UNIVERSIDADE FEDERAL DE MATO GROSSO DO SUL

INSTITUTO DE MATEMÁTICA

PROGRAMA DE PÓS GRADUAÇÃO

MATEMÁTICA EM REDE NACIONAL

MESTRADO PROFISSIONAL

ELIEL GONÇALVES VILLA NOVA

ANÁLISE COMPARATIVA ENTRE AVALIAÇÃO DIAGNÓSTICA E DESEMPENHO ESCOLAR NO COLÉGIO MILITAR DE CAMPO GRANDE - VIA INTERVALOS DE CONFIANÇA *BOOTSTRAP*

CAMPO GRANDE - MS

MAIO DE 2014

UNIVERSIDADE FEDERAL DE MATO GROSSO DO SUL

INSTITUTO DE MATEMÁTICA

PROGRAMA DE PÓS GRADUAÇÃO

MATEMÁTICA EM REDE NACIONAL

MESTRADO PROFISSIONAL

ELIEL GONÇALVES VILLA NOVA

ANÁLISE COMPARATIVA ENTRE AVALIAÇÃO DIAGNÓSTICA E DESEMPENHO ESCOLAR NO COLÉGIO MILITAR DE CAMPO GRANDE - VIA INTERVALOS DE CONFIANÇA *BOOTSTRAP*

Orientador: Prof. Dr. Jair da Silva

Dissertação apresentada ao Programa de Pós-Graduação em Matemática em Rede Nacional do Instituto de Matemática – INMA/UFMS, como parte dos requisitos para obtenção do Título de Mestre.

CAMPO GRANDE - MS

MAIO DE 2014

ANÁLISE COMPARATIVA ENTRE AVALIAÇÃO DIAGNÓSTICA E DESEMPENHO ESCOLAR NO COLÉGIO MILITAR DE CAMPO GRANDE - VIA INTERVALOS DE CONFIANÇA *BOOTSTRAP*

ELIEL GONÇALVES VILLA NOVA

Dissertação submetida ao Programa de Pós-Graduação em Matemática em Rede Nacional, do Instituto de Matemática, da Universidade Federal de Mato Grosso do Sul, como parte dos requisitos para obtenção do título de Mestre.

Aprovado pela Banca Examinadora:

Prof. Dr. Jair da Silva - UFMS

Prof. Dr. Erlandson Ferreira Saraiva - UFMS

Profª. Drª. Maristela Missio- UEMS

CAMPO GRANDE - MS

MAIO DE 2014

Dedico este trabalho a toda a minha família, em particular à minha esposa que sempre me apoiou durante este curso.

Epígrafe

O Temor do SENHOR é o princípio da sabedoria.

Provérbios 1:7

(in) Bíblia Sagrada

AGRADECIMENTOS

Em primeiro lugar, agradeço a Deus pela salvação de minha alma, pela minha família e amigos.

À minha esposa Rebeca, que me apoiou e incentivou durante todo o curso.

Aos meus pais Emanuel e Lenita, que têm suportado a distância para a realização desse sonho.

Ao meu orientador, Prof. Dr. Jair da Silva, pelo apoio e paciência nas orientações.

Aos companheiros do curso, pela amizade construída.

Aos professores do Profmat, pelos conhecimentos transmitidos.

Ao Colégio Militar de Campo Grande, pela oportunidade dada para realizar esse sonho.

À CAPES, pelo apoio financeiro.

Resumo

O presente trabalho tem como objetivo analisar quantitativamente a aprovação do aluno no final do ano escolar com base no resultado que ele obteve na Avaliação Diagnóstica (AD), a qual é um instrumento de avaliação de conteúdos da área cognitiva que visa verificar o nível de absorção de pré-requisitos, em uma ou mais disciplinas indispensáveis à continuidade dos estudos no ano escolar pretendido.

Os dados para essa análise são provenientes do Colégio Militar de Campo Grande um dos doze Colégios Militares do Sistema Colégio Militar do Brasil. Entretanto, para essa análise, a abordagem tradicional da Estatística Inferencial não nos atente. Pois ela se baseia em teorias que não estão disponíveis para pequenas amostras, que é a realidade dos dados que desejamos analisar.

Em meio a esse problema, encontramos uma alternativa a essa abordagem, o método *bootstrap*, introduzido por Efron em 1979. Esse é um método de reamostragem o qual é amplamente aplicável, ele é utilizado pois não necessita de muitas suposições para estimação dos parâmetros das distribuições de interesse. A aplicabilidade desse método é facititada, atualmente, pela enorme capacidade de cálculo de nossos computadores, já que ele necessita de um considerável custo computacional.

Palavras chaves: Colégio Militar de Campo Grande, Avaliação Diagnóstica, Estatística Inferencial, Reamostragem, Intervalos de confiança *Bootstrap*.

Abstract

The main purpose of this research paper is to analyze, in a quantitative way, the rates of a student approval at the end of the school year based on the results he/she has gotten on the Diagnostic Assessment Test (AD), wich is an evaluation instrument that covers contents of a specific field of study (Mathematics and Portuguese) aiming at checking the level of requirements uptaking on one or more school subjeccts that are considered essential to the continuity of his/her studies in the intended school year.

The data used on this analysis comes from Colégio Militar de Campo Grande, one of the twelve schools tha are parto of the Brasilian Militar Schooling System (SCMB). Nevertheless, the traditional approch of the Statistics Inference does not aplly to our study. This analysis approach is sometimes based on dreamed patterns and assumptions, wich usually are based on theories that are not available for small samples, what are the reality of the data we intend to analyze.

Faced whith the issue of this problem, we come to an alternative for approaching, the bootstrap method, proposed by Efron in 1979. This is a resampling method which is widely applicable, it is used because it does not require many assumptions to estimate the parameters of the distributions of interest. Nowadays, the applicability of this method has been made easier by the huge capacity of calculating of computers, since it requires a considerable computational cost.

Key words: Colégio Militar de Campo Grande, Diagnostic Assessment Test, Inference Statistics, Resampling, Bootstrap Confidence Intervals.

Lista de Tabelas

2.1.1 Altura de 20 alunos 6

2.3.1 Princípio *Bootstrap* 14

3.2.1 Resultado ao final do ano letivo do 6°Ano do Ensino Fundamental 25

3.2.2 Resultado ao final do ano letivo do 7°Ano do Ensino Fundamental 25

3.2.3 Resultado ao final do ano letivo do 8°Ano do Ensino Fundamental 26

3.2.4 Resultado ao final do ano letivo do 9°Ano do Ensino Fundamental 26

3.2.5 Resultado ao final do ano letivo do 1°Ano 27

3.2.6 Resultado ao final do ano letivo do 2°Ano do Ensino Médio 27

3.2.7 Resultado ao final do ano letivo do 3°Ano do Ensino Médio 28

4.4.1 Estimativa da média de aprovados a 95% de confiança referente ao ano escolar sem considerarmos os resultados na AD 38

