**Universidade de São Paulo**

# Estudo Quantitativo de Desempenho no ENEM

**São Paulo, Abril de 2009**

# Instituto de Matemática e Estatística - USP

Título: **Estudo Quantitativo de Desempenho no ENEM**

Aluna: **Fanny Almeida Marcondes Nogueira**

Instituição: **Instituto de Matemática e Estatística (IME-USP)**

Curso: **Bacharelado em Matemática Aplicada e Computacional**

Finalidade: **Obtenção do título de Bacharel em Matemática Aplicada e Computacional com Habilitação em Saúde Animal**

Orientadores: **Cláudia Monteiro Peixoto e Sérgio Muniz Oliva Filho**

**Agradecimentos**

**Primeiramente agradeço a Deus por ter me dado força para chegar até aqui.**

Aos meus queridos pais pelo apoio e compreensão em todos os momentos dessa longa trajetória que só está começando. Aos meus orientadores Cláudia e Sérgio pelo apoio que me deram para realização desse trabalho. À professora Sônia que agüentou todos os meus momentos de fraqueza e vontade de desistir da graduação, me ajudou e me aconselhou como uma segunda mãe no IME.

Enfim agradeço a todos que diretamente ou indiretamente me ajudaram.

# Índice

## RESUMO

A proposta desse trabalho é analisar estatisticamente os dados do ENEM (Exame Nacional do Ensino Médio – Brasil) coletando informações para avaliar os fatores associados ao desempenho dos alunos.

A nossa base de dados é composta por 3 anos de prova ,donde selecionamos 8 variáveis dentre as 216 que são disponibilizadas em cada base. Esse estudo visa analisar essas variáveis e verificar o quanto cada uma delas influencia a nota final da prova.

Em uma segunda análise agruparemos os alunos com melhor e pior desempenho e descobriremos quais ou qual variável é decisiva nessa classificação.



## ABSTRACT

The purpose of this study is to analyze statistically the data obtained from ENEM (Test of the National High School - Brazil), gathering information to see if there is any difference in performance among students who attend public and private schools.

Our database is composed of 3 years of exams, and we selected 8 variables from the 216 available in each base. This study aims to examine these variables and see how each of them influence in the final score.

In a second analysis, we group the students with better and worse performance and determine which variables are crucial to this performance.

# 1. INTRODUÇÃO

O que é o ENEM?

O ENEM é uma prova um pouco diferente dos vestibulares, com exigências menores, e por isso os alunos , em geral , tem um desempenho melhor. Muitas faculdades e universidades usam a nota do ENEM em seus processos seletivos. Isso tem feito com que cada vez mais alunos participem anualmente da prova. Em 2005 houve quase 3 milhões de participantes.

Para orientar as políticas educacionais, o ENEM é um instrumento muito útil. Os dados apontados por essas avaliações têm mostrado, por exemplo, a enorme distância entre o nível do ensino público e particular. Mesmo numa prova que avalia habilidades e competências, em detrimento da memorização de conteúdos, a diferença de notas entre alunos de um e outro sistema de ensino (público e privado) é muito grande: 62% de diferença nas notas em 2005.

A prova:
Em cinco horas de prova, os alunos devem responder 63 questões e uma redação , o que daria aproximadamente três minutos para cada questão e uma hora e meia para a redação. Todas as questões são de múltipla escolha com 5 alternativas. Para evitar fraude, a prova é realizada em 4 versões identificadas por cores (amarela, branca, rosa e azul). O que difere uma prova da outra é a ordem das questões e alternativas. De resto, são as mesmas questões e textos.

Por objetivar avaliar competências e não informações, a prova não é dividida em matérias. Também não é indicada a competência a ser avaliada em cada questão. Portanto, as questões são colocadas em uma seqüência sem qualquer tipo de agrupamento.

Este trabalho analisará, através da prova especificada acima, qual é o perfil de um aluno que tem melhor desempenho na prova teste.

# 2. DESCRIÇÃO DO ESTUDO

Este trabalho trata os resultados do ENEM dos anos 2003, 2004 e 2005, retirados do site do INEP– Instituto Nacional de Estudos e Pesquisas Educacionais no endereço:

http://www.inep.gov.br/basica/levantamentos/microdados.asp

Não há identificação do aluno nem endereço residencial nos dados fornecidos.
A estrutura (categorias) da tabela está dividida da seguinte forma:

Variáveis de Controle do Inscrito (7 questões)
Variáveis de Controle da Escola (8 questões)
Variáveis da Prova Objetiva (7 questões)
Variáveis da Prova de Redação (7 questões)
Questionário Socioeconômico do Enem – Você e Sua Família (38 questões)
Você e o Trabalho (30 questões)
Você e os Estudos (82 questões)
Seus Valores (15 questões)
Egressos (19 questões)
Informações da Prova Objetiva (3 questões)

Totalizando 216 variáveis na prova de 2003, variando de 2 a 3 questões para mais ou para menos nos anos posteriores. Temos por volta de 1.200.000 registros por tabela.

## 2.1 Amostra

Restringiremos o estudo às escolas do município de São Paulo. Além disso, selecionamos somente alunos que concluirão o ensino médio no ano em questão e alunos que já concluíram. Dessa forma segue abaixo os filtros descritos acima e também outros que julgamos importantes:

1) Indicador de Conclusão:

   *IN_CONCLUIU = (1,2):= (concluirá no ano em questão, já concluiu)*

   Alunos de 16 a 18 anos.

2) Para selecionar os alunos do município de São Paulo, filtramos os alunos que concluíram o ensino médio em escolas que se localizam no município:

   *MUN_ESC = São Paulo*

3) Exclusão de alunos que fizeram o antigo MOBRAL:

   *Q76 = (A,C):=(Ensino Regular,Ensino Técnico/Profissional)*

Com esses filtros aplicados às bases reduzimos de 1.200.000 para em torno de 65.000 registros. Logo após essa seleção, escolhemos as variáveis que julgamos influenciar o desempenho do aluno, que é justamente o que queremos provar com as correlações que serão feitas mais adiante. Com esses filtros as bases ficaram com os seguintes tamanhos:

2003: 67.662 registros 2004: 58.080 registros 2005: 55.245 registros

## 2.2 Softwares Utilizados

**SAS - Statistical Analysis System**

O **SAS** é um sistema integrado de aplicações para a análise de dados, que consiste em: Recuperação de dados, Gerenciamento de arquivos, Análise estatística, Acesso a Banco de Dados, Geração de gráficos, Geração de relatórios. Trabalha com quatro ações básicas sobre o dado: Acessar, Manipular, Analisar e Apresentar. Pode ser instalado em diversos ambientes operacionais.

Para a filtragem da base foi utilizado o software SAS versão 9.1 Guide, pois é indicado para manipulação de bases de milhões de registros.

**Minitab 14**

Software estatístico amplamente utilizado no IME-USP para realização de trabalhos com análise de dados, relatórios , etc.

Logo após a filtragem no SAS os dados foram transferidos para o Minitab 14, para fazer as análises descritivas e inferenciais.

# 3. DESCRIÇÃO DAS VARIÁVEIS

As variáveis foram escolhidas com base na minha experiência de vida como aluna de escola pública que prestou a prova do ENEM para ingressar na universidade e também em conformidade com os orientadores. As variáveis são as seguintes:

*1) Até quando a sua mãe estudou?*

A = Não estudou;

B = Da 1ª a 4ª série do ensino fundamental (antigo primário);

C = Da 5ª a 8ª série do ensino fundamental (antigo ginásio);

D = Ensino Médio (2°grau) incompleto;

E = Ensino Médio (2°grau) completo;

F = Ensino Superior Incompleto;

G = Ensino Superior Completo;

H = Pós – Graduação;

I = Não sei;

*2) Renda Familiar (Somando a do próprio aluno com a das pessoas que moram com ele).*

A = Até 1 salário mínimo (até R$240,00);

B = De 1 a 2 salários mínimos (R$240,00 a R$480,00);

C = De 2 a 5 salários mínimos (R$480,00 a R$1.200,00);

D = De 5 a 10 salários mínimos (R$1.200,00 a R$2.400,00);

E = De 10 a 30 salários mínimos (R$2.400,00 a R$7.200,00);

F = De 30 a 50 salários mínimos (R$7.200,00 a R$12.000,00);

G = Mais de 50 salários mínimos (mais que R$12.000,00);

H = Nenhuma Renda;

*3) Trabalhou ou teve alguma atividade remunerada durante o ensino médio (segundo grau)?*

A = Sim, todo o tempo;

B = Sim, menos de 1 ano;

C = Sim, de 1 a 2 anos;

D = Sim, de 2 a 3 anos;

E = Não;

*4) Sexo do Inscrito?*

1 =Masculino;

2 = Feminino;

*5) Em que tipo de escola cursou ou esta cursando o ensino médio (segundo Grau)?*

A = Somente em escola pública;

B = Maior parte em escola pública;

C = Somente em escola particular;

D = Maior parte em escola particular;

*6) Fez curso preparatório para o vestibular (cursinho) fora da escola no ensino médio (segundo grau)?*

A = Sim;

B = Não;

# 3. ANÁLISE DESCRITIVA

A Análise Descritiva é a fase inicial deste processo de estudo dos dados coletados. Utilizei métodos de Estatística Descritiva para organizar, resumir e descrever os aspectos importantes de um conjunto de características observadas ou comparar tais características entre dois ou mais conjuntos de dados.

A descrição dos dados também tem como objetivo identificar anomalias, até mesmo resultante do registro incorreto de valores, e dados dispersos, aqueles que não seguem a tendência geral do restante do conjunto. Não só nos artigos técnicos direcionados para pesquisadores, mas também nos artigos de jornais e revistas escritos para o público leigo, é cada vez mais freqüente a utilização destes recursos de descrição para complementar a apresentação de um fato, justificar ou referendar um argumento.

As ferramentas descritivas são os muitos tipos de gráficos e tabelas e também medidas de síntese como porcentagens, índices e médias. Ao sintetizarmos os dados, perdemos informação, pois não se têm as observações originais, entretanto, esta perda de informação é pequena se comparada ao ganho que temos com a clareza da interpretação proporcionada.

## 3.1 Medidas de Tendência Central

A tendência central de uma variável pode ser descrita pela média, mediana ou moda.

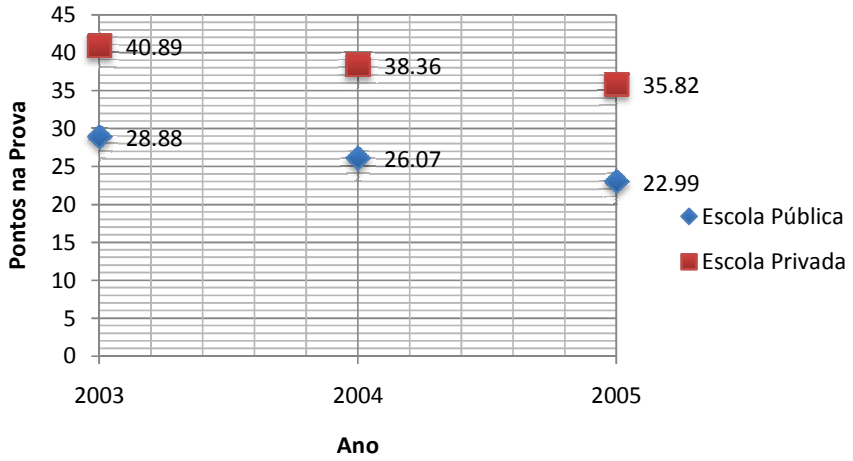
**Média Aritmética Simples:** A média aritmética simples (que chamaremos apenas de média) é a medida de tendência central mais conhecida e usada para o resumo de dados. Essa popularidade pode ser devida à facilidade de cálculo e à idéia simples que ela nos sugere.

**Notação:**

| | |
|---|---|
| $n$ | número de indivíduos no conjunto de dados |
| $x_i$ | valor da i-ésima observação do conjunto de dados, i = 1, 2, 3,..., n |
| $\sum x_i$ | soma de todas as observações da amostra (a letra grega Σ é o símbolo que indica soma). |
| $\overline{X}$ | é o símbolo usado para representar a média aritmética simples. |

$$\overline{X} = \frac{\text{Soma de todas as observações do conjunto de dados}}{\text{tamanho do conjunto de dados}} = \frac{\sum x_i}{n}$$

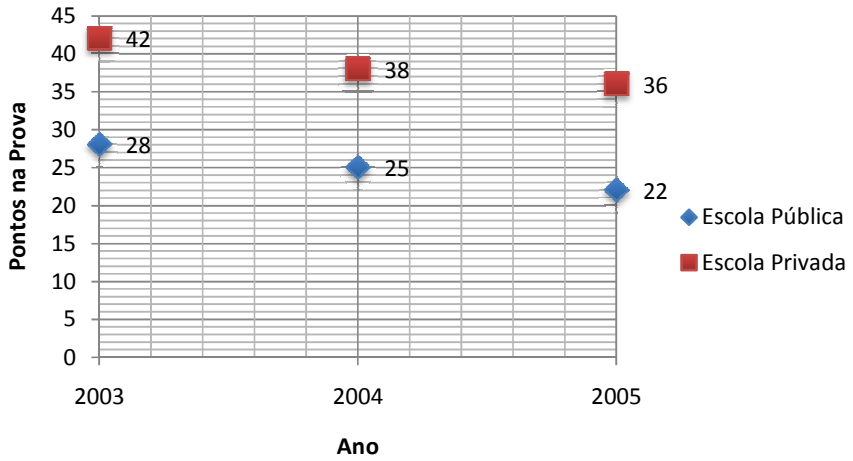
Média das notas para escola Pública e Particular:




Vemos no gráfico acima que a Escola Privada tem um desempenho melhor do que a Escola Pública, entretanto ambas vão piorando o desempenho ao longo dos três anos analisados, que poderá ser um sinal de aumento na dificuldade da prova. Queda de dois pontos para os dois tipos de colégios comparando os anos de 2003 e 2004 e uma queda de 3 pontos comparando o ano de 2004 com o ano de 2005.

**Mediana** : A mediana de um conjunto ordenado de dados é definida como sendo o "valor que divide a amostra em dois grupos, 50% acima e 50% abaixo da mediana" desse conjunto de dados, dispostos em ordem crescente. Como medida de tendência central, a mediana é até mais intuitiva do que a média, pois representa, de fato, o centro (meio) do conjunto de valores ordenados. Assim como a média, o valor da mediana não precisa coincidir com algum dos valores do conjunto de dados. Em particular, quando os dados forem de natureza contínua, essa coincidência dificilmente ocorrerá.

Mediana dos dados analisados :

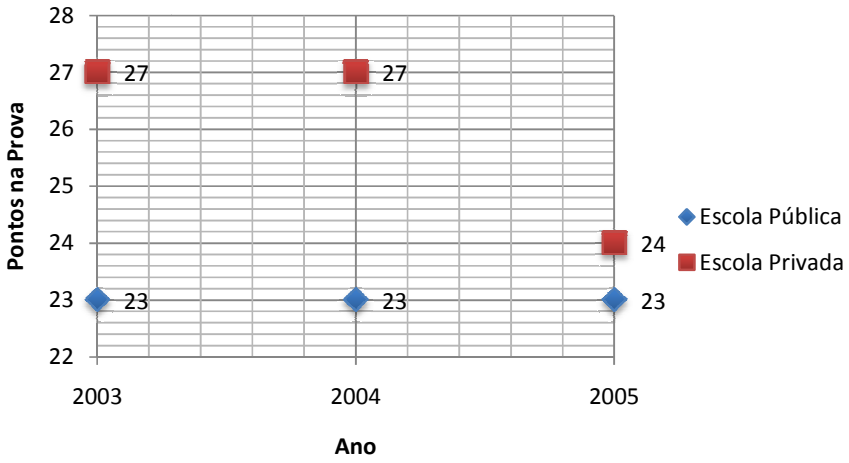



A mediana assim como a média, comporta-se de maneira decrescente ao longo do tempo, tendo uma queda constante para a Escola Pública e uma queda mais acentuada para a Escola Privada, se comparado os anos de 2003 e 2004.Vale ressaltar que apesar de ambas terem quedas no desempenho, a instituição privada sempre tem um desempenho superior à instituição pública.

Outra observação importante é que as médias estão próximas da mediana o que pode indicar uma simetria dos dados

**Moda** : Uma maneira alternativa de representar o que é "típico" é através do valor mais freqüente da variável, chamado de moda.

Moda dos dados analisados:




Na escola pública a moda das notas se matêm ao longo do tempo, já na privada em 2005 temos uma queda que pode novamente caracterizar o aumento da dificuldade da prova.

## 3.2 Medidas de Variabilidade

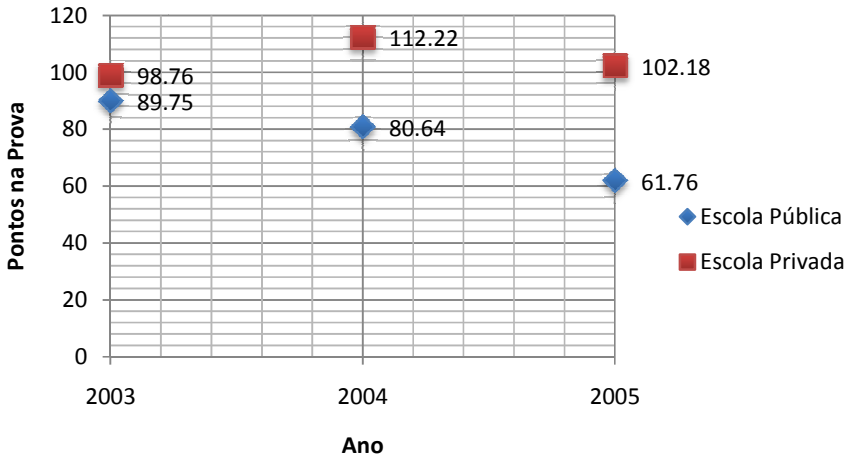
Dois conjuntos de dados podem ter a mesma medida de tendência central (valor típico), porém com uma dispersão diferente em torno desse valor. Desse modo, além de uma medida que nos diga qual é o valor " típico" do conjunto de dados, precisamos de uma medida do grau de dispersão (variabilidade) dos dados em torno do valor típico.O objetivo das medidas de variabilidade é quantificar esse grau de dispersão.

**Variância :** Na teoria da probabilidade e na estatística, a variância de uma variável aleatória é uma medida da sua dispersão estatística, indicando quão longe em geral os seus valores se encontram do valor esperado.

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$$

em que $x_i$ são os dados das amostras, $\bar{x}$ é a média da amostra.No caso estamos utilizando o termo (n-1) no denominador, pois trata-se de uma amostra dos dados.

Variância dos dados analisados:



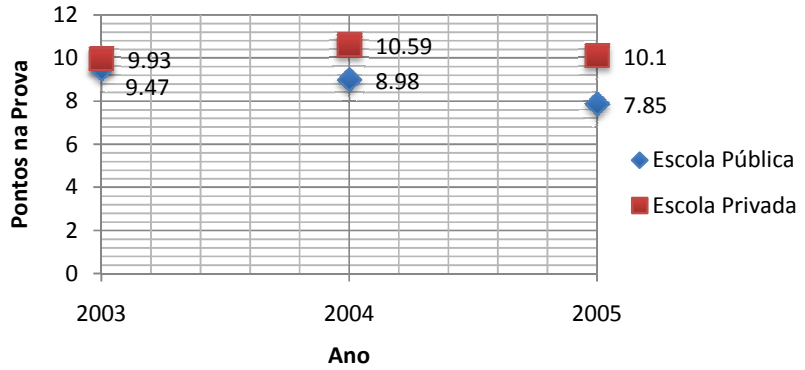

Temos que a variância diminui com o passar dos anos para escola pública sinalizando um desempenho cada vez mais homogêneo em torno da média ao longo do tempo. Para a escola privada existe uma variância alta no ano de 2004, indicando que uma aproximação das notas para a média poderia ser um erro, não expressando o real desempenho dos alunos nesse ano.

**Desvio Padrão** : O desvio-padrão, como o nome já diz, representa o desvio típico dos dados em relação à média, escolhida como medida de tendência central.

$$DP = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

O gráfico abaixo do desvio padrão dos dados indica em média qual será o "erro" (desvio) cometido ao tentar substituir cada observação pela medida resumo do conjunto de dados (no caso a média). Então em 2004 para a escola privada poderíamos estar errando em 10 pontos o desempenho de um aluno caso adotássemos a média como medida padrão de desempenho.

Desvio Padrão dos dados analisados:



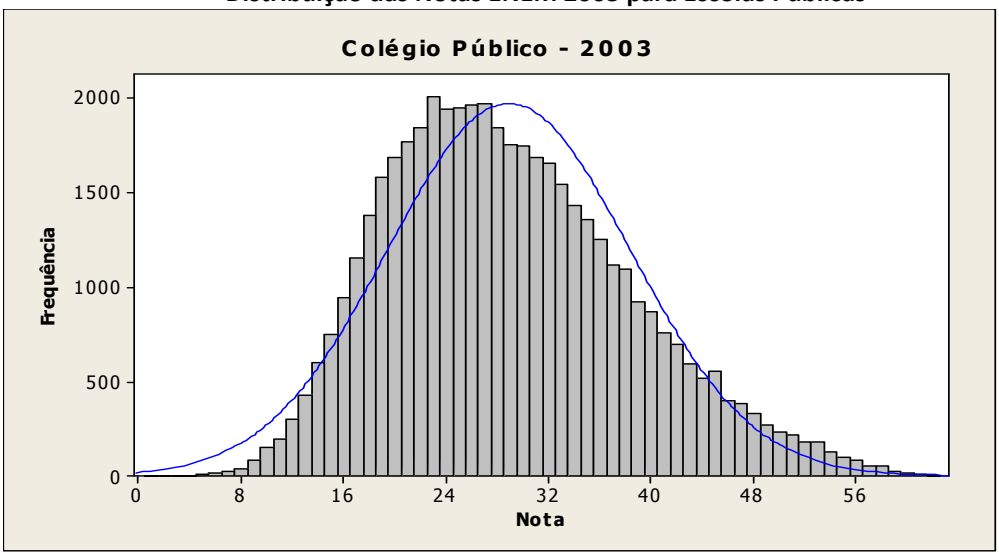
## 3.3 Histograma

Corresponde a um gráfico de barras utilizado para representar a distribuição de um conjunto de dados. Identifica rapidamente os padrões de variabilidade inerentes a um dado processo com o intuito de especular sobre as possíveis causas determinantes da forma da distribuição. A representação gráfica da dispersão de um processo permite conhecer melhor o tipo de distribuição característica dos fenômenos em estudo.

