PONTIFÍCIA UNIVERSIDADE CATÓLICA DE SÃO PAULO
FEA - Faculdade de Economia e Administração
Programa de Estudos Pós-Graduados em Administração

CLASSIFICAÇÃO DO BRASIL
Focando principalmente indicadores relacionados a habitação, trabalho, saúde e muito particularmente EDUCAÇÃO.

MÉTODOS QUANTITATIVOS NA PESQUISA EMPÍRICA

Professor: Dr. Arnoldo Jose de Hoyos

Clarice Santiago

CAP I ANÁLISE DE CONGLOMERADOS

1. INTRODUÇÃO

O presente trabalho tem por objetivo efetuar uma análise comparativa de médias, intervalos de confiança e regressões dos dados da Pesquisa Firjan/FGV sobre o Desenvolvimento dos Municípios nos períodos de 2000 e 2010. Iniciamos com o entendimento dos dados, incluindo a definição dos indivíduos e das variáveis, suas classificações em variáveis categóricas ou quantitativas, os significados e unidades de medida, além da apresentação da tabela de dados.

Na sequência serão geradas análise comparativas dos dados de Educação agrupado por Estado, excluindo o Distrito Federal por ter apenas um Município. Será calculada a Anova do ISDM e da Educação por Estado, serão gerados vários gráficos com as diversas variáveis de Educação. Comparando-se os resultados das médias por estado, poderemos agrupar as linhas de dados pelo nível de desigualdade dos fatores ISDM e Educação.

Por fim, fazemos as considerações finais. O software estatístico utilizado é o **MINITAB16.**

2. ENTENDENDO OS DADOS

2.1 Os Indivíduos

Esta pesquisa ilustra dois rankings lançados no final de 2012, e chegaram a conclusões diferentes sobre quais cidades de maior desenvolvimento do país.

Os indivíduos desta análise são os 5565 municípios brasileiros. Os dados analíticos foram extraídos do IBGE, e possibilitam uma comparação entre os dados colhidos em 2000 com 2010.

2.2 As Variáveis

As variáveis desta pesquisa incluem os 3 principais índices sintéticos que são ISDM, IFDM e IFDF, que são médias ponderadas dos dados analíticos globais da pesquisa, e variáveis analíticas, referente à educação do ensino pré escola, fundamental e médio. Esta pesquisa não utiliza variáveis do ensino superior.

Tabela 1. Comparativo entre as Variáveis ISDM e IFDM

O QUE O ISDM (FGV) MEDE	O QUE O IFDM (Firjan) MEDE
Educação: taxa de analfabetismo e taxa de crianças e jovens que frequentam a escola em cada etapa, desempenho na Prova Brasil (MEC)	**Educação:** taxa de matrícula infantil, abandono, distorção idade-série, desempenho no Ideb, taxa de docentes com ensino superior
Saúde e Segurança: taxa de mortalidade infantil, gravidez precoce e mortalidade por causas evitáveis; homicídios	**Saúde:** número de consultas pré-natal, óbitos por causa mal definidas e óbitos infantis evitáveis
Renda: presença de pobreza e extrema pobreza	**Emprego e renda:** geração, estoque e salários médios dos empregos formais
Trabalho: taxa de ocupação e formalização	
Habitação: coleta de lixo, energia elétrica, água canalizada, esgotamento sanitário, domicílio próprio	

Tabela 2. A definição das Variáveis

Variável	Significado	Tipo	Unidade de Medida
Município	Nome do Município	Texto	Na
Cód. IBGE	Código de referência do Município no IBGE	Numérico	Na
UF	Unidade da Federação	Texto	Na
ISDM	Índice Social de Desenvolvimento Municipal: Média ponderada dos indicadores das dimensões Habitação, Renda, Trabalho, Saúde e Segurança e Educação (H, R, T, S e E) padronizada pela média do Brasil.	Numérico	Percentual
IFDM	Índice Firjan de Desenvolvimento Municipal: Calculado pelo Firjan	Numérico	Percentual
IFGF	Índice Firjan de Gestão Fiscal	Numérico	Percentual
E1_1	Crianças de 0 a 3 anos que freqüentam creches	Numérico	Percentual
E1_2	Percentual de crianças de 4 a 6 anos que frequentam pré-escola.	Numérico	Percentual
E2_1 ...	Percentual de crianças de 8 ou 9 anos sabem ler e escrever.	Numérico	Percentual
E2_2	Crianças de 10 a 14 anos que sabem ler e escrever	Numérico	Percentual
E2_3	Percentual de crianças de 7 a 14 anos que frequentam escola.	Numérico	Percentual
E2_4	Percentual de crianças de 7 a 14 anos que estão na série correta segundo a idade	Numérico	Percentual
E2_5	Índice transformado na escala Ideb de proficiência Português e Matemática Agregado para a quarta série do Ensino Fundamental (5º ano EF)	Numérico	Percentual

E2_6	Índice transformado na escala Ideb de proficiência em Português e Matemática Agregado oitava série do Ensino Fundamental (9º ano EF).	Numérico	Percentual
E3_1	Percentual de jovens de 15 a 17 anos que frequentam escola.	Numérico	Percentual
E3_2	Percentual de jovens de 15 a 17 anos sabem ler e escrever.	Numérico	Percentual
E3_3	Pessoas com mais de 18 anos que sabem ler e escrever	Numérico	Percentual

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 barras.

3.1.1 Variável: “Estado”

Fazem parte desta pesquisa os 27 estados brasileiros e suas cidades. O gráfico abaixo exibe o número de cidades por estado.

A variação no número de cidades por estado é acentuada. Considerando que o Distrito Federal é um estado brasileiro, é o estado com o menor número de cidades (1), enquanto o Mato Grosso possui mais de 852 cidades.

3.1.2 Variável: “REGIÃO”

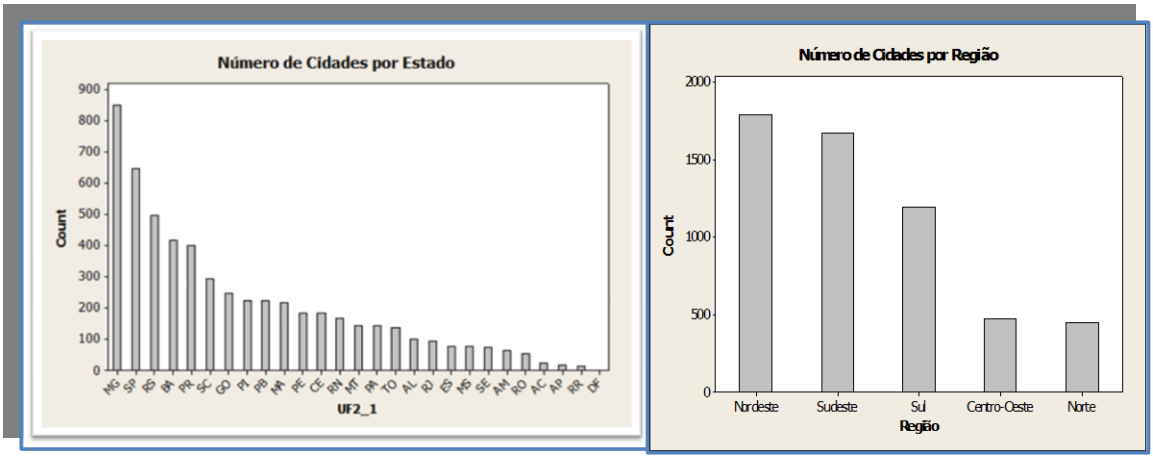


Figura 1. Número de Cidades por Estado e Região do Brasil

Podemos verificar no gráfico acima que a Região Nordeste é a que possui o maior número de cidades do Brasil (1790) e seguido pela Região Sudeste (1669). A Região que possui o menor

número de cidades é a Norte, com 447 cidades, muito próxima da Região Centro-Oeste (468). A Região Sul possui 1191 cidades.

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, gráfico de ramos, box-plot e dot-plot, além de informações numéricas como média, desvio-padrão, mediana, quartis, 5 números, intervalo de confiança e teste de normalidade de Anderson-Darling.

3.2.1. DENDOGRAMA DE EDUCAÇÃO POR ESTADO (-DF)

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

STAT >> MULTIVARIATE >> CLUSTER OBSERVATION

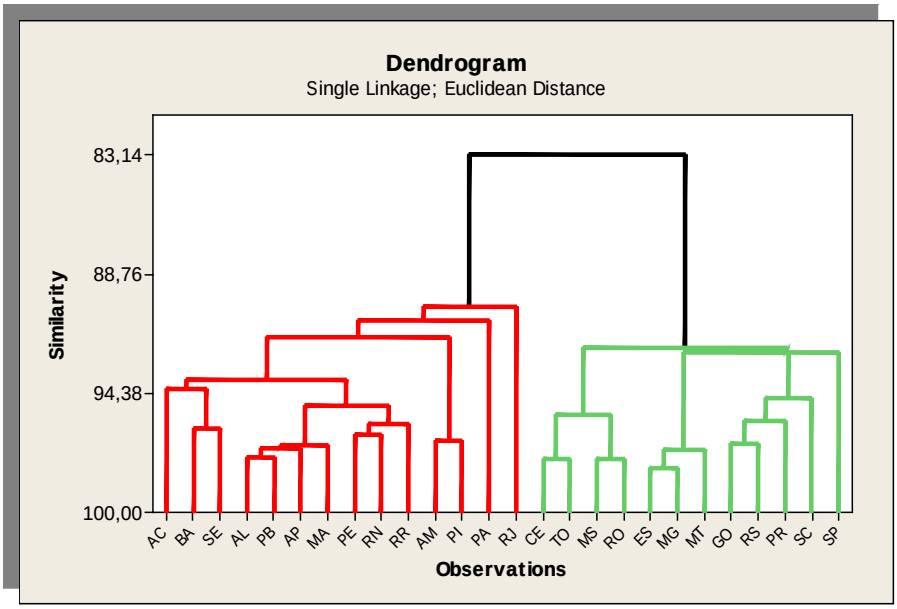


Figura 2. Dendograma da variável Educação por estados do Brasil (classificação não supervisionada)

Na figura acima podemos verificar dois grandes grupos de variáveis, agrupadas pela similaridade dos dados. Os estados que possuem maior similaridade são Alagoas e Paraiba no grupo vermelho e Espírito Santo e Minas Gerais no grupo verde. O nível de similaridade dos dados destes estados está por volta de 98%, conforme indicado na escala apresentada no exio Y do gráfico.