4.4.2 Resultados referentes ao 6°Ano do Ensino Fundamental 38

4.4.3 Resultados referentes ao 7°Ano do Ensino Fundamental 39

4.4.4 Resultados referentes ao 8°Ano do Ensino Fundamental 40

4.4.5 Resultados referentes ao 9°Ano do Ensino Fundamental 40

4.4.6 Resultados referentes ao 1°Ano do Ensino Médio 41

4.4.7 Resultados referentes ao 2°Ano do Ensino Médio 42

4.4.8 Resultados referentes ao 3°Ano do Ensino Médio 42

Lista de Figuras

2.1.1 Gráfico de setores da altura dos 20 alunos 7

2.1.2 Histograma da altura dos 20 alunos 7

2.2.1 Curva normal 11

3.2.1 Ano escolar de ingresso (2008 a 2012) 20

3.2.2 Resultado na AD dos alunos do 6°Ano Ensino Fundamental (2008 a 2012) . . 21

3.2.3 Resultado na AD dos alunos do 7° Ano Ensino Fundamental (2008 a 2012) . 22

3.2.4 Resultado na AD dos alunos do 8° Ano Ensino Fundamental (2008 a 2012) . 22

3.2.5 Resultado na AD dos alunos do 9° Ano Ensino Fundamental (2008 a 2012) . 23

3.2.6 Resultado na AD dos alunos do 1° Ano Ensino Médio (2008 a 2012) 23

3.2.7 Resultado na AD dos alunos do 2° Ano Ensino Médio (2008 a 2012) 24

3.2.8 Resultado na AD dos alunos do 3° Ano Ensino Médio (2008 a 2012) 24

4.1.1 Intervalo de confiança *bootstrap* padrão 32

4.2.1 Intervalo de confiança baseado nos percentis *bootstrap* 34

4.3.1 Intervalo de confiança percentis BC_a 37

A.2.1histograma temperaturas 51

A.2.2Temperaturas 52

A.2.3Temperaturas 53

A.2.4Temperaturas 54

Sumário

1 Introdução **1**

2 Referencial Teórico **4**

2.1 Estatística Descritiva 4

2.1.1 Conceitos básicos 4

2.1.2 Representação gráfica de uma distribuição de frequências 6

2.1.3 Características numéricas de uma distribuição de frequências 7

2.1.3.1 Medidas de posição 8

2.1.3.2 Medidas de dispersão 9

2.1.3.3 Outra medida de separação de dados 9

2.2 Introdução à Estatística Inferencial 10

2.2.1 Distribuição de probabilidades 10

2.2.1.1 Distribuição Normal 11

2.2.2 Noções sobre intervalos de confiança 12

2.3 Método de Reamostragem 12

2.3.1 O princípio do método *Bootstrap* 13

2.3.2 Estimativa do erro padrão 14

2.3.3 Intervalos de confiança *bootstrap* 15

2.3.3.1 Intervalo de confiança *bootstrap* padrão 15

2.3.3.2 Intervalo de confiança baseado nos percentis *bootstrap* . . . 15

2.3.3.3 Intervalo de confiança percentis BC_a 16

3 Metodologia, Apresentação dos Dados **18**

3.1 Referencial Metodológico 18

3.1.1 Tipo de pesquisa 18

3.1.2 Questões de estudo 19

3.1.3 Procedimentos metodológicos 19

3.2 Características dos alunos que realizaram a AD 20

4 Construção dos intervalos de confiança *bootstrap* através do R **29**

4.1 Construção do intervalo de confiança *bootstrap* padrão 30

4.2 Construção do intervalo de confiança baseado nos percentis *bootstrap* 32

4.3 Construção do Intervalo de confiança percentis BC_a 34

4.4 Resultados dos demais dados 37

4.5 Respostas às questões de estudo 43

5 Conclusão **45**

A O programa R **47**

A.1 Comandos utilizados nesse trabalho 47

A.1.1 Operações básicas 48

A.1.2 Vetores com valores numéricos 48

A.1.3 Algumas funções 48

A.1.4 Operações com vetores 50

A.2 Gráficos 50

A.3 Comandos de lógica 54

B Programa Geral **55**

Capítulo 1

Introdução

Anualmente é aplicada pelo Sistema Colégio Militar do Brasil (SCMB) uma avaliação conhecida como: Avaliação Diagnóstica (AD).

O SCMB, subordinado à Diretoria de Educação Preparatória e Assistencial (DEPA), que é um dos subsistemas de ensino do Exército Brasileiro e tem a seu cargo ministrar a educação básica, nos níveis fundamental (6^{o} a 9^{o} ano) e médio. Num total de doze Colégios Militares disseminados pelo país, oferecem educação a mais de 14400 jovens, 37% dos quais oriundos do meio civil, integrados ao sistema através de concurso público federal e os outros 63 % são dependentes de militares.

Os alunos dependentes de militares, ao ingressar no SCMB, realizam a AD compreendendo as disciplinas de Matemática e Português. Há também um teste de nivelamento da disciplina de idioma estrangeiro, Inglês, que não será objeto de estudo desse trabalho.

A AD é um instrumento de avaliação de conteúdos da área cognitiva que visa verificar o nível de absorção de pré-requisitos, em uma ou mais disciplinas indispensáveis à continuidade dos estudos no ano escolar pretendido pelo responsável para o seu dependente [1].

A AD não visa aprovar ou reprovar o ingresso de um aluno no SCMB, ou seja, a matrícula do aluno não está condicionada ao resultado na AD, ela é apenas um parecer com

relação ao conhecimento prévio do aluno que será considerado apto, apto com restrição ou inapto para frequentar o ano escolar em que se deseja matricular. Esses dados servem, dentre outros aspectos, de base para a elaboração dos programas de recuperação da aprendizagem e apoio pedagógico.

Dentro desse contexto da AD, esse trabalho visa, dentre outras questões, que apresentaremos no terceiro capítulo, analisar quantitativamente a aprovação do aluno no final do ano escolar com base no resultado que ele obteve na AD. Os dados para essa análise são provenientes do Colégio Militar de Campo Grande (CMCG) um dos doze Colégios Militares do SCMB.

Com as respostas das perguntas norteadoras desse trabalho, uma das possíveis consequências será direcionar as ações de apoio pedagógico para àqueles casos que encontrarmos uma maior probabilidade de reprovação e também orientar os pais ou responsáveis dos alunos com relação à dificuldade que eles poderão ter, a necessidade de apoio pedagógico, visando sempre o sucesso escolar desse aluno.

Sabemos que para responder essas perguntas são necessários conceitos da Ciência Estatística, que está presente no cotidiano de toda a população brasileira, pois diariamente vemos notícias tais como: 20% dos brasileiros acessam a internet diariamente, 55% da população considera bom o desempenho do presidente.

Por outro lado ouvem-se comentários de cidadãos dizendo: Nunca fui entrevistado! Será que de fato é possível chegar àquela conclusão sem que toda a população seja entrevistada? A resposta é sim, e este é, segundo Costa Neto [2], o objetivo da Estatística Inferencial, a saber: "O objetivo da Estatística Inferencial é tirar conclusões sobre populações com base nos resultados observados em amostras extraídas dessas populações."

Além da Estatística Inferencial, que visa à análise e interpretação de dados, a Ciência Estatística também tem outro ramo, que organiza e faz a descrição de dados experimentais, esta, por sua vez, é chamada de Estatística Descritiva, a qual é geralmente estudada no ensino básico.