✓ 2003

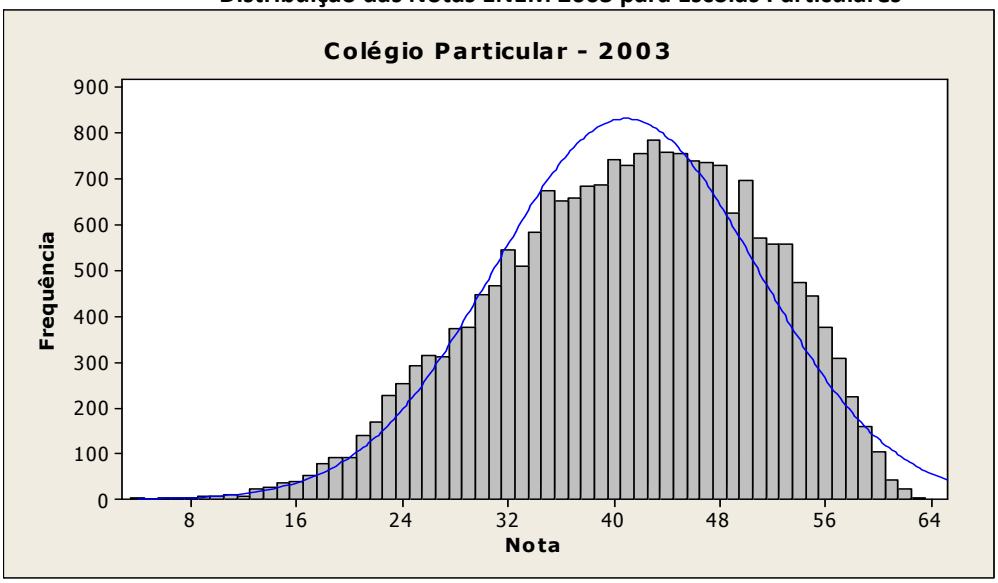
Tanto o ensino público quanto o privado temos um desempenho homogêneo em 2003, que "acompanha" curva azul (distribuição normal). O que difere entre os dois é a média que no caso da escola particular faz com que o gráfico seja deslocado para o lado direito (>nota), predominam notas mais altas. Ao contrário do colégio público, no qual o gráfico é deslocado para a esquerda (< nota) predominando as notas mais baixas.

**Distribuição das Notas ENEM 2003 para Escolas Públicas**




| | |
|---|---|
| Média | 28.89 |
| Desvio Padrão | 9.474 |
| Observações | 46808 |

**Distribuição das Notas ENEM 2003 para Escolas Particulares**



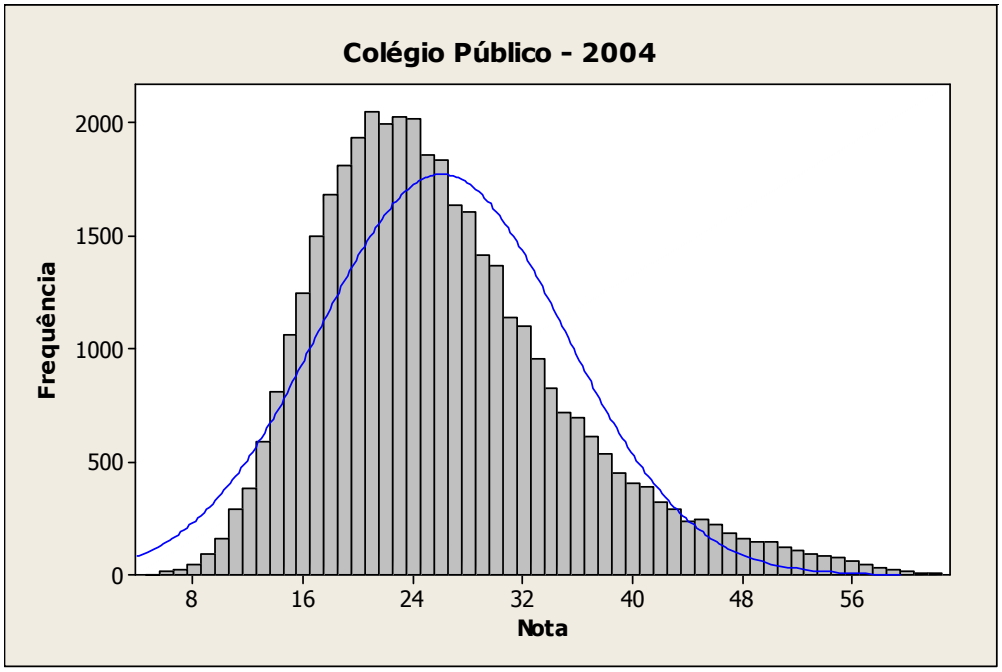

| | |
|---|---|
| Média | 40.9 |
| Desvio Padrão | 9.938 |
| Observações | 20715 |

✓ 2004

Da mesma forma que em 2003 o fato se repete com o colégio público, temos um acentuado volume de notas concentrado do lado direito do gráfico, demonstrando o baixo desempenho dos alunos, até menor do que em 2003 (do ponto de vista da média dos dados).

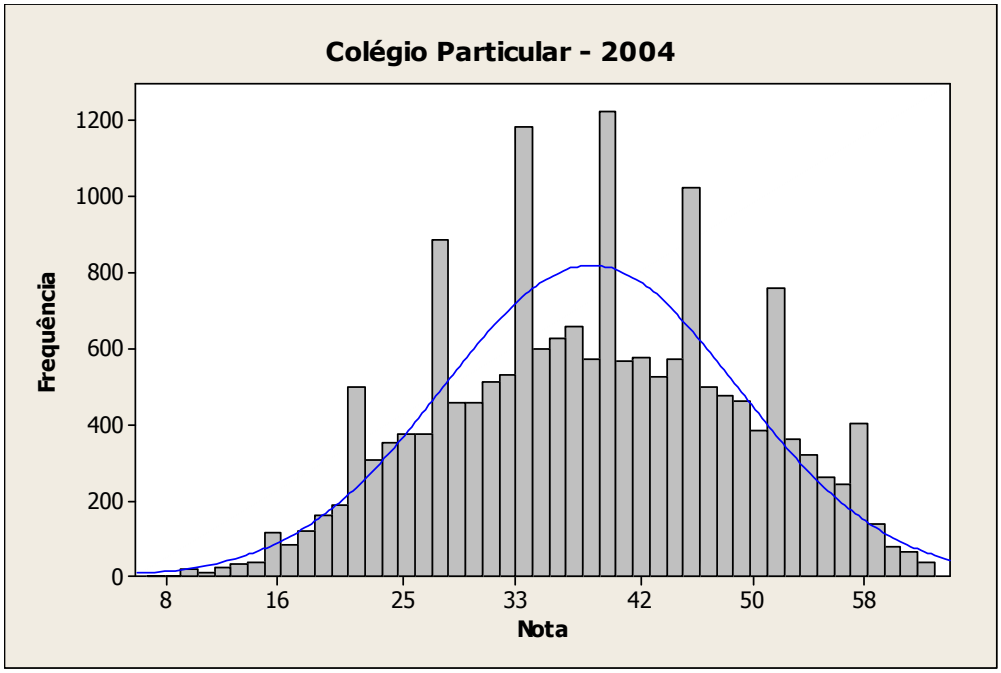
Para a escola particular temos algumas anomalias, concentração de notas em alguns pontos do gráfico, isso nós podemos notar no gráfico da variância, pois é o ano com menor homogeneidade nos dados.

**Distribuição das Notas ENEM 2004 para Escolas Públicas**




| | |
|---|---|
| Média | 26.07 |
| Desvio Padrão | 8.98 |
| Observações | 39911 |

**Distribuição das Notas ENEM 2004 para Escolas Particulares**



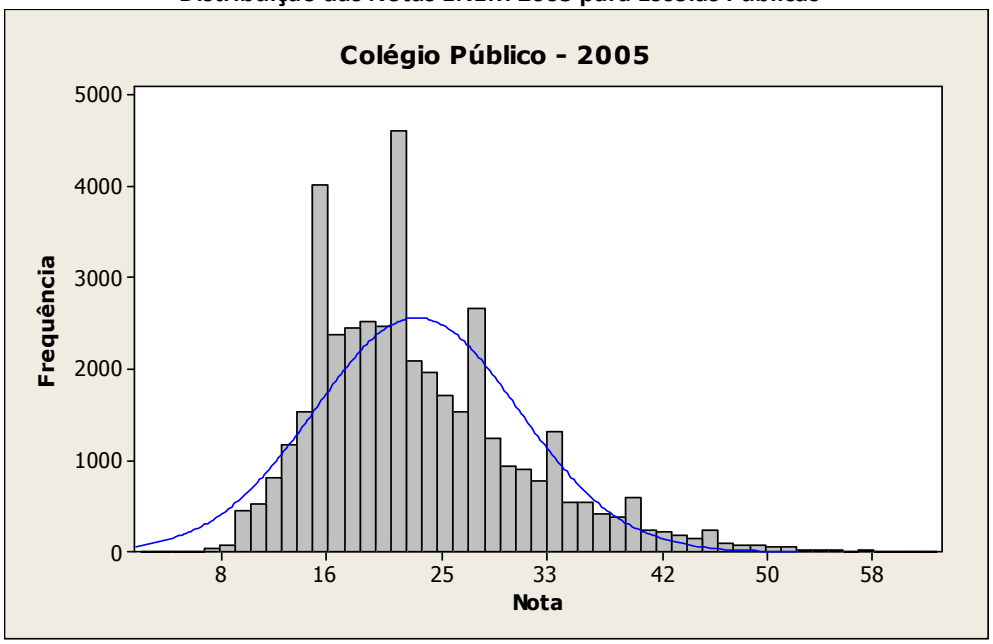

| | |
|---|---|
| Média | 38.37 |
| Desvio Padrão | 10.57 |
| Observações | 18093 |

✓ 2005

No ano de 2005 temos uma alta variabilidade nos dados dos colégios públicos, alta freqüência de determinadas notas como 23 que é mostrado no gráfico da moda dos dados.

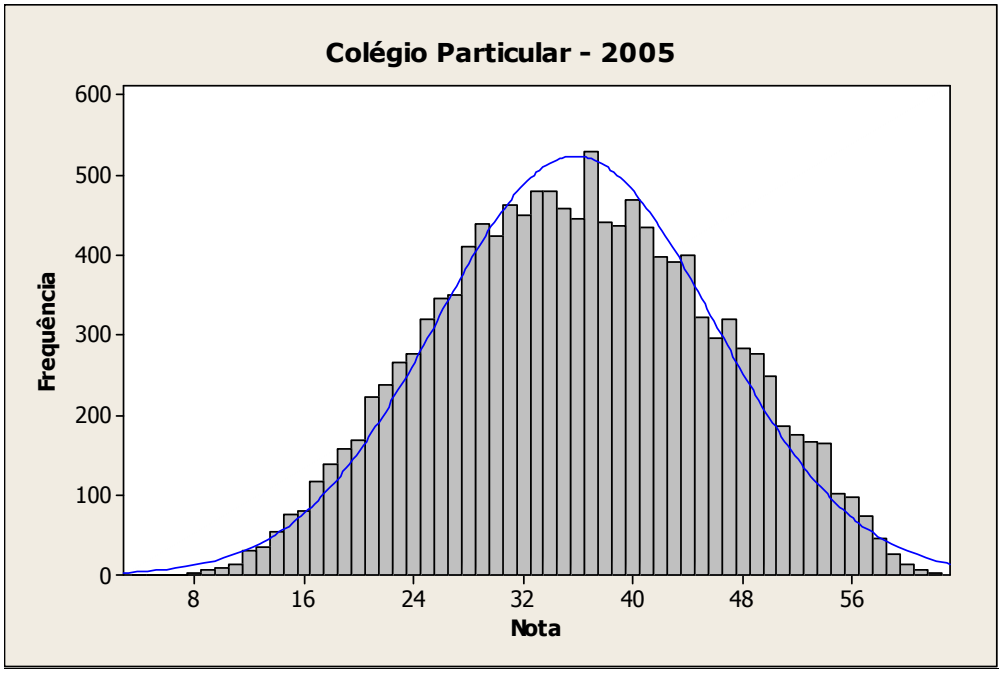
Para o colégio particular temos um desempenho homogêneo dos alunos, entretanto se comparado com o ano anterior temos uma queda de 3 pontos na média geral.

**Distribuição das Notas ENEM 2005 para Escolas Públicas**




| | |
|---|---|
| Média | 22.99 |
| Desvio Padrão | 7.85 |
| Observações | 42007 |

**Distribuição das Notas ENEM 2005 para Escolas Particulares**

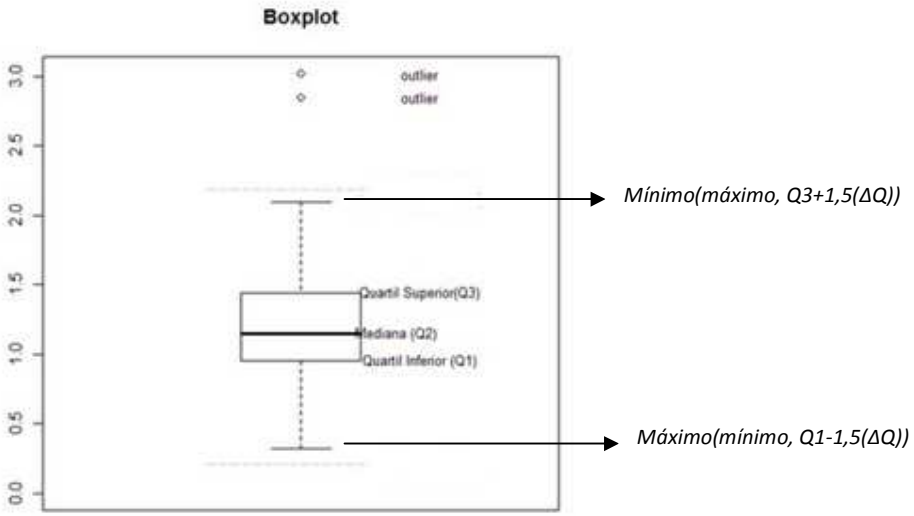



| | |
|---|---|
| Média | 35.82 |
| Desvio Padrão | 10.11 |
| Observações | 13238 |

## 3.4 Boxplot

O boxplot é um gráfico que possibilita representar a distribuição de um conjunto de dados com base em alguns de seus parâmetros descritivos, quais sejam: a mediana (q2), o quartil inferior (q1), o quartil superior (q3) e do intervalo interquartil (IQR = q3 - q1).

A figura a seguir apresenta o boxplot, destacando suas principais características:




A linha central da caixa marca a mediana do conjunto de dados. A parte inferior da caixa é delimitada pelo quartil inferior (Q1) e a parte superior pelo quartil superior (Q3). As hastes inferiores e superiores se estendem, respectivamente, do quartil inferior até o menor valor não inferior a Q1 - 1.5ΔQ e do quartil superior até o maior valor não superior a Q3 + 1.5ΔQ. Os valores inferiores a Q1 - 1.5ΔQ e superiores a Q3 + 1.5ΔQ são representados individualmente no gráfico sendo estes valores caracterizados como outliers.

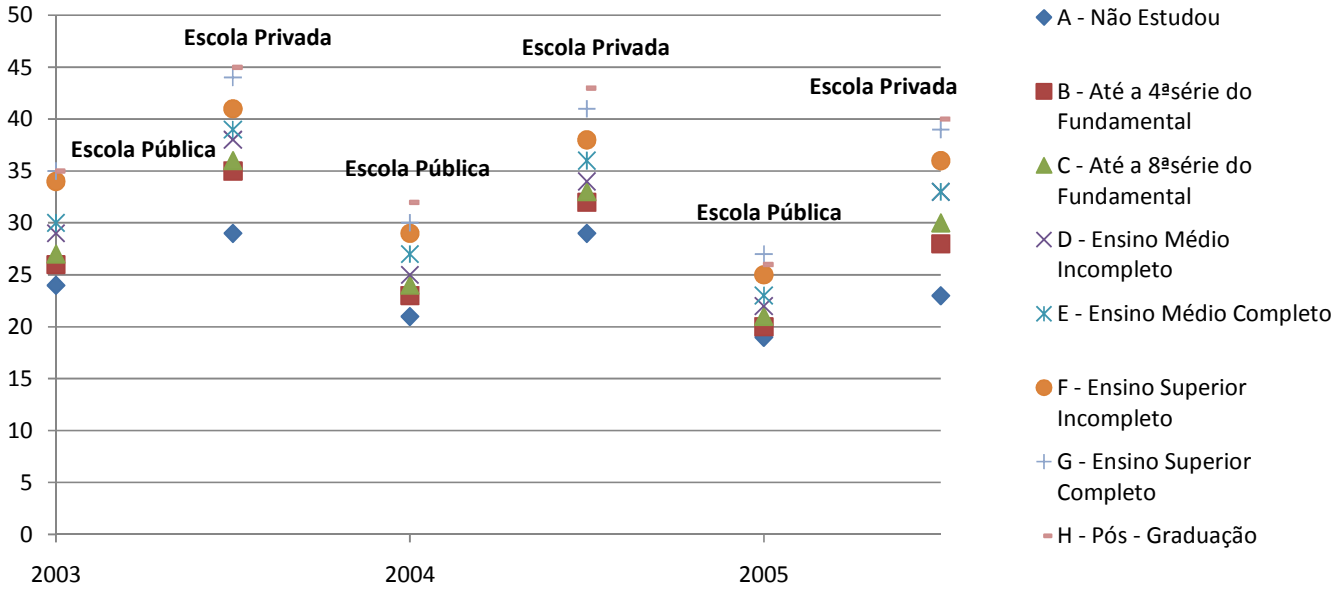
O boxplot permite avaliar a simetria dos dados, sua dispersão e a existência ou não de outliers nos mesmos, sendo especialmente adequado para a comparação de dois ou mais conjuntos de dados correspondentes às categorias de uma variável qualitativa.

- ✓ 2003 – 2004 – 2005

  - *Grau de Instrução da Mãe*

Observando o gráfico abaixo, extraído das tabelas A.1 a A.6 (geradas a partir do Box plot), podemos ver claramente que o fato da mãe ter grau de instrução maior faz com que a nota do aluno seja elevada. Isso ocorre tanto na Instituição pública quanto na privada, sendo que na instituição publica sempre há uma queda na nota. Outro item importante extraído das tabelas no apêndice é que alunos de escolas privadas têm um maior percentual (em média 42%) de mães que concluíram o ensino superior, enquanto que na escola pública o maior percentual são mães que concluíram a 4ª série do ensino fundamental.

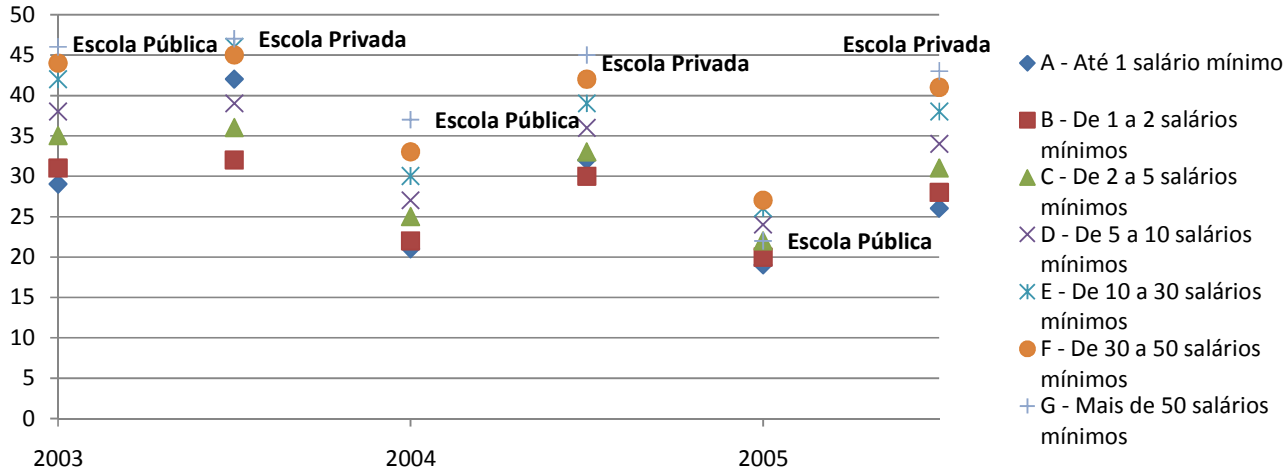
**Nota x Grau de Instrução da Mãe**




- *Renda Familiar*

O gráfico abaixo foi extraído das tabelas B.1 a B.6, podemos verificar que tanto na escola pública quanto na escola particular quanto maior a renda maior a nota do aluno, isso é compatível visto que quanto maior renda, maior o investimento no estudo. Nas tabelas que se encontram no apêndice podemos notar que maior percentual dos alunos que estudam em escola privada tem renda entre **20 e 30 salários mínimos**, enquanto que na escola pública o maior percentual são dos alunos que têm renda entre **2 e 5 salários mínimos.**

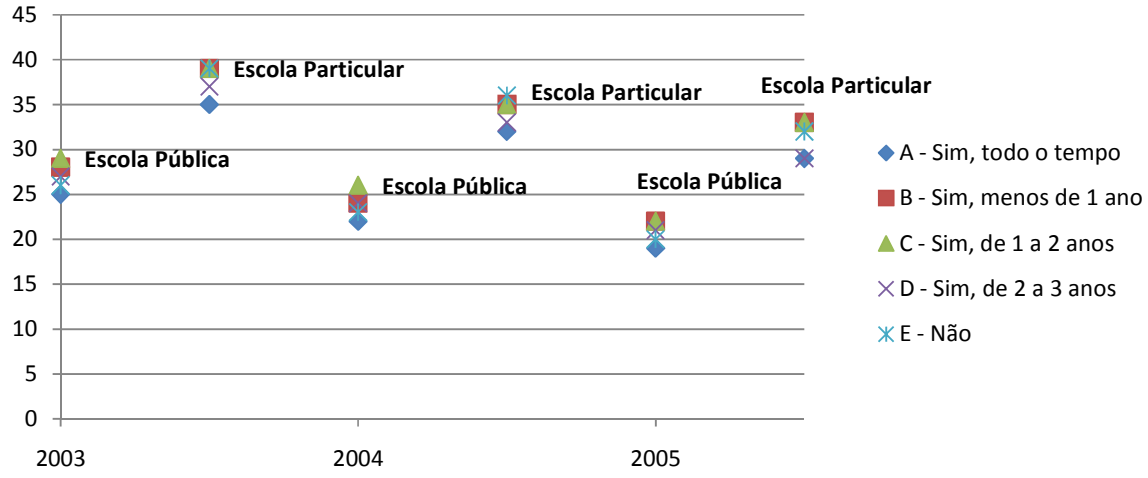
**Nota x Renda Familiar**



- *Exercer Atividade Remunerada no Ensino Médio*

AS tabelas C.1 a C.6 não apontam que ter uma atividade remunerada no ensino médio influência diretamente na nota, pois os alunos que não trabalham têm praticamente o mesmo desempenho dos que exercem algum tipo de atividade extra-escolar. O que parece relativamente estranho, pois o cansaço no momento de assistir aula é um fator preponderante na absorção das informações, quanto maior o cansaço, menor será a concentração. Os dados das escolas públicas se concentram na opção B, que tiveram algum tipo de trabalho que durou menos de 1 ano. Já os alunos de colégios privados se concentram na opção E, não trabalharam nesse período.

