PONTÍFICIA UNIVERSIDADE CATÓLICA DE SÃO PAULO
Faculdade de Economia, Administração, Contabilidade e Atuariais.

ATLAS BRASIL 2013

DIMENSÃO DESENVOLVIMENTO HUMANO E

OUTRAS VARIÁVEIS

Disciplina: Métodos Quantitativos

Professor: Dr. Arnoldo Jose de Hoyos

Luciano Ferreira da Silva

1º Semestre 2014

SUMÁRIO

2. ENTENDENDO OS DADOS.... 4

2.1 – OS INDIVÍDUOS. 4

2.2 AS VARIÁVEIS 5

2.3 A TABELA DE DADOS 7

3. ANÁLISE DAS VARIÁVEIS 8

3.1 VARIÁVEIS CATEGÓRICAS 8

3.1.1 Variável: "Município" 8

3.2 VARIÁVEIS QUANTITATIVAS 11

3.2.1 Variável: "ESPVIDA" (Dimensão Demografia) 11

3.2.2 Variável: "IDHM_R" 12

3.2.3 Variável: "IDHM" 13

3.2.4 Variável: "I_FREQ_PROP" 14

3.2.5 Variável: "IDHM_E" 14

3.2.6 Variável: "T_NESTUDA_NTRAB_MMEIO" 15

3.2.7 Variável: "T_FUNDIN_TODOS_MMEIO" 16

3.2.8 Variável: "MORT_1" 17

3.2.9 Variável: "T_DENS" 18

3.2.10 Variável: "T_FLBAS" 19

3.2.11 Variável: "T_FLFUND" 20

3.2.12 Variável: "RENOCUP" 21

3.2.13 Variável: "PRENTRAB" 22

3.2.14 Variável: "T_DES2529" 23

3.2.15 Variável: "P_FORMAL" 24

3.2.16 Variável: "T_ATIV" 25

4. ANÁLISE COMPARATIVA DA ANÁLISE DESCRITIVA 27

5. CORRELAÇÃO DAS VARIÁVEIS 31

6. DENDROGRAMA 34

7. GRÁFICOS DE DISPERSÃO 36

8. ANÁLISE DE REGRESSÃO DAS VARIÁVEIS COM SIMILARIDADE 45

COMENTÁRIOS DAS ANÁLISES 61

9. REGRESSÃO MULTIVARIADA 62

10. ANÁLISE ANOVA VARIÁVEIS REGIÃO 67

10.1 – VARIÁVEL IDHM POR REGIÃO 67

10.2 – VARIÁVEL ESPVIDA POR REGIÃO 68

10.3 – VARIÁVEL IDHM_R POR REGIÃO 69

10.4 – VARIÁVEL IDHM_E POR REGIÃO 71

10.5 – VARIÁVEL POR I_FREQ_PROP REGIÃO 72

10.6 – VARIÁVEL MORT1 POR REGIÃO 74

10.7 – VARIÁVEL T_NESTUDA_MMEIO POR REGIÃO 75

10.8 – VARIÁVEL T_FUNDIN_TODOS_MMEIO POR REGIÃO 77

10.9 – VARIÁVEL T_DENS POR REGIÃO 78
10.10 – VARIÁVEL T_FLBAS POR REGIÃO. 80
10.11 – VARIÁVEL T_FLFUND POR REGIÃO 81
10.12 – VARIÁVEL RENOCUP POR REGIÃO 83
10.13 – VARIÁVEL T_ATIV POR REGIÃO 84
10.14 – VARIÁVEL PRENTRAB POR REGIÃO 86
10.15 – VARIÁVEL P_FORMA POR REGIÃO 87
10.16 – VARIÁVEL T_DES2529 POR REGIÃO 89
COMENTÁRIOS DA ANÁLISE 92
11. PESQUISA POR AMOSTRAGEM 95
11.1 – VARIÁVEL MORT1 95
11.2 – VARIÁVEL ESPVIDA 100
11.3 – VARIÁVEL IDHM_R 106
12. CORRELAÇÃO LINEAR 113
12.1 CORRELAÇÃO DAS VARIÁVEIS 113
12.2 DENDOGRAMA 115
12.3. PRINCIPAIS COMPONENTES 118
COMENTÁRIOS DA ANÁLISE 122
13. DENDOGRAMA DOS DADOS AGRUPADOS PELO RESULTADO DAS MORT1 x ESPVIDA x IDHM_R x T_NESTUDA_MMEIO x IDHM POR ESTADO (-DF) 123
13.1 DENDOGRAMA DOS DADOS AGRUPADOS PELO RESULTADO DOS DESVIOS PADRÃO ENTRE MORT1 x ESPVIDA x IDHM_R x T_NESTUDA_MMEIO x IDHM POR ESTADO (-DF) 125
13.2 ANÁLISE DAS VARIÂNCIAS DAS VARIÁVEIS POR ESTADO 128
13.2.1 Análise das variâncias da variável MORT1 por estado 129
13.2.2 Análise das variâncias da variável ESPVIDA por estado 130
13.2.3 Análise das variâncias da variável IDHM_R por estado 132
13.2.4 Análise das variâncias da variável T_NESTUDA_MMEIO por estado 134
13.2.4 Análise das variâncias da variável IDHMn por estado 136
COMENTÁRIOS DA ANÁLISE 137
14. ANÁLISE DISCRIMINANTE 138
14.1 ANÁLISE DISCRIMINANTE LINEAR POR REGIÃO 138
14.2 ANÁLISE DISCRIMINANTE LINEAR POR “2 BRASIS” 139
14.3 ANÁLISE DISCRIMINANTE QUADRÁTICA POR “3 BRASIS” 140
14.4 ANÁLISE DISCRIMINANTE LINEAR PARA DADOS AGRUPADOS 141
15. REGRESSÃO LOGÍSTICA ORDINAL PARA AS VARIÁVEIS: MORT1 x ESPVIDA x IDHM_R x T_NESTUDA_MMEIO x IDHM. 145
COMENTÁRIOS DA ANÁLISE 149
16. ARVORE DE DECISÃO PELO SPSS 150
CONSIDERAÇÕES FINAIS 152
17. ANÁLISE DE CORRESPONDÊNCIA DAS VARIÁVEIS 154
COMENTÁRIOS DA ANÁLISE 156

1. INTRODUÇÃO

O presente trabalho tem por objetivo efetuar uma análise exploratória na dimensão Desenvolvimento Humano dos dados apresentados na plataforma Atlas Brasil. Para tanto, o relatório utilizado para análise da dimensão Desenvolvimento Humano no Atlas Brasil 2013, que é apresenta o Índice de Desenvolvimento Humano Municipal – IDHM - de 5.565 municípios brasileiros, além de mais de 180 indicadores de população, educação, habitação, saúde, trabalho, renda e vulnerabilidade, com dados extraídos dos Censos Demográficos de 1991, 2000 e 2010.[1]

Para iniciar o entendimento dos dados que tem como foco o relatório Atlas Brasil 2013 (dados 2010), incluindo a definição das variáveis, suas classificações em variáveis categóricas ou quantitativas, os significados e unidades de medida, além da apresentação da tabela de dados. Na seqüência, analisamos cada uma das variáveis separadamente quanto a sua forma de distribuição, os valores atípicos, medidas de centro e dispersão. Para tal contamos com o auxílio de gráficos (*pie chart*, barras, histogramas, gráficos de ramos, box-plot, dot-plot e curvas de densidade) e de medidas numéricas (média, mediana, quartis, desvio-padrão, variância, intervalo de confiança e teste de normalidade de Anderson-Darling). No final, buscamos comparar as análises efetuadas para cada variável. O software estatístico utilizado é o **MINITAB 16.**

2. ENTENDENDO OS DADOS

2.1 – OS INDIVÍDUOS.

Os indivíduos deste trabalho são os municípios brasileiros, que serão analisados pelos seus indicadores relativos à dimensão Desenvolvimento Humano presentes no relatório Atlas Brasil 2013, dados referentes ao ano de 2010. Este sujeito da análise é composto por um total de 5565 municípios brasileiros e os dados analisados de cada município são as variáveis que serão descritas na próxima seção.

Quanto à dimensão Desenvolvimento Humano, esta está relacionada ao processo de ampliação das liberdades das pessoas, no que tange as suas capacidades e as oportunidades a seu dispor, para que elas possam escolher a vida que desejam ter. O processo de expansão destas liberdades inclui as dinâmicas sociais, econômicas, políticas e ambientais necessárias para garantir uma variedade de oportunidades, bem como o ambiente propício para cada um exercer na plenitude o seu potencial.

Deste modo, o Desenvolvimento Humano deve estar centrado nas pessoas e na ampliação do seu bem-estar. Nesta abordagem, a renda e a riqueza não são fins em si mesmas, mas meios para que as pessoas possam viver a vida que desejam. Assim, o crescimento econômico de uma sociedade não se traduz automaticamente em qualidade de vida e, muitas vezes, o que se observa é o reforço das desigualdades.

Portanto, é preciso que o crescimento econômico seja transformado em conquistas concretas para as pessoas, por meio de ações que proporcionem uma realidade que apresente crianças mais saudáveis, educação universal e de qualidade, ampliação da participação

[1] Cf. http://www.atlasbrasil.org.br/2013/

política dos cidadãos, preservação ambiental, equilíbrio da renda e das oportunidades entre toda a população, maior liberdade de expressão, entre outras. Além disso, ao colocar as pessoas no centro da análise, a abordagem de desenvolvimento humano redefine a maneira com que pensamos e lidamos com o desenvolvimento de forma nacional e local, ou seja, no âmbito dos municípios.

2.2 AS VARIÁVEIS

São 13 as variáveis desta pesquisa, incluindo a Unidade da Federação (UF). As mesmas são melhor explicadas na Tabela 1. Ressalta-se que todos os dados desta pesquisa são referentes ao ano de 2010.

Tabela 1 – Variáveis Utilizadas Atlas Brasil 2013

VARIÁVEL	SIGNIFICADO	TIPO	UNIDADE DE MEDIDA
ESPVIDA	Número médio de anos que as pessoas deverão viver a partir do nascimento, se permanecerem constantes ao longo da vida o nível e o padrão de mortalidade por idade prevalecentes no ano do Censo.	Variável Quantitativa	índice
IDHM_R	Índice da dimensão Renda que é um dos 3 componentes do IDHM. É obtido a partir do indicador Renda per capita, através da fórmula: [ln (valor observado do indicador) - ln (valor mínimo)] / [ln (valor máximo) - ln (valor mínimo)], onde os valores mínimo e máximo são R$ 8,00 e R$ 4.033,00 (a preços de agosto de 2010).	Variável Quantitativa	Índice
IDHM_E	Índice sintético da dimensão Educação que é um dos 3 componentes do IDHM. É obtido através da média geométrica do subíndice de frequência de crianças e jovens à escola, com peso de 2/3, e do subíndice de escolaridade da população adulta, com peso de 1/3.	Variável Quantitativa	Índice
IDHM	Índice de Desenvolvimento Humano Municipal. Média geométrica dos índices das dimensões Renda, Educação e Longevidade, com pesos iguais.	Variável Quantitativa	Índice
I_FREQ_PROP	Subíndice selecionado para compor o IDHMEducação, representando a frequência de crianças e jovens à escola em séries adequadas à sua idade. É obtido através da média aritmética simples de 4 indicadores: % de crianças de 5 a 6 anos na escola, % de crianças de 11 a 13 anos no 2º ciclo do fundamental, % de jovens de 15 a 17 anos com o fundamental completo e % de jovens de 18 a 20 anos com o médio completo.	Variável Quantitativa	Índice
MORT1_np	Número de crianças que não deverão sobreviver ao primeiro ano de vida em cada 1000 crianças nascidas vivas.	Variável Quantitativa	Índice

T_NESTUDA_NTRAB_MMEIO	Razão entre as pessoas de 15 a 24 anos que não estudam nem trabalham e são vulneráveis à pobreza e a população total nesta faixa etária multiplicado por 100. Define-se como vulneráveis à pobreza as pessoas que moram em domicílios com renda per capita inferior a 1/2 salário mínimo de agosto de 2010. São considerados apenas os domicílios particulares permanentes.	Variável Quantitativa	Índice
T_FUNDIN_TODOS_MMEIO	Percentual de pessoas que vivem em domicílios vulneráveis à pobreza (com renda per capita inferior a 1/2 salário mínimo de agosto de 2010) e em que ninguém tem o ensino fundamental completo. São considerados apenas os domicílios particulares permanentes.	Variável Quantitativa	Índice
T_DENS	Razão entre a população que vive em domicílios particulares permanentes com densidade superior a 2 e a população total residente em domicílios particulares permanentes multiplicado por 100. A densidade do domicílio é dada pela razão entre o total de moradores do domicílio e o número total de cômodos usados como dormitório.	Variável Quantitativa	Índice
T_FLBAS	Razão entre o número de pessoas na faixa etária de 6 a 17 anos frequentando o ensino básico (fundamental ou médio - regular ou seriado) e a população total dessa mesma faixa etária multiplicado por 100. As pessoas de 6 a 17 anos frequentando a pré-escola foram consideradas como se estivessem no 1º ano do ensino fundamental. As pessoas de 6 a 17 anos frequentando a 4ª série do ensino médio foram consideradas como já tendo concluído esse nível de ensino.	Variável Quantitativa	Índice
T_FLFUND	Razão entre o número de pessoas na faixa etária de 6 a 14 anos frequentando o ensino fundamental regular seriado e a população total dessa mesma faixa etária multiplicado por 100. As pessoas de 6 a 14 anos frequentando a pré-escola foram consideradas como se estivessem no 1º ano do ensino fundamental.	Variável Quantitativa	Índice
RENOCUP	Média dos rendimentos de todos os trabalhos das pessoas ocupadas de 18 anos ou mais de idade. Valores em reais de agosto de 2010.	Variável Quantitativa	Índice
T_ATIV	Razão entre as pessoas de 10 anos ou mais de idade que eram economicamente ativas, ou seja, que estavam ocupadas ou desocupadas na semana de referência do Censo e o total de pessoas nesta faixa etária multiplicado por 100. Considera-se desocupada a pessoa que, não estando ocupada na semana de referência, havia procurado trabalho no mês anterior a essa pesquisa.	Variável Quantitativa	Índice
PRENTRAB	Participação percentual das rendas provenientes do trabalho (principal e outros) na renda total, considerando-se apenas as pessoas que vivem em domicílios particulares permanentes.	Variável Quantitativa	Índice

P_FORMAL	Razão entre o número de pessoas de 18 anos ou mais formalmente ocupadas e o número total de pessoas ocupadas nessa faixa etária multiplicado por 100. Foram considerados como formalmente ocupados os empregados com carteira de trabalho assinada, os militares do exército, da marinha, da aeronáutica, da polícia militar ou do corpo de bombeiros, os empregados pelo regime jurídico dos funcionários públicos, assim como os empregadores e trabalhadores por conta própria que eram contribuintes de instituto de previdência oficial.	Variável Quantitativa	Índice
T_DES2529	Percentual da população economicamente ativa (PEA) nessa faixa etária que estava desocupada, ou seja, que não estava ocupada na semana anterior à data do Censo mas havia procurado trabalho ao longo do mês anterior à data dessa pesquisa.	Variável Quantitativa	Índice
UF	Código utilizado pelo IBGE para identificação do Estado.	Variável Categórica	n/a

Fonte: Atlas Brasil, 2013.

2.3 A TABELA DE DADOS[2]

Descriptive Statistics: MORT1_np; T_NESTUDA_NT; T_FUNDIN_TOD; T_DENS(np); ...

Variable	N	N*	Mean	Minimum	Median	Maximum
MORT1_np	5564	0	0,71919	0,00000	0,78034	1,00000
T_NESTUDA_NTRAB_MMEIO_np	5564	0	0,73254	0,00000	0,75258	1,00000
T_FUNDIN_TODOS_MMEIO_np	5564	0	0,72383	0,00000	0,75027	1,00000
T_DENS(np)	5564	0	0,72182	0,00000	0,74526	1,00000
ESPVIDAnp	5564	0	0,58383	0,00000	0,61244	1,00000
I_FREQ_PROPnp	5564	0	0,57684	0,00000	0,57925	1,00000
IDHMnp	5564	0	0,54308	0,00000	0,55631	1,00000
IDHM_Enp	5564	0	0,56968	0,00000	0,57120	1,00000
IDHM_Rnp	5564	0	0,49457	0,00000	0,51731	1,00000
T_FLBASnp	5564	0	0,80070	0,00000	0,80948	1,00000
T_FLFUNDnp	5564	0	0,86454	0,000000	0,87260	1,00000
RENOCUPnp	5564	0	0,21158	0,00000	0,20561	1,00000
PRENTRABnp	5564	0	0,60539	0,00000	0,63626	1,00000
P_FORMAnp	5564	0	0,47052	0,00000	0,46291	1,00000
T_ATIVnp	5564	0	0,48060	0,00000	0,48397	1,00000
T_DES2529np	5564	0	0,82568	0,00000	0,84216	1,00000

[2] Destaca-se que os dados são compostos de 5564 municípios, pois excluiu-se o DF.

3. ANÁLISE DAS VARIÁVEIS

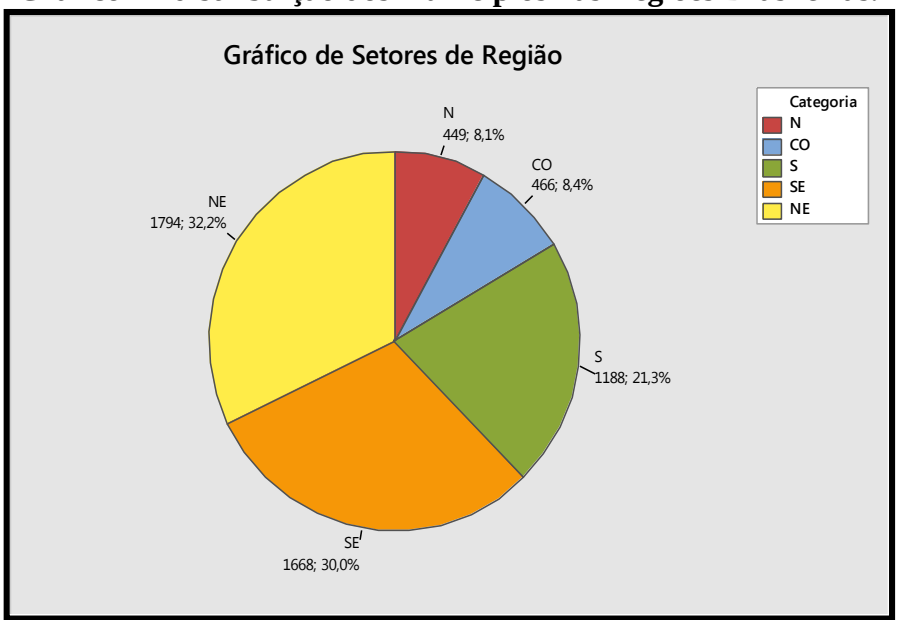
3.1 VARIÁVEIS CATEGÓRICAS

Este tipo de variável indica que o foco de concentração deve ser a análise de gráficos do tipo *pie chart* e/ou barras.

3.1.1 Variável: “Município”

A amostra totaliza 5565 municípios, que pode ser verificada na distribuição no território nacional de acordo com a região no gráfico 1.

Gráfico 1 - distribuição dos municípios nas Regiões Brasileiras.

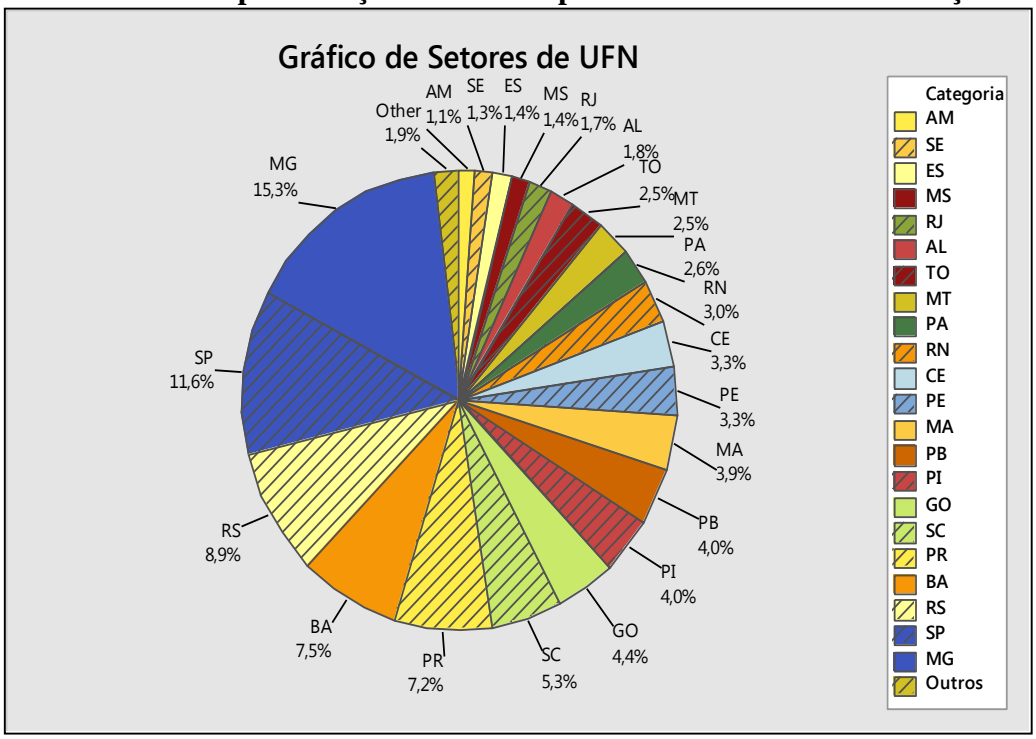


Fonte: elaborado pelo autor, 2014 (Atlas Brasil, 2014)

De acordo com gráfico 1 pode-se observar que as maiores concentrações de municípios brasileiros estão nas regiões do Nordeste com 32,20% e Sudeste com 30% somando juntas mais de 50% dos municípios pesquisados (62,20%).

O Gráfico 2 demonstra a distribuição dos municípios pelas Unidades Federativas do Brasil.

Gráfico 2 – Representação dos municípios nas Unidades da Federação



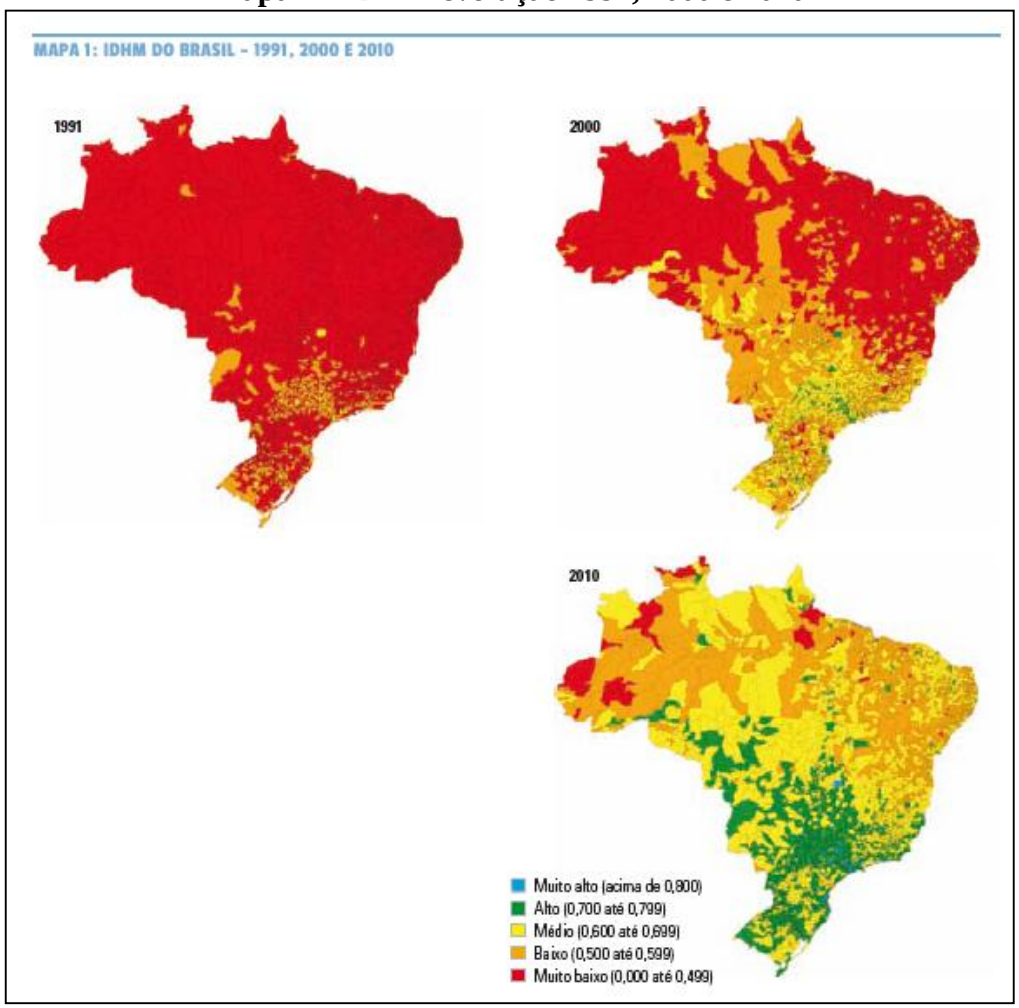
Fonte: elaborado pelo autor, 2014 (Atlas Brasil, 2014)

Conforme pode ser observado no Gráfico 2, as Unidades da Federação mais representativas são Minas Gerais (15,3%), São Paulo (11,6%) e Rio Grande do Sul (8,9%). As menos expressivas são de Amazonas, Sergipe, Espírito Santos, entre outros.

O Mapa 1 (Atlas Brasil, 2014) apresenta o IDHM (Índice de Desenvolvimento Humano Municipal) dos municípios brasileiros em 1991, 2000 e 2010. Com base nesta representação pode se observar que, com relação IDHM, existe uma concentração de índices mais altos nos municípios da região centro-sul do Brasil.

Contudo, também pode ser percebido que a região Norte e Nordeste que concentrava os municípios que apresentavam índices muito baixo de desenvolvimento humano, conseguiram reverter para índices baixos e médios de desenvolvimento.

Mapa 1 - IDHM evolução 1991, 2000 e 2010



Fonte: Atlas Brasil, 2014.

Para entender esta evolução do IDHM dos municípios brasileiros são apresentadas informações na tabela 2, ilustrada pelo gráfico 1. A classificação IDHM proposta pelo Atlas Brasil tem sua variação entre Muito Baixo Desenvolvimento Humano (IDHM inferior a 0,500) a Muito Alto Desenvolvimento Humano (IDHM igual ou superior a 0,800).

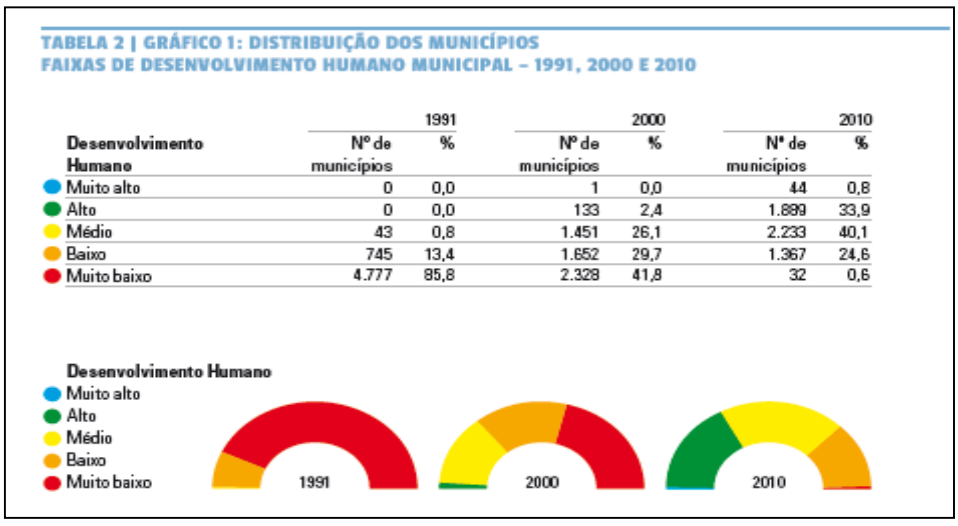


TABELA 2 | GRÁFICO 1: DISTRIBUIÇÃO DOS MUNICÍPIOS
FAIXAS DE DESENVOLVIMENTO HUMANO MUNICIPAL – 1991, 2000 E 2010

Desenvolvimento Humano	1991 Nº de municípios	1991 %	2000 Nº de municípios	2000 %	2010 Nº de municípios	2010 %
Muito alto	0	0,0	1	0,0	44	0,8
Alto	0	0,0	133	2,4	1.889	33,9
Médio	43	0,8	1.451	26,1	2.233	40,1
Baixo	745	13,4	1.652	29,7	1.367	24,6
Muito baixo	4.777	85,8	2.328	41,8	32	0,6

Conforme estas informações pode-se perceber a evolução dos municípios entre o período de 1991 e 2010. Em 1991, mais de 85% dos municípios encontravam-se na faixa de Muito Baixo Desenvolvimento Humano. Já nos anos 2000, pouco mais que 70% deles encontravam-se nas faixas de Baixo e Muito Baixo Desenvolvimento Humano.

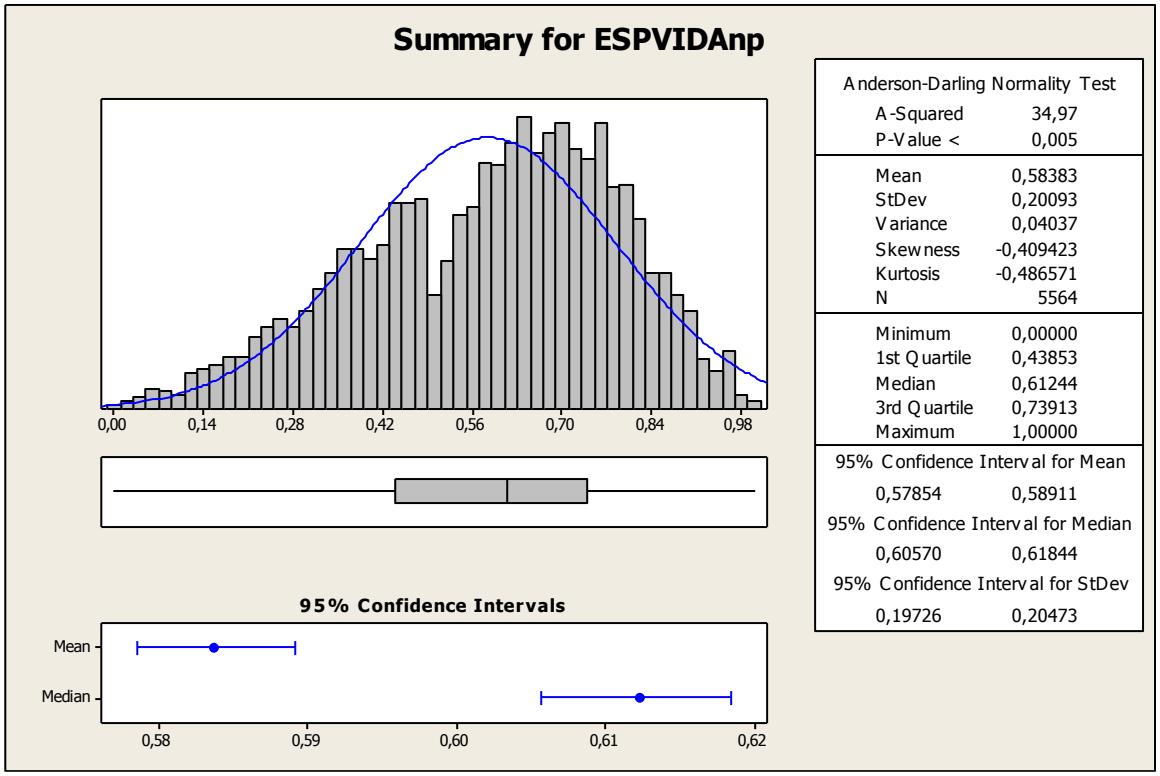
Na última análise referente a 2010, apenas um quarto (25%) dos municípios brasileiros encontravam-se nessas faixas e mais de 70% deles já figuravam nas faixas de Médio e Alto Desenvolvimento Humano. Segundo as informações constantes no Atlas Brasil 2013 isso ilustra os avanços do desenvolvimento humano no país nas últimas duas décadas.

3.2 VARIÁVEIS QUANTITATIVAS

A análise deste tipo de variável permite a utilização de uma maior gama de ferramentas de análise como histogramas, curvas de densidade e box-plot, além de informações numéricas como média, desvio-padrão, mediana, intervalo de confiança e teste de normalidade de Anderson-Darling.

3.2.1 Variável: "ESPVIDA" (Dimensão Demografia)

Segue abaixo quadro contendo Histograma, Curva de Densidade, Box-Plot, Intervalo de confiança da média e mediana, além das medidas numéricas como média, desvio-padrão, variância, quantidade de observações, valores mínimos, máximos e o teste de normalidade de Anderson-Darling (A-Squared e P-Value), para a variável **"ESPVIDA".**

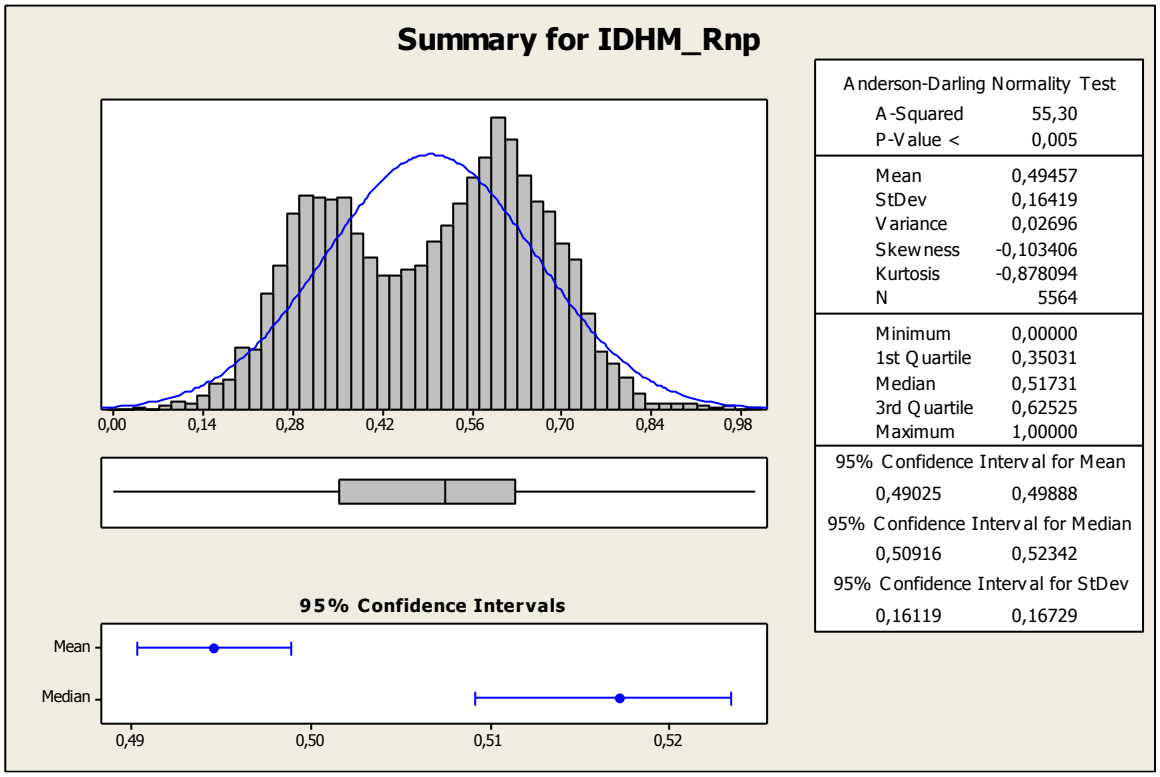


Algumas observações que podemos fazer:

- Forma: O Histograma nos permite analisar uma distribuição concentrada na faixa entre 0,14 a 0,98. O Box-Plot demonstra a concentração na faixa acima citada, o posicionamento da mediana está a direita do referido intervalo.
- Valores Atípicos: não se apresentaram nesta análise.
- Centro e Dispersão: A mediana nos indica que aproximadamente metade dos municípios tem **ESPVIDA** menor do que 0,61244 e a outra metade maior que este valor. A **ESPVIDA** média dos municípios é de 0.58383, tendo um desvio-padrão de 0,20093, não sendo um valor expressivo. A **ESPVIDA** mínima é de 0,0 e a máxima de 1,0. Com 95% de confiança, podemos afirmar que a média encontra-se entre os valores 0,57854 e 0,58911.

3.2.2 Variável: "IDHM_R"

Segue abaixo quadro contendo Histograma, Curva de Densidade, Box-Plot, Intervalo de confiança da média e mediana, além das medidas numéricas como média, desvio-padrão, variância, quantidade de observações, valores mínimos, máximos e o teste de normalidade de Anderson-Darling (A-Squared e P-Value), para a variável "**IDHM_R**".



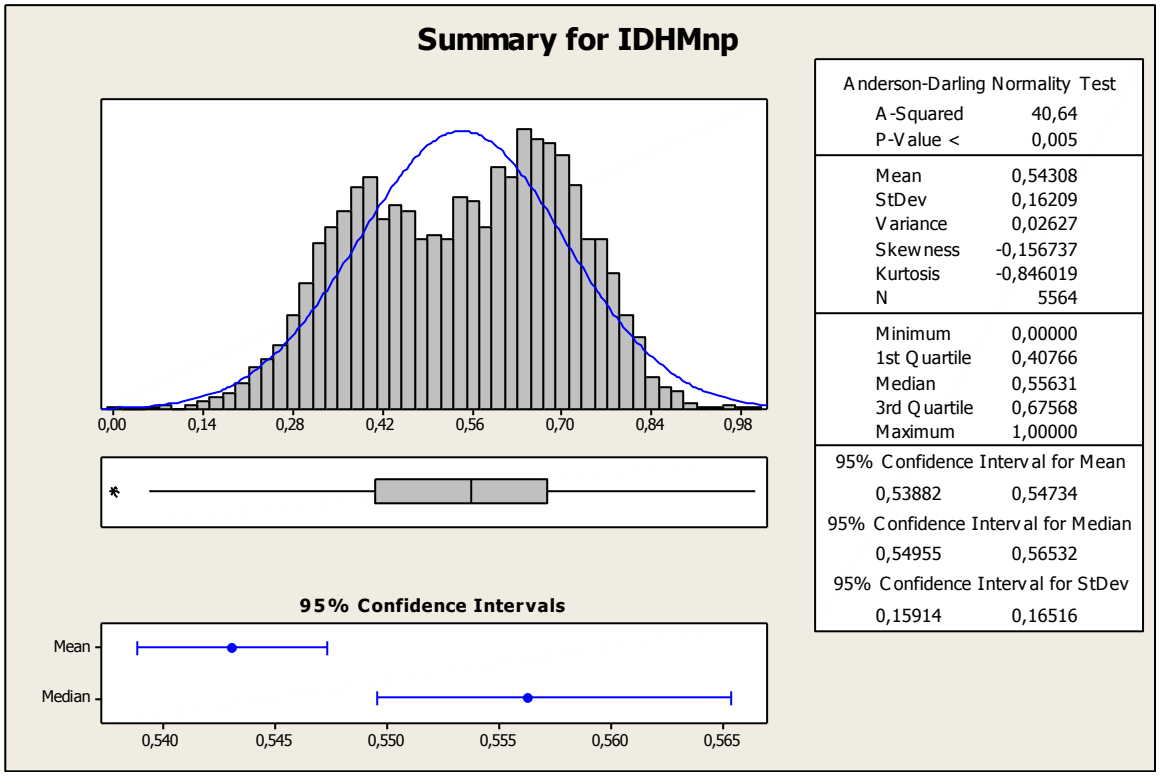
Algumas observações que podemos fazer:

- **Forma**: O Histograma nos apresenta uma distribuição concentrada na faixa entre os índices 0,14 e 0,84. O Box-Plot demonstra a concentração na faixa acima citada, o posicionamento da mediana confirma o posicionamento.
- **Valores Atípicos**: não há valores atípicos nesta análise.

- **Centro e Dispersão**: A mediana nos indica que há uma leve concentração de municípios do lado esquerdo, sendo seu valor de **IDHM_R** 0,51731. O IDHM_R médio dos municípios é de 0,49457, tendo um desvio-padrão de 0,16419. O **IDHM_R** mínimo é de 0,0 e o máximo de 1,0. Com 95% de confiança, podemos afirmar que a média encontra-se entre os índices de 0,49025 e 0,49888.

3.2.3 Variável: "IDHM"

Segue abaixo quadro contendo Histograma, Curva de Densidade, Box-Plot, Intervalo de confiança da média e mediana, além das medidas numéricas como média, desvio-padrão, variância, quantidade de observações, valores mínimos, máximos e o teste de normalidade de Anderson-Darling (A-Squared e P-Value), para a variável "**IDHM**".

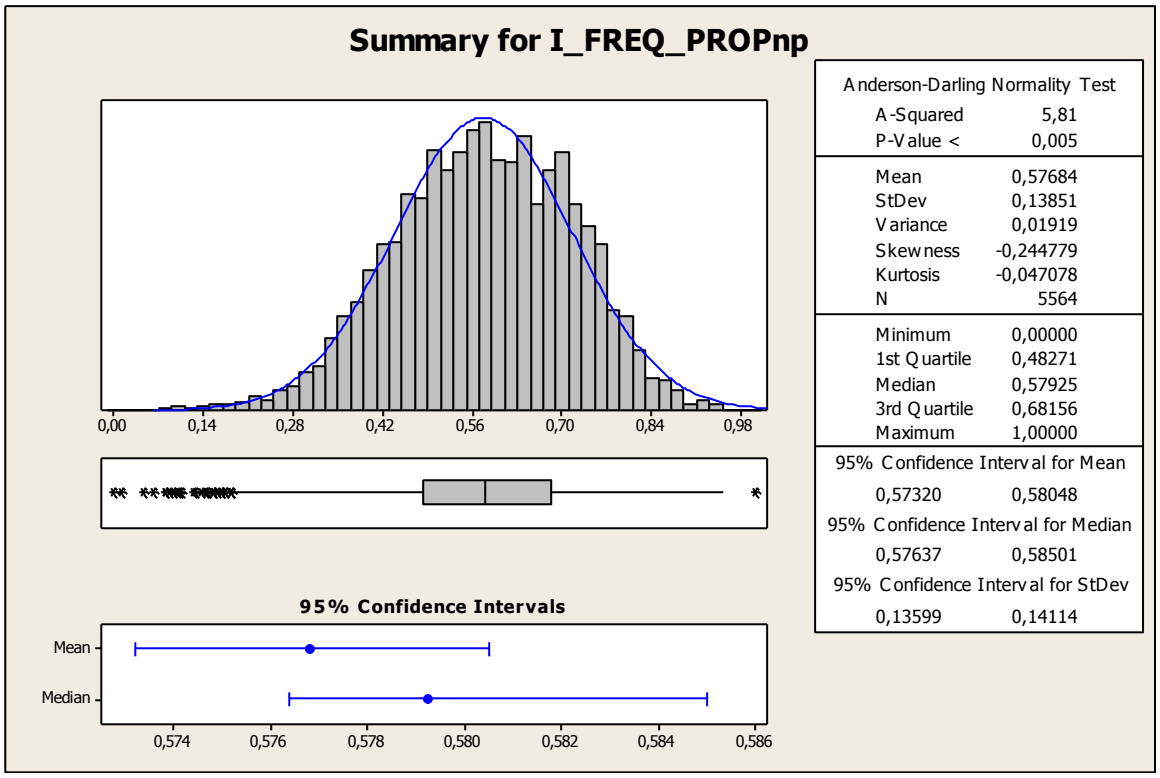


Algumas observações que podemos fazer:

- **Forma**: O Histograma nos apresenta uma distribuição concentrada na faixa entre os índices 0.28 e 0.84. O Box-Plot demonstra a concentração na faixa acima citada, o posicionamento da mediana confirma o posicionamento.
- **Valores Atípicos**: temos um valor atípico que é representado pelo município de Melgaço, do Pará, com um índice de 0,0.
- **Centro e Dispersão**: A mediana nos indica que municípios estão divididos proporcionalmente entre os dois lados, sendo que seu valor para o **IDHM** é de 0,55631. O **IDHM** médio dos municípios é de 0,54308, tendo um desvio-padrão de 0,16209. O **IDHM** mínimo é de 0,0 e o máximo de 1,0. Com 95% de confiança, podemos afirmar que a média encontra-se entre os índices de 0,53882 e 0,54734.

3.2.4 Variável: “I_FREQ_PROP”

Segue abaixo quadro contendo Histograma, Curva de Densidade, Box-Plot, Intervalo de confiança da média e mediana, além das medidas numéricas como média, desvio-padrão, variância, quantidade de observações, valores mínimos, máximos e o teste de normalidade de Anderson-Darling (A-Squared e P-Value), para a variável “**I_FREQ_PROP**”.



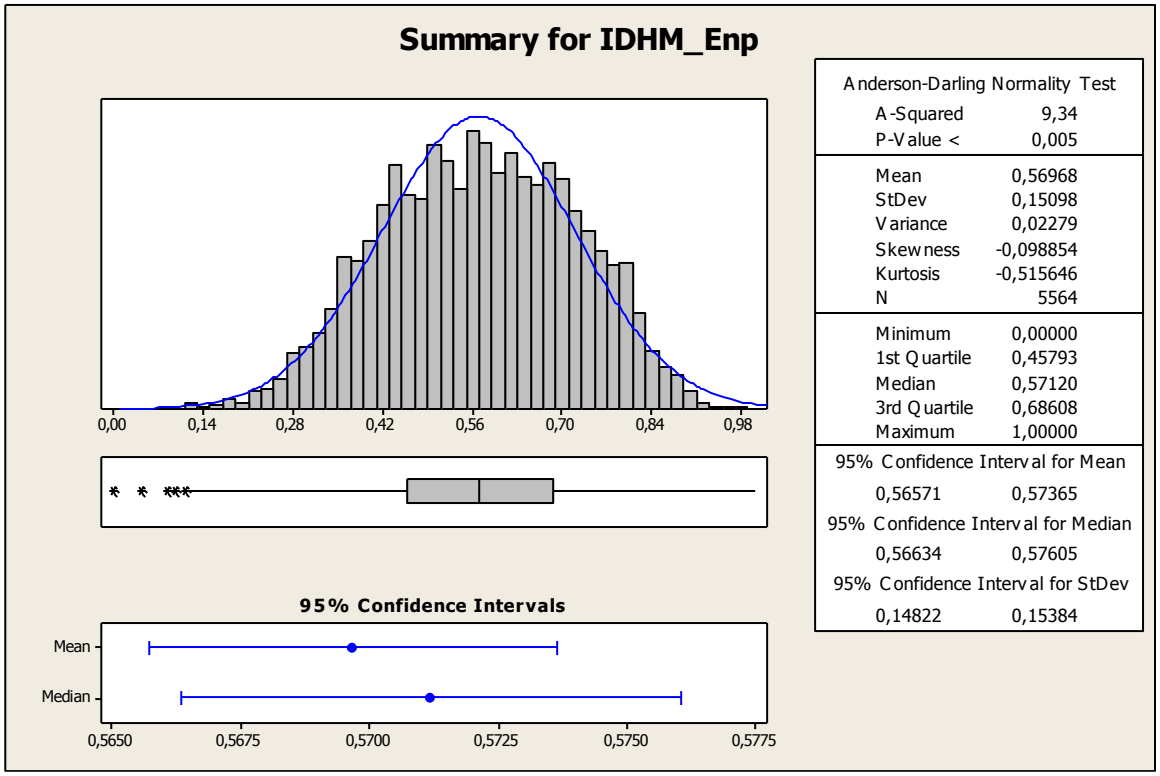
Algumas observações que podemos fazer:

- **Forma**: O Histograma nos apresenta uma distribuição concentrada na faixa entre os índices 0,28 e 0,84. O Box-Plot demonstra a concentração na faixa acima citada, o posicionamento da mediana confirma o posicionamento.
- **Valores Atípicos**: temos alguns valores atípicos que são representados na sua extremidade esquerda pelos municípios de Melgaço, do Pará, com um índice de 0,0, e Atalaia do Norte, do Amazonas, com um índice de 0,01009. Na extremidade direita temos o município de Monções de São Paulo com um índice de 1,0.
- **Centro e Dispersão**: A mediana nos indica que a maior parte dos municípios tem índice **I_FREQ_PROP** acima de 0,57925. O **I_FREQ_PROP** médio dos municípios é de 0,57684, tendo um desvio-padrão de 0,13851. O **I_FREQ_PROP** mínimo é de 0,0 e o máximo de 1,0. Com 95% de confiança, podemos afirmar que a média encontra-se entre os índices de 0,57320 e 0,58048.

3.2.5 Variável: “IDHM_E”

Segue abaixo quadro contendo Histograma, Curva de Densidade, Box-Plot, Intervalo de confiança da média e mediana, além das medidas numéricas como média, desvio-padrão,

variância, quantidade de observações, valores mínimos, máximos e o teste de normalidade de Anderson-Darling (A-Squared e P-Value), para a variável "**IDHM_E**".

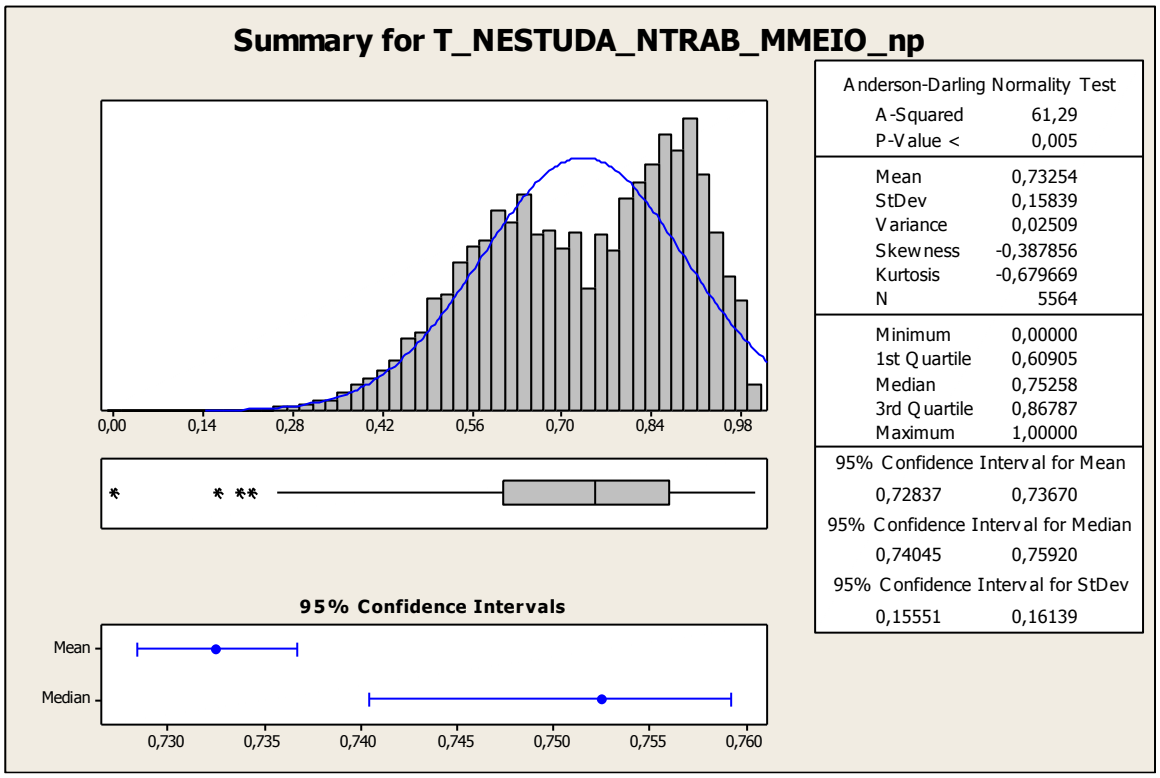


Algumas observações que podemos fazer:

- **Forma**: O Histograma nos apresenta uma distribuição concentrada na faixa entre os percentuais de 0,28 e 0,84. O Box-Plot demonstra a concentração na faixa acima citada, o posicionamento da mediana confirma o posicionamento.
- **Valores Atípicos**: temos alguns valores atípicos que são representados na sua extremidade esquerda, com destaque aqui para os municípios de Melgaço com um índice de 0,000, e Chaves com um índice de 0,04369, ambos do Estado do Pará.
- **Centro e Dispersão**: A mediana nos indica que há uma distribuição proporcional de municípios entre os dois lados, sendo que o seu valor para **IDHM_E é de** 0,57120. O **IDHM_E** médio dos municípios é de 0,05698, tendo um desvio-padrão de 0,15098. O **IDHM_E** mínimo é de 0,0000 e o máximo de 1,0. Com 95% de confiança, podemos afirmar que a média encontra-se entre os índices de 0,56571 e 0,57365.

3.2.6 Variável: "T_NESTUDA_NTRAB_MMEIO"

Segue abaixo quadro contendo Histograma, Curva de Densidade, Box-Plot, Intervalo de confiança da média e mediana, além das medidas numéricas como média, desvio-padrão, variância, quantidade de observações, valores mínimos, máximos e o teste de normalidade de Anderson-Darling (A-Squared e P-Value), para a variável "**T_NESTUDA_NTRAB_MMEIO**".

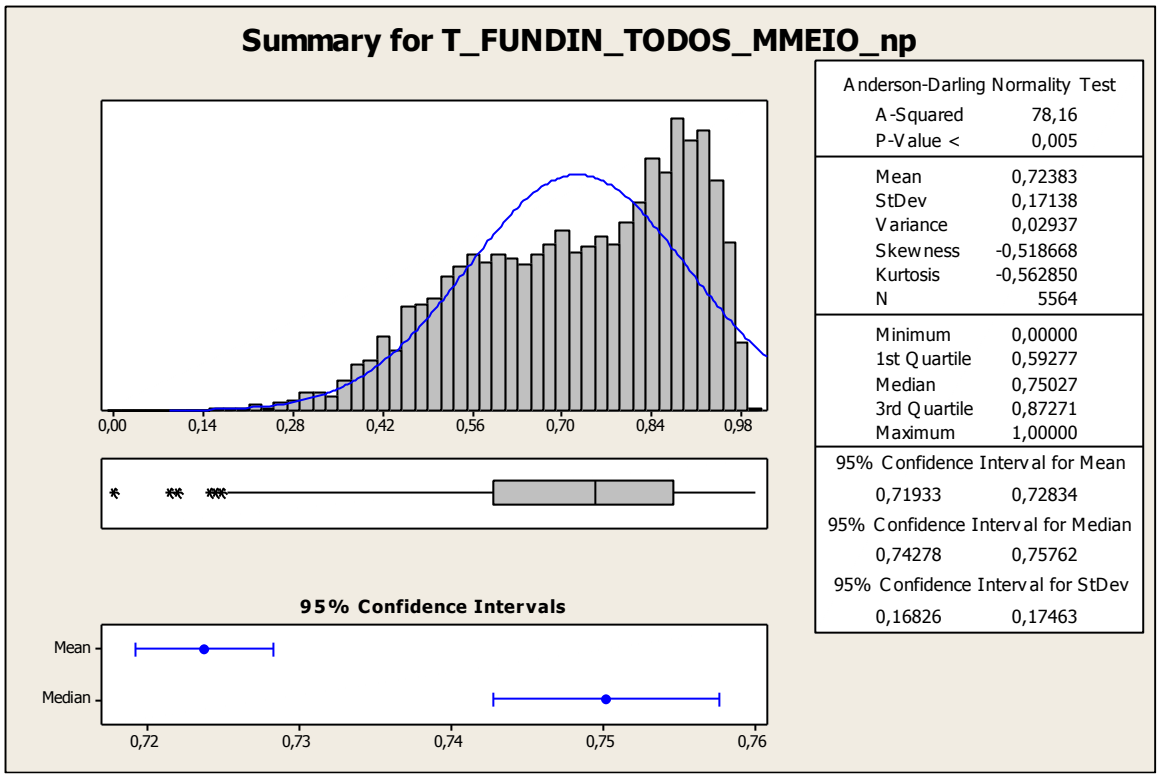


Algumas observações que podemos fazer:

- **Forma**: O Histograma nos apresenta uma distribuição concentrada na faixa entre os índices 0,42 e 0,98. O Box-Plot demonstra a concentração na faixa acima citada, o posicionamento da mediana confirma o posicionamento.
- **Valores Atípicos**: temos quatro valores atípicos que estão representados do esquerdo do gráfico, são os municípios: Amaraji de Roraima com um índice de 0,0; São João Batista do Maranhão com um índice de 0,16018; Recursolândia de Tocantins com índice de 0,19457; Santa Isabel do Rio Negro do Amazonas com índice de 0,21285.
- **Centro e Dispersão**: A mediana nos indica que há uma distribuição dos municípios proporcional entre os lados, sendo que o seu valor para **T_NESTUDA_NTRAB_MMEIO** é de 0,75258. O **T_NESTUDA_NTRAB_MMEIO** médio dos municípios é de 0,73254 tendo um desvio-padrão de 0,15839. Com 95% de confiança, podemos afirmar que a média encontra-se entre os índices de 0,72837 e 0,73670.

3.2.7 Variável: "T_FUNDIN_TODOS_MMEIO"

Segue abaixo quadro contendo Histograma, Curva de Densidade, Box-Plot, Intervalo de confiança da média e mediana, além das medidas numéricas como média, desvio-padrão, variância, quantidade de observações, valores mínimos, máximos e o teste de normalidade de Anderson-Darling (A-Squared e P-Value), para a variável "**T_FUNDIN_TODOS_MMEIO**".

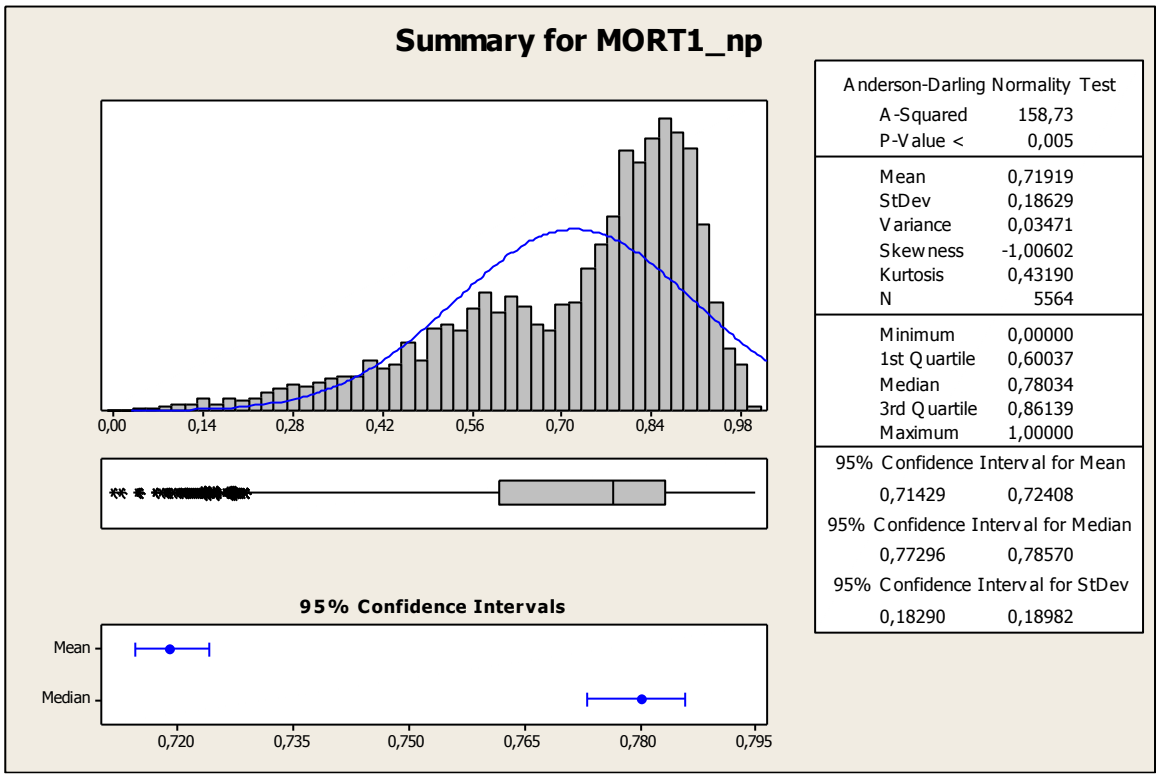


Algumas observações que podemos fazer:

- **Forma**: O Histograma nos apresenta uma distribuição concentrada na faixa entre os 0,33 e 0,99. O Box-Plot demonstra a concentração na faixa acima citada, o posicionamento da mediana confirma o posicionamento.
- **Valores Atípicos**: temos alguns valores atípicos que são representados na sua extremidade esquerda pelos municípios de Melgaço do Pará com um índice de 0,0; Itamarati do Amazonas com um índice de 0,08702; e Marajá do Sena do Maranhão com um índice de 0.09968.
- **Centro e Dispersão**: A mediana nos indica que há um maior número de municípios com **T_FUNDIN_TODOS_MMEIO** maior número de municípios do lado esquerdo do gráfico, sendo seu valor de 0,75027. O **T_FUNDIN_TODOS_MMEIO** médio dos municípios é de 0.72383, tendo um desvio-padrão de 0,17138. Com 95% de confiança, podemos afirmar que a média encontra-se entre os índices de 0,71933 e 0,72834.

3.2.8 Variável: “MORT_1”

Segue abaixo quadro contendo Histograma, Curva de Densidade, Box-Plot, Intervalo de confiança da média e mediana, além das medidas numéricas como média, desvio-padrão, variância, quantidade de observações, valores mínimos, máximos e o teste de normalidade de Anderson-Darling (A-Squared e P-Value), para a variável “**MORT_1**”.

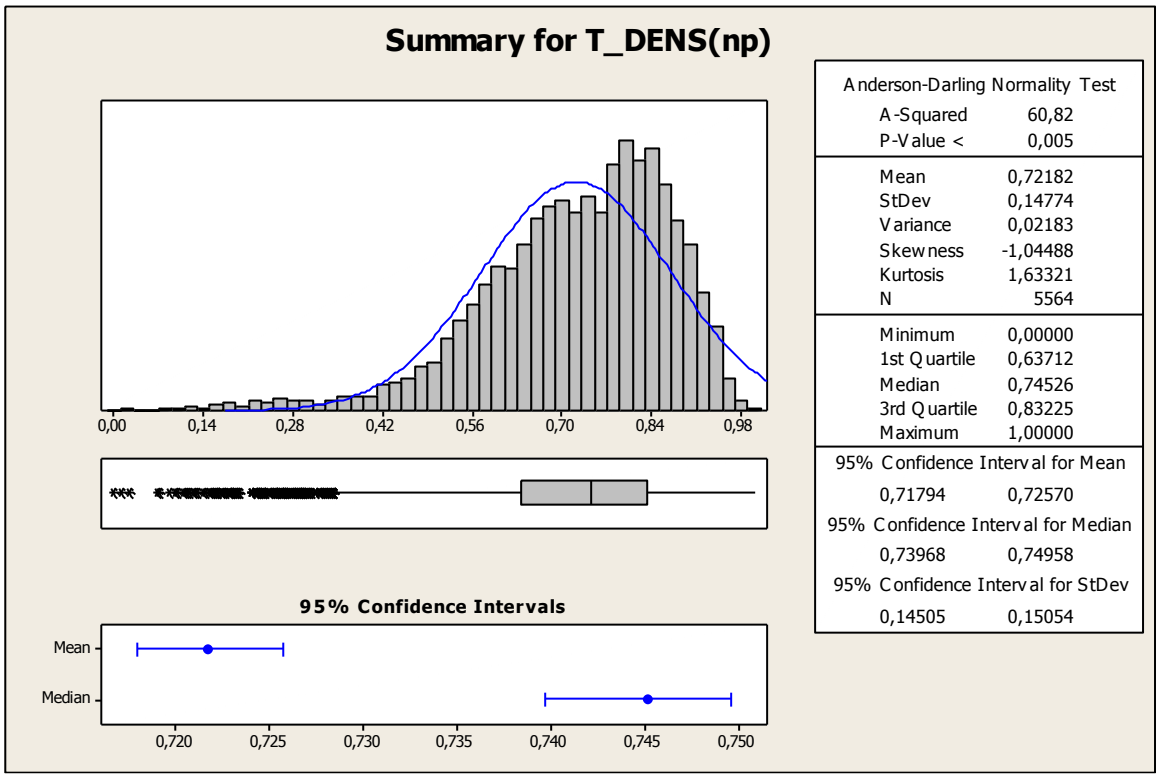


Algumas observações que podemos fazer:

- **Forma**: O Histograma nos apresenta uma distribuição concentrada na faixa entre os percentuais de 0,40 e 0,70. O Box-Plot demonstra a concentração na faixa acima citada, o posicionamento da mediana confirma o posicionamento.
- **Valores Atípicos**: temos alguns valores atípicos que são representados na sua extremidade esquerda, com destaque aqui para os municípios de Roteiro no Alagoas com um índice de 0,0, e Olho D'água Grande também em Alagoas com um índice de 0,0,01044.
- **Centro e Dispersão**: A mediana nos indica que há uma concentração de municípios do lado direito, sendo que o seu valor para **MORT_1 é de** 0,78034. O **MORT_1** médio dos municípios é de 0,71919, tendo um desvio-padrão de 0,18629. Com 95% de confiança, podemos afirmar que a média encontra-se entre os índices de 0,71429 e 0,72408.

3.2.9 Variável: "T_DENS"

Segue abaixo quadro contendo Histograma, Curva de Densidade, Box-Plot, Intervalo de confiança da média e mediana, além das medidas numéricas como média, desvio-padrão, variância, quantidade de observações, valores mínimos, máximos e o teste de normalidade de Anderson-Darling (A-Squared e P-Value), para a variável "**T_DENS**".