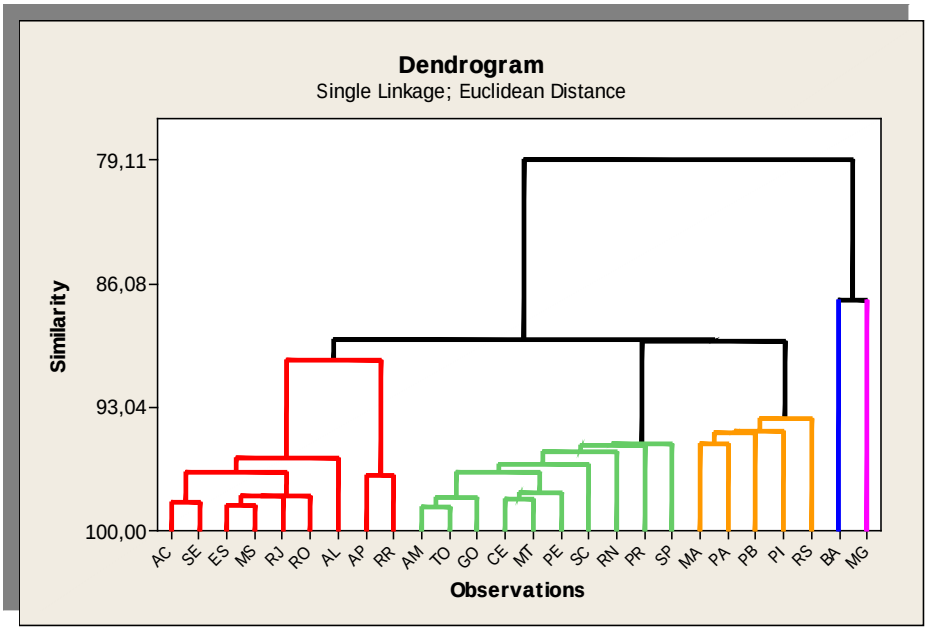


Figura 3. Dendrograma “Desiglalômetro” da Educação dos municípios por estado

No gráfico acima, podemos verificar 3 grandes agrupamento de dados, compostos pelos estados do Brasil, elem de dois estados que ficaram isolados por não ter seus dados em similaridade com os outros estados. Estes estados isolados são Bahia e Minas Gerais.

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

Podemos observar estados com 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 Espírito Santo e Mato Grosso do Sul, que tem um nível de similaridade próximo de 98%.

Quando o nível de desigualdade é baixo poderíamos erroneamente dizer que a situação é boa. **Isso não é verdade.** Baixa desigualdade não significa que as coisas vão bem, e sim que existe um padrão nos municípios do estado em termos de educação, uma maior similaridade entre este municípios, e não é possível responder se esta similaridade é boa ou não.

3.2.2. ANÁLISE DAS VARIANCIAS DE ISDM E EDUCAÇÃO POR ESTADO – DF

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.

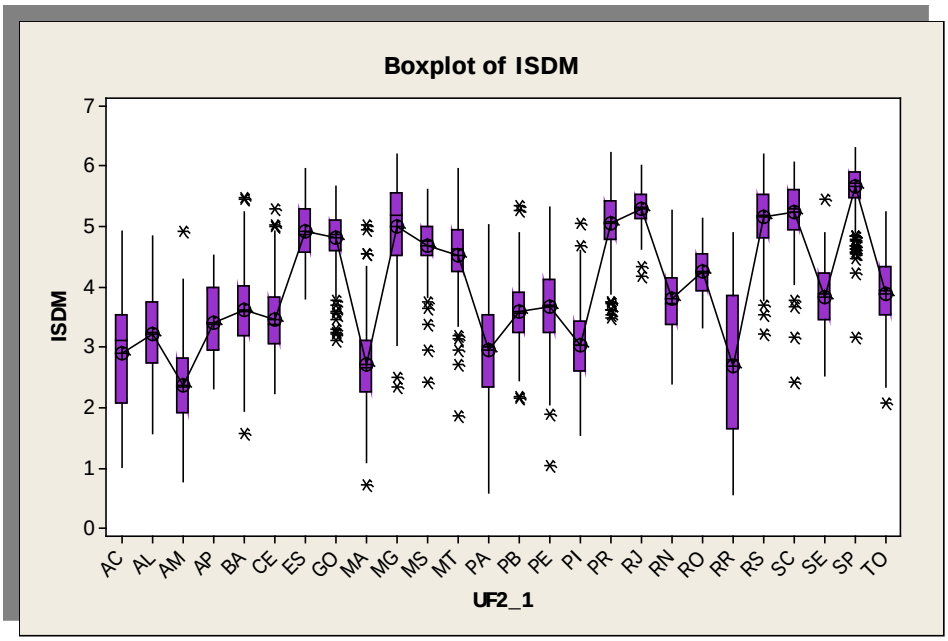


Figura 4. Gráfico BOXPLOT de ISDM por estado

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

Podemos visualizar no gráfico da figura4, uma grande variabilidade sobre as médias de ISDM por estado. O estado que apresenta maior variabilidade dos dados é Roraima. São Paulo apresenta uma baixa variabilidade dos dados de ISDM, 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.

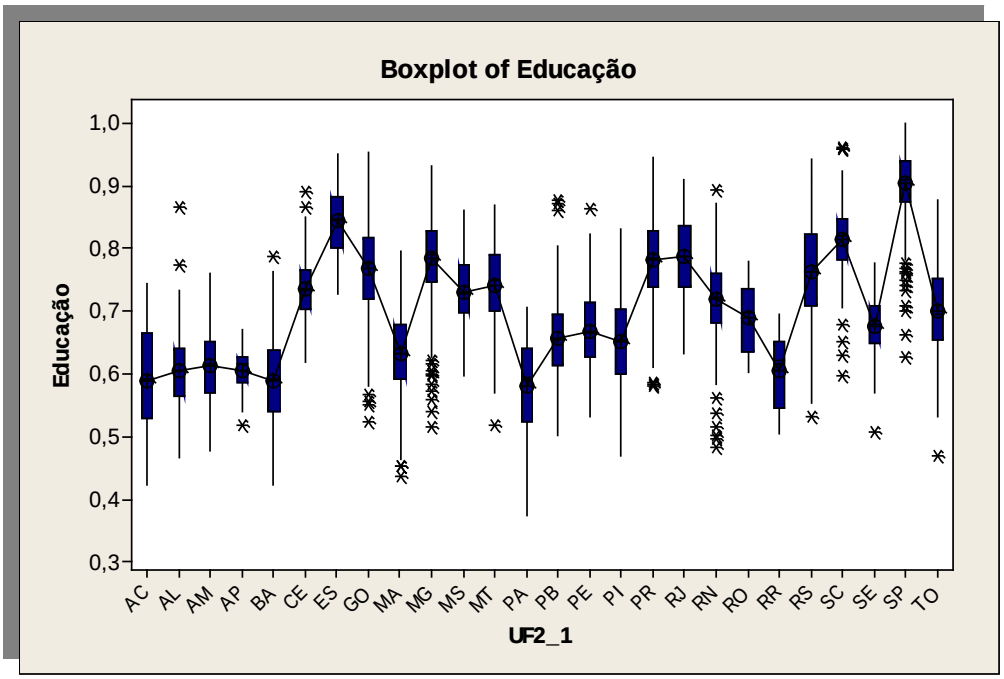


Figura 5. Gráfico BOXPLOT de Educação por estado

Podemos verificar que existe uma variação grande entre as médias dos estados do Brasil, no que diz respeito à educação. O tamanho das caixas de cada estado representa a variância dos dados de educação de cada estado, e os sinais * representam ou *outliers* ou pontos fora da curva, que são dados ou muito acima ou abaixo da média dos dados do estado. O estado que apresenta a maior média de ecucação é São Paulo (acima de 0,9), e o estado que apresenta a menor média é Para, com a média próxima a 0,6.