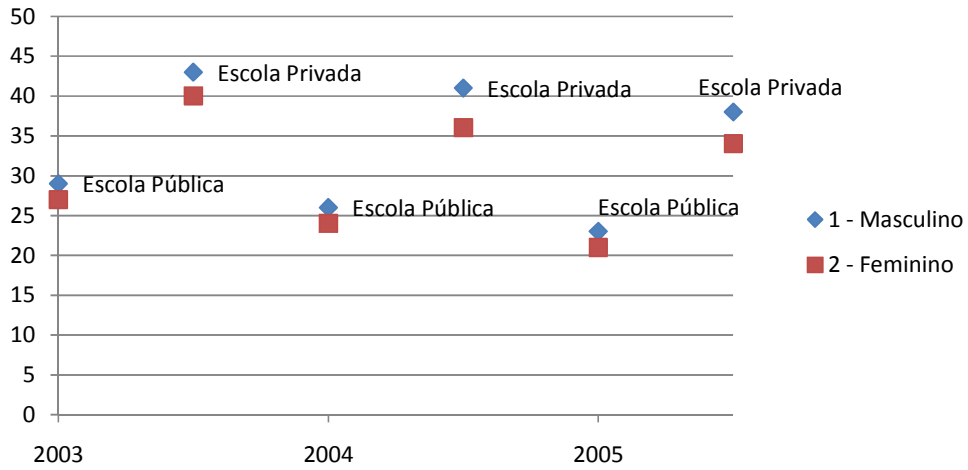
**Nota x Atividade Remunerada**




- *Sexo do Aluno*

Conseguimos verificar no gráfico abaixo que os alunos do sexo masculino têm um desempenho melhor do que o sexo feminino em ambos os tipos de colégios. Em todos os anos o percentual de participação feminina na prova é maior do que o masculino.

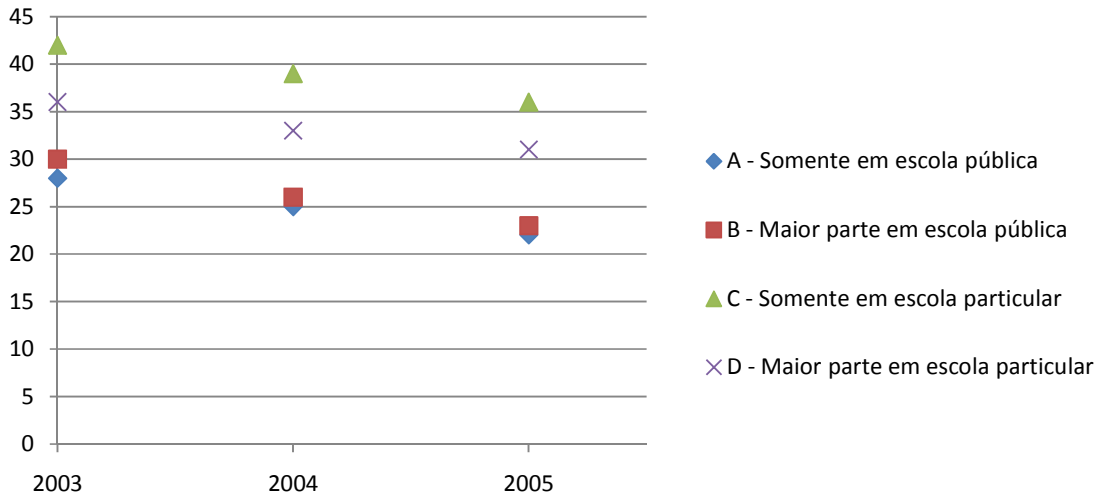
**Nota x Sexo do Aluno**




- *Tipo de Escola no Ensino Médio*

É claro no gráfico abaixo a separação das médias das notas dos alunos que fizeram o Ensino Médio em escola pública e em escola privada. Nos domínios A (somente em escola pública) e C (somente escola particular) temos uma diferença de 13 pontos em média.

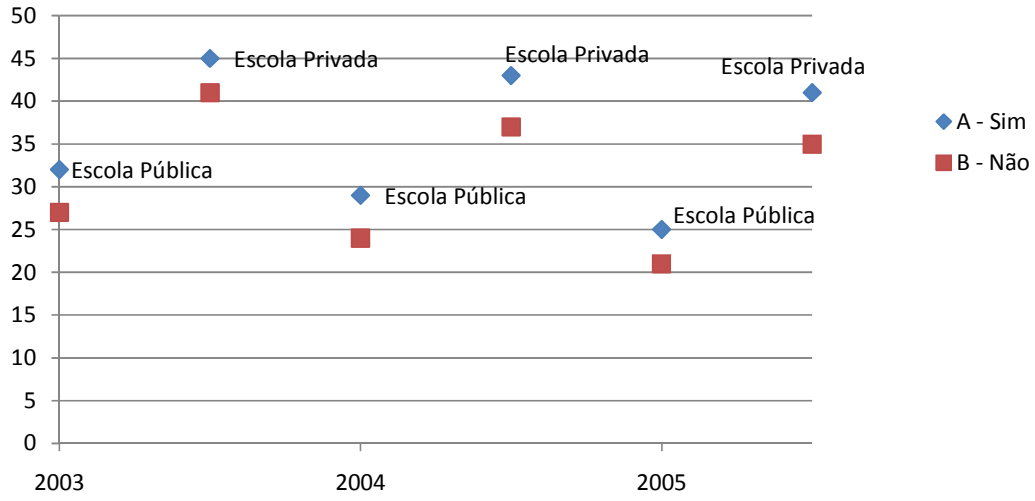
**Nota x Tipo de Escola no Ensino Médio**




- Curso Pré - Vestibular

O fato de fazer um curso pré – vestibular faz com que as notas de ambas as instituições aumentem. Pois se caracteriza como um reforço para o Ensino Médio.

**Nota x Curso Pré - Vestibular**



# 5. ANÁLISE INFERENCIAL

A análise inferencial é um método que torna possível a estimativa de uma característica da população ou a tomada de uma decisão referente à população com base somente em resultados de amostras.

## 5.1 Teste de Independência das Variáveis

O objetivo do teste de independência é verificar se existe independência entre duas variáveis medidas nas mesmas unidades experimentais.

Em geral, os dados referem-se a mensurações de duas características A e B feitas com n unidades experimentais que são apresentadas conforme a tabela abaixo:

| B \ A | $A_1$ | $A_2$ | … | $A_s$ | Total |
|---|---|---|---|---|---|
| $B_1$ | $n_{11}$ | $n_{12}$ | … | $n_{1s}$ | $n_{1\cdot}$ |
| $B_2$ | $n_{21}$ | $n_{22}$ | … | $n_{2s}$ | $n_{1\cdot}$ |
| … | … | … | … | … | … |
| $B_r$ | $n_{r1}$ | $n_{r2}$ | … | $n_{rs}$ | $n_{r\cdot}$ |
| Total | $n_{\cdot 1}$ | $n_{\cdot 2}$ | … | $n_{\cdot s}$ | $n_{\cdot\cdot}$ |

Sendo que o elemento $n_{ij}$ o número de elementos na amostra que apresenta a característica $A_i$ e $B_j$ dentre os n.

Hipóteses a serem testadas – Teste de Independência:

$H_0$ : A e B são variáveis Independentes;

$H_1$ : A e B são Dependentes;

Quantas observações devemos ter em cada casela se A e B forem independentes?
Se A e B forem independentes, temos que, para todos os possíveis (Ai e Bj):

$$P(Ai \cap Bj) = P(Ai) \times P(Bj),\ j = 1, 2, \ldots, r \ \textbf{e}\ i = 1, 2, \ldots, s.$$

Logo, o número esperado de observações com as características (Ai e Bj) entre as n observações sob a hipótese de independência, é dado por:

$$E_{ij} = n_{\cdot\cdot} \times p_{ij} = n_{\cdot\cdot} \times p_{i\cdot} \times p_{\cdot j} = n_{\cdot\cdot} \times \frac{n_{i\cdot}}{n_{\cdot\cdot}} \times \frac{n_{\cdot j}}{n_{\cdot\cdot}}$$

sendo $p_{ij}$ a proporção de observações com as características (Ai e Bj). Assim:

$$E_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{n_{\cdot\cdot}}$$

O processo deve ser repetido para todas as caselas (ij).

A Distância entre os valores observados e os valores esperados sob a suposição de independência:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

em que Oij = nij representa o total de observações na casela ( ij ).

Supondo $H_0$ verdadeira,

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \approx \chi^2_q$$

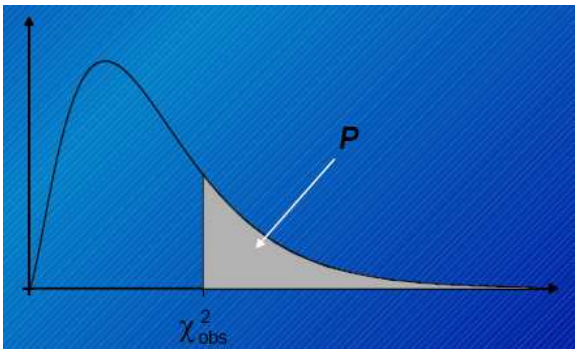
sendo q = ( r – 1) x ( s – 1 ) graus de liberdade.

**Regra de decisão:** Pode ser baseada no nível descritivo P, neste caso

$$P = P(\chi^2_q \geq \chi^2_{obs})$$

Em que $\chi^2_{obs}$ é o valor calculado, a partir dos dados, para a estatística do teste.

Graficamente:




**Se para α fixado obtemos P <= α, rejeitamos a hipótese de Independência.**

Para todas as variáveis e anos o Minitab apontou que nenhuma é independente juntamente com a nota. Para fazer os cálculos as notas foram divididas em faixas da seguinte forma:

**Faixa 0 = Notas de 0 a 10 ;**
**Faixa 1 = Notas de 11 a 20 ;**
**Faixa 2 = Notas de 21 a 30 ;**
**Faixa 3 = Notas de 31 a 40 ;**
**Faixa 4 = Notas de 41 a 50 ;**
**Faixa 5 = Notas de 51 a 60 ;**
**Faixa 6 = Notas de 61 a 63;**

Dado que a distribuição qui-quadrado, é uma distribuição aproximada, precisamos tomar certos cuidados na sua aplicação. Um deles é garantir que todos os valores esperados das caselas não sejam inferiores a cinco. Tivemos alguns casos desse tipo, entretanto conseguimos sanar o problema juntando duas faixas de notas e suas respectivas freqüências.

No apêndice temos todas as tabelas e resultados gerados pelo programa. Abaixo temos um resumo:

| 2003 | |
|---|---|
| **Escola Privada** | Nota |
| **Grau de Instrução da Mãe** | Não são Independentes |
| **Renda Familiar** | Não são Independentes |
| **Trabalho no Ensino Médio** | Não são Independentes |
| **Escola no Ensino Médio** | Não são Independentes |
| **Fez Cursinho?** | Não são Independentes |
| **Sexo** | Não são Independentes |

| | |
|---|---|
| **Escola Pública** | Nota |
| **Grau de Instrução da Mãe** | Não são Independentes |
| **Renda Familiar** | Não são Independentes |
| **Trabalho no Ensino Médio** | Não são Independentes |
| **Escola no Ensino Médio** | Não são Independentes |
| **Fez Cursinho?** | Não são Independentes |
| **Sexo** | Não são Independentes |

| 2004 | |
|---|---|
| **Escola Privada** | Nota |
| **Grau de Instrução da Mãe** | Não são Independentes |
| **Renda Familiar** | Não são Independentes |
| **Trabalho no Ensino Médio** | Não são Independentes |
| **Escola no Ensino Médio** | Não são Independentes |
| **Fez Cursinho?** | Não são Independentes |
| **Sexo** | Não são Independentes |

| | |
|---|---|
| **Escola Pública** | Nota |
| **Grau de Instrução da Mãe** | Não são Independentes |
| **Renda Familiar** | Não são Independentes |
| **Trabalho no Ensino Médio** | Não são Independentes |
| **Escola no Ensino Médio** | Não são Independentes |
| **Fez Cursinho?** | Não são Independentes |
| **Sexo** | Não são Independentes |

| 2005 | |
|---|---|
| **Escola Privada** | Nota |
| **Grau de Instrução da Mãe** | Não são Independentes |
| **Renda Familiar** | Não são Independentes |
| **Trabalho no Ensino Médio** | Não são Independentes |
| **Escola no Ensino Médio** | Não são Independentes |
| **Fez Cursinho?** | Não são Independentes |
| **Sexo** | Não são Independentes |

| | |
|---|---|
| **Escola Pública** | Nota |
| **Grau de Instrução da Mãe** | Não são Independentes |
| **Renda Familiar** | Não são Independentes |
| **Trabalho no Ensino Médio** | Não são Independentes |
| **Escola no Ensino Médio** | Não são Independentes |
| **Fez Cursinho?** | Não são Independentes |
| **Sexo** | Não são Independentes |

Analisando as tabelas acima vemos que todos as variáveis não são independentes mas também não podemos afirmar que são dependentes sem justificativas plausíveis. Para isso faremos um teste de hipótese para as médias das amostras para todas as variáveis

## 5.2 Teste de Hipótese

Após esse resultado de não independência das variáveis, faremos o teste de médias para nos certificar de que a nota do aluno se modifica de acordo com o domínio da variável. Faremos o teste para todos os anos 2003,2004,2005, entretanto só mostremos os testes feitos em 2003.

Fizemos o teste **Two-Sample T-Test** do software Minitab, pois estamos comparando sempre duas amostras. O teste de hipótese para as variáveis é da seguinte forma:

$H_0$: As variáveis têm a média igual
$H_1$: As variáveis têm média diferente

Com o nível de confiança de 95%.

Dessa forma para cada variável separaremos sempre em duas amostras de domínio extremo:

- Grau de Instrução da Mãe: Comparação entre alunos que têm mãe com pós-graduação e alunos que tem mãe com ensino fundamental

| | N | Mean | StDev | SE Mean |
|---|---|---|---|---|
| Pós - Graduação | 2269 | 43.0 | 10.3 | 0.22 |
| Fundamental | 14055 | 27.44 | 8.89 | 0.075 |

Difference = mu (Pós - Graduação) - mu (Fundamental)
Estimate for difference: 15.5868
95% CI for difference: (15.1398; 16.0338)

T-Test of difference = 0 (vs not =): T-Value = 68.37 **P-Value = 0.000** DF = 2844

Como o valor de P = 0 rejeitamos a hipótese H0 e aceitamos a hipótese H1 de que as médias das notas são diferentes para alunos que tem a mãe com grau de instrução Ensino Fundamental e Pós Graduação.

- Renda Familiar: comparação entre alunos que têm renda até um salário mínimo e alunos que tem renda entre 30 e 50 salários mínimos

| | N | Mean | StDev | SE Mean |
|---|---|---|---|---|
| 20 a 50 salarios | 2631 | 43.3 | 10.0 | 0.20 |
| 1 salario | 2067 | 25.03 | 8.77 | 0.19 |

Difference = mu (20 a 50 salarios) - mu (1 salario)
Estimate for difference: 18.2882
95% CI for difference: (17.7504; 18.8260)
T-Test of difference = 0 (vs not =): T-Value = 66.67 **P-Value = 0.000** DF = 4640

Como o valor de P = 0 rejeitamos a hipótese H0 e aceitamos a hipótese H1 de que as médias das notas são diferentes para alunos que tem renda de até 1 salário e alunos que tem renda entre 20 e 50 salários mínimos

- Trabalho no Ensino Médio: comparação entre alunos que trabalharam durante o ensino médio e alunos que não trabalharam.

| | N | Mean | StDev | SE Mean |
|---|---|---|---|---|
| não trabalhou | 9662 | 30.0 | 10.6 | 0.11 |
| trabalhou | 6149 | 27.21 | 8.99 | 0.11 |

Difference = mu (não trabalhou) - mu (trabalhou)
Estimate for difference: 2.77124
95% CI for difference: (2.46240; 3.08008)
T-Test of difference = 0 (vs not =): T-Value = 17.59 **P-Value = 0.000** DF = 14601

Como o valor de P = 0 rejeitamos a hipótese H0 e aceitamos a hipótese H1 de que as médias das notas são diferentes para alunos que trabalharam e para alunos que não trabalharam no ensino médio.

- Escola no Ensino Médio: comparação entre alunos que estudaram em colégios privados e alunos que estudaram em colégios públicos.

| | N | Mean | StDev | SE Mean |
|---|---|---|---|---|
| privada | 19436 | 41.22 | 9.84 | 0.071 |
| publica | 45428 | 28.85 | 9.44 | 0.044 |

Difference = mu (privada) - mu (publica)
Estimate for difference: 12.3705
95% CI for difference: (12.2071; 12.5339)
T-Test of difference = 0 (vs not =): T-Value = 148.41 **P-Value = 0.000** DF = 35413

Como o valor de P = 0 rejeitamos a hipótese H0 e aceitamos a hipótese H1 de que as médias das notas são diferentes para alunos que estudaram o ensino médio em colégio particular e os alunos que estudaram o ensino médio em colégio público.

- Curso Pré – Vestibular: comparação entre alunos que fizeram curso pré-vestibular e alunos que não fizeram

| | N | Mean | StDev | SE Mean |
|---|---|---|---|---|
| sim | 9775 | 37.8 | 11.7 | 0.12 |
| não | 57047 | 31.8 | 10.7 | 0.045 |

Difference = mu (sim) - mu (não)
Estimate for difference: 6.05625
95% CI for difference: (5.80873; 6.30377)
T-Test of difference = 0 (vs not =): T-Value = 47.96 **P-Value = 0.000** DF = 12770

Como o valor de P = 0 rejeitamos a hipótese H0 e aceitamos a hipótese H1 de que as médias das notas são diferentes para alunos que fizeram curso pré-vestibular e de alunos que não fizeram.

- Sexo : Comparação entre alunos do sexo feminino e masculino

| | N | Mean | StDev | SE Mean |
|---|---|---|---|---|
| Masculino | 28322 | 34.2 | 11.6 | 0.069 |
| Feminino | 39340 | 31.4 | 10.5 | 0.053 |

Difference = mu (Masculino) - mu (Feminino)
Estimate for difference: 2.88331
95% CI for difference: (2.71238; 3.05424)
T-Test of difference = 0 (vs not =): T-Value = 33.06 **P-Value = 0.000** DF = 57198

Como o valor de P =0 rejeitamos H0 e aceitamos a hipótese H1 de que as médias são diferentes para alunos do sexo masculino e feminino.

Após as análises podemos concluir que as variáveis escolhidas são **dependentes** da nota.

## 5.3 Análise de Cluster

Nesse capítulo, descreve-se alguns métodos de análise de cluster, tendo como objeto de estudo, os métodos hierárquicos e os não-hierárquicos de agrupamento. Primeiramente, destacam-se medidas de similaridade e o uso da matriz de similaridade. Em seguida, a descrição dos métodos com seus algoritmos, funções distância e algumas características,trazendo um exemplo da formação dos grupos em cada método. Na última seção, apresentam-se,brevemente, outros métodos, como agrupamentos fuzzy e mapas de Kohonen. A análise de cluster busca agrupar elementos de dados baseando-se na similaridade entre eles. Os grupos são determinados de forma a obter-se homogeneidade dentro dos grupos e heterogeneidade entre eles.

A necessidade de classificar elementos em grupos por suas características está presente em várias áreas do conhecimento, como nas ciências biológicas, ciências sociais e comportamentais, ciências da terra, medicina, informática, entre outras. Tendo em vista a dificuldade de se examinar todas as combinações de grupo possíveis em um grande volume de dados, desenvolveram-se diversas técnicas capazes de auxiliar na formação dos agrupamentos.

Uma análise de cluster criteriosa exige métodos que apresentem as seguintes características (ZAIANE, 2003):

· Ser capaz de lidar com dados com alta dimensionalidade;
· Ser "escalável" com o número de dimensões e com a quantidade de elementos a serem agrupados;
· Habilidade para lidar com diferentes tipos de dados;
· Capacidade de definir agrupamentos de diferentes tamanhos e formas;
· Exigir o mínimo de conhecimento para determinação dos parâmetros de entrada;
· Ser robusto à presença de ruído;
· Apresentar resultado consistente independente da ordem em que os dados são apresentados;

Em geral, algoritmo algum atende a todos esses requisitos e, por isso, é importante entender as características de cada algoritmo para a escolha de um método adequado a cada tipo de dado ou problema (HALDIKI, 2001).

### 5.3.1 Medidas de Similaridade

A maioria dos métodos de análise de cluster requer uma medida de similaridade entre os elementos a serem agrupados, normalmente expressa como uma função distância ou métrica.

Seja M um conjunto, uma métrica em M é uma função $d: M \times M \rightarrow \Re$ , tal que para quaisquer $x, y, z \in M$, tenhamos:

1. dxy > 0 - para todo $x \neq y$
2. dxy = 0 $\Leftrightarrow$ x = y
3. dxy = dyx
4. dxy $\leq$ dxz + dzy

### Distância Euclidiana

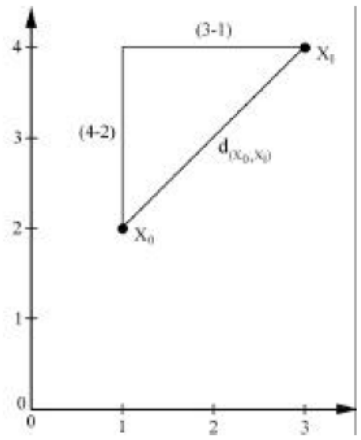
A distância euclidiana é a distância geométrica no espaço multidimensional.

A distância euclidiana entre dois elementos $X = [X_1, X_2, ..., Xp]$ e $Y = [Y_1, Y_2, ..., Yp]$, é definida por:

$$d_{xy} = \sqrt{(X1 - Y1)^2 + (X2 - Y2)^2 + ... + (Xp - Yp)^2} = \sqrt{\sum_{i=1}^{p} (Xi - Yi)^2}$$

**Exemplo.** Considerando-se elementos como pontos no plano (espaço euclidiano $\Re_2$ ), a distância entre os elementos X0 = (1,2) e X1 = (3,4) é dada por:

$$d_{x_0 x_1} = \sqrt{(3-1)^2 + (4-2)^2} = \sqrt{8} = 2{,}83$$



Distância euclidiana entre os pontos X0 e X1 no plano.

## Distância Euclidiana Quadrática

A distância euclidiana quadrática é definida pela expressão:

$$d_{xy} = (X1 - Y1)^2 + (X2 - Y2)^2 + ... + (Xp - Yp)^2 = \sum_{i=1}^{p} (Xi - Yi)^2$$

**Exemplo**. Considerando-se os mesmos pontos X0 e X1 do exemplo anterior, observa-se a intensificação da distância:

$$d_{x_0x_1} = (3-1)^2 + (4-2)^2 = 8$$

## Distância de *Manhattan*

A distância de *Manhattan* é definida pela expressão:

$$d_{xy} = |X1 - Y1| + |X2 - Y2| + ... + |Xp - Yp| = \sum_{i=1}^{p} |Xi - Yi|$$

Em muitos casos, a distância de Manhattan apresenta resultados similares ao da distância Euclidiana. Entretanto, nessa medida, o efeito de uma grande diferença entre uma das dimensões de um elemento é minimizado, já que a mesma não é elevada ao quadrado.