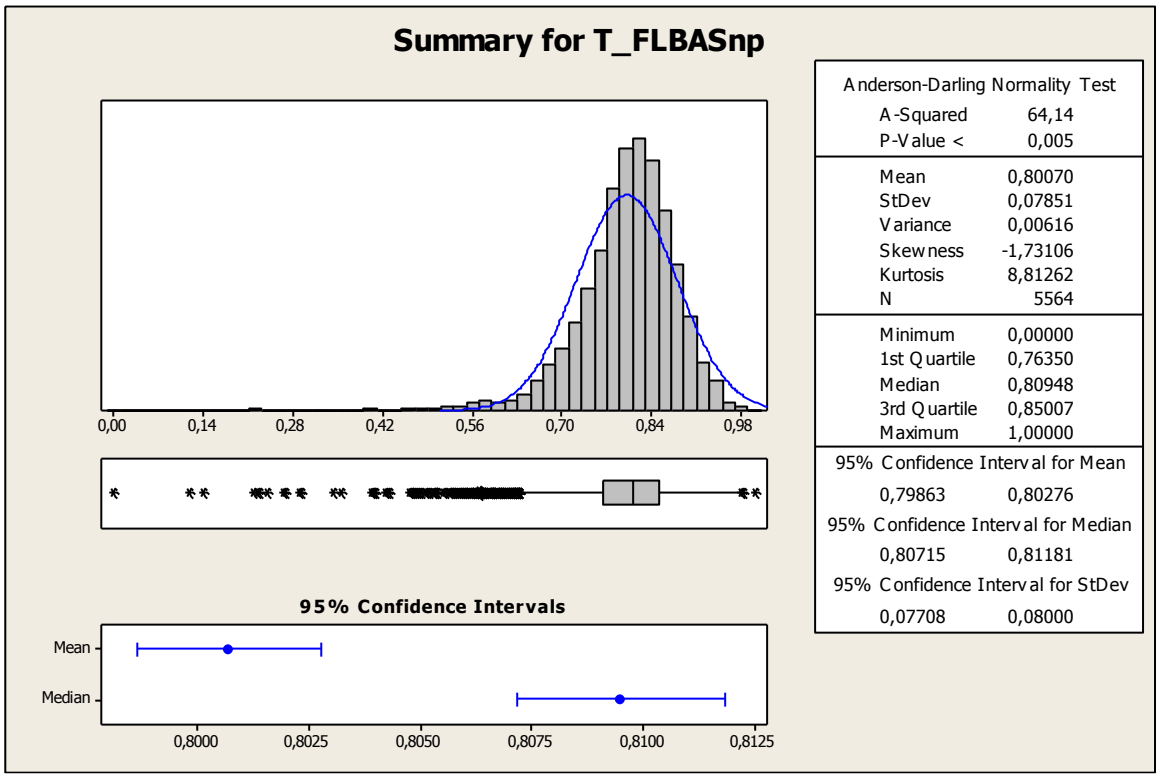


Algumas observações que podemos fazer:

- **Forma**: O Histograma nos apresenta uma distribuição concentrada na faixa entre os percentuais de 0,42 e 0,98. O Box-Plot demonstra a concentração de municípios abaixo da linha da mediana.
- **Valores Atípicos**: temos alguns valores atípicos na sua extremidade esquerda que são representados aqui pelos municípios: Uiramutã, de Roraima, com um índice de 0,0; o município de Santa Isabel do Rio Negro, do Amazonas, com um índice de 0,01193; e, o município de Melgaço, do Pará, com um índice de 0,02512.
- **Centro e Dispersão**: A mediana nos indica que mais da metade dos municípios tem **T_DENS** menor do que 0,74526. O **T_DENS** médio dos municípios é de 0,72182, tendo um desvio-padrão de 0,14774. Com 95% de confiança, podemos afirmar que a média encontra-se entre os índices de 0,71794 e 0,72570.

3.2.10 Variável: “T_FLBAS”

Segue abaixo quadro contendo Histograma, Curva de Densidade, Box-Plot, Intervalo de confiança da média e mediana, além das medidas numéricas como média, desvio-padrão, variância, quantidade de observações, valores mínimos, máximos e o teste de normalidade de Anderson-Darling (A-Squared e P-Value), para a variável “**T_FLBAS**”.

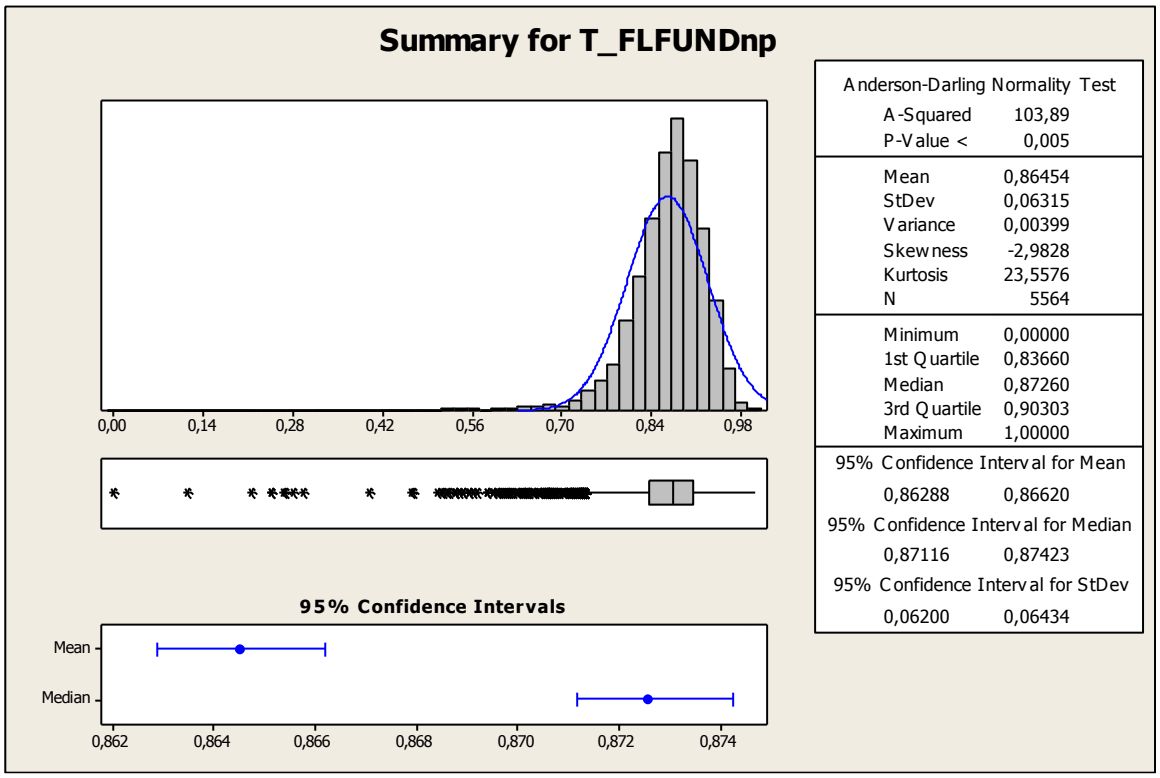


Algumas observações que podemos fazer:

- **Forma**: O Histograma nos apresenta uma distribuição concentrada na faixa entre os percentuais de 0,70 e 0,95. O Box-Plot demonstra a concentração na faixa acima citada, o posicionamento da mediana confirma o posicionamento.
- **Valores Atípicos**: temos alguns valores atípicos na sua extremidade esquerda que são representados aqui pelos municípios: Santa Isabel do Rio Negro, do Amazonas, com um índice de 0,0; Alto Alegre, de Roraima, com um índice de 0,11828; e, Nova Nazaré, do Mato Grosso, com um índice de 0,13809.
- **Centro e Dispersão**: A mediana nos indica que uma maior concentração na extremidade direita com valor de **T_FLBAS** de 0,80276. O **T_FLBAS** médio dos municípios é de 0,80070, tendo um desvio-padrão de 0,07851. Com 95% de confiança, podemos afirmar que a média encontra-se entre os índices de 0,79863 e 0,80279.

3.2.11 Variável: "T_FLFUND"

Segue abaixo quadro contendo Histograma, Curva de Densidade, Box-Plot, Intervalo de confiança da média e mediana, além das medidas numéricas como média, desvio-padrão, variância, quantidade de observações, valores mínimos, máximos e o teste de normalidade de Anderson-Darling (A-Squared e P-Value), para a variável "**T_FLFUND**".

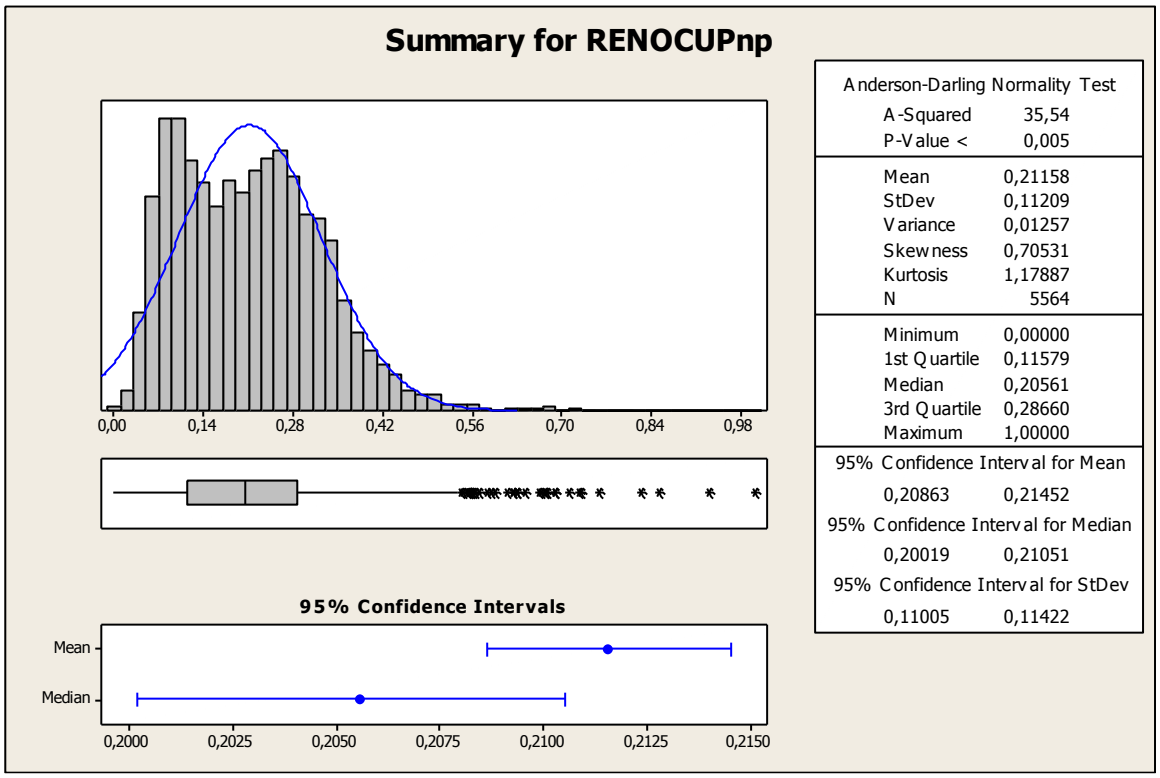


Algumas observações que podemos fazer:

- **Forma**: O Histograma nos apresenta uma distribuição concentrada na faixa entre os índices de 0,70 e 0,98. O Box-Plot demonstra a concentração na faixa acima citada, o posicionamento da mediana confirma o posicionamento.
- **Valores Atípicos**: temos muitos valores atípicos na sua extremidade esquerda que são representados aqui pelos municípios: Santa Isabel do Rio Negro, do Amazonas, com um índice de 0,0, e o município de Alto Alegre, de Roraima, com um índice de 0,11386.
- **Centro e Dispersão**: A mediana nos indica que pelo menos metade dos municípios tem **T_FLFUND** maior que 0,87260. O **T_FLFUND** médio dos municípios é de 0,86454, tendo um desvio-padrão de 0,06315. Com 95% de confiança, podemos afirmar que a média encontra-se entre os índices de 0,86288 e 0,86620.

3.2.12 Variável: "RENOCUP"

Segue abaixo quadro contendo Histograma, Curva de Densidade, Box-Plot, Intervalo de confiança da média e mediana, além das medidas numéricas como média, desvio-padrão, variância, quantidade de observações, valores mínimos, máximos e o teste de normalidade de Anderson-Darling (A-Squared e P-Value), para a variável "**RENOCUP**".

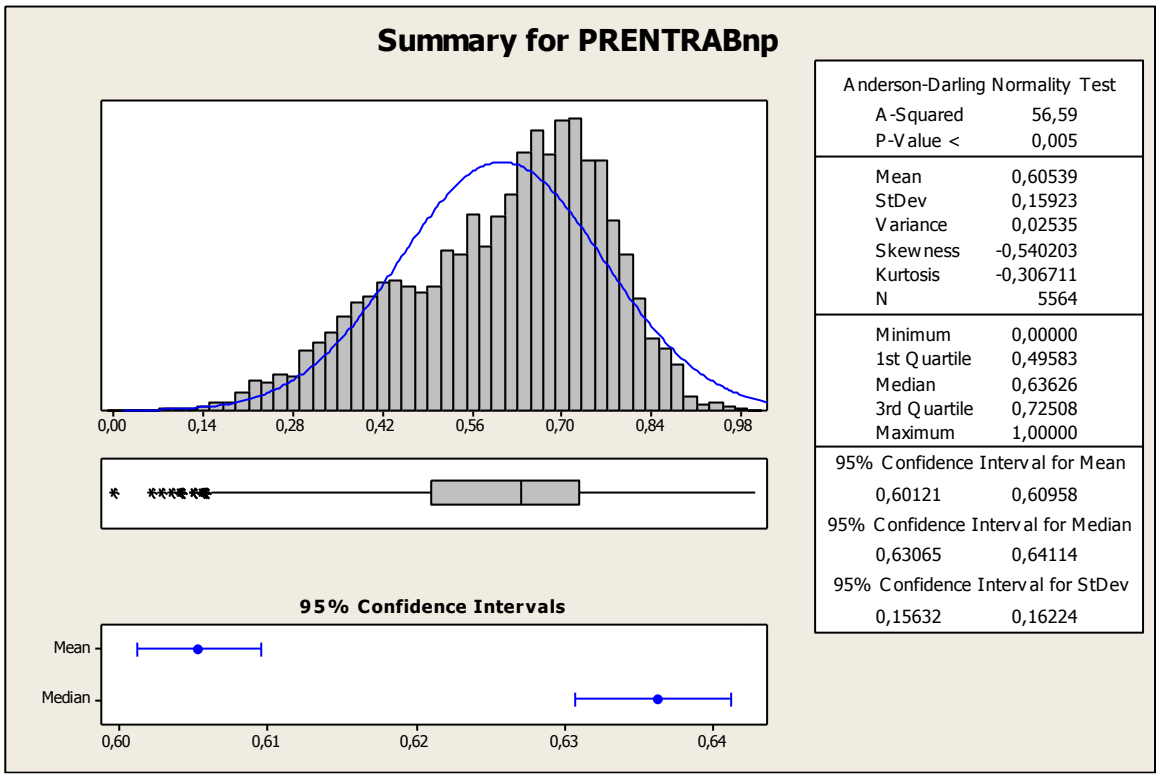


Algumas observações que podemos fazer:

- **Forma**: O Histograma nos apresenta uma distribuição concentrada na faixa entre os percentuais de 0,0 e 0,56. O Box-Plot demonstra a concentração na faixa acima citada, o posicionamento da mediana confirma o posicionamento.
- **Valores Atípicos**: temos alguns valores atípicos que são representados na sua extremidade direita, com destaque aqui para os municípios de Santana do Parnaíba com um índice de 1,0, e São Caetano do Sul com um índice de 0,93055, ambos do Estado de São Paulo.
- **Centro e Dispersão**: A mediana nos indica que há uma distribuição concentrada de municípios na extremidade esquerda, sendo que o seu valor para **RENOCUP** é de 0,20561. O **RENOCUP** médio dos municípios é de 0,21158, tendo um desvio-padrão de 0,11209. Com 95% de confiança, podemos afirmar que a média encontra-se entre os índices de 0,21452 e 0,21452.

3.2.13 Variável: "PRENTRAB"

Segue abaixo quadro contendo Histograma, Curva de Densidade, Box-Plot, Intervalo de confiança da média e mediana, além das medidas numéricas como média, desvio-padrão, variância, quantidade de observações, valores mínimos, máximos e o teste de normalidade de Anderson-Darling (A-Squared e P-Value), para a variável "**PRENTRAB**".

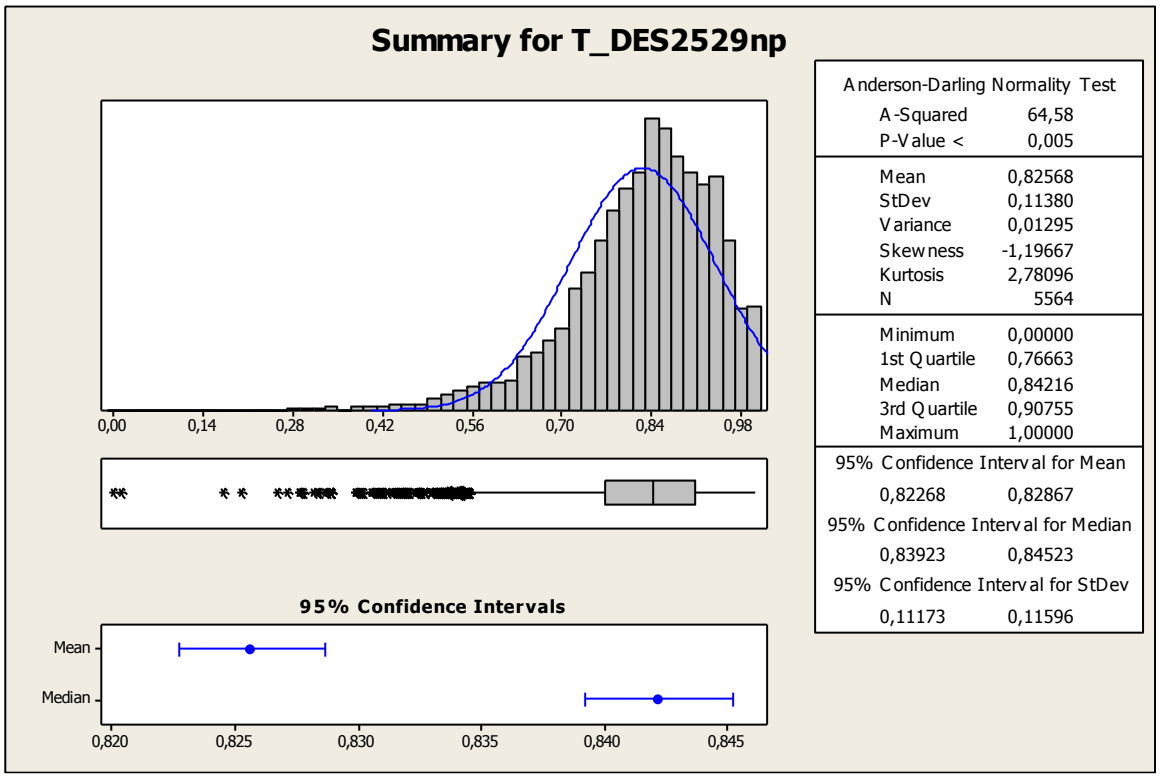


Algumas observações que podemos fazer:

- **Forma**: O Histograma nos apresenta uma distribuição concentrada na faixa entre os percentuais de 0,28 e 0,84. O Box-Plot demonstra a concentração na faixa acima citada, o posicionamento da mediana confirma o posicionamento.
- **Valores Atípicos**: temos alguns valores atípicos que são representados na sua extremidade esquerda, com destaque aqui para os municípios de Jequitibá, de Minas Gerais, com um índice de 0,0, e São José dos Cordeiros, da Paraíba, com um índice de 0,05869.
- **Centro e Dispersão**: A mediana nos indica que há uma leve concentração de municípios do lado esquerdo, sendo que o seu valor para **PRENTRAB** é de 0,63626. O **PRENTRAB** médio dos municípios é de 0,60539, tendo um desvio-padrão de 0,15923. Com 95% de confiança, podemos afirmar que a média encontra-se entre os índices de 0,060121 e 0,60958.

3.2.14 Variável: "T_DES2529"

Segue abaixo quadro contendo Histograma, Curva de Densidade, Box-Plot, Intervalo de confiança da média e mediana, além das medidas numéricas como média, desvio-padrão, variância, quantidade de observações, valores mínimos, máximos e o teste de normalidade de Anderson-Darling (A-Squared e P-Value), para a variável "**T_DES2529**".

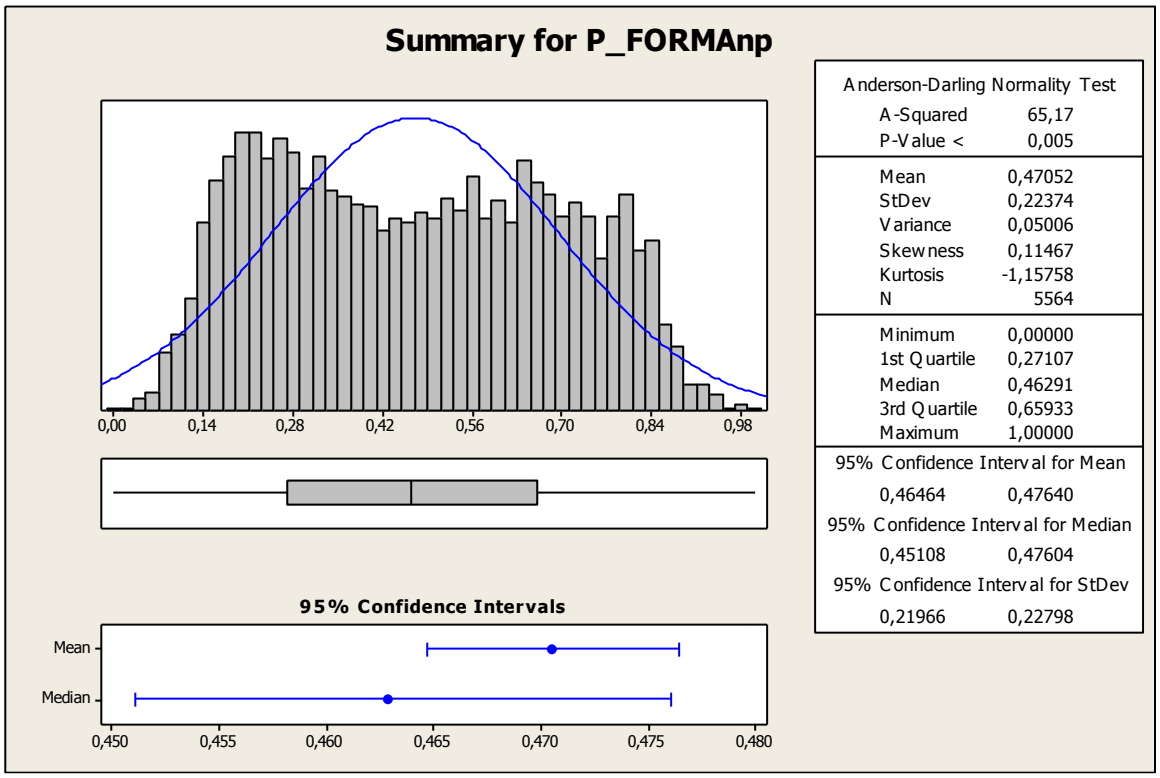


Algumas observações que podemos fazer:

- **Forma**: O Histograma nos apresenta uma distribuição concentrada na faixa entre os percentuais de 0,56 e 0,99. O Box-Plot demonstra a concentração na faixa acima citada, o posicionamento da mediana confirma o posicionamento.
- **Valores Atípicos**: temos alguns valores atípicos que são representados na sua extremidade esquerda, com destaque aqui para os municípios de Riacho da Cruz, do Rio Grande do Norte, com um índice de 0,0, e Capo Alegre de Fidalgo, do Piauí, com um índice de 0,01218.
- **Centro e Dispersão**: A mediana nos indica que há uma distribuição proporcional de municípios entre os dois lados, sendo que o seu valor para **T_DES2529** é de 0,84216. O **T_DES2529** médio dos municípios é de 0,82568, tendo um desvio-padrão de 0,11380. Com 95% de confiança, podemos afirmar que a média encontra-se entre os índices de 0,82268 e 0,82867.

3.2.15 Variável: "P_FORMAL"

Segue abaixo quadro contendo Histograma, Curva de Densidade, Box-Plot, Intervalo de confiança da média e mediana, além das medidas numéricas como média, desvio-padrão, variância, quantidade de observações, valores mínimos, máximos e o teste de normalidade de Anderson-Darling (A-Squared e P-Value), para a variável "**P_FORMAL**".

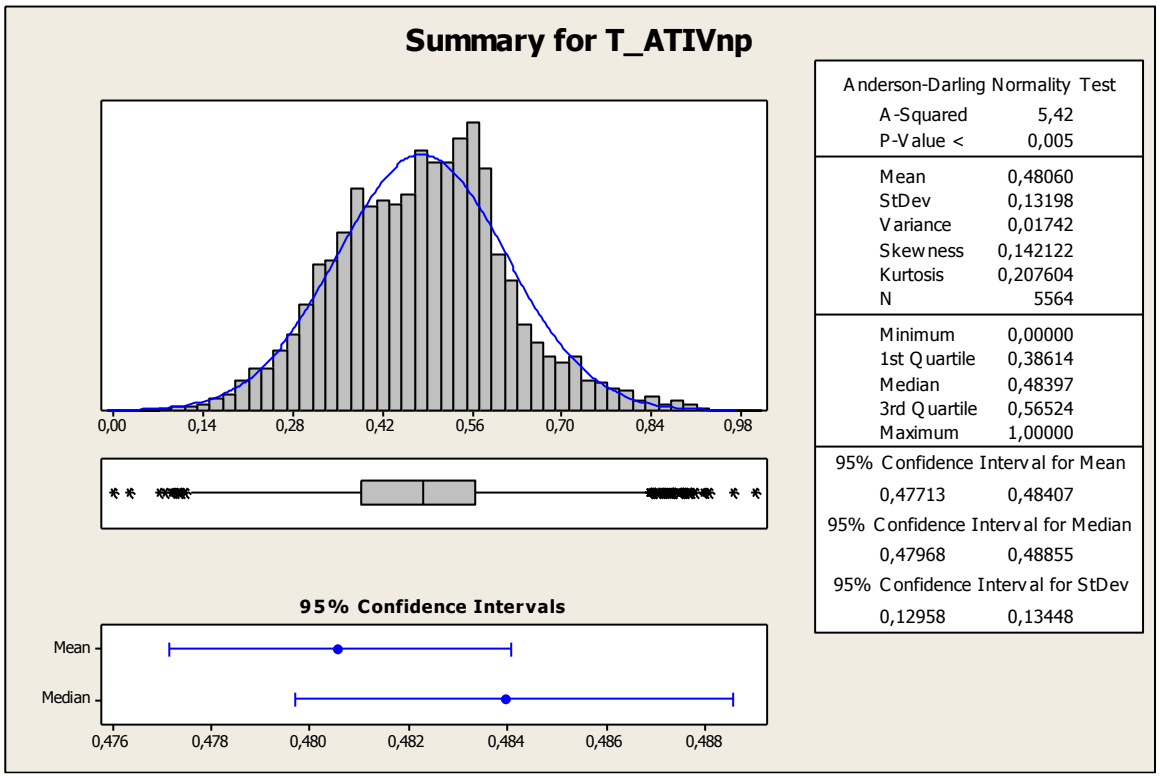


Algumas observações que podemos fazer:

- **Forma**: O Histograma nos apresenta uma distribuição concentrada na faixa entre os percentuais de 0,10 e 0,84. O Box-Plot demonstra a concentração na faixa acima citada, o posicionamento da mediana confirma o posicionamento.
- **Valores Atípicos**: não há dados atípicos.
- **Centro e Dispersão**: A mediana nos indica que há uma distribuição proporcional de municípios entre os dois lados, sendo que o seu valor para **P_FORMAL** é de 0,46291. O **P_FORMAL** médio dos municípios é de 0,47052, tendo um desvio-padrão de 0,22374. Com 95% de confiança, podemos afirmar que a média encontra-se entre os índices de 0,46464 e 0,47640.

3.2.16 Variável: "T_ATIV"

Segue abaixo quadro contendo Histograma, Curva de Densidade, Box-Plot, Intervalo de confiança da média e mediana, além das medidas numéricas como média, desvio-padrão, variância, quantidade de observações, valores mínimos, máximos e o teste de normalidade de Anderson-Darling (A-Squared e P-Value), para a variável "**T_ATIV**".



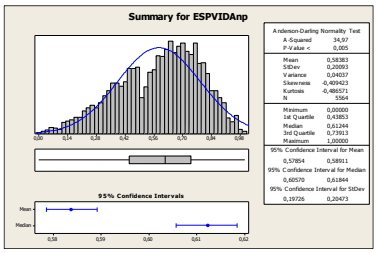
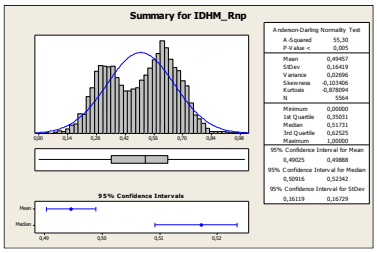
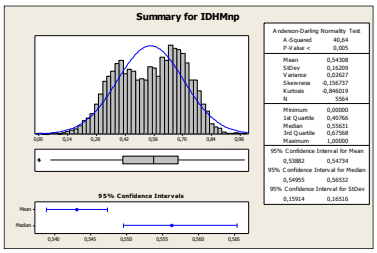
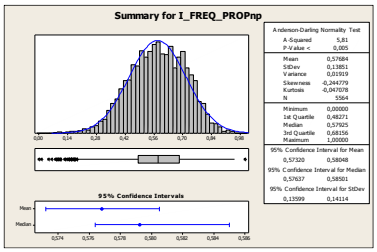
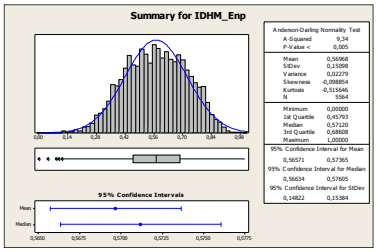
Algumas observações que podemos fazer:

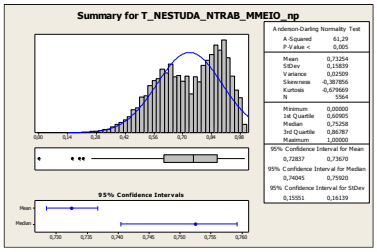
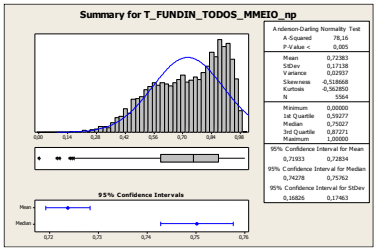
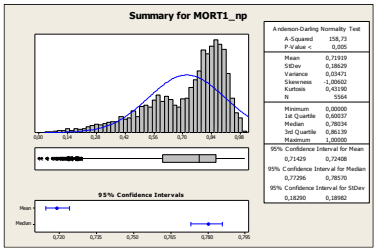
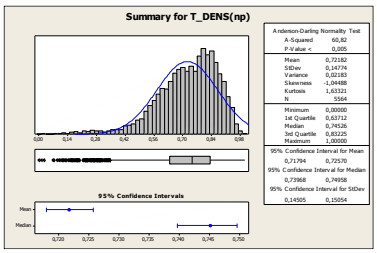
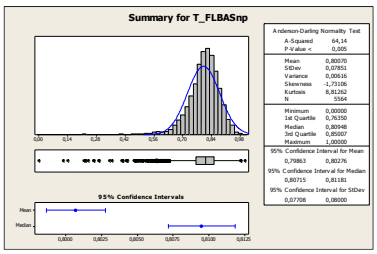
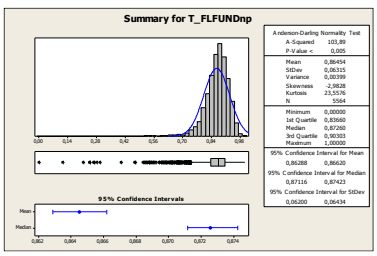
- **Forma**: O Histograma nos apresenta uma distribuição concentrada na faixa entre os percentuais de 0,28 e 0,70. O Box-Plot demonstra a concentração na faixa acima citada, o posicionamento da mediana confirma o posicionamento.
- **Valores Atípicos**: temos alguns valores atípicos que são representados nas duas extremidades. Na extremidade esquerda temos destaque aqui para os municípios de João Dias, do Rio Grande do Norte, com um índice de 0,0, e Santana do Maranhão com um índice de 0,02497. Do lado direito temos destaque para os municípios de Coronel Pilar com um índice de 1, e Itapuca com um índice de 0,96645, ambos do Estado do Rio Grande do Sul.
- **Centro e Dispersão**: A mediana nos indica que há uma distribuição proporcional de municípios entre os dois lados, sendo que o seu valor para **T_ATIV** é de 0,48397. O **T_ATIV** médio dos municípios é de 0,48060, tendo um desvio-padrão de 0,13198. Com 95% de confiança, podemos afirmar que a média encontra-se entre os índices de 0,47713 e 0,48407.

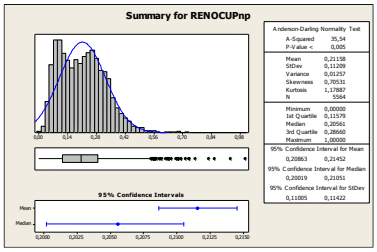
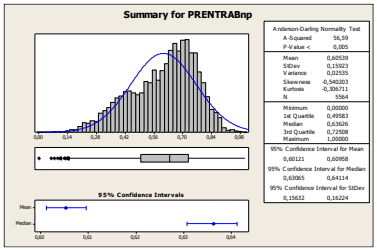
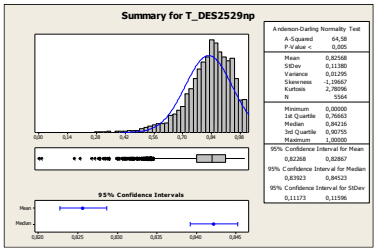
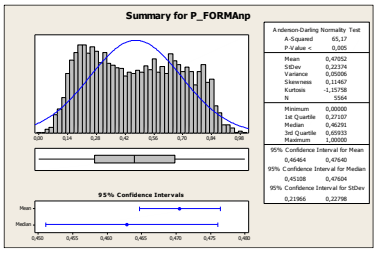
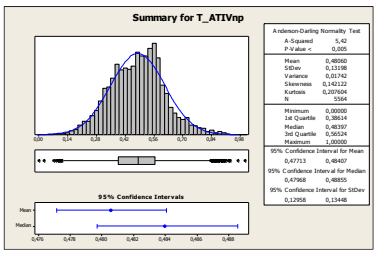
4. ANÁLISE COMPARATIVA DA ANÁLISE DESCRITIVA

Abaixo apresentamos uma tabela comparando Histograma, Box-Plot, Curva de Densidade, média, desvio-padrão, mediana e P-Value do teste de normalidade de Anderson-Darling, das variáveis quantitativas analisadas.

Tabela 2. Analise Comparativa das Variáveis (Dimensões Desenvolvimento Humano)

VARIÁVEL	GRÁFICO	MÉDIA	DESVIO_PADRÃO	MEDIANA	P-VALUE
ESPVIDA		**0,58383**	**0,20093**	**0,61244**	**0,005**
IDHM_R		**0,49457**	**0,16419**	**0,51731**	**0,005**
IDHM		**0,54308**	**0,16209**	**0,55631**	**0,005**
I_FREQ_PROP		**0,57684**	**0,13851**	**0,57925**	**0,005**
IDHM_E		**0,56968**	**0,15098**	**0,57120**	**0,005**

T_NESTUDA_NTRAB_MMEIO		**0,73254**	**0,15839**	**0,75258**	**0,005**
T_FUNDIN_TODOS_MMEIO		**0,72383**	**0,17138**	**0,75027**	**0,005**
MORT_1		**0,71919**	**0,18629**	**0,78034**	**0,005**
T_DENS		**0,72182**	**0,14774**	**0,74526**	**0,005**
T_FLBAS		**0,80070**	**0,07851**	**0,80948**	**0,005**
T_FLFUND		**0,86454**	**0,06315**	**0,87260**	**0,005**

RENOCUP		**0,21158**	**0,11209**	**0,20561**	**0,005**
PRENTRAB		**0,60539**	**0,15923**	**0,63626**	**0,005**
T_DES2529		**0,82568**	**0,11380**	**0,84216**	**0,005**
P_FORMAL		**0,47052**	**0,22374**	**0,46291**	**0,005**
T_ATIV		**0,48060**	**0,13198**	**0,48397**	**0,005**

A tabela 2 nos mostra uma visão geral das dimensões e variáveis já apresentadas e analisadas individualmente nos tópicos anteriores. As variáveis que representam aspectos relacionados a educação demonstram melhores resultados nos primeiros anos de vida, ou seja, T_FUNDIN_TODOS_MMEIO, T_FLBAS e T_FLFUND. A P_FORMAL e ESPVIDA possuem uma distribuição mais simétrica, assim como as variáveis IDHM e IDHM_L.

Vale ressaltar que a simetria ou não das distribuições não necessariamente tem relação com a qualidade ou validade dos dados trabalhados. Distribuições assimétricas podem, por exemplo,

nos indicar onde devemos focar ou concentrar esforços para a obtenção de resultados esperados de forma mais eficiente.

5. CORRELAÇÃO DAS VARIÁVEIS

Os dados abaixo representam a correlação entre as variáveis selecionadas e já trabalhadas anteriormente.

Correlations: MORT1_np; T_NESTUDA_NT; T_FUNDIN_TOD; T_DENS(np); ESPVIDAnp; ...

	MORT1_np	T_NESTUDA_NTRAB_	T_FUNDIN_TODOS_M
T_NESTUDA_NTRAB_	0,741		
	0,000		
T_FUNDIN_TODOS_M	0,796	0,835	
	0,000	0,000	
T_DENS(np)	0,594	0,670	0,662
	0,000	0,000	0,000
ESPVIDAnp	0,967	0,753	0,802
	0,000	0,000	0,000
I_FREQ_PROPnp	0,636	0,698	0,853
	0,000	0,000	0,000
IDHMnp	0,829	0,837	0,962
	0,000	0,000	0,000
IDHM_Enp	0,684	0,722	0,921
	0,000	0,000	0,000
IDHM_Rnp	0,814	0,877	0,906
	0,000	0,000	0,000
T_FLBASnp	-0,074	0,071	0,061
	0,000	0,000	0,000
T_FLFUNDnp	-0,089	0,053	0,032
	0,000	0,000	0,017
RENOCUPnp	0,705	0,721	0,792
	0,000	0,000	0,000
PRENTRABnp	0,593	0,618	0,620
	0,000	0,000	0,000
P_FORMAnp	0,678	0,697	0,802
	0,000	0,000	0,000
T_ATIVnp	0,663	0,877	0,742
	0,000	0,000	0,000
T_DES2529np	0,373	0,526	0,340
	0,000	0,000	0,000

	T_DENS(np)	ESPVIDAnp	I_FREQ_PROPnp
ESPVIDAnp	0,589		
	0,000		
I_FREQ_PROPnp	0,689	0,641	
	0,000	0,000	
IDHMnp	0,646	0,852	0,862
	0,000	0,000	0,000
IDHM_Enp	0,555	0,704	0,914
	0,000	0,000	0,000

IDHM_Rnp	0,663	0,834	0,729
	0,000	0,000	0,000
T_FLBASnp	0,161	-0,068	0,255
	0,000	0,000	0,000
T_FLFUNDnp	0,198	-0,069	0,178
	0,000	0,000	0,000
RENOCUPnp	0,430	0,742	0,589
	0,000	0,000	0,000
PRENTRABnp	0,245	0,585	0,385
	0,000	0,000	0,000
P_FORMAnp	0,470	0,709	0,620
	0,000	0,000	0,000
T_ATIVnp	0,676	0,662	0,642
	0,000	0,000	0,000
T_DES2529np	0,395	0,372	0,327
	0,000	0,000	0,000

	IDHMnp	IDHM_Enp	IDHM_Rnp
IDHM_Enp	0,951		
	0,000		
IDHM_Rnp	0,948	0,819	
	0,000	0,000	
T_FLBASnp	0,068	0,159	-0,017
	0,000	0,000	0,196
T_FLFUNDnp	0,026	0,079	-0,025
	0,055	0,000	0,059
RENOCUPnp	0,869	0,765	0,915
	0,000	0,000	0,000
PRENTRABnp	0,638	0,531	0,689
	0,000	0,000	0,000
P_FORMAnp	0,824	0,761	0,810
	0,000	0,000	0,000
T_ATIVnp	0,732	0,612	0,790
	0,000	0,000	0,000
T_DES2529np	0,334	0,234	0,393
	0,000	0,000	0,000

	T_FLBASnp	T_FLFUNDnp	RENOCUPnp
T_FLFUNDnp	0,800		
	0,000		
RENOCUPnp	-0,101	-0,121	
	0,000	0,000	
PRENTRABnp	-0,221	-0,222	0,753
	0,000	0,000	0,000
P_FORMAnp	-0,043	-0,042	0,802
	0,001	0,002	0,000
T_ATIVnp	0,063	0,066	0,576

	0,000	0,000	0,000
T_DES2529np	-0,022	-0,004	0,247
	0,101	0,755	0,000

	PRENTRABnp	P_FORMAnp	T_ATIVnp
P_FORMAnp	0,631		
	0,000		
T_ATIVnp	0,553	0,534	
	0,000	0,000	
T_DES2529np	0,270	0,185	0,512
	0,000	0,000	0,000

Cell Contents: Pearson correlation
P-Value

Com base nas informações constantes na matriz de correlação pode-se perceber que as variáveis descritas na Tabela 3 apresentam forte relação, vale destacar que isto não significa que elas apresentam causalidade, ou seja, um sentido direto entre elas. Ressalta-se ainda que para esta análise foram selecionadas aquelas com de correlação maior que 0,800.

TABELA 3 – CORRELAÇÃO DAS VARIÁVEIS

Variável	Variável	Grau de Correlação
ESPVIDAnp	MORT1_np	0,967
IDHM_Enp	IDHMnp	0,951
IDHM_Rnp	IDHMnp	0,948
IDHM_Enp	T_FUNDIN_TODOS_M	0,921
RENOCUPnp	IDHM_Rnp	0,915
ESPVIDAnp	T_FUNDIN_TODOS_M	0,913
IDHM_Rnp	T_FUNDIN_TODOS_M	0,906
T_ATIVnp	T_NESTUDA_NTRAB_	0,877
RENOCUPnp	IDHMnp	0,869
I_FREQ_PROPnp	T_FUNDIN_TODOS_M	0,853
IDHMnp	ESPVIDAnp	0,852
T_FUNDIN_TODOS_M	T_NESTUDA_NTRAB	0,835
IDHM_Rnp	ESPVIDAnp	0,834
P_FORMAnp	IDHMnp	0,824
IDHM_Rnp	IDHM_Enp	0,819
P_FORMAnp	RENOCUPnp	0,802

Fonte: elaborado pelo autor, 2014.

6. DENDROGRAMA

Um **Dendrograma** (dendr(o) = árvore) é um tipo específico de diagrama ou representação icônica que organiza determinados fatores e variáveis. Isto quer dizer que sua representação apresenta um diagrama de similaridade.

A interpretação de um Dendrograma de similaridade entre amostras fundamenta-se na intuição: duas amostras próximas devem ter também valores semelhantes para as variáveis medidas. Ou seja, elas devem ser próximas matematicamente no espaço multidimensional.

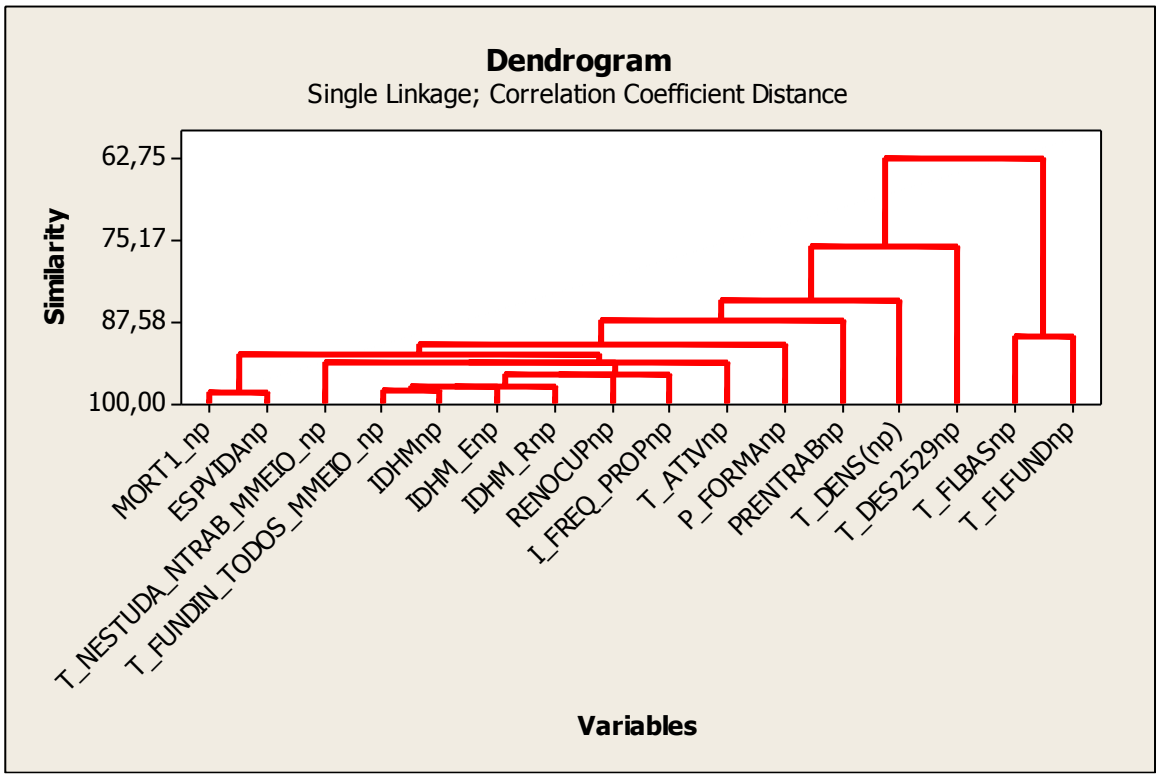
Portanto, quanto maior a proximidade entre as medidas relativas às amostras, maior a similaridade entre elas. O dendrograma hierarquiza esta similaridade de modo que podemos ter uma visão bidimensional da similaridade ou dissimilaridade de todo o conjunto de amostras utilizado no estudo.

Cluster Analysis of Variables: MORT1_np; T_NESTUDA_NT; T_FUNDIN_TOD; ...

Correlation Coefficient Distance, Single Linkage
Amalgamation Steps

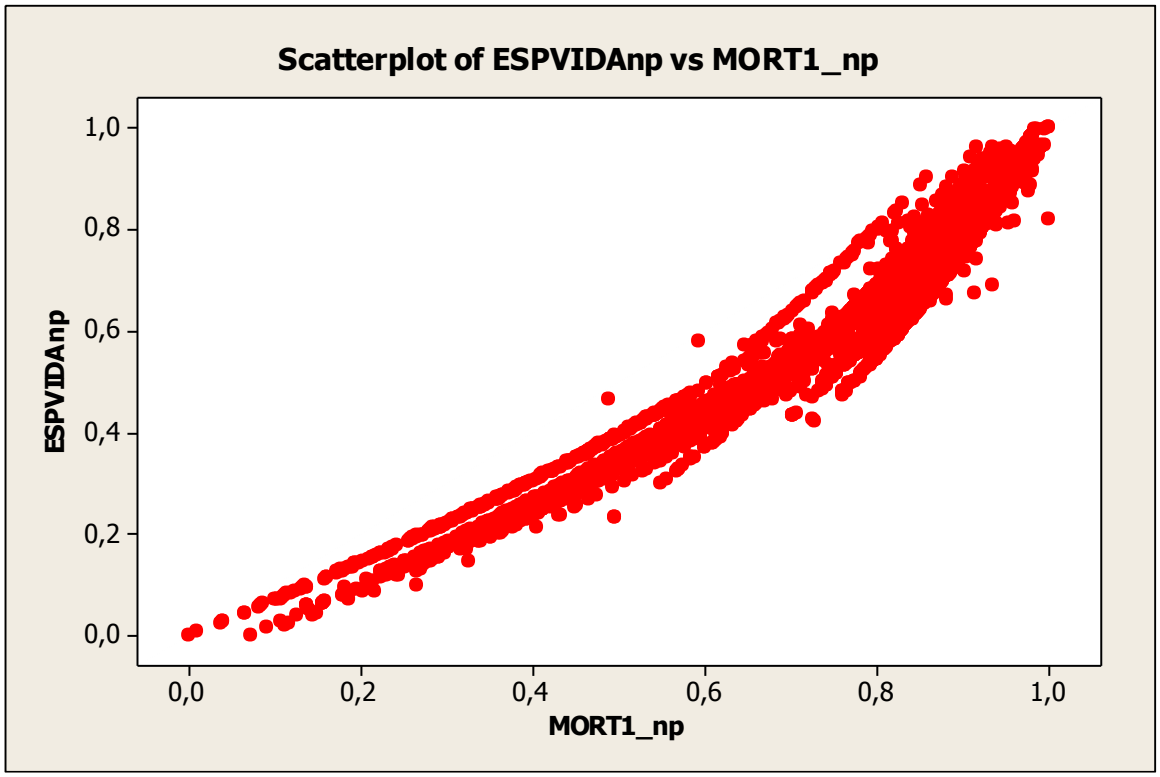
Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	15	98,3572	0,032857	1	5	1	2
2	14	98,0777	0,038445	3	7	3	2
3	13	97,5422	0,049155	3	8	3	3
4	12	97,4140	0,051720	3	9	3	4
5	11	95,7702	0,084596	3	12	3	5
6	10	95,6766	0,086467	3	6	3	6
7	9	93,8677	0,122646	2	3	2	7
8	8	93,8673	0,122655	2	15	2	8
9	7	92,6019	0,147962	1	2	1	10
10	6	91,1857	0,176286	1	14	1	11
11	5	89,9867	0,200266	10	11	10	2
12	4	87,6672	0,246655	1	13	1	12
13	3	84,4718	0,310564	1	4	1	13
14	2	76,3215	0,473569	1	16	1	14
15	1	62,7547	0,744907	1	10	1	16

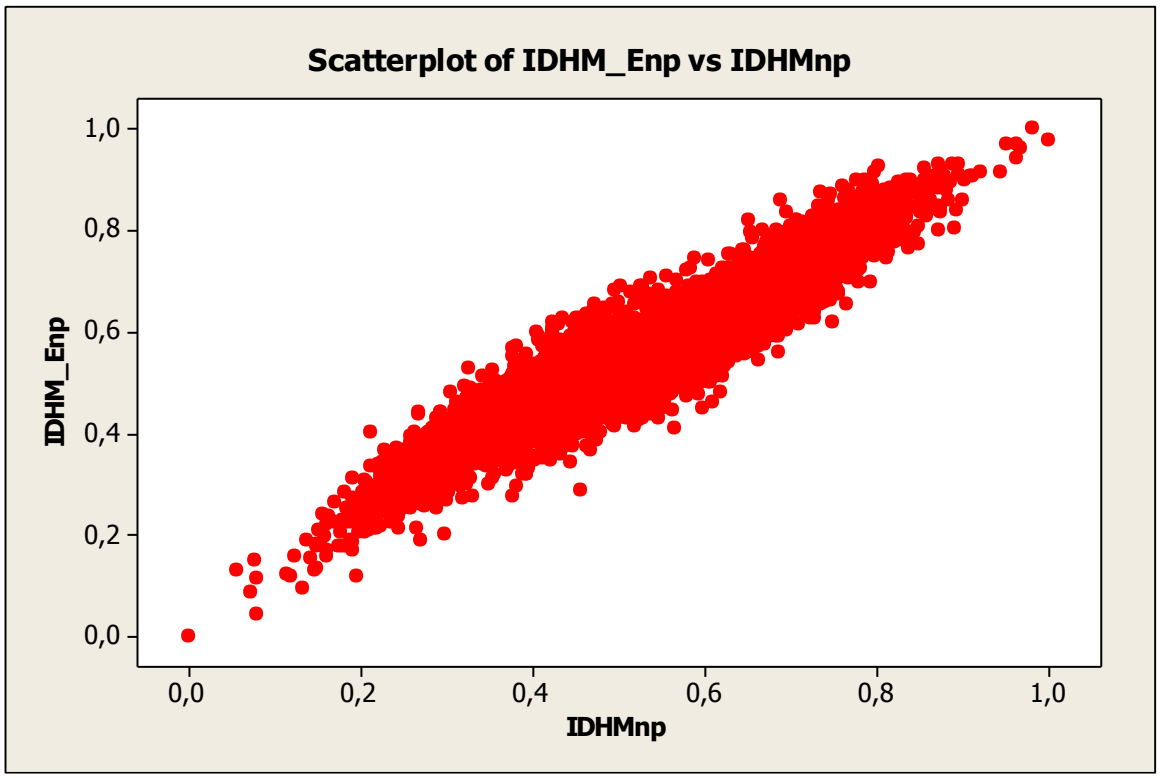
Segue abaixo o Dendrograma das variáveis analisadas:

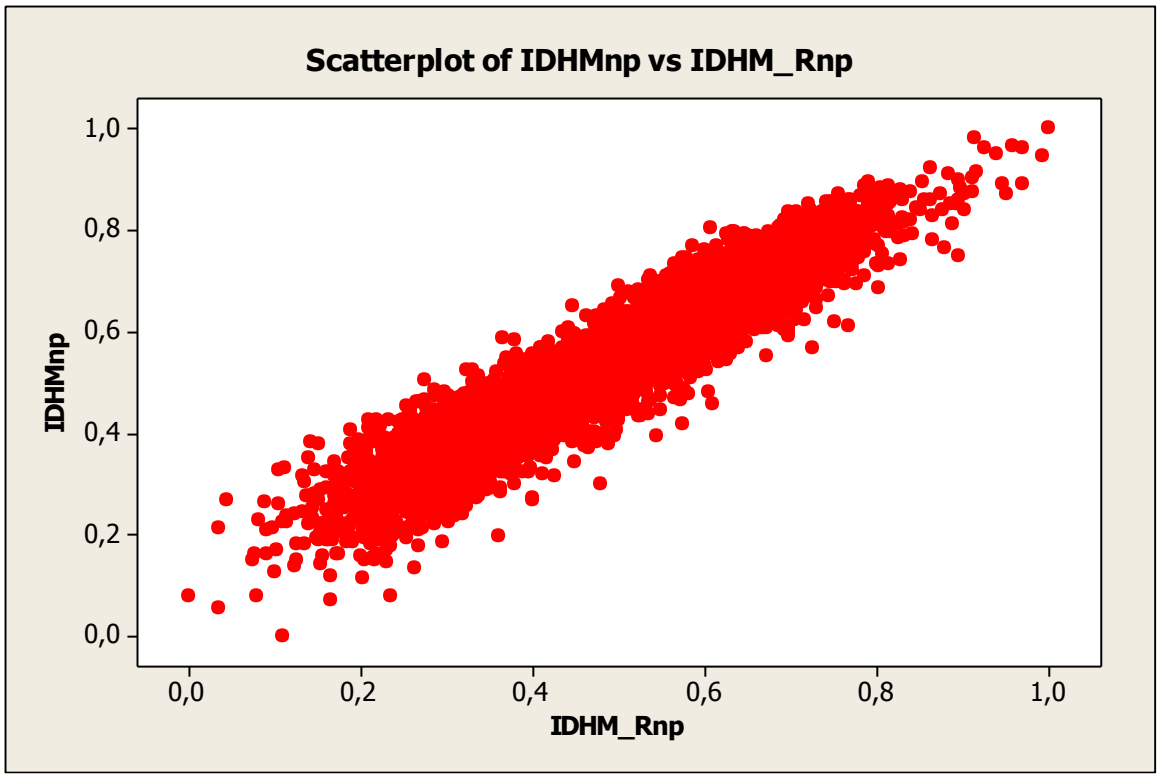

Dendrogram
Single Linkage; Correlation Coefficient Distance
Similarity
62,75
75,17
87,58
100,00
MORT1_np
ESPVIDAnp
T_NESTUDA_NTRAB_MMEIO_np
T_FUNDIN_TODOS_MMEIO_np
IDHMnp
IDHM_Enp
IDHM_Rnp
RENOCUPnp
I_FREQ_PROPnp
T_ATIVnp
P_FORMAnp
PRENTRABnp
T_DENS(np)
T_DES2529np
T_FLBASnp
T_FLFUNDnp
Variables

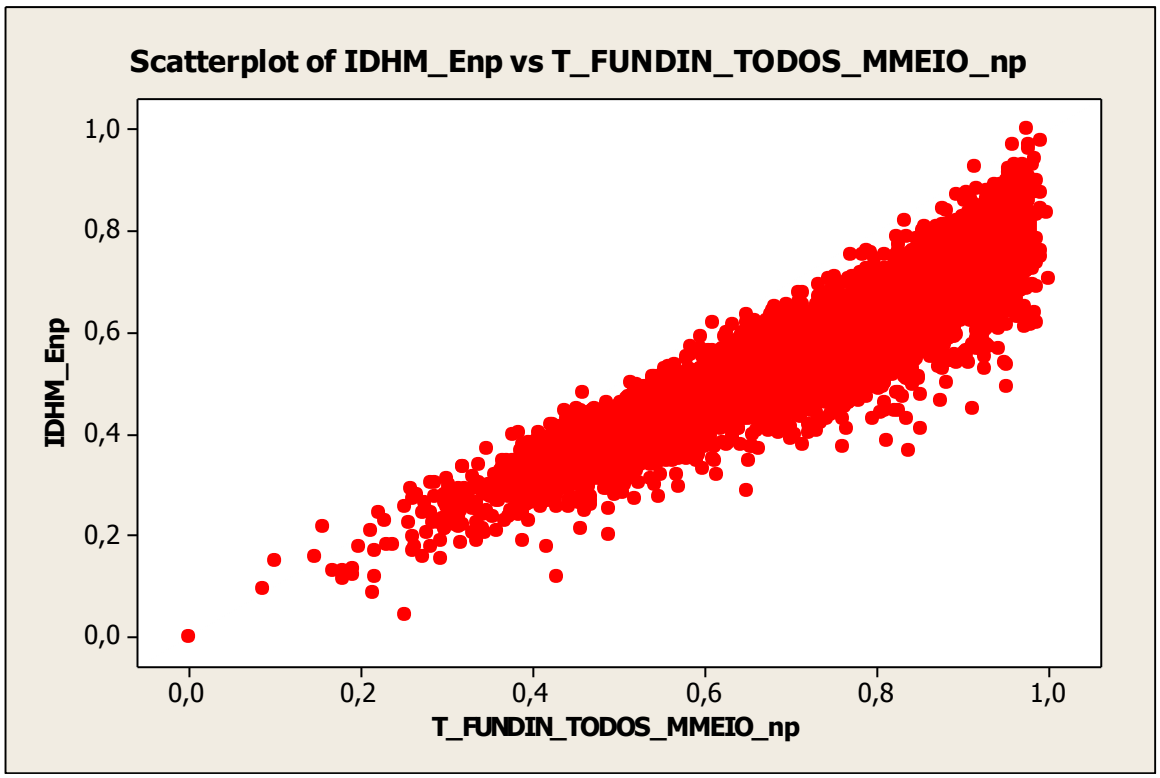
7. GRÁFICOS DE DISPERSÃO

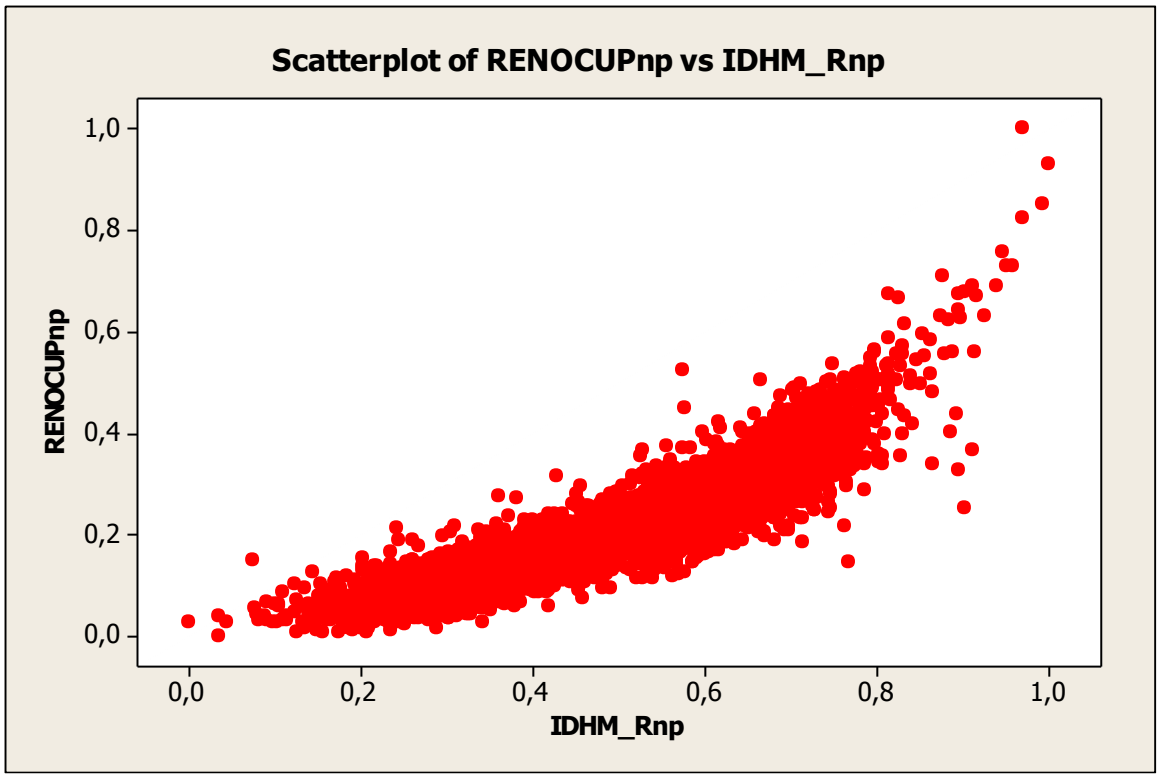
Nos gráficos abaixo são apresentadas as relações entre as variáveis relacionadas na tabela 3.