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

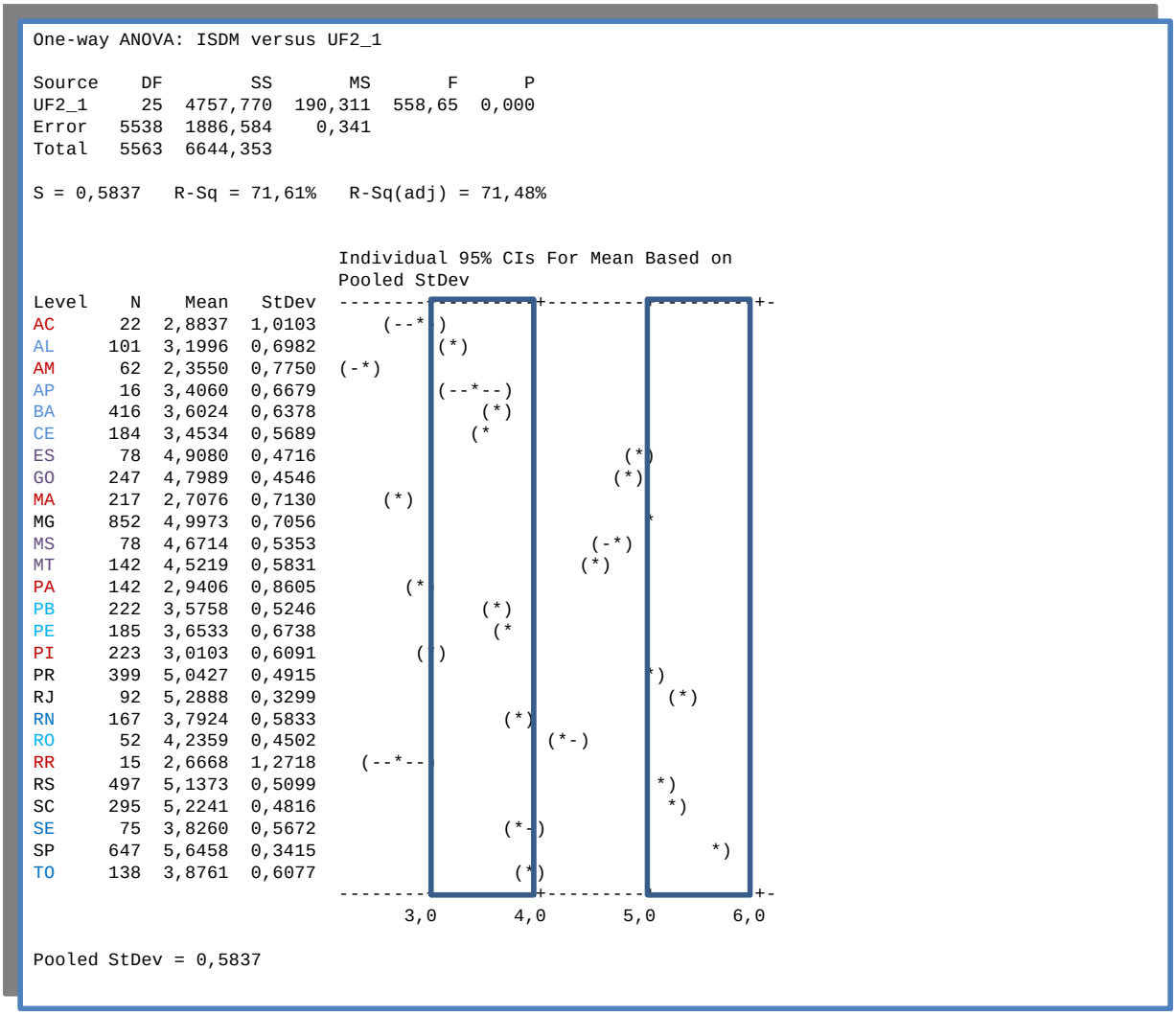


Figura 6. Análise de variância de ISDM por estado

É possível separar os dados das médias de cada estado por quartil. Desta forma teríamos 4 tipos de regiões no Brasil, separados pelos índices de Desenvolvimento de seus Municípios.

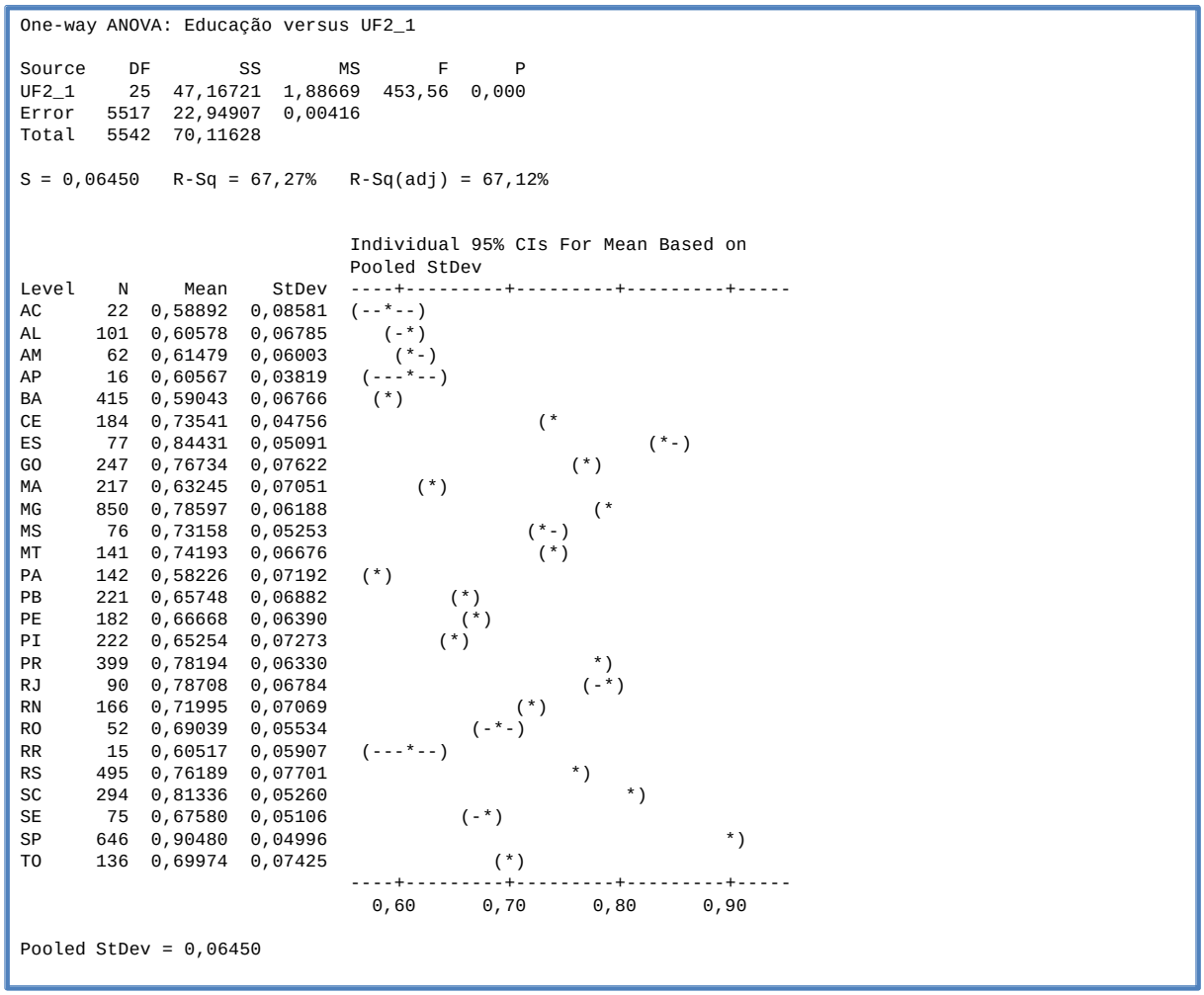


Figura 7. Análise de variância de Educação por estado

Podemos observar que alguns estados possuem alta variabilidade dos dados em relação à média, como Acre, Amapá e Roraima. Já outros tem o desvio padrão com menor variabilidade como Bahia e São Paulo.

Existe uma variação grande entre as médias de educação por estado, por exemplo o estado que apresenta a maior média é São Paulo, com 0,90480, e a menor média está com o Pará, com 0,582 seguido do Acre, com 0,588.