**Exemplo.** Empregando-se os pontos do exemplo anterior, temos:

$$d_{x_0x_1} = |3-1| + |4-2| = |2| + |2| = 4$$

## Distância de *Chebychev*

A distância de *Chebychev* é apropriada no caso em que se deseja definir dois elementos como diferentes, se apenas umas das dimensões é diferente. Ela é definida por:

$$d_{xy} = máximo(|X1 - Y1|, |X2 - Y2|, ..., |Xp - Yp|)$$

**Exemplo.** Considerando-se os pontos X2 = (9,2) e X3 = (2,5), a distância de Chebychev é dada

por:

$$d_{x_2x_3} = máximo(|9-2|,|2-5|) = (|7|,|-3|) = 7$$

As medidas de similaridade são utilizadas na análise de cluster de forma a determinar a distância entre elementos. Essa distância é normalmente representada na forma de matriz, ou seja, em uma matriz de similaridade. A matriz de similaridade é simétrica e utiliza, na maioria dos casos, a distância Euclidiana.

**Exemplo.** Considerando os elementos da tabela abaixo, obtemos a matriz de similaridade D.

| ELEMENTO | X | Y |
|---|---|---|
| 1 | 4 | 3 |
| 2 | 2 | 7 |
| 3 | 4 | 7 |
| 4 | 2 | 3 |
| 5 | 3 | 5 |
| 6 | 6 | 1 |

**Elementos do exemplo para matriz de similaridade.**

$$D = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{bmatrix} 0 & 4,47 & 4 & 2 & 2,24 & 2,83 \\ 4,47 & 0 & 2 & 4 & 2,24 & 7,21 \\ 4 & 2 & 0 & 4,47 & 2,24 & 6,32 \\ 2 & 4 & 4,47 & 0 & 2,24 & 4,47 \\ 2,24 & 2,24 & 2,24 & 2,24 & 0 & 5 \\ 2,83 & 7,21 & 6,32 & 4,47 & 5 & 0 \end{bmatrix}$$

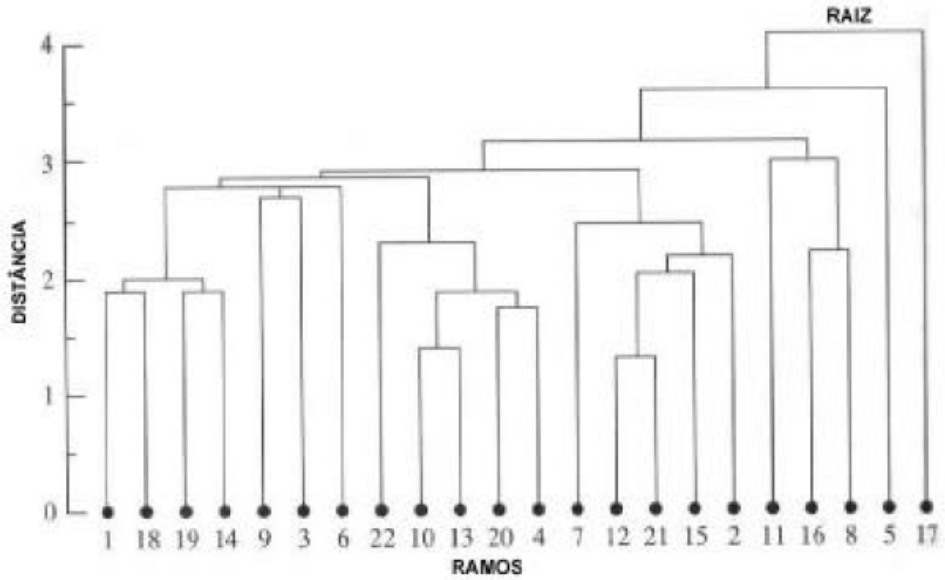
- sendo:
  - 5 a distância Euclidiana entre os elementos 6 e 5;
  - 2,83 a distância Euclidiana entre os elementos 1 e 6;
  - 6,32 a distância Euclidiana entre os elementos 3 e 6.

### 5.3.2 Métodos Hierárquicos

O método hierárquico de cluster consiste em uma série de sucessivos agrupamentos ou sucessivas divisões de elementos, onde os elementos são agregados ou desagregados. Os métodos hierárquicos são subdivididos em métodos aglomerativos e divisivos.

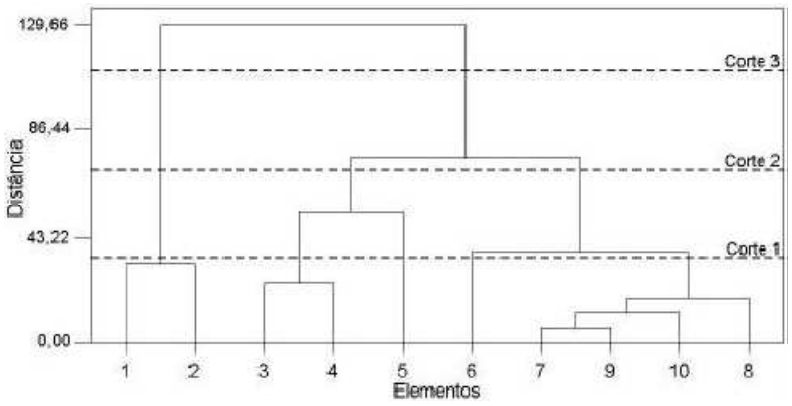
Os grupos, nos métodos hierárquicos, são geralmente representados por um diagrama bi-dimensional chamado de dendograma ou diagrama de árvore. Neste diagrama, cada ramo representa um elemento, enquanto a raiz representa o agrupamento de todos os elementos.

A figura abaixo traz um exemplo de dendograma:




Através do dendograma e do conhecimento prévio sobre a estrutura dos dados, deve-se determinar uma distância de corte para definir quais serão os grupos formados. Essa decisão é subjetiva, e deve ser feita de acordo o objetivo da análise e o número de grupos desejados.

**Exemplo.** Considerando o dendograma da figura acima, pode-se verificar que com três diferentes cortes, obtemos diferentes grupos:




No corte 1, verifica-se a existência de cinco grupos, sendo (1,2), (3,4), (5), (6) e (7,9,10,8). No corte 2, o número de grupos diminui para três, sendo (1,2), (3,4,5) e (6,7,9,10,8). Considerando o corte 3, o número de grupos diminui para dois, sendo (1,2) e (3,4,5,6,7,9,10,8).

Dessa forma, o usuário deverá escolher o corte mais adequado às suas necessidades e à estrutura dos dados.

### 5.3.3 Métodos Aglomerativos

No método aglomerativo, cada elemento inicia-se representando um grupo, e a cada passo, um grupo ou elemento é ligado a outro de acordo com sua similaridade, até o último passo, onde é formado um grupo único com todos os elementos. Existe uma variedade de métodos aglomerativos, que são caracterizados de acordo com o critério utilizado para definir as distâncias entre grupos.

Entretanto, a maioria dos métodos parecem ser formulações alternativas de três grandes conceitos de agrupamento aglomerativo (ANDERBERG, 1973):

1) Métodos de ligação (single linkage, complete linkage, average linkage, median linkage);
2) Métodos de centróide;
3) Métodos de soma de erros quadráticos ou variância (método de Ward).

Os métodos aglomerativos possuem a complexidade de tempo da ordem de $O(n^2 \log n)$ e a complexidade de espaço da ordem de $O(n^2)$, onde $n$ é o número de elementos (JAIN, 1999).

De modo geral, os métodos aglomerativos utilizam os passos de um algoritmo padrão, conforme descrito na figura abaixo. A diferença entre os métodos ocorre no passo 5, onde a função distância é definida de acordo com cada método (JOHNSON, 1992).

**Entrada:** Uma base de dados com ***N*** elementos.
Saída: Um conjunto de grupos.
1. Iniciar com ***N*** grupos, contendo um elemento em cada grupo e uma matriz de similaridade $D_N \times_N$;
2. Repetir;
3. Localizar a menor distância $d_{UV}$ (maior similaridade);
4. Atualizar a matriz ***D***, retirando os elementos ***U*** e ***V***;
5. Atualizar a matriz ***D***, adicionando as novas distâncias do grupo ***(U, V)***;
6. Até $N$-1, quando todos elementos estarão em um único grupo.

**Exemplo.**Considerando-se os elementos da tabela utilizada acima, obteve-se a matriz D abaixo,onde,aplicando uma iteração do algoritmo padrão, temos:

$$D = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{bmatrix} 0 & 4{,}47 & 4 & 2 & 2{,}24 & 2{,}83 \\ 4{,}47 & 0 & 2 & 4 & 2{,}24 & 7{,}21 \\ 4 & 2 & 0 & 4{,}47 & 2{,}24 & 6{,}32 \\ 2 & 4 & 4{,}47 & 0 & 2{,}24 & 4{,}47 \\ 2{,}24 & 2{,}24 & 2{,}24 & 2{,}24 & 0 & 5 \\ 2{,}83 & 7{,}21 & 6{,}32 & 4{,}47 & 5 & 0 \end{bmatrix}$$

A menor distância $d_{UV}$ está entre os elementos 1 e 4, e 3 e 2. Portanto, $d_{14} = d_{32} = 2$.
Como a distância $d_{14}$ é a primeira a aparecer na matriz, os elementos 1 e 4 serão considerados no primeiro agrupamento.

Assim, a matriz resultante após uma iteração do algoritmo será:

$$\begin{array}{c} (1,4) \\ 2 \\ 3 \\ 5 \\ 6 \end{array} \begin{bmatrix} 0 & d(1,4)2 & d(1,4)3 & d(1,4)5 & d(1,4)6 \\ d2(1,4) & 0 & 2 & 2,24 & 7,21 \\ d3(1,4) & 2 & 0 & 2,24 & 6,32 \\ d5(1,4) & 2,24 & 2,24 & 0 & 5 \\ d6(1,4) & 7,21 & 6,32 & 5 & 0 \end{bmatrix}$$

### 5.3.3.1 Método Single Linkage ou ligação por vizinho mais próximo

O método de ligação por vizinho mais próximo emprega a distância de valor mínimo:

$$d(UV)W = \min(dUW, dVW)$$

**Exemplo.** Utilizando a matriz do exemplo anterior, as distâncias mínimas encontradas são:

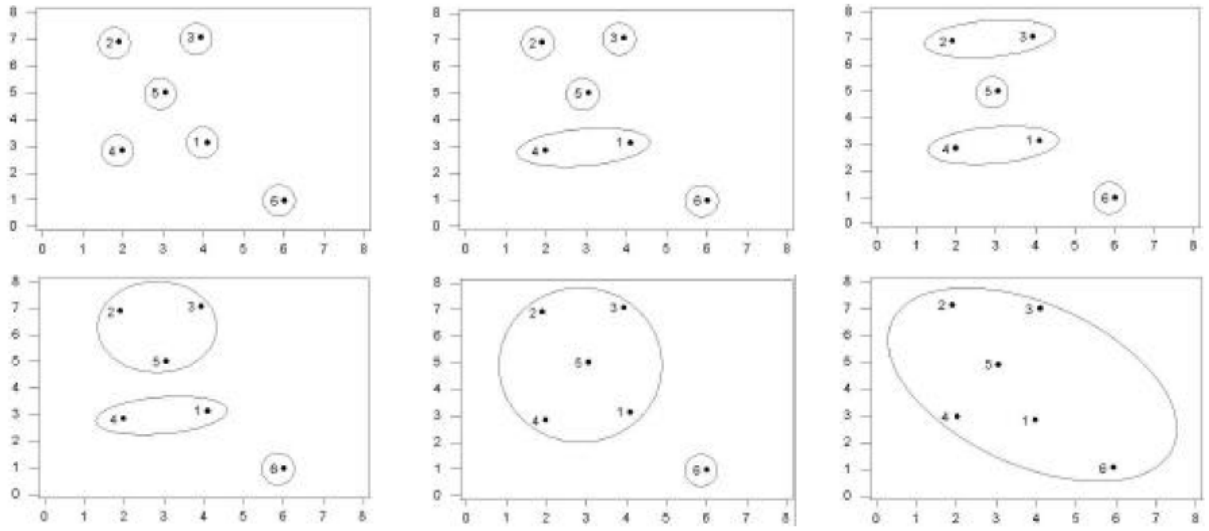
| CÁLCULO DAS DISTÂNCIAS | VALOR |
|---|---|
| $d_{(1,4)2} = \min(d_{12}, d_{42}) = \min(4,47;4)$ | 4 |
| $d_{(1,4)3} = \min(d_{13}, d_{43}) = \min(4;4,47)$ | 4 |
| $d_{(1,4)5} = \min(d_{15}, d_{45}) = \min(2,24;2,24)$ | 2,24 |
| $d_{(1,4)6} = \min(d_{16}, d_{46}) = \min(2,83;4,47)$ | 2,83 |

Assim a matriz resultante será:

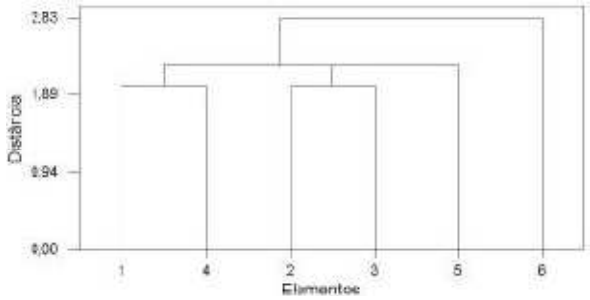
$$\begin{array}{c} (1,4) \\ 2 \\ 3 \\ 5 \\ 6 \end{array} \begin{bmatrix} 0 & 4 & 4 & 2,24 & 2,83 \\ 4 & 0 & 2 & 2,24 & 7,21 \\ 4 & 2 & 0 & 2,24 & 6,32 \\ 2,24 & 2,24 & 2,24 & 0 & 5 \\ 2,83 & 7,21 & 6,32 & 5 & 0 \end{bmatrix}$$

As demais iterações serão repetidas como descrito no algoritmo padrão, sempre calculando, para esse método, as distâncias mínimas entre os elementos ou grupos.

A seqüência abaixo traz os grupos formados a cada iteração do algoritmo:



A figura abaixo traz o dendograma gerado pelo método de ligação do vizinho mais próximo

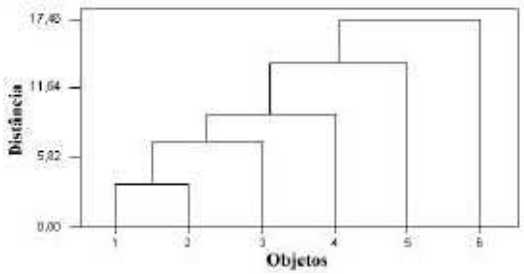



Algumas características desse método são (ANDERBERG, 1973):

- Em geral, grupos muito próximos podem não ser identificados;
- Permite detectar grupos de formas não-elípticas;
- Apresenta pouca tolerância a ruído, pois tem tendência a incorporar os ruídos em um grupo já existente;
- Apresenta bons resultados tanto para distâncias Euclidianas quanto para outras distâncias;
- Tendência a formar longas cadeias (encadeamento).

Encadeamento é um termo que descreve a situação onde há um primeiro grupo de um ou mais elementos que passa a incorporar, a cada iteração, um grupo de apenas um elemento.

Assim, é formada uma longa cadeia, onde torna-se difícil definir um nível de corte para classificar os elementos em grupos (ROMESBURG, 1984), conforme figura a seguir:




Esse fenômeno ocorre em dados com a distribuição mostrada anteriormente, onde cada elemento tem como vizinho mais próximo o grupo formado na iteração anterior.

### 5.3.3.2 Método Complete Linkage ou ligação por vizinho mais distante

Nesse método, é empregada a distância máxima, dada por:

$$d(UV)W = \max(dUW\ ,\ dVW)$$

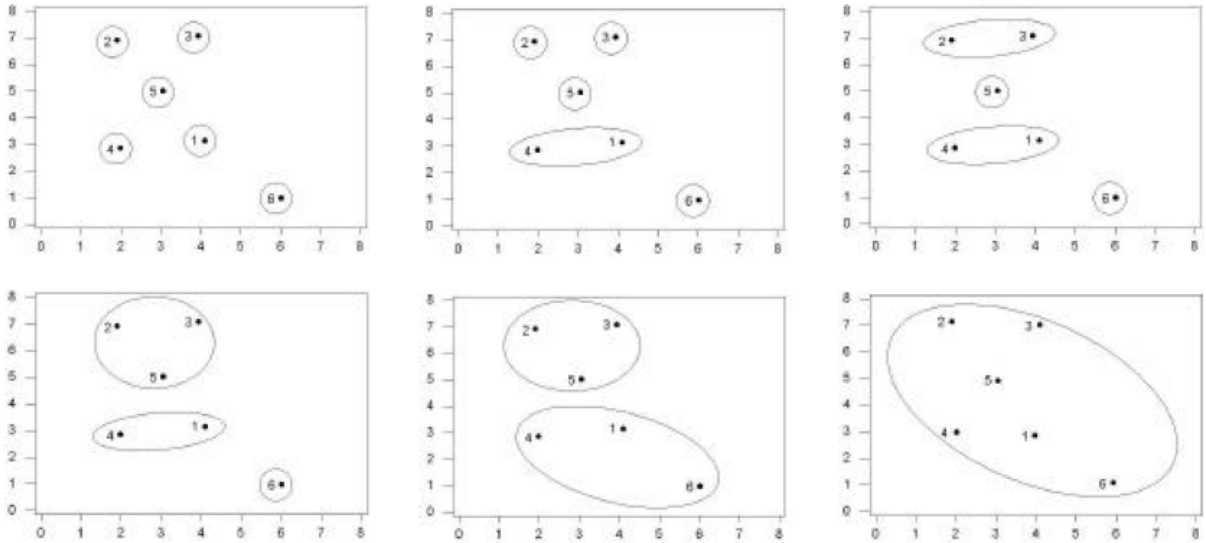
Exemplo. Utilizando a matriz do exemplo anterior, as distâncias máximas encontradas são:

| CÁLCULO DAS DISTÂNCIAS | VALOR |
|---|---|
| $d_{(1,4)2} = \max(d_{12}, d_{42}) = \max(4{,}47;4)$ | 4,47 |
| $d_{(1,4)3} = \max(d_{13}, d_{43}) = \max(4;4{,}47)$ | 4,47 |
| $d_{(1,4)5} = \max(d_{15}, d_{45}) = \max(2{,}24;2{,}24)$ | 2,24 |
| $d_{(1,4)6} = \max(d_{16}, d_{46}) = \max(2{,}83;4{,}47)$ | 4,47 |

Assim, a matriz resultante será:

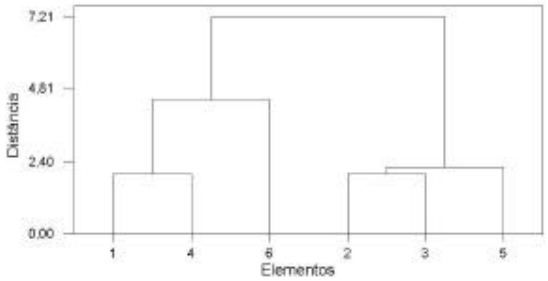
$$\begin{array}{c} (1,4) \\ 2 \\ 3 \\ 5 \\ 6 \end{array} \begin{bmatrix} 0 & 4{,}47 & 4{,}47 & 2{,}24 & 4{,}47 \\ 4{,}47 & 0 & 2 & 2{,}24 & 7{,}21 \\ 4{,}47 & 2 & 0 & 2{,}24 & 6{,}32 \\ 2{,}24 & 2{,}24 & 2{,}24 & 0 & 5 \\ 4{,}47 & 7{,}21 & 6{,}32 & 5 & 0 \end{bmatrix}$$

As demais iterações serão repetidas como descrito no algoritmo padrão, sempre calculando, para esse método, as distâncias máximas entre os elementos ou grupos.
A figura a seguir traz a seqüência dos grupos formados em cada iteração do algoritmo.



De acordo com a figura acima, pode-se verificar que na quarta iteração do algoritmo no método de ligação por vizinho mais distante, os agrupamentos são realizados de maneira diferente ao método de ligação por vizinho mais próximo.

A figura abaixo traz o dendograma gerado pelo do método de ligação por vizinho mais distante.




Algumas características desse método são (KAUFMANN, 1990):

- Apresenta bons resultados tanto para distâncias Euclidianas quanto para outras distâncias;
- Tendência a formar grupos compactos;
- Os ruídos demoram a serem incorporados ao grupo.

Os métodos de ligação por mais próximo e por vizinho mais distante trabalham em direções opostas. Se eles apresentam resultados semelhantes, significa que o grupo está bem definido no espaço, ou seja, o grupo é real. Mas se ocorre o contrário, os grupos provavelmente não existem (ROMESBURG, 1984).

### 5.3.3.3 Método Average Linkage ou ligação por média

Nesse método, a função distância é definida por:

$$d_{(UV)W} = \frac{(N_u . d_{UW} + N_v . d_{VW})}{N_u + N_v}$$

- onde: $N_U$ e $N_V$ são os números de elementos no grupo U e V, respectivamente;
- $d_{UW}$ e $d_{VW}$ são as distâncias entre os elementos UW e VW, respectivamente.

**Exemplo.** Considerando-se a matriz do exemplo anterior, as distâncias médias são calculadas a seguir:

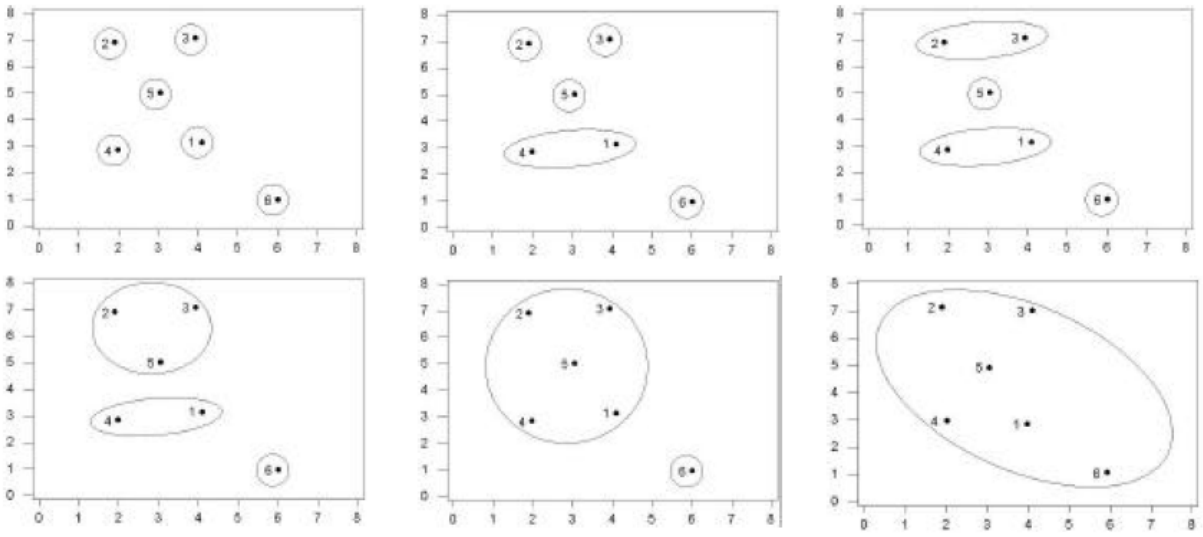
| CÁLCULO DAS DISTÂNCIAS | VALOR |
|---|---|
| $d_{(1,4)2} = \frac{(1.4{,}47 + 1.4)}{1+1}$ | 4,24 |
| $d_{(1,4)3} = \frac{(1.4 + 1.4{,}47)}{1+1}$ | 4,24 |
| $d_{(1,4)5} = \frac{(1.2{,}24 + 1.2{,}24)}{1+1}$ | 2,24 |
| $d_{(1,4)6} = \frac{(1.2{,}83 + 1.4{,}47)}{1+1}$ | 3,65 |

Assim, a matriz resultante será:

$$\begin{array}{c} (1,4) \\ 2 \\ 3 \\ 5 \\ 6 \end{array} \begin{bmatrix} 0 & 4{,}24 & 4{,}24 & 2{,}24 & 3{,}65 \\ 4{,}24 & 0 & 2 & 2{,}24 & 7{,}21 \\ 4{,}24 & 2 & 0 & 2{,}24 & 6{,}32 \\ 2{,}24 & 2{,}24 & 2{,}24 & 0 & 5 \\ 3{,}65 & 7{,}21 & 6{,}32 & 5 & 0 \end{bmatrix}$$

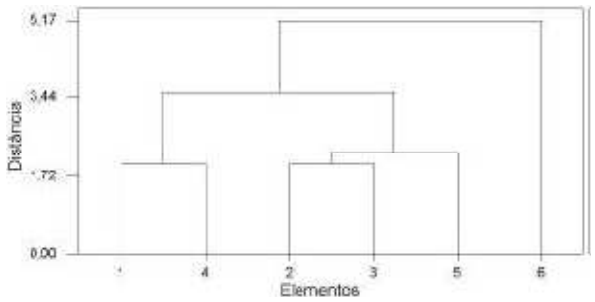
As demais iterações serão repetidas como descrito no algoritmo padrão, sempre calculando, para esse método, as distâncias médias entre os elementos ou grupos.