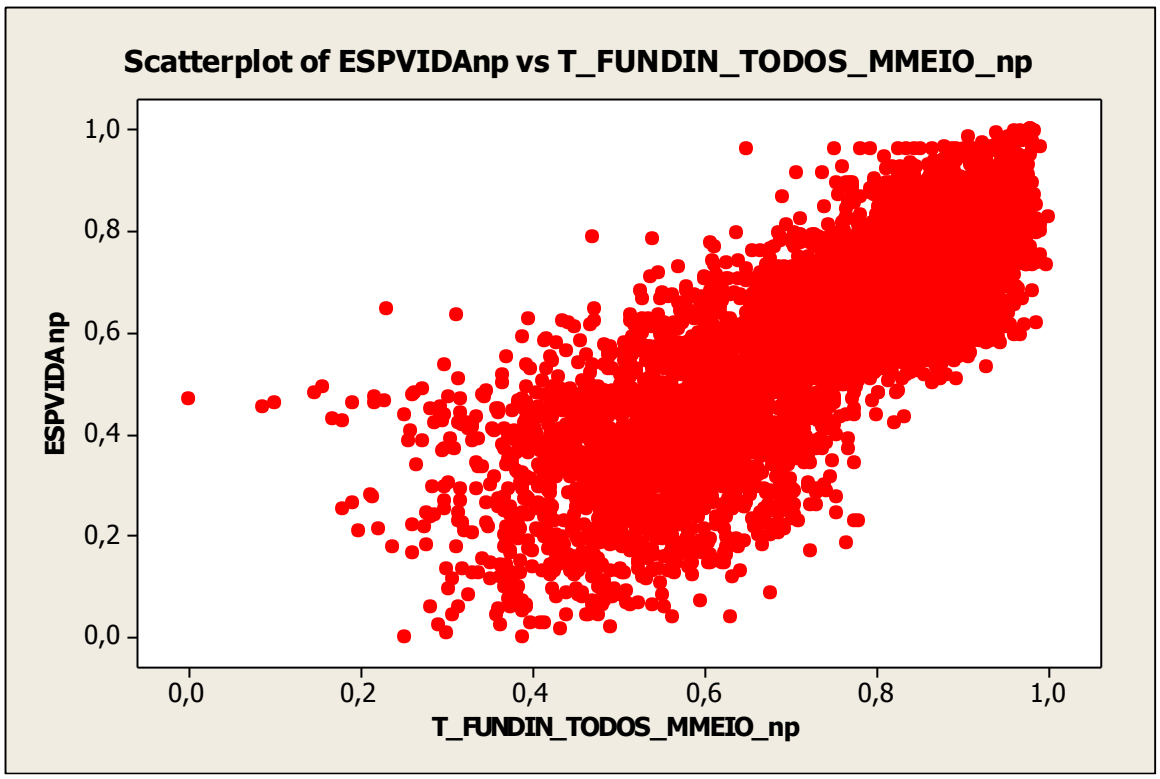


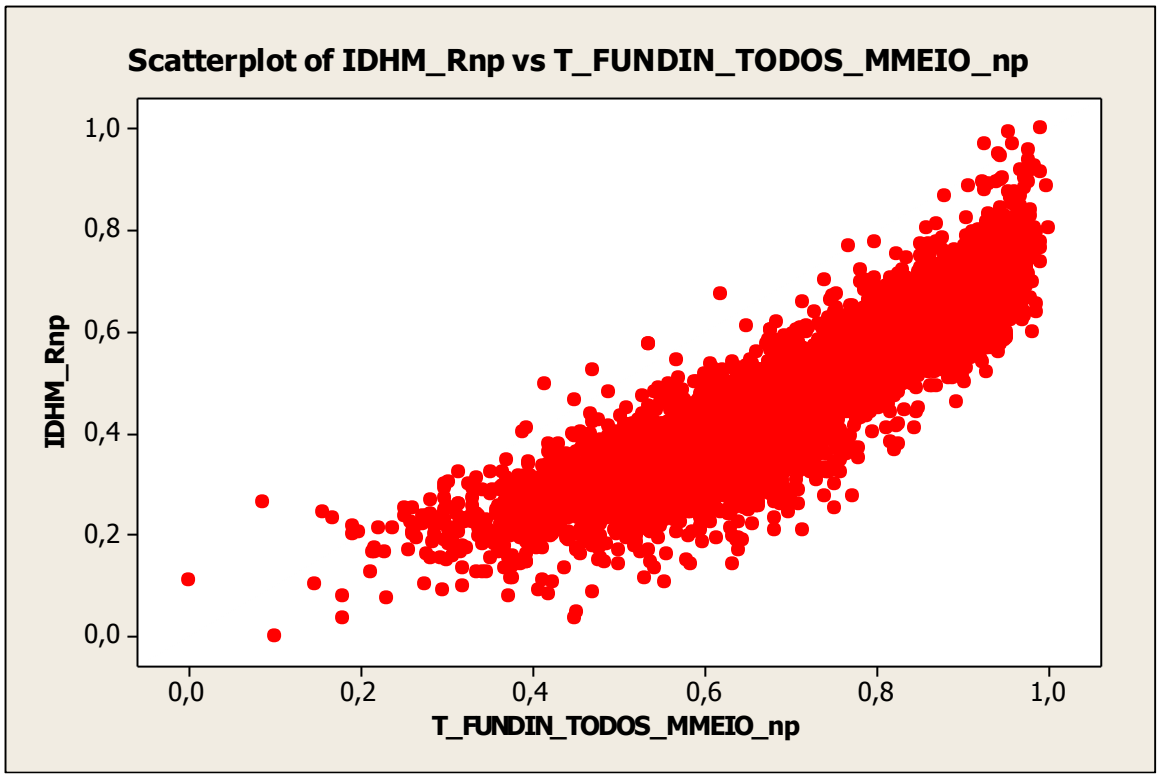


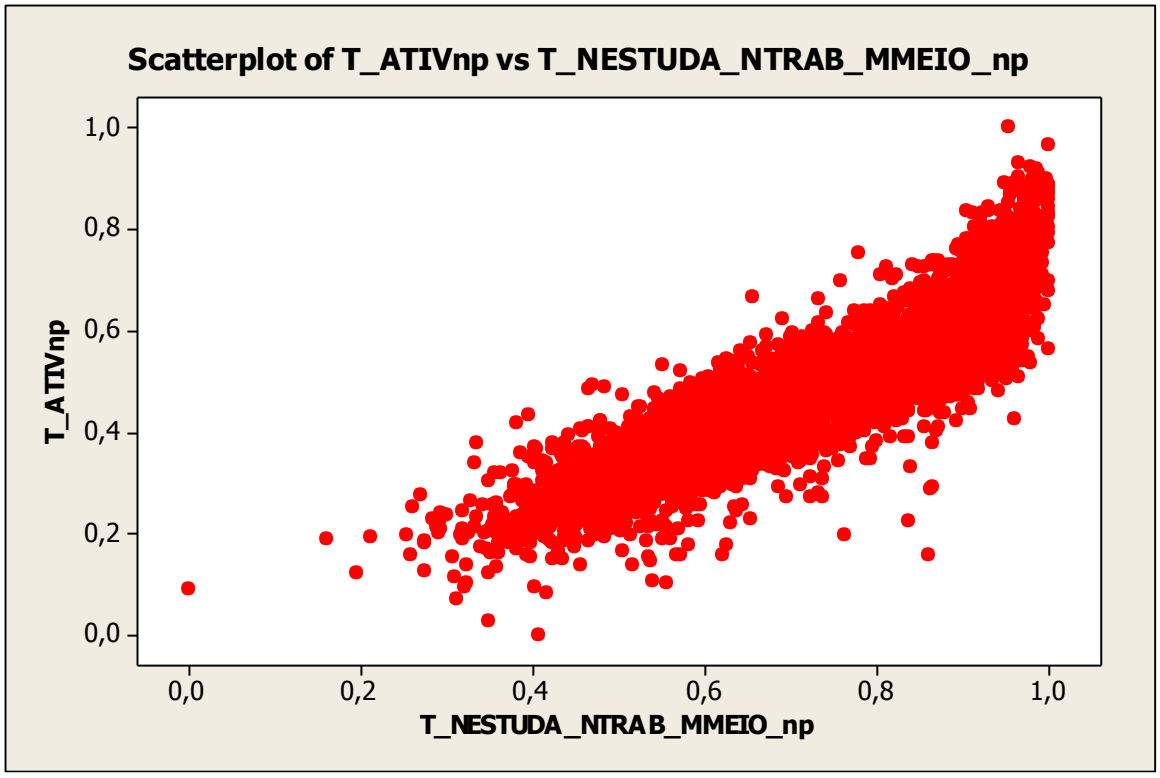

Scatterplot of IDHMnp vs IDHM_Rnp
IDHMnp
IDHM_Rnp
1,0
0,8
0,6
0,4
0,2
0,0
0,0
0,2
0,4
0,6
0,8
1,0

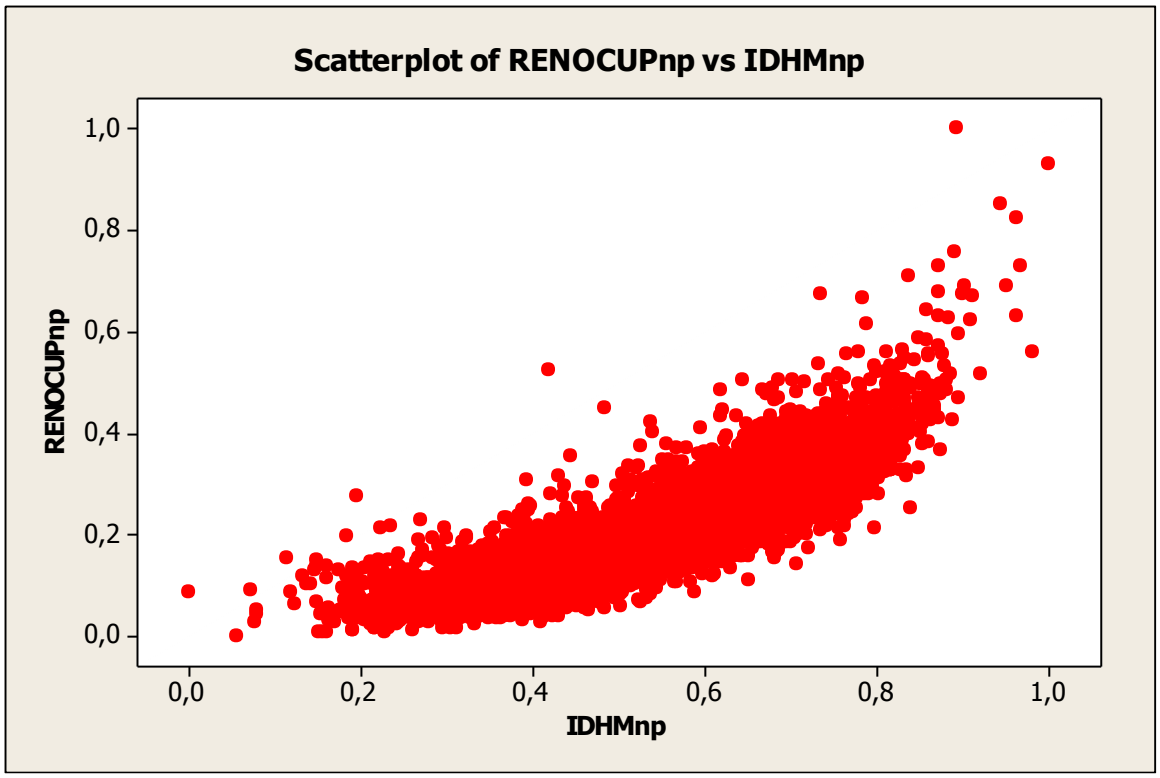

Scatterplot of IDHM_Enp vs T_FUNDIN_TODOS_MMEIO_np
IDHM_Enp
T_FUNDIN_TODOS_MMEIO_np
1,0
0,8
0,6
0,4
0,2
0,0
0,0
0,2
0,4
0,6
0,8
1,0

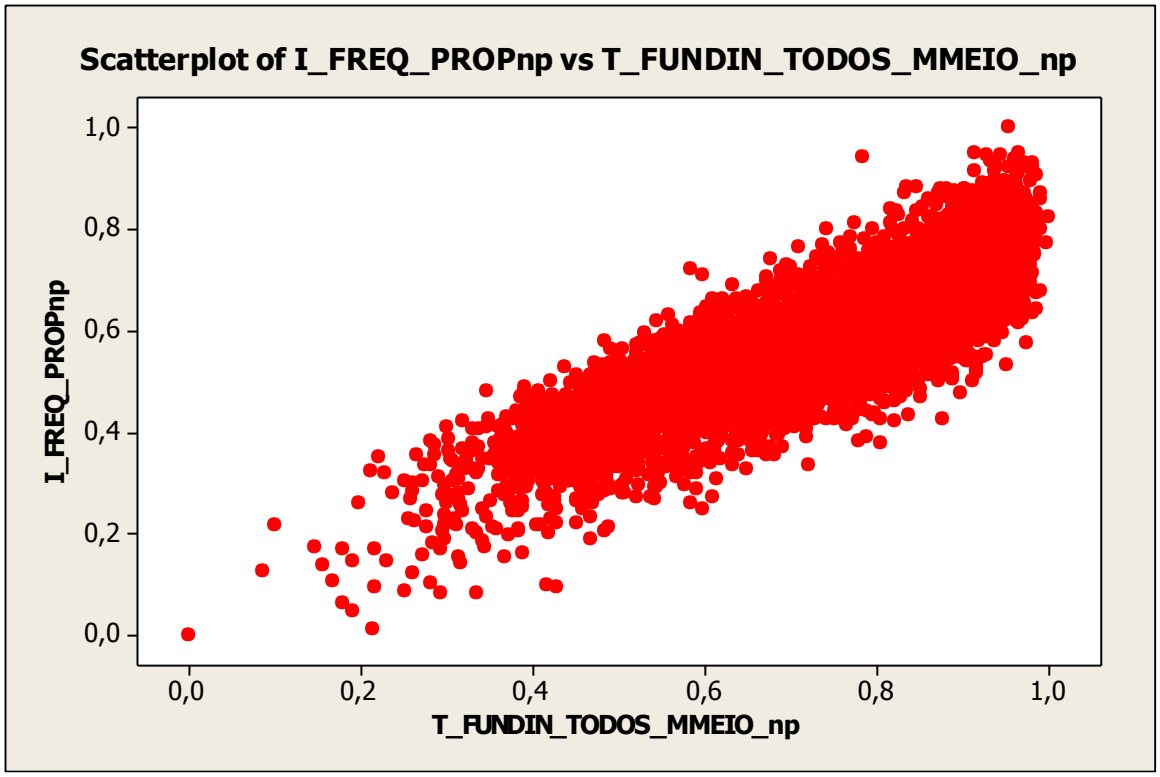

Scatterplot of RENOCUPnp vs IDHM_Rnp
RENOCUPnp
IDHM_Rnp
1,0
0,8
0,6
0,4
0,2
0,0
0,0
0,2
0,4
0,6
0,8
1,0

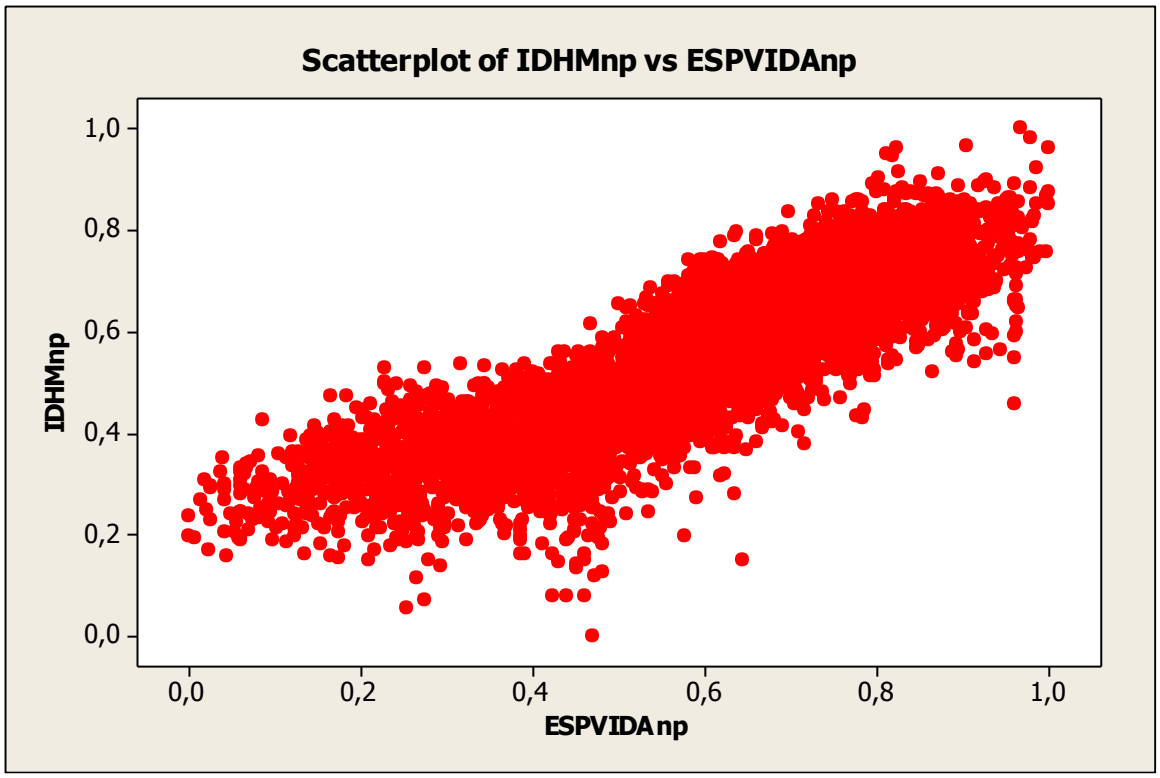

Scatterplot of ESPVIDAnp vs T_FUNDIN_TODOS_MMEIO_np
ESPVIDAnp
T_FUNDIN_TODOS_MMEIO_np
1,0
0,8
0,6
0,4
0,2
0,0
0,0
0,2
0,4
0,6
0,8
1,0

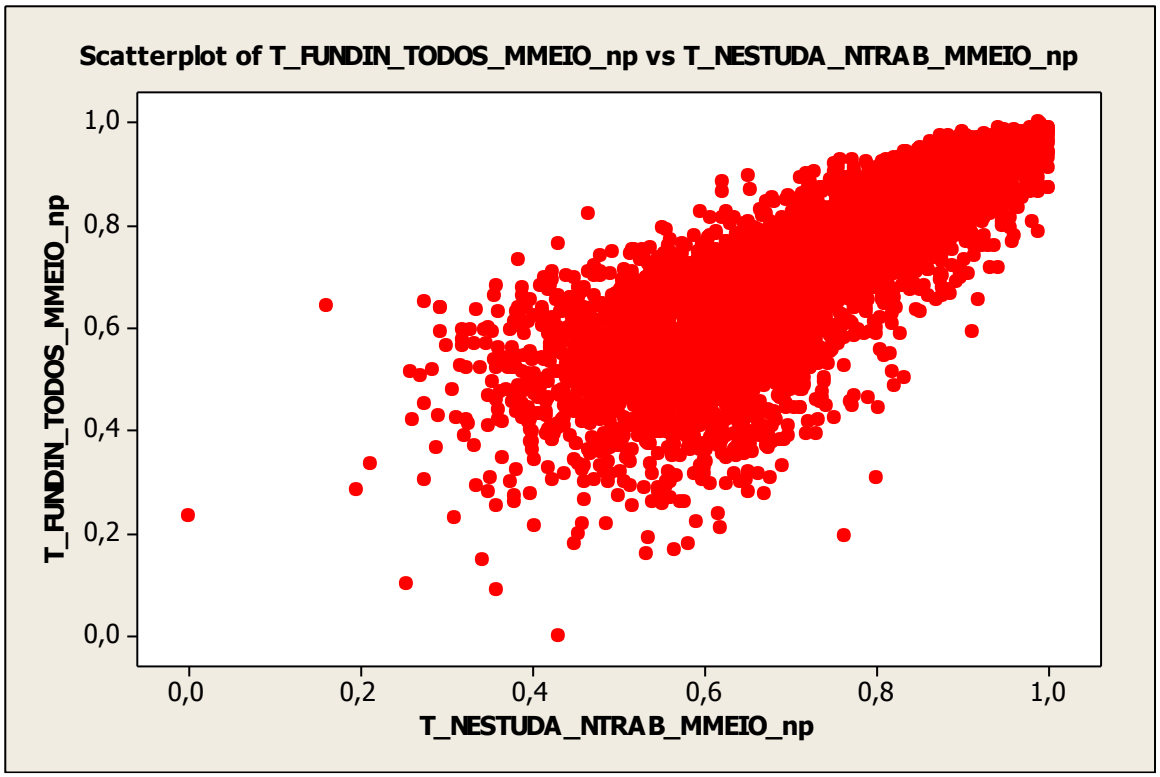

Scatterplot of IDHM_Rnp vs T_FUNDIN_TODOS_MMEIO_np
IDHM_Rnp
1,0
0,8
0,6
0,4
0,2
0,0
0,0
0,2
0,4
0,6
0,8
1,0
T_FUNDIN_TODOS_MMEIO_np

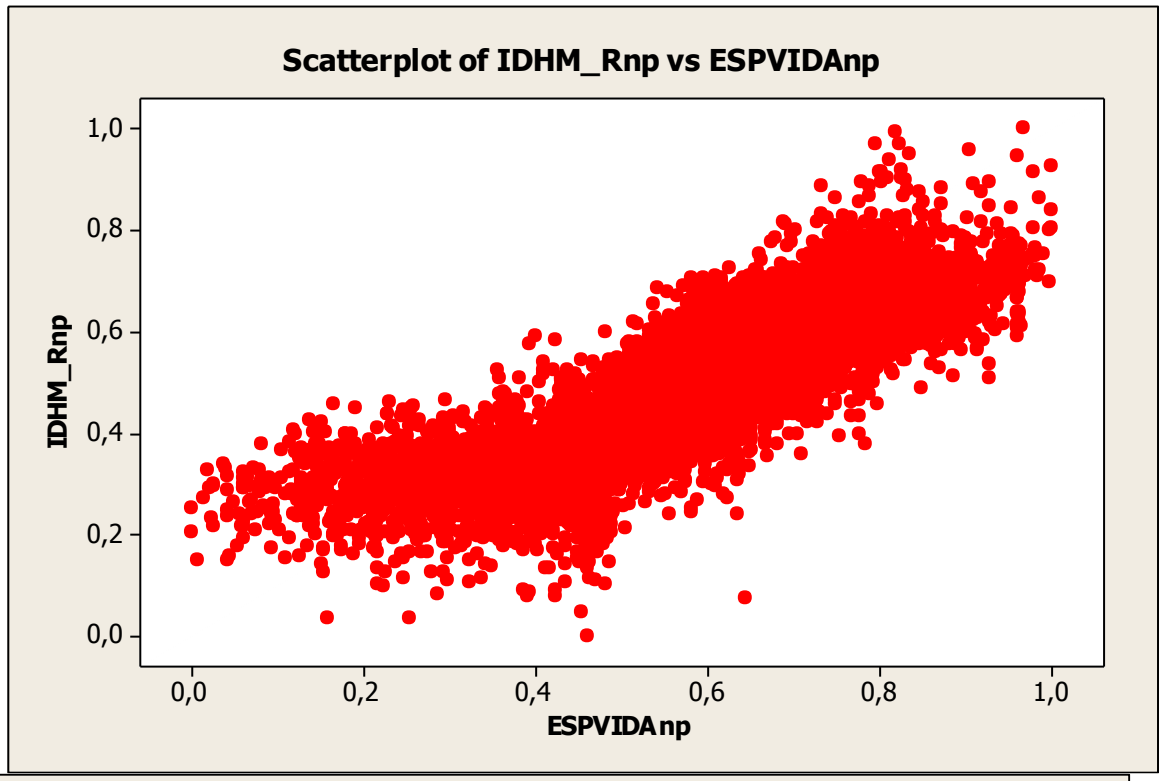

Scatterplot of T_ATIVnp vs T_NESTUDA_NTRAB_MMEIO_np
T_ATIVnp
1,0
0,8
0,6
0,4
0,2
0,0
0,0
0,2
0,4
0,6
0,8
1,0
T_NESTUDA_NTRAB_MMEIO_np

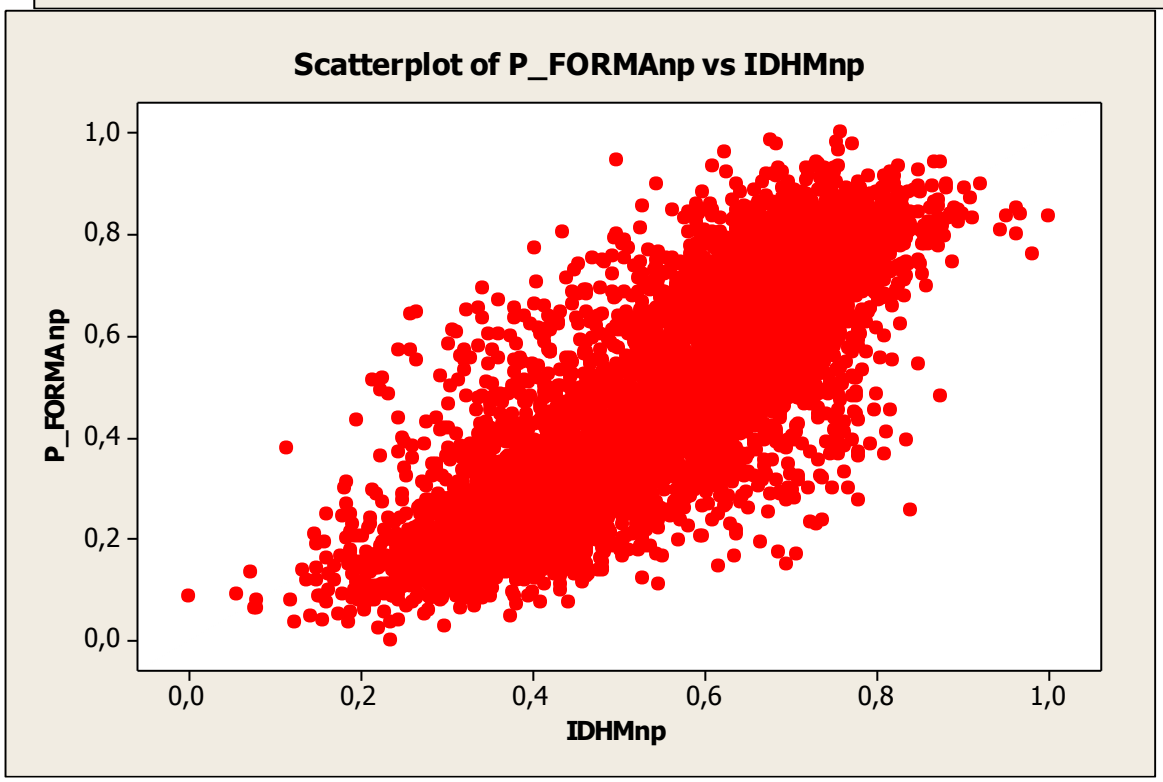

Scatterplot of RENOCUPnp vs IDHMnp
RENOCUPnp
IDHMnp
1,0
0,8
0,6
0,4
0,2
0,0
0,0
0,2
0,4
0,6
0,8
1,0

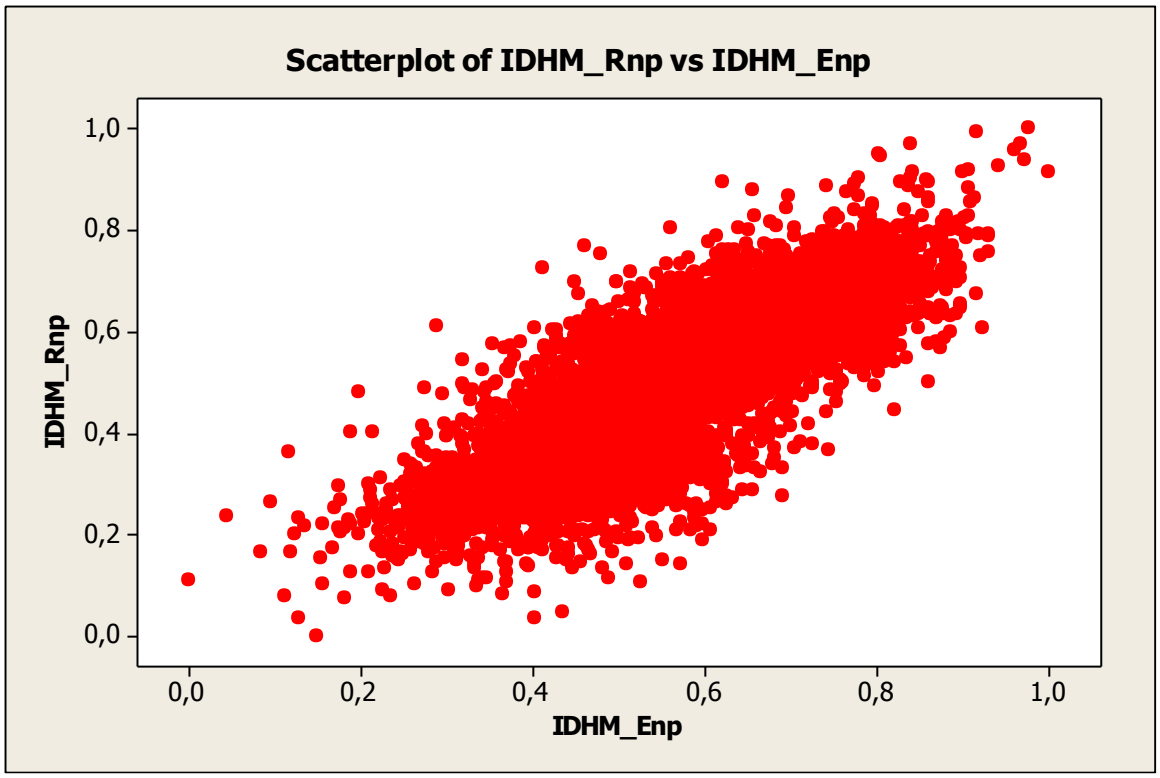

Scatterplot of I_FREQ_PROPnp vs T_FUNDIN_TODOS_MMEIO_np
I_FREQ_PROPnp
T_FUNDIN_TODOS_MMEIO_np
1,0
0,8
0,6
0,4
0,2
0,0
0,0
0,2
0,4
0,6
0,8
1,0

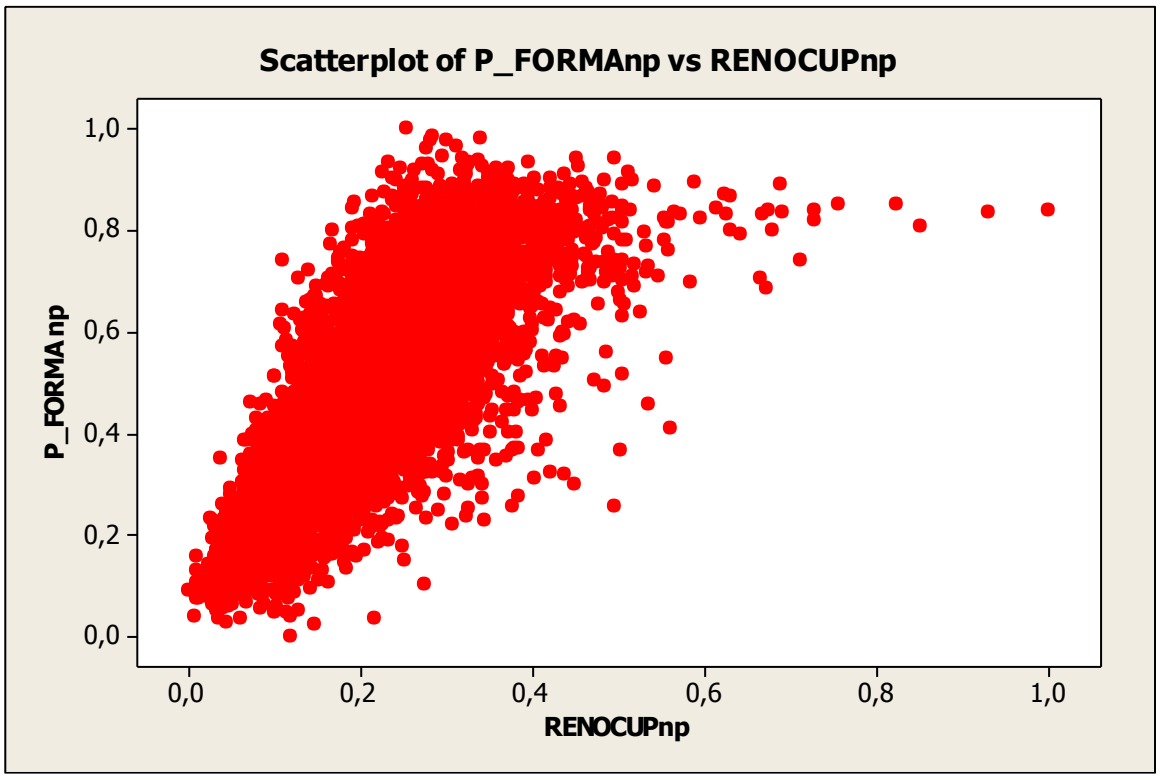

Scatterplot of IDHMnp vs ESPVIDAnp
IDHMnp
ESPVIDAnp
1,0
0,8
0,6
0,4
0,2
0,0
0,0
0,2
0,4
0,6
0,8
1,0


Scatterplot of T_FUNDIN_TODOS_MMEIO_np vs T_NESTUDA_NTRAB_MMEIO_np
T_FUNDIN_TODOS_MMEIO_np
T_NESTUDA_NTRAB_MMEIO_np
1,0
0,8
0,6
0,4
0,2
0,0
0,0
0,2
0,4
0,6
0,8
1,0


Scatterplot of IDHM_Rnp vs ESPVIDAnp
IDHM_Rnp
ESPVIDAnp
0,0
0,2
0,4
0,6
0,8
1,0


Scatterplot of P_FORMAnp vs IDHMnp
P_FORMAnp
IDHMnp
0,0
0,2
0,4
0,6
0,8
1,0


Scatterplot of IDHM_Rnp vs IDHM_Enp
IDHM_Rnp
IDHM_Enp
0,0
0,2
0,4
0,6
0,8
1,0


Scatterplot of P_FORMAnp vs RENOCUPnp
P_FORMAnp
RENOCUPnp
0,0
0,2
0,4
0,6
0,8
1,0

Inicialmente os gráficos de dispersão devem ser analisados quanto a seu padrão geral e seus desvios relativos ao padrão. A descrição do padrão geral pode ser feita pela verificação de sua forma, direção e intensidade.

Direção: Da análise das correlações acima percebemos que quase todas possuem associações positivas, ou seja, o crescimento de uma variável é acompanhado do crescimento da outra. O que nos parece é que não há nenhuma associação negativa, ao menos de evidência visual.

Intensidade: Os gráficos acima apresenta uma relação linear, mas os gráficos que relacionam ESPVIDA x MORT1; IDHM_E x IDHM; IDHM_R x IDHM; e, IDHM_E x T_FUNDIN_TODOS_MEIO possuem uma relação mais forte que as demais.

Forma: Os gráficos apresentam conglomerados que sugerem relações lineares, no entanto vale salientar a relação dos gráficos ESPVIDA x MORT1; IDHM_E x IDHM; IDHM_R x IDHM; e, IDHM_E x T_FUNDIN_TODOS_MEIO que apresentam um agrupamento mais intenso.

Valores Atípicos: Os gráficos indicam a existência de valores atípicos, ou seja, municípios que estão localizados longe dos demais. Com exceção dos gráficos P_FORMA x RENOCUP e IDHM_R x IDHM_E.

8. ANÁLISE DE REGRESSÃO DAS VARIÁVEIS COM SIMILARIDADE

A correlação mede a direção e a intensidade da relação linear (linha reta) entre duas variáveis quantitativas. Se um diagrama de dispersão mostra uma relação linear, é interessante resumirmos esse padrão geral traçando uma reta no diagrama de dispersão. Uma reta de regressão resume a relação entre duas variáveis, mas somente em um contexto específico: quando uma das variáveis ajuda a explicarmos ou predizermos a outra, ou seja, a regressão descreve uma relação entre uma variável explanatória e uma variável resposta. Vale destacar que em nossas análises não foram classificadas as variáveis como sendo de caráter explanatória (variável independente) ou de resposta (variável dependente)

.

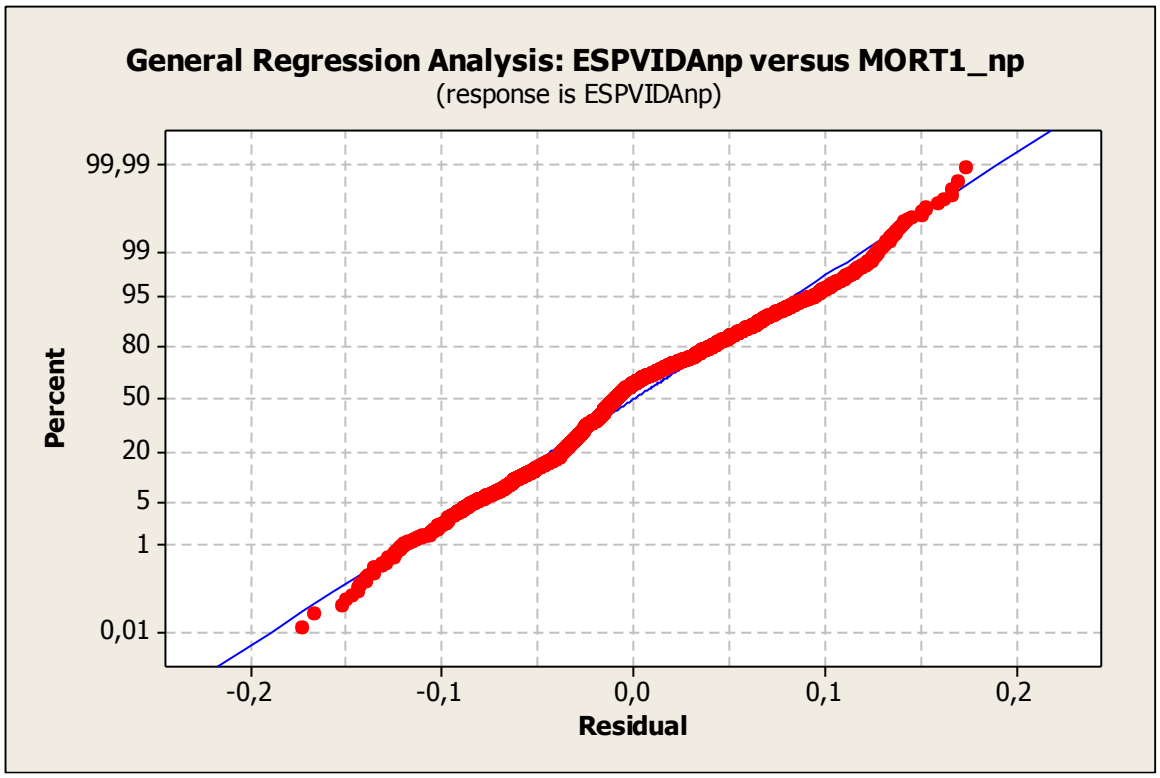
A regressão linear assume sempre a forma de uma equação linear:

Y = a + bx, sendo:

Y= Variável dependente;
a = uma constante, o intercepto;
b = a inclinação na reta;
x = variável independente ou explicativa.

O “b”, ou seja, a declividade é dada pela multiplicação do índice de correlação pela divisão dos desvios-padrão entre as variáveis x e y. E “a” é dada pela média de “Y” menos a multiplicação de “b” pela média de “x”. Assim, percebe-se muito claramente que a regressão depende da correlação entre as variáveis, além de medidas de centro de cada uma das variáveis.

Serão apresentadas as análises de Regressão bem como seus respectivos gráficos:



General Regression Analysis: ESPVIDAnp versus MORT1_np

Regression Equation

ESPVIDAnp = -0,16636 + 1,0431 MORT1_np

Coefficients

Term	Coef	SE Coef	T	P
Constant	-0,16636	0,0027314	-60,905	0,000
MORT1_np	1,04310	0,0036766	283,711	0,000

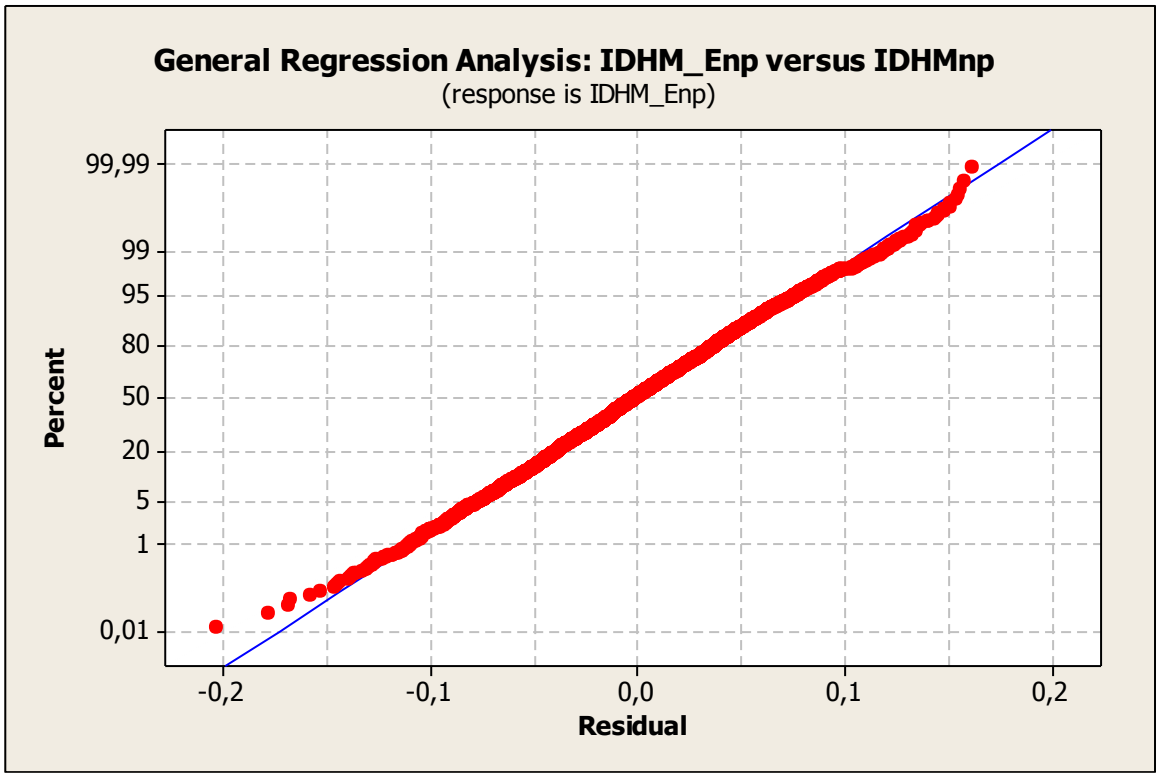
Summary of Model

S = 0,0510863 R-Sq = 93,54% R-Sq(adj) = 93,54%
PRESS = 14,5286 R-Sq(pred) = 93,53%

Analysis of Variance

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	1	210,069	210,069	210,069	80492,0	0
MORT1_np	1	210,069	210,069	210,069	80492,0	0
Error	5562	14,516	14,516	0,003		
Lack-of-Fit	556	6,258	6,258	0,011	6,8	0
Pure Error	5006	8,258	8,258	0,002		
Total	5563	224,584				

Fits and Diagnostics for Unusual Observations



General Regression Analysis: IDHM_Enp versus IDHMnp

Regression Equation

IDHM_Enp = 0,0887083 + 0,885633 IDHMnp

Coefficients

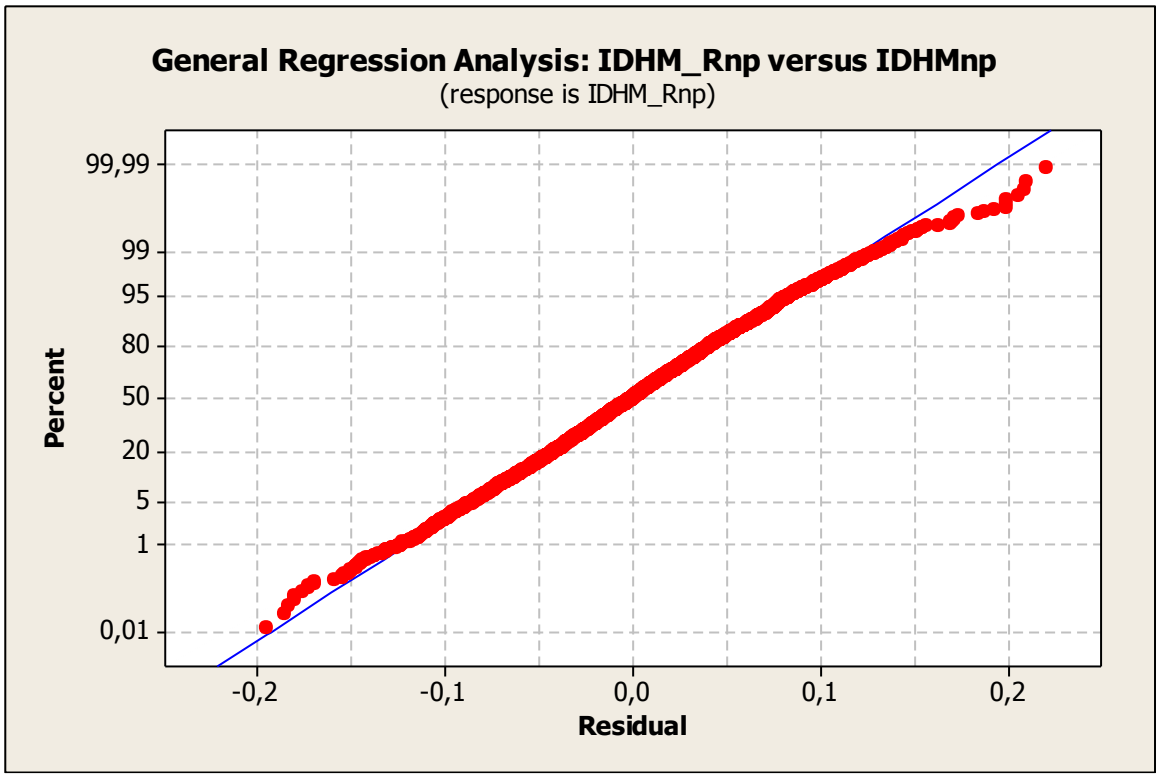
Term	Coef	SE Coef	T	P
Constant	0,088708	0,0021919	40,471	0,000
IDHMnp	0,885633	0,0038674	228,997	0,000

Summary of Model

S = 0,0467569 R-Sq = 90,41% R-Sq(adj) = 90,41%
PRESS = 12,1678 R-Sq(pred) = 90,40%

Analysis of Variance

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	1	114,644	114,644	114,644	52439,6	0
IDHMnp	1	114,644	114,644	114,644	52439,6	0
Error	5562	12,160	12,160	0,002		
Lack-of-Fit	346	1,576	1,576	0,005	2,2	0
Pure Error	5216	10,583	10,583	0,002		
Total	5563	126,804				



General Regression Analysis: IDHM_Rnp versus IDHMnp

Regression Equation

IDHM_Rnp = -0,0270661 + 0,960512 IDHMnp

Coefficients

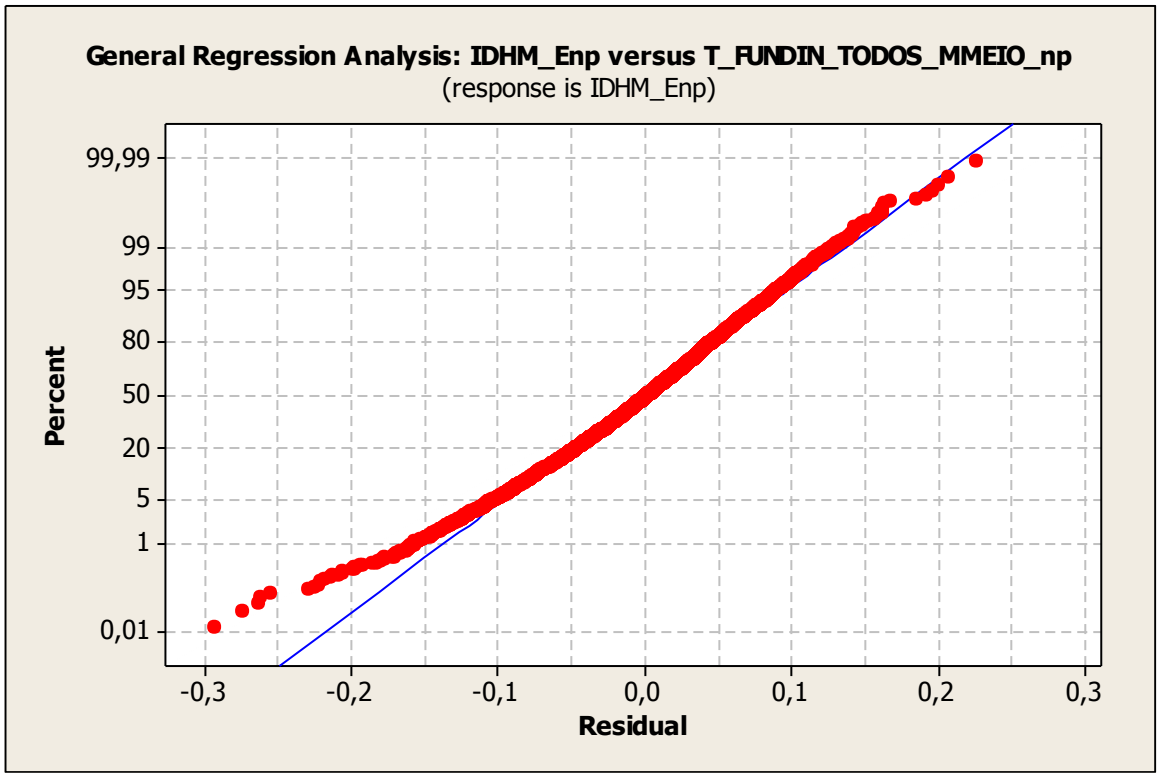
Term	Coef	SE Coef	T	P
Constant	-0,027066	0,0024434	-11,077	0,000
IDHMnp	0,960512	0,0043113	222,790	0,000

Summary of Model

S = 0,0521230 R-Sq = 89,92% R-Sq(adj) = 89,92%
PRESS = 15,1219 R-Sq(pred) = 89,92%

Analysis of Variance

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	1	134,850	134,850	134,850	49635,4	0
IDHMnp	1	134,850	134,850	134,850	49635,4	0
Error	5562	15,111	15,111	0,003		
Lack-of-Fit	346	1,806	1,806	0,005	2,0	0
Pure Error	5216	13,304	13,304	0,003		
Total	5563	149,960				



General Regression Analysis: IDHM_Enp versus T_FUNDIN_TODOS_MMEIO_np

Regression Equation

IDHM_Enp = -0,0178157 + 0,811644 T_FUNDIN_TODOS_MMEIO_np

Coefficients

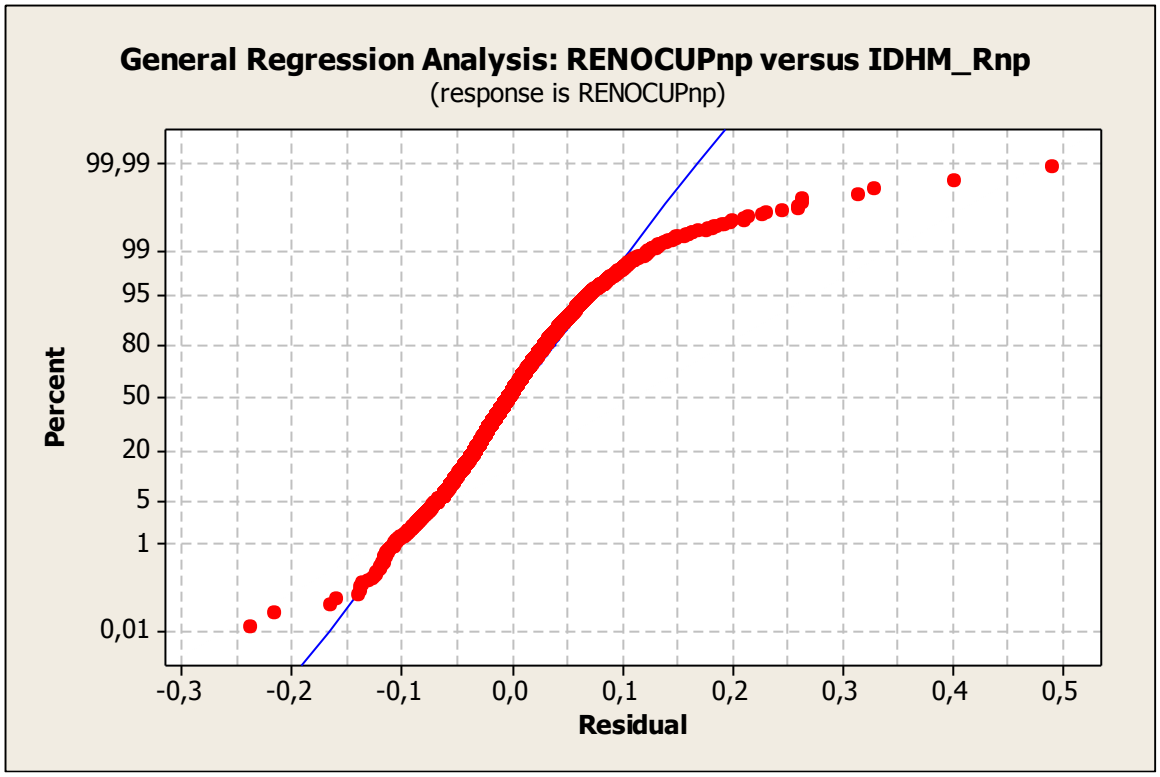
Term	Coef	SE Coef	T	P
Constant	-0,017816	0,0034158	-5,216	0,000
T_FUNDIN_TODOS_MMEIO_np	0,811644	0,0045922	176,746	0,000

Summary of Model

S = 0,0586997 R-Sq = 84,89% R-Sq(adj) = 84,88%
PRESS = 19,1784 R-Sq(pred) = 84,88%

Analysis of Variance

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	1	107,639	107,639	107,639	31239,0	0,000000
T_FUNDIN_TODOS_MMEIO_np	1	107,639	107,639	107,639	31239,0	0,000000
Error	5562	19,165	19,165	0,003		
Lack-of-Fit	3139	10,953	10,953	0,003	1,0	0,223519
Pure Error	2423	8,212	8,212	0,003		
Total	5563	126,804				



General Regression Analysis: RENOCUPnp versus IDHM_Rnp

Regression Equation

RENOCUPnp = -0,0975161 + 0,624973 IDHM_Rnp

Coefficients

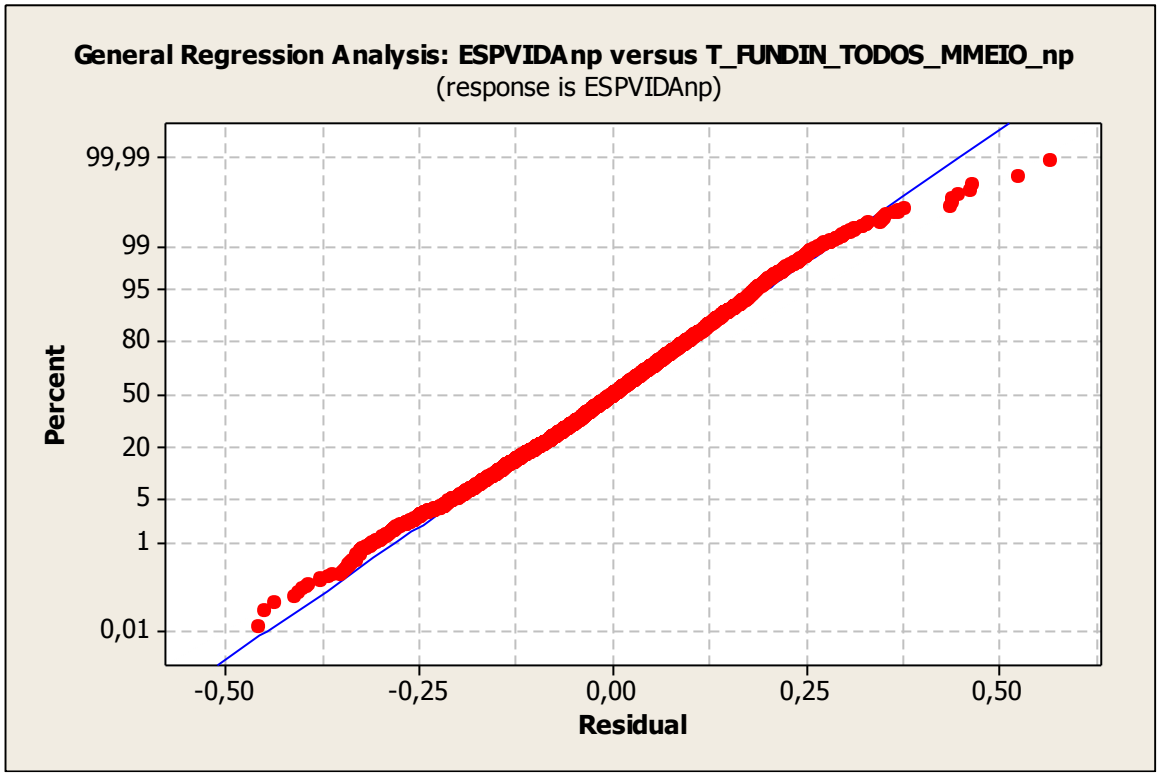
Term	Coef	SE Coef	T	P
Constant	-0,097516	0,0019203	-50,782	0,000
IDHM_Rnp	0,624973	0,0036850	169,599	0,000

Summary of Model

S = 0,0451259 R-Sq = 83,80% R-Sq(adj) = 83,79%
PRESS = 11,3388 R-Sq(pred) = 83,78%

Analysis of Variance

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	1	58,5732	58,5732	58,5732	28763,9	0
IDHM_Rnp	1	58,5732	58,5732	58,5732	28763,9	0
Error	5562	11,3262	11,3262	0,0020		
Lack-of-Fit	387	3,2700	3,2700	0,0084	5,4	0
Pure Error	5175	8,0562	8,0562	0,0016		
Total	5563	69,8994				



General Regression Analysis: ESPVIDAnp versus T_FUNDIN_TODOS_MMEIO_np

Regression Equation

ESPVIDAnp = -0,0963757 + 0,939722 T_FUNDIN_TODOS_MMEIO_np

Coefficients

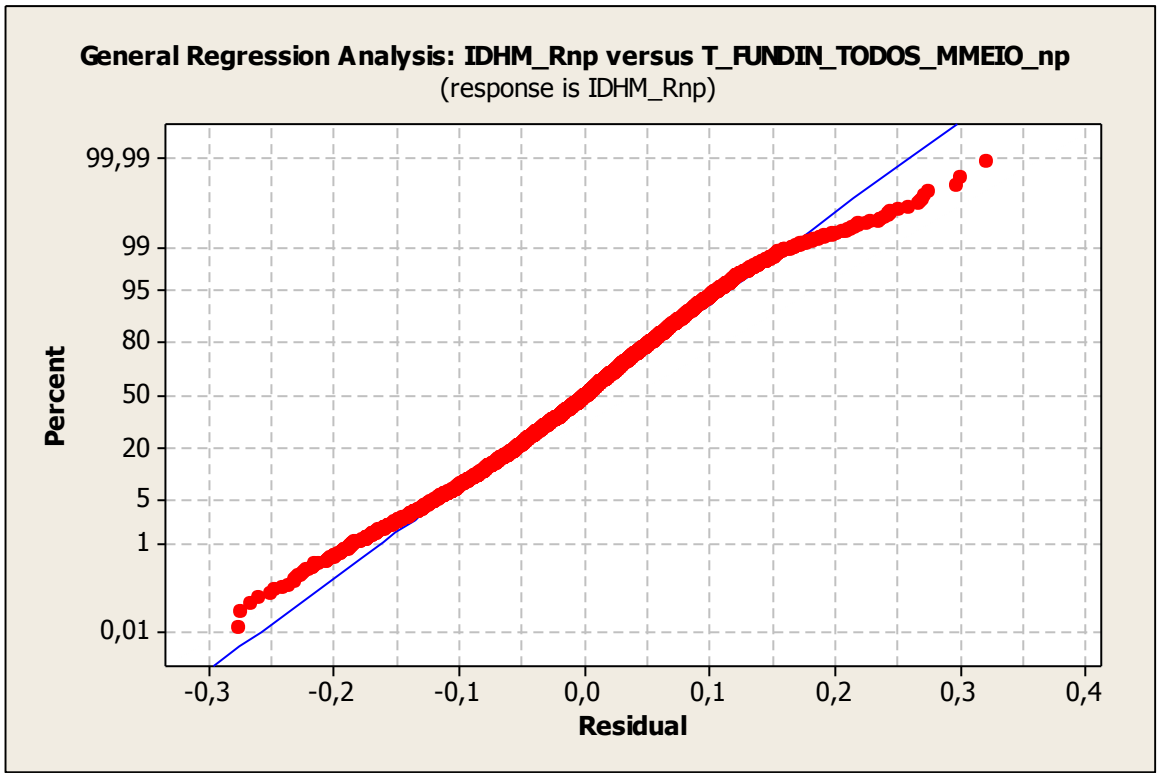
Term	Coef	SE Coef	T	P
Constant	-0,096376	0,0069918	-13,7842	0,000
T_FUNDIN_TODOS_MMEIO_np	0,939722	0,0093995	99,9753	0,000

Summary of Model

S = 0,120151 R-Sq = 64,25% R-Sq(adj) = 64,24%
PRESS = 80,3632 R-Sq(pred) = 64,22%

Analysis of Variance

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	1	144,290	144,290	144,290	9995,05	0,0000000
T_FUNDIN_TODOS_MMEIO_np	1	144,290	144,290	144,290	9995,05	0,0000000
Error	5562	80,294	80,294	0,014		
Lack-of-Fit	3139	50,149	50,149	0,016	1,28	0,0000000
Pure Error	2423	30,145	30,145	0,012		
Total	5563	224,584				



General Regression Analysis: IDHM_Rnp versus T_FUNDIN_TODOS_MMEIO_np

Regression Equation

IDHM_Rnp = -0,133544 + 0,867762 T_FUNDIN_TODOS_MMEIO_np

Coefficients

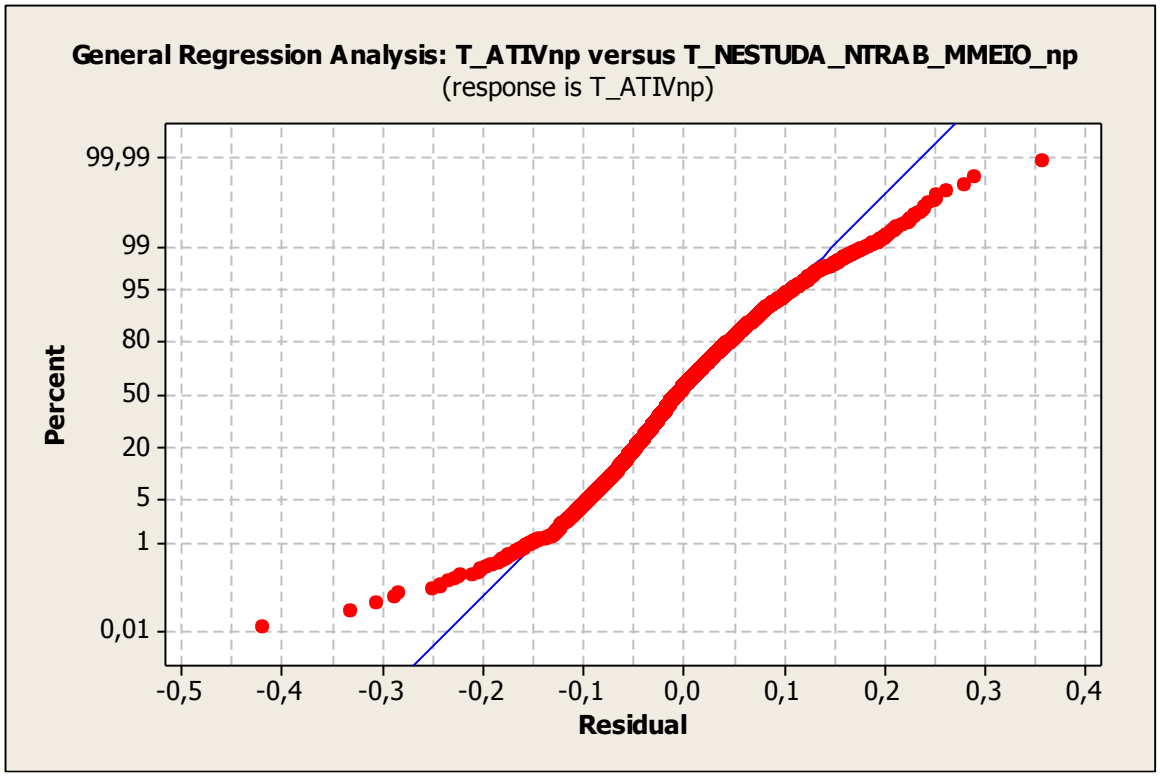
Term	Coef	SE Coef	T	P
Constant	-0,133544	0,0040486	-32,985	0,000
T_FUNDIN_TODOS_MMEIO_np	0,867762	0,0054428	159,433	0,000

Summary of Model

S = 0,0695733 R-Sq = 82,05% R-Sq(adj) = 82,04%
PRESS = 26,9438 R-Sq(pred) = 82,03%

Analysis of Variance

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	1	123,038	123,038	123,038	25418,8	0,0000000
T_FUNDIN_TODOS_MMEIO_np	1	123,038	123,038	123,038	25418,8	0,0000000
Error	5562	26,923	26,923	0,005		
Lack-of-Fit	3139	16,799	16,799	0,005	1,3	0,0000000
Pure Error	2423	10,123	10,123	0,004		
Total	5563	149,960				



General Regression Analysis: T_ATIVnp versus T_NESTUDA_NTRAB_MMEIO_np

Regression Equation

T_ATIVnp = -0,0549291 + 0,731063 T_NESTUDA_NTRAB_MMEIO_np

Coefficients

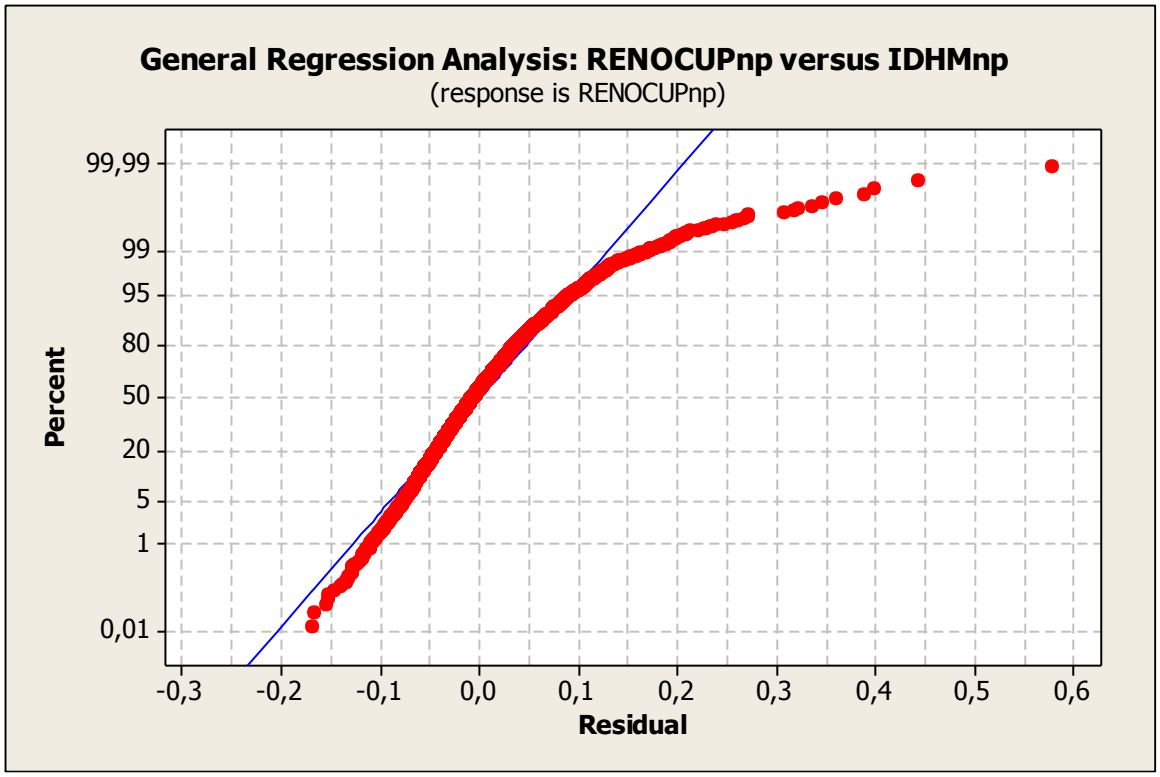
Term	Coef	SE Coef	T	P
Constant	-0,054929	0,0040182	-13,670	0,000
T_NESTUDA_NTRAB_MMEIO_np	0,731063	0,0053615	136,355	0,000

Summary of Model

S = 0,0633399 R-Sq = 76,97% R-Sq(adj) = 76,97%
PRESS = 22,3327 R-Sq(pred) = 76,95%

Analysis of Variance

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	1	74,5929	74,5929	74,5929	18592,7	0,0000000
T_NESTUDA_NTRAB_MMEIO_np	1	74,5929	74,5929	74,5929	18592,7	0,0000000
Error	5562	22,3144	22,3144	0,0040		
Lack-of-Fit	2614	11,5268	11,5268	0,0044	1,2	0,0000004
Pure Error	2948	10,7876	10,7876	0,0037		
Total	5563	96,9073				



General Regression Analysis: RENOCUPnp versus IDHMnp

Regression Equation

RENOCUPnp = -0,114892 + 0,601142 IDHMnp

Coefficients

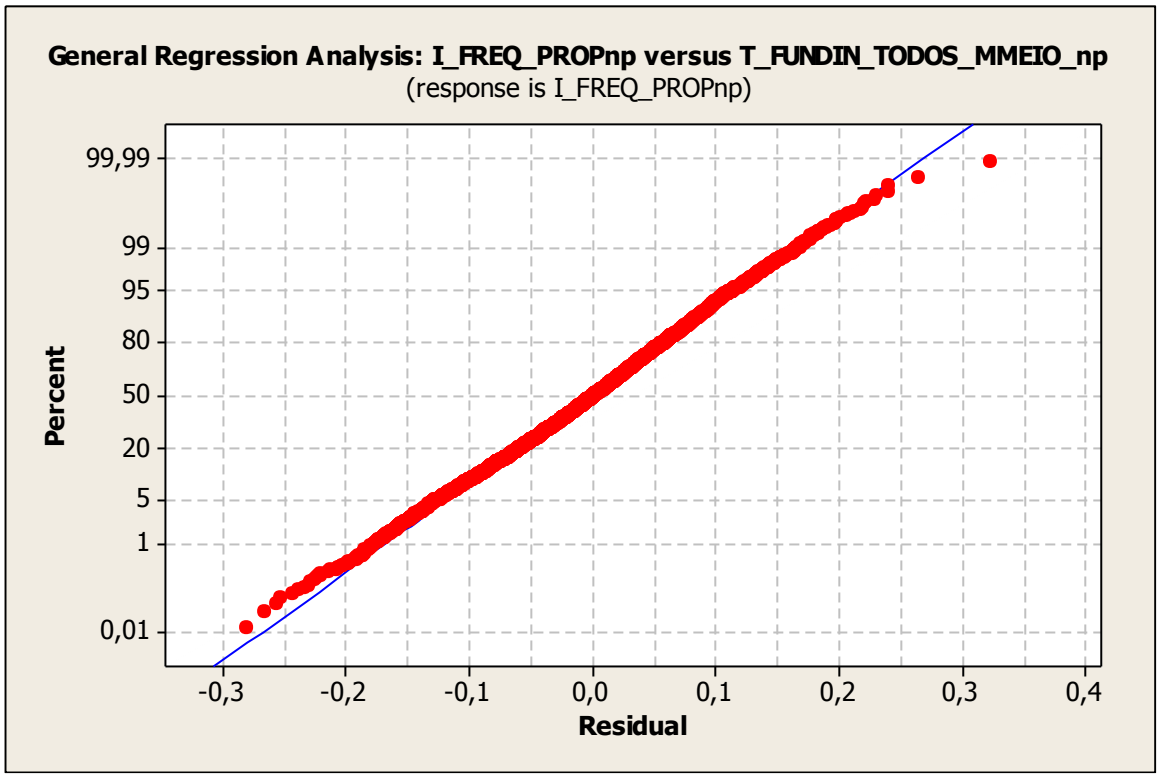
Term	Coef	SE Coef	T	P
Constant	-0,114892	0,0025977	-44,228	0,000
IDHMnp	0,601142	0,0045835	131,153	0,000

Summary of Model

S = 0,0554141 R-Sq = 75,57% R-Sq(adj) = 75,56%
PRESS = 17,0958 R-Sq(pred) = 75,54%

Analysis of Variance

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	1	52,8200	52,8200	52,8200	17201,2	0
IDHMnp	1	52,8200	52,8200	52,8200	17201,2	0
Error	5562	17,0793	17,0793	0,0031		
Lack-of-Fit	346	4,2564	4,2564	0,0123	5,0	0
Pure Error	5216	12,8229	12,8229	0,0025		
Total	5563	69,8994				