Entretanto, somente a Estatística Descritiva não é o bastante para responder as questões de estudo dessa pesquisa, pois ela se baseia, por vezes, em modelos idealizados e suposições, geralmente, as expressões para medidas de precisão, tais como o desvio padrão são baseados em teorias que não estão disponíveis para pequenas amostras, que é a realidade dos dados que desejamos analisar.

Em meio a esse problema, encontramos uma alternativa a essa abordagem, o método *bootstrap*, introduzido por Efron em 1979. Esse é um método de reamostragem o qual é amplamente aplicável a pequenas amostras. A aplicabilidade desse método é facilitada, atualmente, pela enorme capacidade de cálculo de nossos computadores, já que ele necessita de um considerável custo computacional.

Para que se atinjam os objetivos desse trabalho optamos por dividi-lo em cinco capítulos, sendo que no primeiro, tratamos da introdução, apresentamos os objetivos, e que para atingi-los, foi necessário o uso do método *bootstrap*.

No segundo capítulo, desenvolvemos noções de estatística descritiva, estatística inferencial, a teoria do método *bootstrap.*

Já no terceiro capítulo apresentamos o referencial metodológico destacando as questões de estudo dessa pesquisa. Posteriormente, apresentamos os dados das AD aplicadas nos anos de 2008 a 2012 no CMCG.

No quarto capítulo, construímos os intervalos de confiança, conforme a teoria do capítulo dois, buscando responder às questões de estudo que deram origem a essa pesquisa. Ainda nesse capítulo, fizemos as primeiras análises dos resultados encontrados.

Finalmente, no quinto e último capítulo apresentamos a conclusão, bem como relatamos outras observações oriundas da análise dos dados, em seguida, nos apêndices, desenvolvemos a teoria do software computacional R, que utilizamos para construir os intervalos de confiança *bootstrap*. Posteriormente, apresentamos as referências bibliográficas.

Capítulo 2

Referencial Teórico

Esse capítulo subdivide-se em três seções. Na primeira desenvolveremos conceitos atinentes à estatística descritiva, na segunda, noções gerais de intervalos de confiança, posteriormente na terceira seção a teoria do método de reamostragem *bootstrap*.

2.1 Estatística Descritiva

Apresentaremos alguns conceitos fundamentais da Estatística Descritiva, sobretudo aqueles essenciais para o desenvolvimento deste trabalho. Essa seção traz definições e exemplos baseados, principalmente, nos livros: Estatística [2] , Matemática2 [3] e Fundamentos de Matemática Elementar [4].

2.1.1 Conceitos básicos

Inicialmente, é necessário diferenciar as características dos dados a serem pesquisados, isto é, classificando-os em relação aos tipos de variáveis que eles representam. Essas variáveis se classificam em qualitativa nominal ou ordinal, e quantitativa contínua ou discreta.

Uma variável é dita qualitativa nominal quando seus valores representam atributos ou qualidades, mas não têm uma relação de ordem entre eles, por exemplo: sexo, grupo sanguíneo. É dita qualitativa ordinal quando seus valores também representam atributos ou

qualidades, mas têm uma relação de ordem entre eles, por exemplo: classe social, grau de instrução.

Por outro lado, uma variável é dita quantitativa contínua quando seus valores são medidos em escala métrica e em que valores fracionários são possíveis, por exemplo, altura, temperatura. É dita quantitativa discreta quando seus valores são medidos em escala métrica em que só são possíveis valores inteiros, por exemplo, número de filhos, número de alunos.

A seguir, veremos algumas outras importantes definições e, posteriormente, exemplos.

Definição 1. Chama-se de universo estatístico ou população estatística o conjunto formado por todos os elementos que possam oferecer dados pertinentes ao assunto em questão.

Exemplo 1. Altura de todos os 1100 alunos do CMCG.

Definição 2. Chama-se amostra qualquer subconjunto da população.

Exemplo 2. Altura, em metros, de 20 alunos do CMCG: $1.76, 1.56, 1.68, 1.74, 1.68, 1.68, 1.76, 1.68, 1.63, 1.70, 1.70, 1.85, 1.74, 1.64, 1.56, 1.76, 1.68, 1.80, 1.83, 1.71$.

Definição 3. Se organizarmos esses mesmos dados numéricos brutos, em ordem crescente ou decrescente, a lista recebe o nome de rol.

Exemplo 3. Altura dos 20 alunos em ordem crescente: $1.56, 1.56, 1.63, 1.64, 1.68, 1.68, 1.68, 1.68, 1.68, 1.70, 1.70, 1.71, 1.74, 1.74, 1.76, 1.76, 1.76, 1.80, 1.83, 1.85$.

Definição 4. Classe é qualquer intervalo real, aberto, semi-aberto ou fechado que contenha um rol da amostra.

Aqui, nesse texto, não discutiremos critérios para formação de classes.

Observação 1. O símbolo $a \vdash b$ indica o intervalo semi-aberto $[a, b)$, isto é, fechado a e aberto em b .

Exemplo 4. Podemos dividir os dados em cinco classes, a saber: $1.56 \vdash 1.62$, $1.62 \vdash 1.68$, $1.68 \vdash 1.74$, $1.74 \vdash 1.80$, $1.80 \vdash 1.86$.

Definição 5. Frequência Absoluta é o número de vezes que o elemento aparece na amostra ou o número de elementos pertencentes a uma mesma classe. Temos também a Frequência Relativa que é dada pela razão entre a frequência absoluta e o total da amostra.

Classe	Frequência Absoluta	Frequência Relativa
$1.56 \vdash 1.62$	2	$\frac{2}{20} = 0,1 = 10\%$
$1.62 \vdash 1.68$	2	$\frac{2}{20} = 0,1 = 10\%$
$1.58 \vdash 1.74$	8	$\frac{8}{20} = 0,4 = 40\%$
$1.74 \vdash 1.80$	5	$\frac{5}{20} = 0,25 = 25\%$
$1.80 \vdash 1.86$	3	$\frac{3}{20} = 0,15 = 15\%$
$\sum$	20	100%

Tabela 2.1.1: Altura de 20 alunos

2.1.2 Representação gráfica de uma distribuição de frequências

Uma distribuição de frequências pode ser representada graficamente. A seguir, veremos dois tipos de gráficos, os quais são conhecidos como gráficos de informação, que é o histograma e o gráfico de setores.

O gráfico de setores, ou diagrama circular é mais indicado para variáveis qualitativas, ou quando temos classes unitárias, isto é, quando apenas um nome ou um número representa a classe. No exemplo das alturas dos vinte alunos, se cada classe for representado pela altura média da classe. Temos o seguinte gráfico de setores:

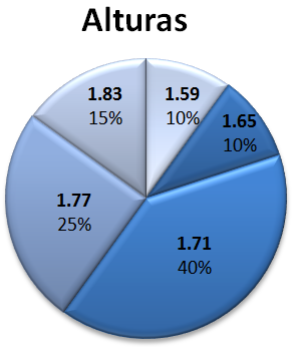

Figura 2.1.1: Gráfico de setores da altura dos 20 alunos

O histograma é um gráfico utilizado para representar uma distribuição de frequência em que as classes não são unitárias, vejamos, a seguir, o histograma referente às alturas dos 20 alunos.

