the Royal Society of London 45 (1888), 135-145.]

"Co-relation or correlation of structure" is a phrase much used in biology, and not least in that branch of it which refers to heredity, and the idea is even more frequently present than the phrase; but 1 am not aware of any previous attempt to define it clearly, to trace its mode of action in detail, or to show how to measure its degree.

Two variable organs are said to be co-related when the variation of the one is accompanied on the average by more or less variation of the other, and in the same direction. Thus the length of the arm is said to be correlated with that of the leg, because a person with a long arm has usually a long log, and conversely. If the correlation be close, then a person with a very long arm would usually have a very long leg; if it be moderately close, then the length of the leg would usually be only long, not very long; and if there were no correlation at all then the length of the leg would on the average be mediocre. It is easy to see that correlation must be the consequence of the variations of the two organs being partly due to common causes. If they were wholly due to common causes, the correlation would be perfect, as is approximately the case with the symmetrically disposed parts of the body. If they were in no respect due to common causes, the co-relation would be *nil*. Between these two extremes are an number of intermediate cases, and it will be shown how the closeness of correlation in any particular case admits of being expressed by a simple number.

To avoid the possibility of misconception, it is well to point out that the subject in hand has nothing whatever to do with the average proportions between the various limbs, in different races, which have been often discussed from early times up to the present day, both by artists and by anthropologists. The fact that the average ratio between the stature and the cubit is as 100 to 37, or thereabouts, does not give the slightest information about the nearness with which they vary together. It would be an altogether erroneous inference to suppose their average proportion to be maintained so that when the cubit might be expected to be one-twentieth longer than the average cubit, the stature might be expected to be one-twentieth greater than the average stature, and conversely. Such a supposition is easily shown to be contradicted both by fact and theory.

The relation between the cubit and the stature will be shown to be such that for every inch, centimetre, or other unit of absolute length that the cubit deviates from the mean length of cubits, cubits, the stature will on the average deviate from the mean length of statures to the amount of 2.5 units, and in the same direction. Conversely, for each unit of deviation of stature, the average deviation of the cubit will he 0.26 unit. These relations are not numerically reciprocal, but the exactness of the co-relation becomes established when we have transmuted the inches or other measurement of the cubit and of the stature into units dependent on their respective scales of variability. We thus cause a long cubit and an equally long stature, as compared to the general run of cubits and statures, to be designated by an

identical scale-value. The particular unit that I shall employ is the value of the probable error of any single measure in its own group. In that of the cubit, the probable error is 0.56 inch = 1.42 cm.; in the stature it is 1.75 inch = 4.44 cm. Therefore the measured lengths of the cubit in inches will be transmuted into terms of a new scale in which each unit = 0.56 inch, and the measured lengths of the stature will be transmuted into terms of another new scale in which each unit is 1.75 inch. After this has been done, we shall find the deviation of the cubit as compared to the mean of the corresponding deviations of the stature, to be as 1 to 0.8. Conversely, the deviation of the stature as compared to the mean of the corresponding deviations of the cubit will also be as I to 0.8. Thus the existence of the co-relation is established, and its measure is found to be 0.8.

Now as to the evidence of all this. The data were obtained at my anthropometric laboratory at South Kensington. They are of males of 21 years and upwards, but as a large proportion of them were students, and barely 21 years of age, they were not wholly full-grown; but neither that fact nor the small number of observations is prejudicial to the conclusions that will be reached. They were measured in various ways, partly for the purpose of this inquiry. It will be sufficient to give some of them as examples. The exact number of 350 is not preserved throughout, as injury to some limb or other reduced the available number by 1, 2, or 3 in different cases. After marshalling the measures of each limb in the order of their magnitudes, I noted the measures in each series that occupied the positions of the first, second and third quarterly divisions. Calling these measures in any one series Q_1, M and Q_3, I take M, which is the median or middlemost value, as that whence the deviations are to be measured, and $[1/2]\{\tilde{Q_3}Q_3\}=Q$ as the probable error of any single measure in the series. This is practically the same as saying that one-half of the deviations fall within the distance of $\square Q$ from the mean value, because the series run with fair symmetry. In this way I obtained the following values of M and Q, in which the second decimal must be taken as only roughly approximate. The M and Q of any particular series may be identified by a suffix, thus M_c, Q_c might stand for those of the cubit, and M_i, Q_i for those of the stature.