General Regression Analysis: I_FREQ_PROPnp versus T_FUNDIN_TODOS_MMEIO_np

Regression Equation

I_FREQ_PROPnp = 0,0776642 + 0,689634 T_FUNDIN_TODOS_MMEIO_np

Coefficients

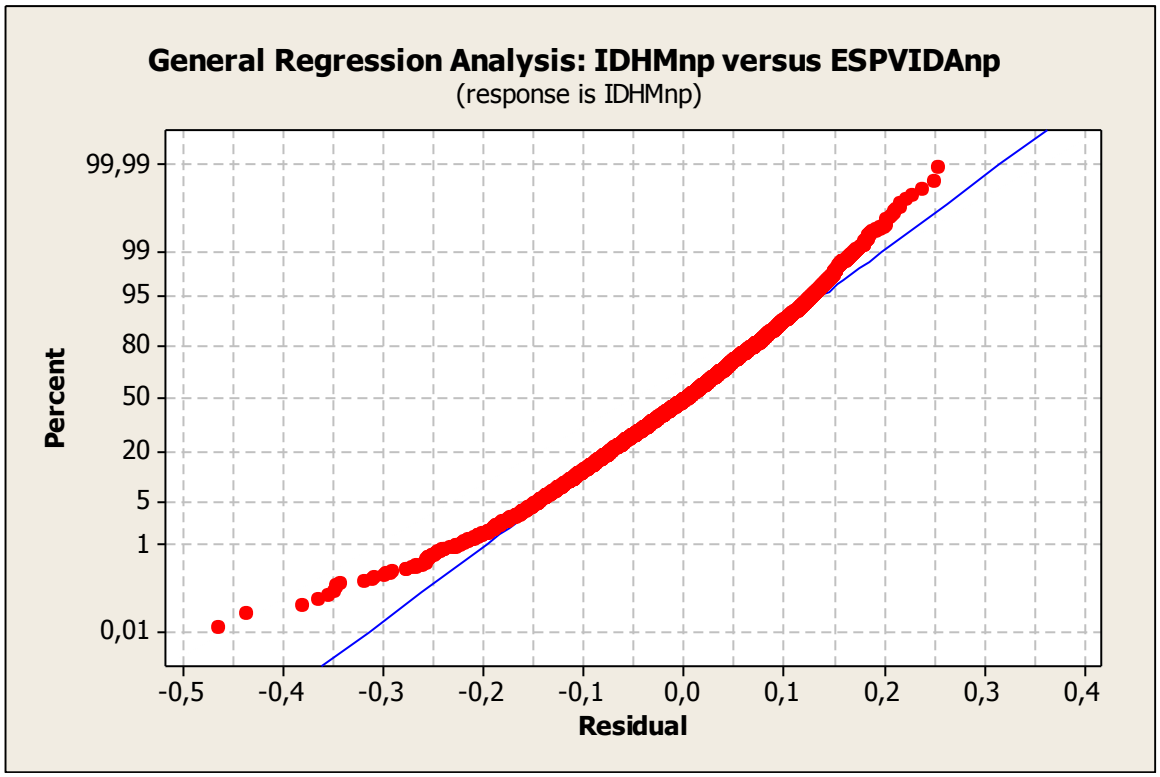
Term	Coef	SE Coef	T	P
Constant	0,077664	0,0042034	18,477	0,000
T_FUNDIN_TODOS_MMEIO_np	0,689634	0,0056509	122,040	0,000

Summary of Model

S = 0,0722329 R-Sq = 72,81% R-Sq(adj) = 72,80%
PRESS = 29,0411 R-Sq(pred) = 72,79%

Analysis of Variance

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	1	77,710	77,7096	77,7096	14893,8	0,000000
T_FUNDIN_TODOS_MMEIO_np	1	77,710	77,7096	77,7096	14893,8	0,000000
Error	5562	29,020	29,0202	0,0052		
Lack-of-Fit	3139	16,278	16,2778	0,0052	1,0	0,643790
Pure Error	2423	12,742	12,7424	0,0053		
Total	5563	106,730				



General Regression Analysis: IDHMnp versus ESPVIDAnp

Regression Equation

IDHMnp = 0,141776 + 0,687371 ESPVIDAnp

Coefficients

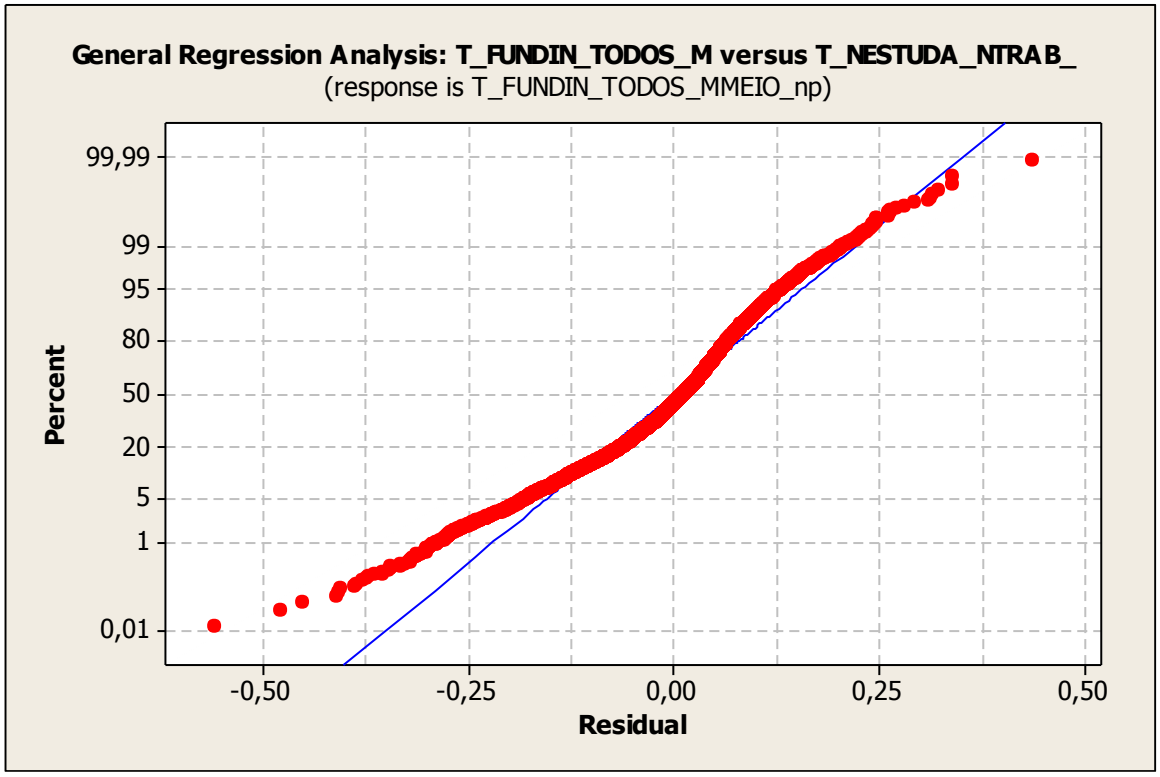
Term	Coef	SE Coef	T	P
Constant	0,141776	0,0034962	40,551	0,000
ESPVIDAnp	0,687371	0,0056626	121,388	0,000

Summary of Model

S = 0,0848606 R-Sq = 72,60% R-Sq(adj) = 72,59%
PRESS = 40,0819 R-Sq(pred) = 72,58%

Analysis of Variance

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	1	106,111	106,111	106,111	14735,0	0
ESPVIDAnp	1	106,111	106,111	106,111	14735,0	0
Error	5562	40,054	40,054	0,007		
Lack-of-Fit	1133	12,728	12,728	0,011	1,8	0
Pure Error	4429	27,326	27,326	0,006		
Total	5563	146,165				



General Regression Analysis: T_FUNDIN_TODOS_M versus T_NESTUDA_NTRAB_

Regression Equation

T_FUNDIN_TODOS_MMEIO_np = 0,06225 + 0,903138 T_NESTUDA_NTRAB_MMEIO_np

Coefficients

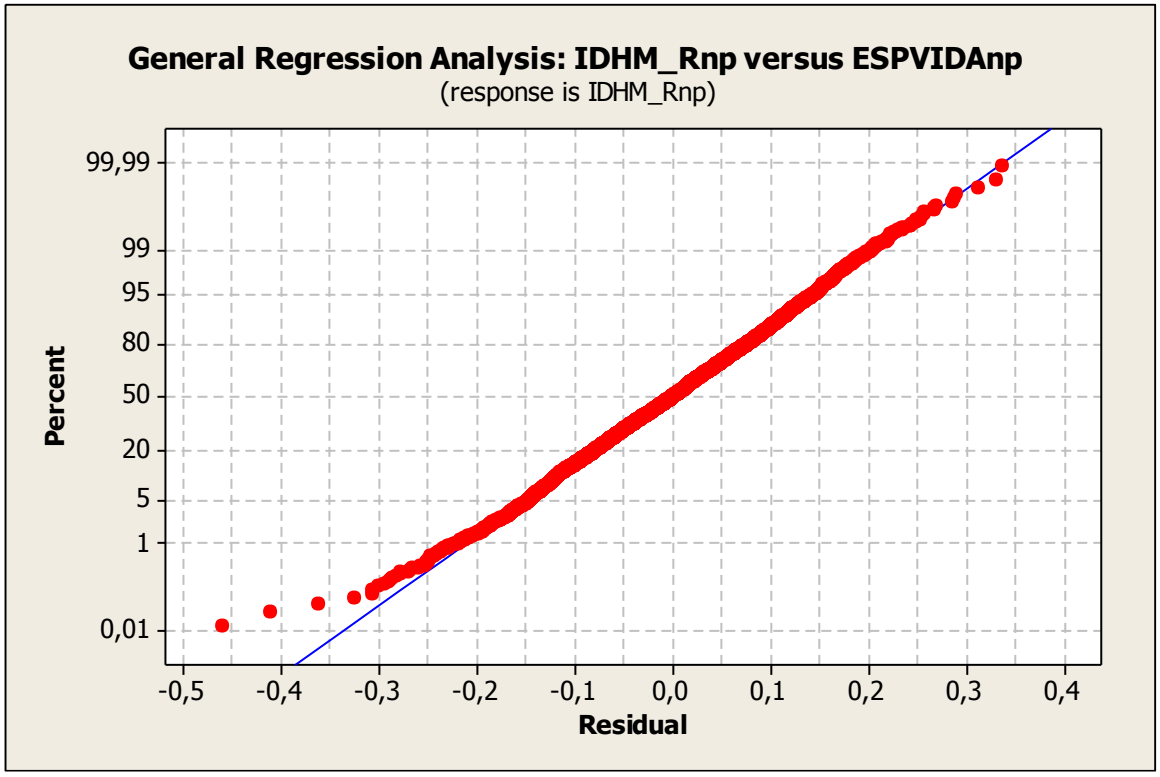
Term	Coef	SE Coef	T	P
Constant	0,062250	0,0059880	10,396	0,000
T_NESTUDA_NTRAB_MMEIO_np	0,903138	0,0079897	113,038	0,000

Summary of Model

S = 0,0943899 R-Sq = 69,67% R-Sq(adj) = 69,67%
PRESS = 49,5937 R-Sq(pred) = 69,65%

Analysis of Variance

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	1	113,840	113,840	113,840	12777,5	0
T_NESTUDA_NTRAB_MMEIO_np	1	113,840	113,840	113,840	12777,5	0
Error	5562	49,554	49,554	0,009		
Lack-of-Fit	2614	27,537	27,537	0,011	1,4	0
Pure Error	2948	22,017	22,017	0,007		
Total	5563	163,395				



General Regression Analysis: IDHM_Rnp versus ESPVIDAnp

Regression Equation

IDHM_Rnp = 0,0967887 + 0,681335 ESPVIDAnp

Coefficients

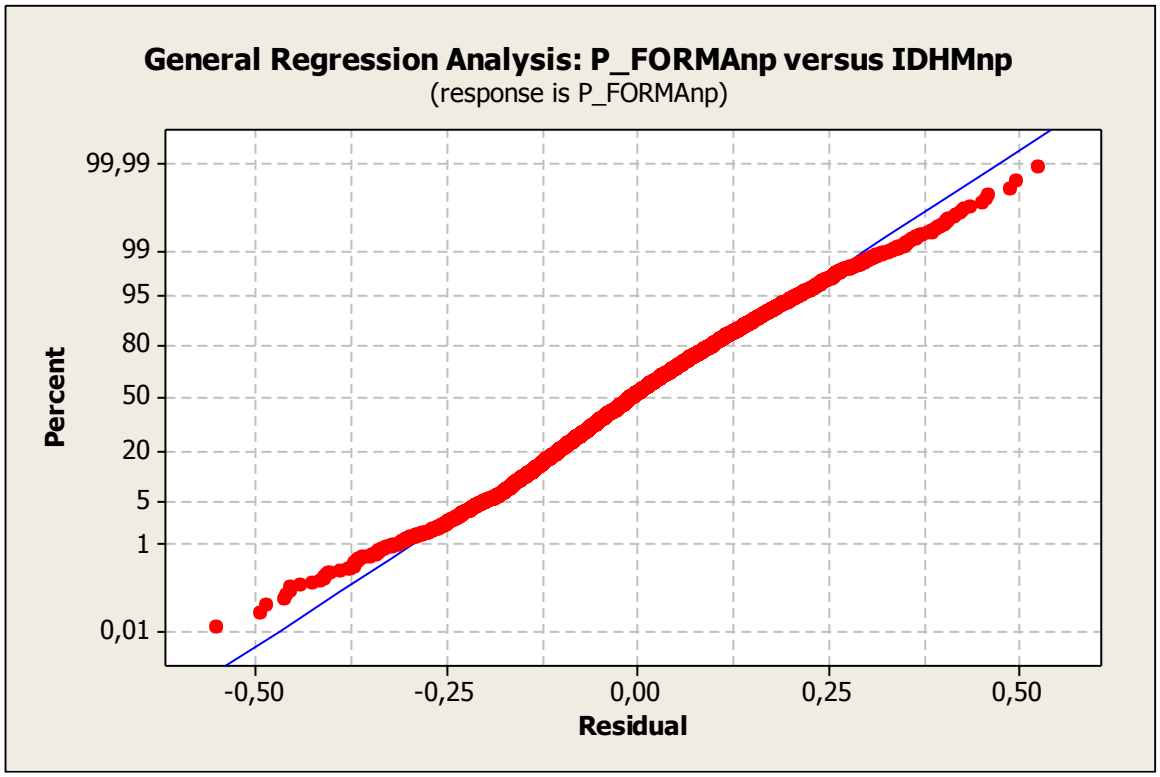
Term	Coef	SE Coef	T	P
Constant	0,096789	0,0037347	25,916	0,000
ESPVIDAnp	0,681335	0,0060489	112,638	0,000

Summary of Model

S = 0,0906492 R-Sq = 69,52% R-Sq(adj) = 69,52%
PRESS = 45,7378 R-Sq(pred) = 69,50%

Analysis of Variance

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	1	104,256	104,256	104,256	12687,4	0
ESPVIDAnp	1	104,256	104,256	104,256	12687,4	0
Error	5562	45,705	45,705	0,008		
Lack-of-Fit	1133	16,378	16,378	0,014	2,2	0
Pure Error	4429	29,327	29,327	0,007		
Total	5563	149,960				



General Regression Analysis: P_FORMAnp versus IDHMnp

Regression Equation

P_FORMAnp = -0,146955 + 1,13698 IDHMnp

Coefficients

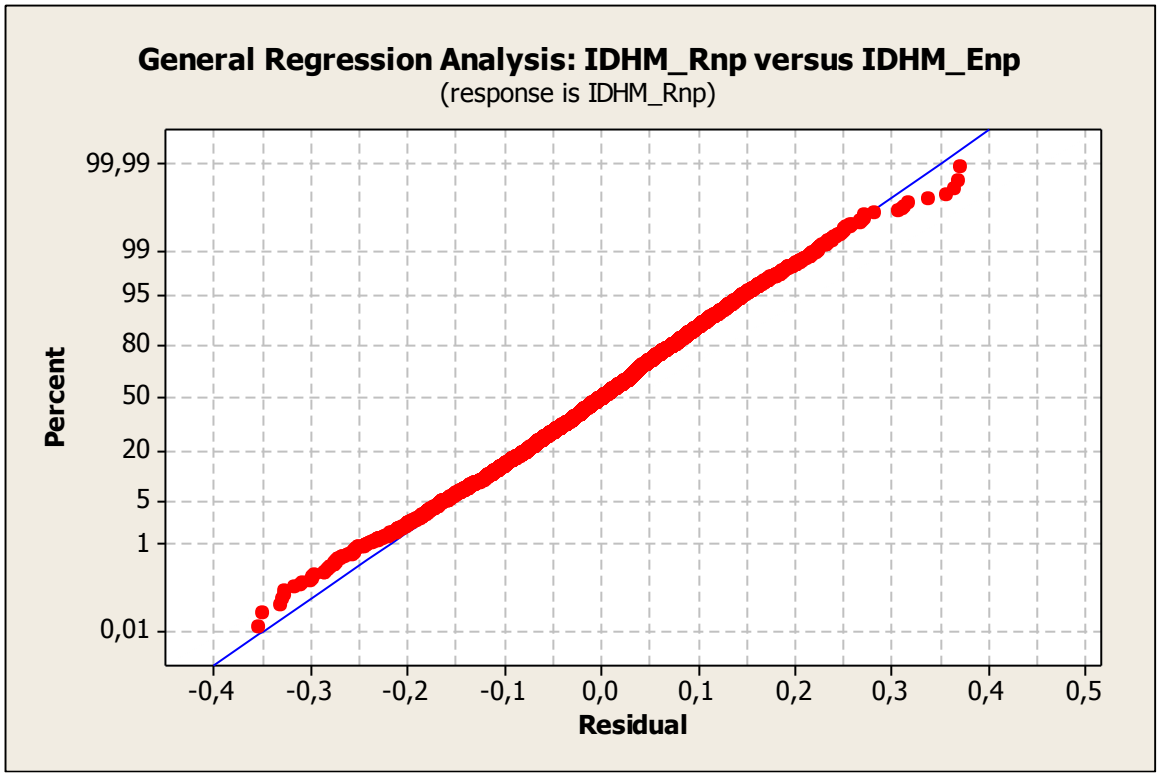
Term	Coef	SE Coef	T	P
Constant	-0,14696	0,0059476	-24,708	0,000
IDHMnp	1,13698	0,0104942	108,344	0,000

Summary of Model

S = 0,126874 R-Sq = 67,85% R-Sq(adj) = 67,84%
PRESS = 89,5923 R-Sq(pred) = 67,83%

Analysis of Variance

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	1	188,952	188,952	188,952	11738,4	0,0000000
IDHMnp	1	188,952	188,952	188,952	11738,4	0,0000000
Error	5562	89,531	89,531	0,016		
Lack-of-Fit	346	7,491	7,491	0,022	1,4	0,0000090
Pure Error	5216	82,041	82,041	0,016		
Total	5563	278,484				



General Regression Analysis: IDHM_Rnp versus IDHM_Enp

Regression Equation

IDHM_Rnp = -0,0131221 + 0,89119 IDHM_Enp

Coefficients

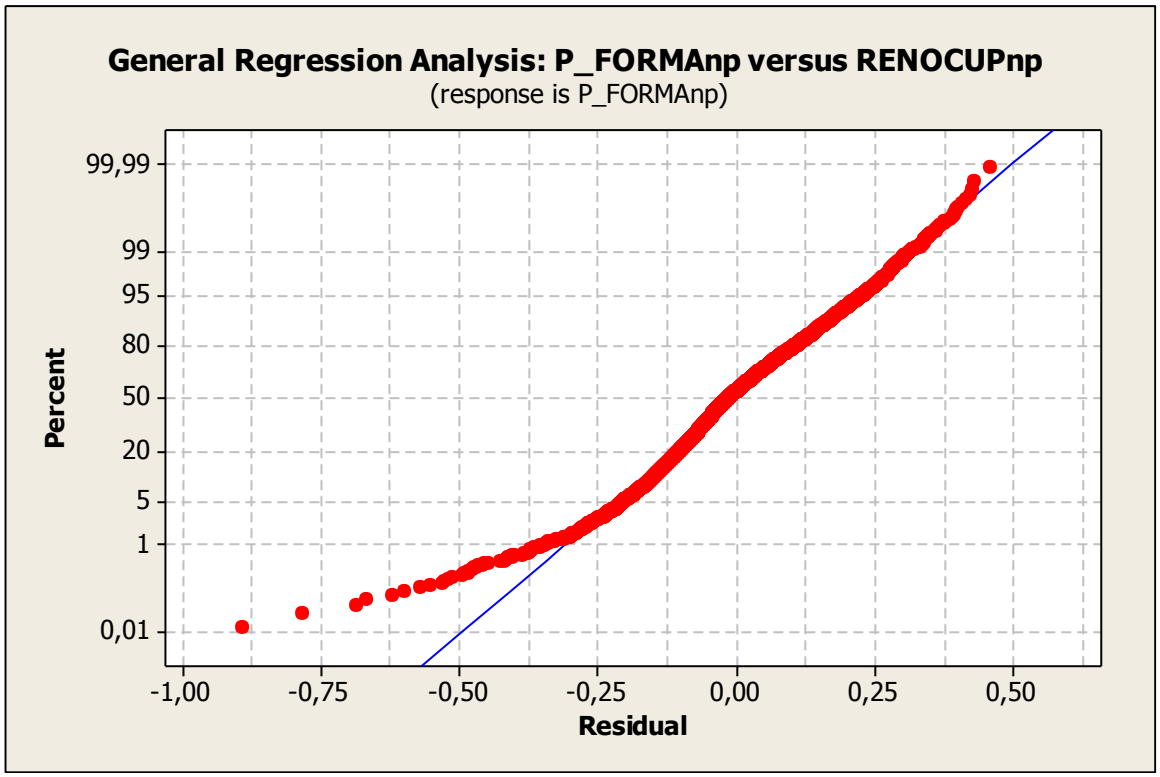
Term	Coef	SE Coef	T	P
Constant	-0,013122	0,0049248	-2,664	0,008
IDHM_Enp	0,891190	0,0083565	106,646	0,000

Summary of Model

S = 0,0941001 R-Sq = 67,16% R-Sq(adj) = 67,15%
PRESS = 49,2827 R-Sq(pred) = 67,14%

Analysis of Variance

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	1	100,710	100,710	100,710	11373,5	0,0000000
IDHM_Enp	1	100,710	100,710	100,710	11373,5	0,0000000
Error	5562	49,251	49,251	0,009		
Lack-of-Fit	464	5,034	5,034	0,011	1,3	0,0003552
Pure Error	5098	44,217	44,217	0,009		
Total	5563	149,960				



General Regression Analysis: P_FORMAnp versus RENOCUPnp

Regression Equation

P_FORMAnp = 0,13189 + 1,6005 RENOCUPnp

Coefficients

Term	Coef	SE Coef	T	P
Constant	0,13189	0,0038290	34,445	0,000
RENOCUPnp	1,60050	0,0159920	100,081	0,000

Summary of Model

S = 0,133703 R-Sq = 64,30% R-Sq(adj) = 64,29%
PRESS = 99,5495 R-Sq(pred) = 64,25%

Analysis of Variance

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	1	179,055	179,055	179,055	10016,3	0,0000000
RENOCUPnp	1	179,055	179,055	179,055	10016,3	0,0000000
Error	5562	99,428	99,428	0,018		
Lack-of-Fit	5428	97,484	97,484	0,018	1,2	0,0519080
Pure Error	134	1,945	1,945	0,015		
Total	5563	278,484				

COMENTÁRIOS DAS ANÁLISES

Considerando as análises de regressão acima pode-se perceber que existe uma relação com grau razoável de explicação entre as varáveis apresentadas. Em alguns casos esta relação se apresenta muito forte, como é o caso da ESPVIDA x IDHM_L, isto pode ser justificado pela característica destas variáveis que carregam em sua composição a esperança de vida ao nascer.

Além disso, vale a pena destacar que as relações que apresentam variáveis como educação (IDHM_E; I_FREQ_PROP...), IDHM e Renda (IDHM_R e RDPC) possuem um alto grau de relação próximos ou acima de 90 % para as análises realizadas. No caso da educação pode-se dizer que quanto mais alta a idade das pessoas menor é o grau de relação.

9. REGRESSÃO MULTIVARIADA

General Regression Analysis: IDHMnp versus MORT1_np; T_NESTUDA_NT; ...

Regression Equation

```
IDHMnp  =  -0,0552517 - 0,00820026 MORT1_np - 0,00302807
           T_NESTUDA_NTRAB_MMEIO_np + 0,00757991 T_DENS(np) + 0,139362
           ESPVIDAnp - 0,00316167 I_FREQ_PROPnp + 0,55975 IDHM_Enp + 0,409695
           IDHM_Rnp - 0,00492184 T_FLBASnp + 0,0173363 T_FLFUNDnp + 0,0120265
           PRENTRABnp - 0,0441913 RENOCUPnp - 0,00100439 P_FORMAnp - 0,0132206
           T_ATIVnp - 0,00268139 T_DES2529np
```

Coefficients

Term	Coef	SE Coef	T	P
Constant	-0,055252	0,0007900	-69,942	0,000
MORT1_np	-0,008200	0,0009818	-8,352	0,000
T_NESTUDA_NTRAB_MMEIO_np	-0,003028	0,0007943	-3,812	0,000
T_DENS(np)	0,007580	0,0005650	13,415	0,000
ESPVIDAnp	0,139362	0,0009329	149,390	0,000
I_FREQ_PROPnp	-0,003162	0,0011098	-2,849	0,004
IDHM_Enp	0,559750	0,0011510	486,321	0,000
IDHM_Rnp	0,409695	0,0015645	261,864	0,000
T_FLBASnp	-0,004922	0,0009840	-5,002	0,000
T_FLFUNDnp	0,017336	0,0011822	14,664	0,000
PRENTRABnp	0,012026	0,0004935	24,368	0,000
RENOCUPnp	-0,044191	0,0016304	-27,105	0,000
P_FORMAnp	-0,001004	0,0003863	-2,600	0,009
T_ATIVnp	-0,013221	0,0008498	-15,558	0,000
T_DES2529np	-0,002681	0,0004891	-5,482	0,000

Summary of Model

```
S = 0,00322640     R-Sq = 99,96%        R-Sq(adj) = 99,96%
PRESS = 0,0583977  R-Sq(pred) = 99,96%
```

Analysis of Variance

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	14	146,107	146,107	10,4362	1002556	0,0000000
MORT1_np	1	100,419	0,001	0,0007	70	0,0000000
T_NESTUDA_NTRAB_MMEIO_np	1	16,181	0,000	0,0002	15	0,0001391
T_DENS(np)	1	0,429	0,002	0,0019	180	0,0000000
ESPVIDAnp	1	2,496	0,232	0,2323	22317	0,0000000
I_FREQ_PROPnp	1	14,827	0,000	0,0001	8	0,0044042
IDHM_Enp	1	9,513	2,462	2,4620	236508	0,0000000
IDHM_Rnp	1	2,230	0,714	0,7138	68573	0,0000000
T_FLBASnp	1	0,001	0,000	0,0003	25	0,0000006
T_FLFUNDnp	1	0,003	0,002	0,0022	215	0,0000000
PRENTRABnp	1	0,002	0,006	0,0062	594	0,0000000
RENOCUPnp	1	0,005	0,008	0,0076	735	0,0000000
P_FORMAnp	1	0,000	0,000	0,0001	7	0,0093420
T_ATIVnp	1	0,002	0,003	0,0025	242	0,0000000
T_DES2529np	1	0,000	0,000	0,0003	30	0,0000000
Error	5549	0,058	0,058	0,0000		
Total	5563	146,165				

Stepwise Regression: IDHMnp versus MORT1_np; T_NESTUDA_NTRAB_; ...

Alpha-to-Enter: 0,15 Alpha-to-Remove: 0,15

Response is IDHMnp on 15 predictors, with N = 5564

Step	1	2	3	4	5
Constant	-0,11521	-0,05843	-0,04406	-0,04346	-0,05625
T_FUNDIN_TODOS_MMEIO_np	0,90945	0,54055	0,09116	0,02569	0,02521
T-Value	261,14	87,98	22,85	25,99	26,45
P-Value	0,000	0,000	0,000	0,000	0,000
IDHM_Rnp		0,42512	0,46454	0,37045	0,37142
T-Value		66,29	164,46	493,01	511,65
P-Value		0,000	0,000	0,000	0,000
IDHM_Enp			0,51153	0,54091	0,53920
T-Value			152,79	663,76	682,52
P-Value			0,000	0,000	0,000
ESPVIDAnp				0,13117	0,13208
T-Value				299,43	311,06
P-Value				0,000	0,000
T_FLFUNDnp					0,01517
T-Value					20,57
P-Value					0,000
T_DENS(np)					
T-Value					
P-Value					
S	0,0445	0,0333	0,0146	0,00353	0,00340
R-Sq	92,46	95,79	99,19	99,95	99,96
R-Sq(adj)	92,46	95,79	99,19	99,95	99,96
Mallows Cp	1112772,5	619132,0	114602,3	1461,0	966,5

Step	6
Constant	-0,05573
T_FUNDIN_TODOS_MMEIO_np	0,02222
T-Value	22,87
P-Value	0,000
IDHM_Rnp	0,36978
T-Value	507,39
P-Value	0,000
IDHM_Enp	0,54125
T-Value	678,74
P-Value	0,000
ESPVIDAnp	0,13175
T-Value	313,74
P-Value	0,000
T_FLFUNDnp	0,01235
T-Value	16,18
P-Value	0,000
T_DENS(np)	0,00541
T-Value	12,23
P-Value	0,000
S	0,00336
R-Sq	99,96

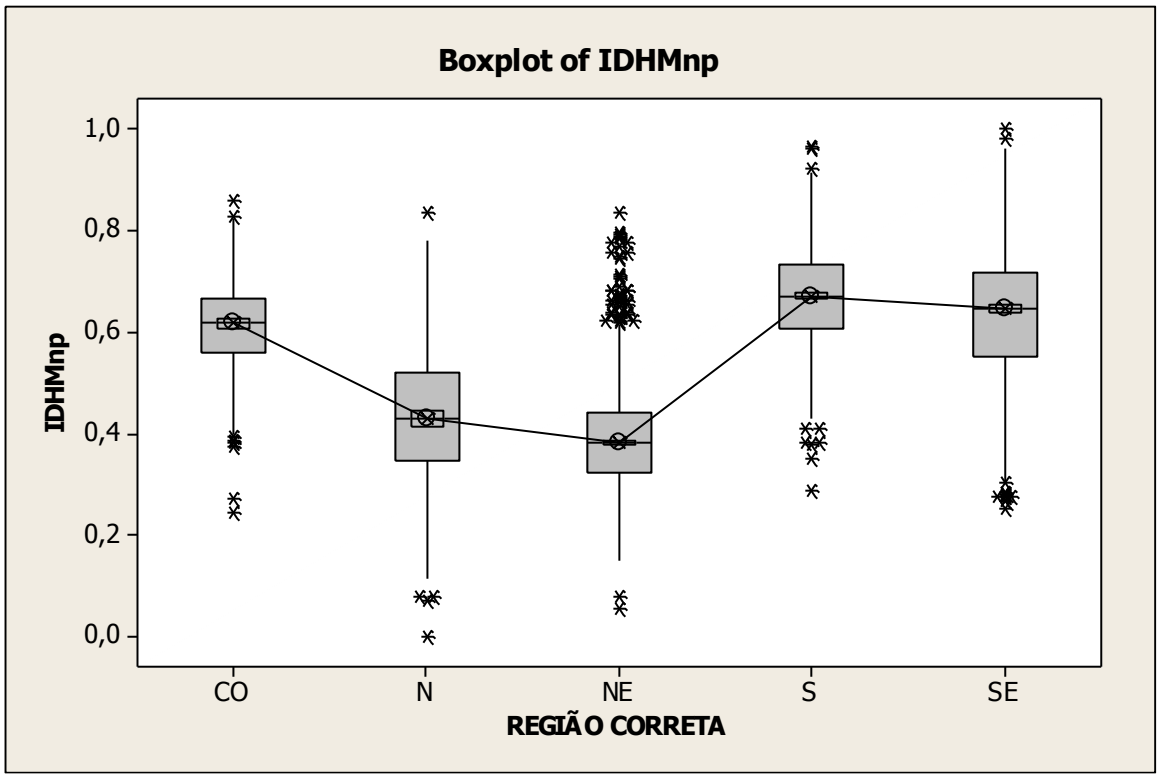
R-Sq(adj)	99,96				
Mallows Cp	797,6				
Step	7	8	9	10	11
Constant	-0,05508	-0,05798	-0,05761	-0,05878	-0,05679
T_FUNDIN_TODOS_MMEIO_np	0,02459	0,02316	0,01870	0,01725	0,01830
T-Value	24,86	23,40	18,15	16,95	18,07
P-Value	0,000	0,000	0,000	0,000	0,000
IDHM_Rnp	0,37371	0,37086	0,38701	0,39689	0,39915
T-Value	458,71	434,87	262,59	246,06	247,73
P-Value	0,000	0,000	0,000	0,000	0,000
IDHM_Enp	0,53969	0,54082	0,54467	0,54505	0,54462
T-Value	671,31	672,71	646,52	657,77	662,97
P-Value	0,000	0,000	0,000	0,000	0,000
ESPVIDAnp	0,13187	0,13183	0,13176	0,13125	0,13982
T-Value	316,92	319,80	324,64	327,55	155,21
P-Value	0,000	0,000	0,000	0,000	0,000
T_FLFUNDnp	0,01328	0,01474	0,01426	0,01439	0,01358
T-Value	17,44	19,21	18,85	19,35	18,34
P-Value	0,000	0,000	0,000	0,000	0,000
T_DENS(np)	0,00607	0,00797	0,00544	0,00576	0,00600
T-Value	13,72	16,76	10,78	11,60	12,18
P-Value	0,000	0,000	0,000	0,000	0,000
T_NESTUDA_NTRAB_MMEIO_np	-0,00650	-0,00737	-0,01065	-0,00579	-0,00631
T-Value	-10,39	-11,77	-16,05	-7,82	-8,60
P-Value	0,000	0,000	0,000	0,000	0,000
PRENTRABnp		0,00448	0,00692	0,00890	0,00999
T-Value		10,31	14,87	18,57	20,58
P-Value		0,000	0,000	0,000	0,000
RENOCUPnp			-0,0194	-0,0302	-0,0340
T-Value			-13,33	-18,56	-20,61
P-Value			0,000	0,000	0,000
T_ATIVnp				-0,01127	-0,01180
T-Value				-13,94	-14,71
P-Value				0,000	0,000
MORT1_np					-0,01014
T-Value					-10,60
P-Value					0,000
T_DES2529np					
T-Value					
P-Value					
S	0,00332	0,00329	0,00324	0,00319	0,00315
R-Sq	99,96	99,96	99,96	99,96	99,96
R-Sq(adj)	99,96	99,96	99,96	99,96	99,96
Mallows Cp	678,6	563,7	376,4	178,3	66,9
Step	12				
Constant	-0,05548				
T_FUNDIN_TODOS_MMEIO_np	0,01798				
T-Value	17,74				
P-Value	0,000				
IDHM_Rnp	0,39945				
T-Value	248,16				

P-Value	0,000		
IDHM_Enp	0,54444		
T-Value	663,22		
P-Value	0,000		
ESPVIDAnp	0,13996		
T-Value	155,56		
P-Value	0,000		
T_FLFUNDnp	0,01335		
T-Value	18,04		
P-Value	0,000		
T_DENS(np)	0,00613		
T-Value	12,45		
P-Value	0,000		
T_NESTUDA_NTRAB_MMEIO_np	-0,00537		
T-Value	-7,07		
P-Value	0,000		
PRENTRABnp	0,01007		
T-Value	20,78		
P-Value	0,000		
RENOCUPnp	-0,0347		
T-Value	-20,99		
P-Value	0,000		
T_ATIVnp	-0,01163		
T-Value	-14,51		
P-Value	0,000		
MORT1_np	-0,01018		
T-Value	-10,65		
P-Value	0,000		
T_DES2529np	-0,00213		
T-Value	-4,63		
P-Value	0,000		
S	0,00315		
R-Sq	99,96		
R-Sq(adj)	99,96		
Mallows Cp	47,4		
Step	13	14	15
Constant	-0,05557	-0,05566	-0,05597
T_FUNDIN_TODOS_MMEIO_np	0,0187	0,0183	0,0182
T-Value	18,29	17,80	17,68
P-Value	0,000	0,000	0,000
IDHM_Rnp	0,4003	0,4002	0,4001
T-Value	247,55	247,67	247,61
P-Value	0,000	0,000	0,000
IDHM_Enp	0,54457	0,54534	0,54741
T-Value	664,22	640,99	414,75
P-Value	0,000	0,000	0,000
ESPVIDAnp	0,14049	0,14019	0,14012
T-Value	155,22	154,33	154,19
P-Value	0,000	0,000	0,000
T_FLFUNDnp	0,01334	0,01633	0,01627
T-Value	18,05	14,18	14,12

P-Value	0,000	0,000	0,000
T_DENS(np)	0,00612	0,00606	0,00658
T-Value	12,45	12,33	11,90
P-Value	0,000	0,000	0,000
T_NESTUDA_NTRAB_MMEIO_np	-0,00475	-0,00440	-0,00445
T-Value	-6,17	-5,66	-5,72
P-Value	0,000	0,000	0,000
PRENTRABnp	0,01040	0,01027	0,01028
T-Value	21,27	20,94	20,97
P-Value	0,000	0,000	0,000
RENOCUPnp	-0,0344	-0,0345	-0,0350
T-Value	-20,87	-20,93	-21,00
P-Value	0,000	0,000	0,000
T_ATIVnp	-0,01261	-0,01265	-0,01255
T-Value	-15,24	-15,31	-15,17
P-Value	0,000	0,000	0,000
MORT1_np	-0,01065	-0,01037	-0,01028
T-Value	-11,10	-10,79	-10,68
P-Value	0,000	0,000	0,000
T_DES2529np	-0,00252	-0,00260	-0,00240
T-Value	-5,39	-5,56	-5,05
P-Value	0,000	0,000	0,000
P_FORMAnp	-0,00175	-0,00187	-0,00195
T-Value	-4,66	-4,95	-5,14
P-Value	0,000	0,000	0,000
T_FLBASnp		-0,00324	-0,00305
T-Value		-3,38	-3,17
P-Value		0,001	0,002
I_FREQ_PROPnp			-0,0022
T-Value			-2,04
P-Value			0,041
S	0,00314	0,00314	0,00314
R-Sq	99,96	99,96	99,96
R-Sq(adj)	99,96	99,96	99,96
Mallows Cp	27,6	18,2	16,0

10. ANÁLISE ANOVA VARIÁVEIS REGIÃO

10.1 – VARIÁVEL IDHM POR REGIÃO



A Região Sul possui o maior IDHM do país, o que indica que esta é a Região mais desenvolvida do Brasil, segundo a pesquisa. A região Sudeste encontra-se próxima a Região Sul, e ocupa o segundo lugar. Vale também destacar que a região Centro Oeste está muito próxima.

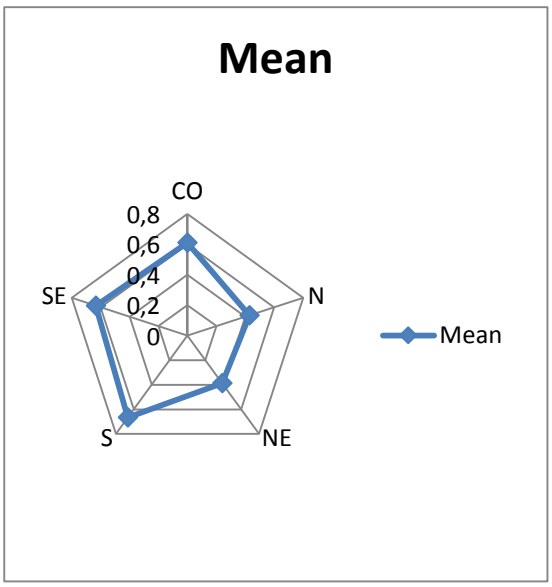
A Região que apresenta o IDHM médio mais baixo do País é a Nordeste, seguida da Norte. Pelo tamanho da caixa do BloxPlot podemos visualizar a amplitude da variância. Podemos afirmar que os dados da Região Norte possuem maior variabilidade que os dados das demais regiões. As Regiões que possuem menor variabilidade dos dados são Centro-Oeste, Norte e Sul.

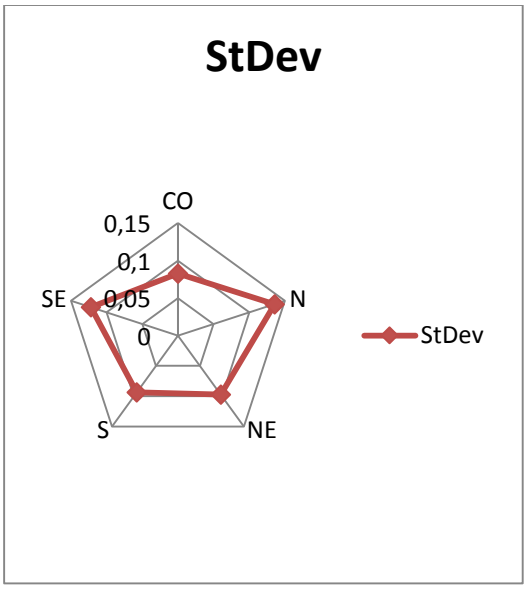
One-way ANOVA: IDHMnp versus REGIÃO CORRETA

```
Source            DF        SS       MS        F      P
REGIÃO CORRETA     4   82,3938  20,5985  1795,58  0,000
Error           5559   63,7714   0,0115
Total           5563  146,1652

S = 0,1071   R-Sq = 56,37%   R-Sq(adj) = 56,34%


                             Individual 95% CIs For Mean Based on
                             Pooled StDev
Level     N    Mean   StDev  --+---------+---------+---------+-------
CO      465  0,6108  0,0829                              (*-)
N       449  0,4278  0,1355      (*-)
NE     1794  0,3889  0,0975  (*
```

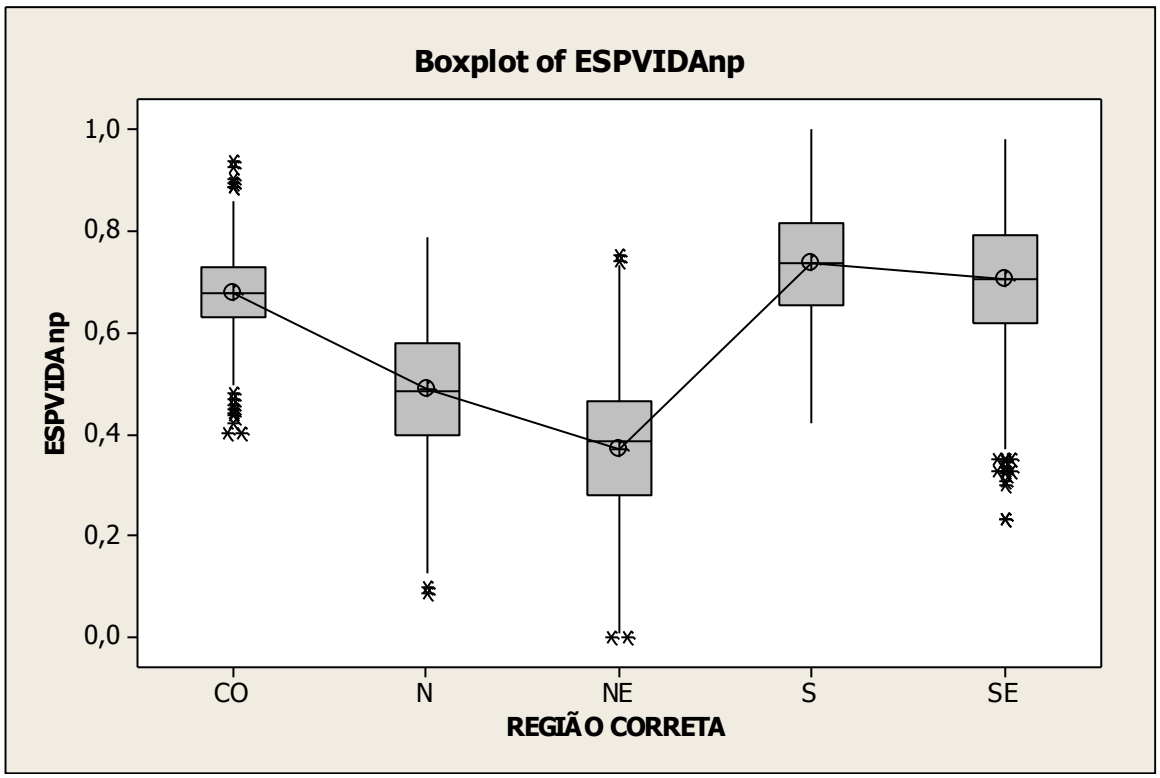
```
S       1188  0,6669  0,0937                                        *)
SE      1668  0,6328  0,1223                                  (*)
                                --+---------+---------+---------+-------
                                0,400     0,480     0,560     0,640

Pooled StDev = 0,1071
```

O grau de variação entre as Regiões é muito alto (1795,58), e o P-value nos indica que a informação é confiável e não existe chance deste valor ser diferente.





10.2 – VARIÁVEL ESPVIDA POR REGIÃO

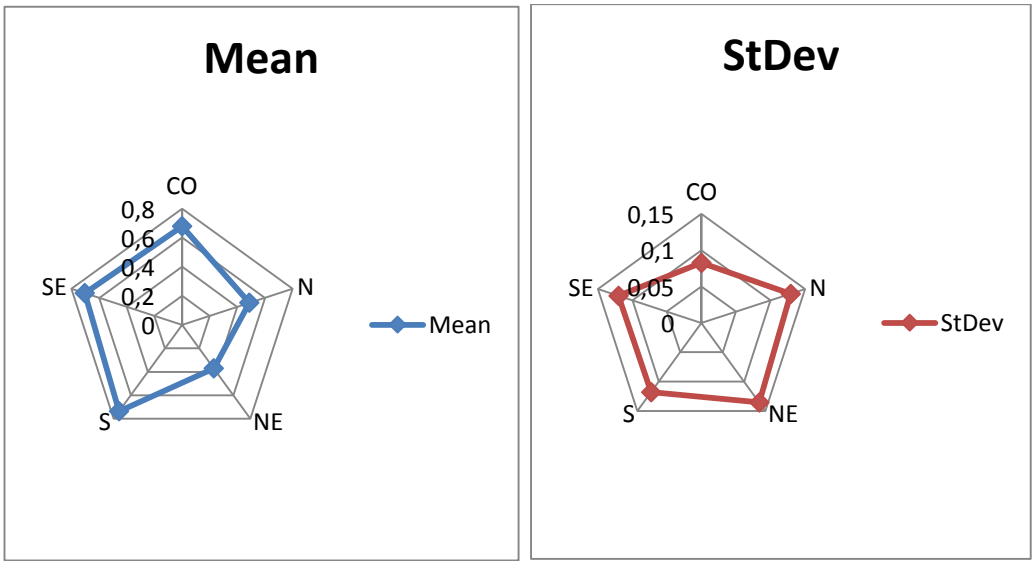


O ESPVIDA é maior para a Região Sul, que fica muito próximo da Região Sudeste, que aponta o Sul com a melhor ESPVIDA. O menor índice de ESPVIDA está para a Região Nordeste.

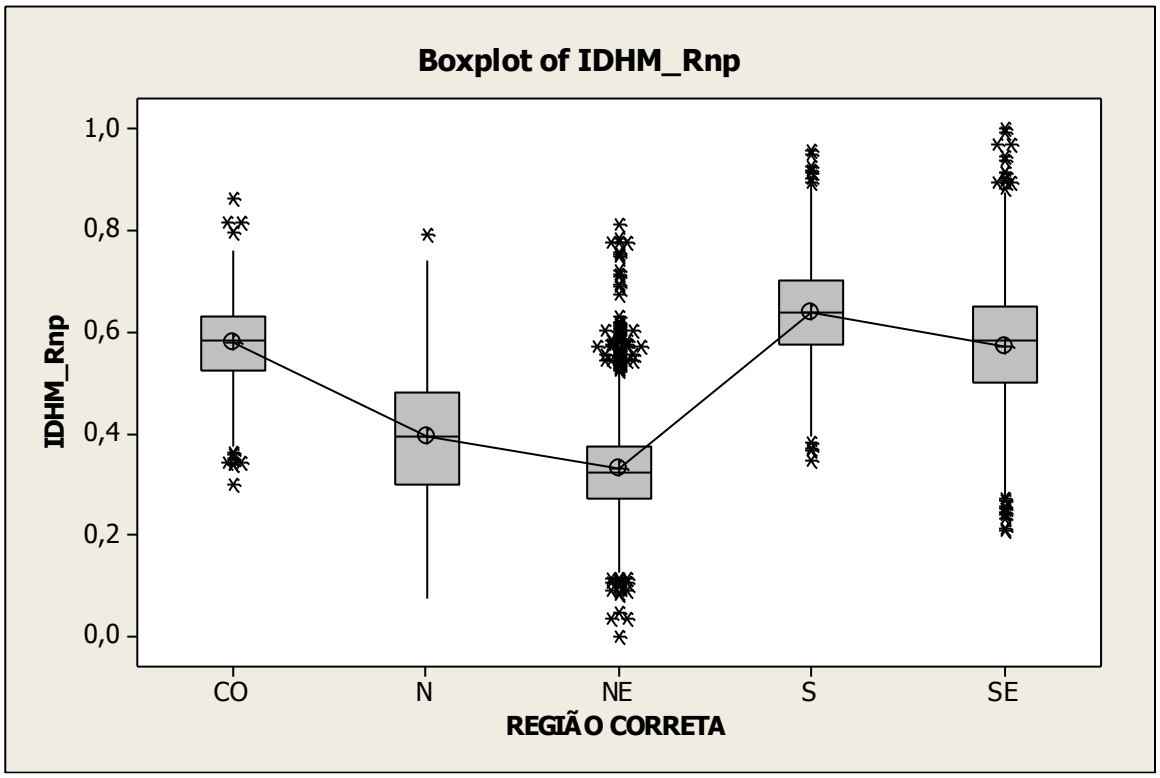
One-way ANOVA: ESPVIDAnp versus REGIÃO CORRETA

```
Source             DF        SS       MS        F      P
REGIÃO CORRETA      4  140,4313  35,1078  2319,16  0,000
Error            5559   84,1530   0,0151
Total            5563  224,5843

S = 0,1230   R-Sq = 62,53%   R-Sq(adj) = 62,50%


                                Individual 95% CIs For Mean Based on
                                Pooled StDev
Level     N    Mean   StDev  ---+---------+---------+---------+------
CO      465  0,6772  0,0825                                  (*)
N       449  0,4886  0,1292             (*)
NE     1794  0,3714  0,1356  *)
S      1188  0,7358  0,1177                                       (*
SE     1668  0,7036  0,1202                                    *)
                                ---+---------+---------+---------+------
                                 0,40      0,50      0,60      0,70

Pooled StDev = 0,1230
```

O grau de variação entre as Regiões é alto para ESPVIDA (2319,16), sendo inclusive maior que o de IDHM, mas o p-value nos indica que a informação é confiável e não existe chance deste valor ser diferente.



10.3 – VARIÁVEL IDHM_R POR REGIÃO

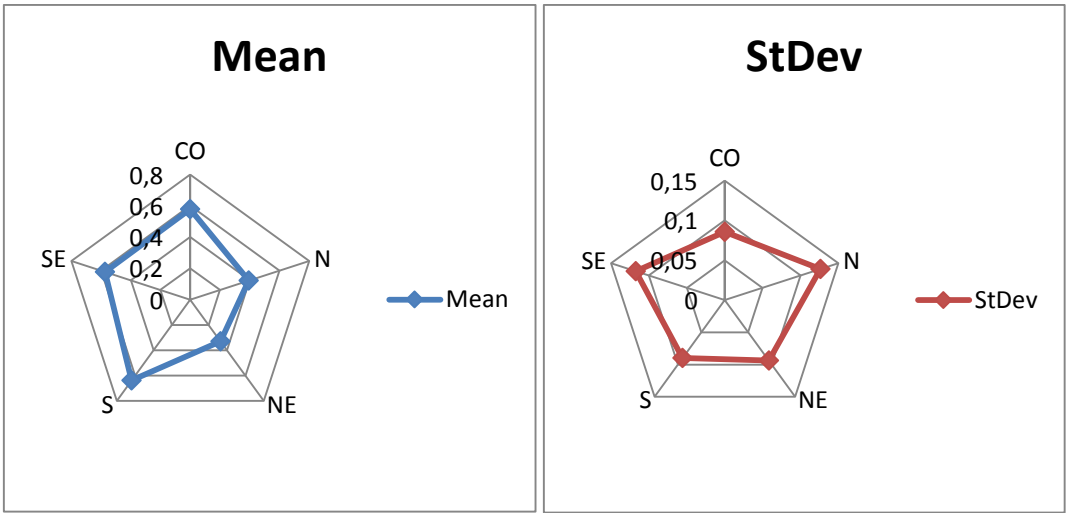


No indicador de **IDHM_R**, podemos verificar que a Região Sul possui o maior índice, que é seguida da Região Sudeste e Sul, e o menor índice é o da Região Nordeste. Vale a pena destacar que a região que apresenta a maior variação é a Norte.

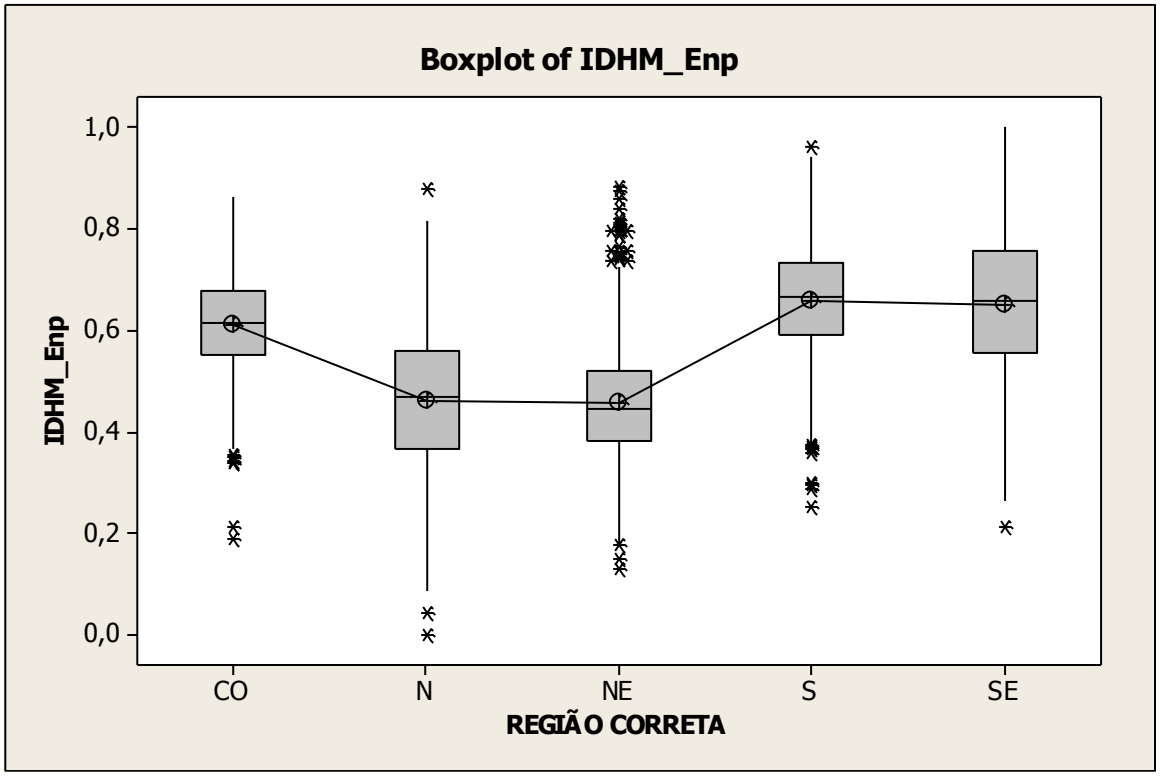
One-way ANOVA: IDHM_Rnp versus REGIÃO CORRETA

```
Source             DF        SS        MS         F       P
REGIÃO CORRETA      4   90,9836   22,7459   2143,97   0,000
Error            5559   58,9768    0,0106
Total            5563  149,9605

S = 0,1030   R-Sq = 60,67%   R-Sq(adj) = 60,64%


                                Individual 95% CIs For Mean Based on Pooled StDev
Level     N    Mean   StDev  ---------+---------+---------+---------+
CO      465  0,5786  0,0857                                  (*)
N       449  0,3927  0,1261         (*)
NE     1794  0,3305  0,0939  *)
S      1188  0,6384  0,0900                                        (*)
SE     1668  0,5726  0,1175                                  (*
                                ---------+---------+---------+---------+
                                       0,400     0,480     0,560     0,640

Pooled StDev = 0,1030
```

O grau de variação entre as Regiões é baixo para **IDHM_R** (2143,97). O p-value nos indica que a informação é confiável e não existe chance deste valor ser diferente.



10.4 – VARIÁVEL IDHM_E POR REGIÃO



Pode-se observar que o índice **IDHM_E** é maior na Regiões Sul, Sudeste e Centro Oeste. O menor índice é o da Região Norte.

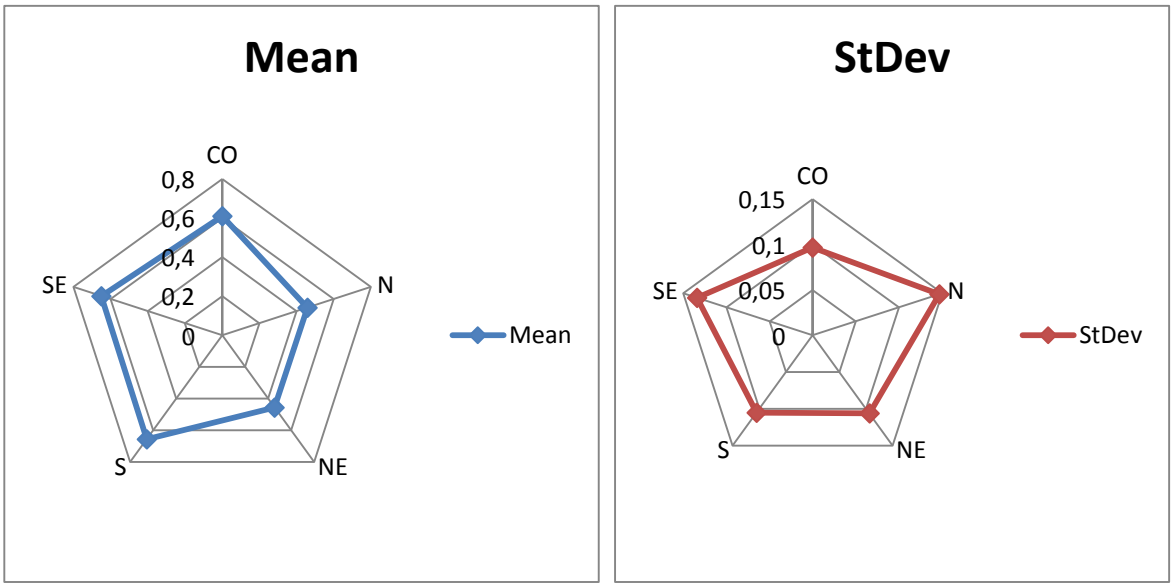
One-way ANOVA: IDHM_Enp versus REGIÃO CORRETA

Source	DF	SS	MS	F	P
REGIÃO CORRETA	4	49,3198	12,3300	884,60	0,000
Error	5559	77,4838	0,0139		
Total	5563	126,8037			

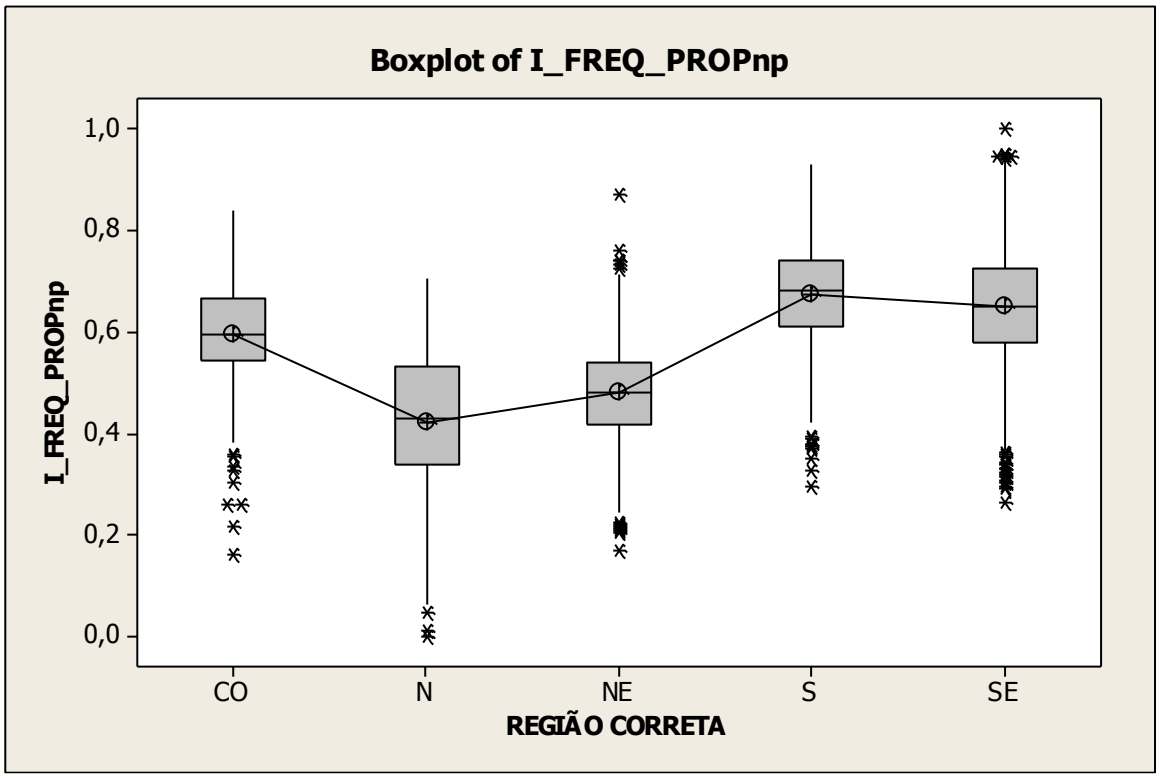
S = 0,1181 R-Sq = 38,89% R-Sq(adj) = 38,85%

```
                                Individual 95% CIs For Mean Based on
                                Pooled StDev
Level     N    Mean   StDev  -----+---------+---------+---------+----
CO      465  0,6097  0,0967                          (-*)
N       449  0,4586  0,1467  (*-)
NE     1794  0,4554  0,1064  (*)
S      1188  0,6570  0,1055                                    (*-)
SE     1668  0,6492  0,1341                                  (*)
                               -----+---------+---------+---------+----
                                  0,480     0,540     0,600     0,660

Pooled StDev = 0,1181
```

Existe uma variação maior na distribuição dos dados nos municípios da região Norte. Já as regiões Sul, Sudeste e Centro Oeste possuem as menores variações dos dados. O grau de variação entre as Regiões é alto (884,60) e o p-value nos indica que a informação é confiável e não existe chance deste valor ser diferente.



10.5 – VARIÁVEL POR I_FREQ_PROP REGIÃO



Pode-se observar que o valor de **I_FREQ_PROP** é maior nas Regiões Sul, Sudeste e Centro Oeste. Vale a pena destacar que a região Norte apresenta a maior variação.

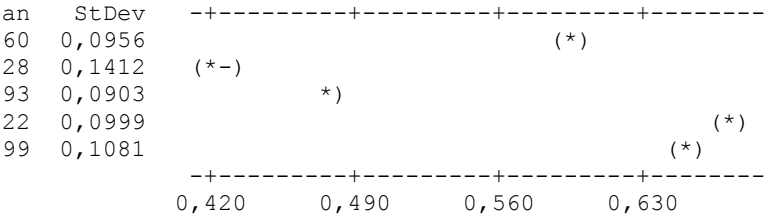
One-way ANOVA: I_FREQ_PROPnp versus REGIÃO CORRETA

Source	DF	SS	MS	F	P
REGIÃO CORRETA	4	47,5969	11,8992	1118,63	0,000
Error	5559	59,1330	0,0106		
Total	5563	106,7299			

S = 0,1031 R-Sq = 44,60% R-Sq(adj) = 44,56%

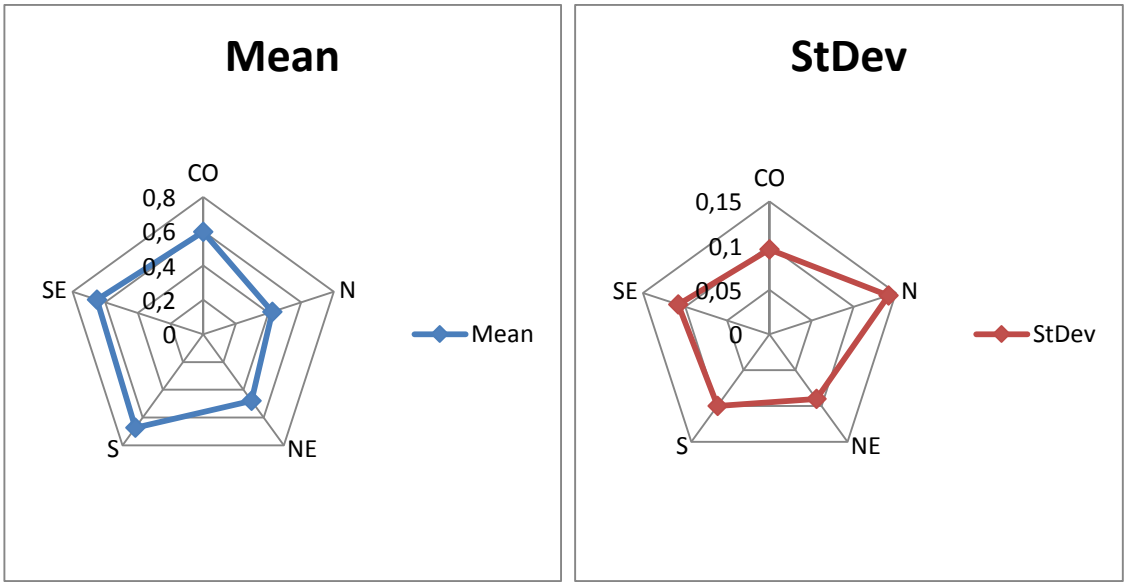
Level	N	Mean	StDev
CO	465	0,5960	0,0956
N	449	0,4228	0,1412
NE	1794	0,4793	0,0903
S	1188	0,6722	0,0999
SE	1668	0,6499	0,1081

Individual 95% CIs For Mean Based on Pooled StDev

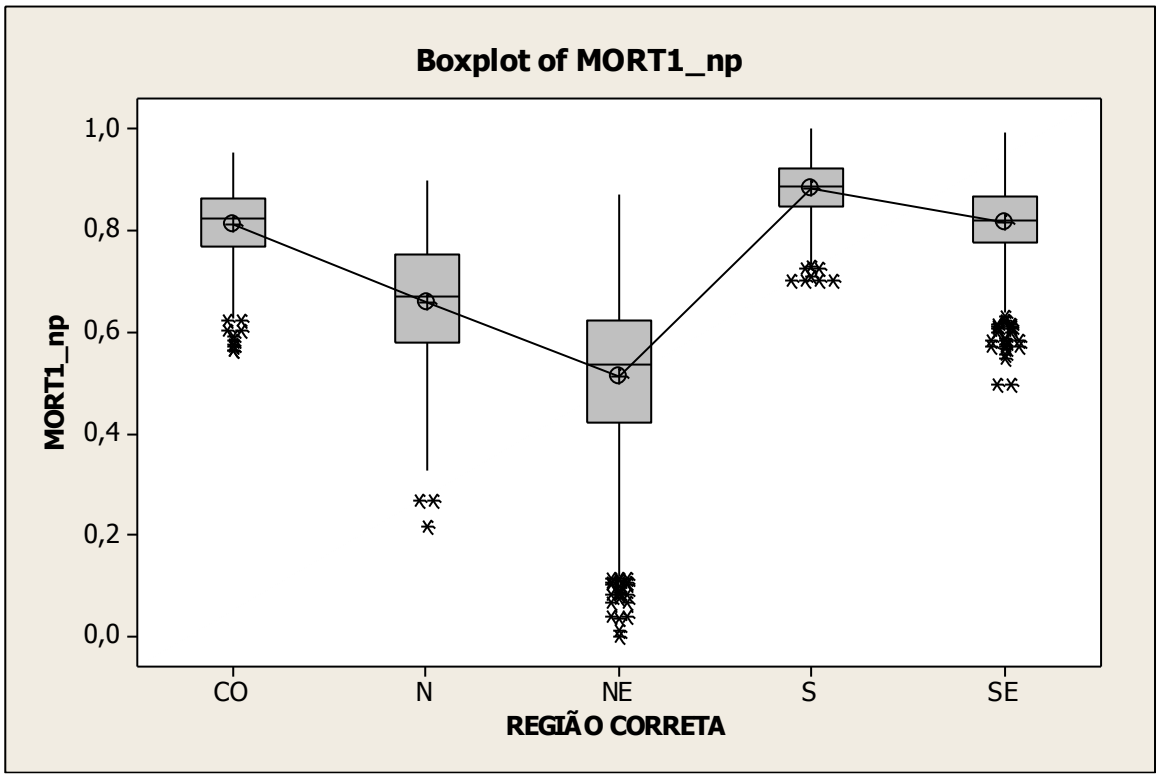


Pooled StDev = 0,1031

O grau de variação entre as Regiões é alto (1118,63) e o p-value nos indica que a informação é confiável e não existe chance deste valor ser diferente.