3.2.3. DENDOGRAMA DOS DADOS AGRUPADOS PELO RESULTADO DAS MÉDIAS

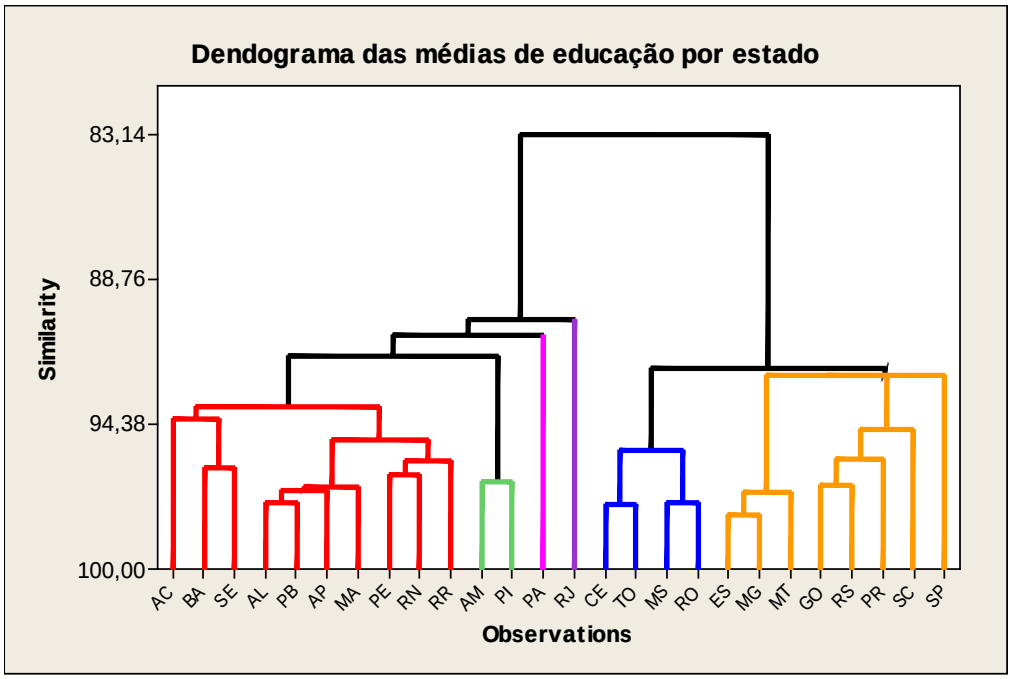
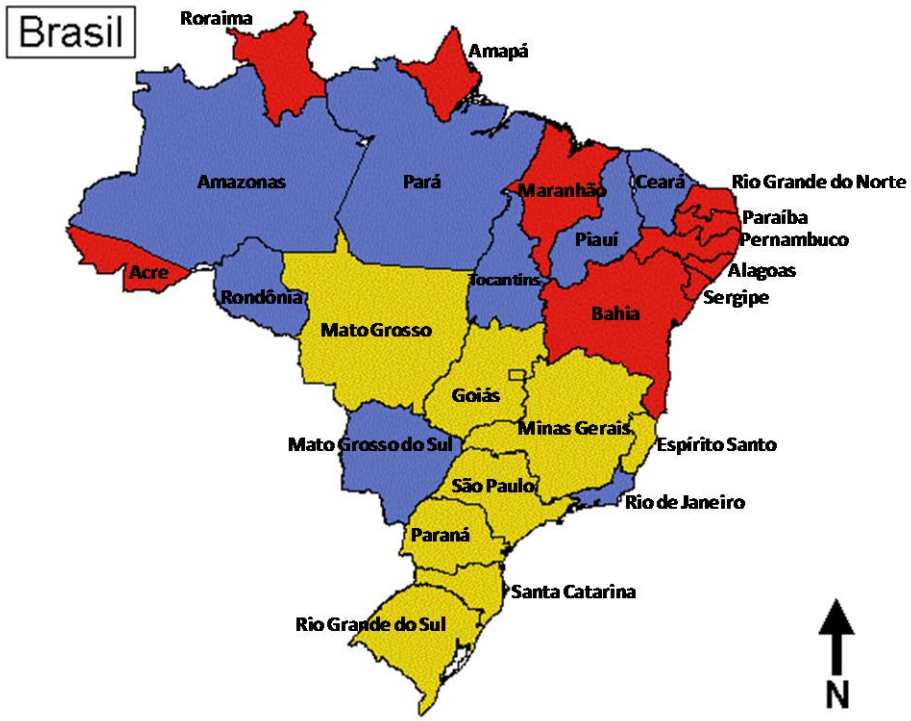


Figura 8. Dendograma dos dados das médias de ISDM, Educação E2_4, E2_5 e E2_6 dos municípios dos estados.

Podemos observar que existem dois grandes grupos de similaridade e podemos considerar um terceiro composto pelas cores verde, rosa, roxo e azul. Estes estado tem baixo nível de similaridade com os outros mais para efeito de análise iremos agrupá-los para maior entendimento da situação da educção nos municípios do Brasil. Podemos observar no mapa do Brasil os estados que tem média semelhante em educação.



Os 3 Brasis

3.2.5. CONSIDERAÇÕES FINAIS

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.

CAP II ANALISE DISCRIMINANTE

1. INTRODUÇÃO

A análise discriminante é uma técnica da estatística multivariada utilizada para discriminar e classificar objetos. É uma técnica da estatística multivariada que estuda a separação de objetos de uma população em duas ou mais classes. A discriminação ou separação é a primeira etapa, sendo a parte exploratória da análise e consiste em se procurar características capazes de serem utilizadas para alocar objetos em diferentes grupos previamente definidos. A classificação ou alocação pode ser definida como um conjunto de regras que serão usadas para alocar novos objetos.

O presente trabalho tem por objetivo efetuar uma análise comparativa de médias, intervalos de confiança e regressões de dados de indicadores relacionados ao desenvolvimento humano dos municípios do Brasil. Utilizamos a análise discriminante para tentar predizer ou explicar os indicadores relacionados ao desenvolvimento da educação dos municípios do Brasil.

Contudo, a função que separa objetos pode também servir para alocar, e o inverso, regras que alocam objetos podem ser usadas para separar. Normalmente, discriminação e classificação se sobrepõem na análise, e a distinção entre separação e alocação é confusa. O problema da discriminação entre dois ou mais grupos, visando posterior classificação consiste em obter funções matemáticas capazes de classificar um indivíduo X (uma observação X) em uma de várias populações, com base em medidas de um número p de características, buscando minimizar a probabilidade de má classificação.

Os dados são originários da pesquisa da FGV / FIRJAM sobre o desenvolvimento dos municípios do Brasil. Neste trabalho abordaremos as variáveis referentes à educação dos municípios. O software estatístico utilizado é o **MINITAB16.**

2. ENTENDENDO OS DADOS

2.1 Os Indivíduos

Esta pesquisa ilustra dois rankings lançados no final de 2012, e chegaram a conclusões diferentes sobre quais cidades de maior desenvolvimento do país.

Os indivíduos desta análise são os 5565 municípios brasileiros. Os dados analíticos foram extraídos do IBGE, e possibilitam uma comparação entre os dados colhidos em 2000 com 2010.

2.2 As Variáveis

As variáveis desta pesquisa incluem os 3 principais índices sintéticos que são ISDM, IFDM e IFDF, que são médias ponderadas dos dados analíticos globais da pesquisa, e variáveis analíticas, referente à educação do ensino pré escola, fundamental e médio. Esta pesquisa não utiliza variáveis do ensino superior.

Tabela 1. Comparativo entre as Variáveis ISDM e IFDM

O QUE O ISDM (FGV) MEDE	O QUE O IFDM (Firjan) MEDE
Educação: taxa de analfabetismo e taxa de crianças e jovens que frequentam a escola em cada etapa, desempenho na Prova Brasil (MEC)	**Educação:** taxa de matrícula infantil, abandono, distorção idade-série, desempenho no Ideb, taxa de docentes com ensino superior
Saúde e Segurança: taxa de mortalidade infantil, gravidez precoce e mortalidade por causas evitáveis; homicídios	**Saúde:** número de consultas pré-natal, óbitos por causa mal definidas e óbitos infantis evitáveis
Renda: presença de pobreza e extrema pobreza	**Emprego e renda:** geração, estoque e salários médios dos empregos formais
Trabalho: taxa de ocupação e formalização	
Habitação: coleta de lixo, energia elétrica, água canalizada, esgotamento sanitário, domicílio próprio	

Tabela 2. A definição das Variáveis

Variável	Significado	Tipo	Unidade de Medida
Município	Nome do Município	Texto	Na
Cód. IBGE	Código de referência do Município no IBGE	Numérico	Na
UF	Unidade da Federação	Texto	Na
ISDM	Índice Social de Desenvolvimento Municipal: Média ponderada dos indicadores das dimensões Habitação, Renda, Trabalho, Saúde e Segurança e Educação (H, R, T, S e E) padronizada pela média do Brasil.	Numérico	Percentual
IFDM	Índice Firjan de Desenvolvimento Municipal: Calculado pelo Firjan	Numérico	Percentual
IFGF	Índice Firjan de Gestão Fiscal	Numérico	Percentual
E1_1	Crianças de 0 a 3 anos que freqüentam creches	Numérico	Percentual
E1_2	Percentual de crianças de 4 a 6 anos que frequentam pré-escola.	Numérico	Percentual
E2_1 ...	Percentual de crianças de 8 ou 9 anos sabem ler e escrever.	Numérico	Percentual
E2_2	Crianças de 10 a 14 anos que sabem ler e escrever	Numérico	Percentual
E2_3	Percentual de crianças de 7 a 14 anos que frequentam escola.	Numérico	Percentual
E2_4	Percentual de crianças de 7 a 14 anos que estão na série correta segundo a idade	Numérico	Percentual
E2_5	Índice transformado na escala Ideb de proficiência Português e Matemática Agregado para a quarta série do Ensino Fundamental (5º ano EF)	Numérico	Percentual
E2_6	Índice transformado na escala Ideb de proficiência em Português e Matemática Agregado oitava série do Ensino Fundamental (9º ano EF).	Numérico	Percentual
E3_1	Percentual de jovens de 15 a 17 anos que frequentam escola.	Numérico	Percentual
E3_2	Percentual de jovens de 15 a 17 anos sabem ler e escrever.	Numérico	Percentual
E3_3	Pessoas com mais de 18 anos que sabem ler e escrever	Numérico	Percentual

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 barras.

3.1.1 Variável: “Estado”

Fazem parte desta pesquisa os 27 estados brasileiros e suas cidades. O gráfico abaixo exibe o número de cidades por estado.

A variação no número de cidades por estado é acentuada. Considerando que o Distrito Federal é um estado brasileiro, é o estado com o menor número de cidades (1), enquanto o Mato Grosso possui mais de 852 cidades.