Table I.

	M		Q	
	Inch.	Cubit.	Inch.	Cubit.
Head length	7.62	19.35	0.19	0.48
Head breadth	6.00	15.24	0.18	0.46
Stature	67.20	170.69	1.75	4.44
Left middle finger	4.54	11.53	0.15	0.38
Left cubit	18.05	45.70	0.56	1.42
Height of right knee	20.50	52.00	0.80	2.03

NOTE.-The head length is its maximum length measured from the notch between and just below the eyebrows. The cubit is measured from the hand prone and without taking off the coat; it is the distance between the elbow of the bent left arm and the tip of the middle finger. The height of the knee is taken sitting when the knee is bent at right angles, less the measured thickness of the heel of the boot.

Tables were then constructed, each referring to a different pair the above elements, like Tables II and III, which will suffice as examples of the whole of them. It will be understood that the Q value is a universal unit applicable to the most varied measurements, such as breathing capacity, strength, memory, keenness of eyesight, and enables them to be compared together on equal terms

notwithstanding their intrinsic diversity. It does not only refer to measures of length, though partly for the sake of compactness, it is only those of length that will be here given as examples. It is unnecessary to extend the limits of Table II, as it includes every line and column in my MS table that contains not less than twenty entries. None of the entries lying within the flanking lines and columns of Table II were used.

Table II.

	Length of left cubit in inches, 348 adult males.								
Stature in		16.5	17.0	17.5	18.0	18.5	19.0		Total
inches.	Under	and	and	and	and	and	and	19.5	cases.
	16.5	under	under	under	under	under	under	and	
		17.0	17.5	18.0	18.5	19.0	19.5	above	
71 and above	..	..	..	1	3	4	15	7	30
70	..	..	..	1	5	13	11	..	30
69	..	1	1	2	25	15	6	..	50
68	..	1	3	7	14	7	4	2	48
67	..	1	7	15	28	8	2	..	61
66	..	1	7	18	15	6	..	..	48
65	..	4	10	12	8	2	..	..	36
64	..	5	11	2	3	..	..	..	21
Below 64	9	12	10	3	1	..	..	..	34
Totals	9	25	49	61	102	55	38	9	348

The measures were made and recorded to the nearest tenth of an inch. The heading of 70 inches of stature includes all records between 69.5 and 70.4 inches; that of 69 includes all between 68.5 and 69.4, and so on.

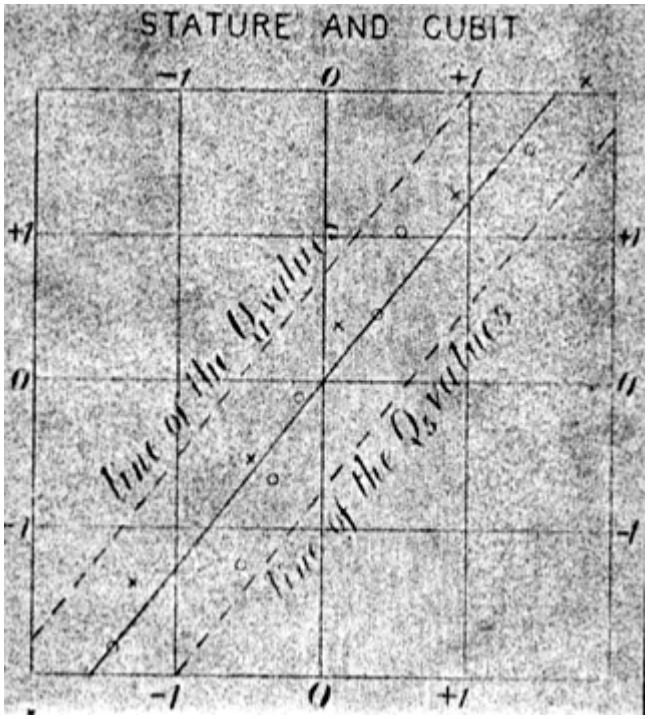


Table III.

Stature Ms = 67.2 inches; Qs = 1.75 inch. Left Cubit Mc = 18.05 inches; Qc = 0.56 inch.