10.6 – VARIÁVEL MORT1 POR REGIÃO



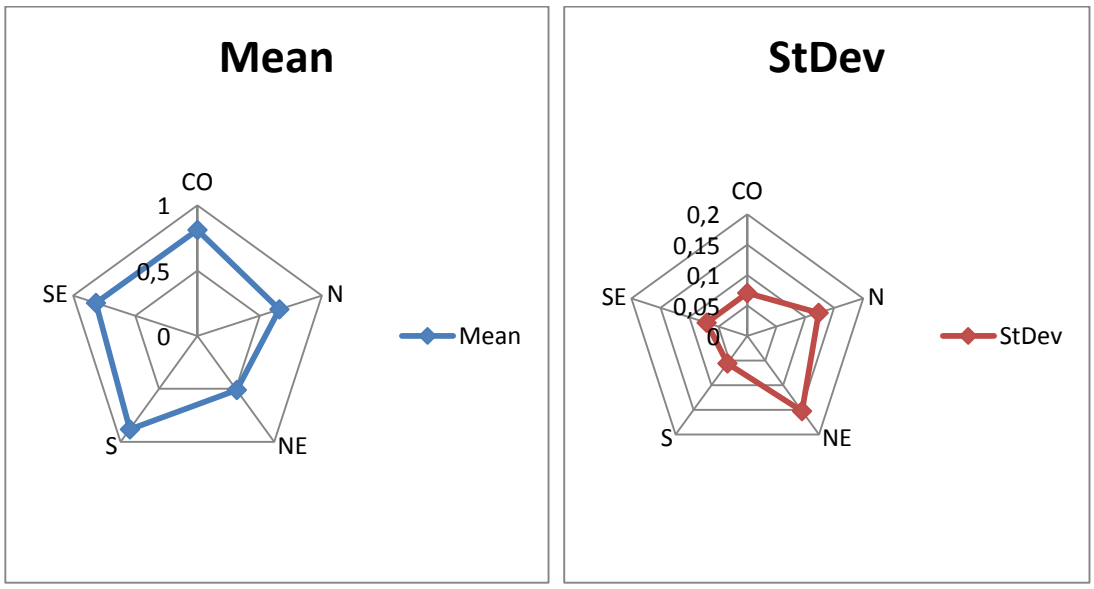
Pode-se observar que o índice **MORT1** é maior nas Regiões Sul, Sudeste e Centro Oeste. O menor índice é o da Região Nordeste. Vale a pena destacar que apresentam maior varião são Nordeste e Norte. Além disso, os municípios que apresentam resultados piores são considerados nesta análise como atípicos.

A maior variabilidade de dados se encontra na Região Sudeste, e a Região que apresenta menor variabilidade é a Nordeste.

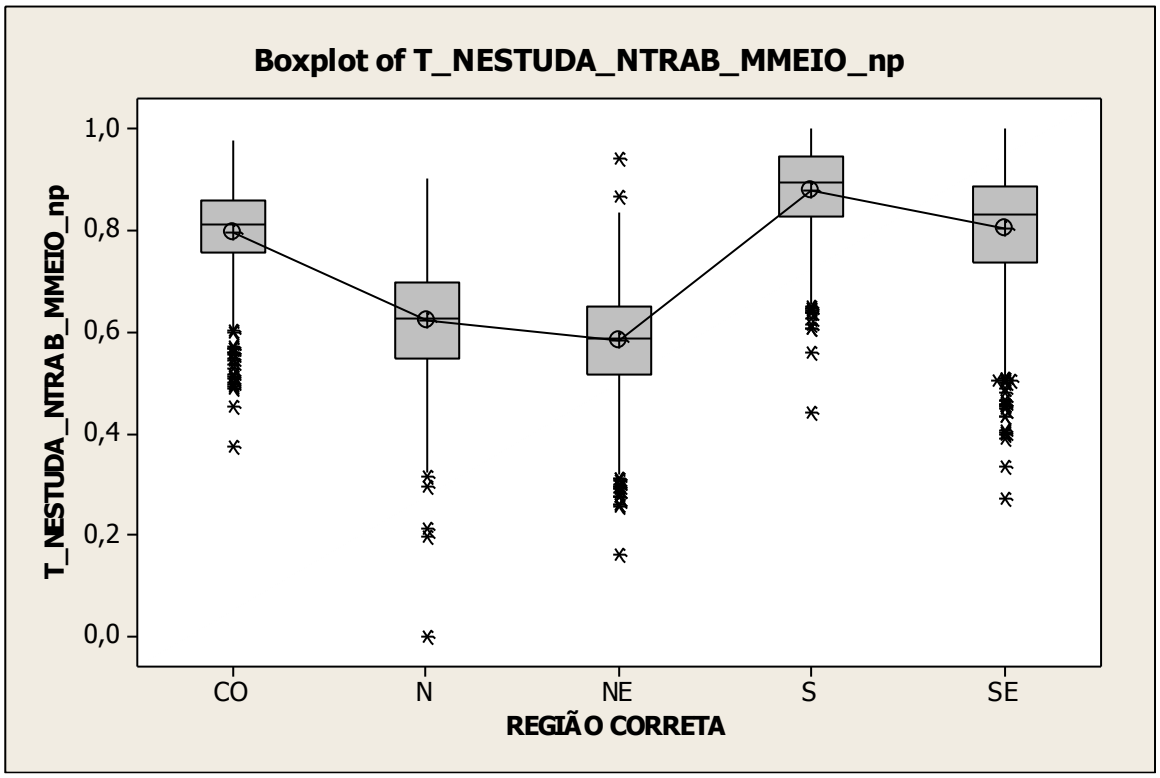
One-way ANOVA: MORT1_np versus REGIÃO CORRETA

```
Source             DF        SS       MS        F      P
REGIÃO CORRETA      4  130,1073  32,5268  2871,93  0,000
Error            5559   62,9600   0,0113
Total            5563  193,0673

S = 0,1064   R-Sq = 67,39%   R-Sq(adj) = 67,37%


                                Individual 95% CIs For Mean Based on
                                Pooled StDev
Level     N    Mean   StDev  ---------+---------+---------+---------+
CO      465  0,8110  0,0706                                    (*)
N       449  0,6584  0,1230                          (*)
NE     1794  0,5119  0,1529  *)
S      1188  0,8823  0,0558                                        *)
SE     1668  0,8167  0,0705                                   (*
                                ---------+---------+---------+---------+
                                       0,60      0,70      0,80      0,90

Pooled StDev = 0,1064
```

O grau de variação entre as Regiões é alto (2871,93) e o p-value = 0 nos indica que a informação é confiável e não existe chance deste valor ser diferente.



10.7 – VARIÁVEL T_NESTUDA_MMEIO POR REGIÃO

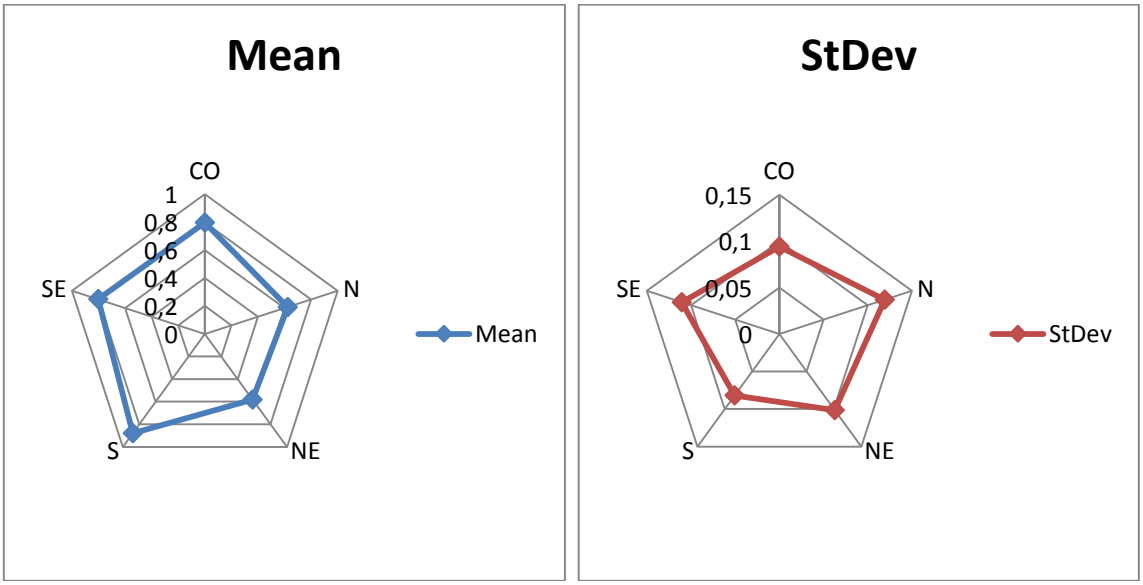


Pode-se observar que o índice **T_NESTUDA_MMEIO** é maior nas Regiões Sul, Sudeste e Centro Oeste. Os índices mais baixos estão nas regiões Norte e Nordeste. Vale a pena destacar que as regiões apresentam uma média ou baixa variação. Além disso, os municípios que apresentam resultados piores são considerados nesta análise como atípicos, exceto no Nordeste.

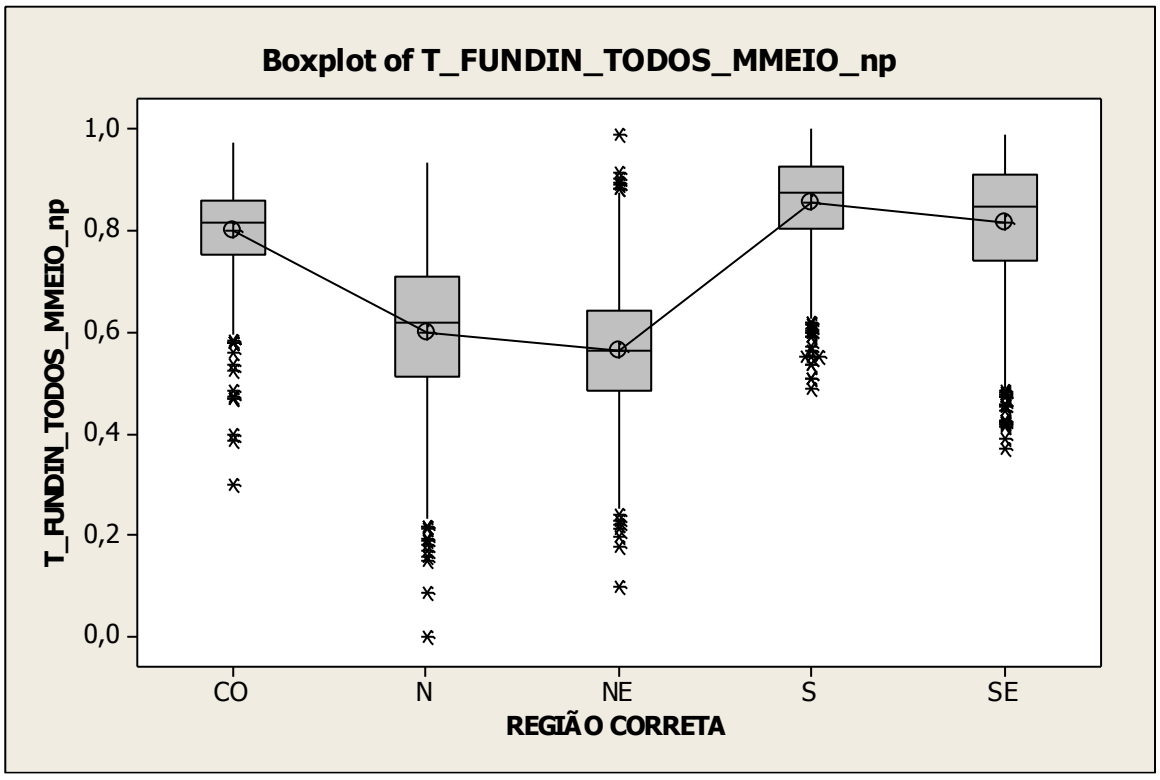
One-way ANOVA: T_NESTUDA_NTRAB_MMEIO_np versus REGIÃO CORRETA

```
Source             DF        SS       MS        F      P
REGIÃO CORRETA      4   82,1968  20,5492  1991,10  0,000
Error            5559   57,3719   0,0103
Total            5563  139,5687

S = 0,1016   R-Sq = 58,89%   R-Sq(adj) = 58,86%


                            Individual 95% CIs For Mean Based on
                            Pooled StDev
Level     N    Mean   StDev  --------+---------+---------+---------+-
CO      465  0,7958  0,0941                            (*-)
N       449  0,6222  0,1194       (*)
NE     1794  0,5813  0,1016  (*
S      1188  0,8794  0,0820                                      (*)
SE     1668  0,8027  0,1105                              *)
                            --------+---------+---------+---------+-
                                  0,640     0,720     0,800     0,880

Pooled StDev = 0,1016
```

O grau de variação entre as Regiões é alto (1991,10) e o p-value igual a zero nos indica que a informação é confiável e não existe chance deste valor ser diferente.



10.8 – VARIÁVEL T_FUNDIN_TODOS_MMEIO POR REGIÃO

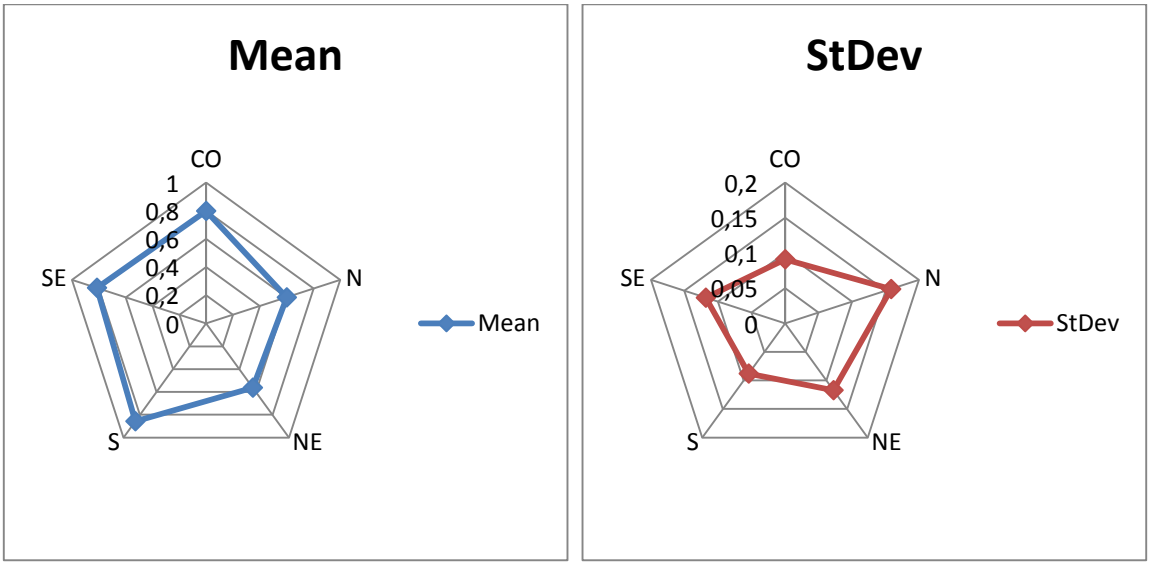


Pode-se observar que o valor de **T_FUNDIN_TODOS_MMEIO** é maior nas Regiões Sul, Sudeste e Centro Oeste. Os índices mais baixos estão nas regiões Norte e Nordeste. Vale a pena destacar que a região Norte apresenta a maior variação. Além disso, os municípios que apresentam resultados piores nas cinco regiões são considerados nesta análise como atípicos, inclusive estes valores são muito distantes da média.

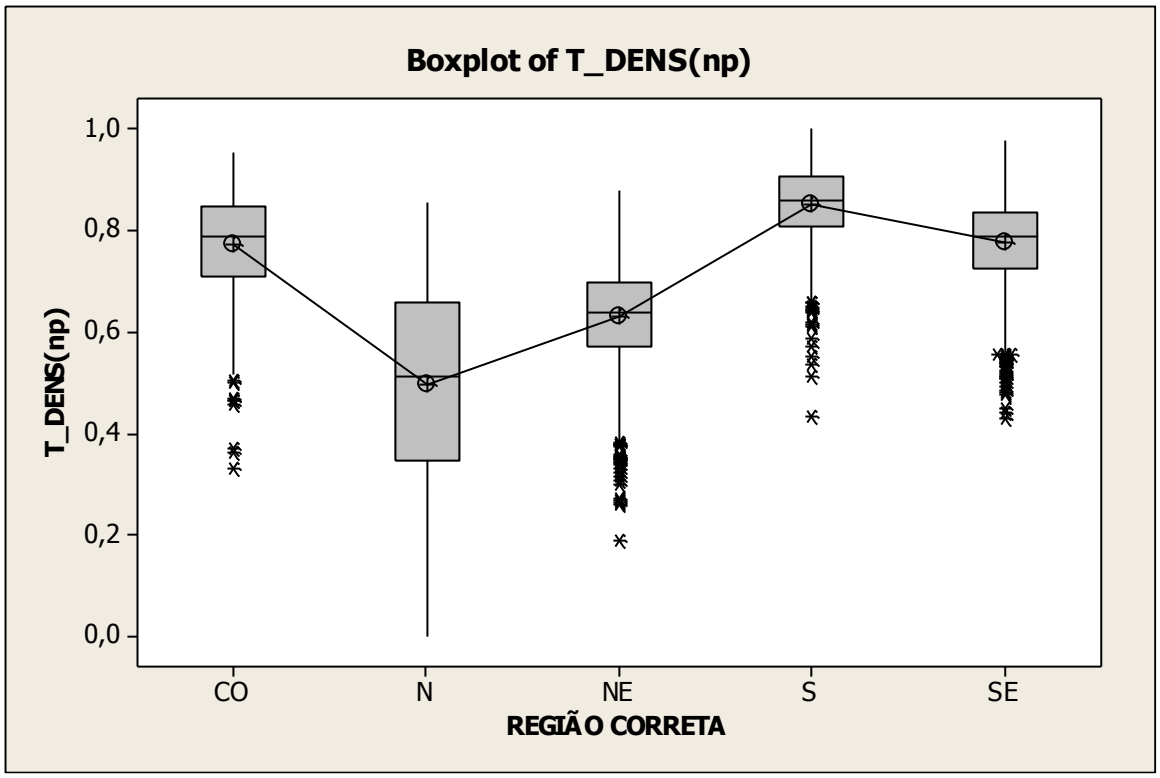
One-way ANOVA: T_FUNDIN_TODOS_MMEIO_np versus REGIÃO CORRETA

```
Source             DF         SS        MS         F      P
REGIÃO CORRETA      4    91,1520   22,7880   1753,51  0,000
Error            5559    72,2428    0,0130
Total            5563   163,3948

S = 0,1140    R-Sq = 55,79%    R-Sq(adj) = 55,75%


                                 Individual 95% CIs For Mean Based on Pooled StDev
Level      N    Mean   StDev    +---------+---------+---------+---------
CO       465  0,7978  0,0910                                (-*)
N        449  0,5992  0,1581         (*)
NE      1794  0,5626  0,1174    *)
S       1188  0,8569  0,0878                                       (*)
SE      1668  0,8154  0,1183                                   (*)
                                 +---------+---------+---------+---------
                                0,560     0,640     0,720     0,800

Pooled StDev = 0,1140
```

O grau de variação entre as Regiões é alto (1753,51) e o p-value igual zero nos indica que a informação é confiável e não existe chance deste valor ser diferente.



10.9 – VARIÁVEL T_DENS POR REGIÃO

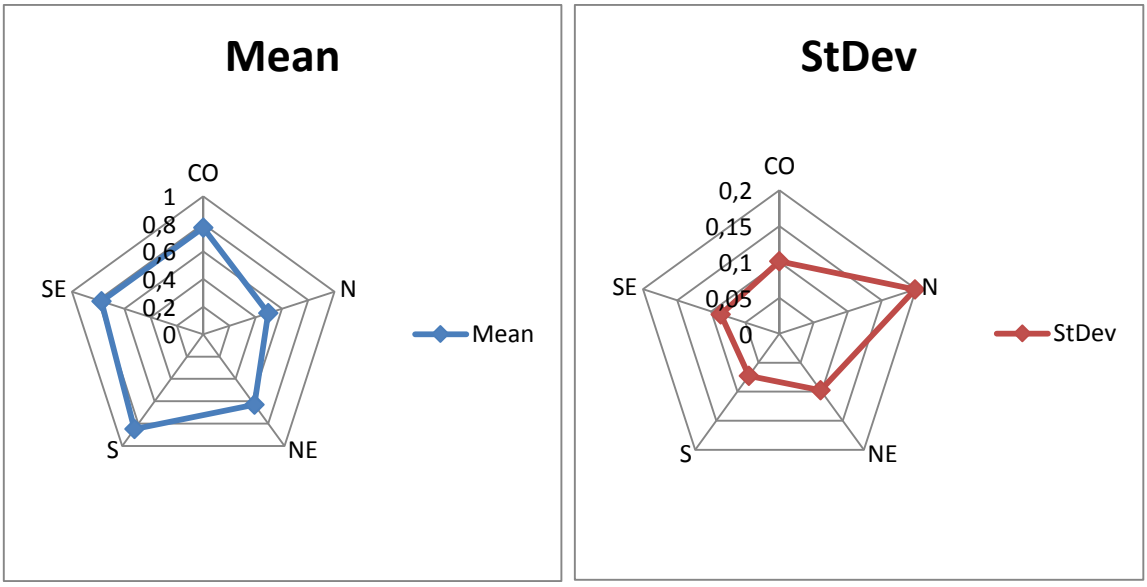


Pode-se observar que o índice de **T_DENS** é maior nas Regiões Sul e Sudeste. Já as regiões Norte e Nordeste apresentam os piores resultados. Vale destacar que a região Norte apresenta uma alta variação nos resultados de seus municípios

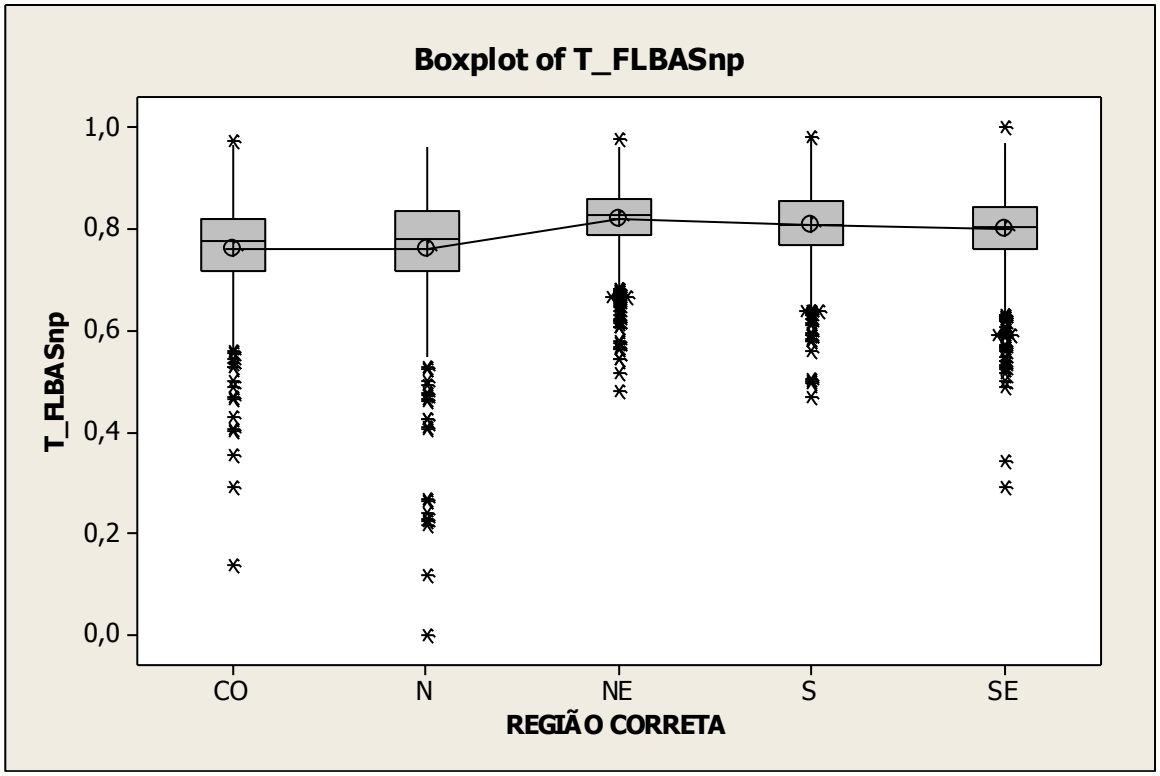
One-way ANOVA: T_DENS(np) versus REGIÃO CORRETA

```
Source             DF        SS       MS        F      P
REGIÃO CORRETA      4   63,2432  15,8108  1510,67  0,000
Error            5559   58,1811   0,0105
Total            5563  121,4244

S = 0,1023   R-Sq = 52,08%   R-Sq(adj) = 52,05%


                              Individual 95% CIs For Mean Based on
                              Pooled StDev
Level     N    Mean   StDev  --+---------+---------+---------+-------
CO      465  0,7712  0,1009                              (*)
N       449  0,4944  0,1988  (*)
NE     1794  0,6316  0,0978                *)
S      1188  0,8502  0,0726                                      (*)
SE     1668  0,7749  0,0861                                   *)
                              --+---------+---------+---------+-------
                              0,50      0,60      0,70      0,80

Pooled StDev = 0,1023
```

O grau de variação entre as Regiões é alto (1510,67) e o p-value nos indica que a informação é confiável e não existe chance deste valor ser diferente.



10.10 – VARIÁVEL T_FLBAS POR REGIÃO.



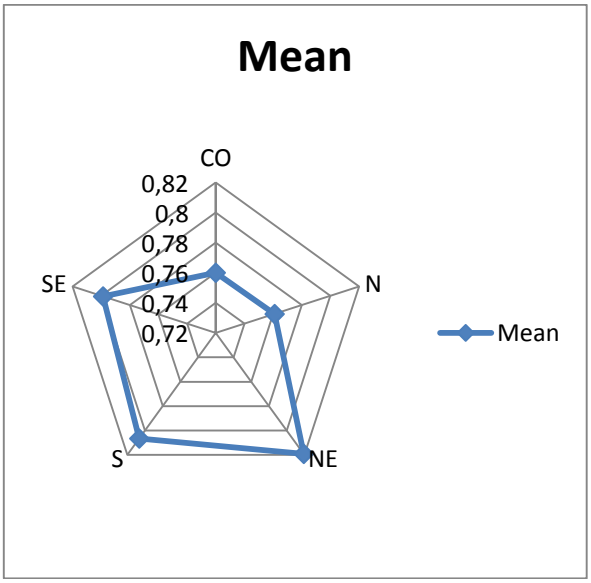
Pode-se observar que o índice de **T_FLBAS** é maior nas Regiões Nordeste e Sul, mas todas as regiões apresentam valores para esta variável bem próximos (0,80).

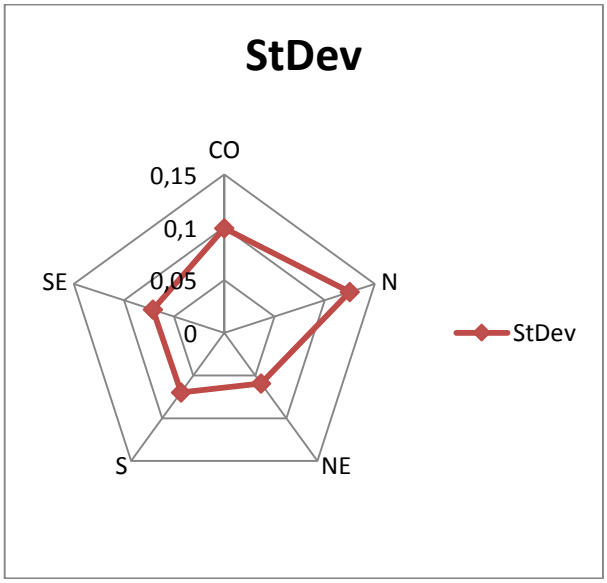
One-way ANOVA: T_FLBASnp versus REGIÃO CORRETA

Source	DF	SS	MS	F	P
REGIÃO CORRETA	4	2,13722	0,53430	92,37	0,000
Error	5559	32,15483	0,00578		
Total	5563	34,29205			

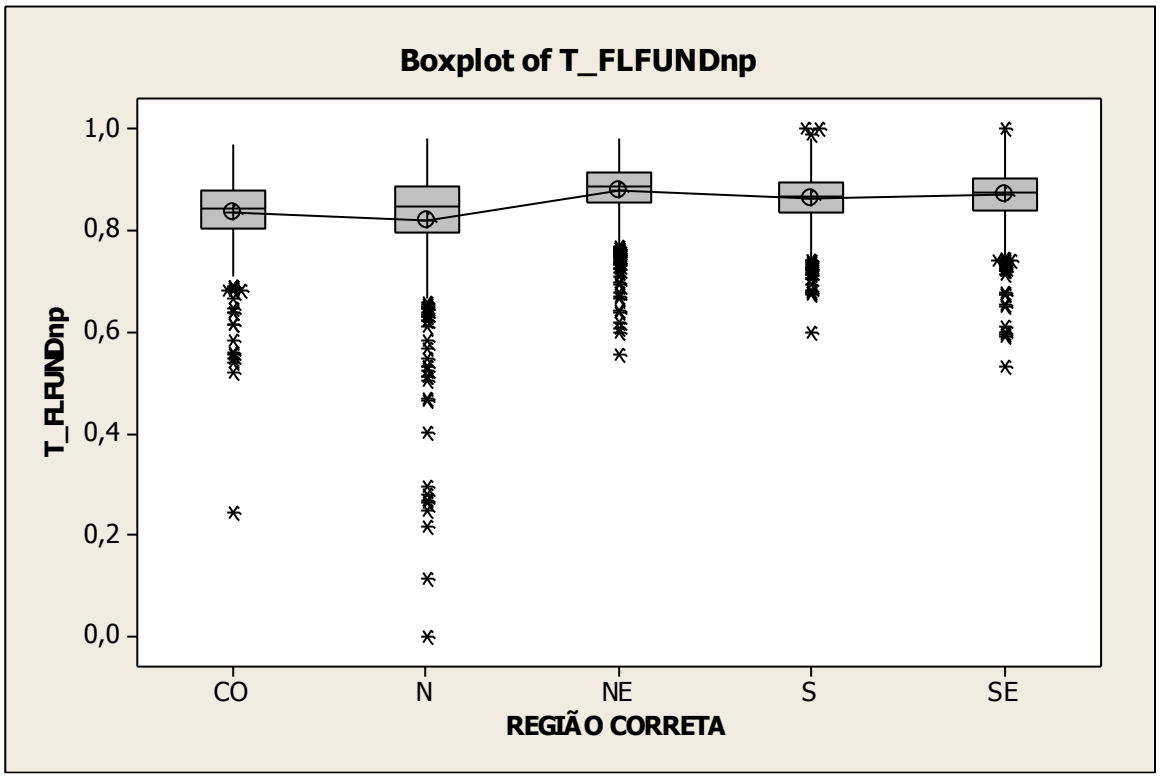
```
S = 0,07605   R-Sq = 6,23%   R-Sq(adj) = 6,16%


                                 Individual 95% CIs For Mean Based on
                                 Pooled StDev
Level     N     Mean    StDev    ---+---------+---------+---------+------
CO      465  0,76002  0,09902    (--*--)
N       449  0,76091  0,12513    (--*---)
NE     1794  0,81911  0,05939                                    (-*)
S      1188  0,80666  0,06996                              (-*-)
SE     1668  0,79870  0,07123                         (*-)
                                 ---+---------+---------+---------+------
                                  0,760     0,780     0,800     0,820

Pooled StDev = 0,07605
```

O grau de variação entre as Regiões é baixo (92,37) e o p-value nos indica que a informação é confiável e não existe chance deste valor ser diferente.





10.11 – VARIÁVEL T_FLFUND POR REGIÃO

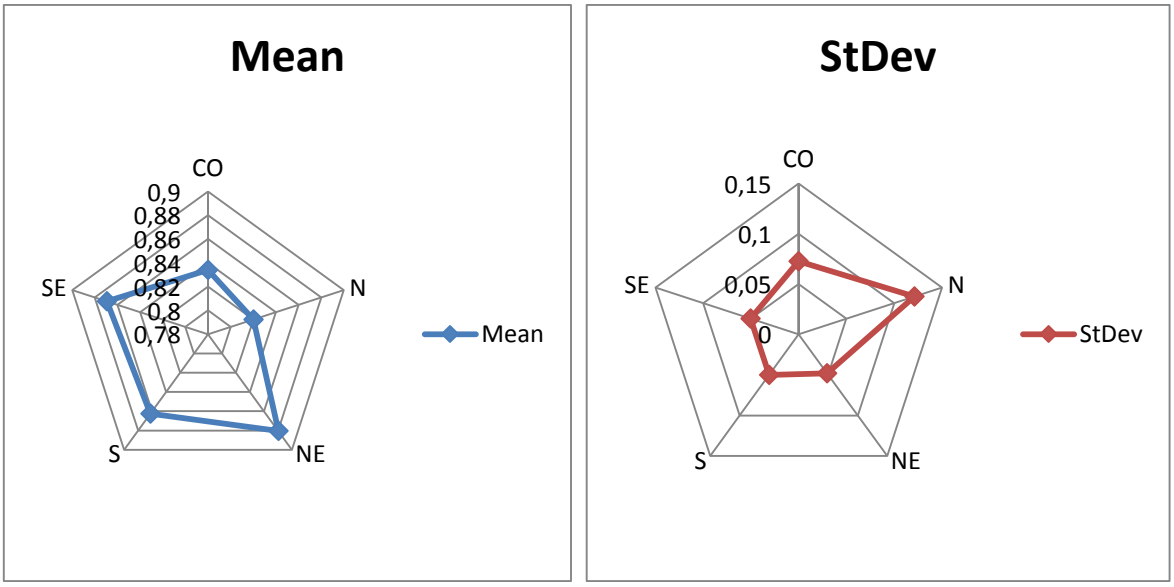


Pode-se observar que o percentual de **T_FLFUND** é é maior nas Regiões Nordeste e Sul, mas todas as regiões apresentam valores para esta variável bem próximos (0,80). Vale destacar que os valores atípicos em todas as regiões são representados pelos municípios com piores resultados.

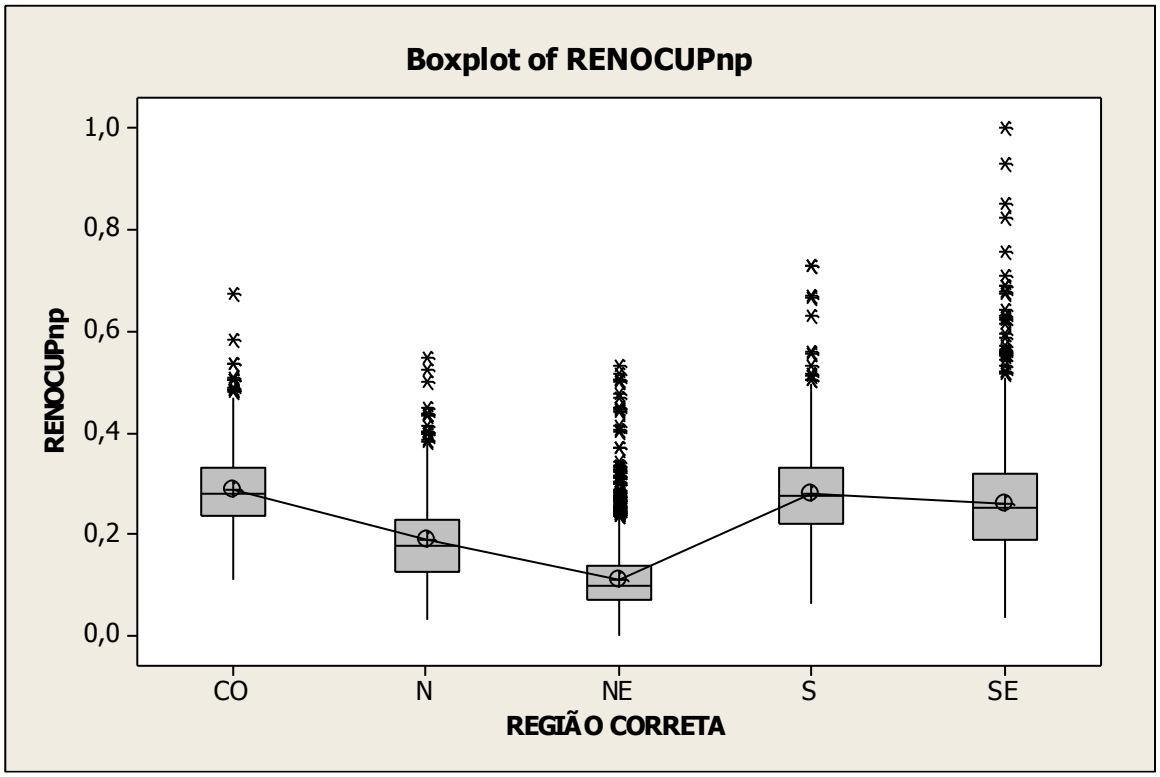
One-way ANOVA: T_FLFUNDnp versus REGIÃO CORRETA

```
Source            DF        SS       MS       F      P
REGIÃO CORRETA     4   1,81941  0,45485  124,16  0,000
Error           5559  20,36457  0,00366
Total           5563  22,18398

S = 0,06053   R-Sq = 8,20%   R-Sq(adj) = 8,14%


                              Individual 95% CIs For Mean Based on
                              Pooled StDev
Level     N     Mean    StDev ---+---------+---------+---------+------
CO      465  0,83394  0,07271        (--*--)
N       449  0,82012  0,12118  (--*--)
NE     1794  0,88043  0,04804                                    (*-)
S      1188  0,86246  0,05009                           (*-)
SE     1668  0,86942  0,05029                              (-*)
                              ---+---------+---------+---------+------
                               0,820     0,840     0,860     0,880

Pooled StDev = 0,06053
```

O grau de variação entre as Regiões é alto (124,16) e o p-value nos indica que a informação é confiável e não existe chance deste valor ser diferente.



10.12 – VARIÁVEL RENOCUP POR REGIÃO

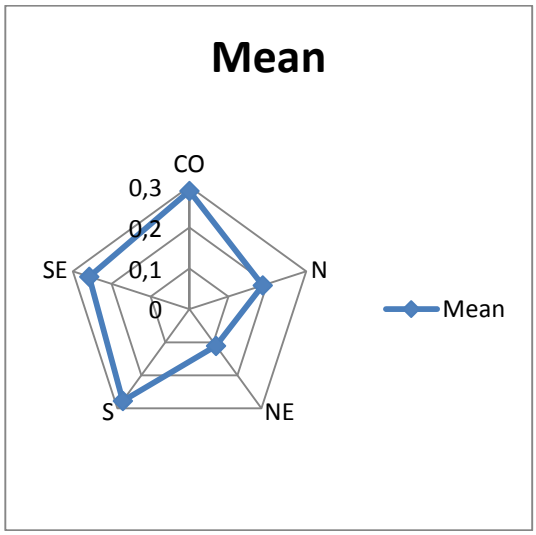


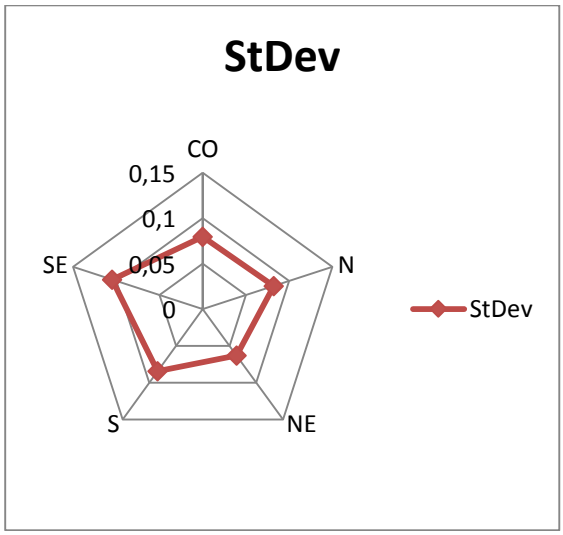
Pode-se observar que o percentual de **RENOCUP** é maior nas Regiões Sul, Sudeste e Centro Oeste. Já as regiões Norte e Nordeste apresentam índices baixos para esta variável, com destaque para o Nordeste. Vale destacar que os municípios que apresentam os melhores resuldados para esta variável são considerados atípicos em todas as regiões.

One-way ANOVA: RENOCUPnp versus REGIÃO CORRETA

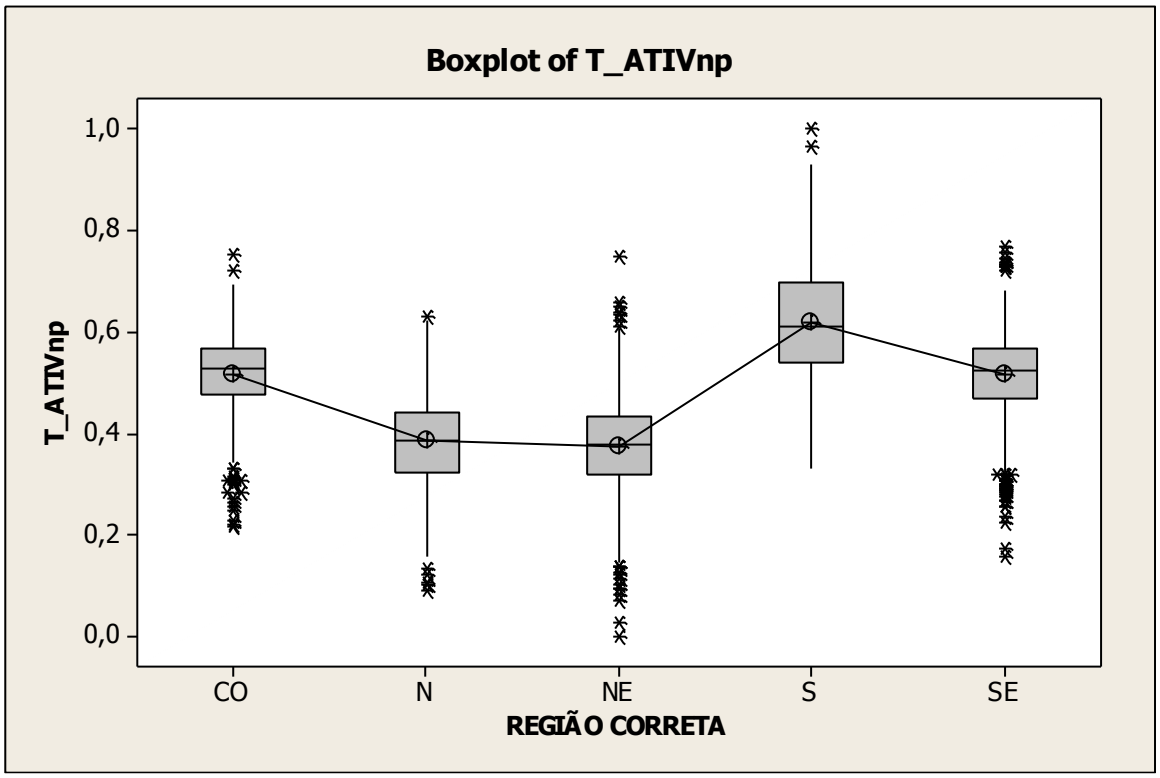
```
Source             DF        SS        MS         F      P
REGIÃO CORRETA      4  29,91656  7,47914  1039,86  0,000
Error            5559  39,98282  0,00719
Total            5563  69,89938

S = 0,08481   R-Sq = 42,80%   R-Sq(adj) = 42,76%



Level      N     Mean    StDev
CO       465  0,28898  0,07945
N        449  0,18801  0,08223
NE      1794  0,11091  0,06318
S       1188  0,27744  0,08463
SE      1668  0,25770  0,10497

        Individual 95% CIs For Mean Based on Pooled StDev
Level   ---------+---------+---------+---------+
CO                                        (-*)
N                         (-*)
NE      (*)
S                                       *)
SE                                   (*
        ---------+---------+---------+---------+
               0,150     0,200     0,250     0,300

Pooled StDev = 0,08481
```

O grau de variação entre as Regiões é alto (1039,86) e o p-value nos indica que a informação é confiável e não existe chance deste valor ser diferente.





10.13 – VARIÁVEL T_ATIV POR REGIÃO

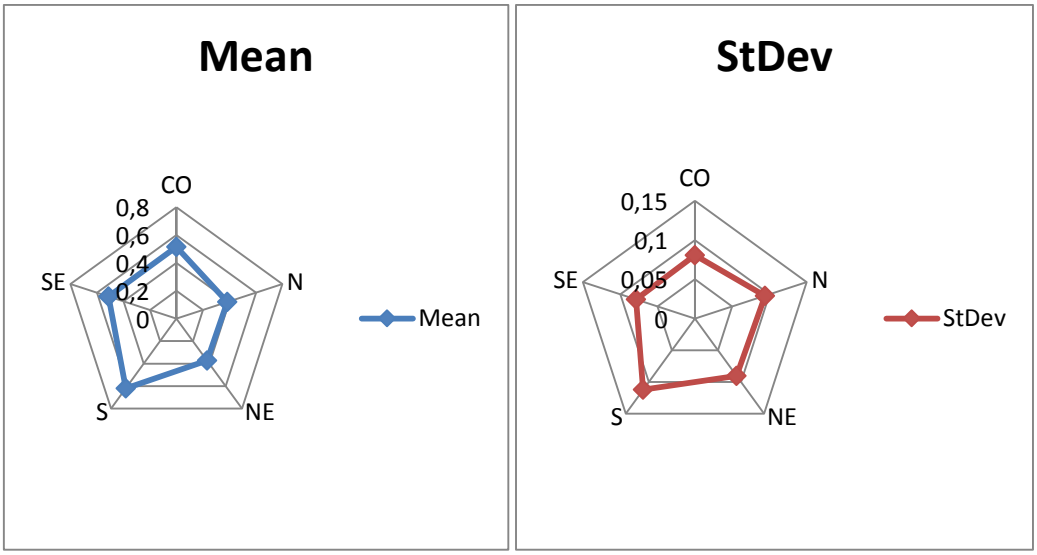


No indicador de **T_ATIV**, podemos verificar que a Região Sul possui o maior índice, que é seguida da Região Sudeste e Centro Oeste, e o menor índice é o da Região Norte.

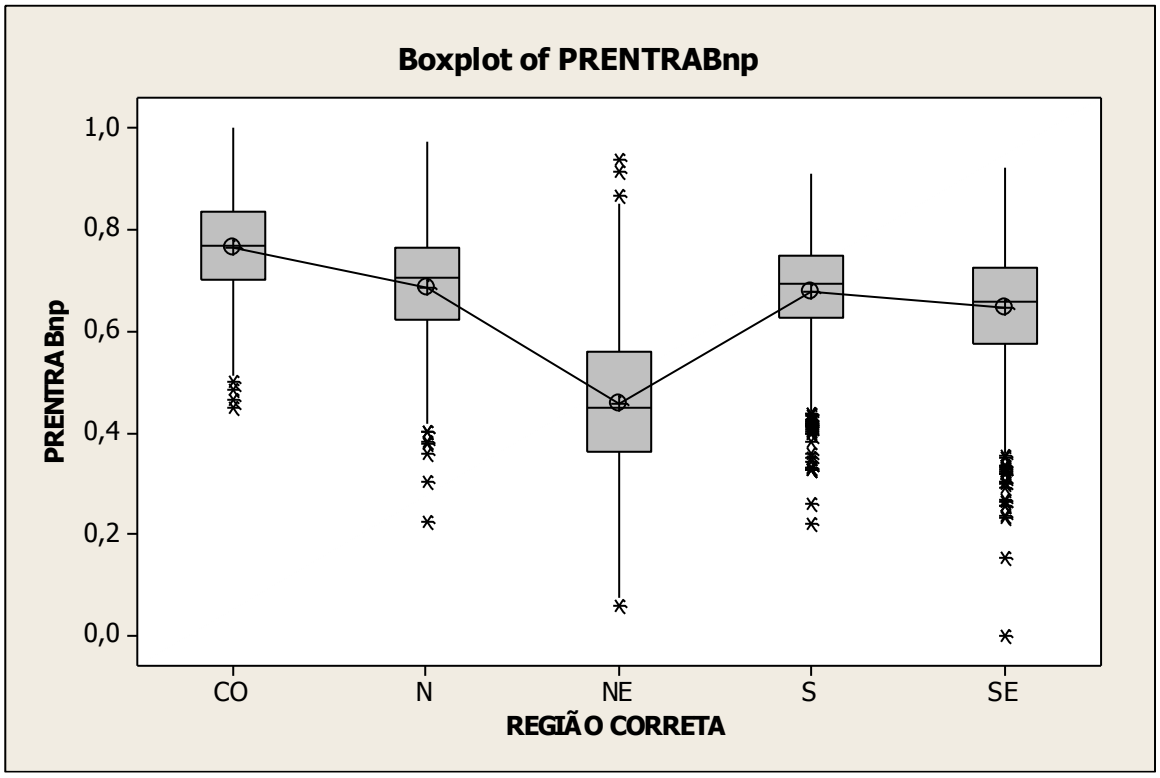
One-way ANOVA: T_ATIVnp versus REGIÃO CORRETA

```
Source            DF        SS        MS        F      P
REGIÃO CORRETA     4  49,84347  12,46087  1471,83  0,000
Error           5559  47,06386   0,00847
Total           5563  96,90733

S = 0,09201   R-Sq = 51,43%   R-Sq(adj) = 51,40%


                                Individual 95% CIs For Mean Based on
                                Pooled StDev
Level     N     Mean    StDev  -------+---------+---------+---------+--
CO      465  0,51470  0,08064                    (-*)
N       449  0,38417  0,09416   (*)
NE     1794  0,37347  0,09055  *)
S      1188  0,61882  0,11229                                    *)
SE     1668  0,51383  0,07900                     *)
                                -------+---------+---------+---------+--
                                     0,420     0,490     0,560     0,630

Pooled StDev = 0,09201
```

O grau de variação entre as Regiões é alto (1471,83) e o p-value nos indica que a informação é confiável e não existe chance deste valor ser diferente.



10.14 – VARIÁVEL PRENTRAB POR REGIÃO



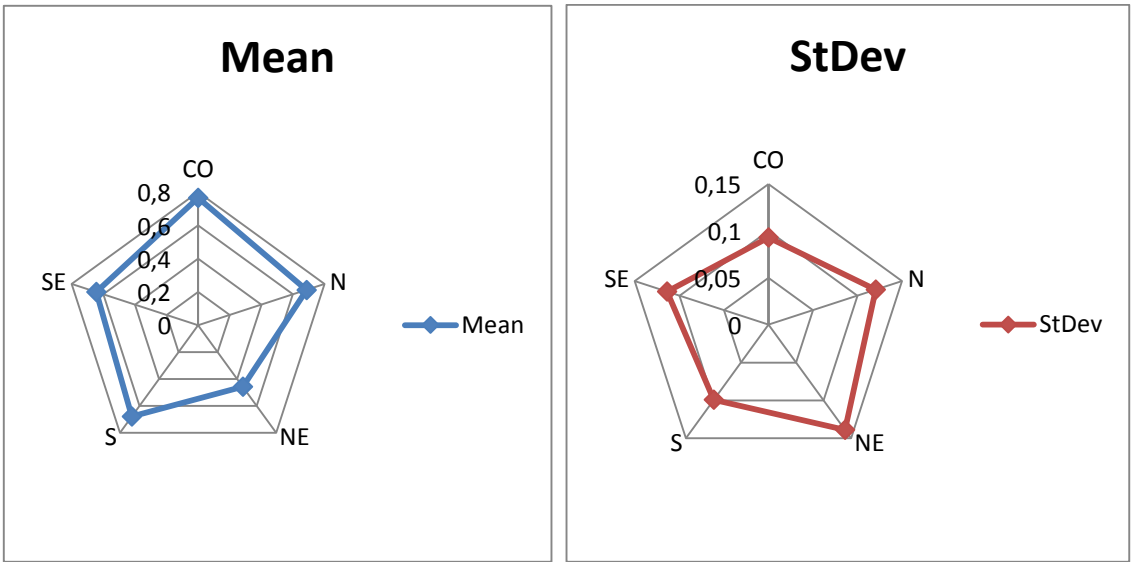
No indicador de **PRENTRAB**, podemos verificar que a Região Nordeste possui o menor índice, inclusive se distanciando das demais regiões. Além disso, a Região Nordeste ainda apresenta uma variação grande entre seus municípios.

One-way ANOVA: PRENTRABnp versus REGIÃO CORRETA

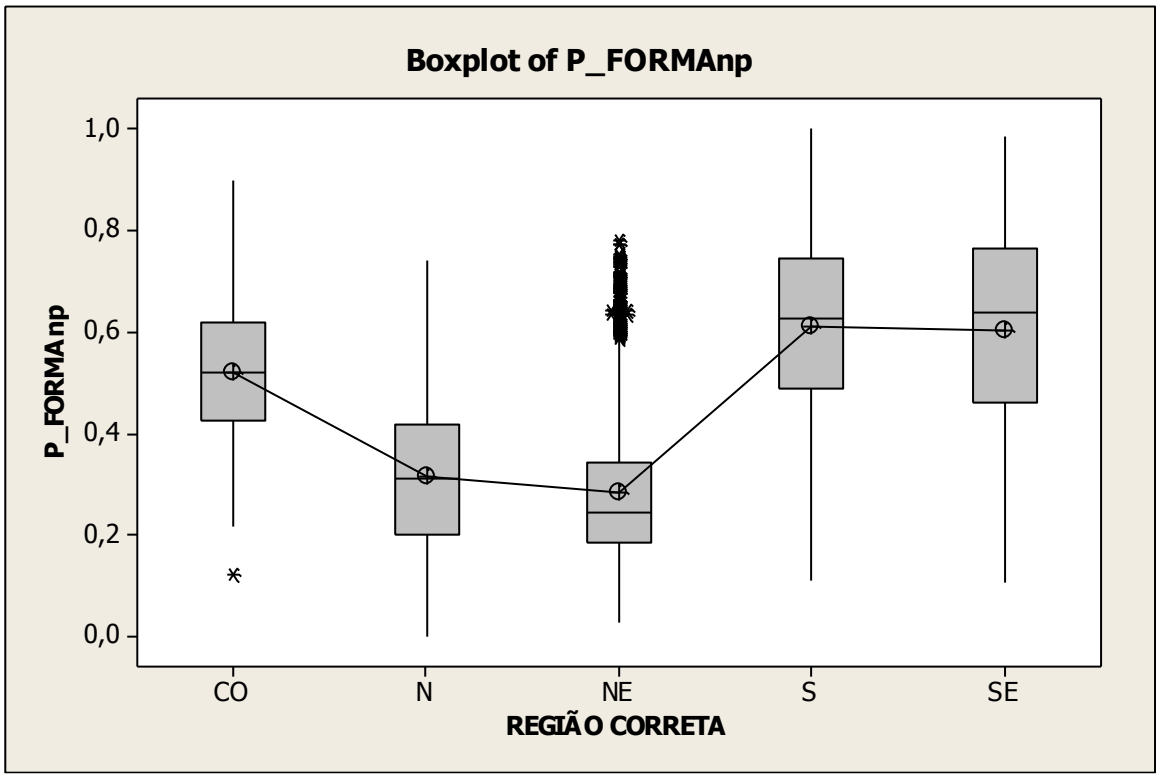
Source	DF	SS	MS	F	P
REGIÃO CORRETA	4	62,6591	15,6648	1111,02	0,000

```
Error            5559   78,3786   0,0141
Total            5563  141,0376

S = 0,1187   R-Sq = 44,43%   R-Sq(adj) = 44,39%


                                Individual 95% CIs For Mean Based on
                                Pooled StDev
Level     N    Mean   StDev  -----+---------+---------+---------+----
CO      465  0,7659  0,0931                                    (*)
N       449  0,6865  0,1203                              (*)
NE     1794  0,4585  0,1390  (*
S      1188  0,6792  0,0991                             (*)
SE     1668  0,6443  0,1137                          *)
                               -----+---------+---------+---------+----
                                  0,50      0,60      0,70      0,80

Pooled StDev = 0,1187
```

O grau de variação entre as Regiões é alto (1111,02) e o p-value nos indica que a informação é confiável e não existe chance deste valor ser diferente.



10.15 – VARIÁVEL P_FORMA POR REGIÃO

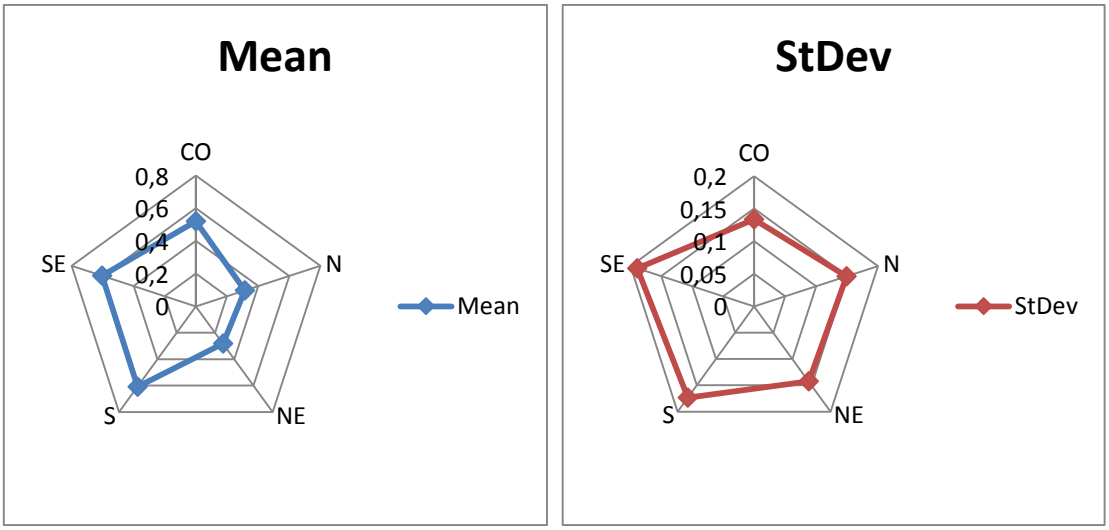


No indicador de **P_FORMA**, podemos verificar que a Região Nordeste possui o menor índice, seguida pela região Norte. Vale a pena destacar que para esta variável as regiões apresentam uma alta variação, exceto no Nordeste, sendo que esta região também apresenta dados atípicos como municípios com melhores índices.

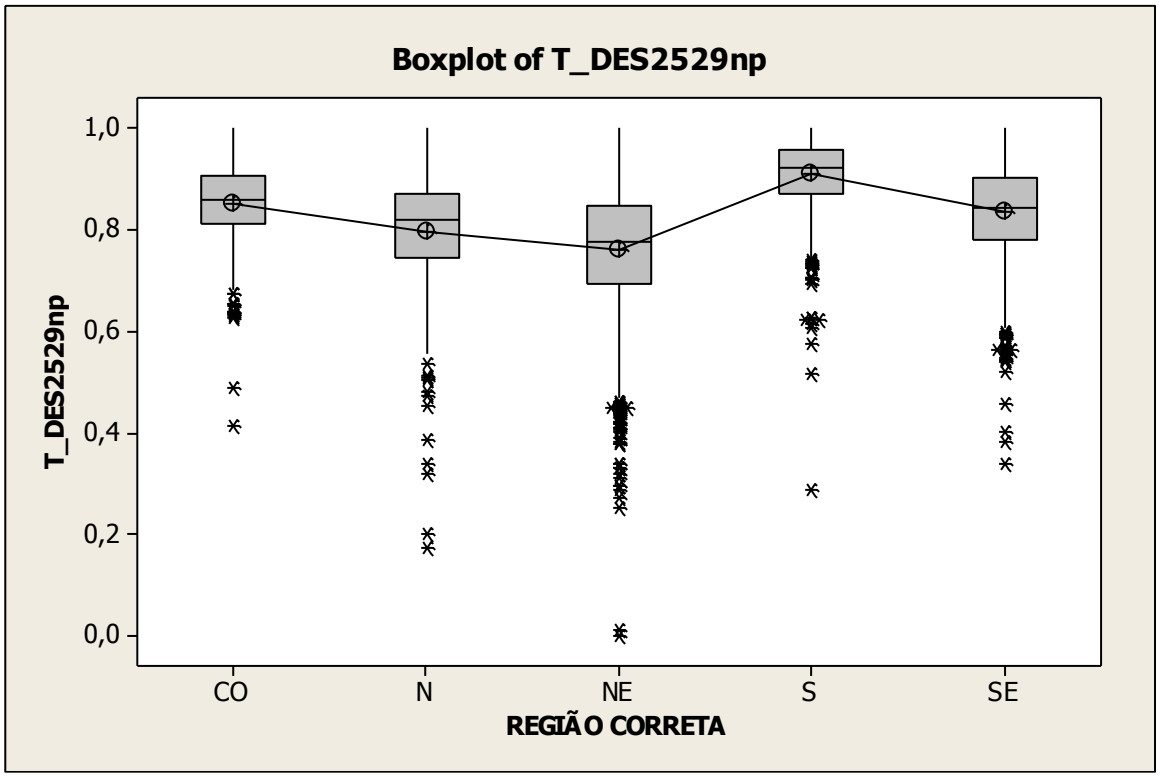
One-way ANOVA: P_FORMAnp versus REGIÃO CORRETA

```
Source            DF        SS       MS        F      P
REGIÃO CORRETA     4  128,3849  32,0962  1188,70  0,000
Error           5559  150,0988   0,0270
Total           5563  278,4837

S = 0,1643   R-Sq = 46,10%   R-Sq(adj) = 46,06%


                              Individual 95% CIs For Mean Based on
                              Pooled StDev
Level     N    Mean   StDev  ---+---------+---------+---------+------
CO      465  0,5192  0,1334                          (-*)
N       449  0,3141  0,1495     (*-)
NE     1794  0,2816  0,1425  (*)
S      1188  0,6090  0,1736                                    (*)
SE     1668  0,6036  0,1891                                    *)
                              ---+---------+---------+---------+------
                               0,30      0,40      0,50      0,60

Pooled StDev = 0,1643
```

O grau de variação entre as Regiões é alto (1188,70) e o p-value nos indica que a informação é confiável e não existe chance deste valor ser diferente.



10.16 – VARIÁVEL T_DES2529 POR REGIÃO



No indicador de **T_DES2529**, podemos verificar que a Região Sul possui o maior índice, mas que as demais regiões apresentam valores próximos (0,80). Vale a pena destacar que para esta variável os valores atípicos em todas as regiões estão nos municípios que apresentam os piores valores

One-way ANOVA: T_DES2529np versus REGIÃO CORRETA

Source	DF	SS	MS	F	P
REGIÃO CORRETA	4	17,32863	4,33216	440,14	0,000

```
Error            5559  54,71560  0,00984
Total            5563  72,04423

S = 0,09921   R-Sq = 24,05%   R-Sq(adj) = 24,00%



Level     N     Mean    StDev
CO      465  0,85268  0,07901
N       449  0,79659  0,11297
NE     1794  0,75948  0,12355
S      1188  0,91048  0,06958
SE     1668  0,83677  0,08827

       Individual 95% CIs For Mean Based on Pooled StDev
Level  ---------+---------+---------+---------+
CO                       (-*)
N             (-*-)
NE     (*)
S                                      (*)
SE                    (*)
       ---------+---------+---------+---------+
              0,800     0,850     0,900     0,950

Pooled StDev = 0,09921
```

O grau de variação entre as Regiões é baixo (440,14) e o p-value nos indica que a informação é confiável e não existe chance deste valor ser diferente.