3.1.2 Variável: “REGIÃO”

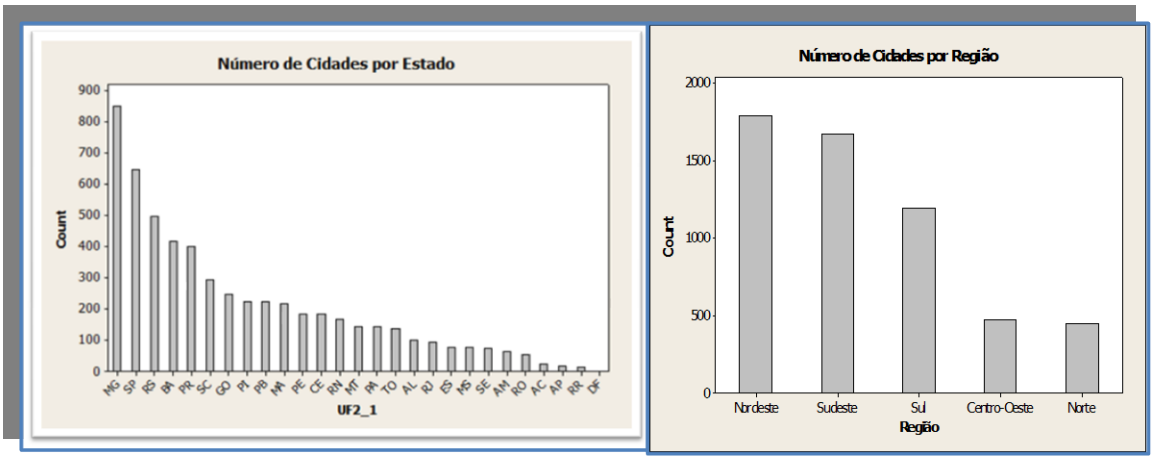


Figura 1. Número de Cidades por Estado e Região do Brasil

Podemos verificar no gráfico acima que a Região Nordeste é a que possui o maior número de cidades do Brasil (1790) e seguido pela Região Sudeste (1669). A Região que possui o menor número de cidades é a Norte, com 447 cidades, muito próxima da Região Centro-Oeste (468). A Região Sul possui 1191 cidades.

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, gráfico de ramos, box-plot e dot-plot, além de informações numéricas como média, desvio-padrão, mediana, quartis, 5 números, intervalo de confiança e teste de normalidade de Anderson-Darling. Também podemos fazer classificações supervisionadas das variáveis quantitativas, através da análise discriminante.

3.2.1. 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 de educação dos municípios 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 versus ISDM; Educação; E2_4; E2_5; E2_6

Linear Method for Response: Região

Predictors: ISDM; Educação; E2_4; E2_5; E2_6

Group	Centro-Oeste	Nordeste	Norte	Sudeste	Sul
Count	464	1783	445	1663	1188

5543 cases used, 21 cases contain missing values

Summary of classification

	True Group				
Put into Group	Centro-Oeste	Nordeste	Norte	Sudeste	Sul
Centro-Oeste	267	142	100	226	279
Nordeste	16	1132	124	17	6
Norte	28	485	205	46	8
Sudeste	63	13	2	1102	272
Sul	90	11	14	272	623
Total N	464	1783	445	1663	1188
N correct	267	1132	205	1102	623
Proportion	0,575	0,635	0,461	0,663	0,524

N = 5543 N Correct = 3329 Proportion Correct = 0,601

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

A região que acertou mais é Sudeste (0,663) e a que errou mais é o Norte (0,461). O gráfico exibe o cruzamento de dados entre as regiões. Por exemplo, a região Sudeste possui 1663 municípios e apenas 1102 correspondem a região, sendo que 272 são semelhantes aos dados da região Sul. 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. A % de acertos ficou em 60% para 5 Brasis utilizando nossas variáveis focando a Educação.

3.2.2. ANÁLISE DISCRIMINANTE LINEAR DOS MUNICIPIOS POR "3 BRASIS"

Esta segunda análise está interessada em verificar os possíveis agrupamento de dados utilizando a variável 3 Brasis, calculada no exercício anterior, e demonstra os agrupamentos do Brasil segundo sua proximidade de dados de educação.

```
Discriminant Analysis: 3 Brasis versus ISDM; Educação; E2_4; E2_5; E2_6

Linear Method for Response: 3 Brasis


Predictors: ISDM; Educação; E2_4; E2_5; E2_6


Group  Centro-Oeste   Nor    Su
Count           464  2228  2851


5543 cases used, 21 cases contain missing values


Summary of classification

                        True Group
Put into Group  Centro-Oeste    Nor     Su
Centro-Oeste             314    313    731
Nor                       35   1897     60
Su                       115     18   2060
Total N                  464   2228   2851
N correct                314   1897   2060
Proportion             0,677  0,851  0,723

N = 5543           N Correct = 4271           Proportion Correct = 0,771
```

Existem duas possibilidades análise discriminante que são a linear e a quadrática. Dependendo da variável deve-se dar mais peso e mais atenção a um método que outro. Neste caso utilizamos a linear. Podemos observar que alguns estados e municípios da região centro-oeste tem características das regiões Sul, visto pelo número 731 municípios foram encontrados na intersecção entre sul e centro-oeste.

No caso de 3 Brasis melhorou para 77% a percentagem de acertos no caso de Analise Discriminante Linear.

3.2.3. ANÁLISE DISCRIMINANTE QUADRÁTICA POR "3 BRASIS"

Uma boa classificação deve resultar em pequenos erros, isto é, deve haver pouca probabilidade de má classificação, e para que isso ocorra a regra de classificação deve considerar as probabilidades a priori e os custos de má classificação. 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 3 Brasis.

```
Discriminant Analysis: 3 Brasis versus ISDM; Educação; E2_4; E2_5; E2_6

Quadratic Method for Response:   3 Brasis


Predictors: ISDM; Educação; E2_4; E2_5; E2_6


Group  Centro-Oeste   Nor    Su
Count           464  2228  2851


5543 cases used, 21 cases contain missing values


Summary of classification

                         True Group
Put into Group  Centro-Oeste    Nor     Su
Centro-Oeste             344    254    703
Nor                       44   1948     89
Su                        76     26   2059
Total N                  464   2228   2851
N correct                344   1948   2059
Proportion             0,741  0,874  0,722

N = 5543          N Correct = 4351          Proportion Correct = 0,785
```

No modelo quadrático a proporção foi alterada em apenas 1% (de 0,77 para 0,78). Seguindo o pensamento da Parsimonia (simplicidade), vamos escolher o método linear pois é o mais simples.

Em Ciência, 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 a diferença entre o método linear e o quadrático é pequena e não justifica a utilização do método quadrático.