No. of Cases	Stature.	Deviation from Ms reckoned in		Mean of Corresponding left cubits.	Deviation from Ms reckoned in			Smoothed Values Multiplied by Qc.	Added to Mc.
		Inches.	Units of $Q_{,}$.		Inches.	Units of Qc.			
					Observed.	Smoothed.			
	inches.			inches.					
30	70.0	+2.8	+1.60	18.8	+0.8	+1.42	+1.30	+0.73	18.8
50	69.0	+1.8	+1.03	18.3	+0.3	+0.53	+0.84	+0.47	18.5
38	68.0	+0.8	+0.46	18.2	+0.2	+0.36	+0.38	+0.21	18.3
61	67.0	-0.2	-0.11	18.1	+0.1	+0.18	-0.08	-0.04	18.0
48	66.0	-1.2	-0.69	17.8	-0.2	-0.36	-0.54	-0.30	17.8
36	65.0	-2.2	-1.25	17.7	-0.3	-0.53	-1.00	-0.56	17.5
21	64.0	-3.2	-1.83	17.2	-0.8	-1.46	-1.46	-0.80	17.2

No. of cases.	Left cubit.	Deviation from Mc reckoned in		Mean of corresponding statures.	Deviation from Ms, reckoned in			Smoothed values Multiplied by Qs.	Added to Ms
		Inches.	Units of Qc.		Inches	Units of Qs.			
					Observed.	Smoothed.			
	inches.			inches.					
38	19.25	+1.20	+2.14	70.3	+3.1	+1 8	+1.70	+3.0	70.2
55	18 75	+0.70	+1.25	68.7	+1.5	+0.9	+1.00	+1.8	69.0
102	18.25	+0.20	+0.36	67.4	+0.8	+0.1	+0.28	+0.5	67.7
61	17.75	-0.30	-0.53	86.3	-0.9	-0.5	-0.43	-0.8	66.4
98	17.25	-0.80	-1.43	66.0	-2.2	-1.3	-1.15	-2.0	65.2
26	18.75	-1.30	-2.31	63.7	-3.S	-2.0	-1.85	-3.2	64.0

The values derived from Table II, and from other similar tables, are entered in Table III, where they occupy all the columns up to the three last, the first of which is headed "smoothed." These smoothed values were obtained by plotting the observed values, after transmuting them as above described into their respective Q units, upon a diagram such as is shown in the figure. The deviations of the "subject" are measured parallel to the axis of y in the figure, and those of the mean of the corresponding values of the "relative" are measured parallel to the axis of x. When the stature is taken as the subject, the median positions of the corresponding cubits, which are given in the successive lines of Table III, are marked with small circles. When the cubit is the subject, the mean positions of the corresponding statures are marked with crosses. The firm line in the figure is drawn to represent the general run of the small circles and crosses. It is here seen to be a straight line, and it was similarly found to be straight in every other figure drawn from the different pairs of co-related variables that I have as yet tried. But the inclination of the line to the vertical differs considerably in different cases. In the present one the inclination is such that a deviation of 1 on the part of the subject, whether it be stature or cubit, is accompanied by a mean deviation on the part of the relative, whether it be cubit or stature, of 0.8. This decimal fraction is consequently the measure of the closeness of the correlation. We easily retransmute it into inches. If the stature be taken as the subject, then Q_s is associated with $Q_c \times 0.8$; that is, a deviation of 1.75 inches in the one with 0.56 ×0.8 of the other. This is the same as 1 inch of stature being associated with a mean length of cubit equal to 0.26 inch.

Conversely, if the cubit he taken as the subject, then Q_c is associated with $Q_s \times 0.8$; that is, a deviation of 0.56 inch in the one with 1.75×0.8 of the other. This is the same as 1 inch of cubit being associated with a mean length of 2.5 inches of stature. If centimetre be read for inch the same holds true. Six other tables are now given in a summary form, to show how well calculation on the above principle agrees with observation.

Table IV.