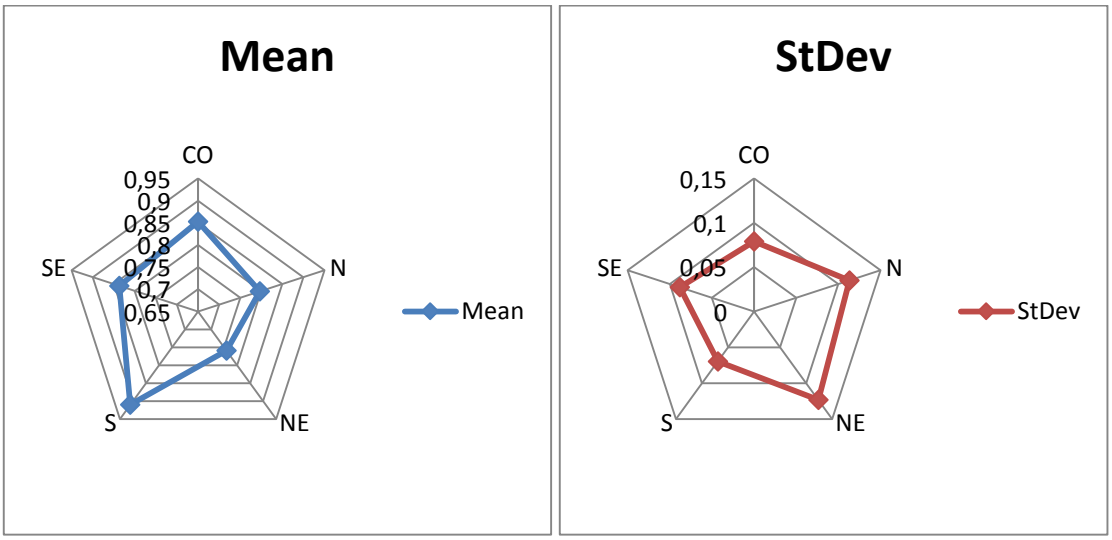
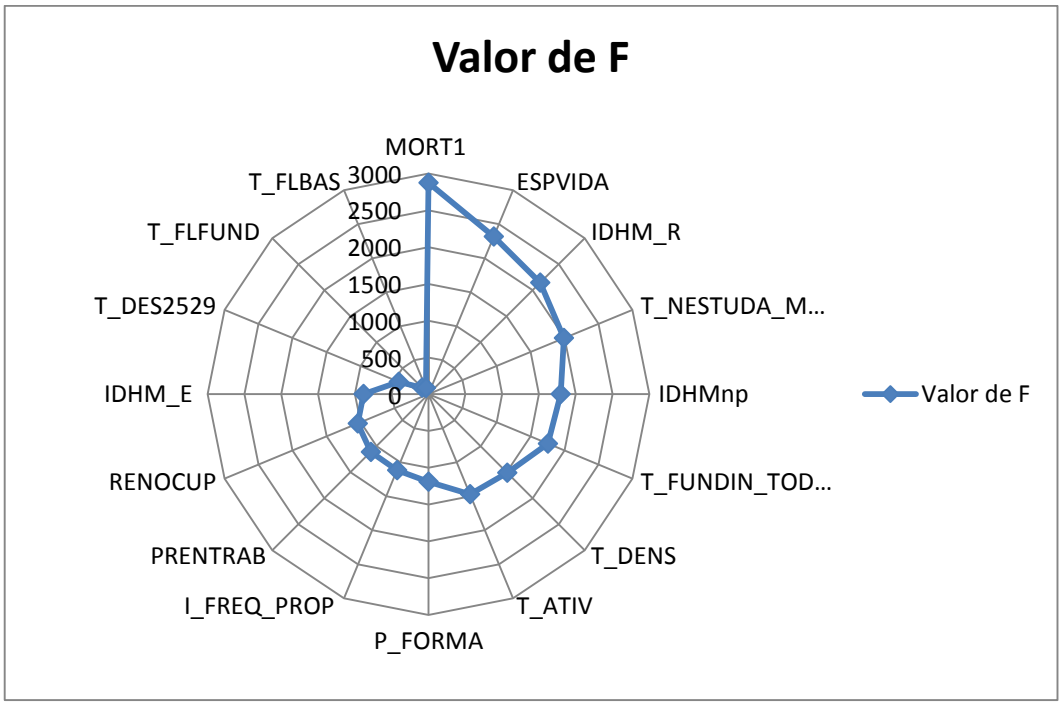


Tabela – Valor de F

Variável	Valor de F
MORT1	2871,93
ESPVIDA	2319,16
IDHM_R	2143,97
T_NESTUDA_MMEIO	1991,1
IDHMnp	1795,58
T_FUNDIN_TODOS_MMEIO	1753,51
T_DENS	1510,67
T_ATIV	1471,83
P_FORMA	1188,7
I_FREQ_PROP	1118,63
PRENTRAB	1111,02
RENOCUP	1039,86
IDHM_E	884,6
T_DES2529	440,14
T_FLFUND	124,16
T_FLBAS	92,37

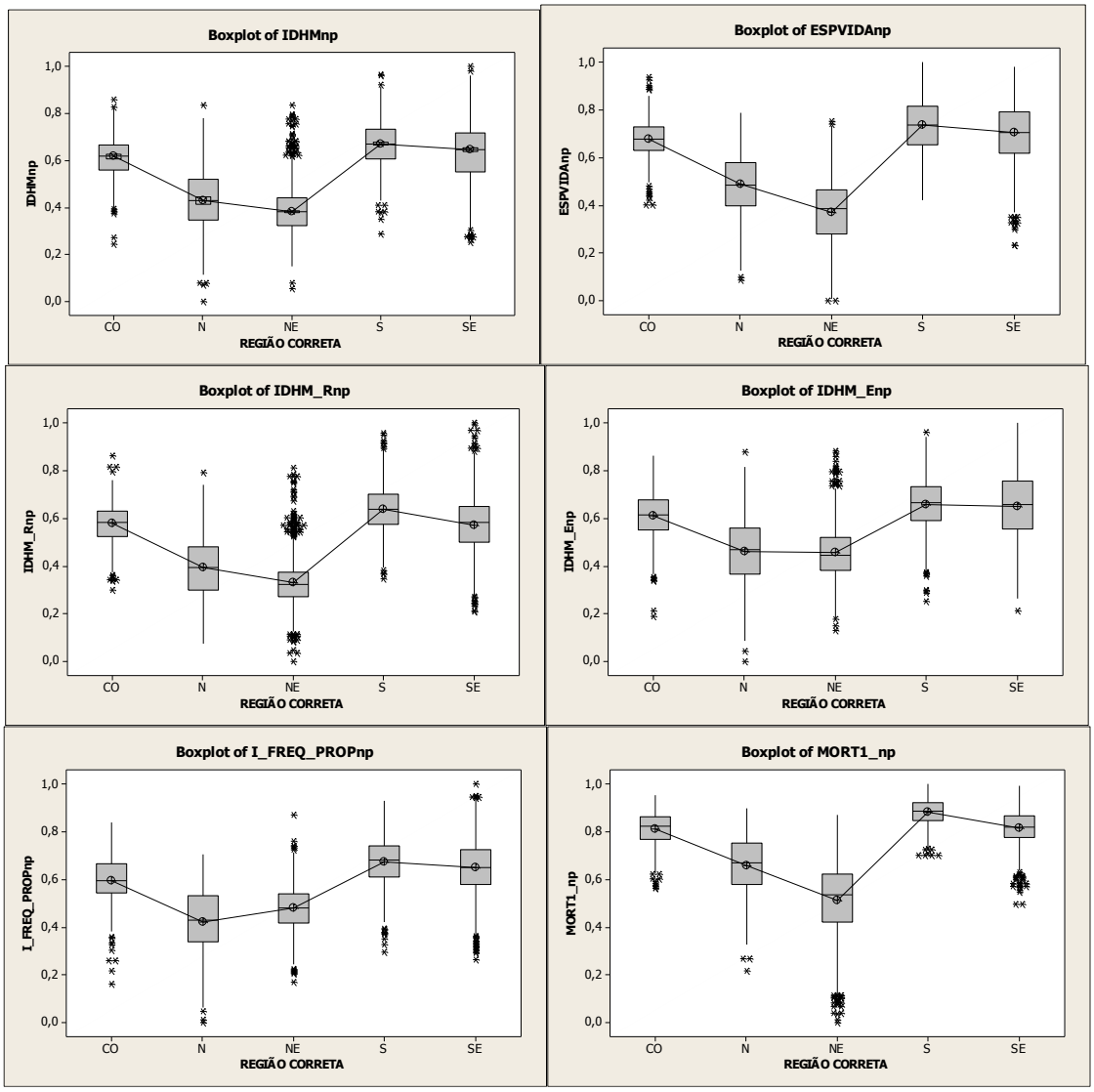
Fonte: dados da pesquisa, 2014.

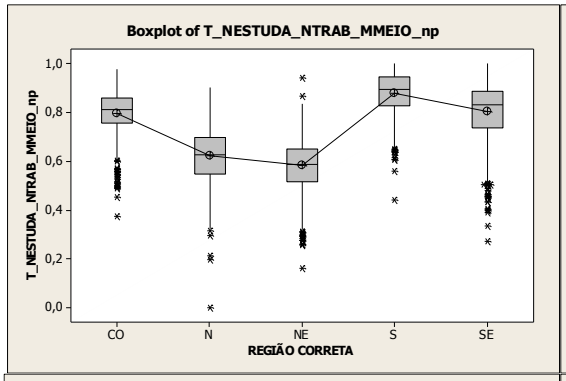
Para uma melhor compreensão da variabilidade nas análises comparativas segue gráfico de radar para as varáveis analisadas na dimensão desenvolvimento humano.

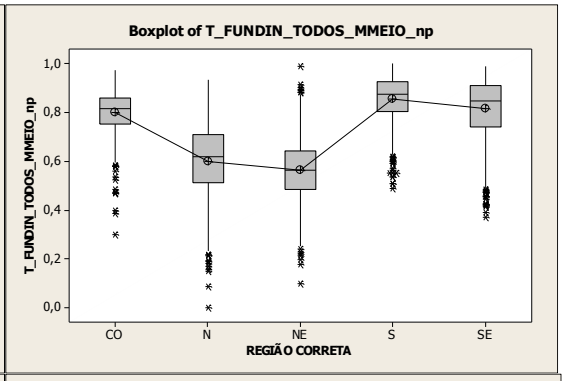


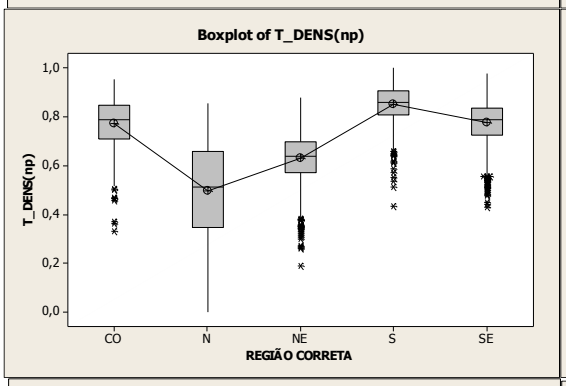
COMENTÁRIOS DA ANÁLISE

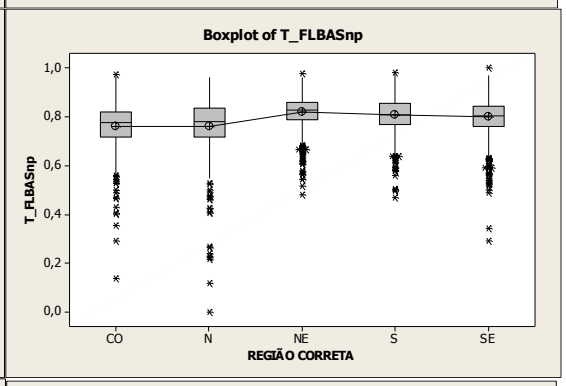
Esta parte do trabalho teve por objetivo comparar as médias dos indicadores das variáveis dos dados apresentados no relatório Atlas Brasil 2013.
Estas comparações indicam que os municípios do Brasil apresentam disparidades quanto as variáveis. Ressalta-se que isto ocorre principalmente com relação as regiões norte e nordeste das demais.
Podemos observar em quase todos os gráficos que existem dois *Brasis*, ou seja, os dados das Regiões Sudeste e Sul, e quase sempre acompanhadas pela região Centro Oeste, são muito próximos e apresentam resultados melhores. Já os dados das Regiões Norte e Nordeste são próximos também, porém apresentam os piores resultados.

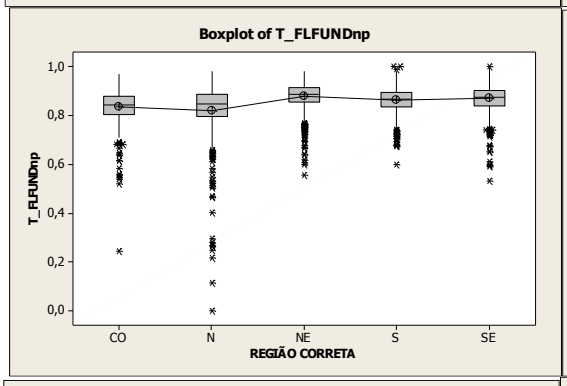


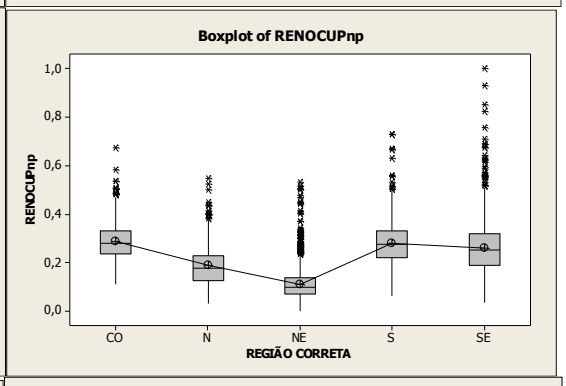

Boxplot of T_NESTUDA_NTRAB_MMEIO_np
T_NESTUDA_NTRAB_MMEIO_np
1,0
0,8
0,6
0,4
0,2
0,0
CO
N
NE
S
SE
REGIÃO CORRETA

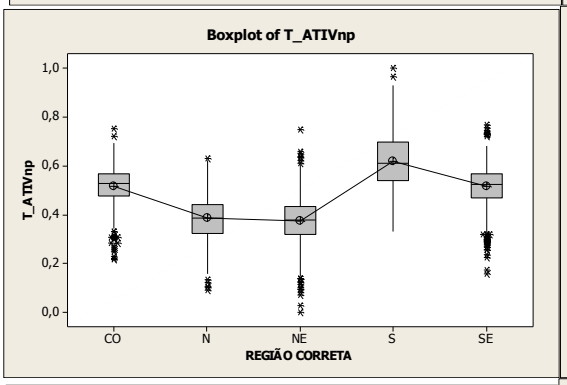

Boxplot of T_FUNDIN_TODOS_MMEIO_np
T_FUNDIN_TODOS_MMEIO_np
1,0
0,8
0,6
0,4
0,2
0,0
CO
N
NE
S
SE
REGIÃO CORRETA

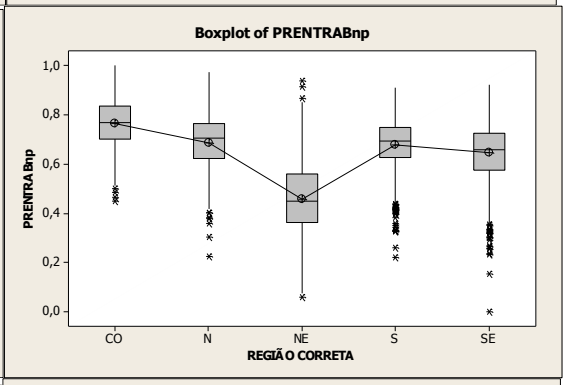

Boxplot of T_DENS(np)
T_DENS(np)
1,0
0,8
0,6
0,4
0,2
0,0
CO
N
NE
S
SE
REGIÃO CORRETA

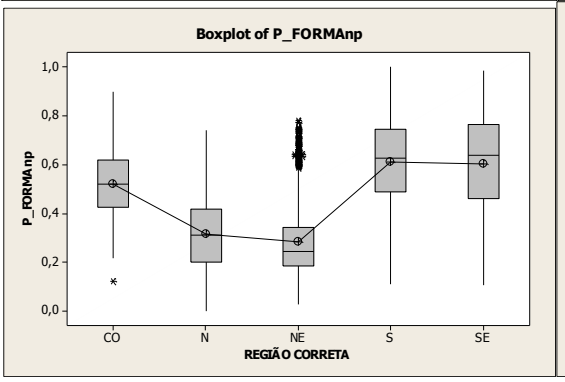

Boxplot of T_FLBASnp
T_FLBASnp
1,0
0,8
0,6
0,4
0,2
0,0
CO
N
NE
S
SE
REGIÃO CORRETA

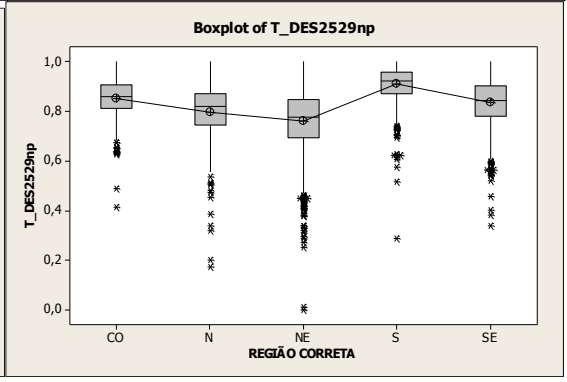

Boxplot of T_FLFUNDnp
T_FLFUNDnp
1,0
0,8
0,6
0,4
0,2
0,0
CO
N
NE
S
SE
REGIÃO CORRETA


Boxplot of RENOCUPnp
RENOCUPnp
1,0
0,8
0,6
0,4
0,2
0,0
CO
N
NE
S
SE
REGIÃO CORRETA


Boxplot of T_ATIVnp
T_ATIVnp
1,0
0,8
0,6
0,4
0,2
0,0
CO
N
NE
S
SE
REGIÃO CORRETA


Boxplot of PRENTRABnp
PRENTRABnp
1,0
0,8
0,6
0,4
0,2
0,0
CO
N
NE
S
SE
REGIÃO CORRETA


Boxplot of P_FORMAnp
P_FORMAnp
1,0
0,8
0,6
0,4
0,2
0,0
CO
N
NE
S
SE
REGIÃO CORRETA


Boxplot of T_DES2529np
T_DES2529np
1,0
0,8
0,6
0,4
0,2
0,0
CO
N
NE
S
SE
REGIÃO CORRETA

Para entendermos o quanto cada região vem se desenvolvendo em termos de Educação, Renda ou Expectativa de Vida é necessário comparar os dados de 1991, 2001 com os de 2010.

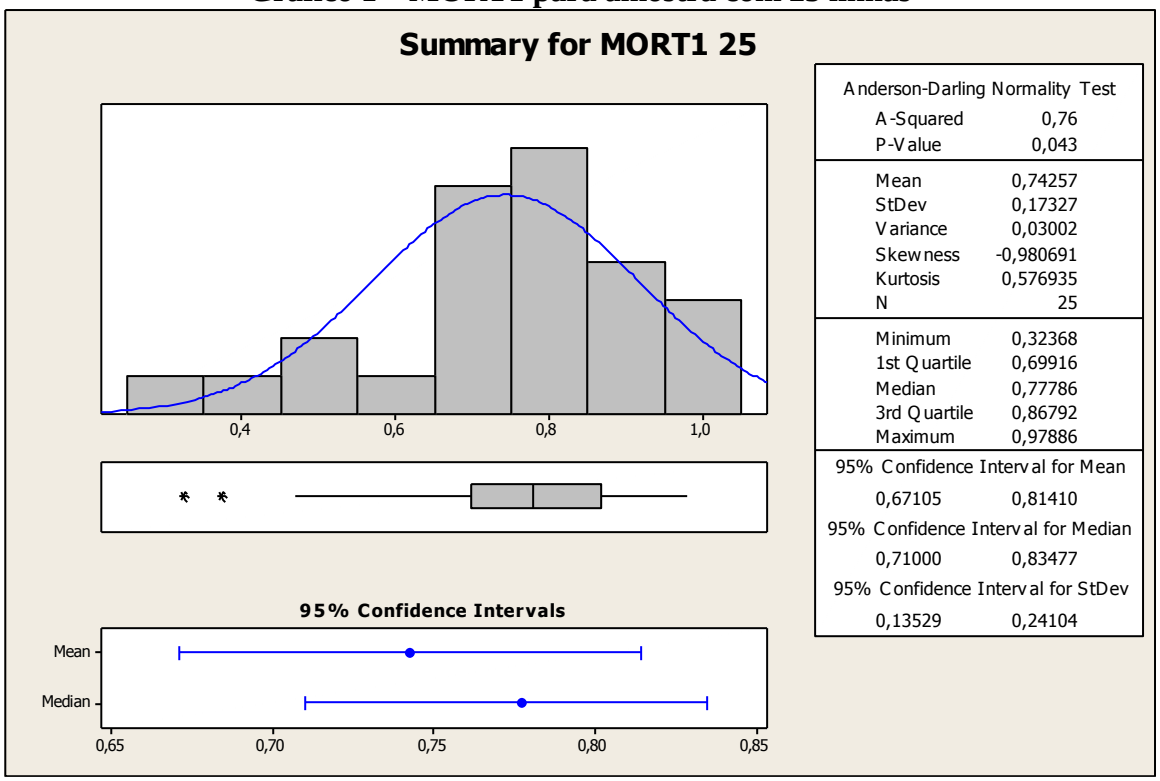
11. PESQUISA POR AMOSTRAGEM

A pesquisa por amostragem foi feita em três amostras, uma de 25, 100 e 400 indivíduos, ou municípios. Para cada amostra foram efetuadas análises do tipo exploratória de dados, as correlações e os dendrogramas. As variáveis utilizadas nesta pesquisa são MORT1, ESPVIDA e IDHM_R. A amostragem aleatória foi feita através do Minitab16, utilizando a função:

Calc >> Random Data >> Sample form columns

11.1 – VARIÁVEL MORT1

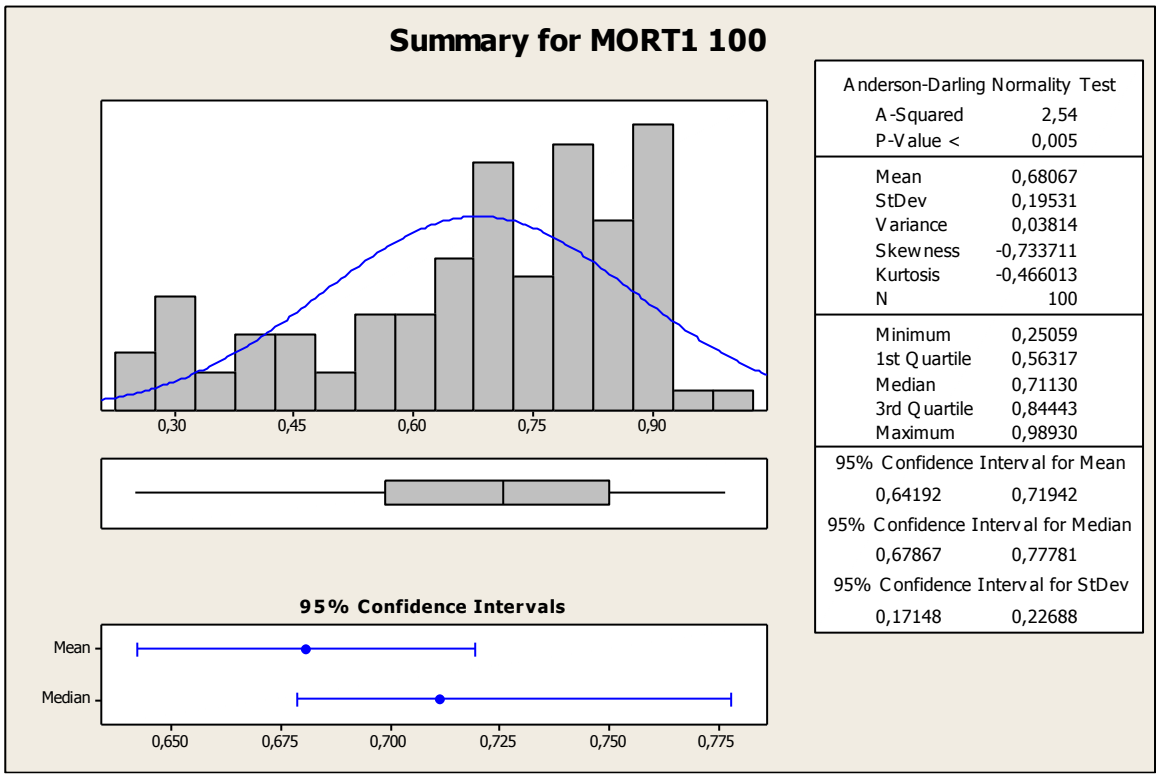
Gráfico 1 – MORT1 para amostra com 25 linhas



Este é o resultado da análise exploratória dos dados de MORT1, para uma população de 25 indivíduos ou municípios.

A curva possui uma calda maior para a esquerda, o que demonstra que os dados possuem maior variabilidade para este lado em relação à média. Podemos verificar que a média dos dados é de 0,74257 e o desvio padrão 0,17327. Existe 95% de confiança de que a média está entre o intervalo de 0,67105 e 0,81410.

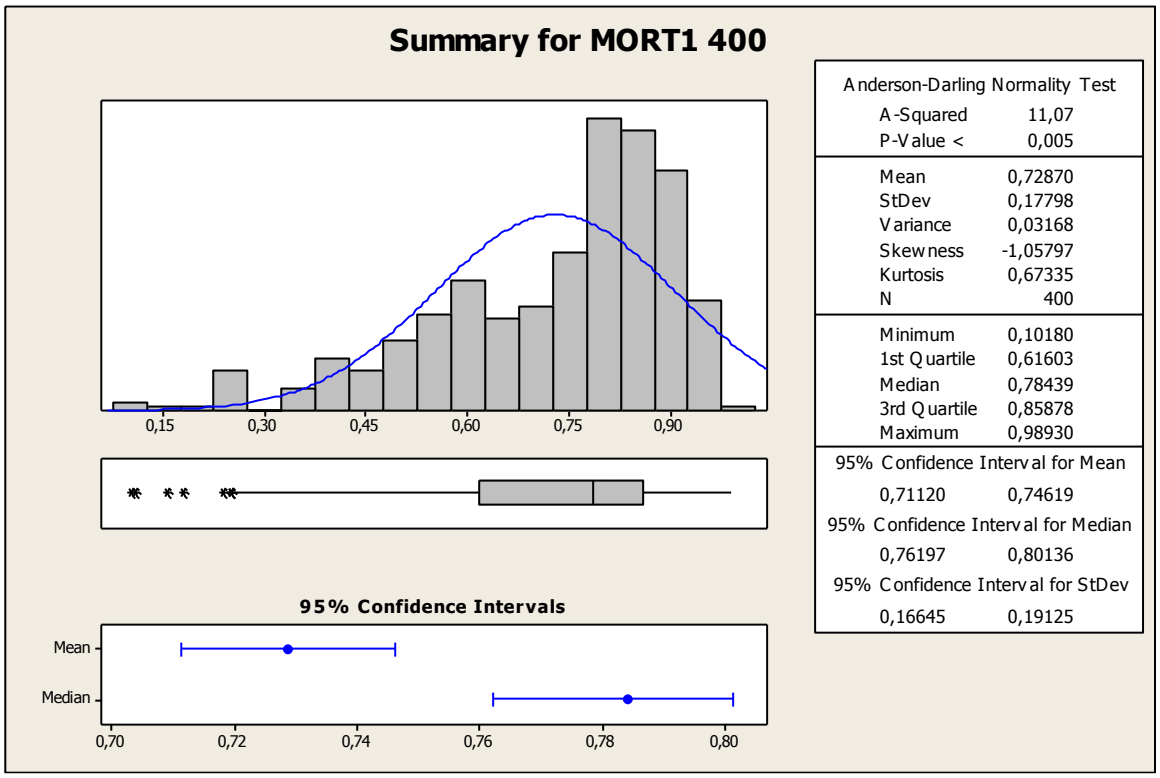
Gráfico 2 – MORT1 para amostra com 100 linhas



Este é o resultado da análise exploratória dos dados de MORT1, para uma população de 100 indivíduos ou municípios.

A curva possui uma leve calda para a esquerda, o que demonstra que os dados possuem maior variabilidade para este lado em relação à média. Podemos verificar que a média dos dados é de 0,68067 e o desvio padrão 0,19531. Existe 95% de confiança de que a média está entre o intervalo de 0,64192 e 0,71942.

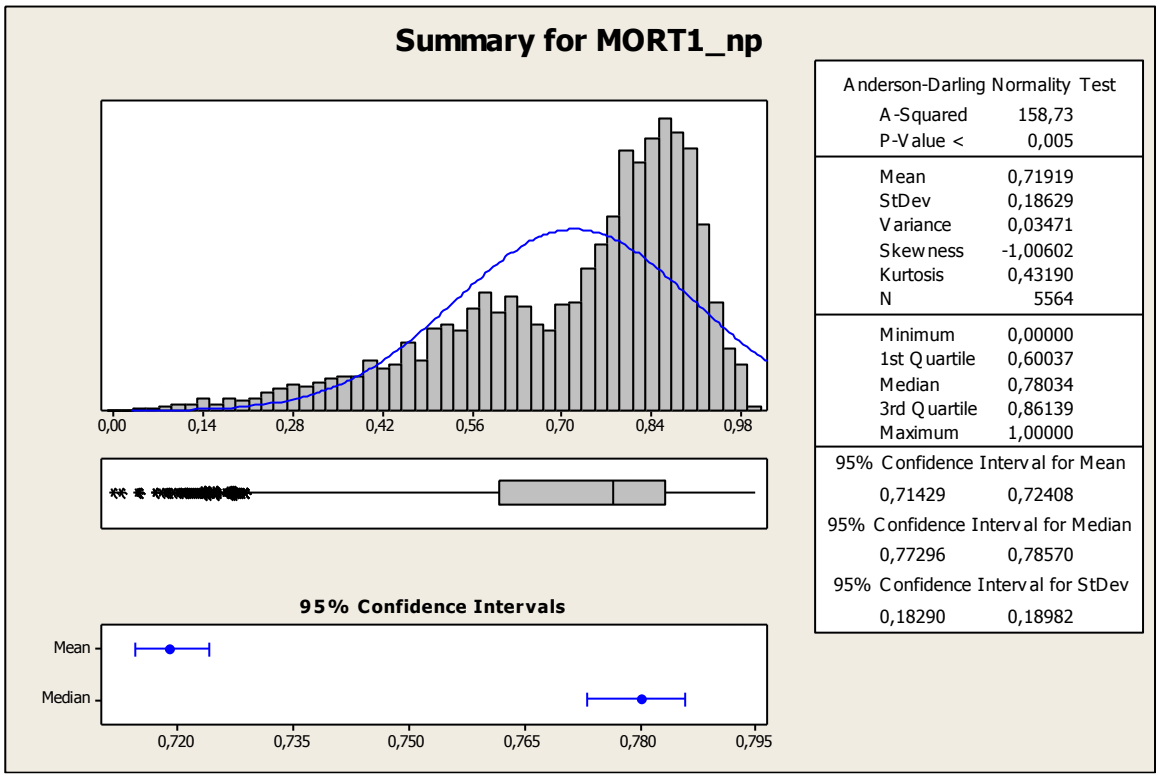
Gráfico 3 – MORT1 para amostra com 400 linhas



Este é o resultado da análise exploratória dos dados de MORT1, para uma população de 400 indivíduos ou municípios.

A curva possui uma calda maior para a esquerda, o que demonstra que os dados possuem maior variabilidade para este lado em relação à média, inclusive apresentando também dados atípicos. Podemos verificar que a média dos dados é de 0,77870 e o desvio padrão 0,17798. Existe 95% de confiança de que a média está entre o intervalo de 0,71120 e 0,74619.

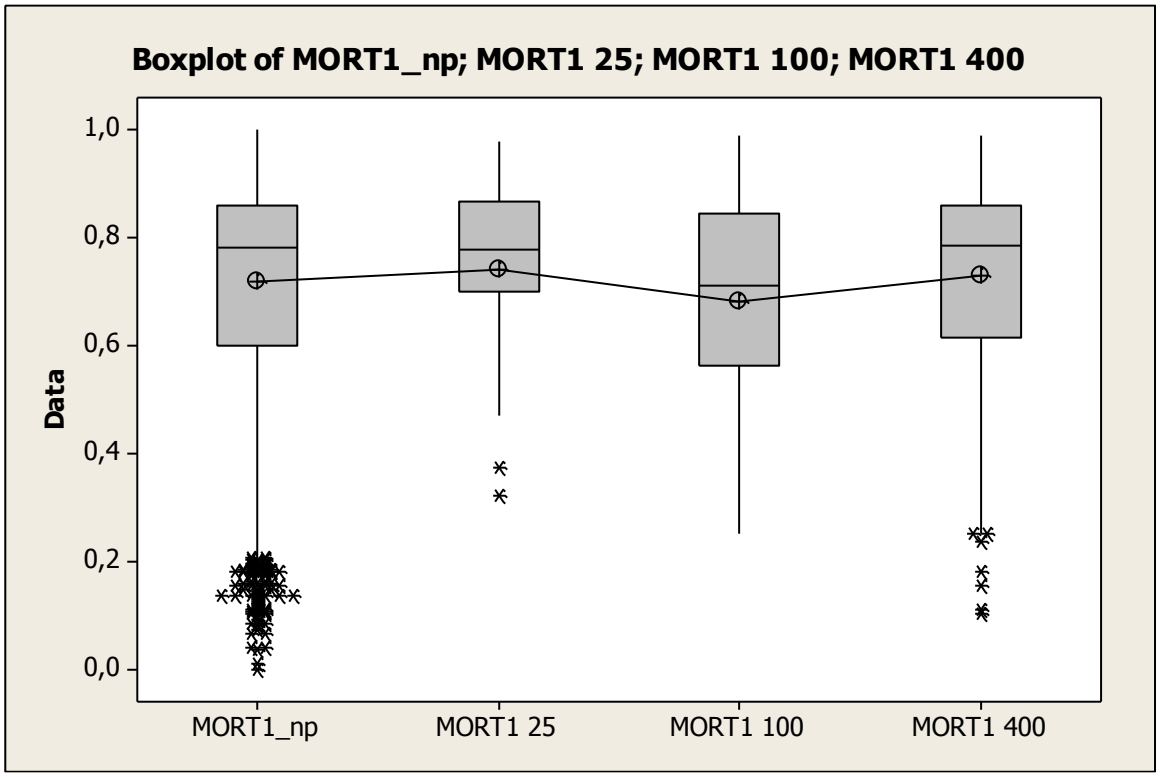
Gráfico 4 – MORT1 para população com 5664 linhas



Podemos observar no gráfico 4, resultado da análise exploratória dos dados de MORT1, para toda a população de 5564 municípios.

A curva possui uma calda maior para a esquerda mesmo utilizando toda a população, o que demonstra que os dados possuem maior variabilidade para este lado em relação à média. Podemos verificar que a média dos dados é de 0,71919 e o desvio padrão 0,18629. Existe 95% de confiança de que a média está entre o intervalo de 0,71429 e 0,72408.

Gráfico 5 – Boxplot de MORT1 para amostras de 25, 100, 400 e população

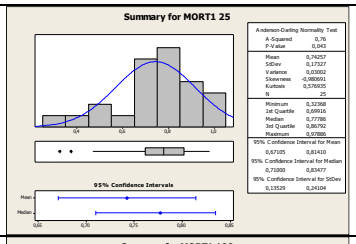
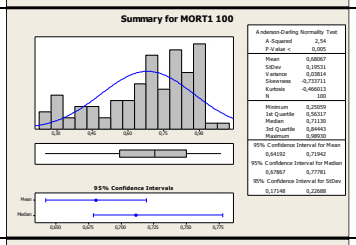
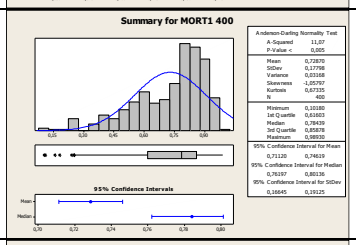
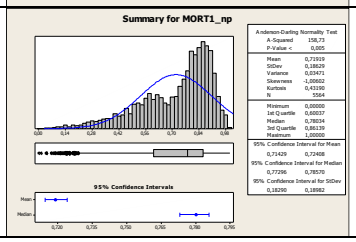


O Boxplot confirma a semelhança e diferenças dos resultados das três amostras comparadas com a população total. Existe uma variabilidade dos dados em todas as análises para esta variável.

One-way ANOVA: MORT1_np; MORT1 25; MORT1 100; MORT1 400

```
Source     DF        SS      MS     F      P
Factor      3    0,1982  0,0661  1,91  0,125
Error    6085  210,2038  0,0345
Total    6088  210,4020

S = 0,1859   R-Sq = 0,09%   R-Sq(adj) = 0,04%


                                  Individual 95% CIs For Mean Based on
                                  Pooled StDev
Level          N    Mean   StDev   -+---------+---------+---------+--------
MORT1_np    5564  0,7192  0,1863                 (*)
MORT1 25      25  0,7426  0,1733        (--------------*-------------)
MORT1 100    100  0,6807  0,1953   (------*------)
MORT1 400    400  0,7287  0,1780                (---*--)
                                   -+---------+---------+---------+--------
                                  0,650     0,700     0,750     0,800

Pooled StDev = 0,1859
```

A análise de variância entre as amostras e a população nos indica que a variação é muito pequena (F=1,91) e o P-Value alto (0,125), ou seja, que existe alguma chance de outros resultados interferirem nestes resultados.

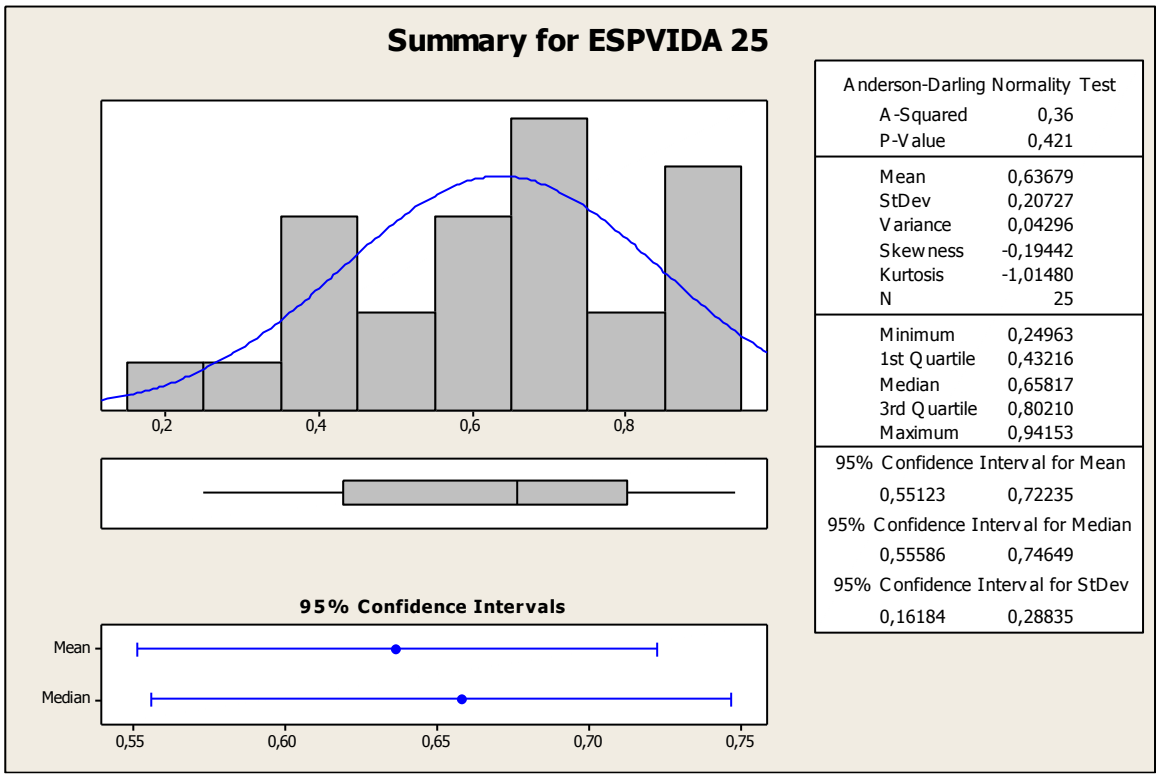
VARIÁVEL	HISTOGRAMA	MEDIANA	MÉDIA	D.PADRÃO	P_VALUE
MORT1 25		0,77786	0,74257	0,17327	0,043
MORT1 100		0,71130	0,68067	0,19531	0,005
MORT1 400		0,78439	0,72870	0,17798	0,005
MORT1		0,78034	0,71919	0,18629	0,005

A tabela acima resume os resultados da análise exploratória dos dados das amostragens e da população da variável MORT1. Os valores de P-values são idênticos para a amostra de 100, 400 indivíduos e a população, ou seja, não existe chance dos resultados acima apresentarem valores diferentes. A mediana obteve uma diferença semelhante entre as amostras com uma aproximação gradual dos valores da população conforme aumentava-se o número de indivíduos, exceto quando da análise de 100 indivíduos. Esta diferença também pôde ser percebida com relação a média das amostras e a média da população.

Portanto, pode-se dizer que na amostra de 400 indivíduos os valores são mais eficientes de se trabalhar, e possuem uma boa precisão em relação aos resultados da população.

11.2 – VARIÁVEL ESPVIDA

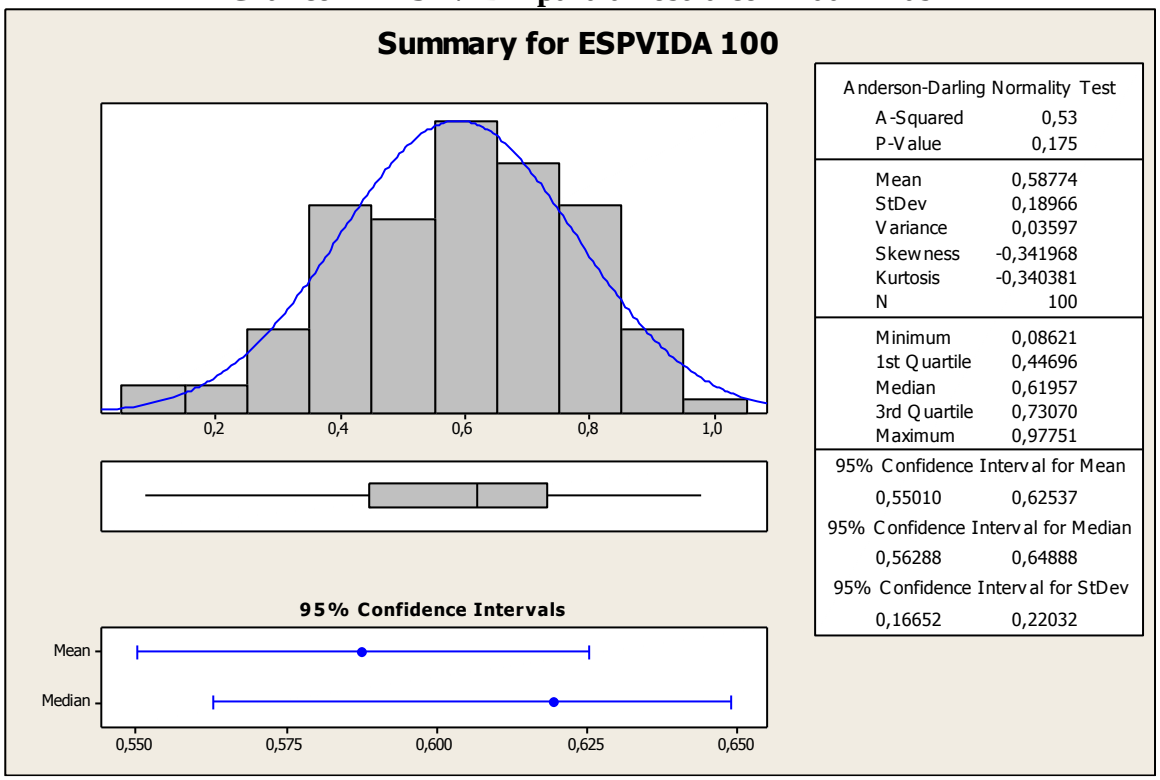
Gráfico 1 – ESPVIDA para amostra com 25 linhas



Este é o resultado da análise exploratória dos dados de ESPVIDA, para uma população de 25 indivíduos ou municípios.

A curva possui uma pequena calda para a esquerda, o que demonstra que os dados possuem maior variabilidade para este lado em relação à média. Podemos verificar que a média dos dados é de 0,63679 e o desvio padrão 0,20727. Existe 95% de confiança de que a média está entre o intervalo de 0,55123 e 0,72235.

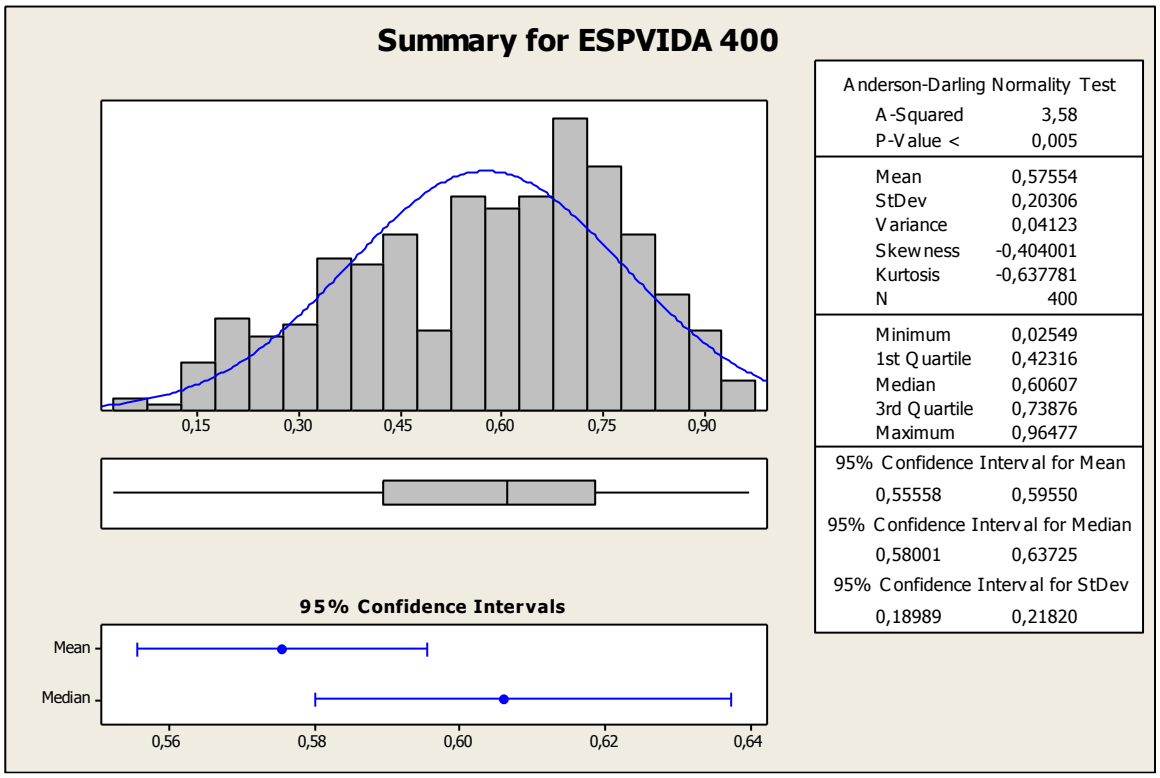
Gráfico 2 – ESPVIDA para amostra com 100 linhas



Este é o resultado da análise exploratória dos dados de ESPVIDA, para uma população de 100 indivíduos ou municípios.

A curva possui uma normalidade na distribuição dos dados em relação à média. Podemos verificar que a média dos dados é de 0,58774 e o desvio padrão 0,18966. Existe 95% de confiança de que a média está entre o intervalo de 0,55010 e 0,62537.

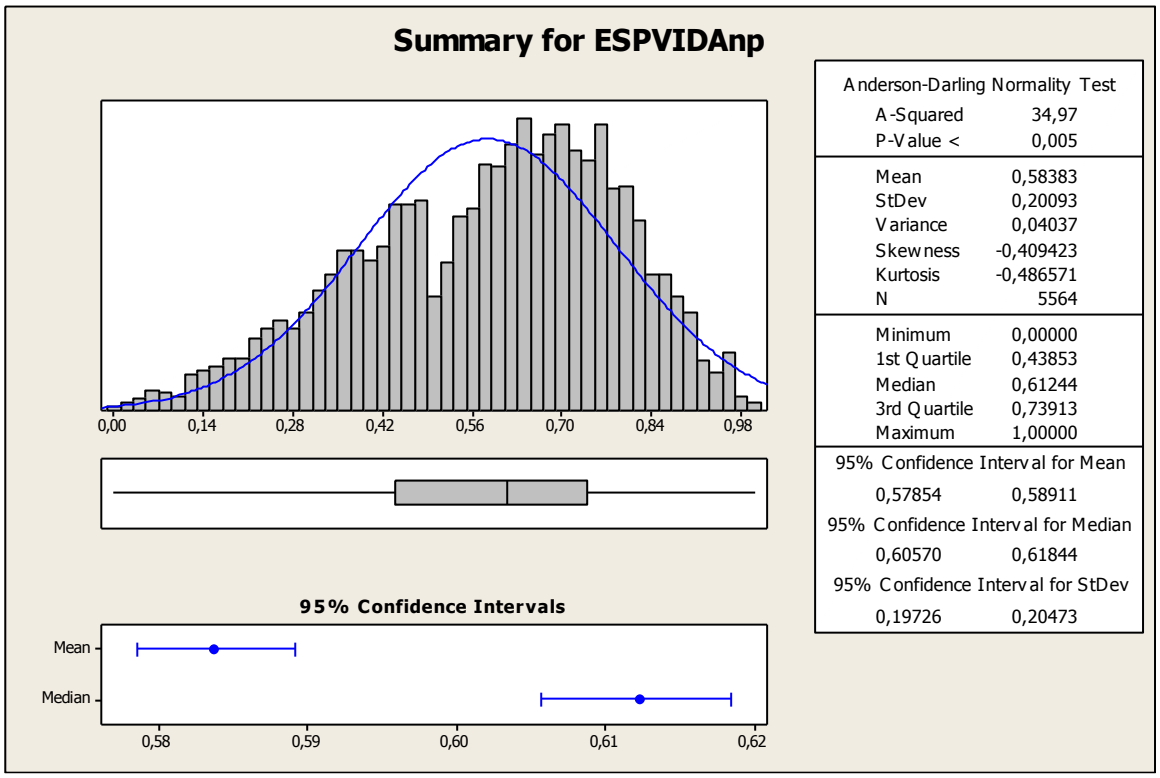
Gráfico 3 – ESPVIDA para amostra com 400 linhas



Este é o resultado da análise exploratória dos dados de ESPVIDA, para uma população de 400 indivíduos ou municípios.

A curva possui uma normalidade na distribuição dos dados em relação à média. Podemos verificar que a média dos dados é de 0,57554 e o desvio padrão 0,20306. Existe 95% de confiança de que a média está entre o intervalo de 0,55558 e 0,59550.

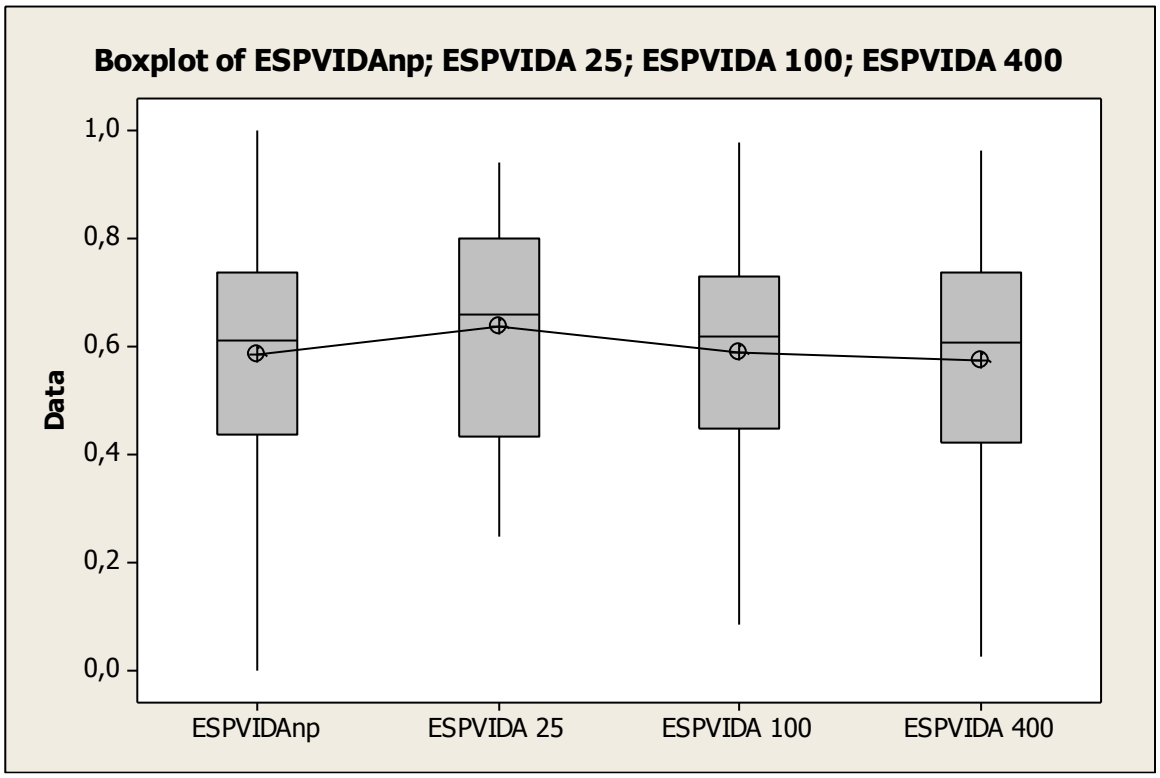
Gráfico 4 – ESPVIDA para população com 5664 linhas



Podemos observar no gráfico 4, resultado da análise exploratória dos dados de ESPVIDA, para toda a população de 5564 municípios.

A curva possui uma calda maior para a esquerda mesmo utilizando toda a população, o que demonstra que os dados possuem maior variabilidade para este lado em relação à média. Podemos verificar que a média dos dados é de 0,58383 e o desvio padrão 0,20093. Existe 95% de confiança de que a média está entre o intervalo de 0,57854 e 0,58911.

Gráfico 5 – Boxplot de ESPVIDA para amostras de 25, 100, 400 e população



O Boxplot confirma a semelhança e diferenças dos resultados das três amostras comparadas com a população total. Existe uma variabilidade dos dados em todas as análises para esta variável.

One-way ANOVA: ESPVIDAnp; ESPVIDA 25; ESPVIDA 100; ESPVIDA 400

Source	DF	SS	MS	F	P
Factor	3	0,0987	0,0329	0,82	0,485
Error	6085	245,6282	0,0404		
Total	6088	245,7269			

S = 0,2009 R-Sq = 0,04% R-Sq(adj) = 0,00%

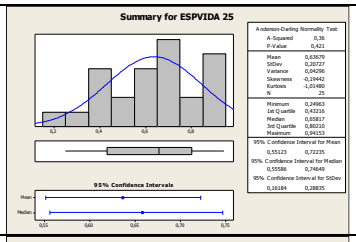
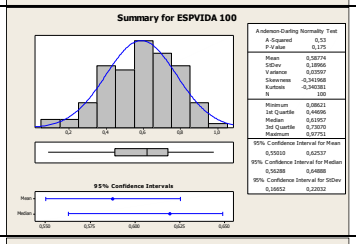
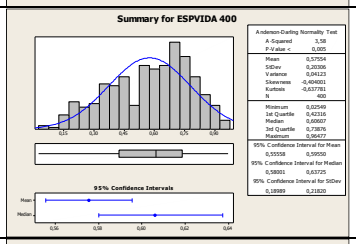
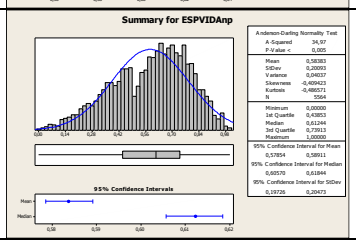
Level	N	Mean	StDev
ESPVIDAnp	5564	0,5838	0,2009
ESPVIDA 25	25	0,6368	0,2073
ESPVIDA 100	100	0,5877	0,1897
ESPVIDA 400	400	0,5755	0,2031

Individual 95% CIs For Mean Based on Pooled StDev

```
Level          +---------+---------+---------+---------
ESPVIDAnp             (*)
ESPVIDA 25       (--------------*---------------)
ESPVIDA 100    (-------*------)
ESPVIDA 400     (---*---)
               +---------+---------+---------+---------
             0,550     0,600     0,650     0,700
```

Pooled StDev = 0,2009

A análise de variância entre as amostras e a população nos indica que a variação é muito pequena (F=0,82) e o P-Value alto (0,485), ou seja que existe alguma chance de outros resultados.

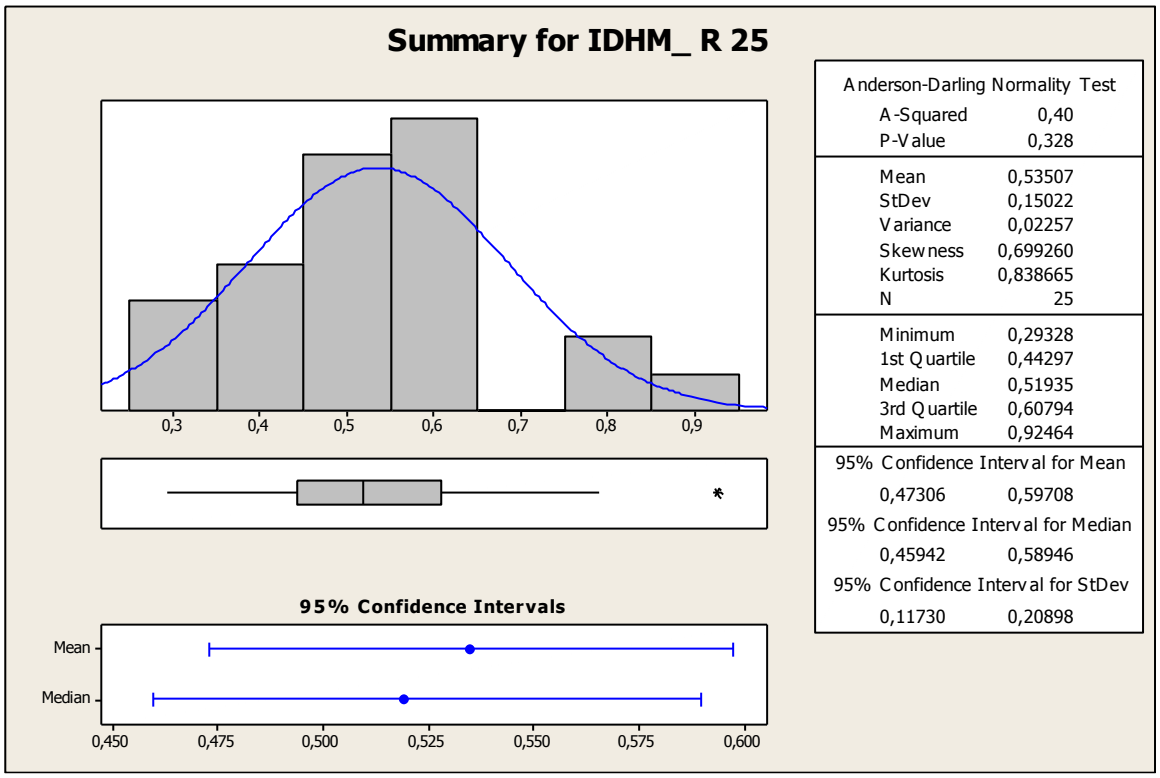
VARIÁVEL	HISTOGRAMA	MEDIANA	MÉDIA	D.PADRÃO	P_VALUE
ESPVIDA 25		0,65817	0,63679	0,20727	0,421
ESPVIDA 100		0,61957	0,58774	0,18966	0,175
ESPVIDA 400		0,60607	0,57554	0,20306	0,005
ESPVIDA		0,61244	0,58383	0,20093	0,005

A tabela acima resume os resultados da análise exploratória dos dados das amostragens e da população da variável ESPVIDA. Os valores de P-values são idênticos para a amostra de 400 indivíduos e a população, ou seja, não existe chance dos resultados acima apresentarem valores diferentes. A mediana obteve uma diferença grande entre as amostras com o número menor de indivíduos como já era esperado. Esta diferença também pôde ser percebida com relação a média das amostras e a média da população, no entanto, nos valores de média a diferença foi menor para as amostras baixas do que nos valores apresentados na mediana.

Portanto, pode-se dizer que na amostra de 400 indivíduos os valores são mais eficientes de se trabalhar, e possuem uma boa precisão em relação aos resultados da população.

11.3 – VARIÁVEL IDHM_R

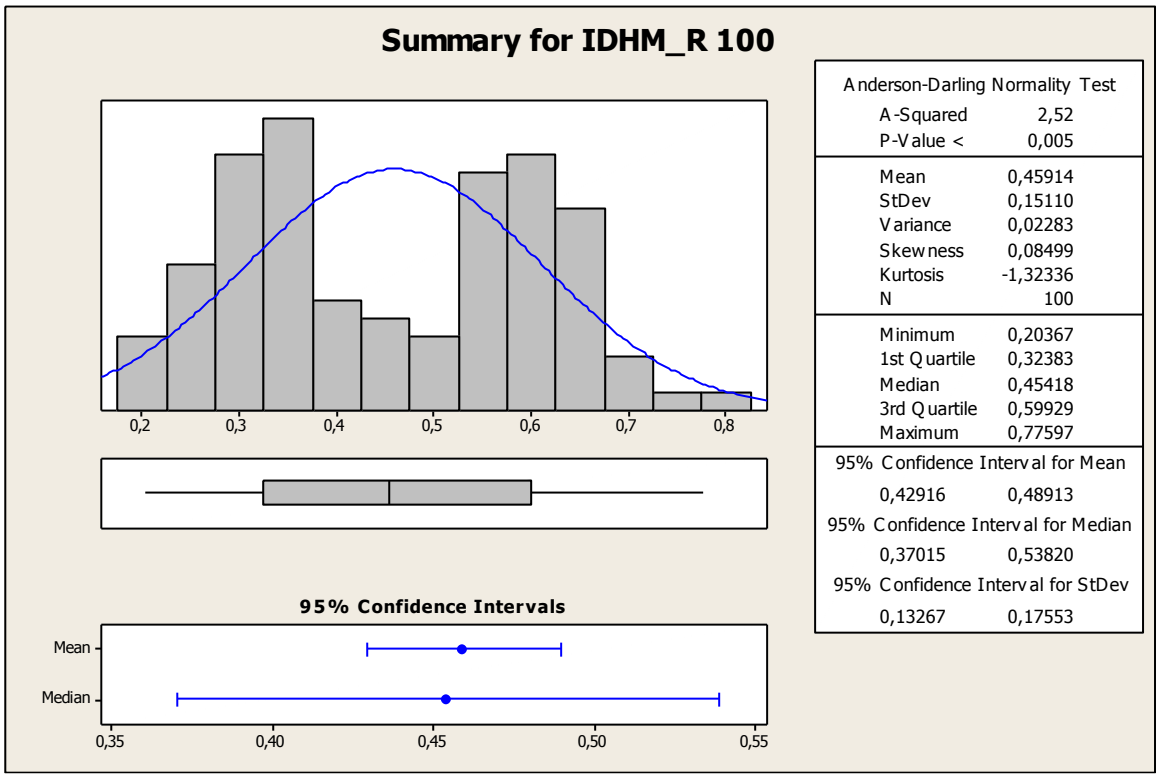
Gráfico 1 – IDHM_R para amostra com 25 linhas



Este é o resultado da análise exploratória dos dados de IDHM_R, para uma população de 25 indivíduos ou municípios.

A curva possui uma calda maior para a direira, o que demonstra que os dados possuem maior variabilidade para este lado em relação à média. Podemos verificar que a média dos dados é de 0,53507 e o desvio padrão 0,15022. Existe 95% de confiança de que a média está entre o intervalo de 0,47306 e 0,59708.

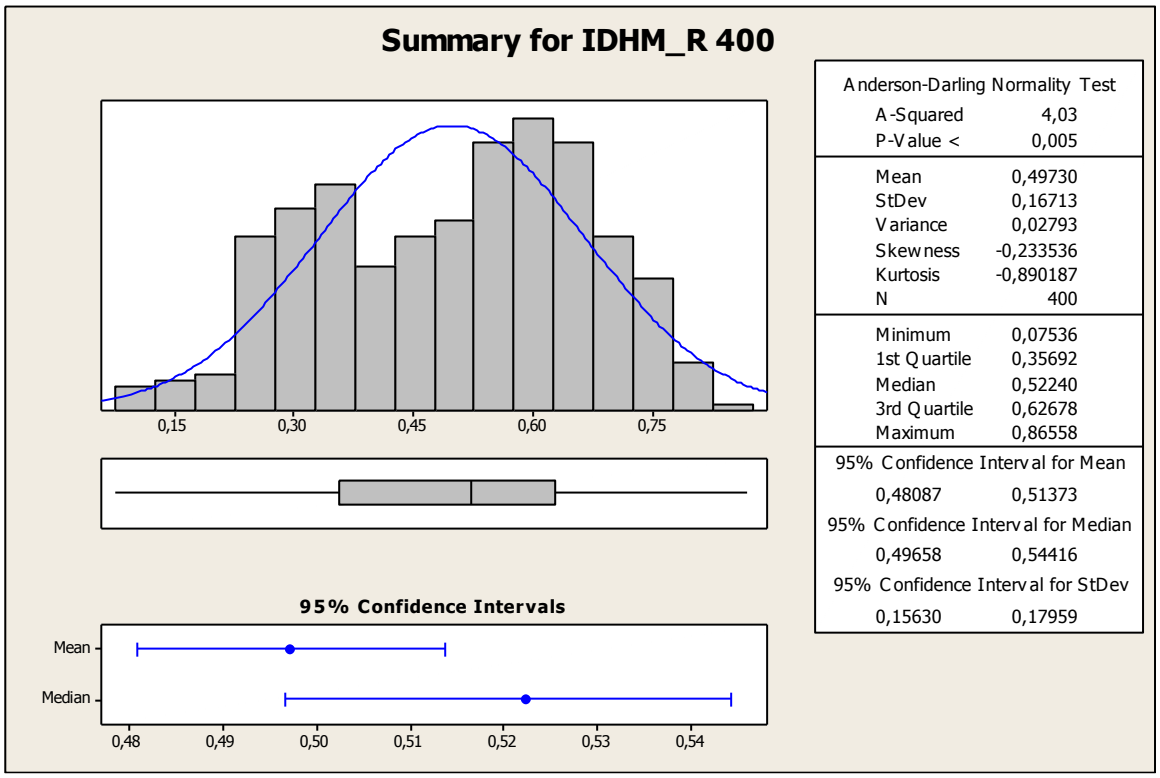
Gráfico 2 – IDHM_R para amostra com 100 linhas



Este é o resultado da análise exploratória dos dados de IDHM_R, para uma população de 100 indivíduos ou municípios.

A curva possui uma pequena calda para a direita, o que demonstra que os dados possuem maior variabilidade para este lado em relação à média. Podemos verificar que a média dos dados é de 0,45914 e o desvio padrão 0,15110. Existe 95% de confiança de que a média está entre o intervalo de 0,42916 e 0,48913.

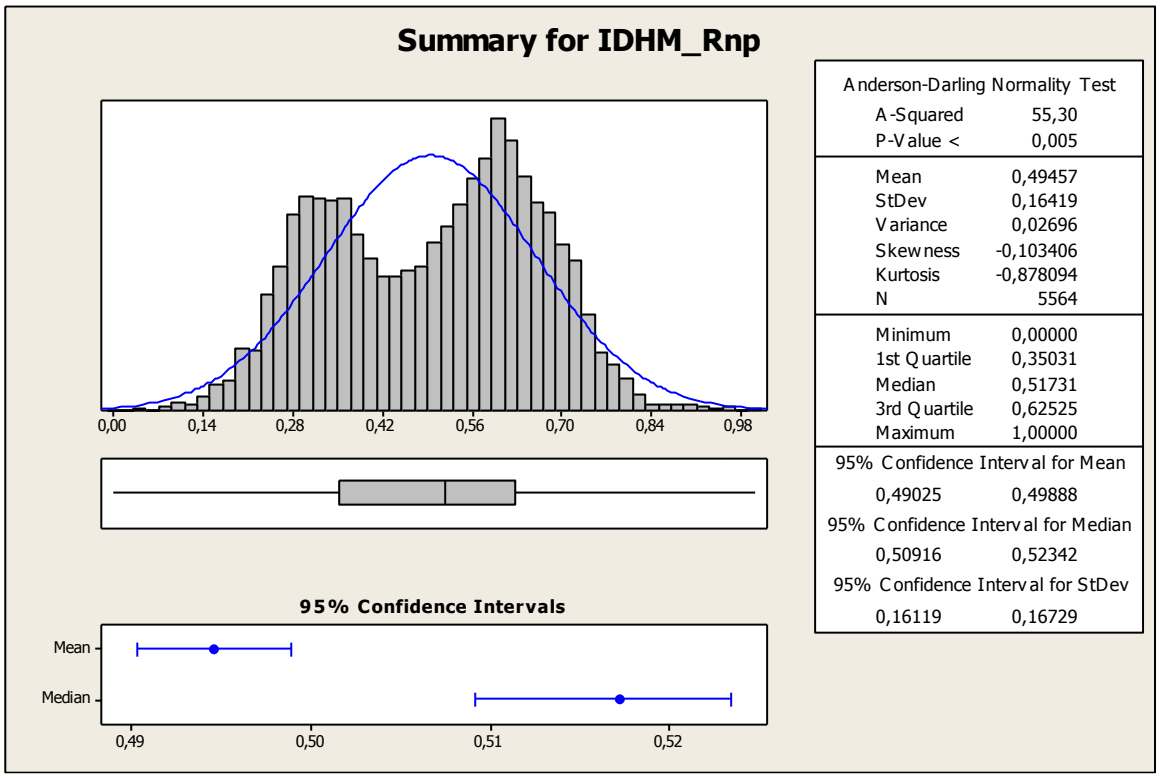
Gráfico 3 – IDHM_R para amostra com 400 linhas



Este é o resultado da análise exploratória dos dados de IDHM_R, para uma população de 400 indivíduos ou municípios.

A curva possui certa normalidade na distribuição dos dados em relação à média. Podemos verificar que a média dos dados é de 0,49730 e o desvio padrão 0,16713. Existe 95% de confiança de que a média está entre o intervalo de 0,48087 e 0,51373.