3.2.4. ANÁLISE DISCRIMINANTE LINEAR PARA DADOS AGRUPADOS POR ESTADO UTILIZANDO DISPARIDADES (sd)

Neste exemplo abaixo vamos através do dendograma pesquisar o grau de similaridade das variáveis de desvio padrão da educação nos municípios do Brasil. Com base na similaridade poderemos definimos agrupamento de dados e após utilizamos a análise discriminante para verificar a proporção correta dos agrupamentos.

```
Discriminant Analysis: 4 Brasis G versus ISDM sdn; Edu sdn; ...

Linear Method for Response: 4 Brasis G


Predictors: ISDM sdn; Edu sdn; E2_4 sdn; E2_5 sdn; E2_6 sdn


Group         B1         B2         B3         B4
Count          9         10          5          2


Summary of classification

                          True Group
Put into Group      B1      B2      B3      B4
B1                   9       0       0       0
B2                   0      10       0       0
B3                   0       0       5       0
B4                   0       0       0       2
Total N              9      10       5       2
N correct            9      10       5       2
Proportion       1,000   1,000   1,000   1,000

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

Neste caso a proporção correta é de 100%, ou seja, os agrupamentos gerados anteriormente pelo agrupamento em 4 Brasis gerou a mesma proporção do método linear utilizado na análise discriminante.

4. CONSIDERAÇÕES FINAIS

A tarefa da análise discriminante é encontrar a melhor função discriminante linear 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.

Nos nossos exemplos com as variáveis da educação, utilizamos a análise discriminante linear e conseguimos um resultado de 0,77 de proporção correta.

CAP III REGRESSÃO LOGÍSTICA

1. INTRODUÇÃO

A regressão logística é uma técnica estatística que tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica, frequentemente binária, a partir de uma série de variáveis explicativas contínuas e/ou binárias1 2. A regressão logística é amplamente usada em ciências médicas e sociais, e tem outras denominações, como modelo logístico e classificador de máxima entropia.

No domínio dos seguros, permite encontrar frações da clientela que sejam sensíveis a determinada política securitária em relação a um dado risco particular, em instituições financeiras, pode detectar os grupos de risco para a subscrição de um crédito e em econometria, permite explicar uma variável discreta, como por exemplo as intenções de voto em atos eleitorais.

O êxito da regressão logística assenta sobretudo nas numerosas ferramentas que permitem interpretar de modo aprofundado os resultados obtidos. Em comparação com as técnicas conhecidas em regressão, em especial a regressão linear, a regressão logística distingue-se essencialmente pelo fato de a variável resposta ser categórica.

Enquanto método de predição para variáveis categóricas, a regressão logística é comparável às técnicas supervisionadas propostas em aprendizagem automática (árvores de decisão, redes neuronais, etc.), ou ainda a análise discriminante preditiva em estatística exploratória. É possível de as colocar em concorrência para escolha do modelo mais adaptado para um certo problema preditivo a resolver.

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 factores. Os dados são originários da pesquisa da FGV / FIRJAM sobre o desenvolvimento dos municípios do Brasil. Neste trabalho abordaremos as variáveis referentes à educação dos municípios. O software estatístico utilizado é o **MINITAB16.**

2. ENTENDENDO OS DADOS

2.1 Os Indivíduos

Esta pesquisa ilustra dois rankings lançados no final de 2012, e chegaram a conclusões diferentes sobre quais cidades de maior desenvolvimento do país.

Os indivíduos desta análise são os 5565 municípios brasileiros. Os dados analíticos foram extraídos do IBGE, e possibilitam uma comparação entre os dados colhidos em 2000 com 2010.

2.2 As Variáveis

As variáveis desta pesquisa incluem os 3 principais índices sintéticos que são ISDM, IFDM e IFDF, que são médias ponderadas dos dados analíticos globais da pesquisa, e variáveis analíticas, referente à educação do ensino pré escola, fundamental e médio. Esta pesquisa não utiliza variáveis do ensino superior.

Tabela 1. Comparativo entre as Variáveis ISDM e IFDM

O QUE O ISDM (FGV) MEDE	O QUE O IFDM (Firjan) MEDE
Educação: taxa de analfabetismo e taxa de crianças e jovens que frequentam a escola em cada etapa, desempenho na Prova Brasil (MEC)	**Educação:** taxa de matrícula infantil, abandono, distorção idade-série, desempenho no Ideb, taxa de docentes com ensino superior
Saúde e Segurança: taxa de mortalidade infantil, gravidez precoce e mortalidade por causas evitáveis; homicídios	**Saúde:** número de consultas pré-natal, óbitos por causa mal definidas e óbitos infantis evitáveis
Renda: presença de pobreza e extrema pobreza	**Emprego e renda:** geração, estoque e salários médios dos empregos formais
Trabalho: taxa de ocupação e formalização	
Habitação: coleta de lixo, energia elétrica, água canalizada, esgotamento sanitário, domicílio próprio	

Tabela 2. A definição das Variáveis

Variável	Significado	Tipo	Unidade de Medida
Município	Nome do Município	Texto	Na
Cód. IBGE	Código de referência do Município no IBGE	Numérico	Na
UF	Unidade da Federação	Texto	Na
ISDM	Índice Social de Desenvolvimento Municipal: Média ponderada dos indicadores das dimensões Habitação, Renda, Trabalho, Saúde e Segurança e Educação (H, R, T, S e E) padronizada pela média do Brasil.	Numérico	Percentual
IFDM	Índice Firjan de Desenvolvimento Municipal: Calculado pelo Firjan	Numérico	Percentual
IFGF	Índice Firjan de Gestão Fiscal	Numérico	Percentual
E1_1	Crianças de 0 a 3 anos que freqüentam creches	Numérico	Percentual
E1_2	Percentual de crianças de 4 a 6 anos que frequentam pré-escola.	Numérico	Percentual
E2_1 ...	Percentual de crianças de 8 ou 9 anos sabem ler e escrever.	Numérico	Percentual
E2_2	Crianças de 10 a 14 anos que sabem ler e escrever	Numérico	Percentual
E2_3	Percentual de crianças de 7 a 14 anos que frequentam escola.	Numérico	Percentual
E2_4	Percentual de crianças de 7 a 14 anos que estão na série correta segundo a idade	Numérico	Percentual
E2_5	Índice transformado na escala Ideb de proficiência Português e Matemática Agregado para a quarta série do Ensino Fundamental (5º ano EF)	Numérico	Percentual