A figura abaixo traz a seqüência dos grupos formados em cada iteração do algoritmo:



De acordo com a figura acima, pode-se verificar que na quarta iteração do algoritmo no método de ligação por média, os agrupamentos são realizados de maneira diferente do método de ligação por vizinho mais distante e igual ao método de ligação por vizinho mais próximo.

A figura abaixo traz o dendograma gerado pelo do método de ligação por média.




Algumas características desse método são (KAUFMANN, 1990):

- Menor sensibilidade à ruídos que o os métodos de ligação por vizinho mais próximo e por vizinho mais distante;
- Apresenta bons resultados tanto para distâncias Euclidianas quanto para outras distâncias;

- Tendência a formar grupos com número de elementos similares.

### 5.3.3.4 Método Centroid Linkage ou ligação por centróide

Nesse método, a função distância é definida por:

$$d_{(UV)W} = \frac{N_U . d_{UW} + N_V . d_{VW}}{N_U + N_V} - \frac{N_U . N_V . d_{UV}}{(N_U + N_V)^2}$$

- onde: $N_U$ e $N_V$ são os números de elementos no grupo U e V, respectivamente;
- $d_{UW}$ e $d_{VW}$ são as distâncias entre os elementos UW e VW, respectivamente.

Exemplo. Utilizando a matriz do exemplo anterior, as distâncias são calculadas a seguir:

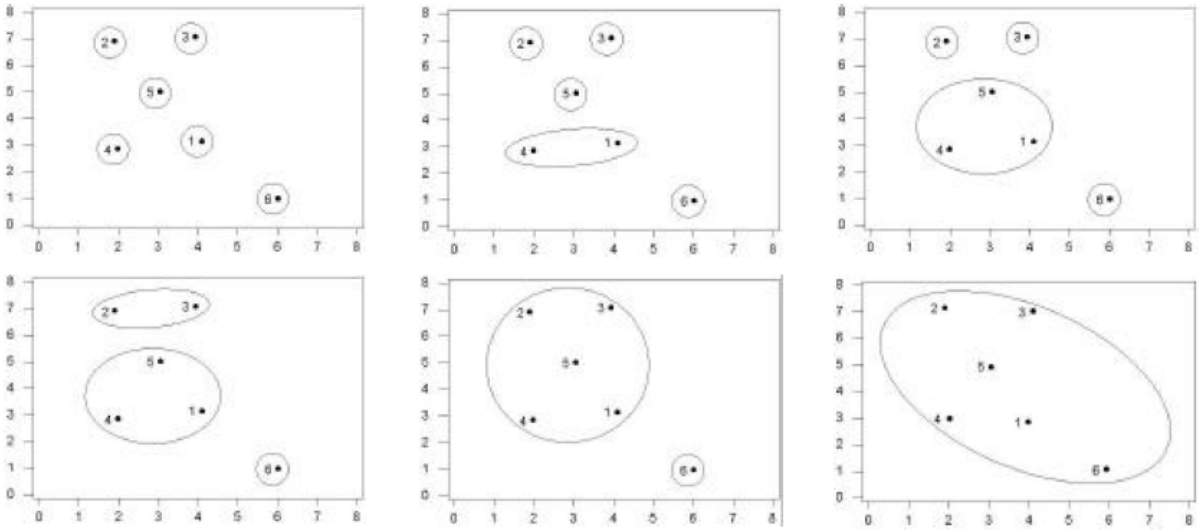
| CÁLCULO DAS DISTÂNCIAS | VALOR |
|---|---|
| $d_{(1,4)2} = \frac{1.4,47 + 1.4}{1+1} - \frac{1.1.1}{(1+1)^2}$ | 3,99 |
| $d_{(1,4)3} = \frac{1.4 + 1.4,47}{1+1} - \frac{1.1.1}{(1+1)^2}$ | 3,99 |
| $d_{(1,4)5} = \frac{1.2,24 + 1.2,24}{1+1} - \frac{1.1.1}{(1+1)^2}$ | 1,99 |
| $d_{(1,4)6} = \frac{1.2,83 + 1.4,47}{1+1} - \frac{1.1.1}{(1+1)^2}$ | 3,4 |

Assim, a matriz resultante será:

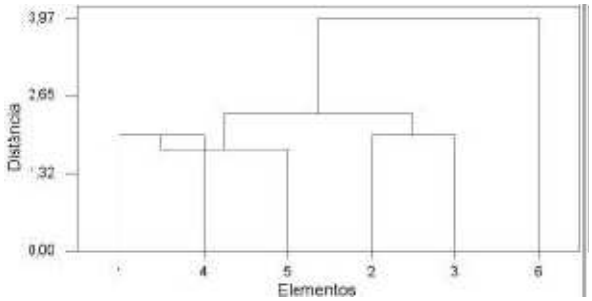
$$\begin{array}{c} (1,4) \\ 2 \\ 3 \\ 5 \\ 6 \end{array} \begin{bmatrix} 0 & 3,99 & 3,99 & 1,99 & 3,4 \\ 3,99 & 0 & 2 & 2,24 & 7,21 \\ 3,99 & 2 & 0 & 2,24 & 6,32 \\ 1,99 & 2,24 & 2,24 & 0 & 5 \\ 3,4 & 7,21 & 6,32 & 5 & 0 \end{bmatrix}$$

As demais iterações serão repetidas como descrito no algoritmo padrão, sempre calculando, para esse método, as distâncias entre os centróides.

A figura abaixo traz a seqüência dos grupos formados em cada iteração do algoritmo:



A figura abaixo mostra o dendograma gerado pelo do método de ligação por centróide:

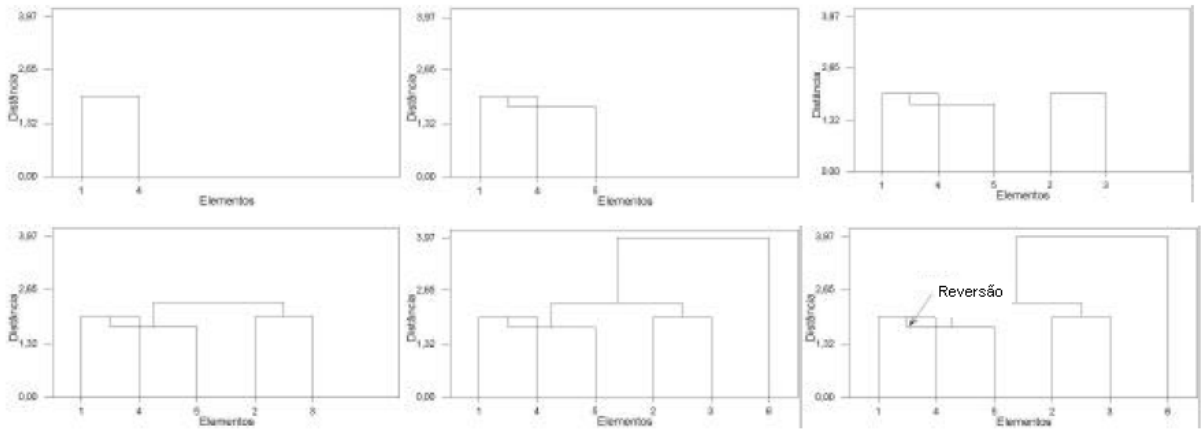



Como características desse método, encontram-se:
- Robustez à presença de ruídos;
- Devido ao fenômeno da reversão, o método não é muito utilizado.

O fenômeno da reversão ocorre quando a distância entre centróides é menor que a distância entre grupos já formados, isso fará com que os novos grupos se formem ao um nível inferior aos grupos existentes, tornando o dendograma confuso.

O centróide é o ponto médio no espaço multidimensional e representa o centro de gravidade do respectivo grupo.

No exemplo, observa-se o fenômeno da reversão, pois na primeira ligação, entre os elementos 1 e 4, a distância entre centróides foi maior que na segunda ligação, entre o grupo (1,4) e o elemento 5.




### 5.3.3.5 Método Median Linkage ou ligação por mediana

Nesse método, a função distância é dada por:

$$d_{(UV)W} = \frac{d_{UW} + d_{VW}}{2} - \frac{d_{UV}}{4}$$

- onde: $d_{UW}$ e $d_{VW}$ são as distâncias entre os elementos UW e VW, respectivamente.

**Exemplo.** Utilizando a matriz do exemplo anterior, as distâncias são calculadas a seguir:

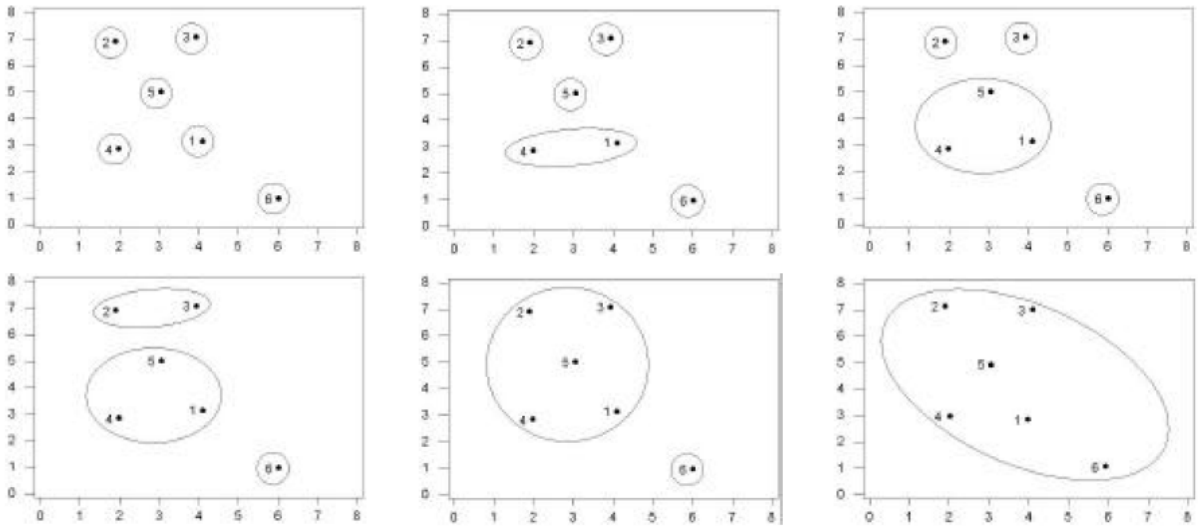
| CÁLCULO DAS DISTÂNCIAS | VALOR |
|---|---|
| $d_{(1,4)2} = \frac{4,47 + 4}{2} - \frac{1}{4}$ | 3,99 |
| $d_{(1,4)3} = \frac{4 + 4,47}{2} - \frac{1}{4}$ | 3,99 |
| $d_{(1,4)5} = \frac{2,4 + 2,4}{2} - \frac{1}{4}$ | 1,99 |
| $d_{(1,4)6} = \frac{2,83 + 4,47}{2} - \frac{1}{4}$ | 3,4 |

Assim, a matriz resultante será:

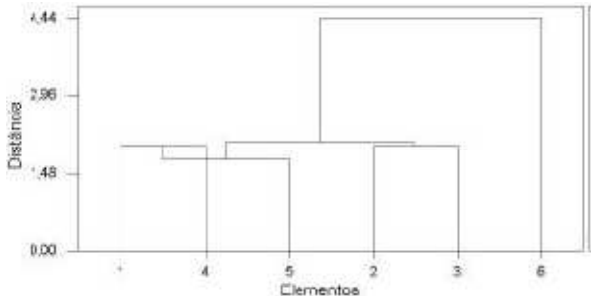
$$\begin{array}{c} (1,4) \\ 2 \\ 3 \\ 5 \\ 6 \end{array} \begin{bmatrix} 0 & 3{,}99 & 3{,}99 & 1{,}99 & 3{,}4 \\ 3{,}99 & 0 & 2 & 2{,}24 & 7{,}21 \\ 3{,}99 & 2 & 0 & 2{,}24 & 6{,}32 \\ 1{,}99 & 2{,}24 & 2{,}24 & 0 & 5 \\ 3{,}4 & 7{,}21 & 6{,}32 & 5 & 0 \end{bmatrix}$$

As demais iterações serão repetidas como descrito no algoritmo padrão, sempre calculando, para esse método, as distâncias entre os elementos ou grupos de acordo com a equação da distância descrita acima.

A figura abaixo traz a seqüência dos grupos formados em cada iteração do algoritmo.



Abaixo mostra o dendograma gerado pelo do método de ligação por mediana:




Algumas características referentes a esse método são:

- Apresenta resultado satisfatório quando os grupos possuem tamanhos diferentes;
- Pode apresentar resultado diferente quando permutado os elementos na matriz de similaridade;
- Robustez à presença de outliers.

### 5.3.3.6 Método de ligação de Ward

Nesse método, a função distância é dada por:

$$d_{(UV)W} = \frac{((N_W + N_U).d_{UW} + (N_W + N_V).d_{VW} - N_W.d_{UV})}{N_W + N_U + N_V}$$

- onde: $N_U$ e $N_W$ são os números de elementos no grupo U e V, respectivamente;
- $d_{UW}$ e $d_{VW}$ são as distâncias entre os elementos UW e VW, respectivamente.

**Exemplo.** Utilizando a mesma matriz do exemplo anterior, as distâncias são calculadas a seguir:

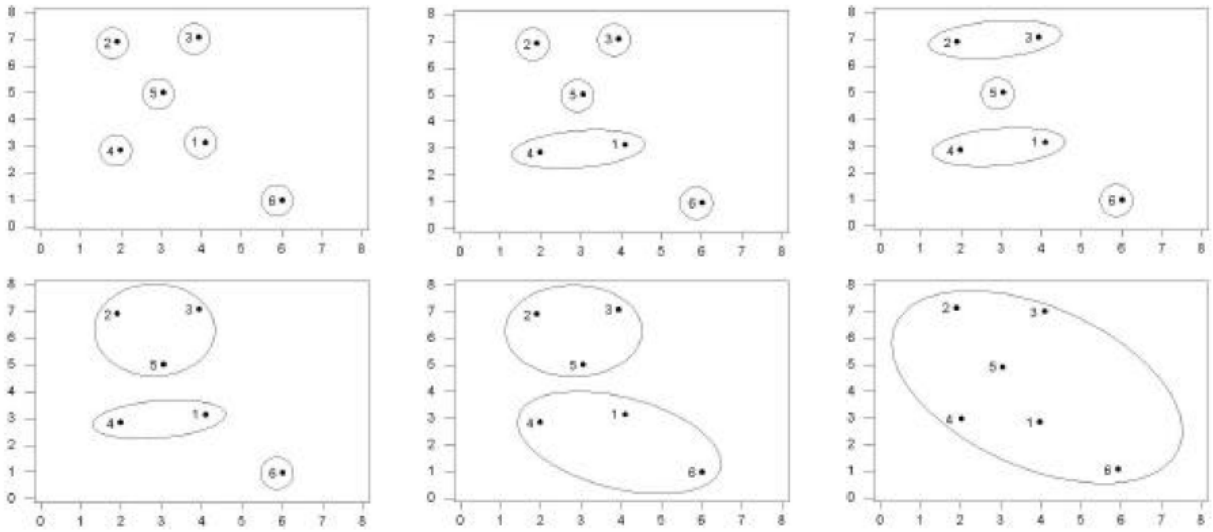
| CÁLCULO DAS DISTÂNCIAS | VALOR |
|---|---|
| $d_{(1,4)2} = \frac{((1+1).4{,}47 + (1+1).4 - 1.1)}{3}$ | 5,31 |
| $d_{(1,4)3} = \frac{((1+1).4 + (1+1).4{,}47 - 1.1)}{3}$ | 5,31 |
| $d_{(1,4)5} = \frac{((1+1).2{,}24 + (1+1).2{,}24 - 1.1)}{3}$ | 2,65 |
| $d_{(1,4)6} = \frac{((1+1).2{,}83 + (1+1).4{,}47 - 1.1)}{3}$ | 4,53 |

Assim, a matriz resultante será:

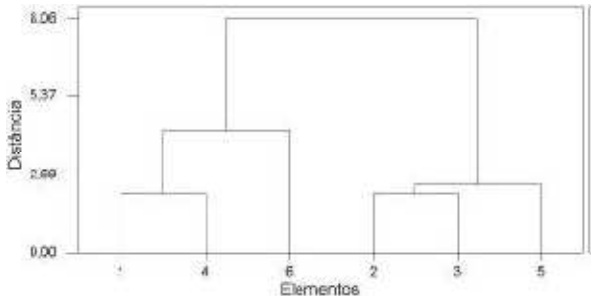
$$\begin{array}{c} (1,4) \\ 2 \\ 3 \\ 5 \\ 6 \end{array} \begin{bmatrix} 0 & 5{,}31 & 5{,}31 & 2{,}65 & 4{,}53 \\ 5{,}31 & 0 & 2 & 2{,}24 & 7{,}21 \\ 5{,}31 & 2 & 0 & 2{,}24 & 6{,}32 \\ 2{,}65 & 2{,}24 & 2{,}24 & 0 & 5 \\ 4{,}53 & 7{,}21 & 6{,}32 & 5 & 0 \end{bmatrix}$$

Os demais passos serão repetidos como descrito no algoritmo padrão, sempre calculando, para esse método, as distâncias entre os elementos ou grupos de acordo com a equação acima.

A figura abaixo traz a seqüência dos grupos formados em cada iteração do algoritmo.



A figura abaixo mostra o dendograma gerado pelo do método de ligação de Ward.




Algumas características desse método são:

- Apresenta bons resultados tanto para distâncias euclidianas quanto para outras distâncias;
- Pode apresentar resultados insatisfatórios quando o número de elementos em cada grupo é praticamente igual;
- Tem tendência a combinar grupos com poucos de elementos;
- Sensível à presença de outliers.

Segue abaixo um resumo dos métodos aglomerativos:

| MÉTODO | DISTÂNCIA | CARACTERÍSTICAS |
|---|---|---|
| Ligação por vizinho mais próximo | $d_{(UV)W} = \min(d_{UW}, d_{VW})$ | Sensibilidade à ruídos. Encadeamento. |
| Ligação por vizinho mais distante | $d_{(UV)W} = \max(d_{UW}, d_{VW})$ | Tendência a formar grupos compactos. |
| Ligação por média | $d_{(UV)W} = \frac{(N_u . d_{UW} + N_v . d_{VW})}{N_u + N_v}$ | Tendência a formar grupos com número de elementos similares. |
| Ligação por centróide | $d_{(UV)W} = \frac{N_U . d_{UW} + N_V . d_{VW}}{N_U + N_V} - \frac{N_U . N_V . d_{UV}}{(N_U + N_V)^2}$ | Robustez à ruídos. Reversão. |
| Ligação por mediana | $d_{(UV)W} = \frac{d_{UW} + d_{VW}}{2} - \frac{d_{UV}}{4}$ | Robustez à ruídos. |
| ***Ligação de* Ward** | $d_{(UV)W} = \frac{((N_W + N_U).d_{UW} + (N_W + N_V).d_{VW} - N_W . d_{UV})}{N_W + N_U + N_V}$ | Sensibilidade à ruídos. |

### 5.3.4 Métodos não hierárquicos ou por particionamento

Os métodos não-hierárquicos, ou por particionamento, foram desenvolvidos para agrupar elementos em K grupos, onde K é a quantidade de grupos definida previamente.Nem todos valores de K apresentam grupos satisfatórios, sendo assim, aplica-se o método várias vezes para diferentes valores de K, escolhendo os resultados que apresentem melhor interpretação dos grupos ou uma melhor representação gráfica (BUSSAB, 1990).

A idéia central da maioria dos métodos por particionamento é escolher uma partição inicial dos elementos e, em seguida, alterar os membros dos grupos para obter-se a melhor partição (ANDERBERG, 1973).Quando comparado com o método hierárquico, o método por particionamento é mais rápido porque não é necessário calcular e armazenar, durante o processamento, a matriz de similaridade.Em geral, os métodos por particionamento diferem entre si pela maneira que constituem a melhor partição. Os métodos por particionamento mais conhecidos são o método k-means (k-médias) e o método k-medoid (k-medóides), e são descritos a seguir.

#### 5.3.4.1 Método *k – Means*

O método k-means toma um parâmetro de entrada, K, e particiona um conjunto de N elementos em K grupos, conforme figura abaixo:

**Entrada:** O número de grupos, ***K***, e a base de dados com ***N*** elementos.
**Saída:** Um conjunto de ***K*** grupos.
1. Escolher arbitrariamente ***K*** elementos da base de dados como os centros iniciais dos grupos;
2. Repetir;
3. (re)Atribua cada elemento ao grupo ao qual o elemento é mais similar, de acordo com o valor médio dos elementos no grupo;
4. Atualizar as médias dos grupos, calculando o valor médio dos elementos para cada grupo;
5. Até que não haja mudanças de elementos de um grupo para outro.

Esse método possui uma complexidade de tempo da ordem de O(nkl) e uma complexidade de espaço é da ordem de O(k + n) , onde n é o número de elementos, k é o número de grupos e l é o número de iterações do algoritmo (JAIN et al, 1999).