		Mean of corresponding				Mean of corresponding	
No.	Length	statures.		No.		lengths of head.	
of	of			of	Height		
cases.	head.			cases.			
		Observed.	Calculated.			Observed.	Calculated.
32	7.90	68.5	68.1	26	70.5	7.72	7.75
41	7.80	67.2	67.8	30	69.5	7.70	7.72
46	7.70	67.6	67.5	50	68.5	7.65	7.68
52	7.60	66.7	67.2	49	67.5	7.65	7.64
58	7.50	66.8	66.8	56	66.5	7.57	7.60
34	7.40	66.0	66.5	43	65.5	7.57	7.69
26	7.30	66.7	66.2	31	64.5	7.54	7.65
		Mean of corresponding			Length	Mean of corresponding	
No.		lengths of left		No.	of left	statures.	
of	Height.	middle finger.		of	middle		
cases.				cases.	finger.		
		Observed.	Calculated.			Observed.	Calculated.
30	70.5	4.71	4.74	23	4.80	70.2	69.4
50	69.5	4.55	4.68	49	4.70	68.1	68.5
37	68.5	4.57	4.62	62	4.60	68.0	67.7
62	67.5	4.58	4.56	63	4.50	67.3	66.9
48	66.5	4.59	4.50	57	4.40	66.0	66.1
37	65.5	4.47	4.44	35	4.30	65.7	65.3
20	64.5	4.33	4.38				
		Mean of corresponding				Mean of corresponding	
No.	Left	lengths of left cubit.		No.	Length	lengths of left middle	
of	middle			of	of left	finger.	
cases.	finger.			cases.	cubit.		
		Observed.	Calculated.			Observed.	Calculated.
23	4.80	18.97	18.80	29	19.00	4.76	4.75
50	4.70	18.55	18.49	32	18.70	4.64	4.69
62	4.60	18.24	18.18	48	18.40	4.60	4.62
62	4.50	18.00	17.87	70	18.10	4.56	4.55
57	4.40	17.72	17.55	37	17.80	4.49	4.48
34	4.30	17.27	17.24	31	17.50	4.40	4.41
				28	17.20	4.37	4.34
				24	16.90	4.32	4.28

		Mean of corresponding				Mean of corresponding	
No.	Length	breadths of head.		No.	Breadth	lengths of head.	
of	of			of	of		
cases.	head.			cases.	head.		
		Observed.	Calculated.			Observed.	Calculated.
32	7.90	6.14	6.12	27	6.30	7.72	7.84
41	7.80	6.05	6.08	36	6.20	7.72	7.75
46	7.70	6.14	6.04	53	6.10	7.65	7.65
52	7.60	5.98	6.00	58	6.00	7.68	7.60
34	7.40	5.96	5.91	37	5.80	7.55	7.50
26	7.30	5.85	5.87	30	5.70	7.45	7.46
		Mean of corresponding				Mean of corresponding	
No.		heights of knee.		No.	Height	statures.	
of	Stature.			of	of		
cases.				cases.	knee.		
		Observed.	Calculated.			Observed.	Calculated.
30	70.0	21.7	21.7	23	22.2	70.5	70.6
50	69.0	21.1	21.3	32	21.7	69.8	69.6
38	68.0	20.7	20.9	50	21.2	68.7	68.6
61	67.0	20.5	20.5	68	20.7	67.3	67.7
49	66.0	20.2	20.1	74	20.2	66.2	66.7
36	65.0	19.7	19.7	41	19.7	65.5	65.7
				26	19.2	64.3	64.7
		Mean of corresponding				Mean of corresponding	
No.		heights of knee.		No.	Height	left cubit.	
of	Left			of	of		
cases.	cubit.			cases.	knee.		
		Observed.	Calculated.			Observed.	Calculated.
29	19.0	21.5	21.6	23	22.25	18.98	18.97
32	18.7	21.4	21.2	30	21.75	18.68	18.70
48	18.4	20.8	20.9	52	21.25	18.38	18.44
70	17.1	20.7	20.6	69	20.75	18.15	18.17
37	17.8	20.4	20.2	70	20.25	17.75	17.90
31	17.5	20.0	19.9	41	19.75	17.55	17.63
28	17.2	19.8	19.6	27	19.25	17.02	17.36
23	16.9	19.3	19.2				

From Table IV the deductions given in Table V can be made; but they may be made directly from tables of the form of Table III, whence Table IV was itself derived.

Table V.