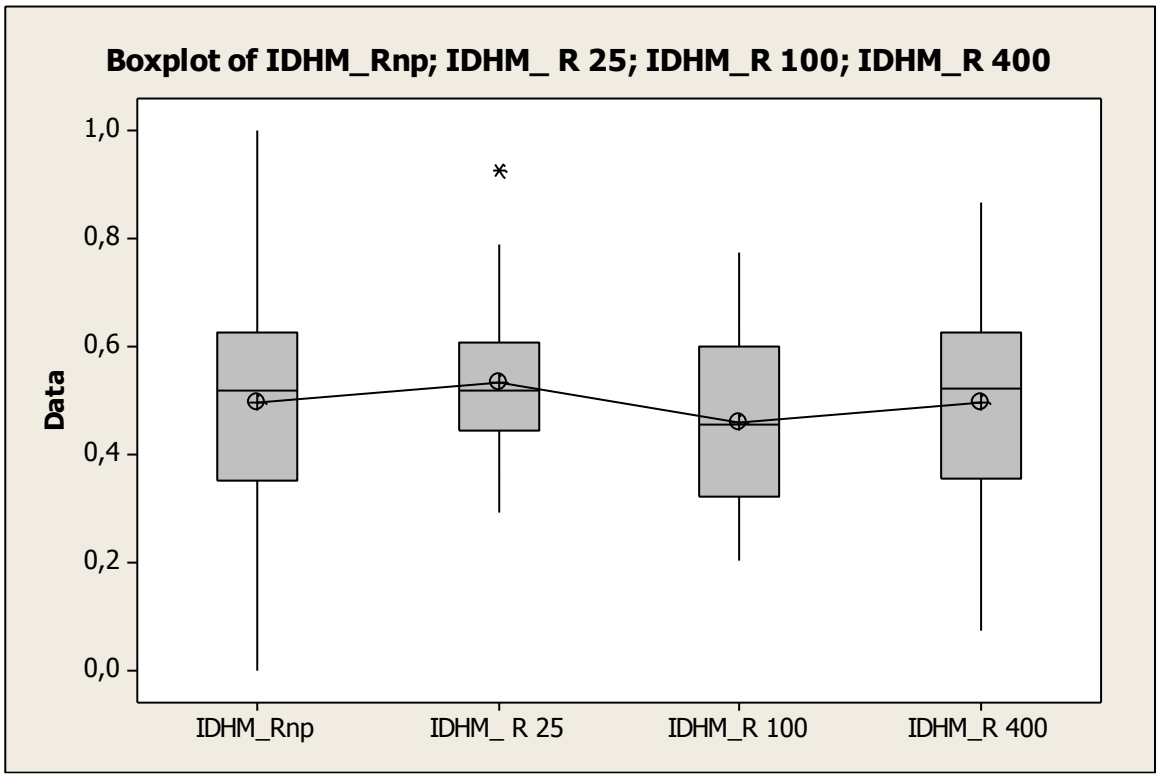
Gráfico 4 – IDHM_R para população com 5664 linhas



Podemos observar no gráfico 4, resultado da análise exploratória dos dados de IDHM_R, para toda a população de 5564 municípios.

A curva possui certa normalidade na distribuição dos dados em relação à média. Podemos verificar que a média dos dados é de 0,49457 e o desvio padrão 0,16419. Existe 95% de confiança de que a média está entre o intervalo de 0,49025 e 0,49888.

Gráfico 5 – Boxplot de IDHM_R para amostras de 25, 100, 400 e população

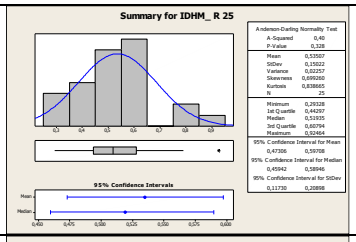
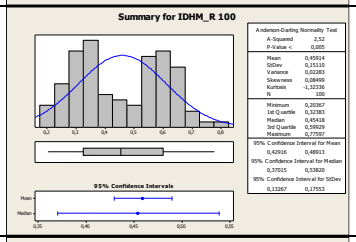
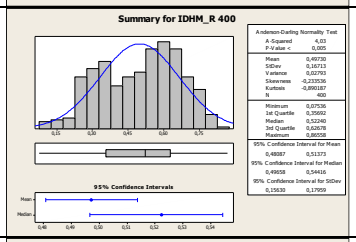
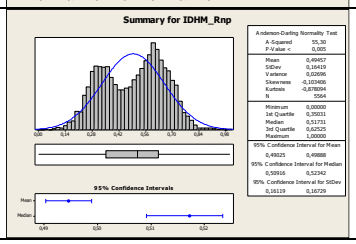


O Boxplot confirma a semelhança e diferenças dos resultados das três amostras comparadas com a população total. Existe uma variabilidade dos dados em todas as análises para esta variável.

One-way ANOVA: IDHM_Rnp; IDHM_ R 25; IDHM_R 100; IDHM_R 400

```
Source     DF        SS      MS     F      P
Factor      3    0,1691  0,0564  2,09  0,099
Error    6085  163,9076  0,0269
Total    6088  164,0768

S = 0,1641   R-Sq = 0,10%   R-Sq(adj) = 0,05%


                                   Individual 95% CIs For Mean Based on
                                   Pooled StDev
Level          N    Mean   StDev  -----+---------+---------+---------+----
IDHM_Rnp    5564  0,4946  0,1642              (*)
IDHM_ R 25    25  0,5351  0,1502          (------------*------------)
IDHM_R 100   100  0,4591  0,1511  (------*-----)
IDHM_R 400   400  0,4973  0,1671            (--*---)
                                   -----+---------+---------+---------+----
                                      0,450     0,500     0,550     0,600

Pooled StDev = 0,1641
```

A análise de variância entre as amostras e a população nos indica que a variação é muito pequena (F=2,09) e o P-Value alto (0,099), ou seja que existe alguma chance de outros resultados influenciarem no resultado desta análise.

VARIÁVEL	HISTOGRAMA	MEDIANA	MÉDIA	D.PADRÃO	P_VALUE
IDHM_R 25		0,51935	0,53507	0,15022	0,328
IDHM_R 100		0,45418	0,45914	0,15110	0,005
IDHM_R 400		0,52240	0,49730	0,16713	0,005
IDHM_R		0,51731	0,49457	0,16419	0,005

A tabela acima resume os resultados da análise exploratória dos dados das amostragens e da população da variável IDHM_R. Os valores de P-values são idênticos para a amostra de 100, 400 indivíduos e a população, ou seja, não existe chance dos resultados acima apresentarem valores diferentes. Vale destacar que a análise com 25 indivíduos apresentou resultados semelhantes aos da população, mas com um P-value de 0,328. A mediana obteve uma variação gradual entre as amostras. Esta diferença também pôde ser percebida com relação a média das amostras com relação a população.

Portanto, pode-se dizer que na amostra de 400 indivíduos os valores são mais eficientes de se trabalhar, e possuem uma boa precisão em relação aos resultados da população.

12. CORRELAÇÃO LINEAR

Segue abaixo tabela descritiva dos dados e a matriz de correlação incluindo o teste de significância p-value. Para a correlação foi utilizado o índice de Pearson. Vale ressaltar que o índice de correlação entre as variáveis não requer que exista uma relação de causa-efeito entre ambas.

A Tabela de Dados[3]

Descriptive Statistics: T_NESTUDA_NT; MORT1_np; T_FUNDIN_TOD; T_DENS(np); ...

Variable	N	N*	Mean	Minimum	Median	Maximum
T_NESTUDA_NTRAB_MMEIO_np	5564	0	0,73254	0,00000	0,75258	1,00000
MORT1_np	5564	0	0,71919	0,00000	0,78034	1,00000
T_FUNDIN_TODOS_MMEIO_np	5564	0	0,72383	0,00000	0,75027	1,00000
T_DENS(np)	5564	0	0,72182	0,00000	0,74526	1,00000
ESPVIDAnp	5564	0	0,58383	0,00000	0,61244	1,00000
I_FREQ_PROPnp	5564	0	0,57684	0,00000	0,57925	1,00000
IDHMnp	5564	0	0,54308	0,00000	0,55631	1,00000
IDHM_Enp	5564	0	0,56968	0,00000	0,57120	1,00000
IDHM_Rnp	5564	0	0,49457	0,00000	0,51731	1,00000
T_FLBASnp	5564	0	0,80070	0,00000	0,80948	1,00000
T_FLFUNDnp	5564	0	0,86454	0,000000	0,87260	1,00000
RENOCUPnp	5564	0	0,21158	0,00000	0,20561	1,00000
PRENTRABnp	5564	0	0,60539	0,00000	0,63626	1,00000
P_FORMAnp	5564	0	0,47052	0,00000	0,46291	1,00000
T_ATIVnp	5564	0	0,48060	0,00000	0,48397	1,00000
T_DES2529np	5564	0	0,82568	0,00000	0,84216	1,00000

12.1 CORRELAÇÃO DAS VARIÁVEIS

Os dados abaixo representam a correlação entre as variáveis selecionadas e já trabalhadas anteriormente.

Correlations: ESPVIDAn; T_FUND11A13n; T_FUND15A17n; T_FUND18Mn; ...

	ESPVIDAn	T_FUND11A13n	T_FUND15A17n	T_FUND18Mn
T_FUND11A13n	0,517			
	0,000			
T_FUND15A17n	0,666	0,726		
	0,000	0,000		
T_FUND18Mn	0,632	0,446	0,601	
	0,000	0,000	0,000	
T_MED18A20n	0,660	0,651	0,833	0,656
	0,000	0,000	0,000	0,000
RDPCn	0,784	0,525	0,671	0,757
	0,000	0,000	0,000	0,000
I_FREQ_PROPn	0,641	0,812	0,927	0,633
	0,000	0,000	0,000	0,000
IDHMn	0,852	0,682	0,832	0,857

[3] Para as análises foram normalizados todos os dados, sendo que o valor que se aplica é: "quanto mais próximo de 1 melhor".

	0,000	0,000	0,000	0,000
IDHM_En	0,704	0,705	0,855	0,892
	0,000	0,000	0,000	0,000
IDHM_Ln	1,000	0,517	0,666	0,632
	0,000	0,000	0,000	0,000
IDHM_Rn	0,834	0,586	0,721	0,757
	0,000	0,000	0,000	0,000
T_FREQ5A6n	0,016	0,206	0,197	0,187
	0,229	0,000	0,000	0,000

	T_MED18A20n	RDPCn	I_FREQ_PROPn	IDHMn
RDPCn	0,717			
	0,000			
I_FREQ_PROPn	0,917	0,686		
	0,000	0,000		
IDHMn	0,851	0,908	0,862	
	0,000	0,000	0,000	
IDHM_En	0,877	0,791	0,913	0,951
	0,000	0,000	0,000	0,000
IDHM_Ln	0,660	0,784	0,641	0,852
	0,000	0,000	0,000	0,000
IDHM_Rn	0,748	0,962	0,729	0,948
	0,000	0,000	0,000	0,000
T_FREQ5A6n	0,250	0,108	0,434	0,224
	0,000	0,000	0,000	0,000

	IDHM_En	IDHM_Ln	IDHM_Rn
IDHM_Ln	0,704		
	0,000		
IDHM_Rn	0,820	0,834	
	0,000	0,000	
T_FREQ5A6n	0,351	0,016	0,095
	0,000	0,228	0,000

Cell Contents: Pearson correlation
P-Value

A correlação é sempre um número entre zero e um e mede a intensidade de relações lineares. A correlação entre as variáveis analisadas é positiva na maior parte dos dados, mas de fraca intensidade. Contudo, com base nas informações constantes na matriz de correlação pode-se perceber que as variáveis descritas na Tabela 3 apresentam forte relação, vale destacar que isto não significa que elas apresentam causalidade, ou seja, um sentido direto entre elas.

TABELA 3 – CORRELAÇÃO DAS VARIÁVEIS

Variável	Variável	Grau de Correlação
ESPVIDAnp	MORT1_np	0,967
IDHM_Enp	IDHMnp	0,951

IDHM_Rnp	IDHMnp	0,948
IDHM_Enp	T_FUNDIN_TODOS_M	0,921
RENOCUPnp	IDHM_Rnp	0,915
ESPVIDAnp	T_FUNDIN_TODOS_M	0,913
IDHM_Rnp	T_FUNDIN_TODOS_M	0,906
T_ATIVnp	T_NESTUDA_NTRAB_	0,877
RENOCUPnp	IDHMnp	0,869
I_FREQ_PROPnp	T_FUNDIN_TODOS_M	0,853
IDHMnp	ESPVIDAnp	0,852
T_FUNDIN_TODOS_M	T_NESTUDA_NTRAB	0,835
IDHM_Rnp	ESPVIDAnp	0,834
P_FORMAnp	IDHMnp	0,824
IDHM_Rnp	IDHM_Enp	0,819
P_FORMAnp	RENOCUPnp	0,802

Fonte: elaborado pelo autor, 2014.

12.2 DENDOGRAMA

Um **Dendrograma** (dendr(o) = árvore) é um tipo específico de diagrama ou representação icônica que organiza determinados fatores e variáveis. Isto quer dizer que sua representação apresenta um diagrama de similaridade.

A interpretação de um Dendrograma de similaridade entre amostras fundamenta-se na intuição: duas amostras próximas devem ter também valores semelhantes para as variáveis medidas. Ou seja, elas devem ser próximas matematicamente no espaço multidimensional.

Portanto, quanto maior a proximidade entre as medidas relativas às amostras, maior a similaridade entre elas. O dendrograma hierarquiza esta similaridade de modo que podemos ter uma visão bidimensional da similaridade ou dissimilaridade de todo o conjunto de amostras utilizado no estudo.

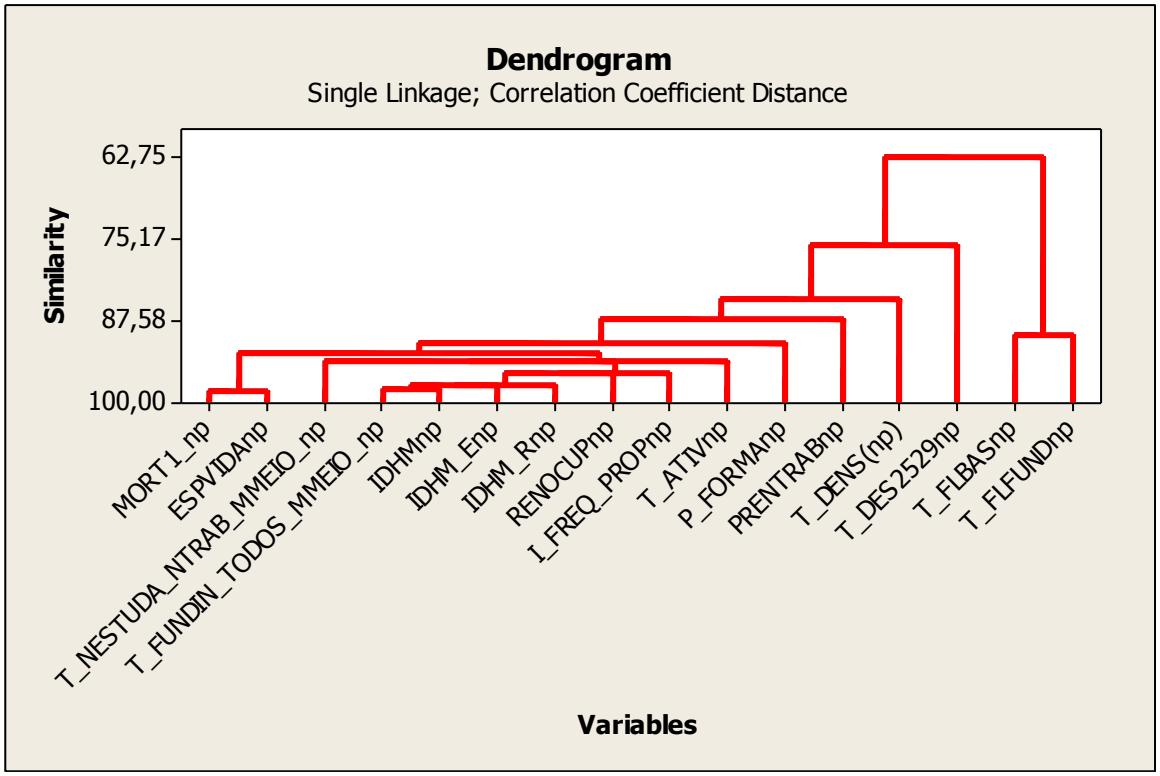
Cluster Analysis of Variables: MORT1_np; T_NESTUDA_NT; T_FUNDIN_TOD; ...

Correlation Coefficient Distance, Single Linkage
Amalgamation Steps

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	15	98,3572	0,032857	1	5	1	2
2	14	98,0777	0,038445	3	7	3	2
3	13	97,5422	0,049155	3	8	3	3
4	12	97,4140	0,051720	3	9	3	4
5	11	95,7702	0,084596	3	12	3	5
6	10	95,6766	0,086467	3	6	3	6
7	9	93,8677	0,122646	2	3	2	7
8	8	93,8673	0,122655	2	15	2	8
9	7	92,6019	0,147962	1	2	1	10
10	6	91,1857	0,176286	1	14	1	11
11	5	89,9867	0,200266	10	11	10	2
12	4	87,6672	0,246655	1	13	1	12
13	3	84,4718	0,310564	1	4	1	13
14	2	76,3215	0,473569	1	16	1	14
15	1	62,7547	0,744907	1	10	1	16

Segue abaixo o Dendrograma das variáveis analisadas:

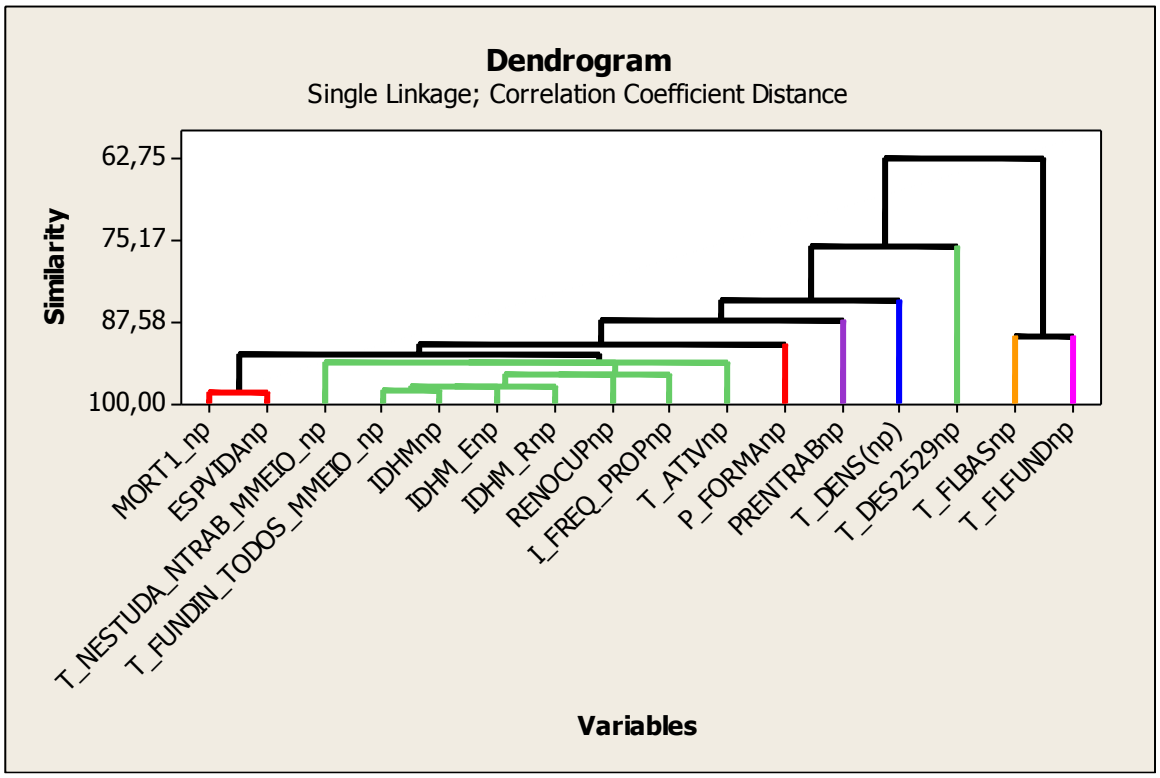
Gráfico - Dendrograma das variáveis



Podemos concluir pelo Dendrograma que existem vários grupos de variáveis semelhantes.

STAT >> MULTIVARIATE >> CLUSTER VARIABLE (number of cluster = 1)

Figura 3 – Dendograma dos agrupamentos das variáveis por similaridade



Podemos concluir pelo Dendrograma que existem oito grupos de variáveis semelhantes.

STAT >> MULTIVARIATE >> CLUSTER VARIABLE (number of cluster = 8)

Cluster Analysis of Variables: MORT1_np; T_NESTUDA_NT; T_FUNDIN_TOD; ...

Correlation Coefficient Distance, Single Linkage
Amalgamation Steps

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	15	98,3572	0,032857	1	5	1	2
2	14	98,0777	0,038445	3	7	3	2
3	13	97,5422	0,049155	3	8	3	3
4	12	97,4140	0,051720	3	9	3	4
5	11	95,7702	0,084596	3	12	3	5
6	10	95,6766	0,086467	3	6	3	6
7	9	93,8677	0,122646	2	3	2	7
8	8	93,8673	0,122655	2	15	2	8
9	7	92,6019	0,147962	1	2	1	10
10	6	91,1857	0,176286	1	14	1	11
11	5	89,9867	0,200266	10	11	10	2
12	4	87,6672	0,246655	1	13	1	12
13	3	84,4718	0,310564	1	4	1	13
14	2	76,3215	0,473569	1	16	1	14
15	1	62,7547	0,744907	1	10	1	16

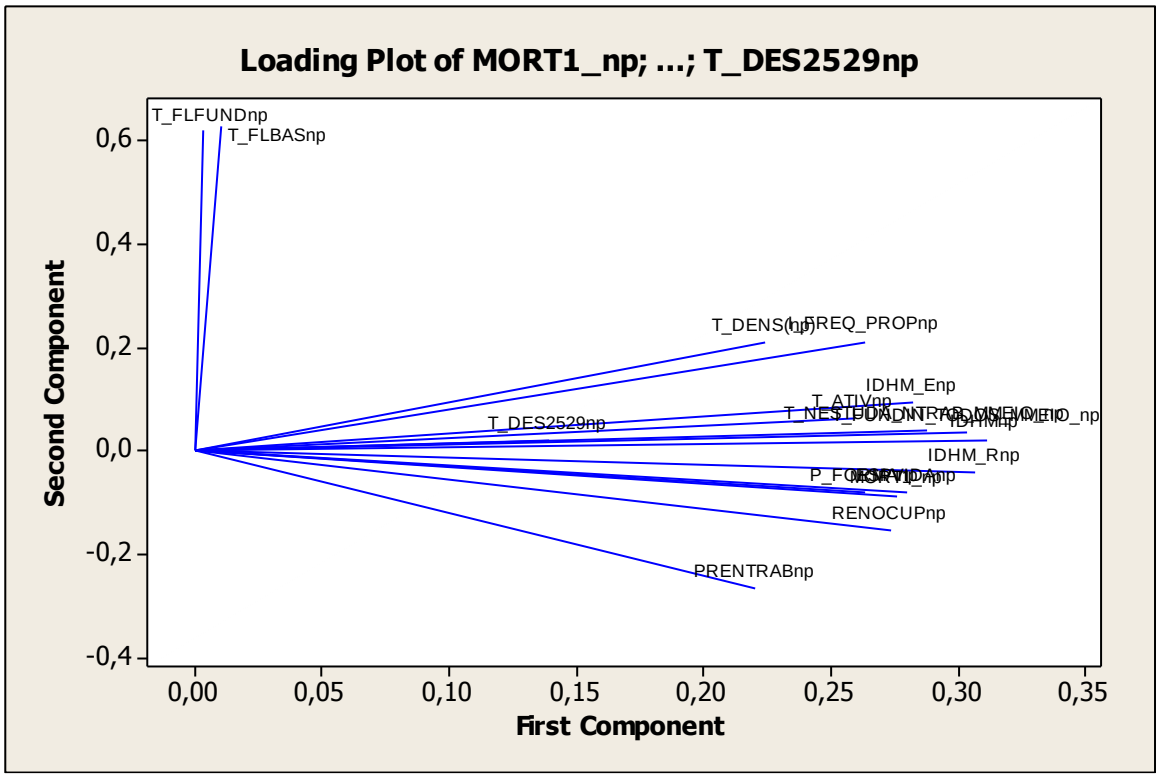
Final Partition

```
Cluster 1
  MORT1_np  ESPVIDAnp
Cluster 2
  T_NESTUDA_NTRAB_MMEIO_np  T_FUNDIN_TODOS_MMEIO_np  I_FREQ_PROPnp  IDHMnp
     IDHM_Enp  IDHM_Rnp  RENOCUPnp  T_ATIVnp
Cluster 3
  T_DENS(np)
Cluster 4
  T_FLBASnp
Cluster 5
  T_FLFUNDnp
Cluster 6
  PRENTRABnp
Cluster 7
  P_FORMAnp
Cluster 8
  T_DES2529np
```

12.3. PRINCIPAIS COMPONENTES

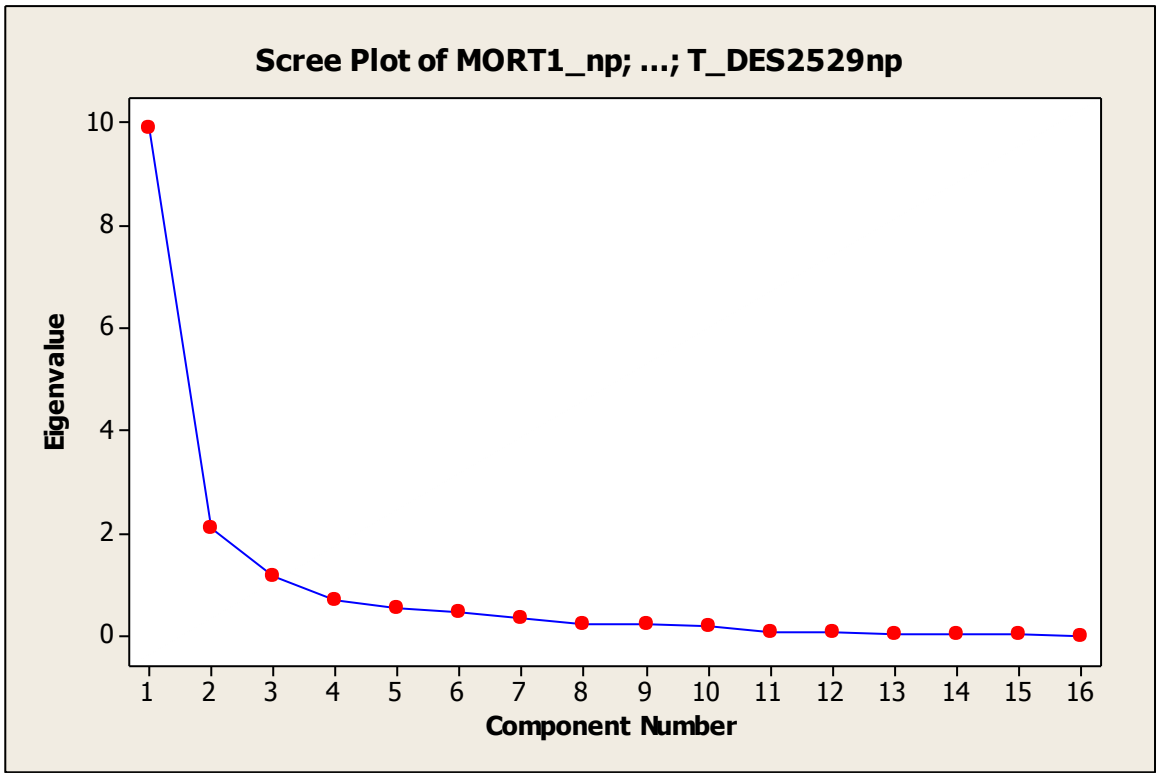
>> STAT >> MULTIVARIATE >> Principal Components

Figura 4 – Gráfico Loadin Plot das variáveis



Podemos observar 2 grupos de dados sendo o primeiro composto pelas seguintes variáveis: T_FLFUND e T_FLBAS. Já o segundo é formado pelo agrupamento das variáveis: MORT1, T_DESNUDA_MMEIO, T_FUNDIN_TODOS_MMEIO, T_DENS, RENOCUP, PRENTRAB, P_FORMA, T_ATIV, IDHM_R, IDHM, IDHM_E, I_FREQ_PROP, T_DES2529 e ESPVIDA.

Gráfico 5 – Scree Plot das variáveis



Existe um peso muito grande da primeira variável e as demais estão bem distantes. As variáveis 2 e 3 possuem peso maior que 1, e as demais não dá para aproveitar pois estão abaixo de 1.

Principal Component Analysis: MORT1_np; T_NESTUDA_NT; T_FUNDIN_TOD; T_DENS(np);

Eigenanalysis of the Correlation Matrix

Eigenvalue	9,8850	2,1133	1,1567	0,6908	0,5231	0,4456	0,3238	0,2407
Proportion	0,618	0,132	0,072	0,043	0,033	0,028	0,020	0,015
Cumulative	0,618	0,750	0,822	0,865	0,898	0,926	0,946	0,961
Eigenvalue	0,2081	0,1782	0,0865	0,0695	0,0340	0,0262	0,0181	0,0003
Proportion	0,013	0,011	0,005	0,004	0,002	0,002	0,001	0,000
Cumulative	0,974	0,985	0,991	0,995	0,997	0,999	1,000	1,000

Variable	PC1	PC2
MORT1_np	0,276	-0,087
T_NESTUDA_NTRAB_MMEIO_np	0,288	0,038
T_FUNDIN_TODOS_MMEIO_np	0,304	0,034
T_DENS(np)	0,224	0,208
ESPVIDAnp	0,280	-0,082
I_FREQ_PROPnp	0,264	0,210
IDHMnp	0,311	0,021
IDHM_Enp	0,282	0,092
IDHM_Rnp	0,307	-0,044
T_FLBASnp	0,010	0,628
T_FLFUNDnp	0,003	0,621
RENOCUPnp	0,273	-0,154
PRENTRABnp	0,221	-0,266
P_FORMAnp	0,263	-0,082
T_ATIVnp	0,259	0,062

T_DES2529np	0,139	0,019

Os gráficos abaixo apresentam uma visão dos agrupamentos das colunas em 2 variáveis PC1 PC2, e está agrupado por região e por Estado. Como os dados dos municípios são muito grandes (5564), dificulta um pouco a visualização. As colunas PC1 e PC2 são armazenadas como resultado do comando:

Figura 6 – Visão das variáveis C39 e C40 por Estado.

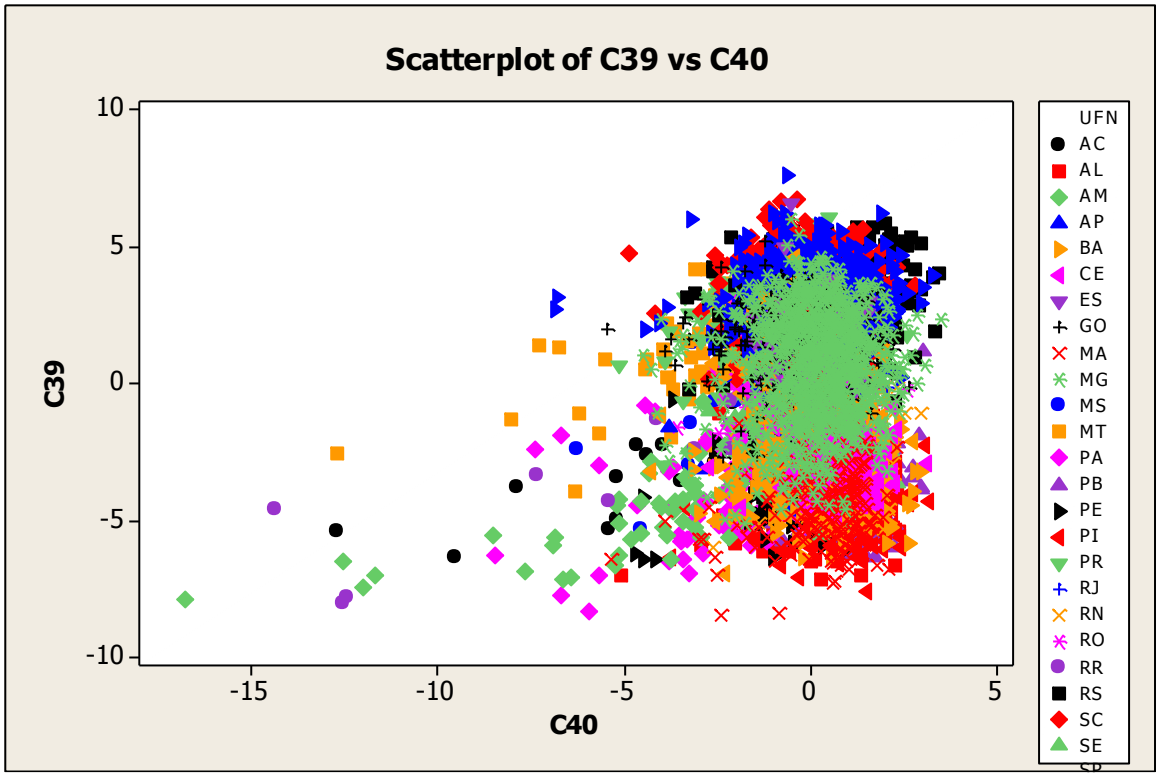
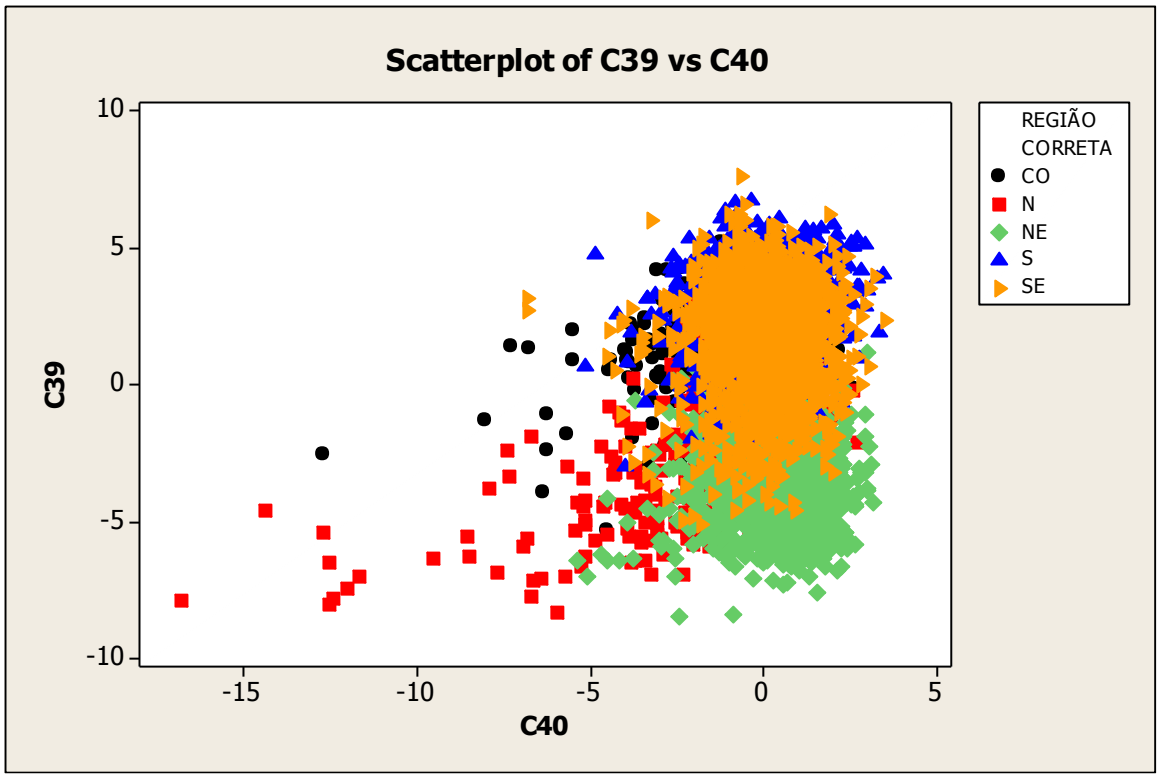


Figura 7 – Visão das variáveis C39 e C40, por Região.



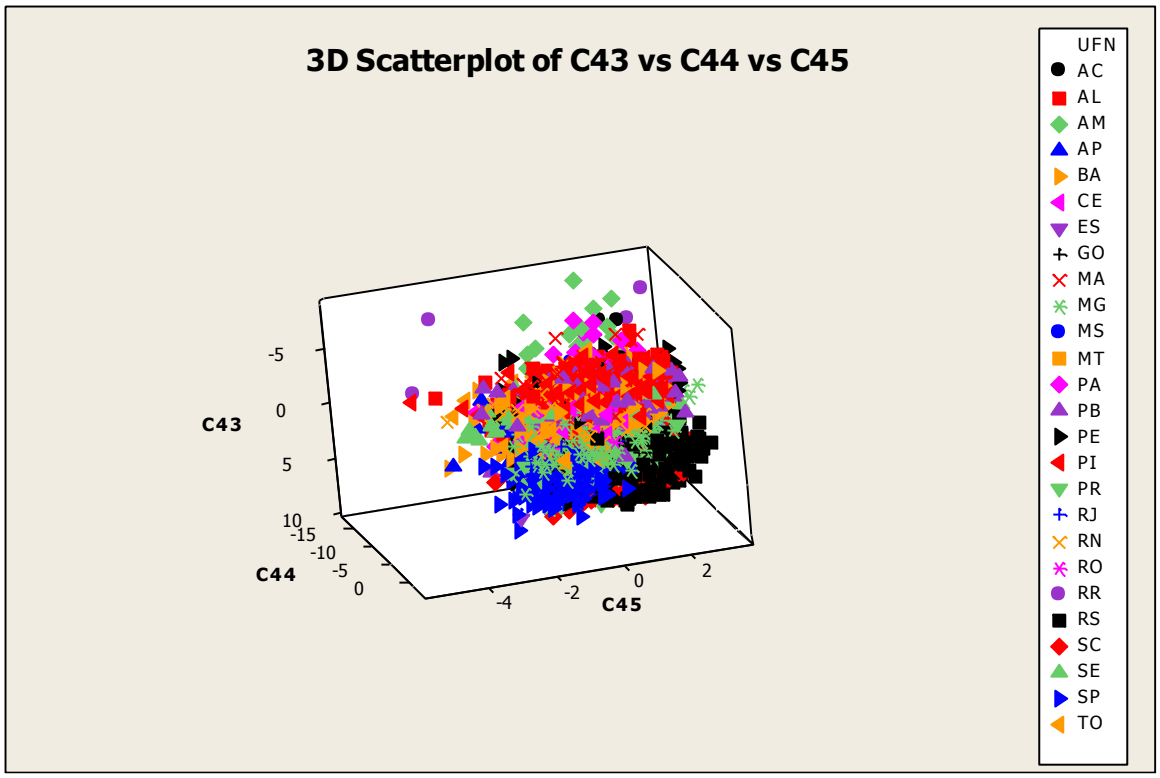
Principal Component Analysis: MORT1_np; T_NESTUDA_NT; T_FUNDIN_TOD; T_DENS(np);

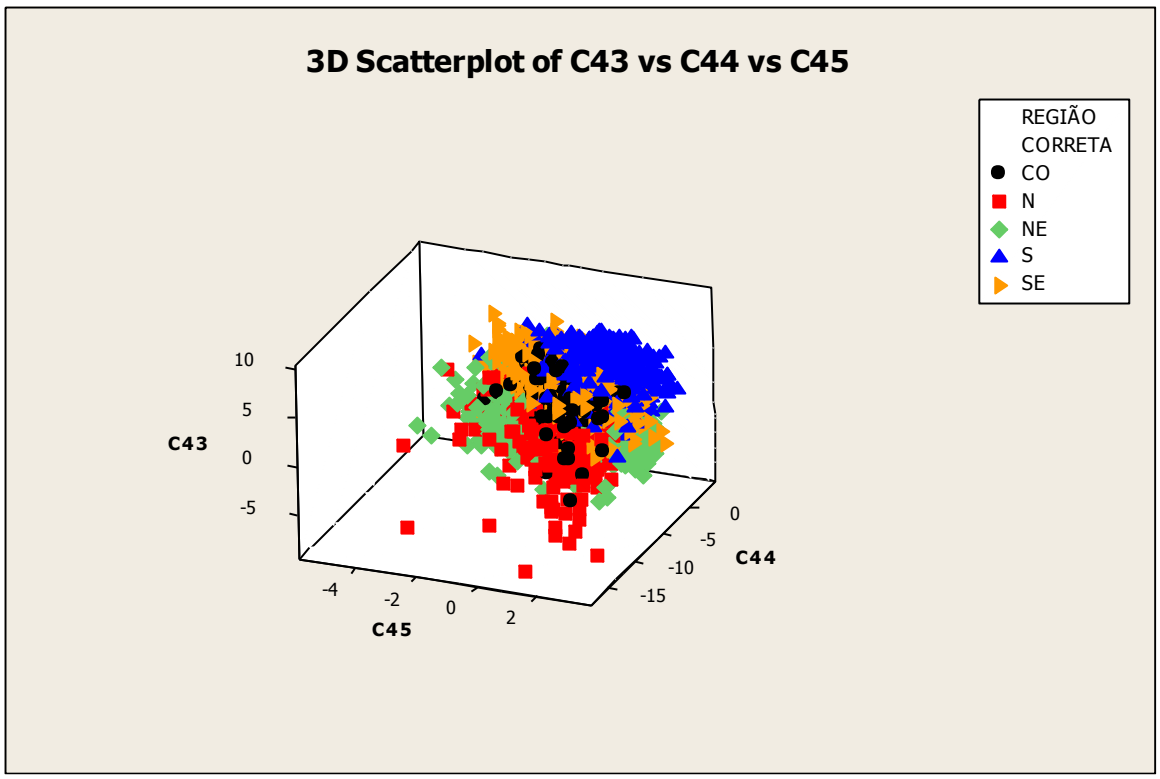
Eigenanalysis of the Correlation Matrix

Eigenvalue	9,8850	2,1133	1,1567	0,6908	0,5231	0,4456	0,3238	0,2407
Proportion	0,618	0,132	0,072	0,043	0,033	0,028	0,020	0,015
Cumulative	0,618	0,750	0,822	0,865	0,898	0,926	0,946	0,961

Eigenvalue	0,2081	0,1782	0,0865	0,0695	0,0340	0,0262	0,0181	0,0003
Proportion	0,013	0,011	0,005	0,004	0,002	0,002	0,001	0,000
Cumulative	0,974	0,985	0,991	0,995	0,997	0,999	1,000	1,000

Variable	PC1	PC2	PC3
MORT1_np	0,276	-0,087	0,055
T_NESTUDA_NTRAB_MMEIO_np	0,288	0,038	0,203
T_FUNDIN_TODOS_MMEIO_np	0,304	0,034	-0,098
T_DENS(np)	0,224	0,208	0,273
ESPVIDAnp	0,280	-0,082	0,030
I_FREQ_PROPnp	0,264	0,210	-0,074
IDHMnp	0,311	0,021	-0,136
IDHM_Enp	0,282	0,092	-0,257
IDHM_Rnp	0,307	-0,044	-0,023
T_FLBASnp	0,010	0,628	-0,133
T_FLFUNDnp	0,003	0,621	-0,071
RENOCUPnp	0,273	-0,154	-0,226
PRENTRABnp	0,221	-0,266	-0,087
P_FORMAnp	0,263	-0,082	-0,279
T_ATIVnp	0,259	0,062	0,325
T_DES2529np	0,139	0,019	0,718





COMENTÁRIOS DA ANÁLISE

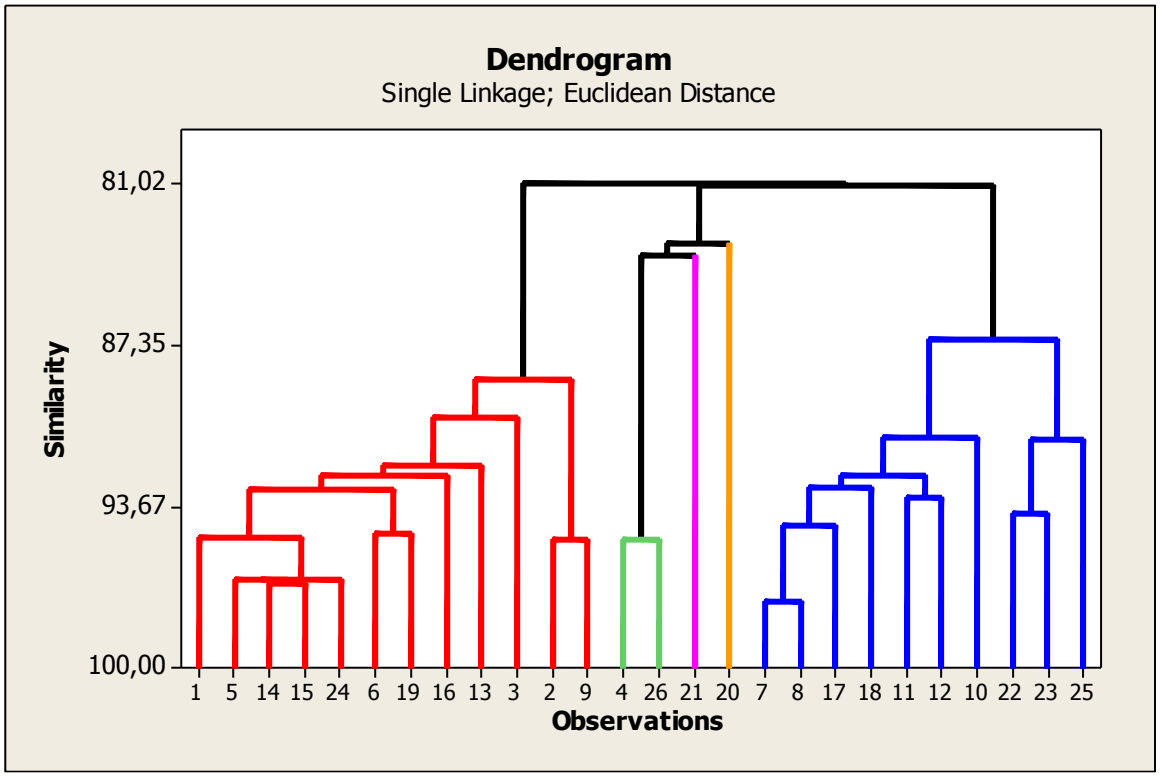
Pelo resultado das análises da correlação linear, dendrograma e principais componentes, os dados podem ser reduzidos para duas ou três variáveis, o que torna o trabalho com os números mais fácil e prático de serem manuseados.

13. DENDOGRAMA DOS DADOS AGRUPADOS PELO RESULTADO DAS MORT1 x ESPVIDA x IDHM_R x T_NESTUDA_MMEIO x IDHM POR ESTADO (-DF)

O Dendograma permite uma análise do grau de similaridade dos dados para uma determinada variável. Em seguida geramos o Dendograma Das variáveis por Estado

STAT >> MULTIVARIATE >> CLUSTER OBSERVATION

Gráfico2. Dendograma da variáveis **MORT1 x ESPVIDA x IDHM_R x T_NESTUDA_MMEIO x IDHM** por estados do Brasil (classificação não supervisionada)



Na figura 2 acima podem-se verificar cinco grupos de variáveis, agrupadas pela similaridade dos dados. Abaixo segue análise:

Cluster Analysis of Observations: MORT1 MEDIA; ESPVIDA MEDI; IDHM_R MEDIA; ...

Euclidean Distance, Single Linkage
Amalgamation Steps

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	25	97,4117	0,024644	7	8	7	2
2	24	96,7082	0,031343	14	15	14	2
3	23	96,5842	0,032523	5	14	5	3
4	22	96,5718	0,032641	5	24	5	4
5	21	95,0148	0,047466	4	26	4	2
6	20	94,9603	0,047985	2	9	2	2
7	19	94,8656	0,048887	1	5	1	5
8	18	94,7183	0,050289	6	19	6	2
9	17	94,4131	0,053195	7	17	7	3

10	16	93,9794	0,057325	22	23	22	2
11	15	93,3559	0,063262	11	12	11	2
12	14	93,0321	0,066345	1	6	1	7
13	13	92,9428	0,067195	7	18	7	4
14	12	92,4565	0,071825	7	11	7	6
15	11	92,4461	0,071924	1	16	1	8
16	10	92,0959	0,075258	1	13	1	9
17	9	91,1002	0,084739	22	25	22	3
18	8	91,0067	0,085629	7	10	7	7
19	7	90,1718	0,093578	1	3	1	10
20	6	88,7366	0,107244	1	2	1	12
21	5	87,1696	0,122163	7	22	7	10
22	4	83,8419	0,153848	4	21	4	3
23	3	83,4177	0,157887	4	20	4	4
24	2	81,0721	0,180220	4	7	4	14
25	1	81,0188	0,180728	1	4	1	26

Final Partition
Number of clusters: 5

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	12	0,124371	0,0890007	0,179539
Cluster2	2	0,001127	0,0237332	0,023733
Cluster3	10	0,111658	0,0965849	0,168335
Cluster4	1	0,000000	0,0000000	0,000000
Cluster5	1	0,000000	0,0000000	0,000000

Cluster Centroids

Variable	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Grand centroid
MORT1 MEDIA	0,529994	0,703225	0,829407	0,70387	0,76083	0,674044
ESPVIDA MEDIA	0,384633	0,517600	0,699930	0,52660	0,56850	0,528662
IDHM_R MEDIA	0,332672	0,437240	0,600770	0,52581	0,36225	0,452396
T_NESTUDA_MMEIO MEDIA	0,580943	0,620415	0,824346	0,75404	0,53725	0,682573
IDHM MEDIA	0,380952	0,503085	0,638431	0,50910	0,43288	0,496303

Distances Between Cluster Centroids

	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Cluster1	0,000000	0,274041	0,621673	0,366162	0,304257
Cluster2	0,274041	0,000000	0,368530	0,160679	0,152898
Cluster3	0,621673	0,368530	0,000000	0,270354	0,451158
Cluster4	0,366162	0,160679	0,270354	0,000000	0,290791
Cluster5	0,304257	0,152898	0,451158	0,290791	0,000000

No mapa abaixo pode ser percebido a divisão por cores dos estados de acordo com seu agrupamento por similaridade. Nesta representação vale destacar há certa coerência com as particularidades de cada estado, com o exemplo do agrupamento dos estados na cor verde se

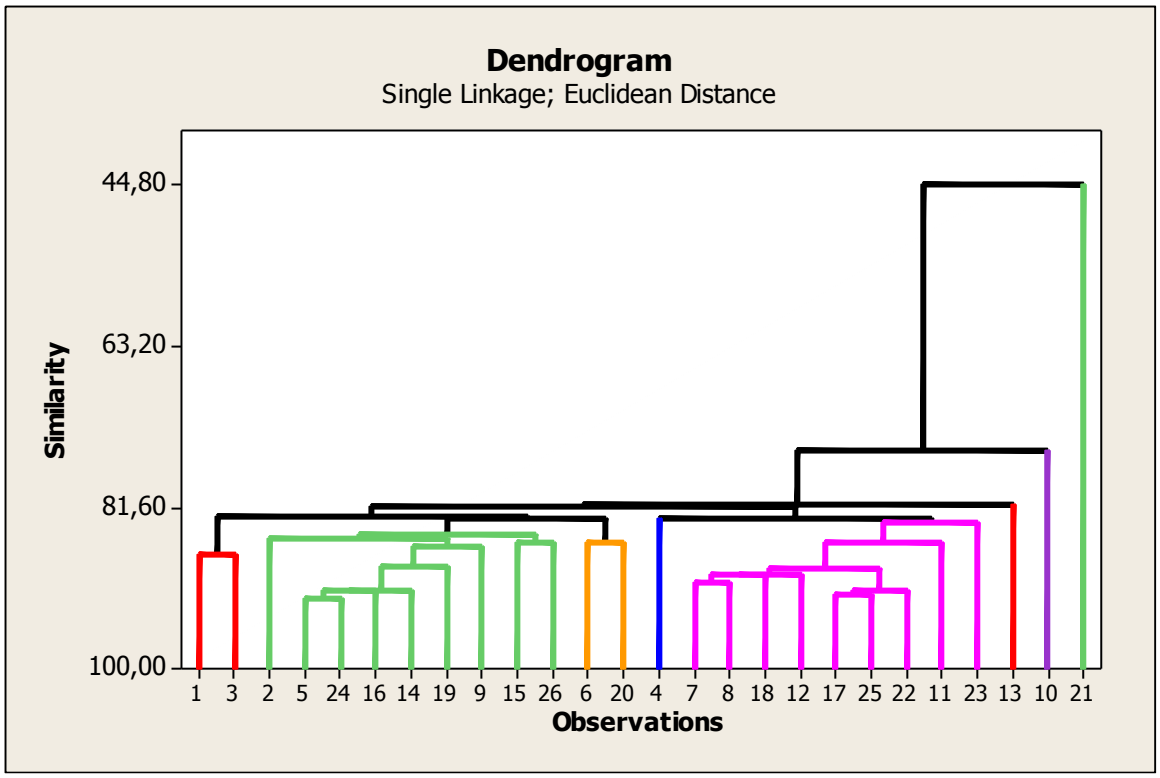
justifica por aparentemente apresentarem baixa capacidade de infraestrutura entre outras particularidades.

13.1 DENDOGRAMA DOS DADOS AGRUPADOS PELO RESULTADO DOS DESVIOS PADRÃO ENTRE MORT1 x ESPVIDA x IDHM_R x T_NESTUDA_MMEIO x IDHM POR ESTADO (-DF)

O Dendograma permite uma análise do grau de similaridade dos dados para uma determinada variável. Em seguida geramos o Dendograma de desvio padrão por Estado

STAT >> MULTIVARIATE >> CLUSTER OBSERVATION

Gráfico 3. Dendograma “Desigualômetro” da variáveis **MORT1 x ESPVIDA x IDHM_R x T_NESTUDA_MMEIO x IDHM** por Estado



No gráfico acima, podemos verificar oito agrupamentos de dados, que são compostos pelos Estados do Brasil. Então, o dendograma é construído por dois grandes agrupamentos e dois grupos com dois Estados, além dos quatro estados que ficaram isolados por não terem seus dados em similaridade com os outros estados.

Na classificação não supervisionada não se tem informações prévias sobre estes grupos. Não se tem informações sobre os porquês ou os critérios de agrupamento utilizados neste agrupamento.

Podemos observar que alguns estados possuem um alto nível de similaridade, o que significa que a desigualdade é baixa. O menor nível de desigualdade se encontra nos estados mais próximos do eixo X, por exemplo.

Cluster Analysis of Observations: MORT1 DESV; ESPVIDA DESV; IDHM_R DESV; ...

Euclidean Distance, Single Linkage
Amalgamation Steps

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	25	91,9064	0,015801	5	24	5	2
2	24	91,4546	0,016684	17	25	17	2
3	23	91,1236	0,017330	5	16	5	3
4	22	91,0898	0,017396	5	14	5	4
5	21	91,0027	0,017566	17	22	17	3
6	20	90,1732	0,019185	7	8	7	2
7	19	89,3134	0,020864	7	18	7	3
8	18	89,1897	0,021105	7	12	7	4
9	17	88,6171	0,022223	7	17	7	7
10	16	88,3844	0,022677	5	19	5	5

11	15	86,9868	0,025406	1	3	1	2
12	14	86,0131	0,027307	5	9	5	6
13	13	85,6880	0,027942	15	26	15	2
14	12	85,6362	0,028043	6	20	6	2
15	11	85,6190	0,028076	7	11	7	8
16	10	85,0921	0,029105	2	5	2	7
17	9	84,6852	0,029900	2	15	2	9
18	8	83,4424	0,032326	7	23	7	9
19	7	82,9434	0,033300	2	6	2	11
20	6	82,8465	0,033489	4	7	4	10
21	5	82,5994	0,033972	1	2	1	13
22	4	81,4595	0,036197	1	4	1	23
23	3	81,2598	0,036587	1	13	1	24
24	2	75,2414	0,048337	1	10	1	25
25	1	44,7976	0,107773	1	21	1	26

Final Partition
Number of clusters: 8

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	2	0,0003227	0,0127030	0,0127030
Cluster2	9	0,0057074	0,0238169	0,0343270
Cluster3	1	0,0000000	0,0000000	0,0000000
Cluster4	2	0,0003932	0,0140215	0,0140215
Cluster5	9	0,0056704	0,0233166	0,0408071
Cluster6	1	0,0000000	0,0000000	0,0000000
Cluster7	1	0,0000000	0,0000000	0,0000000
Cluster8	1	0,0000000	0,0000000	0,0000000

Cluster Centroids

Variable	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
MORT1 DESV	0,123325	0,145660	0,07617	0,097295	0,0502733
ESPVIDA DESV	0,121950	0,136722	0,08910	0,104600	0,0938444
IDHM_R DESV	0,110870	0,089354	0,10836	0,075010	0,0849878
T_NESTUDA_MMEIO DESV	0,098735	0,101307	0,08418	0,083540	0,0802367
IDHM DESV	0,126845	0,093412	0,09620	0,081005	0,0851344

Variable	Cluster6	Cluster7	Cluster8	Grand centroid
MORT1 DESV	0,07623	0,09034	0,05681	0,096315
ESPVIDA DESV	0,13420	0,09550	0,06900	0,112154
IDHM_R DESV	0,11292	0,11832	0,15795	0,093784
T_NESTUDA_MMEIO DESV	0,11859	0,09442	0,17585	0,095057
IDHM DESV	0,11179	0,12608	0,16592	0,097023

Distances Between Cluster Centroids

	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6	Cluster7
Cluster1	0,000000	0,048004	0,066783	0,067799	0,094225	0,054709	0,043155
Cluster2	0,048004	0,000000	0,088085	0,063611	0,107092	0,077580	0,081934
Cluster3	0,066783	0,088085	0,000000	0,045056	0,037113	0,059008	0,036587
Cluster4	0,067799	0,063611	0,045056	0,000000	0,049540	0,070237	0,064475
Cluster5	0,094225	0,107092	0,037113	0,049540	0,000000	0,072554	0,067800
Cluster6	0,054709	0,077580	0,059008	0,070237	0,072554	0,000000	0,050143
Cluster7	0,043155	0,081934	0,036587	0,064475	0,067800	0,050143	0,000000
Cluster8	0,130069	0,167332	0,128461	0,159741	0,147145	0,113422	0,107773

	Cluster8
Cluster1	0,130069
Cluster2	0,167332
Cluster3	0,128461

```
Cluster4  0,159741
Cluster5  0,147145
Cluster6  0,113422
Cluster7  0,107773
Cluster8  0,000000
```

O mapa abaixo representa a divisão por cores dos estados levando em conta o desvio padrão relacionado às variáveis **MORT1 x ESPVIDA x IDHM_R x T_NESTUDA_MMEIO x IDHM**.

Para a leitura dos gráficos se faz necessário entender que quando o nível de desigualdade se apresenta baixo, isto não representa uma situação boa, pois esta inferência é errônea. Portanto, salienta-se que os agrupamentos são feitos por similaridade. Assim, a baixa desigualdade não significa que as coisas vão bem ou mal, mas sim que existe um padrão nos municípios do estado em termos das variáveis selecionadas, uma maior similaridade entre estes municípios.

13.2 ANÁLISE DAS VARIÂNCIAS DAS VARIÁVEIS POR ESTADO

A análise das variâncias permite a verificação e visualização das médias e desvios padrões da variável a ser analisada. O gráfico BOXPLOT ilustra os agrupamentos, o seu tamanho varia de acordo com a quantidade de dados de cada grupo, e também é possível visualizar as ocorrências de *outliers* dentro de um grupo de dados.

13.2.1 Análise das variâncias da variável MORT1 por estado

Podemos visualizar no gráfico, uma grande variabilidade sobre as médias de MORT1 por estado. O estado que apresenta maior variabilidade dos dados é Alagoas. Goiás apresenta uma baixa variabilidade dos dados de MORT1, embora tenha alguns *outliers* que são os dados muito distantes das médias.

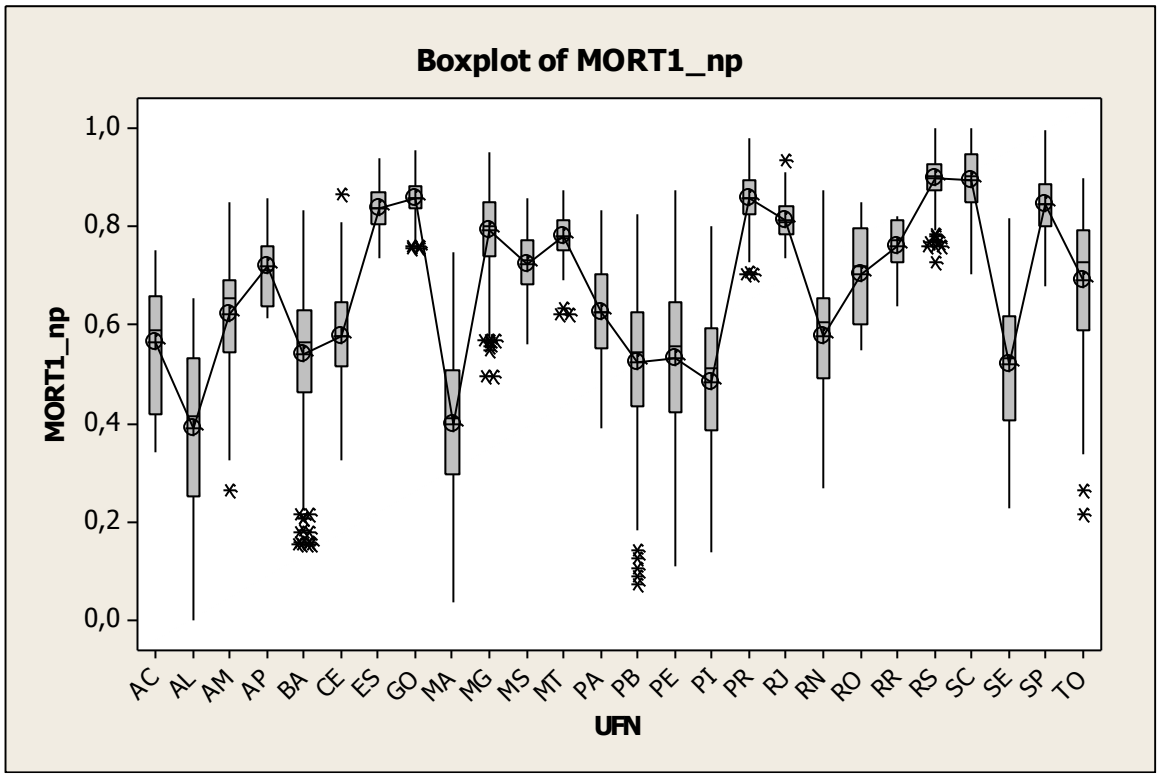


Figura 4. Gráfico BOXPLOT de MORT1 por estado

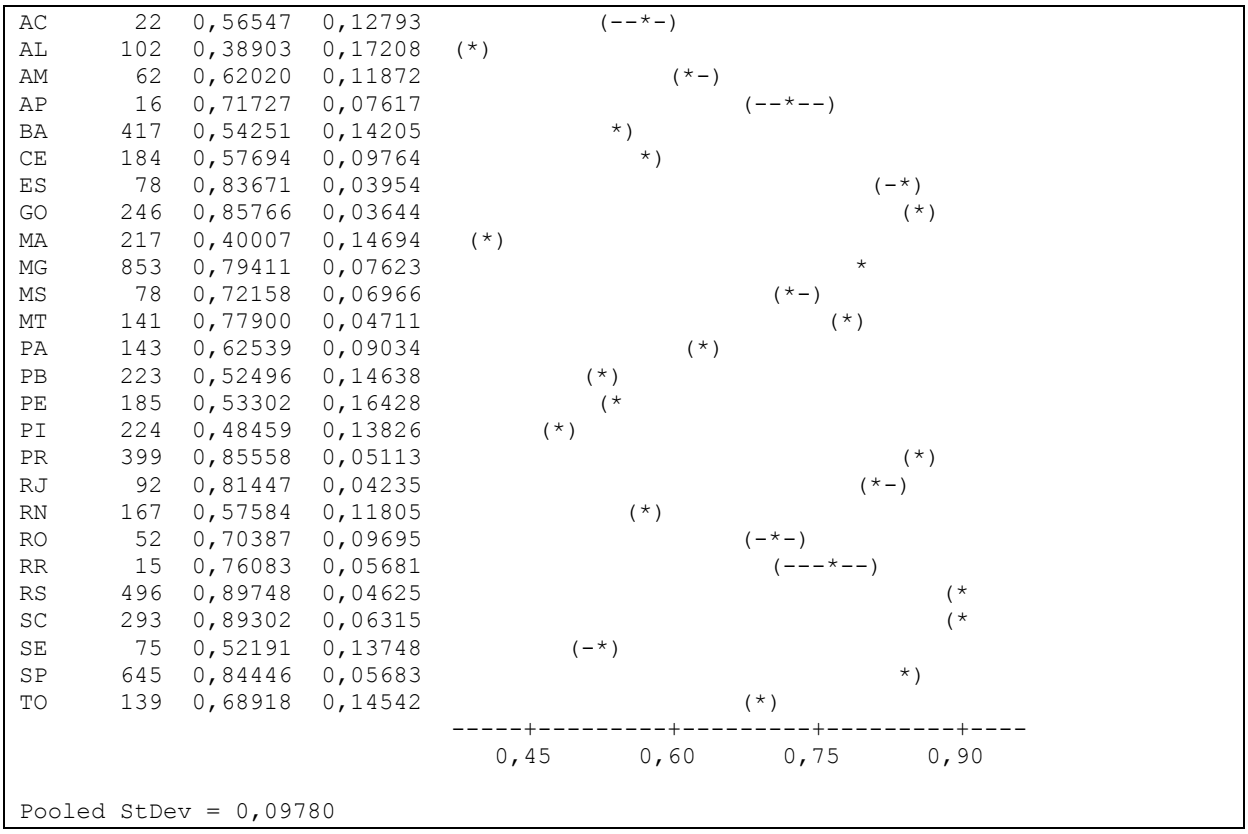
O resultado deste comando não fica armazenado na base de dados, é necessário copiar da área *session* para a área *worksheet*, para cada variável gerada. Com isso temos os dados dos 5564 municípios do Brasil, resumidos pela média e pelo desvio padrão. A partir destes dados resumidos, fica mais fácil trabalhar os dados, uma vez que estando resumido se torna mais simples a sua manipulação e análise.

Comando para gerar os dados agrupados → STAT>> ANOVA >> ONEWAY

Abaixo podemos visualizar os dados descritivos gerados pelo comando, para a variável IDHMn.