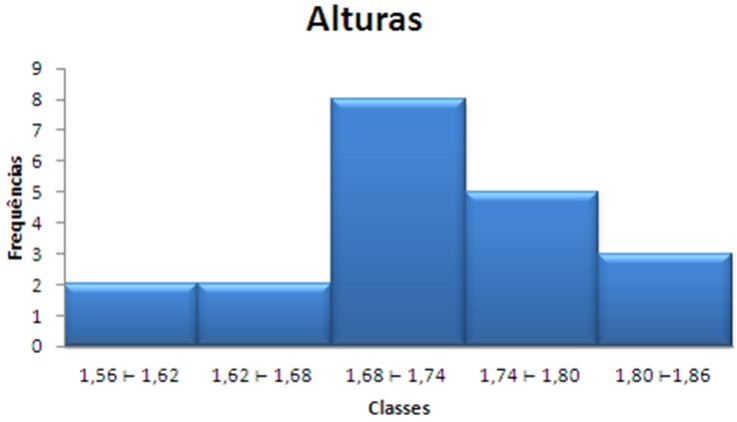

Figura 2.1.2: Histograma da altura dos 20 alunos

2.1.3 Características numéricas de uma distribuição de frequências

Além da representação gráfica, muitas vezes é necessário sumariar certas características das distribuições de frequências, por meio de quantidades denominadas medidas

da distribuição de frequências, as quais procuram quantificar alguns de seus aspectos de interesse. Temos, assim, as chamadas medidas de posição e de dispersão.

2.1.3.1 Medidas de posição

As medidas de posição servem para localizarmos a distribuição de frequências sobre o eixo de variação da variável em questão, a seguir, veremos três dessas medidas: a média, a mediana e a moda.

Definição 6. Sendo $x_i\,(i = 1, 2, \ldots, n)$ um conjunto de dados, a sua média aritmética ou, simplimente, média, é dada por:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.1.1}$$

Exemplo 5. Certo aluno obteve em sete provas as seguintes notas: $3, 5, 7.5, 6, 9, 8.5$ e 10. Assim a média das notas é igual a: $\overline{x} = \frac{3+5+7.5+6+9+8.5+10}{7} = 7$.

Note que esse valor não representa uma nota que ele obteve e, sim, uma tendência central das notas.

Definição 7. Sendo $x_i\,(i = 1, 2, \ldots, n)$ um conjunto de dados, a sua mediana é o termo central do rol desses dados. Caso a quantidade n dos dados seja ímpar a mediana é valor de ordem $(n+1)/2$, caso contrário, a mediana é o valor médio entre os valores de ordem $n/2$ e $(n/2)+1$ do conjunto de dados.

Exemplo 6. A mediana das notas será o quarto termo, pois, $(7+1)/2 = 4$. Logo a mediana é a nota 7.5.

Definição 8. A moda (ou modas) de um conjunto de valores é o valor (valores) de máxima frequência.

Exemplo 7. No exemplo 3, a moda é a altura 1.68 e, no caso da tabela 2.1.1 a classe modal é $1.58 \vdash 1.74$.

2.1.3.2 Medidas de dispersão

Geralmente, a informação fornecida pelas medidas de posição necessita em geral de ser complementada pelas medias de dispersão, como o próprio nome diz, elas servem para indicar o quanto os dados se apresentam dispersos em torno da região central.

Definição 9. A variância de um conjunto de dados é dada por:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} \tag{2.1.2}$$

A variância é, então, o desvio quadrático médio, ou a média dos quadrados das diferenças dos valores em relação à sua própria média. Nessa definição, estamos considerando implicitamente que os dados se referem a uma amostra, razão pela qual, utilizamos $n-1$ no denominador, mais detalhes podem ser vistos em Costa Neto, p.57.

Definição 10. O desvio padrão é a raiz quadrada da variância. Assim, o cálculo do desvio padrão é dado por:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}} \tag{2.1.3}$$

O desvio padrão se expressa na mesma unidade da variável, sendo, por isso, de maior interesse que a variância nas aplicações práticas. Além disso, ele é mais realístico para efeito de comparação de dispersões.

2.1.3.3 Outra medida de separação de dados

Vimos, na definição 7, que a mediana é um valor que divide um conjunto de dados em duas partes iguais. Agora, veremos outra medida de separação de dados.

Definição 11. O $n-$ésimo percentil $(n = 1, 2, ..., 99)$ é o valor que divide um conjunto de dados em duas partes tais $n\,\%$ dos valores da distribuição são menores ou iguais a ele e $(100 - n)\%$ são maiores ou iguais a ele.

Exemplo 8. O décimo quarto percentil é o número que divide os dados de tal forma que 14% são menores que ele e 86% são maiores. Note que a mediana equivale ao quinquagésimo percentil.

2.2 Introdução à Estatística Inferencial

Nessa seção apresentaremos noções gerais sobre a inferência estatística, sobretudo as definições que são necessárias para uma melhor compreensão do método de reamostragem *bootstrap*. As definições dessa seção provém, principalmente, das referências [5, 6, 7].

A inferência estatística é o processo de se obter informações sobre uma população a partir de resultados observados na amostra.

A estatística não paramétrica, como o próprio nome sugere, independe dos parâmetros populacionais (média, desvio padrão) bem como suas estimativas que são dadas pelas amostras.

2.2.1 Distribuição de probabilidades

Definição 12. Sejam E um experimento aleatório e S o espaço associado ao experimento. Uma função X, que associe a cada elemento $s \in S$ um número real $X(s)$ é denominado variável aleatória.

Exemplo 9. E : lançamento de duas moedas; X : número de caras (k) obtidas nas duas moedas; $S = \{(c,c),(c,k),(k,c),(k,k)\}$. $X = 0$, correspode ao evento (c,c) com probabilidade $\frac{1}{4}$.

Seja X uma variável aleatória contínua, temos a seguinte definição:

Definição 13. Uma função $f(x)$, $Rx \longmapsto \mathbb{R}$, é dita função densidade de probabilidade se para todo $x \in Rx$ temos $f(x) \geqslant 0$ e $\int_{Rx} f(x)dx = 1$.

Além disso, define-se, para qualquer $a < b$ em Rx a probabilidade de X estar entre a e b como: $P(a < X < b) = \int_a^b f(x)dx$.

2.2.1.1 Distribuição Normal

Dizemos que X tem distribuição normal com média μ e variância σ^2, que denotaremos por $X \sim N(\mu, \sigma^2)$, se a função de densidade de probabilidade X é dada por:

$$f\left(x \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \; -\infty < x < \infty \tag{2.2.1}$$

em que $-\infty < \mu < \infty$ e $\sigma^2 > 0$. Nesse caso, μ e σ^2 são denominados parâmetros da distribuição.