		In units of Q.		In units of ordinary measure.	
Subject.	Relative.	r.	$\sqrt{\{(1-\rho^2)\}}$ =ϕ.	As 1 to to	f.
Stature	Cubit	0.8	0.6	0.26	0.45
Cubit	Stature			2.5	1.4
Stature	Head length	0.35	0.93	0.38	1.63
Head length	Stature			3.2	0.17
Stature	Middle finger	0.7	0.72	0.06	0.10
Middle finger	Stature			8.2	1.26
Middle finger	Cubit	0.85	0.61	3.13	0.34
Cubit	Middle finger			0.21	0.09
Head length	Head breadth	0.45	0.89	0.43	0.16
Head breadth	Head length			0.48	0.17
Stature	Height of knee	0.9	0.44	0.41	0.35
Height of knee	Stature			1.20	0.77
Cubit	Height of knee	0.8	0.60	1.14	0.64
Height of knee	Cubit			0.56	0.45

When the deviations of the subject and those of the mean of the relatives are severally measured in units of their own Q, there is always a regression in the value of the latter. This is precisely analogous to what was observed in kinship, as I showed in my paper read before this Society on ``Hereditary Stature'' (`Roy. Soc. Proc.,' vol. 40, 1886, p. 42). The statures of kinsmen are co-related variables; thus, the stature of the father is correlated to that of the adult son, and the stature of the adult son to that of the father; the stature of the uncle to that of the adult nephew, and the stature of the adult nephew to that of the uncle, and so on; but the index of correlation which is what I there called ``regression,'' is different in the different cases. In dealing with kinships there is usually no need to reduce the measures to units of Q, because the Q values are alike in all the kinsmen, being of the same value as that of the population at large. It however happened that the very first case that I analysed was different in this respect. It was the reciprocal relation between the statures of what I called the ``mid-parent'' and the son. The mid-parent is an ideal progenitor, whose stature is the average of that of the father on the one hand and of that of the mother on the other, after her stature had been transmuted into its male equivalent by the multiplication of the factor of 1.08. The Q of the mid-parental stature was found to be 1.2, that of the population dealt with was 1.7. Again, the mean deviation measured in inches of the statures of the sons was found to be two-thirds of the deviation of the mid-parents, while the mean deviation in inches of the mid-parent was one-third of the deviation of the sons. Here the regression,

when calculated in Q units, is in the first case from [1/1.2] to [2/3]×1.7=1 to 0.47, and in the second case from [1/1.7] to [1/3]×[1/1.2]=1 to 0.44 which is practically the same.

The *rationale* of all this will be found discussed in the paper on ``Hereditary Stature,'' to which reference has already been made, and in the appendix to it by Mr. J. D. Hamilton Dickson. The entries in any table, such as Table II, may be looked upon as the values of the vertical ordinates to a surface of frequency, whose mathematical properties were discussed in the above-mentioned appendix, therefore I need not repeat them here. But there is always room for legitimate doubt whether conclusions based on the strict properties of the ideal law of error would be sufficiently correct to be serviceable in actual cases of correlation between variables that conform only approximately to that law. It is therefore exceedingly desirable to put the theoretical conclusions to frequent test, as has been done with these anthropometric data. The result is that anthropologists may now have much less hesitation than before, in availing themselves of the properties of the law of frequency of error.

I have given in Table V a column headed $\sqrt{\{(1-r^2)\}}$=f. The meaning of f is explained in the paper on ``Hereditary Stature.'' It is the Q value of the distribution of any system of x values, as x_1, x_2, x_3, &c., round the mean of all of them, which we may call X. The knowledge of f enables dotted lines to be drawn, as in the figure above, parallel to the line of M values, between which one half of the x observations, for each value of y, will be included. This value of f has much anthropological interest of its own, especially in connexion with M. Bertillon's system of anthropometric identification, to which I will not call attention now.

It is not necessary to extend the list of examples to show how to measure the degree in which one variable may be correlated with the combined effect of n other variables, whether these be themselves correlated or not. To do so, we begin by reducing each measure into others, each having the Q of its own system for a unit. We thus obtain a set of values that can be treated exactly in the same way as the measures of a single variable were treated in Tables II and onwards. Neither is it necessary to give examples of a method by which the degree may be measured, in which the variables in a series each member of which is the summed effect of n variables, may be modified by their partial correlation. After transmuting the separate measures as above, and then summing them, we should find the probable error of any one of them to be $\sqrt{n}$ if the variables were perfectly independent, and n if they were rigidly and perfectly co-related. The observed value would be almost always somewhere intermediate between these extremes, and would give that information that is wanted.