E2_6	Índice transformado na escala Ideb de proficiência em Português e Matemática Agregado oitava série do Ensino Fundamental (9º ano EF).	Numérico	Percentual
E3_1	Percentual de jovens de 15 a 17 anos que frequentam escola.	Numérico	Percentual
E3_2	Percentual de jovens de 15 a 17 anos sabem ler e escrever.	Numérico	Percentual
E3_3	Pessoas com mais de 18 anos que sabem ler e escrever	Numérico	Percentual

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 barras.

3.1.1 Variável: “Estado”

Fazem parte desta pesquisa os 27 estados brasileiros e suas cidades. O gráfico abaixo exibe o número de cidades por estado.

A variação no número de cidades por estado é acentuada. Considerando que o Distrito Federal é um estado brasileiro, é o estado com o menor número de cidades (1), enquanto o Mato Grosso possui mais de 852 cidades.

3.1.2 Variável: “REGIÃO”

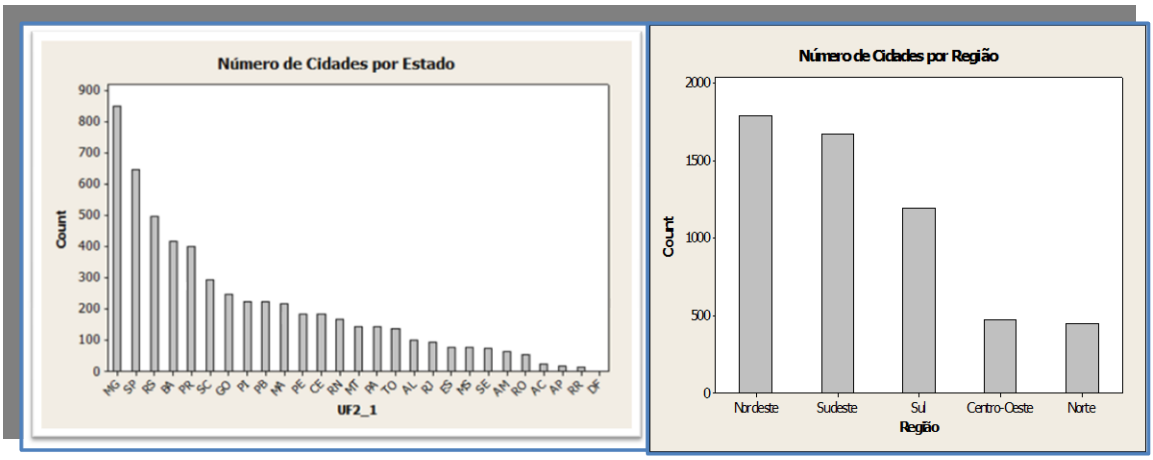


Figura 1. Número de Cidades por Estado e Região do Brasil

Podemos verificar no gráfico acima que a Região Nordeste é a que possui o maior número de cidades do Brasil (1790) e seguido pela Região Sudeste (1669). A Região que possui o menor número de cidades é a Norte, com 447 cidades, muito próxima da Região Centro-Oeste (468). A Região Sul possui 1191 cidades.

Nominal Logistic Regression: Região versus ISDM; Educação; ...

Response Information

Variable	Value	Count	
Região	Sul	1188	(Reference Event)
	Sudeste	1663	
	Norte	445	
	Nordeste	1783	
	Centro-Oeste	464	
	Total	5543	

* NOTE * 5543 cases were used
* NOTE * 21 cases contained missing values

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio
Logit 1: (Sudeste/Sul)					
Constant	8,78885	1,07818	8,15	0,000	
ISDM	-0,296517	0,0984560	-3,01	0,003	0,74
Educação	16,6786	0,894744	18,64	0,000	17514843,94
E2_4	-0,223293	0,0129363	-17,26	0,000	0,80
E2_5	1,06238	0,100491	10,57	0,000	2,89
E2_6	-1,20109	0,119227	-10,07	0,000	0,30
Logit 2: (Norte/Sul)					
Constant	41,3895	1,72849	23,95	0,000	
ISDM	-2,45494	0,146345	-16,78	0,000	0,09
Educação	6,63188	1,45260	4,57	0,000	758,91
E2_4	-0,270238	0,0182621	-14,80	0,000	0,76
E2_5	-0,937882	0,194188	-4,83	0,000	0,39
E2_6	-1,78640	0,229141	-7,80	0,000	0,17
Logit 3: (Nordeste/Sul)					
Constant	51,9835	1,68667	30,82	0,000	
ISDM	-1,94930	0,133440	-14,61	0,000	0,14
Educação	14,8119	1,31942	11,23	0,000	2708440,36
E2_4	-0,342867	0,0172735	-19,85	0,000	0,71
E2_5	-2,11373	0,178716	-11,83	0,000	0,12
E2_6	-3,03067	0,208939	-14,51	0,000	0,05
Logit 4: (Centro-Oeste/Sul)					
Constant	18,8337	1,37949	13,65	0,000	
ISDM	-0,879542	0,120406	-7,30	0,000	0,41
Educação	7,32685	1,11881	6,55	0,000	1520,58
E2_4	-0,118848	0,0161211	-7,37	0,000	0,89
E2_5	-0,268781	0,140343	-1,92	0,055	0,76
E2_6	-1,84735	0,167916	-11,00	0,000	0,16

	95% CI	
Predictor	Lower	Upper
Logit 1: (Sudeste/Sul)		
Constant		
ISDM	0,61	0,90
Educação	3032487,68	1,01161E+08
E2_4	0,78	0,82
E2_5	2,38	3,52
E2_6	0,24	0,38
Logit 2: (Norte/Sul)		
Constant		
ISDM	0,06	0,11
Educação	44,03	13081,05
E2_4	0,74	0,79
E2_5	0,27	0,57
E2_6	0,11	0,26
Logit 3: (Nordeste/Sul)		
Constant		
ISDM	0,11	0,18
Educação	203997,53	35959500,56
E2_4	0,69	0,73
E2_5	0,09	0,17
E2_6	0,03	0,07
Logit 4: (Centro-Oeste/Sul)		
Constant		
ISDM	0,33	0,53
Educação	169,70	13625,24
E2_4	0,86	0,92
E2_5	0,58	1,01
E2_6	0,11	0,22

Log-Likelihood = -4505,276

Test that all slopes are zero: G = 7244,614, DF = 20, P-Value = 0,000

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	360728	22148	0,000
Deviance	9011	22148	1,000

O Algoritmo de Regressão Logística não convergiu devido possivelmente a grande a grande variabilidade dos dados.

4. CONSIDERAÇÕES FINAIS

Enquanto método de predição para variáveis categóricas, a regressão logística é comparável às técnicas supervisionadas propostas em aprendizagem automática (árvores de decisão, redes neuronais, etc.), ou ainda a análise discriminante preditiva em estatística exploratória. É possível de as colocar em concorrência para escolha do modelo mais adaptado para um certo problema preditivo a resolver. **No entanto neste caso utilizando 5 Brasis a Regressão Logística não convergiu.**

CAP IV ÁRVORES DE CLASSIFICAÇÃO

1. INTRODUÇÃO

Nascida na década de 1960, a técnica árvore de classificação alcançou o segmento de negócios através da utilização em pesquisas de mercado. Tendo como pontos fortes a simplicidade de sua representação gráfica baseado em árvores e a facilidade de entender as regras e perfis derivados de cada segmento (nós), rapidamente foi adotada por outras áreas de marketing, sobretudo aquelas formadas por gestores com menor grau de sofisticação analítica.

Um dos vários algoritmos criados (CHAID) tornou-se popular em marketing direto, sobretudo para selecionar grupos de consumidores e prever a taxa de resposta de uma campanha em função do perfil determinado pelo algoritmo.

Os dados são originários da pesquisa da FGV / FIRJAM sobre o desenvolvimento dos municípios do Brasil. Neste trabalho abordaremos as variáveis referentes à educação dos municípios. O software estatístico utilizado é o **SPSS21.**

2. ENTENDENDO OS DADOS

2.1 Os Indivíduos

Esta pesquisa ilustra dois rankings lançados no final de 2012, e chegaram a conclusões diferentes sobre quais cidades de maior desenvolvimento do país.

Os indivíduos desta análise são os 5565 municípios brasileiros. Os dados analíticos foram extraídos do IBGE, e possibilitam uma comparação entre os dados colhidos em 2000 com 2010.

2.2 As Variáveis

As variáveis desta pesquisa incluem os 3 principais índices sintéticos que são ISDM, IFDM e IFDF, que são médias ponderadas dos dados analíticos globais da pesquisa, e variáveis analíticas, referente à educação do ensino pré escola, fundamental e médio. Esta pesquisa não utiliza variáveis do ensino superior.

Tabela 1. Comparativo entre as Variáveis ISDM e IFDM

O QUE O ISDM (FGV) MEDE	O QUE O IFDM (Firjan) MEDE
Educação: taxa de analfabetismo e taxa de crianças e jovens que frequentam a escola em cada etapa, desempenho na Prova Brasil (MEC)	**Educação:** taxa de matrícula infantil, abandono, distorção idade-série, desempenho no Ideb, taxa de docentes com ensino superior
Saúde e Segurança: taxa de mortalidade infantil, gravidez precoce e mortalidade por causas evitáveis; homicídios	**Saúde:** número de consultas pré-natal, óbitos por causa mal definidas e óbitos infantis evitáveis