**Exemplo.** Considerando os elementos e variáveis da tabela abaixo, assumindo K=2, tais como (1,2,3) e (4,5,6), obtemos:

| ELEMENTO | VARIÁVEIS | |
|---|---|---|
| | X | Y |
| 1 | 4 | 3 |
| 2 | 2 | 7 |
| 3 | 4 | 7 |
| 4 | 2 | 3 |
| 5 | 3 | 5 |
| 6 | 6 | 1 |

Calculando a média dos grupos (1,2,3) e (4,5,6), temos:

| GRUPO | VARIÁVEIS | |
|---|---|---|
| | X | Y |
| (1,2,3) | $\frac{4+2+4}{3} = 3{,}33$ | $\frac{3+7+7}{3} = 5{,}67$ |
| (4,5,6) | $\frac{2+3+6}{3} = 3{,}67$ | $\frac{3+5+1}{3} = 3$ |

Calculando a distância Euclidiana de cada objeto ao centróide dos grupos, obtém-se:

| DISTÂNCIA DE CADA ELEMEMTO AO CENTRÓIDE DOS GRUPOS | VALOR |
|---|---|
| $de_{(1(1,2,3))} = \sqrt{(4-4)^2 + (3-3)^2}$ | 0 |
| $de_{(1(4,5,6))} = \sqrt{(4-2)^2 + (3-7)^2}$ | 4,47 |
| $de_{(2(1,2,3))} = \sqrt{(2-4)^2 + (7-3)^2}$ | 4,47 |
| $de_{(2(4,5,6))} = \sqrt{(2-2)^2 + (7-7)^2}$ | 0 |
| $de_{(3(1,2,3))} = \sqrt{(4-4)^2 + (7-3)^2}$ | 4 |
| $de_{(3(4,5,6))} = \sqrt{(4-2)^2 + (7-7)^2}$ | 2 |

Na tabela acima, verifica-se que os elementos 2 e 3 estão mais próximos do grupo (4,5,6). Assim, eles serão retirado do grupo (1,2,3) e associados ao grupo (4,5,6).

Recalculando o centróide dos grupos, temos:

| GRUPO | COORDENADAS | |
|---|---|---|
| | X | Y |
| 1 | 4 | 3 |
| (2,3,4,5,6) | $\frac{(2+4+2+3+6)}{5} = 3{,}4$ | $\frac{(7+7+3+5+1)}{5} = 4{,}6$ |

Calculando a distância Euclidiana de cada objeto ao centróide dos grupos, temos:

| DISTÂNCIA DE CADA ELEMEMTO AO CENTRÓIDE DOS GRUPOS | RESULTADO |
|---|---|
| $de_{(4(1))} = \sqrt{(2-4)^2 + (3-3)^2}$ | 2 |
| $de_{(4(2,3,4,5,6))} = \sqrt{(2-3{,}4)^2 + (3-4{,}6)^2}$ | 2,17 |
| $de_{(5(1))} = \sqrt{(3-4)^2 + (5-3)^2}$ | 2,24 |
| $de_{(5(2,3,4,5,6))} = \sqrt{(3-3{,}4)^2 + (5-4{,}6)^2}$ | 0,57 |
| $de_{(6(1))} = \sqrt{(6-4)^2 + (1-3)^2}$ | 2,83 |
| $de_{(6(2,3,4,5,6))} = \sqrt{(6-3{,}4)^2 + (1-4{,}6)^2}$ | 4,44 |

Na tabela acima, verifica-se que os elementos 4 e 6 estão mais próximos do grupo (1). Assim, eles serão retirados do grupo (2,3,4,5,6) e associados ao grupo (1).

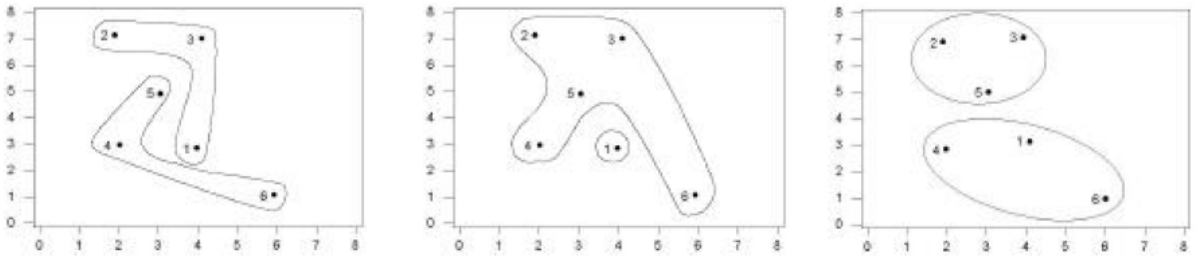
Recalculando o centróide dos grupos, obtém-se:

| GRUPO | COORDENADAS | |
|---|---|---|
| | $\bar{x}$ | $\bar{Y}$ |
| (1,4,6) | $\frac{(4+2+6)}{3} = 4$ | $\frac{(3+3+1)}{3} = 2{,}33$ |
| (2,3,,5) | $\frac{(2+4+3)}{3} = 3$ | $\frac{(7+7+5)}{3} = 6{,}33$ |

Calculando a distância Euclidiana de cada elemento ao centróide dos grupos, temos:

| DISTÂNCIA DE CADA ELEMEMTO AO CENTRÓIDE DOS GRUPOS | RESULTADO |
|---|---|
| $de_{(1(1,4,6))} = \sqrt{(4-4)^2 + (3-2{,}33)^2}$ | 0,45 |
| $de_{(1(2,3,5))} = \sqrt{(4-3)^2 + (3-6{,}33)^2}$ | 12,09 |
| $de_{(2(1,4,6))} = \sqrt{(2-4)^2 + (7-2{,}33)^2}$ | 25,81 |
| $de_{(2(2,3,5))} = \sqrt{(2-3)^2 + (7-6{,}33)^2}$ | 1,45 |
| $de_{(3(1,4,6))} = \sqrt{(4-4)^2 + (7-2{,}33)^2}$ | 21,81 |
| $de_{(3(2,3,5))} = \sqrt{(4-3)^2 + (7-6{,}33)^2}$ | 1,45 |
| $de_{(4(1,4,6))} = \sqrt{(2-4)^2 + (3-2{,}33)^2}$ | 4,45 |
| $de_{(4(2,3,5))} = \sqrt{(2-3)^2 + (3-6{,}33)^2}$ | 12,09 |
| $de_{(5(1,4,6))} = \sqrt{(3-4)^2 + (5-2{,}33)^2}$ | 8,13 |
| $de_{(5(2,3,5))} = \sqrt{(3-3)^2 + (5-6{,}33)^2}$ | 1,77 |
| $de_{(6(1,4,6))} = \sqrt{(6-4)^2 + (1-2{,}33)^2}$ | 5,77 |
| $de_{(6(2,3,5))} = \sqrt{(6-3)^2 + (1-6{,}33)^2}$ | 37,41 |

Verificando a necessidade de realocação, observa-se que cada elemento está corretamente associado ao grupo com o centróide mais próximo, ou seja, cada elemento possui menor distância em relação ao grupo no qual faz parte do que em relação ao outro grupo. Sendo assim, o processo encerra-se com os grupos (1,4,6) e (2,3,5) conforme figura abaixo :



Algumas características desse método são:

- Sensibilidade a ruídos, uma vez que um elemento com um valor extremamente alto pode distorcer a distribuição dos dados;
- Tendência a formar grupos esféricos;
- O número de grupos é o mesmo durante todo o processo;
- Inadequado para descobrir grupos com formas não convexas ou de tamanhos muito diferentes.

#### 5.3.4.2 Método *k – Medoid*

O método k-medoid utiliza o valor médio dos elementos em um grupo como um ponto referência, chamado de medóide. Esse é o elemento mais centralmente localizado em um grupo. A estratégia básica é encontrar K grupos em N elementos e, arbitrariamente, encontrar um elemento representativo (medóide) para cada grupo. Cada elemento remanescente é agrupado com o medóide ao qual ele é mais similar. A estratégia, então, iterativamente, troca um dos medóides por um dos não medóides enquanto a qualidade do agrupamento resultante é melhorada. Abaixo o algoritmo:

**Entrada:** O número de grupos, ***K***, e a base de dados com ***N*** elementos.
**Saída:** Um conjunto de ***K*** grupos.
1. Escolher, arbitrariamente, ***K*** elementos da base de dados como os medóides iniciais dos grupos;
2. Repetir;
3. atribua cada elemento remanescente ao grupo com o medóide mais próximo;
4. aleatoriamente, selecione um elemento que não esteja como medóide, ***r***;
5. calcule o custo total, $S$, de trocar o medóide $O_j$ pelo elemento ***r***;
6. se $S < 0$ então troque $O_j$ por ***r*** para formar o novo conjunto de k-medóides;
7. Até que não haja mudança de objetos de um grupo para outro.

**Exemplo.** Considerando os mesmos elementos do exemplo anterior, assumindo K=2 e escolhendo, aleatoriamente, os elementos 1 e 4 como medóides iniciais, temos:

| ELEMENTOS (i) | $d_{i1}$ | $d_{i4}$ | $Min(d_{i1},d_{i4})$ | MEDÓIDE MAIS PRÓXIMO |
|---|---|---|---|---|
| 1 | 0 | 2 | 0 | 1 |
| 2 | 4,47 | 4 | 4 | 4 |
| 3 | 4 | 4,47 | 4 | 1 |
| 4 | 2 | 0 | 0 | 4 |
| 5 | 2,24 | 2,24 | 2,24 | 1 |
| 6 | 2,83 | 4,47 | 2,83 | 1 |
| | | | $Média_{1,4} = 2,18$ | |

Com base na tabela acima, verificamos que os elementos 2, 5 e 6 são agrupados ao elemento 1, pois estão mais próximos desse medóide. O único elemento mais próximo ao medóide 4 é o elemento 3, portanto será agrupado a esse.

A média das similaridades mínimas, calculada acima, representa a qualidade dos grupos encontrados. Quanto menor esse valor, melhor é a qualidade dos grupos. Essa média é utilizada para encontrar o custo, S, na mudança de medóide. Para verificar a necessidade de mudança de medóide, selecionamos aleatoriamente o elemento 6 e calculamos o custo de trocar o medóide 1 por 6.

| ELEMENTOS (i) | $d_{i6}$ | $d_{i4}$ | $min(d_{i6},d_{i4})$ | MEDÓIDE MAIS PRÓXIMO |
|---|---|---|---|---|
| 1 | 2,83 | 2 | 2 | 4 |
| 2 | 7,21 | 4 | 4 | 4 |
| 3 | 6,32 | 4,47 | 4,47 | 4 |
| 4 | 4,47 | 0 | 0 | 4 |
| 5 | 5 | 2,24 | 2,24 | 4 |
| 6 | 0 | 4,47 | 0 | 6 |
| | | | $Média_{6,4} = 2,12$ | |

Na tabela acima, verificamos que os elementos 1, 2, 3 e 5 são agrupados ao elemento 4, pois estão mais próximos desse medóide. Nenhum elemento é agrupado ao medóide 6.Calculando o custo de troca do medóide 1 pelo 6, temos:

$$S_{1,6} = Média_{6,4} - Média_{1,4} = 2,12-2,18 = - 0,06$$

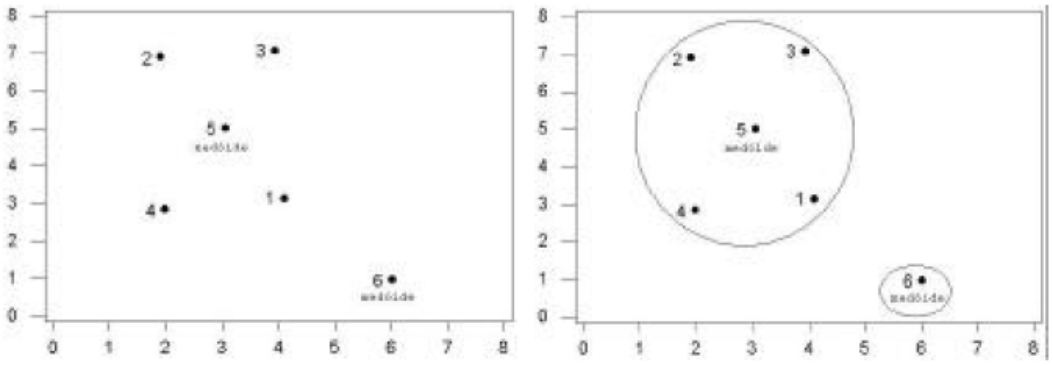
Como o custo é menor que zero, o medóide 1 é substituído pelo medóide 6.

O algoritmo prossegue selecionando novos não-medóides verificando a necessidade de substituir os medóides. Essa análise é feita para todos os pares de elementos.

Na tabela abaixo, temos um resumo dos resultados:

| MEDÓIDES | MÉDIAS |
|---|---|
| (1-4) | 2,18 |
| (4-6) | 2,12 |
| (2-6) | 1,85 |
| (1-2) | 1,51 |
| (1-5) | 1,55 |
| (1-6) | 2,12 |
| (1-3) | 1,51 |
| (2-3) | 2,76 |
| (2-4) | 1,79 |
| (2-5) | 1,91 |
| (3-5) | 1,91 |
| (3-4) | 1,79 |
| (4-5) | 1,83 |
| (4-6) | 2,12 |
| (3-6) | 1,92 |
| (5-6) | 1,49 |

Verificando a tabela acima, os medóides 5 e 6 possuem a menor média e, portanto, são os medóides finais e serão utilizados para formar os grupos.Os elementos 1, 2, 3 e 4 são agrupados ao medóide 5. Nenhum elemento é agrupado ao medóide 6.
Sendo assim, o processo encerra-se com os grupos (1,2,3,4,5) e (6), conforme figura abaixo:




Algumas características desse método são:

- Independente da ordem, os resultados serão os mesmos;
- Tendência a encontrar grupos esféricos;
- Processamento mais custoso que o k-means;
- Não aplicável à grandes bases de dados, pois o custo de processamento é alto;
- Mais robusto do que o k-means na presença de ruídos porque o medóide é menos influenciado pelos ruídos do que a média.

Uma forma de otimizar o método k-medoid para grandes bases de dados é considerar uma porção dos dados como uma amostra representativa, e escolher os medóides dessa amostra. Se a amostra é selecionada aleatoriamente, ela deverá representar bem o conjunto de dados originais, apresentando bons resultados (KAUFMAN, 1990).

## 5.3.5 Outros Métodos

Além das técnicas estatísticas de análise de cluster hierárquica e não-hierárquica, outras técnicas como algoritmos evolutivos (JAIN, 1999), agrupamentos fuzzy, redes neurais (mapas de Kohonen), entre outras, podem ser empregadas para formação de agrupamentos.

Aqui, ilustramos, brevemente, as técnicas de agrupamento fuzzy e mapas de Kohonen, indicando referências mais adequadas a essas técnicas.

### 5.3.5.1 Agrupamento Fuzzy

O agrupamento fuzzy é uma generalização dos métodos por particionamento e, assim como nos métodos por particionamento, também é necessário indicar o número inicial de grupos.

Nos métodos por particionamento, definem-se claramente em qual grupo ficará cada elemento, ou seja, são definidos agrupamentos rígidos (crisp cluster). Já os agrupamentos fuzzy permitem visualizar a grau de associação de cada elemento em cada grupo, que geralmente se verifica em domínios de dados reais, onde um elemento pertence a diferentes grupos, com diferentes graus de associação. A principal vantagem dos agrupamentos fuzzy em relação aos outros métodos por particionamento, é que ele fornece informações mais detalhadas sobre a estrutura dos dados, pois são apresentados os graus de associação de cada elemento a cada grupo, não tendo, portanto, a formação de agrupamentos rígidos. A desvantagem desse método é que a quantidade de coeficientes de associação cresce rapidamente com o aumento do número de elementos e de grupos. Entretanto, trata-se de uma técnica válida, pois ela associa graus de incerteza aos elementos nos grupos e, essa situação, em geral, se aproxima mais das características reais dos dados (KAUFMAN, 1990).

Um algoritmo de agrupamentos fuzzy bastante utilizado é o fuzzy c-means. Trata-se de um algoritmo iterativo que inicia com c valores arbitrários, e com base nesses valores, associa cada elemento ao valor ao qual possui menor distância, formando c grupos. Em seguida, calcula-se o centro de cada grupo formado, e os elementos são reassociados ao centro mais próximo. Assim, os cálculos prosseguem, iterativamente, até que as diferenças entre os centros do passo atual e do anterior sejam mínimas.

- ✓ **Aplicação do Método:**

Para caracterizar e agrupar os alunos que tem o melhor desempenho na prova usaremos método de *K-means* do software minitab com 8 clusters.Não nos preocuparemos em mostrar o dendograma dos dados, pois após a retirada das observações que tinham algum dado (resposta) em branco ficamos com em torno de 35 mil registros para cada ano.E esse numero de dados faz com que o dendograma fique impossível de ver as possíveis separações dos grupos gerados.

Após a geração dos clusters segue abaixo as informações das distâncias entre eles:

| | Cluster1 | Cluster2 | Cluster3 | Cluster4 | Cluster5 | Cluster6 | Cluster7 | Cluster8 |
|---|---|---|---|---|---|---|---|---|
| Cluster1 | 0,0000 | 37,8906 | 15,6358 | 25,3865 | 20,4335 | 10,6562 | 5,7828 | 30,9524 |
| Cluster2 | 37,8906 | 0,0000 | 22,2843 | 12,5134 | 17,4989 | 27,2579 | 32,1300 | 6,9571 |
| Cluster3 | 15,6358 | 22,2843 | 0,0000 | 9,7717 | 4,7982 | 4,9811 | 9,8584 | 15,3334 |
| Cluster4 | 25,3865 | 12,5134 | 9,7717 | 0,0000 | 4,9956 | 14,7460 | 19,6200 | 5,5664 |
| Cluster5 | 20,4335 | 17,4989 | 4,7982 | 4,9956 | 0,0000 | 9,7793 | 14,6565 | 10,5442 |
| Cluster6 | 10,6562 | 27,2579 | 4,9811 | 14,7460 | 9,7793 | 0,0000 | 4,8776 | 20,3105 |
| Cluster7 | 5,7828 | 32,1300 | 9,8584 | 19,6200 | 14,6565 | 4,8776 | 0,0000 | 25,1853 |
| Cluster8 | 30,9524 | 6,9571 | 15,3334 | 5,5664 | 10,5442 | 20,3105 | 25,1853 | 0,0000 |

Dessa forma conseguimos caracterizar o grupo que tem o melhor desempenho e também o que tem o pior desempenho, segue abaixo as características:

O maior número de alunos dentro do grupo de pior desempenho tem as seguintes características:

- Mãe tem ensino fundamental;
- A renda familiar é entre 1 e 2 salários mínimos;
- O aluno trabalhou entre 1 e 2 anos durante o ensino médio;
- Sexo feminino;
- Fez o ensino médio somente em escola pública;
- Não fez cursinho pré-vestibular;
- O grupo tem uma média de 15.33 pontos na prova e uma nota máxima de 18 pontos.

O maior número de alunos dentro do grupo de melhor desempenho tem as seguintes características:

- A mãe tem ensino superior completo;
- A renda familiar é entre 10 e 30 salários mínimos;
- O aluno não trabalhou durante o ensino médio;
- Sexo masculino;
- Fez o ensino médio somente em escola particular;
- Fez cursinho pré – vestibular;
- O grupo tem uma média de 53 pontos na prova e uma nota máxima de 62 pontos;

Podemos notar a partir das distâncias entre os clusters que o grupo de melhor desempenho (Cluster 2) está bem mais distante dos outros grupos, se diferenciando realmente.

# 6. CONCLUSÃO

Após todas as análises conseguimos extrair características importantes dos alunos que se destacam na prova do Enem.

✓ Análise Descritiva

- Alunos de escola privada sempre têm um desempenho melhor do que alunos de escola pública, acusando o quanto o ensino de escola pública na época estava bem longe de se igualar ao da escola privada;
- Temos uma diferença de em média 12 pontos (a mais para escola privada) entre as duas instituições;
- A partir da construção dos box plot´s também retiramos conclusões importantes de como as variáveis que escolhemos influenciam na nota da prova. Em ambas as instituições podemos afirmar que:

**Quanto maior o grau de instrução da mãe, maior será a nota;**

Os alunos que tem mãe com pós graduação tem um desempenho com em torno de 15 pontos a mais do que alunos que tem mãe com ensino fundamental. A maioria dos alunos de colégios públicos têm mães com o ensino fundamental completo (30%), enquanto que nos colégios particulares a maioria dos alunos tem mães com o ensino superior completo (42%)

**Quanto maior a renda, maior será a nota;**

Alunos que tem renda de até 50 salários mínimos também tem uma diferença de até 15 pontos a mais do que alunos que tem renda de até 1 salário mínimo.A maioria dos alunos que estudaram em colégios públicos (45%), tem uma renda até 5 salários mínimos, enquanto que alunos de colégios privados têm uma renda de até 30 salários mínimos.

**Os homens têm melhor desempenho;**

No gráfico podemos notar claramente que os homens tem uma diferença na nota de até 5 pontos a mais do que as mulheres, essa diferença se acentua mais nos colégios privados do que nos públicos. O público feminino é maior que o masculino, em 2004 tivemos 64% de mulheres do público total.

**Os alunos de escola particular têm melhor desempenho na prova;**

Assim como já foi falado acima, é nítida a diferença de nota entre alunos que estudaram em colégio público e os que estudaram em colégios privados, no gráfico retirado do box plot não é diferente, temos uma diferença de até 12 pontos. Também é bem nítido o desempenho intermediário (entre o desempenho de alunos de colégios públicos e o desempenho de alunos de colégios privados) dos alunos que fizeram a maior parte do ensino médio em colégios particulares. Nos 3 anos analisados nesse trabalho temos em torno de 75% de alunos de escola pública realizando a prova em cada ano.

**Curso Pré – Vestibular ajuda no desempenho;**

Em ambas as instituições o curso pré – vestibular ajuda no desempenho do aluno na prova, podemos ver que quem fez o curso tem um desempenho de até 5 pontos a mais do que alunos que não o fizeram.Entretanto o percentual de alunos que fazem o curso é bem pequeno, no máximo 20% na instituição privada e no máximo 13% na instituição pública.

- ✓ Análise Inferencial

Através da análise inferencial conseguimos provar que as variáveis escolhidas para análise não são independentes da nota do aluno, isto é, realmente influenciam no desempenho.Conseguimos através de um método específico de cluster caracterizar os grupos que realmente têm um desempenho diferenciado no ENEM.