O gráfico da distribuição normal padrão, isto é, $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$, é apresentado na figura a seguir:

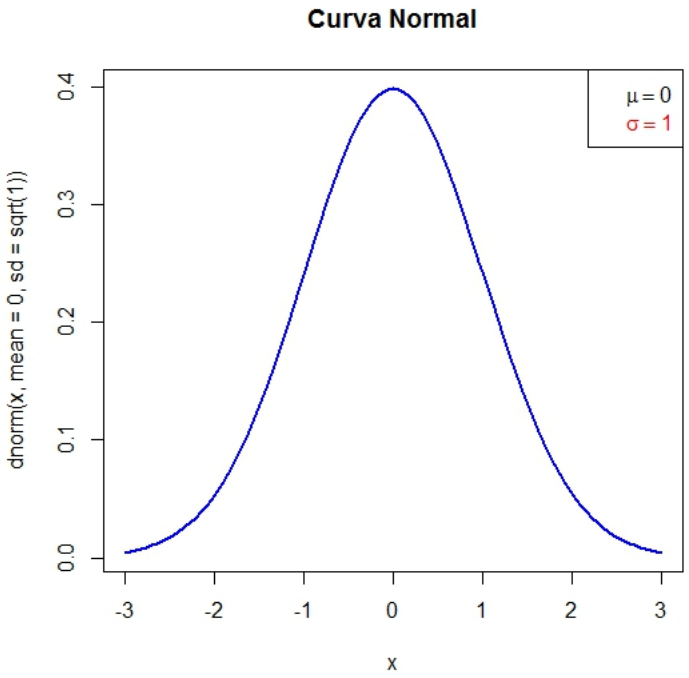

Figura 2.2.1: Curva normal

2.2.2 Noções sobre intervalos de confiança

Por vezes, a indicação de um único valor como estimativa de um parâmetro não nos dá a informação sobre a precisão de tal valor. Por isso, em muitas situações, interessa-nos dar uma medida de erro, ϵ, para indicar que o verdadeiro valor do parâmetro está muito provavelmente entre $\hat{\theta}-\epsilon$ e $\hat{\theta}+\epsilon$.

Definição 14. Um intervalo de estimação de um parâmetro θ é um intervalo da forma $\hat{\theta}_1 < \theta < \hat{\theta}_2$, em que $\hat{\theta}_1$ e $\hat{\theta}_2$ são dois valores assumidos pelo estimador $\hat{\theta}$ do parâmetro θ populacional , face a uma amostra concreta.

A medida da confiança com que aquele intervalo conterá o verdadeiro valor do parâmetro é feita em termos de probabilidades.

Definição 15. O intervalo $\hat{\theta}_1 < \theta < \hat{\theta}_2$, calculado para uma amostra concreta chama-se intervalo de confiança a $100(1-\alpha)\%$, em que $1-\alpha$ denomina-se coeficiente de confiança e, α, nível de significância.

Sumariando a definição acima, se $\alpha = 0.05$, siginifica que temos uma confiança de 95% de que nosso intervalo contenha o verdadeiro valor do parâmetro.

Diferentes amostras, conduzem a diferentes estimadores, produzindo portanto diferentes intervalos de confiança.

Quanto maior for o intervalo, maior é grau de confiança que temos de que ele contenha o verdadeiro valor do parâmetro desconhecido, mas não há interesse em se ter um intervalo muito largo. O ideal é um intervalo curto como probabilidade elevada.

2.3 Método de Reamostragem

A reamostragem de um conjunto de dados tem por finalidade criar replicas dos dados, a partir das quais podemos avaliar a variabilidade de quantidades de interesse, sem usar cálculos analíticos.

A amostra de uma população representa a população da qual ela foi extraída. Dessa maneira, as reamostras obtidas a partir dessa amostra representam o que obteríamos se retirássemos diversas amostras da população.

A ideia básica do método de reamostragem é que, na ausência de qualquer outra informação sobre a distribuição, a amostra observada contém toda a informação disponível sobre a distribuição subjacente, e, portanto, uma nova amostra da amostra é o melhor guia para o que pode ser esperado da distribuição em questão.

As definições das subseções a seguir provém das referências [8, 9, 10, 11, 12].

2.3.1 O princípio do método *Bootstrap*

Observada uma amostra aleatória x_1,x_2,...,x_n de uma distribuição F, defini-se uma função de distribuição empírica $\hat{F}$ como uma distribuição discreta, que atribui probabilidade $\frac{1}{n}$ a cada valor X_i, $i = 1,2, \ldots, n$. Assim, $X = (x_1,x_2,\ldots,x_n)$ representa o vetor dos dados, para os quais se calcula o estimador $\hat{\theta} = s(X)$ de um parâmetro de interesse da população.

Uma amostra *bootstrap* $X^* = (x_1^*,x_2^*,\ldots,x_n^*)$ é obtida reamostrando aleatoriamente n vezes, com reposição, as observações elementos $X = (x_1,x_2,\ldots,x_n)$. Se forem geradas B amostras *bootstrap* X^{*1},X^{*2},...,X^{*B}, de forma independente, então estima-se θ em cada ua dessas amostras por

$$\hat{\theta^*}(b) = s(X^{*B}),b = 1, 2, \ldots, B \tag{2.3.1}$$

O princípio de *bootstrap* pode ser resumido pela seguinte tabela

Amostra real			**Amostra *Bootstrap***	
Distribuição de probabilidade desconhecida	Amostra aleatória observada		Distribuição empírica	Amostra *bootstrap*
F	$\rightarrow X = (x_1, x_2, \ldots, x_n)$	$\Rightarrow$	$\hat{F} \rightarrow X^* = (x_1^*, x_2^*, \ldots, x_n^*)$	
$\downarrow$			$\downarrow$	
$\hat{\theta} = s(X)$			$\hat{\theta^*} = s(X^*)$	
Parâmetro de interesse			Replicações *bootstrap*	

Tabela 2.3.1: Princípio *Bootstrap*

2.3.2 Estimativa do erro padrão

Definição 16. A expressão para o estimador *bootstrap* do erro padrão é dada por

$$\hat{ep}_{boot}\left(\hat{\theta}\right) = \sqrt{\sum_{b=1}^{B} \frac{\left[s(x^{*b}) - s(.)\right]^2}{B-1}}, \tag{2.3.2}$$

em que

$$s(.) = \sum_{b=1}^{B} \frac{s(x^{*b})}{B}, \tag{2.3.3}$$

ou seja, o estimador do erro padrão é o desvio-padrão de suas replicações.

Efron & Tibshirani [8] chamam de estimador erro-padrão *bootstrap* ideal para a distribuição F o limite de $\hat{ep}_{boot}$ quando B vai para o infinito, ou seja, $\lim_{B\to\infty} \hat{ep}_{boot} = ep_{\hat{F}}\left(\hat{\theta^*}\right)$.

O estimador *bootstrap* ideal e sua aproximação $\hat{ep}_{boot}$ são chamados estimadores *bootstrap* não paramétricos, já que se baseiam em $\hat{F}$, um estimador não paramétrico de F.

Um estimador *bootstrap* paramétrico do erro padrão é baseado em um estimador $\hat{F}$, de F derivado de um modelo paramétrico. Por exemplo, ao invés de estimarmos F pela função distribuição empírica $\hat{F}$, podemos assumir que a população tem distribuição normal.

2.3.3 Intervalos de confiança *bootstrap*

Com o uso do método de reamostragem *bootstrap* podemos construir intervalos de confiança com $100(1-\alpha)\%$ de certeza para o parâmetro de interesse $\hat{\theta}$. Descrevemos nas próximas subseções, três diferentes métodos de construção de intervalos de confiança *bootstrap* chamados de *bootstrap* padrão (2.3.3.1), percentis *bootstrap* (2.3.3.2) e *bias-corrected and acelerated* que tem como abreviação padrão BC_a (2.3.3.3).