Renda: presença de pobreza e extrema pobreza	**Emprego e renda:** geração, estoque e salários médios dos empregos formais
Trabalho: taxa de ocupação e formalização	
Habitação: coleta de lixo, energia elétrica, água canalizada, esgotamento sanitário, domicílio próprio	

Tabela 2. A definição das Variáveis

Variável	Significado	Tipo	Unidade de Medida
Município	Nome do Município	Texto	Na
Cód. IBGE	Código de referência do Município no IBGE	Numérico	Na
UF	Unidade da Federação	Texto	Na
ISDM	Índice Social de Desenvolvimento Municipal: Média ponderada dos indicadores das dimensões Habitação, Renda, Trabalho, Saúde e Segurança e Educação (H, R, T, S e E) padronizada pela média do Brasil.	Numérico	Percentual
IFDM	Índice Firjan de Desenvolvimento Municipal: Calculado pelo Firjan	Numérico	Percentual
IFGF	Índice Firjan de Gestão Fiscal	Numérico	Percentual
E1_1	Crianças de 0 a 3 anos que freqüentam creches	Numérico	Percentual
E1_2	Percentual de crianças de 4 a 6 anos que frequentam pré-escola.	Numérico	Percentual
E2_1 ...	Percentual de crianças de 8 ou 9 anos sabem ler e escrever.	Numérico	Percentual
E2_2	Crianças de 10 a 14 anos que sabem ler e escrever	Numérico	Percentual
E2_3	Percentual de crianças de 7 a 14 anos que frequentam escola.	Numérico	Percentual
E2_4	Percentual de crianças de 7 a 14 anos que estão na série correta segundo a idade	Numérico	Percentual
E2_5	Índice transformado na escala Ideb de proficiência Português e Matemática Agregado para a quarta série do Ensino Fundamental (5º ano EF)	Numérico	Percentual

E2_6	Índice transformado na escala Ideb de proficiência em Português e Matemática Agregado oitava série do Ensino Fundamental (9º ano EF).	Numérico	Percentual
E3_1	Percentual de jovens de 15 a 17 anos que frequentam escola.	Numérico	Percentual
E3_2	Percentual de jovens de 15 a 17 anos sabem ler e escrever.	Numérico	Percentual
E3_3	Pessoas com mais de 18 anos que sabem ler e escrever	Numérico	Percentual

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 barras.

3.1.1 Variável: “Estado”

Fazem parte desta pesquisa os 27 estados brasileiros e suas cidades. O gráfico abaixo exibe o número de cidades por estado.

A variação no número de cidades por estado é acentuada. Considerando que o Distrito Federal é um estado brasileiro, é o estado com o menor número de cidades (1), enquanto o Mato Grosso possui mais de 852 cidades.

3.1.2 Variável: “REGIÃO”

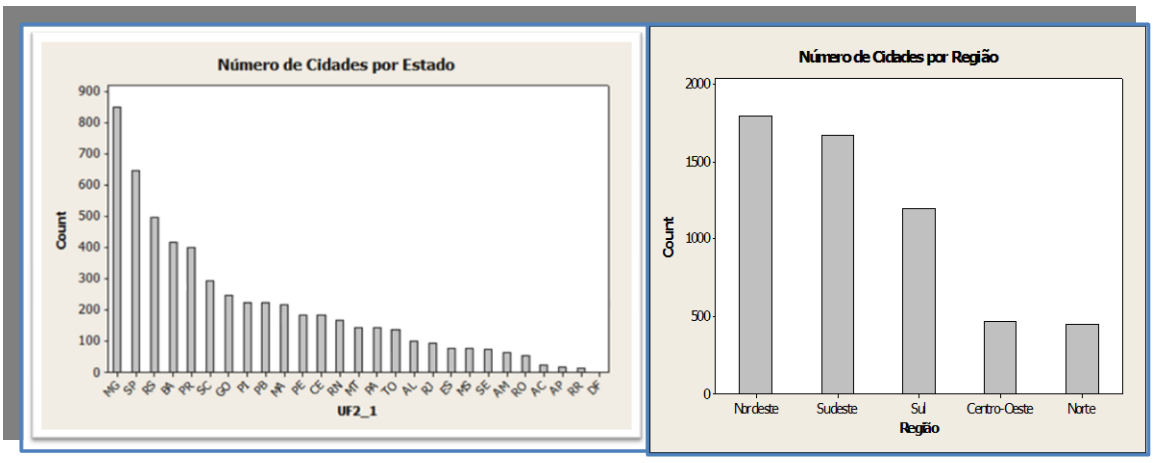


Figura 1. Número de Cidades por Estado e Região do Brasil

Podemos verificar no gráfico acima que a Região Nordeste é a que possui o maior número de cidades do Brasil (1790) e seguido pela Região Sudeste (1669). A Região que possui o menor número de cidades é a Norte, com 447 cidades, muito próxima da Região Centro-Oeste (468). A Região Sul possui 1191 cidades.

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, gráfico de ramos, box-plot e dot-plot, além de informações numéricas como média, desvio-padrão, mediana, quartis, 5 números, intervalo de confiança e teste de normalidade de Anderson-Darling. Também podemos fazer classificações supervisionadas das variáveis quantitativas, através da análise discriminante.

3.2.1. ÁRVORES DE CLASSIFICAÇÃO DAS VARIÁVEIS DE EDUCAÇÃO

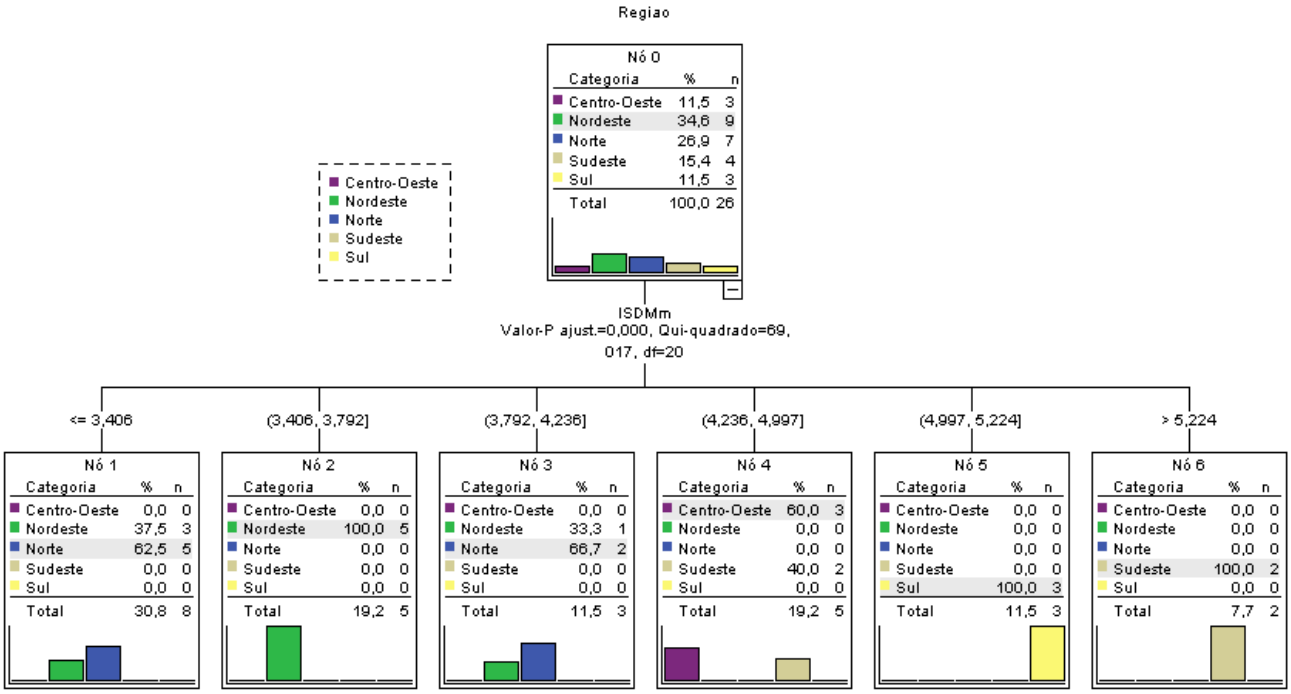
COMANDO SPSS:

ANALISAR >> CLASSIFICAR >> ARVORE

Este resultado se refere à variável dependente REGIAO e as medias por estado referentes as variáveis: ISDM, EduM, E24M, E25M e E26M.

Resumo do modelo

Especificações	Método de crescimento	CHAID
	Variável dependente	Regiao
	Variáveis independentes	ISDMm, Edum, E24m, E25m, E26m
	Validação	Nenhum
	Profundidade de árvore máxima	3
	Casos mínimos em nó pai	2
	Casos mínimos em nó filho	1
Resultados	Variáveis independentes incluídas	ISDMm
	Número de nós	7
	Número de nós de terminal	6
	Profundidade	1



Risco

Estimativas	Modelo padrão
,231	,083

Método de crescimento: CHAID
Variável dependente: Regiao

Posto

Observado	Previsto					
	Centro-Oeste	Nordeste	Norte	Sudeste	Sul	Porcentagem Correta
Centro-Oeste	3	0	0	0	0	100,0%
Nordeste	0	5	4	0	0	55,6%
Norte	0	0	7	0	0	100,0%
Sudeste	2	0	0	2	0	50,0%
Sul	0	0	0	0	3	100,0%
Porcentagem global	19,2%	19,2%	42,3%	7,7%	11,5%	76,9%

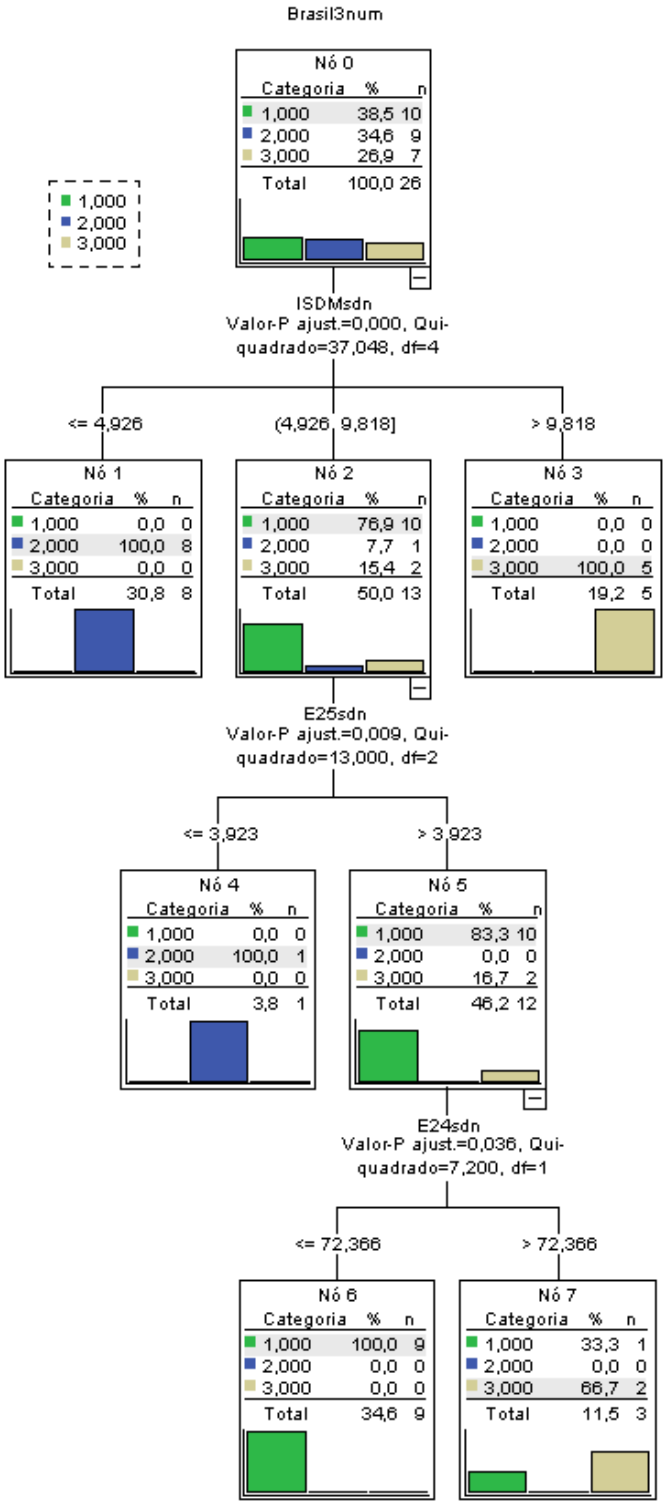
Método de crescimento: CHAID
Variável dependente: Regiao

Esta % de acertos de 76,9% se refere aos 5 Brasis e as medias por Estado das variáveis : ISDM, Edu, E24, E25 e E26.

A Continuação a arvore referente as disparidades por estado

Resumo do modelo

Especificações	Método de crescimento	CHAID
	Variável dependente	Brasil3num
	Variáveis independentes	ISDMsdn, Edusdn, E24sdn, E25sdn, E26sdn
	Validação	Nenhum
	Profundidade de árvore máxima	3
	Casos mínimos em nó pai	2
	Casos mínimos em nó filho	1
Resultados	Variáveis independentes incluídas	ISDMsdn, E25sdn, E24sdn
	Número de nós	8
	Número de nós de terminal	5
	Profundidade	3


Brasil3num
1,000
2,000
3,000
Nó 0
Categoria % n
1,000 38,5 10
2,000 34,6 9
3,000 26,9 7
Total 100,0 26
ISDMsdn
Valor-P ajust.=0,000, Qui-quadrado=37,048, df=4
<= 4,926
(4,926, 9,818]
> 9,818
Nó 1
Categoria % n
1,000 0,0 0
2,000 100,0 8
3,000 0,0 0
Total 30,8 8
Nó 2
Categoria % n
1,000 76,9 10
2,000 7,7 1
3,000 15,4 2
Total 50,0 13
Nó 3
Categoria % n
1,000 0,0 0
2,000 0,0 0
3,000 100,0 5
Total 19,2 5
E25sdn
Valor-P ajust.=0,009, Qui-quadrado=13,000, df=2
<= 3,923
> 3,923
Nó 4
Categoria % n
1,000 0,0 0
2,000 100,0 1
3,000 0,0 0
Total 3,8 1
Nó 5
Categoria % n
1,000 83,3 10
2,000 0,0 0
3,000 16,7 2
Total 46,2 12
E24sdn
Valor-P ajust.=0,036, Qui-quadrado=7,200, df=1
<= 72,366
> 72,366
Nó 6
Categoria % n
1,000 100,0 9
2,000 0,0 0
3,000 0,0 0
Total 34,6 9
Nó 7
Categoria % n
1,000 33,3 1
2,000 0,0 0
3,000 66,7 2
Total 11,5 3

Risco

Estimativas	Modelo padrão
,038	,038

Método de crescimento: CHAID
Variável dependente: Brasil3num

Posto

Observado	Previsto			
	1,00	2,00	3,00	Porcentagem Correta
1,00	9	0	1	90,0%
2,00	0	9	0	100,0%
3,00	0	0	7	100,0%
Porcentagem global	34,6%	34,6%	30,8%	96,2%

Método de crescimento: CHAID
Variável dependente: Brasil3num

Esta alta % de acertos para 96,2% se refere a 3 Brasis e utilizando as disparidade por estado das variáveis : ISDM, Edu, E24, E25 e E26.