O grupo com as maiores notas são:

- ✓ Alunos que a mãe tem ensino superior completo;
- ✓ Alunos que têm renda entre 10 e 30 salários mínimos;
- ✓ Alunos que não trabalham;
- ✓ Alunos homens;
- ✓ Alunos que estudam todo o ensino médio em colégio particular;
- ✓ Alunos que fazem curso pré-vestibular;

# Apêndice

# Tabelas

## ✓ Box Plot

Após a execução dos Box Plot´s obtemos as seguintes tabelas:
Q1 = quartil inferior/ Q3 = quartil superior/N = número de observações

### GRAU DE INSTRUÇÃO DA MÃE

A.1

| Escola Pública | Notas 2003 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Grau de Instrução da Mãe | Q1 | Média | Q3 | N | % |
| A - Não Estudou | 19 | 24 | 30 | 1985 | 4,38 |
| B - Até a 4ª série do Fundamental | 21 | 26 | 33 | 13254 | 29,27 |
| C - Até a 8ª série do Fundamental | 21 | 27 | 34 | 11658 | 25,75 |
| D - Ensino Médio Incompleto | 23 | 29 | 35 | 4143 | 9,15 |
| E - Ensino Médio Completo | 23 | 30 | 37 | 9295 | 20,53 |
| F - Ensino Superior Incompleto | 26 | 34 | 41 | 1758 | 3,88 |
| G - Ensino Superior Completo | 27 | 35 | 43 | 2850 | 6,29 |
| H - Pós - Graduação | 27 | 35 | 45 | 332 | 0,73 |

A.2

| Escola Privada | Notas 2003 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Grau de Instrução da Mãe | Q1 | Média | Q3 | N | % |
| A - Não Estudou | 21 | 29 | 40 | 40 | 0,20 |
| B - Até a 4ª série do Fundamental | 27 | 35 | 42 | 767 | 3,76 |
| C - Até a 8ª série do Fundamental | 30 | 36 | 43 | 1346 | 6,61 |
| D - Ensino Médio Incompleto | 31 | 38 | 45 | 997 | 4,89 |
| E - Ensino Médio Completo | 32 | 39 | 46 | 4579 | 22,48 |
| F - Ensino Superior Incompleto | 35 | 41 | 48 | 2110 | 10,36 |
| G - Ensino Superior Completo | 37 | 44 | 50 | 8598 | 42,20 |
| H - Pós - Graduação | 38 | 45 | 51 | 1935 | 9,50 |

A.3

| Escola Pública | Notas 2004 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Grau de Instrução da Mãe | Q1 | Média | Q3 | N | % |
| A - Não Estudou | 17 | 21 | 26 | 1507 | 3,90 |
| B - Até a 4ª série do Fundamental | 19 | 23 | 28 | 10464 | 27,05 |
| C - Até a 8ª série do Fundamental | 19 | 24 | 29 | 9810 | 25,36 |
| D - Ensino Médio Incompleto | 20 | 25 | 31 | 3659 | 9,46 |
| E - Ensino Médio Completo | 21 | 27 | 33 | 8850 | 22,88 |
| F - Ensino Superior Incompleto | 23 | 29 | 37 | 1527 | 3,95 |
| G - Ensino Superior Completo | 23 | 30 | 40 | 2516 | 6,51 |
| H - Pós - Graduação | 24 | 32 | 43 | 344 | 0,89 |

A.4

| Escola Privada | Notas 2004 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Grau de Instrução da Mãe | Q1 | Média | Q3 | N | % |
| A - Não Estudou | 18 | 29 | 40 | 28 | 0,16 |
| B - Até a 4ª série do Fundamental | 25 | 32 | 39 | 537 | 3,03 |
| C - Até a 8ª série do Fundamental | 25 | 33 | 40 | 1094 | 6,17 |
| D - Ensino Médio Incompleto | 27 | 34 | 41 | 800 | 4,51 |
| E - Ensino Médio Completo | 29 | 36 | 43 | 3971 | 22,40 |
| F - Ensino Superior Incompleto | 31 | 38 | 45 | 1811 | 10,22 |
| G - Ensino Superior Completo | 33 | 41 | 49 | 7573 | 42,72 |
| H - Pós - Graduação | 34 | 43 | 50 | 1912 | 10,79 |

A.5

| Escola Pública | Notas 2005 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Grau de Instrução da Mãe | Q1 | Média | Q3 | N | % |
| A - Não Estudou | 15 | 19 | 23 | 1618 | 3,96 |
| B - Até a 4ª série do Fundamental | 16 | 20 | 25 | 10908 | 26,67 |
| C - Até a 8ª série do Fundamental | 17 | 21 | 26 | 10686 | 26,12 |
| D - Ensino Médio Incompleto | 18 | 22 | 28 | 3814 | 9,32 |
| E - Ensino Médio Completo | 19 | 23 | 30 | 9897 | 24,20 |
| F - Ensino Superior Incompleto | 20 | 25 | 32 | 1446 | 3,54 |
| G - Ensino Superior Completo | 21 | 27 | 35 | 2304 | 5,63 |
| H - Pós - Graduação | 20 | 26 | 34 | 231 | 0,56 |

A.6

| Escola Privada | Notas 2005 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Grau de Instrução da Mãe | Q1 | Média | Q3 | N | % |
| A - Não Estudou | 17 | 23 | 29 | 38 | 0,29 |
| B - Até a 4ª série do Fundamental | 22 | 28 | 35 | 427 | 3,27 |
| C - Até a 8ª série do Fundamental | 24 | 30 | 38 | 824 | 6,32 |
| D - Ensino Médio Incompleto | 26 | 33 | 39 | 630 | 4,83 |
| E - Ensino Médio Completo | 26 | 33 | 40 | 3113 | 23,86 |
| F - Ensino Superior Incompleto | 29 | 36 | 43 | 1361 | 10,43 |
| G - Ensino Superior Completo | 31 | 39 | 46 | 5557 | 42,60 |
| H - Pós - Graduação | 33 | 40 | 47 | 1096 | 8,40 |

## RENDA FAMILIAR

B.1

| Escola Pública | Notas 2003 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Renda Familiar | Q1 | Média | Q3 | N | % |
| A - Até 1 salário mínimo | 19 | 23 | 29 | 1981 | 4,34 |
| B - De 1 a 2 salários mínimos | 20 | 25 | 31 | 9657 | 21,17 |
| C - De 2 a 5 salários mínimos | 22 | 28 | 35 | 21048 | 46,14 |
| D - De 5 a 10 salários mínimos | 24 | 31 | 38 | 9601 | 21,05 |
| E - De 10 a 30 salários mínimos | 26 | 33 | 42 | 3098 | 6,79 |
| F - De 30 a 50 salários mínimos | 26 | 34 | 44 | 165 | 0,36 |
| G - Mais de 50 salários mínimos | 22 | 32 | 46 | 67 | 0,15 |

B.2

| Escola Privada | Notas 2003 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Renda Familiar | Q1 | Média | Q3 | N | % |
| A - Até 1 salário mínimo | 35 | 42 | 50 | 80 | 0,40 |
| B - De 1 a 2 salários mínimos | 25 | 32 | 40 | 317 | 1,57 |
| C - De 2 a 5 salários mínimos | 30 | 36 | 43 | 2326 | 11,54 |
| D - De 5 a 10 salários mínimos | 32 | 39 | 46 | 4784 | 23,74 |
| E - De 10 a 30 salários mínimos | 35 | 46 | 49 | 8430 | 41,84 |
| F - De 30 a 50 salários mínimos | 37 | 45 | 51 | 2464 | 12,23 |
| G - Mais de 50 salários mínimos | 40 | 47 | 52 | 1748 | 8,68 |

B.3

| Escola Pública | Notas 2004 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Renda Familiar | Q1 | Média | Q3 | N | % |
| A - Até 1 salário mínimo | 17 | 21 | 25 | 1925 | 4,92 |
| B - De 1 a 2 salários mínimos | 18 | 22 | 27 | 10009 | 25,57 |
| C - De 2 a 5 salários mínimos | 20 | 25 | 31 | 16936 | 43,26 |
| D - De 5 a 10 salários mínimos | 22 | 27 | 34 | 7421 | 18,96 |
| E - De 10 a 30 salários mínimos | 23 | 30 | 40 | 2611 | 6,67 |
| F - De 30 a 50 salários mínimos | 24 | 33 | 45 | 181 | 0,46 |
| G - Mais de 50 salários mínimos | 23 | 37 | 48 | 67 | 0,17 |

B.4

| Escola Privada | Notas 2004 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Renda Familiar | Q1 | Média | Q3 | N | % |
| A - Até 1 salário mínimo | 22 | 32 | 42 | 67 | 0,38 |
| B - De 1 a 2 salários mínimos | 23 | 30 | 38 | 349 | 1,99 |
| C - De 2 a 5 salários mínimos | 26 | 33 | 40 | 1946 | 11,10 |
| D - De 5 a 10 salários mínimos | 29 | 36 | 43 | 4095 | 23,36 |
| E - De 10 a 30 salários mínimos | 32 | 39 | 47 | 7227 | 41,22 |
| F - De 30 a 50 salários mínimos | 34 | 42 | 50 | 2294 | 13,08 |
| G - Mais de 50 salários mínimos | 37 | 45 | 52 | 1554 | 8,86 |

B.5

| Escola Pública | Notas 2005 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Renda Familiar | Q1 | Média | Q3 | N | % |
| A - Até 1 salário mínimo | 16 | 19 | 23 | 3110 | 7,54 |
| B - De 1 a 2 salários mínimos | 17 | 20 | 25 | 12897 | 31,28 |
| C - De 2 a 5 salários mínimos | 18 | 22 | 28 | 18743 | 45,46 |
| D - De 5 a 10 salários mínimos | 19 | 24 | 31 | 5333 | 12,94 |
| E - De 10 a 30 salários mínimos | 20 | 26 | 35 | 1027 | 2,49 |
| F - De 30 a 50 salários mínimos | 18 | 27 | 37 | 80 | 0,19 |
| G - Mais de 50 salários mínimos | 16 | 22 | 32 | 37 | 0,09 |

B.6

| Escola Privada | Notas 2005 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Renda Familiar | Q1 | Média | Q3 | N | % |
| A - Até 1 salário mínimo | 21 | 26 | 35 | 72 | 0,56 |
| B - De 1 a 2 salários mínimos | 21 | 28 | 34 | 375 | 2,89 |
| C - De 2 a 5 salários mínimos | 25 | 31 | 38 | 1989 | 15,4 |
| D - De 5 a 10 salários mínimos | 28 | 34 | 41 | 3985 | 30,8 |
| E - De 10 a 30 salários mínimos | 31 | 38 | 45 | 4709 | 36,4 |
| F - De 30 a 50 salários mínimos | 33 | 41 | 47 | 1181 | 9,12 |
| G - Mais de 50 salários mínimos | 36 | 43 | 49 | 643 | 4,96 |

## ATIVIDADE REMUNERADA

C.1

| Escola Pública | Notas 2003 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Atividade Remunerada | Q1 | Média | Q3 | N | % |
| A - Sim, todo o tempo | 20 | 25 | 32 | 5463 | 18,78 |
| B - Sim, menos de 1 ano | 22 | 28 | 34 | 10628 | 36,54 |
| C - Sim, de 1 a 2 anos | 23 | 29 | 35 | 4736 | 16,28 |
| D - Sim, de 2 a 3 anos | 21 | 27 | 35 | 1207 | 4,15 |
| E - Não | 20 | 26 | 33 | 7052 | 24,25 |

C.2

| Escola Privada | Notas 2003 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Atividade Remunerada | Q1 | Média | Q3 | N | % |
| A - Sim, todo o tempo | 28 | 35 | 42 | 666 | 9,85 |
| B - Sim, menos de 1 ano | 31 | 39 | 45 | 2326 | 34,41 |
| C - Sim, de 1 a 2 anos | 31 | 39 | 45 | 971 | 14,36 |
| D - Sim, de 2 a 3 anos | 30 | 37 | 44 | 205 | 3,03 |
| E - Não | 31 | 39 | 46 | 2592 | 38,34 |

C.3

| Escola Pública | Notas 2004 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Atividade Remunerada | Q1 | Média | Q3 | N | % |
| A - Sim, todo o tempo | 18 | 22 | 28 | 4321 | 17,42 |
| B - Sim, menos de 1 ano | 20 | 24 | 30 | 8875 | 35,78 |
| C - Sim, de 1 a 2 anos | 21 | 26 | 31 | 3705 | 14,94 |
| D - Sim, de 2 a 3 anos | 19 | 24 | 29 | 757 | 3,05 |
| E - Não | 19 | 23 | 29 | 7147 | 28,81 |

C.4

| Escola Privada | Notas 2004 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Atividade Remunerada | Q1 | Média | Q3 | N | % |
| A - Sim, todo o tempo | 24 | 32 | 38 | 528 | 9,24 |
| B - Sim, menos de 1 ano | 28 | 35 | 43 | 1947 | 34,07 |
| C - Sim, de 1 a 2 anos | 29 | 35 | 43 | 736 | 12,88 |
| D - Sim, de 2 a 3 anos | 26 | 33 | 40 | 128 | 2,24 |
| E - Não | 28 | 36 | 43 | 2376 | 41,57 |

C.5

| Escola Pública | Notas 2005 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Atividade Remunerada | Q1 | Média | Q3 | N | % |
| A - Sim, todo o tempo | 16 | 19 | 24 | 4461 | 16,07 |
| B - Sim, menos de 1 ano | 17 | 22 | 27 | 9795 | 35,29 |
| C - Sim, de 1 a 2 anos | 18 | 22 | 28 | 4470 | 16,10 |
| D - Sim, de 2 a 3 anos | 17 | 21 | 26 | 851 | 3,07 |
| E - Não | 16 | 20 | 25 | 8180 | 29,47 |

C.6

| Escola Privada | Notas 2005 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Atividade Remunerada | Q1 | Média | Q3 | N | % |
| A - Sim, todo o tempo | 23 | 29 | 36 | 371 | 8,23 |
| B - Sim, menos de 1 ano | 26 | 33 | 40 | 1501 | 33,30 |
| C - Sim, de 1 a 2 anos | 26 | 33 | 40 | 669 | 14,84 |
| D - Sim, de 2 a 3 anos | 22 | 29 | 38 | 100 | 2,22 |
| E - Não | 26 | 32 | 40 | 1866 | 41,40 |

## SEXO

D.1

| Escola Pública | Notas 2003 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Sexo | Q1 | Média | Q3 | N | % |
| 1 - Masculino | 22 | 29 | 37 | 18784 | 40,13 |
| 2 - Feminino | 21 | 27 | 34 | 28024 | 59,87 |

D.2

| Escola Privada | Notas 2003 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Sexo | Q1 | Média | Q3 | N | % |
| 1 - Masculino | 36 | 43 | 50 | 9484 | 45,78 |
| 2 - Feminino | 33 | 40 | 47 | 11231 | 54,22 |

D.3

| Escola Pública | Notas 2004 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Sexo | Q1 | Média | Q3 | N | % |
| 1 - Masculino | 20 | 26 | 36 | 15026 | 38,22 |
| 2 - Feminino | 19 | 24 | 30 | 24885 | 63,30 |

D.4

| Escola Privada | Notas 2004 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Sexo | Q1 | Média | Q3 | N | % |
| 1 - Masculino | 33 | 41 | 49 | 8011 | 44,28 |
| 2 - Feminino | 29 | 36 | 44 | 10082 | 55,72 |

D.5

| Escola Pública | Notas 2005 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Sexo | Q1 | Média | Q3 | N | % |
| 1 - Masculino | 18 | 23 | 29 | 15761 | 37,52 |
| 2 - Feminino | 17 | 21 | 26 | 26246 | 62,48 |

D.6

| Escola Privada | Notas 2005 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Sexo | Q1 | Média | Q3 | N | % |
| 1 - Masculino | 30 | 38 | 45 | 5717 | 43,19 |
| 2 - Feminino | 28 | 34 | 42 | 7521 | 56,81 |

**TIPO DE ESCOLA NO ENSINO MÉDIO**

E.1

| Ensino Médio | Notas 2003 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| | Q1 | Média | Q3 | N | % |
| A - Somente em escola pública | 22 | 28 | 35 | 45428 | 67,28 |
| B - Maior parte em escola pública | 22 | 30 | 38 | 1380 | 2,04 |
| C - Somente em escola particular | 34 | 42 | 49 | 19436 | 28,78 |
| D - Maior parte em escola particular | 29 | 36 | 43 | 1279 | 1,89 |

E.2

| Ensino Médio | Notas 2004 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| | Q1 | Média | Q3 | N | % |
| A - Somente em escola pública | 20 | 25 | 31 | 38659 | 66,65 |
| B - Maior parte em escola pública | 20 | 26 | 34 | 1252 | 2,16 |
| C - Somente em escola particular | 31 | 39 | 47 | 16989 | 29,29 |
| D - Maior parte em escola particular | 26 | 33 | 41 | 1104 | 1,90 |

E.3

| Ensino Médio | Notas 2004 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| | Q1 | Média | Q3 | N | % |
| A - Somente em escola pública | 17 | 22 | 27 | 40894 | 74,00 |
| B - Maior parte em escola pública | 18 | 23 | 30 | 1130 | 2,04 |
| C - Somente em escola particular | 29 | 36 | 44 | 12457 | 22,54 |
| D - Maior parte em escola particular | 24 | 31 | 37 | 781 | 1,41 |

**CURSO PRÉ – VESTIBULAR**

F.1

| Escola Pública | Notas 2003 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Curso Pré-Vestibular | Q1 | Média | Q3 | N | % |
| A - Sim | 25 | 32 | 40 | 5395 | 11,71 |
| B - Não | 22 | 27 | 34 | 40673 | 88,29 |

F.2

| Escola Privada | Notas 2003 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Curso Pré-Vestibular | Q1 | Média | Q3 | N | % |
| A - Sim | 38 | 45 | 52 | 4361 | 21,13 |
| B - Não | 33 | 41 | 47 | 16274 | 78,87 |

F.3

| Escola Pública | Notas 2004 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Curso Pré-Vestibular | Q1 | Média | Q3 | N | % |
| A - Sim | 22 | 29 | 36 | 5112 | 12,95 |
| B - Não | 19 | 24 | 30 | 34377 | 87,05 |

F.4

| Escola Privada | Notas 2004 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Curso Pré-Vestibular | Q1 | Média | Q3 | N | % |
| A - Sim | 35 | 43 | 51 | 3502 | 19,48 |
| B - Não | 30 | 37 | 45 | 14477 | 80,52 |

F.5

| Escola Pública | Notas 2005 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Curso Pré-Vestibular | Q1 | Média | Q3 | N | % |
| A - Sim | 19 | 25 | 33 | 3601 | 8,69 |
| B - Não | 17 | 21 | 27 | 37825 | 91,31 |

F.6

| Escola Privada | Notas 2005 | | | | |
|---|---|---|---|---|---|
| | Medidas de Posição | | | | |
| Curso Pré-Vestibular | Q1 | Média | Q3 | N | % |
| A - Sim | 33 | 41 | 48 | 2761 | 20,45 |
| B - Não | 28 | 35 | 42 | 10739 | 79,55 |

## ✓ Independência das Variáveis

As notas foram divididas nas seguintes faixas :

**Faixa 0 = Notas de 0 a 10 ;**
**Faixa 1 = Notas de 11 a 20 ;**
**Faixa 2 = Notas de 21 a 30 ;**
**Faixa 3 = Notas de 31 a 40 ;**
**Faixa 4 = Notas de 41 a 50 ;**
**Faixa 5 = Notas de 51 a 60 ;**
**Faixa 6 = Notas de 61 a 63;**

Para 2003 temos as seguintes estatísticas:

Q1)Qual é o grau de instrução da sua mãe?

| |
|---|
| A - Não Estudou |
| B - Até a 4ª série do Fundamental |
| C - Até a 8ª série do Fundamental |
| D - Ensino Médio Incompleto |
| E - Ensino Médio Completo |
| F - Ensino Superior Incompleto |

| |
|---|
| G - Ensino Superior Completo |
| H - Pós - Graduação |

| PRIVADA | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Faixa de Nota/Q1** | A | B | C | D | E | F | G | H | I | total |
| **0-1** | 9 | 54 | 85 | 38 | 136 | 32 | 95 | 25 | 13 | 487 |
| **2** | 14 | 227 | 291 | 204 | 848 | 285 | 805 | 153 | 48 | 2875 |
| **3** | 8 | 272 | 495 | 356 | 1543 | 645 | 2316 | 453 | 82 | 6170 |
| **4** | 9 | 167 | 378 | 295 | 1501 | 788 | 3293 | 766 | 71 | 7268 |
| **6-5** | 6 | 49 | 99 | 104 | 555 | 361 | 2094 | 540 | 27 | 3835 |
| **total** | 46 | 769 | 1348 | 997 | 4583 | 2111 | 8603 | 1937 | 241 | 20635 |

**Chi-Sq = 1586.444; DF = 32; P-Value = 0.000**

| PÚBLICA | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Faixa de Nota/Q1** | A | B | C | D | E | F | G | H | I | total |
| **0-1** | 628 | 3255 | 2498 | 703 | 1343 | 155 | 245 | 32 | 399 | 9258 |
| **2** | 915 | 5734 | 4985 | 1654 | 3453 | 528 | 749 | 92 | 500 | 18610 |
| **3** | 378 | 3252 | 3143 | 1310 | 2914 | 600 | 921 | 90 | 232 | 12840 |
| **4** | 63 | 899 | 916 | 416 | 1273 | 355 | 673 | 72 | 34 | 4701 |
| **6-5** | 3 | 114 | 114 | 61 | 302 | 120 | 263 | 46 | 10 | 1033 |
| **total** | 1987 | 13254 | 11656 | 4144 | 9285 | 1758 | 2851 | 332 | 1175 | 46442 |

**Chi-Sq = 3912.692; DF = 32; P-Value = 0.000**

Q2)Qual é a renda familiar?

| |
|---|
| A - Até 1 salário mínimo |
| B - De 1 a 2 salários mínimos |
| C - De 2 a 5 salários mínimos |
| D - De 5 a 10 salários mínimos |
| E - De 10 a 30 salários mínimos |
| F - De 30 a 50 salários mínimos |
| G - Mais de 50 salários mínimos |

| PRIVADA | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Faixa de Nota/Q2** | A | B | C | D | E | F | G | H | total |
| **0-1** | 6 | 36 | 104 | 130 | 140 | 24 | 18 | 5 | 463 |
| **2** | 8 | 107 | 548 | 842 | 958 | 215 | 119 | 20 | 2817 |
| **3** | 21 | 95 | 863 | 1651 | 2465 | 592 | 363 | 37 | 6087 |
| **4** | 25 | 63 | 632 | 1594 | 3146 | 954 | 704 | 35 | 7153 |
| **5-6** | 20 | 16 | 179 | 567 | 1721 | 679 | 544 | 25 | 3751 |
| **total** | 80 | 317 | 2326 | 4784 | 8430 | 2464 | 1748 | 122 | 20271 |

**Chi-Sq = 1516.076; DF = 32; P-Value = 0.000**

| PÚBLICA | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Faixa de Nota/Q2** | A | B | C | D | E | F | G | H | total |
| **0-1** | 694 | 2798 | 3822 | 1307 | 336 | 24 | 8 | 174 | 9163 |
| **2** | 856 | 4366 | 8739 | 3361 | 920 | 42 | 23 | 194 | 18501 |
| **3** | 359 | 1992 | 6200 | 3110 | 935 | 43 | 13 | 92 | 12744 |
| **4** | 66 | 461 | 1986 | 1468 | 621 | 32 | 16 | 24 | 4674 |
| **5-6** | 6 | 40 | 301 | 356 | 286 | 24 | 7 | 7 | 1027 |
| **total** | 1981 | 9657 | 21048 | 9602 | 3098 | 165 | 67 | 491 | 46109 |

**Chi-Sq = 3982.597; DF = 32; P-Value = 0.000**

Q3)Por quanto tempo trabalhou no ensino médio?

| |
|---|
| A - Sim, todo o tempo |
| B - Sim, menos de 1 ano |
| C - Sim, de 1 a 2 anos |
| D - Sim, de 2 a 3 anos |
| E - Não |

| PRIVADA | | | | | | |
|---|---|---|---|---|---|---|
| **Faixa de Nota/Q3** | A | B | C | D | E | total |
| **0-1** | 48 | 78 | 38 | 10 | 94 | 268 |
| **2** | 175 | 445 | 190 | 42 | 490 | 1342 |
| **3** | 253 | 795 | 323 | 81 | 878 | 2330 |
| **4** | 151 | 761 | 326 | 57 | 838 | 2133 |
| **5-6** | 39 | 247 | 94 | 15 | 292 | 687 |
| **total** | 666 | 2326 | 971 | 205 | 2592 | 6760 |