To conclude, the prominent characteristics of any two correlated variables, so far at least as I have as yet tested them, are four in number. It is supposed that their respective measures have been first transmuted into others of which the unit is in each case equal to the probable error of a since single measure in its own series. Let y=the deviation of the subject, whichever of the two variables may be taken in that capacity; and let x_1, x_2, x_3, &c., be the corresponding deviations of the relative, and let the mean of these be X. Then we find: (1) that y=rX for all values of y; (2) that r is the same, whichever of the two variables is taken for the subject; (3) that r is always less than 1; (4) that r measures the closeness of correlation.

ANEXO 2 - VALORES CRÍTICOS DO COEFICIENTE DE CORRELAÇÃO

QUADRO A.2.1 - VALORES CRÍTICOS DO COEFICIENTE DE CORRELAÇÃO SEGUNDO NÍVEIS DE SIGNIFICÂNCIA E TAMANHO DA AMOSTRA

TAMANHO DA AMOS-TRA	NÍVEIS DE SIGNIFICÂNCIA				TAMANHO DA AMOS-TRA	NÍVEIS DE SIGNIFICÂNCIA			
	0,010	0,025	0,050	0,100		0,010	0,025	0,050	0,100
3	0,869	0,872	0,879	0,891	32	0,949	0,959	0,966	0,972
4	0,822	0,845	0,868	0,894	33	0,950	0,960	0,967	0,973
5	0,822	0,855	0,879	0,902	34	0,951	0,960	0,967	0,973
6	0,835	0,868	0,890	0,911	35	0,952	0,961	0,968	0,974
7	0,847	0,876	0,899	0,916	36	0,953	0,962	0,968	0,974
8	0,859	0,886	0,905	0,924	37	0,955	0,962	0,968	0,974
9	0,868	0,893	0,912	0,929	38	0,956	0,964	0,970	0,975
10	0,876	0,900	0,917	0,934	39	0,957	0,965	0,971	0,976
11	0,883	0,906	0,922	0,938	40	0,958	0,966	0,972	0,977
12	0,889	0,912	0,926	0,941	41	0,958	0,967	0,972	0,977
13	0,895	0,917	0,931	0,944	42	0,959	0,967	0,973	0,978
14	0,901	0,921	0,934	0,947	43	0,959	0,967	0,973	0,978
15	0,907	0,925	0,937	0,950	44	0,960	0,968	0,973	0,978
16	0,912	0,928	0,940	0,952	45	0,961	0,969	0,974	0,978
17	0,916	0,931	0,942	0,954	46	0,962	0,969	0,974	0,979
18	0,919	0,934	0,945	0,956	47	0,963	0,970	0,974	0,979
19	0,923	0,937	0,947	0,958	48	0,963	0,970	0,975	0,980
20	0,925	0,939	0,950	0,960	49	0,964	0,971	0,975	0,980
21	0,928	0,942	0,952	0,961	50	0,965	0,972	0,977	0,981
22	0,930	0,944	0,954	0,962	55	0,967	0,974	0,978	0,982
23	0,933	0,947	0,955	0,964	60	0,970	0,976	0,980	0,983
24	0,936	0,949	0,957	0,965	65	0,972	0,977	0,981	0,984
25	0,937	0,950	0,958	0,966	70	0,974	0,978	0,982	0,985
26	0,939	0,952	0,959	0,967	75	0,975	0,979	0,983	0,986
27	0,941	0,933	0,960	0,968	80	0,976	0,980	0,984	0,987
28	0,943	0,955	0,962	0,969	85	0,977	0,981	0,985	0,987
29	0,945	0,956	0,962	0,969	90	0,978	0,982	0,985	0,988
30	0,947	0,957	0,964	0,970	95	0,979	0,983	0,986	0,989
31	0,948	0,958	0,965	0,971	100	0,981	0,984	0,987	0,989

FONTE: FILLIBEN (1975)