2.3.3.1 Intervalo de confiança *bootstrap* padrão

Com os valores para o estimador *bootstrap* do erro padrão $\hat{ep}_{boot}$ e o valor de $\hat{\theta} = s(x)$ da amostra original, o intervalo com probabilidade de confiança $100(1-\alpha)\%$ é dado por,

$$\left(\hat{\theta} - Z_{.\left(1-\frac{\alpha}{2}\right)}\,\hat{ep}_{boot}\,, \hat{\theta} + Z_{.\left(1-\frac{\alpha}{2}\right)}\,.\,\hat{ep}_{boot}\right) \tag{2.3.4}$$

Sendo que Z_α é o 100α-ésimo percentil de uma distribuição normal padrão. Esse método torna-se vantajoso pela simplicidade algébrica, note que a definição (2.3.4) é uma consequência de $Z = \frac{\hat{\theta}-\theta}{\hat{ep}_{boot}} \sim N(0,1)$, isto é, é aproximadamente a distribuição normal padrão.

2.3.3.2 Intervalo de confiança baseado nos percentis *bootstrap*

Após serem realizadas replicações X^* de $X = (x_1,x_2,\ldots,x_n)$, e, posteriormente estimadas as replicações *bootstrap* do parâmetro de interesse (2.3.1), o intervalo de confiança de probabilidade $100(1-\alpha)\%$ construído pelo método percentil é obtido pelos $\frac{\alpha}{2}$-ésimo e $\left(1-\frac{\alpha}{2}\right)$-ésimo percentis de $\hat{G}$, que é definida como a função distribuição acumulada de $\hat{\theta^*}$. Uma expressão para o intervalo ora mencionado é dada por

$$\left[\hat{\theta}_{\%,\,inf}\,,\,\hat{\theta}_{\%,\,sup}\right] = \left[\hat{G}^{-1}_{\left(\frac{\alpha}{2}\right),}\,\hat{G}^{-1}_{\left(1-\frac{\alpha}{2}\right)}\right]. \tag{2.3.5}$$

Como $\hat{G}^{-1}(\alpha) = \hat{\theta}^{*(\alpha)}$, o $100(1-\alpha)\%$-ésimo percentil de $\hat{\theta^*}$, podemos reescrever

os intervalos percentis na seguinte forma

$$\left[\hat{\theta}_{\%,inf}\,,\,\hat{\theta}_{\%,sup}\right] = \left[\hat{\theta}^{*}_{\left(\frac{\alpha}{2}\right)},\,\hat{\theta}^{*}_{(1-\alpha)}\right]. \tag{2.3.6}$$

As expressões (2.3.5) e (2.3.6) referem-se à situação ideal do *bootstrap* na qual o número de replicações é infinito.

Na prática devemos usar um número finito B de replicações. Para o processo, geramos B conjuntos de dados *bootstrap* X^{*1},X^{*2},$\ldots$, X^{*B} e calculamos as replicações *bootstrap* $\hat{\theta^{*}}(b) = s(X^{*B})$,$b = 1, 2, \ldots, B$.

Seja $\hat{\theta}^{*\mathbf{B}}_{(\alpha)}$o 100α-ésimo percentil empírico dos valores $\hat{\theta^{*}}(b)$, ou seja, o valor $(B.\alpha)$-ésimo na lista ordenada das B replicações de $\hat{\theta^{*}}$. Assim, se $B = 100$ e $\alpha = 0,05$ então $\hat{B}^{*100}_{(0,05)}$é o quinto termo dos valores ordenados de $\hat{\theta^{*}}$. Se $(B.\alpha)$ não é um número inteiro, utiliza-se o maior inteiro menor ou igual a $(B+1).\alpha$.

Como a distribuição *bootstrap* de $\hat{\theta^{*}}$ é aproximada, melhores resultados serão obtidos de tamanho n grande, e quanto maior for B, melhores serão os intervalos estimados. Assim, o intervalo percentil aproximado de $100(1-\alpha)\%$ de confiança é dado por

$$\left[\hat{\theta}_{\%,inf}\,,\,\hat{\theta}_{\%,sup}\right] = \left[\hat{\theta}^{*B}_{\left(\frac{\alpha}{2}\right)},\,\hat{\theta}^{*B}_{\left(1-\frac{\alpha}{2}\right)}\right]. \tag{2.3.7}$$

2.3.3.3 Intervalo de confiança percentis BC_a

O método BC_a também utiliza os percentis da distribuição *bootstrap* para a construção dos intervalos de confiança para parâmetros de interesse, este método utiliza percentis que dependem de duas constantes, $\hat{a}$ que é denominado aceleração e $\hat{z}_0$ que é a correção para tendência, daí vem a abreviatura BC_a, isto é, *bias-corrected e acceleration*.

O intervalo BC_a de desejada probabilidade$100(1-\alpha)\%$ é dado por

$$\left[\hat{\theta}_{\%,inf}\,,\,\hat{\theta}_{\%,sup}\right] = \left[\hat{\theta}^{*}_{(\alpha_1)},\,\hat{\theta}^{*}_{(\alpha_2)}\right] \tag{2.3.8}$$

sendo,

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + Z_{\left(\frac{\alpha}{2}\right)}}{1 - \hat{a}\left(\hat{z}_0 + Z_{\left(\frac{\alpha}{2}\right)}\right)}\right) \tag{2.3.9}$$

e

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + Z_{\left(1-\frac{\alpha}{2}\right)}}{1 - \hat{a}\left(\hat{z}_0 + Z_{\left(1-\frac{\alpha}{2}\right)}\right)}\right) \tag{2.3.10}$$

em que Φ é a função distribuição acumulada de uma normal padrão e $Z_{(\alpha)}$ é o 100α-ésimo percentil de uma distribuição normal padrão. Note que se $\hat{a}$ e $\hat{z}_0$ são iguais a zero, (2.3.8) é similar a (2.3.7).

Para calcularmos $\hat{z}_0$ utilizamos a seguinte expressão

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\left\{\hat{\theta}^*(b) < \hat{\theta}\right\}}{B}\right), \tag{2.3.11}$$

e , dentre as várias possibilidades para se obter $\hat{a}$ utilizamos, em termos de valores *jackknife* de $\hat{\theta} = s(X)$,

$$\hat{a} = \frac{\sum_{i=1}^{n}\left(\hat{\theta}_{(.)} - \hat{\theta}_{(i)}\right)^3}{6\left\{\sum_{i=1}^{n}\left(\hat{\theta}_{(.)} - \hat{\theta}_{(i)}\right)^2\right\}^{\frac{3}{2}}} \tag{2.3.12}$$

em que $\hat{\theta_{(i)}} = s(X_{(i)})$, $i = 1, 2, \ldots, n$ e com $\hat{\theta}_{(.)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)}$.

Capítulo 3

Metodologia, Apresentação dos Dados

3.1 Referencial Metodológico

Nessa seção explicitaremos a metodologia científica empregada nesse trabalho, para tanto, nos basearemos na referência [13]. Descreveremos o tipo de pesquisa realizada, classificando-a quanto ao seu tipo, suas questões de estudo e, por fim, explicitaremos os procedimentos metodológicos empregados.