One-way ANOVA: MORT1_np versus UFN

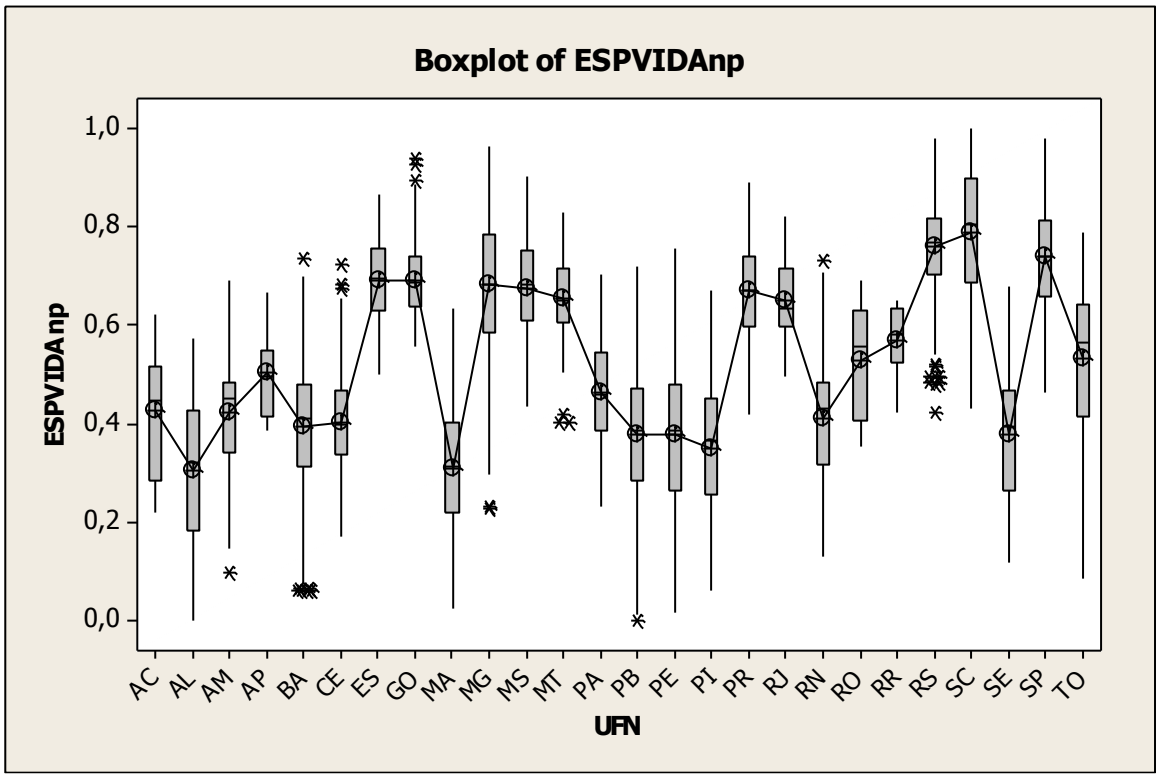
```
Source    DF        SS      MS       F      P
UFN       25  140,0967  5,6039  585,88  0,000
Error   5538   52,9706  0,0096
Total   5563  193,0673

S = 0,09780   R-Sq = 72,56%   R-Sq(adj) = 72,44%


                              Individual 95% CIs For Mean Based on
                              Pooled StDev
Level    N     Mean    StDev  -----+---------+---------+---------+----
```



13.2.2 Análise das variâncias da variável ESPVIDA por estado

Podemos visualizar no gráfico 5, uma grande variabilidade sobre as médias de ESPVIDA por estado. O estado que apresenta maior variabilidade dos dados é Acre. Ceara apresenta uma baixa variabilidade dos dados de ESPVIDA, embora tenha muitos *outliers* que são os dados muito distantes das médias.

Figura 5. Gráfico BOXPLOT de ESPVIDA por estado



Podemos visualizar no gráfico, uma grande variabilidade sobre as médias de ESPVIDA por estado. Destaca-se que diversos estados apresentam um grau semelhante de variabilidade como Acre, Alagoas, Tocantins, Sergipe, entre outros. Por outro lado, destaca-se os estados de Mato Grosso e Goiás que apresentam uma baixa variabilidade dos dados de ESPVIDA.

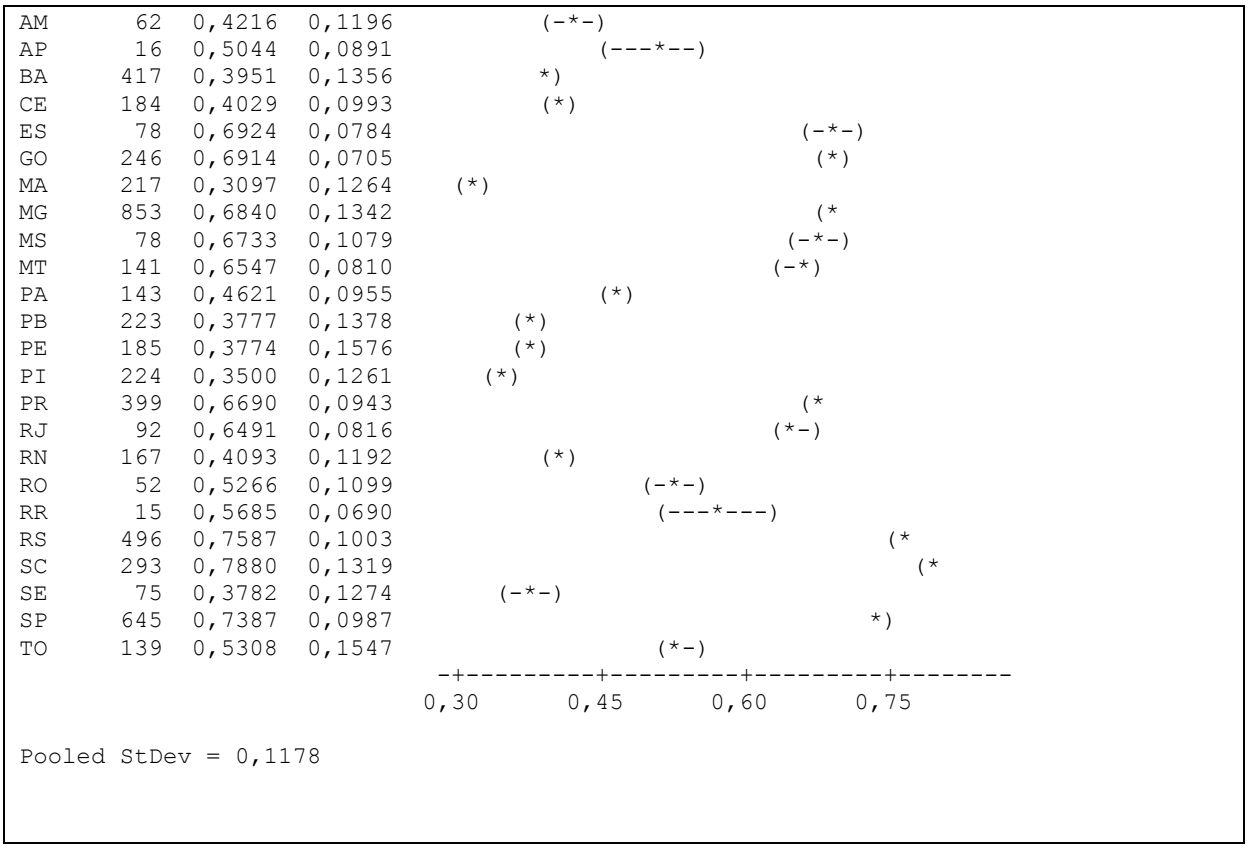
O resultado deste comando não fica armazenado na base de dados, é necessário copiar da área *session* para a área *worksheet*, para cada variável gerada. Com isso temos os dados dos 5564 municípios do Brasil, resumidos pela média e pelo desvio padrão. A partir destes dados resumidos, fica mais fácil trabalhar os dados, uma vez que estando resumido se torna mais simples a sua manipulação e análise.

Comando para gerar os dados agrupados → STAT>> ANOVA >> ONEWAY

Abaixo podemos visualizar os dados descritivos gerados pelo comando, para a variável ESPVIDA.

One-way ANOVA: ESPVIDAnp versus UFN

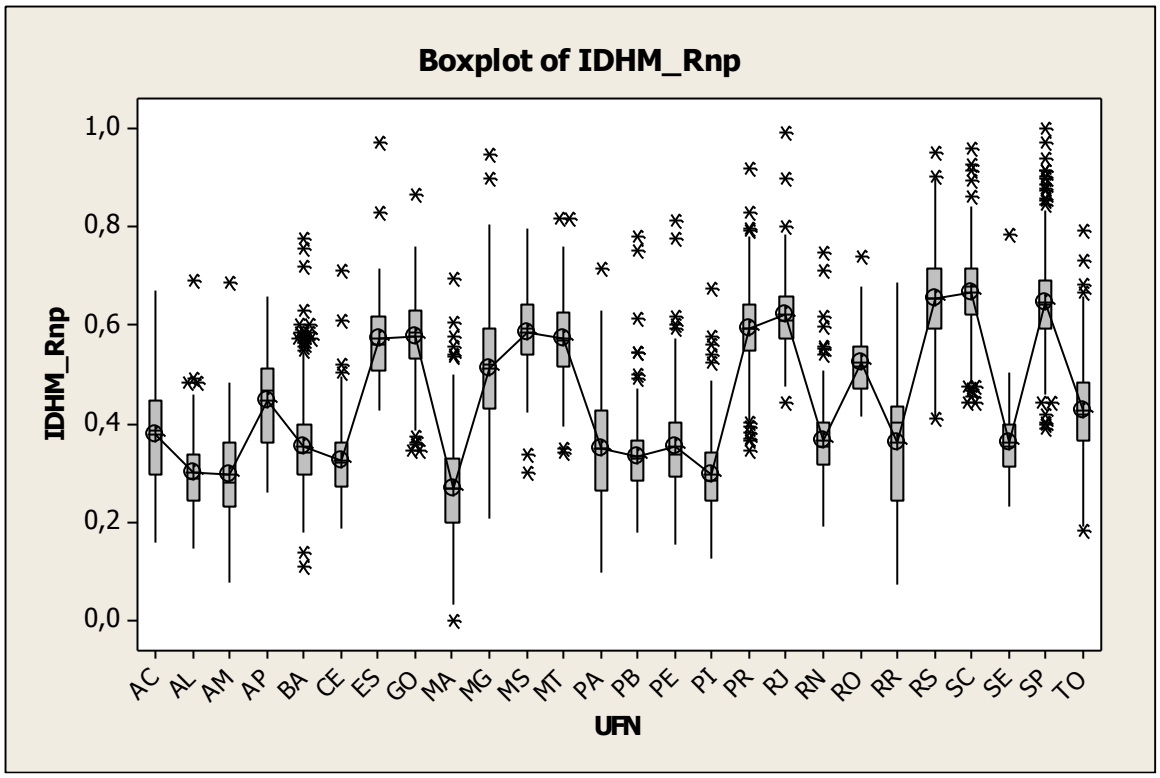
```
Source    DF        SS      MS       F      P
UFN       25  147,7345  5,9094  425,85  0,000
Error   5538   76,8498  0,0139
Total   5563  224,5843

S = 0,1178   R-Sq = 65,78%   R-Sq(adj) = 65,63%


                           Individual 95% CIs For Mean Based on
                           Pooled StDev
Level    N    Mean   StDev   -+---------+---------+---------+--------
AC      22  0,4267  0,1243         (--*---)
AL     102  0,3049  0,1457   (*-)
```



Podemos observar que alguns estados possuem baixa variabilidade dos dados em relação à média, como Paraiba, Pernambuco e Goiás. Já outros apresentam um desvio padrão com maior variabilidade como Roraima e Amapá.

13.2.3 Análise das variâncias da variável IDHM_R por estado

Figura. Gráfico BOXPLOT de IDHM_Rn por estado



Podemos visualizar no gráfico, uma grande variabilidade sobre as médias de IDHM_R por estado. O estado que apresenta maior variabilidade dos dados é Roraima, sendo que o Ceara, Rio Grande do Norte, Sergipe, entre outros possuem baixa variabilidade dos dados de IDHM_R.

O resultado deste comando não fica armazenado na base de dados, é necessário copiar da área *session* para a área *worksheet*, para cada variável gerada. Com isso temos os dados dos 5564 municípios do Brasil, resumidos pela média e pelo desvio padrão. A partir destes dados resumidos, fica mais fácil trabalhar os dados, uma vez que estando resumido se torna mais simples a sua manipulação e análise.

Comando para gerar os dados agrupados → STAT>> ANOVA >> ONEWAY

Abaixo podemos visualizar os dados descritivos gerados pelo comando, para a variável IDHM.

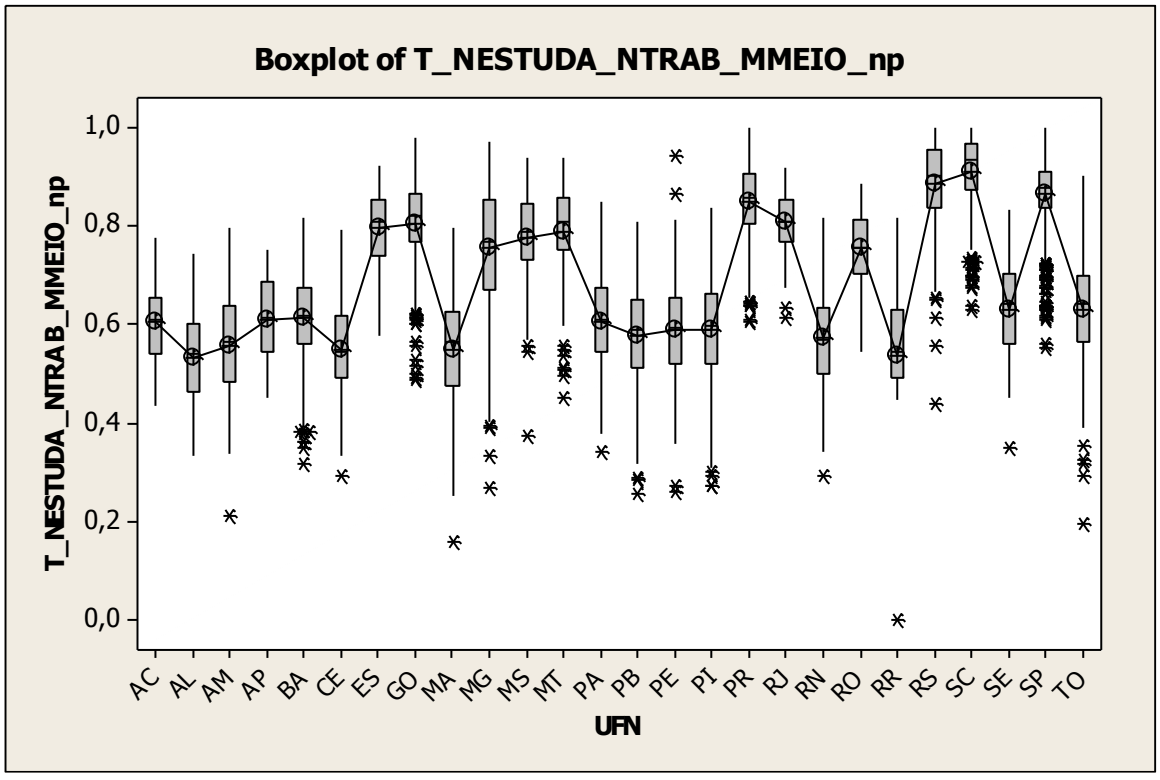
One-way ANOVA: IDHM_Rnp versus UFN

```
Source     DF        SS      MS       F      P
UFN        25  102,5672  4,1027  479,41  0,000
Error    5538   47,3933  0,0086
Total    5563  149,9605

S = 0,09251   R-Sq = 68,40%   R-Sq(adj) = 68,25%


                              Individual 95% CIs For Mean Based on
                              Pooled StDev
Level     N     Mean    StDev ---------+---------+---------+---------+
AC       22  0,37771  0,11764         (--*---)
AL      102  0,30216  0,08115      (*-)
AM       62  0,29555  0,10410     (-*-)
AP       16  0,44577  0,10836               (---*---)
```

```
BA      417  0,35589  0,09107           (*
CE      184  0,32443  0,07624        (*)
ES       78  0,57154  0,08769                            (-*)
GO      246  0,57849  0,08338                             (*)
MA      217  0,26815  0,10735   (*)
MG      853  0,51317  0,11292                         (*
MS       78  0,58624  0,08443                              (-*-)
MT      141  0,57464  0,09063                             (*)
PA      143  0,35103  0,11832          (*-)
PB      223  0,33479  0,07809         (*)
PE      185  0,35515  0,09589          (-*)
PI      224  0,29901  0,08516      (*)
PR      399  0,59460  0,07792                                (*
RJ       92  0,62030  0,08529                                 (-*)
RN      167  0,36740  0,08304           (-*)
RO       52  0,52581  0,07378                         (-*-)
RR       15  0,36225  0,15795        (---*---)
RS      496  0,65604  0,08930                                     (*
SC      293  0,66810  0,08423                                      (*)
SE       75  0,36079  0,08216          (-*-)
SP      645  0,64458  0,08202                                  (*
TO      139  0,42871  0,10028                  (-*)
                                 ---------+---------+---------+---------+
                                        0,36      0,48      0,60      0,72

Pooled StDev = 0,09251
```

13.2.4 Análise das variâncias da variável T_NESTUDA_MMEIO por estado

Figura 4. Gráfico BOXPLOT de IDHM por estado



Podemos visualizar no gráfico, uma grande variabilidade sobre as médias de **T_NESTUDA_MMEIO** por estado. Nota-se que os Estados apresentam uma variabilidade de dodos semelhante. Os Estados de São Paulo, Goiás e Rio de Janeiro apresentam uma baixa variabilidade dos dados de **T_NESTUDA_MMEIO**, embora tenha muitos *outliers* que são os dados muito distantes das médias.

O resultado deste comando não fica armazenado na base de dados, é necessário copiar da área *session* para a área *worksheet*, para cada variável gerada. Com isso temos os dados dos 5565 municípios do Brasil, resumidos pela média e pelo desvio padrão. A partir destes dados resumidos, fica mais fácil trabalhar os dados, uma vez que estando resumido se torna mais simples a sua manipulação e análise.

Comando para gerar os dados agrupados → STAT>> ANOVA >> ONEWAY

Abaixo podemos visualizar os dados descritivos gerados pelo comando, para a variável **T_NESTUDA_MMEIO**.

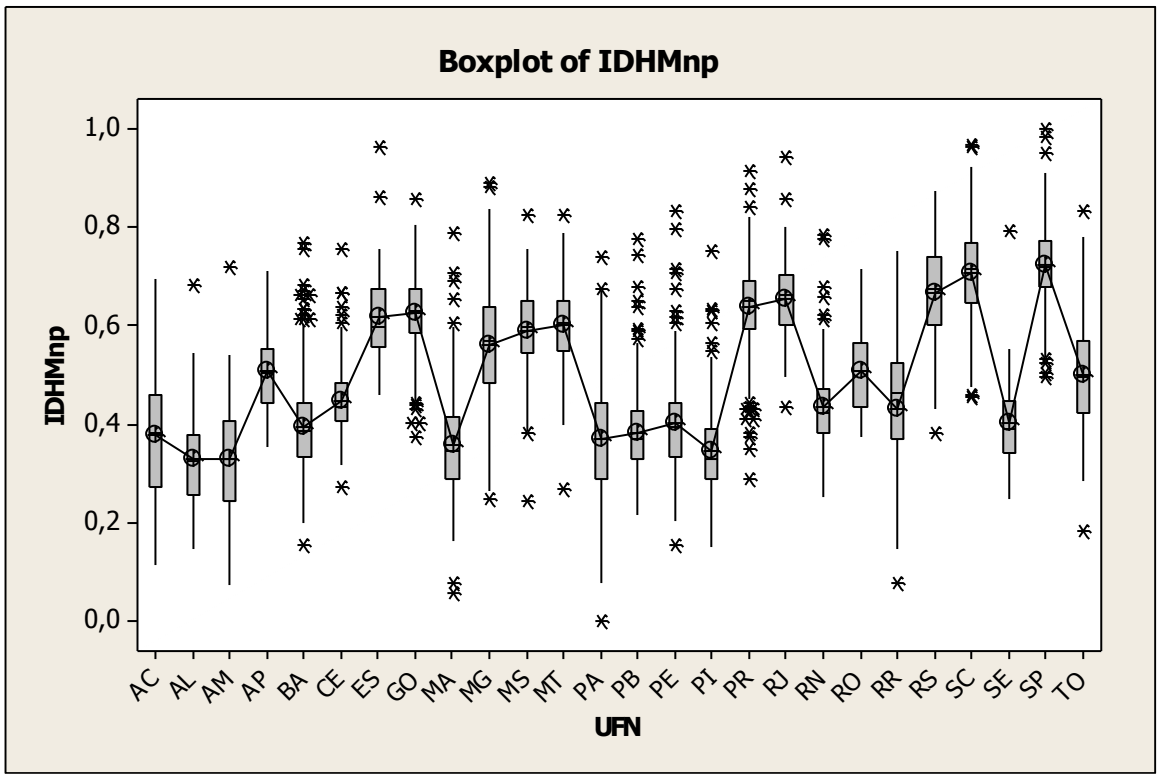
One-way ANOVA: T_NESTUDA_NTRAB_MMEIO_np versus UFN

```
Source    DF        SS      MS       F      P
UFN       25   90,1156  3,6046  403,66  0,000
Error   5538   49,4531  0,0089
Total   5563  139,5687

S = 0,09450   R-Sq = 64,57%   R-Sq(adj) = 64,41%


                                Individual 95% CIs For Mean Based on
                                Pooled StDev
Level    N     Mean    StDev  ---------+---------+---------+---------+
AC      22  0,60545  0,09113        (--*---)
AL     102  0,53362  0,09125     (*-)
AM      62  0,55598  0,10634      (-*-)
AP      16  0,61000  0,08418         (---*---)
BA     417  0,61430  0,08928            (*)
CE     184  0,54973  0,09210       (*)
ES      78  0,79712  0,07751                            (*-)
GO     246  0,80604  0,09049                             (*)
MA     217  0,54721  0,10507       (*)
MG     853  0,75496  0,11859                          (*
MS      78  0,77728  0,09605                           (-*-)
MT     141  0,78814  0,09768                            (-*)
PA     143  0,60738  0,09442           (-*)
PB     223  0,57795  0,10120         (*)
PE     185  0,58928  0,10374          (*)
PI     224  0,58780  0,10714          (*)
PR     399  0,84867  0,07432                                  (*
RJ      92  0,80964  0,05745                               (*-)
RN     167  0,57143  0,09953        (-*)
RO      52  0,75404  0,07498                         (-*-)
RR      15  0,53725  0,17585  (---*---)
RS     496  0,88581  0,08345                                    (*)
SC     293  0,91024  0,07552                                      (*)
SE      75  0,63118  0,09133             (-*)
SP     645  0,86556  0,06966                                  *)
TO     139  0,63083  0,12322             (-*)
                                ---------+---------+---------+---------+
                                       0,60      0,72      0,84      0,96
```

```
Pooled StDev = 0,09450
```

13.2.4 Análise das variâncias da variável IDHMn por estado

Figura 4. Gráfico BOXPLOT de IDHM por estado



Podemos visualizar no gráfico, uma grande variabilidade sobre as médias de IDHM por estado. O estado que apresenta maior variabilidade dos dados é Acre. Ceara apresenta uma baixa variabilidade dos dados de IDMH, embora tenha muitos *outliers* que são os dados muito distantes das médias.

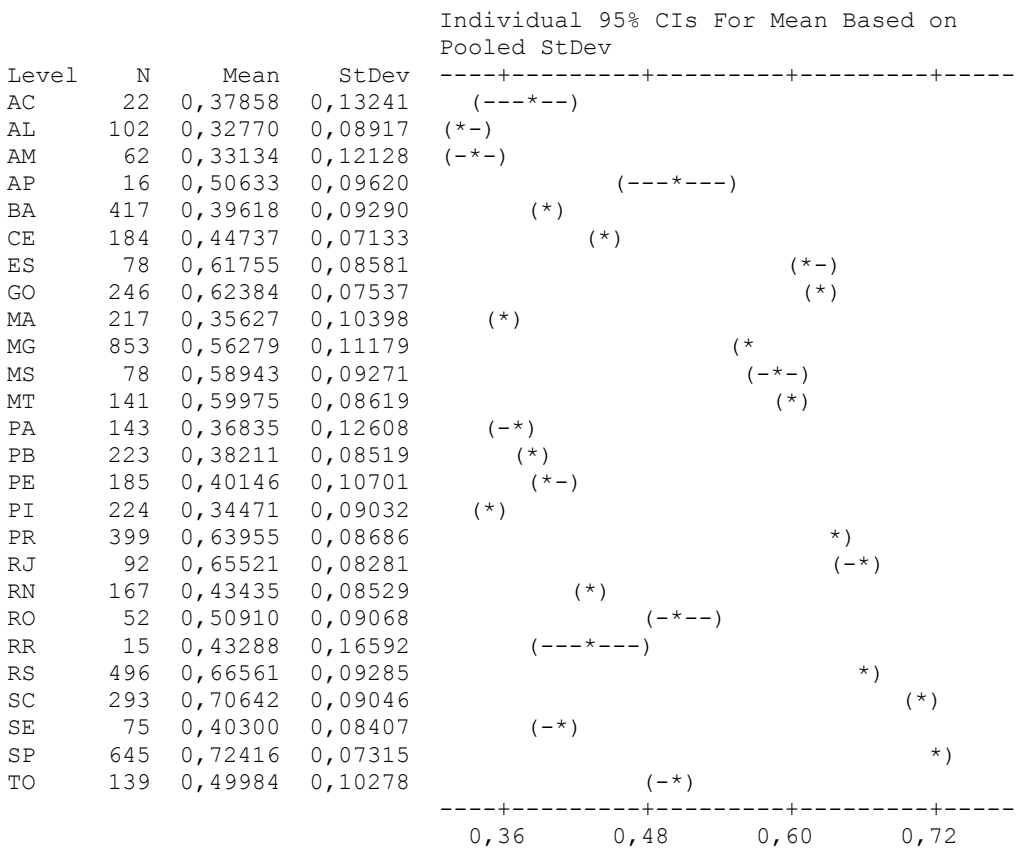
O resultado deste comando não fica armazenado na base de dados, é necessário copiar da área *session* para a área *worksheet*, para cada variável gerada. Com isso temos os dados dos 5565 municípios do Brasil, resumidos pela média e pelo desvio padrão. A partir destes dados resumidos, fica mais fácil trabalhar os dados, uma vez que estando resumido se torna mais simples a sua manipulação e análise.

Comando para gerar os dados agrupados → STAT>> ANOVA >> ONEWAY

Abaixo podemos visualizar os dados descritivos gerados pelo comando, para a variável IDHM.

One-way ANOVA: IDHMnp versus UFN

```
Source    DF        SS      MS       F      P
UFN       25   97,2750  3,8910  440,75  0,000
Error   5538   48,8902  0,0088
Total   5563  146,1652

S = 0,09396   R-Sq = 66,55%   R-Sq(adj) = 66,40%
```




```
Pooled StDev = 0,09396
```

COMENTÁRIOS DA ANÁLISE

As análise comparativas dos dados nos permitem um resumo dos dados através de cálculos específicos como médias e desvios padrões, tornando a análise dos dados mais fácil e simples. Os gráficos de Boxplot e Dendograma são excelentes figuras visuais para podermos analisar e interpretar os diferentes comportamentos dos dados. No dendograma podemos analisar as similaridades dos dados e no Boxplot podemos ver as relações entre as médias e as variâncias dos agrupamentos analisados. Trata-se de ferramentas úteis para análise de grandes volumes de dados.

14. ANÁLISE DISCRIMINANTE

14.1 ANÁLISE DISCRIMINANTE LINEAR POR REGIÃO

A análise discriminante é uma técnica da estatística multivariada utilizada para discriminar e classificar objetos, e estuda a separação de objetos de uma população em duas ou mais classes. Neste caso queremos discriminar os valores das variáveis **MORT1 x ESPVIDA x IDHM_R x T_NESTUDA_MMEIO x IDHM** dos municípios[4] do Brasil, e utilizaremos inicialmente a variável categórica Região. Para geração de análise discriminante utilizaremos o comando do Minitab:

STAT >> MULTIVARIATE >> DISCRIMINANT ANALISYS

Discriminant Analysis: REGIÃO CORRE versus MORT1_np; T_NESTUDA_NT; ...

Linear Method for Response: REGIÃO CORRETA

Predictors: MORT1_np; T_NESTUDA_NTRAB_MMEIO_np; ESPVIDAnp; IDHMnp; IDHM_Rnp

Group	CO	N	NE	S	SE
Count	465	449	1794	1188	1668

Summary of classification

	True Group				
Put into Group	CO	N	NE	S	SE
CO	124	57	35	110	267
N	47	257	464	14	276
NE	2	88	1250	0	3
S	149	13	3	870	196
SE	143	34	42	194	926
Total N	465	449	1794	1188	1668
N correct	**124**	**257**	**1250**	**870**	**926**
Proportion	0,267	0,572	0,697	0,732	0,555

N = 5564 N Correct = 3427 Proportion Correct = 0,616

Squared Distance Between Groups

	CO	N	NE	S	SE
CO	0,0000	4,0564	10,8573	1,1653	0,5258
N	4,0564	0,0000	3,2052	8,5421	4,7610
NE	10,8573	3,2052	0,0000	18,1824	10,1413
S	1,1653	8,5421	18,1824	0,0000	2,1044
SE	0,5258	4,7610	10,1413	2,1044	0,0000

Linear Discriminant Function for Groups

	CO	N	NE	S	SE
Constant	-60,26	-45,53	-34,21	-72,57	-59,05
MORT1_np	172,16	167,09	134,35	188,40	161,83
T_NESTUDA_NTRAB_MMEIO_np	81,50	72,61	71,40	90,15	83,06
ESPVIDAnp	-111,71	-113,63	-97,27	-122,74	-102,13
IDHMnp	-5,68	-5,42	11,87	-8,39	5,31
IDHM_Rnp	-8,35	-16,02	-31,30	-6,97	-21,38

[4] Para está análise excluiu-se o DF – Distrito Federal.

Com base nas informações apresentadas na figura 2 pode ser notado que a região que acertou mais é Sul (0,732) e a que errou mais foi a região Centro Oeste (0,267). As informações ainda exibem o cruzamento de dados entre as regiões, por exemplo, a região Nordeste possui 1794 municípios e apenas 1250 correspondem a região. O nome desta matriz é *confusion matrix* ou matriz de confusão. Podemos concluir que o agrupamento por região não é uma boa escolha segundo esta avaliação.

14.2 ANÁLISE DISCRIMINANTE LINEAR POR "2 BRASIS"

Esta segunda análise está interessada em verificar os possíveis agrupamentos dos dados utilizando a variável 2 Brasis, calculada a partir do exercício anterior, e demonstra os agrupamentos do Brasil segundo sua proximidade de dados de educação. Para esta análise foram agrupadas as regiões de Sul, Sudeste e Centro-Oeste como COSSE, e as regiões de Norte e Nordeste como NNE.

Discriminant Analysis: REAGRUPAMENT versus MORT1_np; T_NESTUDA_NT; ...

Linear Method for Response: REAGRUPAMENTO DE REGIÕES

Predictors: MORT1_np; T_NESTUDA_NTRAB_MMEIO_np; ESPVIDAnp; IDHMnp; IDHM_Rnp

Group	COSSE	NNE
Count	3321	2243

Summary of classification

	True Group	
Put into Group	COSSE	NNE
COSSE	3102	235
NNE	219	2008
Total N	3321	2243
N correct	**3102**	**2008**
Proportion	0,934	0,895

N = 5564 N Correct = 5110 Proportion Correct = 0,918

Squared Distance Between Groups

	COSSE	NNE
COSSE	0,00000	8,70544
NNE	8,70544	0,00000

Linear Discriminant Function for Groups

	COSSE	NNE
Constant	-50,52	-29,19
MORT1_np	119,85	102,56
T_NESTUDA_NTRAB_MMEIO_np	75,02	63,93
ESPVIDAnp	-76,17	-75,19
IDHMnp	24,90	27,25
IDHM_Rnp	-39,45	-46,45

Existem duas possibilidades de realizar a análise discriminante que são a linear e a quadrática. Dependendo da variável deve-se dar mais peso e mais atenção a um método em detrimento do outro. Neste caso a linear já nos apresenta informações satisfatórias. Podemos observar que alguns estados e municípios da região COSSE tem características das região NNE, visto pelo número 235 municípios foram encontrados na intersecção entre COSSE e NNE.

14.3 ANÁLISE DISCRIMINANTE QUADRÁTICA POR "3 BRASIS"

Uma boa classificação deve resultar em pequenos erros, isto é, deve haver pouca probabilidade de classificação inadequada, e para que isso ocorra a regra de classificação deve considerar as probabilidades a priori e os custos de classificação errada. Outro fator que uma regra de classificação deve considerar é se as variâncias das populações são iguais ou não. Quando a regra de classificação assume que as variâncias das populações são iguais, as funções discriminantes são ditas lineares e quando não são funções discriminantes quadráticas. Vamos agora verificar a função quadrática para os 3 Brasis apresentado na análise anterior.

Discriminant Analysis: REAGRUPAMENT versus MORT1_np; T_NESTUDA_NT; ...

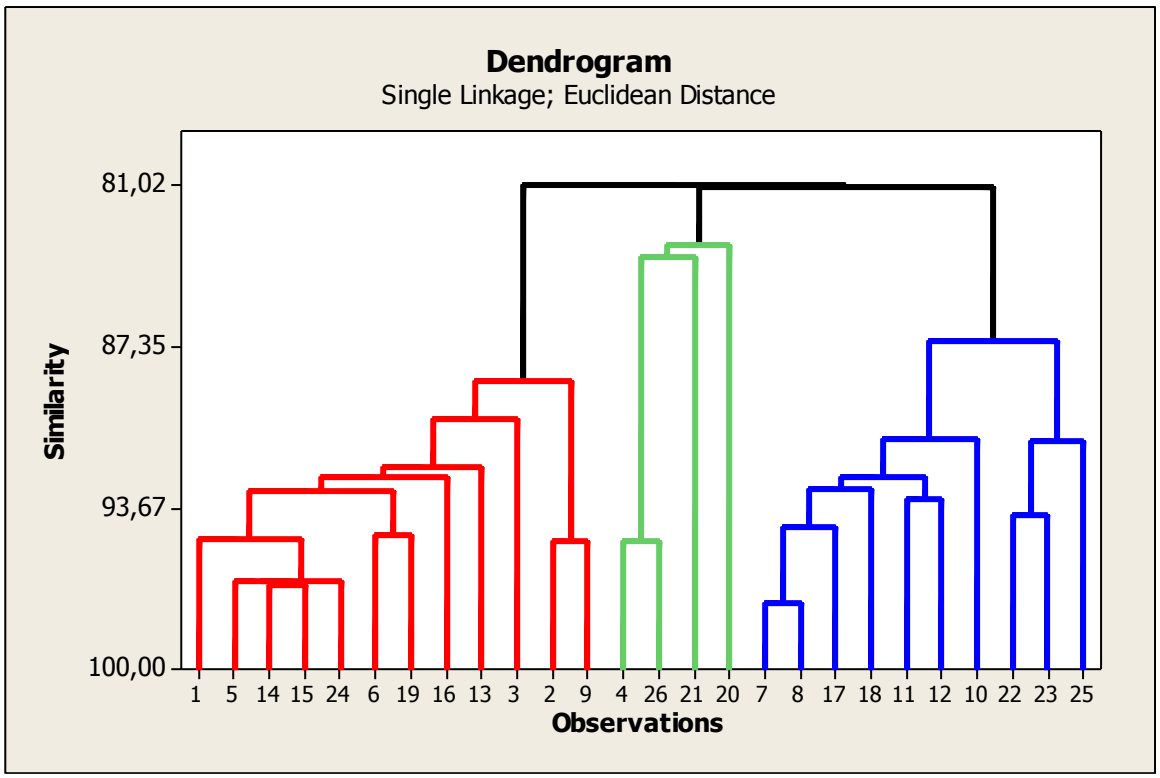
```
Quadratic Method for Response:   REAGRUPAMENTO DE REGIÕES


Predictors: MORT1_np; T_NESTUDA_NTRAB_MMEIO_np; ESPVIDAnp; IDHMnp; IDHM_Rnp


Group      COSSE      NNE
Count       3321     2243


Summary of classification

                  True Group
Put into Group  COSSE    NNE
COSSE            3054    183
NNE               267   2060
Total N          3321   2243
N correct        3054   2060
Proportion      0,920  0,918

N = 5564           N Correct = 5114           Proportion Correct = 0,919


From  Generalized Squared Distance to Group

Group    COSSE     NNE
COSSE  -27,91  -16,28
NNE     -6,67  -26,75
```

No modelo quadrático a proporção não foi alterada permanecendo em 0,919. Seguindo o princípio da simplicidade, vamos escolher o método linear, pois este é o mais simples.

Em Ciência, a parcimônia é a preferência pela explicação mais simples para uma observação. Esta geralmente é considerada a melhor maneira de julgar as hipóteses. Parcimônia também é um conceito utilizado na sistemática moderna que estabelece que ao construir e selecionar árvores filogenéticas, ou seja, os dados, o melhor critério é baseado em seus princípios: normalmente é correto o relacionamento mais simples encontrado entre dois indivíduos, aquele que apresente o menor número de passos intermediários ou mudanças evolucionárias. Portanto, não há diferença entre o método linear e o quadrático, o que não justifica a utilização do método quadrático.

14.4 ANÁLISE DISCRIMINANTE LINEAR PARA DADOS AGRUPADOS

A análise discriminante é uma técnica da estatística multivariada utilizada para discriminar e classificar objetos, e estuda a separação de objetos de uma população em duas ou mais classes. Inicialmente foram transfomadas as cinco regiões anteriormente divididas em três, pois esta análise somente pode ser realizada com mais de um caso (minicípio) por agrupamento.



Este agrupamento pode ser melhor representado no mapa abaixo:

Cluster Analysis of Observations: MORT1 MEDIA; ESPVIDA MEDI; IDHM_R MEDIA; ...

Euclidean Distance, Single Linkage
Amalgamation Steps

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	25	97,4117	0,024644	7	8	7	2
2	24	96,7082	0,031343	14	15	14	2
3	23	96,5842	0,032523	5	14	5	3
4	22	96,5718	0,032641	5	24	5	4
5	21	95,0148	0,047466	4	26	4	2
6	20	94,9603	0,047985	2	9	2	2
7	19	94,8656	0,048887	1	5	1	5
8	18	94,7183	0,050289	6	19	6	2
9	17	94,4131	0,053195	7	17	7	3
10	16	93,9794	0,057325	22	23	22	2
11	15	93,3559	0,063262	11	12	11	2
12	14	93,0321	0,066345	1	6	1	7
13	13	92,9428	0,067195	7	18	7	4
14	12	92,4565	0,071825	7	11	7	6
15	11	92,4461	0,071924	1	16	1	8
16	10	92,0959	0,075258	1	13	1	9
17	9	91,1002	0,084739	22	25	22	3

```
  18          8     91,0067  0,085629   7    10        7        7
  19          7     90,1718  0,093578   1     3        1       10
  20          6     88,7366  0,107244   1     2        1       12
  21          5     87,1696  0,122163   7    22        7       10
  22          4     83,8419  0,153848   4    21        4        3
  23          3     83,4177  0,157887   4    20        4        4
  24          2     81,0721  0,180220   4     7        4       14
  25          1     81,0188  0,180728   1     4        1       26


Final Partition
Number of clusters: 3


                                                    Maximum
                             Within        Average  distance
                Number of  cluster sum  distance from      from
          observations    of squares       centroid  centroid
Cluster1            12     0,124371      0,0890007  0,179539
Cluster2             4     0,046864      0,0929527  0,150381
Cluster3            10     0,111658      0,0965849  0,168335


Cluster Centroids

                                                       Grand
Variable               Cluster1  Cluster2  Cluster3  centroid
MORT1 MEDIA            0,529994  0,717787  0,829407  0,674044
ESPVIDA MEDIA          0,384633  0,532575  0,699930  0,528662
IDHM_R MEDIA           0,332672  0,440635  0,600770  0,452396
T_NESTUDA_MMEIO MEDIA  0,580943  0,633030  0,824346  0,682573
IDHM MEDIA             0,380952  0,487037  0,638431  0,496303


Distances Between Cluster Centroids

          Cluster1  Cluster2  Cluster3
Cluster1  0,000000  0,287709  0,621673
Cluster2  0,287709  0,000000  0,354446
Cluster3  0,621673  0,354446  0,000000
```

Neste caso queremos discriminar os valores das variáveis **MORT1 x ESPVIDA x IDHM_R x T_NESTUDA_MMEIO x IDHM** dos municípios[5] do Brasil, e utilizaremos inicialmente a variável categórica Região. Para geração de análise discriminante utilizaremos o comando do Minitab:

STAT >> MULTIVARIATE >> DISCRIMINANT ANALISYS

Discriminant Analysis: 3 AGRUPAMENT versus MORT1 MEDIA; ESPVIDA MEDI; ...

```
Linear Method for Response: 3 AGRUPAMENTOS DE ESTADOS


Predictors: MORT1 MEDIA; ESPVIDA MEDIA; IDHM_R MEDIA; T_NESTUDA_MMEIO MEDIA;
            IDHM MEDIA


Group         1         2         3
Count        12         4        10
```

[5] Para está análise excluiu-se o DF – Distrito Federal.

```
Summary of classification

                        True Group
Put into Group      1       2       3
1                  12       0       0
2                   0       4       0
3                   0       0      10
Total N            12       4      10
N correct          12       4      10
Proportion      1,000   1,000   1,000

N = 26           N Correct = 26           Proportion Correct = 1,000


Squared Distance Between Groups

          1         2         3
1   0,0000   17,4688   69,8341
2  17,4688    0,0000   29,7555
3  69,8341   29,7555    0,0000


Linear Discriminant Function for Groups

                              1         2         3
Constant                -118,42   -137,90   -221,63
MORT1 MEDIA               47,19     52,45    -33,77
ESPVIDA MEDIA             68,34    126,71    286,18
IDHM_R MEDIA            -504,08   -395,89   -447,75
T_NESTUDA_MMEIO MEDIA    475,63    382,42    445,98
IDHM MEDIA               201,91    211,55    269,90
```

Figura 2. Resultado do comando STAT >> MULTIVARIATE >> DISCRIMINANT ANALISYS

Com base nas informações apresentadas na figura 2 pode ser notado que os Estados se enquadraram nos respectivos agrupamentos. As informações ainda exibem o cruzamento de dados entre as regiões, por exemplo, a região Nordeste possui 1794 municípios e apenas 1255 correspondem a região. O nome desta matriz é *confusion matrix* ou matriz de confusão. Podemos concluir que o agrupamento por região não é uma boa escolha segundo esta avaliação.

15. REGRESSÃO LOGÍSTICA ORDINAL PARA AS VARIÁVEIS: MORT1 x ESPVIDA x IDHM_R x T_NESTUDA_MMEIO x IDHM.

Inicialmente foram classificadas pela análise ANOVA as regiões para as variáveis: **MORT1 x ESPVIDA x IDHM_R x T_NESTUDA_MMEIO x IDHM**.

One-way ANOVA: MORT1_np versus REGIÃO CORRETA

```
Source            DF        SS       MS        F      P
REGIÃO CORRETA     4  130,1073  32,5268  2871,93  0,000
Error           5559   62,9600   0,0113
Total           5563  193,0673

S = 0,1064   R-Sq = 67,39%   R-Sq(adj) = 67,37%


                           Individual 95% CIs For Mean Based on
                           Pooled StDev
Level     N    Mean   StDev  ---------+---------+---------+---------+
CO      465  0,8110  0,0706                              (*)
N       449  0,6584  0,1230                   (*)
NE     1794  0,5119  0,1529  *)
S      1188  0,8823  0,0558                                        *)
SE     1668  0,8167  0,0705                                 (*
                           ---------+---------+---------+---------+
                                  0,60      0,70      0,80      0,90

Pooled StDev = 0,1064
```

One-way ANOVA: ESPVIDAnp versus REGIÃO CORRETA

```
Source            DF        SS       MS        F      P
REGIÃO CORRETA     4  140,4313  35,1078  2319,16  0,000
Error           5559   84,1530   0,0151
Total           5563  224,5843

S = 0,1230   R-Sq = 62,53%   R-Sq(adj) = 62,50%


                           Individual 95% CIs For Mean Based on
                           Pooled StDev
Level     N    Mean   StDev  ---+---------+---------+---------+------
CO      465  0,6772  0,0825                              (*)
N       449  0,4886  0,1292                 (*)
NE     1794  0,3714  0,1356  *)
S      1188  0,7358  0,1177                                      (*
SE     1668  0,7036  0,1202                                   *)
                           ---+---------+---------+---------+------
                            0,40      0,50      0,60      0,70

Pooled StDev = 0,1230
```

One-way ANOVA: T_NESTUDA_NTRAB_MMEIO_np versus REGIÃO CORRETA

```
Source            DF        SS       MS        F      P
REGIÃO CORRETA     4   82,1968  20,5492  1991,10  0,000
Error           5559   57,3719   0,0103
```

```
Total           5563  139,5687

S = 0,1016   R-Sq = 58,89%   R-Sq(adj) = 58,86%


                              Individual 95% CIs For Mean Based on
                              Pooled StDev
Level     N    Mean   StDev  --------+---------+---------+---------+-
CO      465  0,7958  0,0941                            (*-)
N       449  0,6222  0,1194       (*)
NE     1794  0,5813  0,1016  (*
S      1188  0,8794  0,0820                                      (*)
SE     1668  0,8027  0,1105                          *)
                              --------+---------+---------+---------+-
                                    0,640     0,720     0,800     0,880

Pooled StDev = 0,1016
```

One-way ANOVA: IDHMnp versus REGIÃO CORRETA

```
Source            DF        SS       MS        F      P
REGIÃO CORRETA     4   82,3938  20,5985  1795,58  0,000
Error           5559   63,7714   0,0115
Total           5563  146,1652

S = 0,1071   R-Sq = 56,37%   R-Sq(adj) = 56,34%


                              Individual 95% CIs For Mean Based on
                              Pooled StDev
Level     N    Mean   StDev  --+---------+---------+---------+-------
CO      465  0,6108  0,0829                            (*-)
N       449  0,4278  0,1355       (*-)
NE     1794  0,3889  0,0975  (*
S      1188  0,6669  0,0937                                    *)
SE     1668  0,6328  0,1223                               (*)
                              --+---------+---------+---------+-------
                              0,400     0,480     0,560     0,640

Pooled StDev = 0,1071
```

One-way ANOVA: IDHM_Rnp versus REGIÃO CORRETA

```
Source            DF        SS       MS        F      P
REGIÃO CORRETA     4   90,9836  22,7459  2143,97  0,000
Error           5559   58,9768   0,0106
Total           5563  149,9605

S = 0,1030   R-Sq = 60,67%   R-Sq(adj) = 60,64%


                              Individual 95% CIs For Mean Based on Pooled StDev
Level     N    Mean   StDev  ---------+---------+---------+---------+
CO      465  0,5786  0,0857                               (*)
N       449  0,3927  0,1261       (*)
NE     1794  0,3305  0,0939  *)
S      1188  0,6384  0,0900                                        (*)
SE     1668  0,5726  0,1175                               (*
                              ---------+---------+---------+---------+
                                     0,400     0,480     0,560     0,640

Pooled StDev = 0,1030
```

Após esta análise chegou-se a classificação das regiões de acordo com as médias: NE (1); N (2); CO (3); SE (4); S (5). Neste momento é realizado a Regressão Logística Ordinal.

Ordinal Logistic Regression: Ordem das re versus MORT1_np; T_NESTUDA_NT; ...

Link Function: Logit

Response Information

Variable	Value	Count
Ordem das regiões	1	1794
	2	449
	3	465
	4	1668
	5	1188
	Total	5564

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio
Const(1)	19,2788	0,404218	47,69	0,000	
Const(2)	20,6062	0,421213	48,92	0,000	
Const(3)	21,6228	0,433010	49,94	0,000	
Const(4)	24,5006	0,465277	52,66	0,000	
MORT1_np	-38,0686	0,990763	-38,42	0,000	0,00
T_NESTUDA_NTRAB_MMEIO_np	-7,37356	0,445274	-16,56	0,000	0,00
ESPVIDAnp	17,1762	0,679741	25,27	0,000	28807840,69
IDHMnp	7,83558	0,640200	12,24	0,000	2528,99
IDHM_Rnp	-4,70879	0,678867	-6,94	0,000	0,01

	95% CI	
Predictor	Lower	Upper
Const(1)		
Const(2)		
Const(3)		
Const(4)		
MORT1_np	0,00	0,00
T_NESTUDA_NTRAB_MMEIO_np	0,00	0,00
ESPVIDAnp	7601764,98	1,09171E+08
IDHMnp	721,12	8869,24
IDHM_Rnp	0,00	0,03

Log-Likelihood = -4590,830
Test that all slopes are zero: G = 7135,598, DF = 5, **P-Value = 0,000**

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	100336	22247	0,000
Deviance	9182	22247	1,000

Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs	Number	Percent	Summary Measures	
Concordant	10609355	91,7	Somers' D	0,84
Discordant	936520	8,1	Goodman-Kruskal Gamma	0,84

```
Ties            18258      0,2  Kendall's Tau-a        0,63
Total        11564133    100,0
```

Destaca-se que esta análise é confiável, pois o valor de P foi de “0”. O modelo apresentou nível de concordância de 91,7% (acerto).

Foi aplicada também a análise de Regressão Logística Ordinal para os dados agrupados pela média dos Estados em ordem por região, no entanto, este não se mostrou confiável por causa do número de dados analisados serem muito baixos.

Ordinal Logistic Regression: ORDEM REGIÕE versus MORT1 MEDIA; ESPVIDA MEDI; ...

```
Link Function: Logit


Response Information

Variable       Value  Count
ORDEM REGIÕES  1          9
               2          7
               3          3
               4          4
               5          3
               Total     26


Logistic Regression Table

                                                                                  95%
                                                                                   CI
Predictor                   Coef  SE Coef       Z      P   Odds Ratio  Lower
Const(1)                 51,6529  22,7932    2,27  0,023
Const(2)                 67,5840  29,2076    2,31  0,021
Const(3)                 71,9625  29,9831    2,40  0,016
Const(4)                 75,7796  31,0832    2,44  0,015
MORT1 MEDIA             -48,8235  26,1111   -1,87  0,062         0,00    0,00
ESPVIDA MEDIA           -33,5987  24,7839   -1,36  0,175         0,00    0,00
IDHM_R MEDIA            -37,0794  52,2848   -0,71  0,478         0,00    0,00
T_NESTUDA_MMEIO MEDIA   -36,8682  48,2092   -0,76  0,444         0,00    0,00
IDHM MEDIA               65,4508  49,6282    1,32  0,187  2,66023E+28    0,00



Predictor                     Upper
Const(1)
Const(2)
Const(3)
Const(4)
MORT1 MEDIA                   10,52
ESPVIDA MEDIA           3194327,66
IDHM_R MEDIA           2,52060E+28
T_NESTUDA_MMEIO MEDIA  1,05702E+25
IDHM MEDIA             4,66184E+70


Log-Likelihood = -7,971
Test that all slopes are zero: G = 62,412, DF = 5, P-Value = 0,000
```

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	39,0239	95	1,000
Deviance	15,9429	95	1,000

Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs	Number	Percent	Summary Measures	
Concordant	252	98,4	Somers' D	0,97
Discordant	4	1,6	Goodman-Kruskal Gamma	0,97
Ties	0	0,0	Kendall's Tau-a	0,76
Total	256	100,0		

COMENTÁRIOS DA ANÁLISE

A tarefa da análise discriminante é encontrar a melhor função discriminante linear ou quadrática de um conjunto de variáveis que reproduza, tanto quanto possível, um agrupamento a priori de casos considerados.

Um procedimento em passos é utilizado nesse programa, e em cada passo a variável mais poderosa é introduzida na função discriminante. A função critério para selecionar a próxima variável depende do número de grupos especificados (o número de grupos varia de 2 a 20).

Quando o número de variáveis é maior do que dois, então o critério de seleção de variáveis é o traço do produto da matriz de covariância para as variáveis envolvidas e a matriz de covariância interclasse em um passo particular.

Os cálculos podem ser realizados em toda a população ou em amostra de dados ou mesmo em dados previamente agrupados.

Em nossas análises com as variáveis IDHMn, IDHM_Rn e ESPVIDAn, utilizamos a análise discriminante linear e conseguimos um resultado de 0,903 de proporção correta. Isto demonstra coerência na divisão em dois grupos. Além disso, é relevante ressaltar a similaridade destes grupos (municípios) com base nestas variáveis, levando em conta inclusive sua situação geográfica.

Na outra análise realizada com base no agrupamento apresentado no dendograma, onde pode ser percebido 4 “Brasis”, a proporcionalidade ficou em 100%.

16. ARVORE DE DECISÃO PELO SPSS

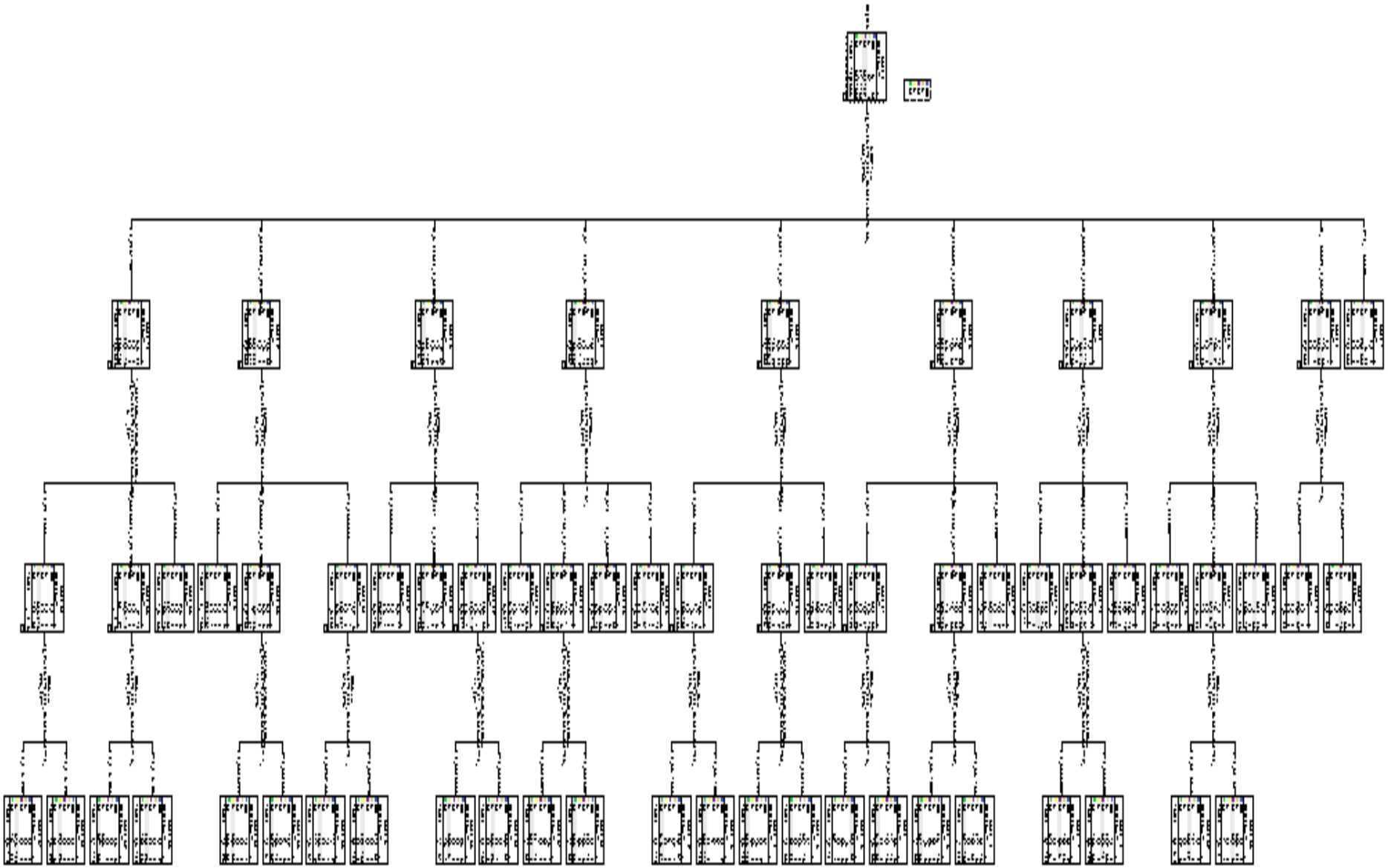
Classification Tree

Warnings

One or more values specified on the DEPCATEGORIES subcommand USEVALUES keyword do not exist in the training sample.
Gain summary Tables are not displayed because profits are undefined.
Target category gains tables are not displayed because target categories are undefined.

Model Summary

Specifications	Growing Method	CHAID
	Dependent Variable	Região
	Independent Variables	MORT1np, T_NESTUDA_NTRAB_MMEIOnp, ESPVIDAnp, IDHMnp, IDHM_Rnp
	Validation	None
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	MORT1np, ESPVIDAnp, IDHMnp, T_NESTUDA_NTRAB_MMEIOnp, IDHM_Rnp
	Number of Nodes	62
	Number of Terminal Nodes	40
	Depth	3



Risk

Estimate	Std. Error
,262	,006

Growing Method: CHAID

Dependent Variable: Região

Classification

Observed	Predicted					
	CO	N	NE	S	SE	Percent Correct
CO	52	0	55	208	150	11,2%
N	5	0	285	25	134	,0%
NE	6	0	1683	2	103	93,8%
S	0	0	0	1052	136	88,6%
SE	41	0	141	166	1320	79,1%
Overall Percentage	1,9%	,0%	38,9%	26,1%	33,1%	73,8%

Growing Method: CHAID

Dependent Variable: Região

CONSIDERAÇÕES FINAIS

A tarefa da análise discriminante é encontrar a melhor função discriminante linear ou quadrática de um conjunto de variáveis que reproduza, tanto quanto possível, um agrupamento a priori de casos considerados.

Um procedimento em passos é utilizado nesse programa, e em cada passo a variável mais poderosa é introduzida na função discriminante. A função critério para selecionar a próxima variável depende do número de grupos especificados (o número de grupos varia de 2 a 20).

Quando o número de variáveis é maior do que dois, então o critério de seleção de variáveis é o traço do produto da matriz de covariância para as variáveis envolvidas e a matriz de covariância interclasse em um passo particular.

Os cálculos podem ser realizados em toda a população ou em amostra de dados ou mesmo em dados previamente agrupados.

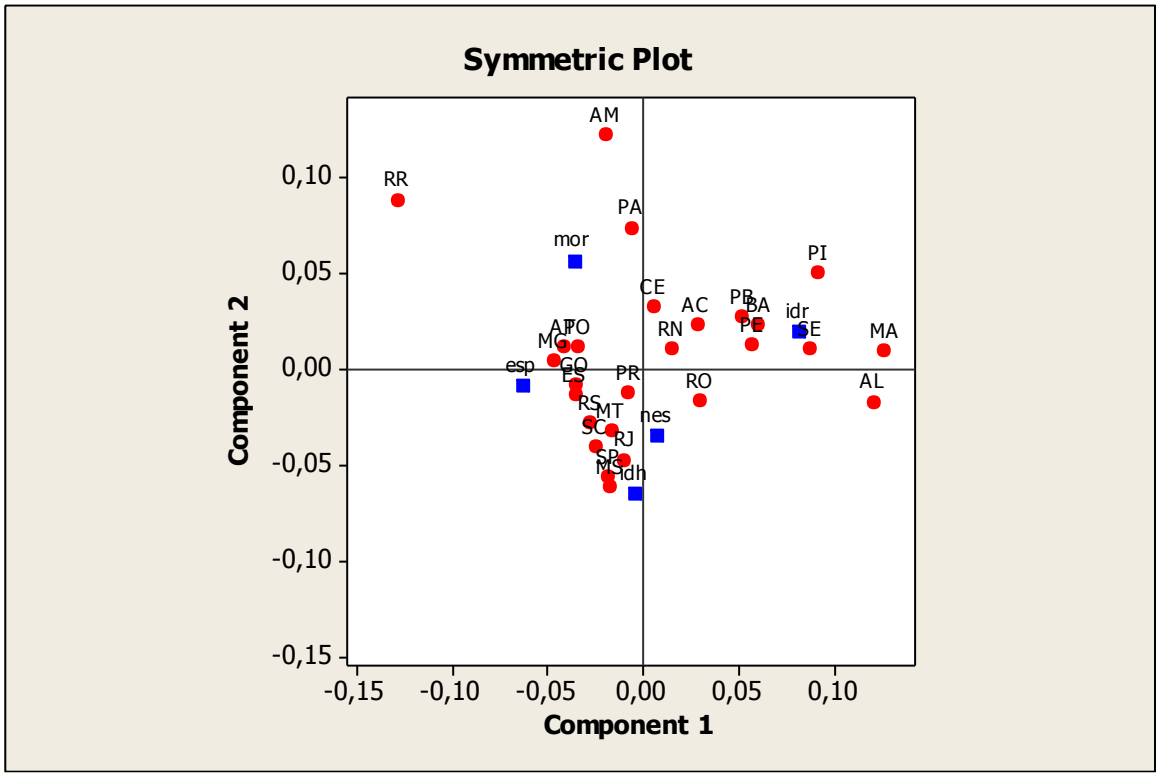
Em nossas análises com as variáveis **MORT1 x ESPVIDA x IDHM_R x T_NESTUDA_MMEIO x IDHM**, utilizamos a análise discriminante linear e conseguimos um resultado de 0,903 de proporção correta. Isto demonstra coerência na divisão em dois grupos. Além disso, é relevante ressaltar a similaridade destes grupos (municípios) com base nestas variáveis, levando em conta inclusive sua situação geográfica.

Na outra análise realizada com base no agrupamento apresentado no dendograma, onde pode ser percebido 4 “Brasis”, a proporcionalidade ficou em 100%.

17. ANÁLISE DE CORRESPONDÊNCIA DAS VARIÁVEIS

Nesta análise serão trabalhados os estados e as médias de educação por estado. Na análise de correspondência será gerado um mapa contendo quais estados estão mais próximos e quais variáveis tem a ver entre si. O comando para gerar o gráfico é:

STAT >> MULTIVARIATE >> SIMPLE CORRESPONDENCE ANALISYS



Nesta análise trabalhamos com as variáveis **MORT1 (mor); ESPVIDA (esp); IDHM_R (idr); T_NESTUDA_MMEIO (nes); IDHM (idh)**, representadas pelos quadrados azuis. Os círculos vermelhos são as médias das variáveis por estado, e podem ser identificados pela sigla.

Simple Correspondence Analysis: MORT1 MEDIA; ESPVIDA MEDI; IDHM_R MEDIA; T_NEST

Relative Inertias

	mor	esp	idh	idr	nes	Total
AC	0,000	0,001	0,000	0,007	0,007	0,015
AL	0,017	0,013	0,000	0,044	0,000	0,074
AM	0,041	0,000	0,026	0,002	0,023	0,093
AP	0,012	0,001	0,000	0,014	0,002	0,030
BA	0,000	0,007	0,001	0,017	0,000	0,026
CE	0,004	0,004	0,013	0,000	0,013	0,035
ES	0,000	0,005	0,001	0,008	0,000	0,014
GO	0,000	0,003	0,001	0,008	0,000	0,012
MA	0,013	0,013	0,009	0,051	0,006	0,092
MG	0,000	0,019	0,001	0,006	0,001	0,027
MS	0,018	0,010	0,013	0,003	0,000	0,045
MT	0,003	0,002	0,005	0,003	0,000	0,013
PA	0,012	0,001	0,008	0,003	0,018	0,042

PB	0,000	0,007	0,002	0,012	0,000	0,020
PE	0,000	0,012	0,000	0,010	0,000	0,023
PI	0,000	0,009	0,008	0,043	0,002	0,062
PR	0,000	0,000	0,002	0,001	0,000	0,003
RJ	0,003	0,001	0,013	0,006	0,005	0,028
RN	0,001	0,006	0,001	0,000	0,003	0,010
RO	0,001	0,006	0,010	0,003	0,002	0,022
RR	0,068	0,027	0,024	0,044	0,006	0,170
RS	0,001	0,005	0,007	0,006	0,000	0,019
SC	0,007	0,008	0,005	0,006	0,001	0,027
SE	0,003	0,015	0,000	0,029	0,000	0,048
SP	0,012	0,003	0,005	0,008	0,012	0,040
TO	0,003	0,001	0,001	0,006	0,001	0,012
Total	0,220	0,179	0,157	0,339	0,104	1,000

Analysis of Contingency Table

Axis	Inertia	Proportion	Cumulative	Histogram
1	0,0027	0,5182	0,5182	******************************
2	0,0018	0,3417	0,8598	******************
3	0,0005	0,0930	0,9529	*****
4	0,0002	0,0471	1,0000	**
Total	0,0052			

Row Contributions

					Component 1			Component 2		
ID	Name	Qual	Mass	Inert	Coord	Corr	Contr	Coord	Corr	Contr
1	AC	0,539	0,032	0,015	0,029	0,334	0,010	0,022	0,206	0,009
2	AL	0,984	0,025	0,074	0,121	0,963	0,137	-0,018	0,021	0,004
3	AM	0,966	0,030	0,093	-0,019	0,024	0,004	0,122	0,942	0,256
4	AP	0,448	0,038	0,030	-0,041	0,416	0,024	0,011	0,032	0,003
5	BA	0,995	0,031	0,026	0,061	0,873	0,043	0,023	0,121	0,009
6	CE	0,185	0,031	0,035	0,006	0,006	0,000	0,032	0,179	0,018
7	ES	0,975	0,048	0,014	-0,035	0,849	0,022	-0,014	0,126	0,005
8	GO	0,998	0,048	0,012	-0,035	0,938	0,022	-0,009	0,060	0,002
9	MA	0,867	0,026	0,092	0,127	0,863	0,154	0,009	0,004	0,001
10	MG	0,690	0,045	0,027	-0,046	0,685	0,036	0,004	0,005	0,000
11	MS	0,819	0,045	0,045	-0,017	0,056	0,005	-0,062	0,763	0,099
12	MT	0,949	0,046	0,013	-0,016	0,182	0,004	-0,033	0,766	0,028
13	PA	0,813	0,033	0,042	-0,006	0,005	0,000	0,073	0,808	0,100
14	PB	0,990	0,030	0,020	0,053	0,781	0,031	0,027	0,209	0,013
15	PE	0,905	0,031	0,023	0,057	0,864	0,038	0,012	0,041	0,003
16	PI	0,970	0,028	0,062	0,092	0,752	0,089	0,050	0,217	0,039
17	PR	0,674	0,049	0,003	-0,007	0,154	0,001	-0,013	0,520	0,005
18	RJ	0,799	0,048	0,028	-0,010	0,032	0,002	-0,048	0,768	0,063
19	RN	0,206	0,032	0,010	0,015	0,140	0,003	0,010	0,066	0,002
20	RO	0,430	0,041	0,022	0,030	0,327	0,014	-0,017	0,103	0,007
21	RR	0,994	0,036	0,170	-0,128	0,676	0,222	0,088	0,318	0,158
22	RS	0,826	0,052	0,019	-0,028	0,400	0,015	-0,028	0,426	0,024
23	SC	0,895	0,054	0,027	-0,024	0,228	0,012	-0,041	0,666	0,052
24	SE	0,997	0,031	0,048	0,088	0,985	0,090	0,010	0,012	0,002
25	SP	0,889	0,052	0,040	-0,018	0,080	0,006	-0,057	0,809	0,095
26	TO	0,749	0,038	0,012	-0,033	0,679	0,016	0,011	0,070	0,002

Column Contributions

					Component 1			Component 2		
ID	Name	Qual	Mass	Inert	Coord	Corr	Contr	Coord	Corr	Contr
1	mor	0,945	0,238	0,220	-0,036	0,275	0,117	0,057	0,669	0,431
2	esp	0,808	0,187	0,179	-0,063	0,793	0,275	-0,009	0,015	0,008
3	idh	0,847	0,160	0,157	-0,004	0,003	0,001	-0,065	0,844	0,388
4	idr	0,975	0,241	0,339	0,082	0,923	0,604	0,020	0,053	0,052
5	nes	0,410	0,175	0,104	0,007	0,016	0,003	-0,035	0,394	0,120

COMENTÁRIOS DA ANÁLISE

A análise de correspondência pode ser considerada como um caso especial da análise de componentes principais (TRABALHO número 7), porém dirigida a dados categóricos organizados em tabelas de contingência e não a dados contínuos. O problema é análogo a encontrar o maior componente principal de um conjunto de I observações e J variáveis, com modificações devido à ponderação das observações e à métrica ponderada.
Trata-se de um modelo de regressão para variáveis dependentes ou de resposta binomialmente distribuídas. É útil para modelar a probabilidade de um evento ocorrer como função de outros fatores.