**Chi-Sq = 79.285; DF = 16; P-Value = 0.000**

| PÚBLICA | | | | | | |
|---|---|---|---|---|---|---|
| **Faixa de Nota/Q3** | A | B | C | D | E | total |
| **0** | 56 | 73 | 19 | 14 | 69 | 231 |
| **1** | 1400 | 2011 | 765 | 243 | 1816 | 6235 |
| **2** | 2419 | 4364 | 1884 | 482 | 2929 | 12078 |
| **3** | 1239 | 3071 | 1497 | 346 | 1646 | 7799 |
| **4** | 320 | 979 | 501 | 113 | 485 | 2398 |
| **5-6** | 29 | 130 | 70 | 9 | 107 | 345 |
| **total** | 5463 | 10628 | 4736 | 1207 | 7052 | 29086 |

**Chi-Sq = 490.782; DF = 20; P-Value = 0.000**

Q4)Em qual tipo de escola cursou o ensino médio?

| |
|---|
| A - Somente em escola pública |
| B - Maior parte em escola pública |

| |
|---|
| C - Somente em escola particular |
| D - Maior parte em escola particular |

| PRIVADA | | | |
|---|---|---|---|
| **Faixa de Nota/Q4** | C | D | total |
| **0-1** | 403 | 67 | 470 |
| **2** | 2579 | 326 | 2905 |
| **3** | 5751 | 450 | 6201 |
| **4** | 6977 | 327 | 7304 |
| **5-6** | 3726 | 109 | 3835 |
| **Total** | 19436 | 1279 | 20715 |

**Chi-Sq = 303.119; DF = 4; P-Value = 0.000**

| PÚBLICA | | | |
|---|---|---|---|
| **Faixa de Nota/Q4** | A | B | total |
| **0** | 305 | 13 | 318 |
| **1** | 8755 | 260 | 9015 |
| **2** | 18337 | 447 | 18784 |
| **3** | 12517 | 408 | 12925 |
| **4** | 4512 | 215 | 4727 |
| **5-6** | 1002 | 37 | 1039 |
| **total** | 45428 | 1380 | 46808 |

**Chi-Sq = 68.419; DF = 5; P-Value = 0.000**

Q5)Fez curso pré-vestibular?

| |
|---|
| A - Sim |
| B - Não |

| PRIVADA | | | |
|---|---|---|---|
| **Faixa de Nota/Q5** | A | B | total |
| **0** | 6 | 14 | 20 |
| **1** | 64 | 378 | 442 |
| **2** | 354 | 2535 | 2889 |
| **3** | 969 | 5209 | 6178 |
| **4** | 1706 | 5571 | 7277 |
| **5** | 1231 | 2529 | 3760 |
| **6** | 31 | 38 | 69 |
| **total** | 4361 | 16274 | 20635 |

**Chi-Sq = 610.012; DF = 6; P-Value = 0.000**

| PÚBLICA | | | |
|---|---|---|---|
| **Faixa de Nota/Q5** | A | B | total |
| **0** | 22 | 280 | 302 |
| **1** | 663 | 8065 | 8728 |
| **2** | 1716 | 16776 | 18492 |
| **3** | 1713 | 11101 | 12814 |
| **4** | 933 | 3766 | 4699 |
| **5** | 343 | 681 | 1024 |
| **6** | 7 | 9 | 16 |
| **total** | 5397 | 40678 | 46075 |

**Chi-Sq = 1074.075; DF = 6; P-Value = 0.000**

Q6)Sexo do candidato.

| |
|---|
| 1 - Masculino |
| 2 - Feminino |

| PRIVADA | | | |
|---|---|---|---|
| **Faixa de Nota/Q6** | 1 | 2 | total |
| **0** | 9 | 12 | 21 |
| **1** | 195 | 254 | 449 |
| **2** | 1104 | 1801 | 2905 |
| **3** | 2444 | 3757 | 6201 |
| **4** | 3524 | 3780 | 7304 |
| **5** | 2161 | 1605 | 3766 |
| **6** | 47 | 22 | 69 |
| **total** | 9485 | 11233 | 20715 |

**Chi-Sq = 409.121; DF = 6; P-Value = 0.000**

| PÚBLICA | | | |
|---|---|---|---|
| **Faixa de Nota/Q6** | 1 | 2 | total |
| **0** | 131 | 187 | 318 |
| **1** | 3359 | 5656 | 9015 |
| **2** | 6681 | 12103 | 18784 |
| **3** | 5499 | 7426 | 12925 |
| **4** | 2464 | 2263 | 4727 |
| **5** | 645 | 385 | 1030 |
| **6** | 6 | 9 | 15 |
| **total** | 18786 | 28031 | 46814 |

**Chi-Sq = 725.595; DF = 6; P-Value = 0.000**

Para 2004 temos as seguintes estatísticas:

Q1)Qual é o grau de instrução da sua mãe?

| PRIVADA | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Faixa de Nota/Q1** | A | B | C | D | E | F | G | H | I | total |
| **0-1** | 55 | 104 | 68 | 207 | 75 | 198 | 37 | 27 | 6 | 777 |
| **2** | 7 | 185 | 349 | 241 | 1025 | 367 | 1152 | 252 | 76 | 3654 |
| **3** | 6 | 174 | 373 | 286 | 1436 | 668 | 2301 | 530 | 93 | 5867 |
| **4-5-6** | 6 | 123 | 265 | 209 | 1296 | 709 | 3919 | 1094 | 77 | 7698 |
| **total** | 74 | 586 | 1055 | 943 | 3832 | 1942 | 7409 | 1903 | 252 | 17996 |

**Chi-Sq = 3326.769; DF = 24; P-Value = 0.000**

| PÚBLICA | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Faixa de Nota/Q1** | A | B | C | D | E | F | G | H | I | Total |
| **0** | 33 | 100 | 86 | 30 | 42 | 7 | 7 | 5 | 20 | 330 |
| **1** | 644 | 3566 | 3061 | 947 | 1955 | 217 | 381 | 49 | 414 | 11234 |
| **2** | 631 | 4910 | 4562 | 1702 | 3851 | 614 | 887 | 101 | 460 | 17718 |
| **3** | 172 | 1551 | 1680 | 707 | 2071 | 411 | 627 | 93 | 117 | 7429 |
| **5;6;4** | 27 | 337 | 421 | 273 | 931 | 278 | 398 | 59 | 31 | 2755 |
| **total** | 1507 | 10464 | 9810 | 3659 | 8850 | 1527 | 2300 | 307 | 1042 | 39466 |

**Chi-Sq = 2498.220; DF = 32; P-Value = 0.000**

Q2)Qual é a renda familiar?

| PRIVADA | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Faixa de Nota/Q2** | A | B | C | D | E | F | G | H | total |
| **0-1** | 16 | 48 | 157 | 202 | 243 | 53 | 30 | 13 | 762 |
| **2** | 13 | 132 | 623 | 1063 | 1286 | 311 | 143 | 23 | 3594 |
| **3** | 18 | 106 | 702 | 1478 | 2370 | 666 | 384 | 35 | 5759 |
| **4** | 16 | 51 | 362 | 989 | 2199 | 746 | 544 | 26 | 4933 |
| **5-6** | 5 | 12 | 102 | 363 | 1191 | 518 | 453 | 16 | 2660 |
| **total** | 68 | 349 | 1946 | 4095 | 7289 | 2294 | 1554 | 113 | 17708 |

**Chi-Sq = 1479.673; DF = 28; P-Value = 0.000**

| PÚBLICA | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Faixa de Nota/Q2** | A | B | C | D | E | F | G | H | total |
| **0-1** | 923 | 3865 | 4569 | 1535 | 405 | 22 | 11 | 165 | 11495 |
| **2** | 794 | 4660 | 7902 | 3145 | 929 | 58 | 14 | 133 | 17635 |
| **3** | 174 | 1245 | 3421 | 1791 | 664 | 41 | 14 | 34 | 7384 |
| **4-5-6** | 34 | 239 | 1044 | 950 | 613 | 60 | 28 | 19 | 2987 |
| **total** | 1925 | 10009 | 16936 | 7421 | 2611 | 181 | 67 | 351 | 39501 |

**Chi-Sq = 3496.934; DF = 21; P-Value = 0.000**

Q3) Por quanto tempo trabalhou no ensino médio?

| PRIVADA | | | | | | |
|---|---|---|---|---|---|---|
| **Faixa de Nota/Q3** | A | B | C | D | E | total |
| **0-1** | 59 | 121 | 40 | 12 | 165 | 397 |
| **2** | 187 | 522 | 178 | 40 | 601 | 1528 |
| **3** | 178 | 678 | 288 | 45 | 831 | 2020 |
| **4** | 82 | 474 | 165 | 21 | 552 | 1294 |
| **5-6** | 22 | 152 | 65 | 10 | 227 | 476 |
| **total** | 528 | 1947 | 736 | 128 | 2376 | 5715 |

**Chi-Sq = 74.752; DF = 16; P-Value = 0.000**

| PÚBLICA | | | | | | |
|---|---|---|---|---|---|---|
| **Faixa de Nota/Q3** | A | B | C | D | E | total |
| **0** | 69 | 66 | 24 | 7 | 85 | 251 |
| **1** | 1662 | 2586 | 845 | 229 | 2485 | 7807 |
| **2** | 1923 | 4119 | 1804 | 360 | 3198 | 11404 |
| **3** | 548 | 1678 | 798 | 121 | 1025 | 4170 |
| **4** | 104 | 365 | 197 | 30 | 293 | 989 |
| **5-6** | 15 | 61 | 37 | 10 | 61 | 184 |
| **total** | 4321 | 8875 | 3705 | 757 | 7147 | 24805 |

**Chi-Sq = 438.993; DF = 20; P-Value = 0.000**

Q4) Em qual tipo de escola cursou o ensino médio?

| PRIVADA | | | |
|---|---|---|---|
| **Faixa de Nota/Q4** | C | D | total |
| **0-1** | 683 | 98 | 781 |
| **2** | 3337 | 358 | 3695 |
| **3** | 5535 | 358 | 5893 |
| **4** | 4858 | 213 | 5071 |
| **5-6** | 2576 | 77 | 2653 |
| **total** | 16989 | 1104 | 18093 |

**Chi-Sq = 219.025; DF = 4; P-Value = 0.000**

| PÚBLICA | | | |
|---|---|---|---|
| **Faixa de Nota/Q4** | A | B | total |
| **0** | 316 | 14 | 330 |
| **1** | 10994 | 309 | 11303 |
| **2** | 17310 | 496 | 17806 |
| **3** | 7160 | 292 | 7452 |
| **4** | 2242 | 108 | 2350 |
| **6;5** | 637 | 33 | 670 |
| **Total** | 38659 | 1252 | 39911 |

**Chi-Sq = 53.096; DF = 5; P-Value = 0.000**

Q5) Fez curso pré-vestibular?

| PRIVADA | | | |
|---|---|---|---|
| **Faixa de Nota/Q5** | A | B | total |
| **0** | 6 | 14 | 20 |
| **1** | 81 | 672 | 753 |
| **2** | 396 | 3268 | 3664 |
| **3** | 988 | 4870 | 5858 |
| **4** | 1154 | 3889 | 5043 |
| **5** | 842 | 1700 | 2542 |
| **6** | 35 | 64 | 99 |
| **Total** | 3502 | 14477 | 17979 |

**Chi-Sq = 593.992; DF = 6; P-Value = 0.000**

| PÚBLICA | | | |
|---|---|---|---|
| **Faixa de Nota/Q5** | A | B | total |
| **0** | 34 | 289 | 323 |
| **1** | 1007 | 10100 | 11107 |
| **2** | 1905 | 15736 | 17641 |
| **3** | 1311 | 6102 | 7413 |
| **4** | 574 | 1763 | 2337 |
| **5** | 270 | 380 | 650 |
| **6** | 11 | 7 | 18 |
| **total** | 5112 | 34377 | 39489 |

**Chi-Sq = 1158.304; DF = 6; P-Value = 0.000**

Q6)Sexo do candidato.

| PRIVADA | | | |
|---|---|---|---|
| **Faixa de Nota/Q6** | 1 | 2 | total |
| **0** | 11 | 9 | 20 |
| **1** | 292 | 469 | 761 |
| **2** | 1186 | 2509 | 3695 |
| **3** | 2403 | 3490 | 5893 |
| **4** | 2471 | 2600 | 5071 |
| **5** | 1571 | 981 | 2552 |
| **6** | 77 | 24 | 101 |
| **total** | 8012 | 10084 | 18093 |

**Chi-Sq = 654.592; DF = 6; P-Value = 0.000**

| PÚBLICA | | | |
|---|---|---|---|
| **Faixa de Nota/Q6** | 1 | 2 | total |
| **0** | 128 | 202 | 330 |
| **1** | 3695 | 7608 | 11303 |
| **2** | 6134 | 11672 | 17806 |
| **3** | 3195 | 4257 | 7452 |
| **4** | 1382 | 968 | 2350 |
| **5-6** | 492 | 178 | 670 |
| **total** | 15026 | 24885 | 39911 |

**Chi-Sq = 1100.700; DF = 6; P-Value = 0.000**

Para 2005 temos as seguintes estatísticas:

Q1)Qual é o grau de instrução da sua mãe?

| PRIVADA | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Faixa de Nota/Q1** | A | B | C | D | E | F | G | H | I | total |
| **0-1** | 13 | 77 | 122 | 63 | 266 | 65 | 209 | 40 | 25 | 880 |
| **2** | 17 | 172 | 292 | 205 | 1003 | 323 | 1030 | 177 | 51 | 3270 |
| **3** | 5 | 126 | 273 | 236 | 1150 | 520 | 1930 | 354 | 36 | 4630 |
| **4-5-6** | 3 | 52 | 137 | 126 | 684 | 453 | 2388 | 525 | 35 | 4403 |
| **total** | 38 | 427 | 824 | 630 | 3103 | 1361 | 5557 | 1096 | 147 | 13183 |

**Chi-Sq = 1188.463; DF = 32; P-Value = 0.000**

| PÚBLICA | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Faixa de Nota/Q1** | A | B | C | D | E | F | G | H | I | total |
| **0** | 49 | 202 | 143 | 33 | 98 | 10 | 23 | 2 | 21 | 581 |
| **1** | 976 | 5469 | 4920 | 1452 | 3393 | 388 | 530 | 61 | 556 | 17745 |
| **2** | 509 | 4159 | 4274 | 1677 | 4197 | 617 | 875 | 88 | 293 | 16689 |
| **3** | 74 | 921 | 1146 | 539 | 1745 | 320 | 574 | 49 | 60 | 5428 |
| **4-5-6** | 10 | 157 | 203 | 113 | 464 | 111 | 302 | 31 | 12 | 1403 |
| **total** | 1618 | 10908 | 10686 | 3814 | 9897 | 1446 | 2304 | 231 | 942 | 41846 |

**Chi-Sq = 2970.955; DF = 32; P-Value = 0.000**

Q2) Qual é a renda familiar?

| PRIVADA | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Faixa de Nota/Q2** | A | B | C | D | E | F | G | H | total |
| **0-1** | 16 | 73 | 237 | 282 | 188 | 32 | 9 | 16 | 853 |
| **2** | 28 | 153 | 677 | 1143 | 965 | 169 | 75 | 20 | 3230 |
| **3** | 19 | 104 | 727 | 1475 | 1678 | 379 | 174 | 23 | 4579 |
| **4** | 6 | 38 | 288 | 891 | 1396 | 425 | 256 | 9 | 3309 |
| **5-6** | 2 | 2 | 55 | 189 | 478 | 174 | 125 | 6 | 1031 |
| **total** | 71 | 370 | 1984 | 3980 | 4705 | 1179 | 639 | 74 | 13002 |

**Chi-Sq = 1182.614; DF = 24; P-Value = 0.000**

| PÚBLICA | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Faixa de Nota/Q2** | A | B | C | D | E | F | G | H | total |
| **0-1** | 1887 | 6654 | 7332 | 1717 | 275 | 28 | 16 | 250 | 18159 |
| **2** | 1011 | 4982 | 0 | 2149 | 377 | 20 | 9 | 92 | 8640 |
| **3** | 174 | 1091 | 2821 | 1024 | 248 | 16 | 6 | 18 | 5398 |
| **4-5-6** | 38 | 170 | 643 | 391 | 127 | 16 | 6 | 4 | 1395 |
| **total** | 3110 | 12897 | 10796 | 5281 | 1027 | 80 | 37 | 364 | 33592 |

**Chi-Sq = 8080.542; DF = 28; P-Value = 0.000**

Q3) Por quanto tempo trabalhou no ensino médio?

| PRIVADA | | | | | | |
|---|---|---|---|---|---|---|
| **Faixa de Nota/Q3** | A | B | C | D | E | total |
| **0-1** | 64 | 137 | 64 | 21 | 201 | 487 |
| **2** | 142 | 452 | 207 | 35 | 594 | 1430 |
| **3** | 111 | 555 | 250 | 31 | 639 | 1586 |
| **4-5-6** | 55 | 357 | 148 | 13 | 432 | 1005 |
| **total** | 372 | 1501 | 669 | 100 | 1866 | 4508 |

**Chi-Sq = 73.302; DF = 16; P-Value = 0.000**

| PÚBLICA | | | | | | |
|---|---|---|---|---|---|---|
| **Faixa de Nota/Q3** | A | B | C | D | E | total |
| **0** | 101 | 117 | 45 | 19 | 159 | 441 |
| **1** | 2417 | 4147 | 1664 | 388 | 4144 | 12760 |
| **2** | 1594 | 4139 | 1978 | 330 | 2977 | 11018 |
| **3** | 306 | 1194 | 666 | 100 | 743 | 3009 |
| **4,5,6** | 43 | 198 | 117 | 14 | 157 | 529 |
| **total** | 4461 | 9795 | 4470 | 851 | 8180 | 27757 |

**Chi-Sq = 534.720; DF = 16; P-Value = 0.000**

Q4) Em qual tipo de escola cursou o ensino médio?

| PRIVADA | | | |
|---|---|---|---|
| **Faixa de Nota/Q4** | C | D | total |
| **0** | 12 | 6 | 18 |
| **1** | 751 | 117 | 868 |
| **2** | 3024 | 262 | 3286 |
| **3** | 4381 | 263 | 4644 |
| **4** | 3255 | 112 | 3367 |
| **5-6** | 1034 | 21 | 1055 |
| **total** | 12457 | 781 | 13238 |

**Chi-Sq = 209.343; DF = 5; P-Value = 0.000**

| PÚBLICA | | | |
|---|---|---|---|
| **Faixa de Nota/Q4** | A | B | total |
| **0** | 566 | 16 | 582 |
| **1** | 17434 | 408 | 17842 |
| **2** | 16318 | 417 | 16735 |
| **3** | 5244 | 197 | 5441 |
| **4** | 1218 | 66 | 1284 |
| **5,6** | 114 | 9 | 123 |
| **Total** | 40894 | 1113 | 42007 |

**Chi-Sq = 71.904; DF = 5; P-Value = 0.000**

Q5)Fez curso pré-vestibular?

| PRIVADA | | | |
|---|---|---|---|
| **Faixa de Nota/Q5** | A | B | total |
| **0-1** | 101 | 771 | 872 |
| **2** | 417 | 2833 | 3250 |
| **3** | 839 | 3772 | 4611 |
| **4** | 1010 | 2342 | 3352 |
| **5-6** | 394 | 661 | 1055 |
| **Total** | 2761 | 10379 | 13140 |

**Chi-Sq = 537.364; DF = 4; P-Value = 0.000**

| PÚBLICA | | | |
|---|---|---|---|
| **Faixa de Nota/Q5** | A | B | total |
| **0** | 46 | 518 | 564 |
| **1** | 1121 | 16382 | 17503 |
| **2** | 1279 | 15283 | 16562 |
| **3** | 798 | 4596 | 5394 |
| **4** | 302 | 978 | 1280 |
| **5-6** | 55 | 68 | 123 |
| **Total** | 3601 | 37825 | 41426 |

**Chi-Sq = 947.486; DF = 5; P-Value = 0.000**

Q6)Sexo do candidato.

| PRIVADA | | | |
|---|---|---|---|
| **Faixa de Nota/Q6** | 1 | 2 | total |
| **0-1** | 315 | 571 | 886 |
| **2** | 1179 | 2107 | 3286 |
| **3** | 1936 | 2708 | 4644 |
| **4** | 1632 | 1735 | 3367 |
| **6-5** | 655 | 400 | 1055 |
| **total** | 5717 | 7521 | 13238 |

**Chi-Sq = 288.687; DF = 4; P-Value = 0.000**

| PÚBLICA | | | |
|---|---|---|---|
| **Faixa de Nota/Q6** | 1 | 2 | total |
| **0** | 207 | 375 | 582 |
| **1** | 5915 | 11927 | 17842 |
| **2** | 6184 | 10551 | 16735 |
| **3** | 2615 | 2826 | 5441 |
| **4** | 753 | 531 | 1284 |
| **5,6** | 87 | 36 | 123 |
| **Total** | 15762 | 26248 | 42007 |

**Chi-Sq = 708.646; DF = 5; P-Value = 0.000**

## Referências Bibliográficas


BUSSAB, Wilton de Oliveira;MORETTIN, Pedro A. Estatística Básica, 5ª Edição 2003.Editora Saraiva.

MAGALHÃES, Marcos Nascimento;LIMA, Antônio Carlos Pedroso. Noções de Probabilidade e Estatística. 6ª Edição 2004. Editora EDUSP

ZAIANE, Osmar R. et al. *On data clustering analysis*: scalability, constraints and validation.
Edmonton Alberta, University of Alberta, 2003.

ANDERBERG, Michael R. *Cluster analysis for applications*. New York: Academic Press, 1973.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. *ACM Computing Surveys*, New York, v. 31, n. 3, p. 265-323, Sept., 1999.

JOHNSON, Richard. A.; WICHERN, Dean W. *Applied multivariate statistical analysis.* 4th ed. New Jersey: Prentice Hall, 1992.

ROMESBURG, Charles H. *Cluster analysis for researchers.* Belmont: Lifetime Learning Publications, 1984.

WITTEN, Ian H.; FRANK, Eibe. *Data mining*: practical machine learning tools and techniques with Java implementations. San Francisco: Morgan Kaufmann, 1999.

BUSSAB, Wilton de Oliveira; MIAZAKI, Édina Shizue; ANDRADE, Dalton Francisco de. *Introdução à análise de agrupamentos*. São Paulo: Associação Brasileira de Estatística, 1990.


## Bibliografia Complementar


PEIXOTO, Cláudia Monteiro.Testes Qui-Quadrado:Aderência e Independência – 2006.Site : http://www.ime.usp.br/~mae116/aula/2006/c-2006-aula-12-testeQuiQuad.pdf

HALKIDI, Maria; BATISTAKIS, Yannis; VAZIRGIANNIS, Michalis. On clustering validation techniques. *Journal of Intelligent Information Systems*, v. 17, n. 2-3, p. 107-145,Dec. 2001.