3.1.1 Tipo de pesquisa

Essa pesquisa caracterizou-se quanto à natureza como aplicada uma vez que se realizou no âmbito do CMCG, na qual se objetivou gerar conhecimentos para aplicações práticas dirigidos à solução de problemas específicos do aluno em seu primeiro ano no CMCG. Sua forma de abordagem foi quantititativa mensurando dados numéricos, isto é, a porcentagem de aprovação em cada ano escolar, sem considerarmos as situações qualitativas envolvidas.

Com relação aos procedimentos técnicos, foi feito levantamentos de dados numéricos cedidos pelo CMCG, e uma revisão dos conceitos matemáticos necessários para a análise dos dados coletados.

3.1.2 Questões de estudo

Nos primeiros anos em que atuei como professor no CMCG participei da banca de correção das avaliações diagnósticas aplicadas nessa instituição de ensino. Devido a essa participação, surgiu o desejo de se fazer uma análise dos resultados obtidos pelos alunos comparando-os com o seu desempenho ao término do ano letivo. Assim, o programa do mestrado veio ao encontro dessa aspiração.

Fruto desse desejo, elaboramos quatro questões de estudo que são o objeto da pesquisa desse trabalho, a saber:

1. Em qual ano, do fundamental ou médio, um aluno mesmo considerado inapto na AD tem maior probabilidade de ser aprovado?

2. O oposto, ou seja, mesmo considerado apto na AD, em qual ano tem maior probabilidade de ser reprovado?

3. Ainda, de acordo com o resultado na AD, em cada ano escolar, qual a probabilidade de aprovação?

4. Finalmente, desejamos conhecer de modo geral a probabilidade de aprovação de um aluno novo em cada ano escolar, sem considerarmos o resultado na AD.

3.1.3 Procedimentos metodológicos

A pesquisa foi conduzida de modo que inicialmente coletamos os dados na Seção Técnica de Ensino do CMCG, posteriormente, fizemos uma primeira análise, separando-os de acordo com interesse da pesquisa, isto é, de acordo com ano escolar de ingresso e o resultado na AD.

Após essa primeira análise, verificamos que eram poucos os dados, era então necessário, antes de uma análise, usarmos um método de reamostragem. Para tanto, escolhemos

o método de reamostragem *bootstrap*. Findo o estudo desse método, buscamos analisar os resultados encontrados buscando responder as questões de estudo.

3.2 Características dos alunos que realizaram a AD

Inicialmente foram recebidos os dados brutos de aproximadamente 900 alunos que ingressaram no CMCG entre janeiro de 2008 e fevereiro de 2012. No momento em que os dados brutos foram organizados constatou-se que apenas 823 deles continham todas as informações pertinentes a este trabalho, a saber, dados referentes ao resultado na AD de Português e de Matemática e o resultado final do aluno (aprovação/reprovação) no ano letivo em questão.

De posse, então, desses 823 dados, ora mencionados, os mesmos foram separados por ano escolar de ingresso, isto é, alunos que ingressaram no 6° ano do Ensino Fundamental, aqueles que ingressaram no 7°ano e assim por diante, até o 3°ano do Ensino Médio. A porcentagem do total de alunos por ano escolar de ingresso pode ser vista no gráfico de setores abaixo

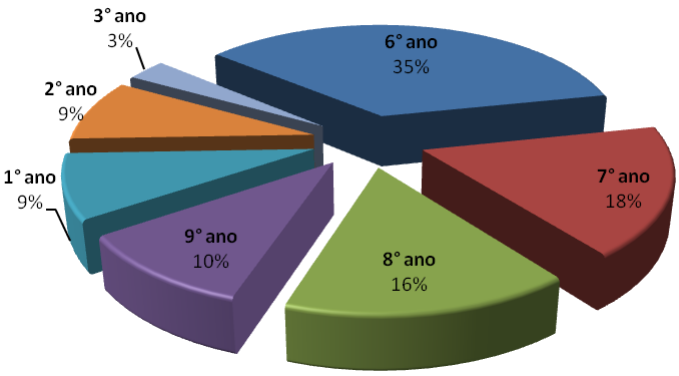

Figura 3.2.1: Ano escolar de ingresso (2008 a 2012)

Inicialmente percebemos, por meio da figura 3.2.1, que a maioria dos dados, isto é, alunos que ingressaram no CMCG entre janeiro de 2008 e fevereiro de 2012, referem-se aos

alunos do 6° ano do Ensino Fundamental e diminuem gradativamente com os anos escolares subsequentes.

Em cada ano escolar de ingresso, foram separados esses dados em nove categorias, cada uma de acordo com o resultado na AD de Português e de Matemática. Ressalta-se que em cada uma dessas avaliações o aluno poderia ser considerado Apto (AP), Apto com restrição (AP c R) e Inapto (IN).

Em todos os gráficos a seguir utilizamos a mesma notação, a saber: o primeiro dado da sigla se refere à AD de Português e o segundo à AD de Matemática, por exemplo, a sigla (AP / AP c P) mostra as porcentagens daqueles que foram aptos em Português e aptos com restrição em Matemática.

Assim sendo, os resultados obtidos nas AD pelos alunos em cada ano escolar de ingresso de 2008 a 2012 podem ser vistas nos sete gráficos a seguir.

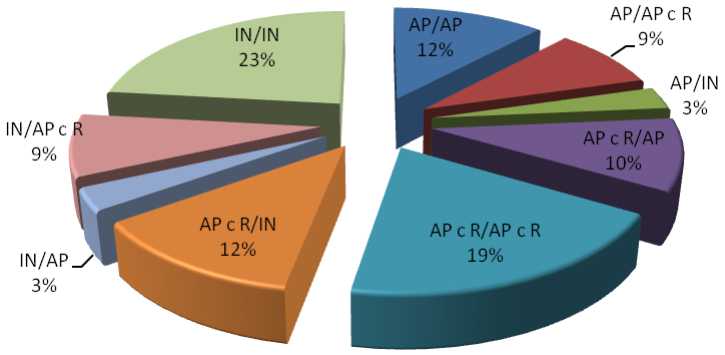

Figura 3.2.2: Resultado na AD dos alunos do 6°Ano Ensino Fundamental (2008 a 2012)

A maior parte dos alunos que ingressaram no 6° ano do Ensino Fundamental, no período em questão, foram considerados inaptos em Português e em Matemática, figura 3.2.2.

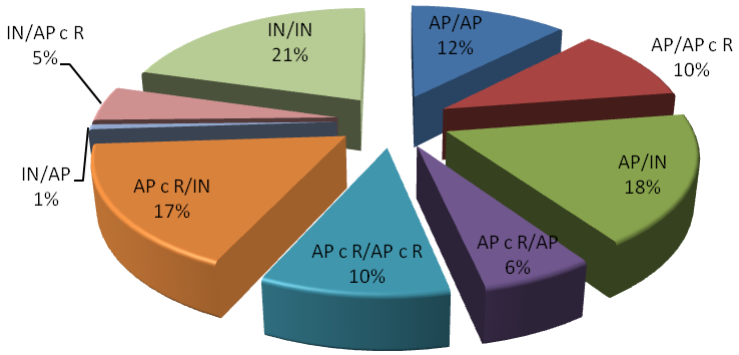

Figura 3.2.3: Resultado na AD dos alunos do 7° Ano Ensino Fundamental (2008 a 2012)

Em relação aos ingressantes no 7° ano do Ensino Fundamental, novamente a maior parte dos alunos foram considerados inaptos em Português e em Matemática, figura 3.2.3.

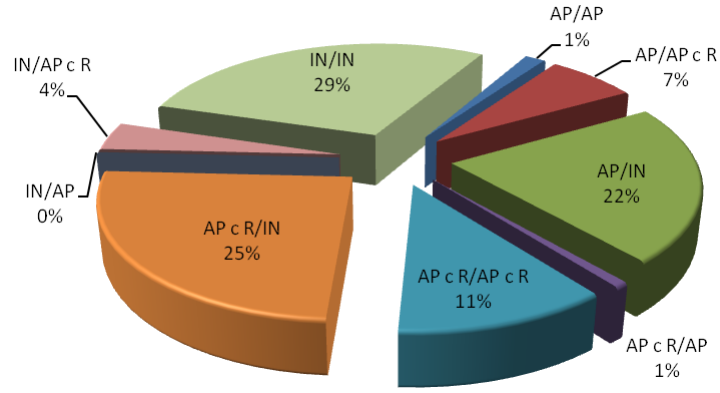

Figura 3.2.4: Resultado na AD dos alunos do 8° Ano Ensino Fundamental (2008 a 2012)

No mesmo viés dos alunos ingressantes do 6° e 7° anos, os ingressantes no 8° ano do Ensino Fundamental foram considerados, em sua maior parte, inaptos em ambas disciplinas avaliadas, tabela 3.2.4.

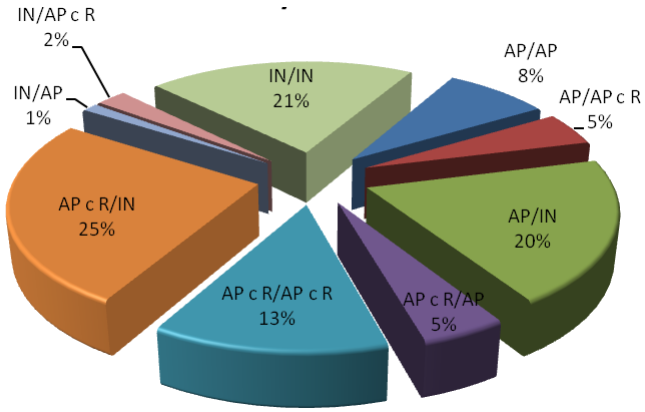

Figura 3.2.5: Resultado na AD dos alunos do 9° Ano Ensino Fundamental (2008 a 2012)

Verificamos na figura 3.2.5 que em relação aos alunos do último ano do Ensino Fundamental, isto é, 9° ano, uma mudança na característica dos alunos, a maior parte dos ingressantes continua sendo considerada inapta em Matemática, porém apta com restrição em Português.

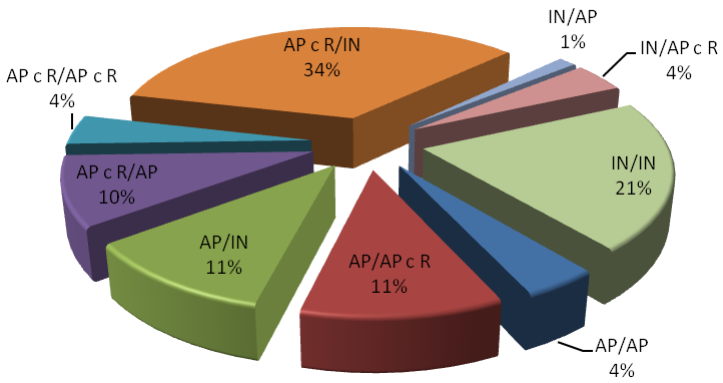

Figura 3.2.6: Resultado na AD dos alunos do 1° Ano Ensino Médio (2008 a 2012)

Na figura 3.2.6 verificamos as características dos alunos ingressantes no 1°ano do Ensino Médio. Destacamos nessa figura que 66% foi considerado inapto em Matemática. Mais de um terço do total foi considerado AP c R/IN.

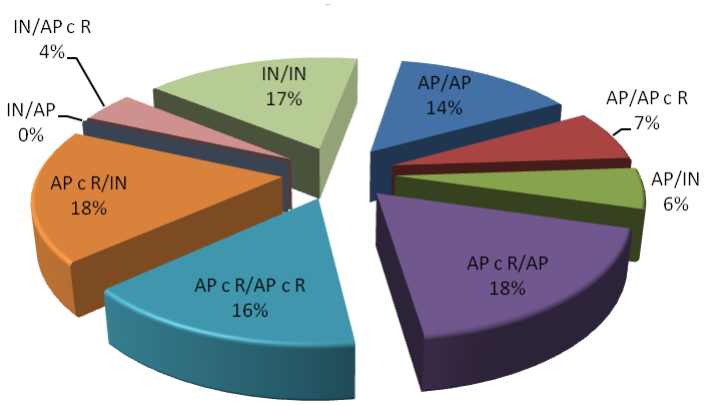

Figura 3.2.7: Resultado na AD dos alunos do 2° Ano Ensino Médio (2008 a 2012)

Os alunos novos no 2°ano do Ensino Médio, conforme figura 3.2.7, em sua maior parte foram considerados AP c R/IN e AP c R/AP cada desses um representam 18% do total, os inaptos em Matemática totalizam mais de 40%.

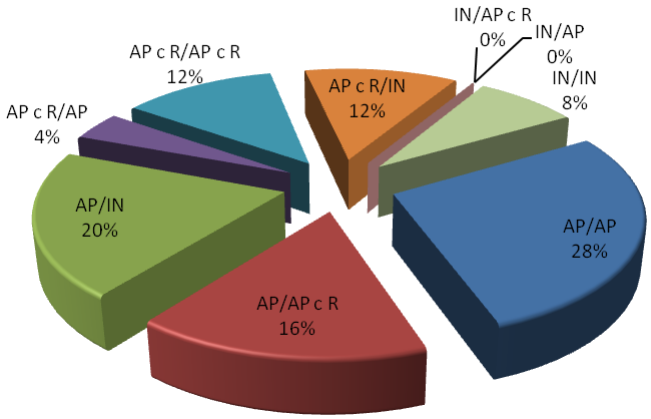

Figura 3.2.8: Resultado na AD dos alunos do 3° Ano Ensino Médio (2008 a 2012)

Finalmente, em relação aos alunos ingressantes no 3°ano do Ensino Médio, verificamos, conforme figura 3.2.8 que a maioria, isto é, 28%, é considerada AP/AP, se considerarmos apenas os aptos em Português esse valor chega a 64%.

Como notado em todos gráficos anteriores, ao realizar a AD de Português e Matemática cada aluno poderia ser considerado Apto (AP), Apto com restrição (AP c R) e Inapto (IN) e cada uma dessas disciplinas, totalizando nove classes possíveis de enquadramento. Os resultados finais, isto é, aprovação ou repovação, obtidos por esses alunos baseado em cada classe possível de enquadramento, no período de 2008 a 2012, podem ser vistos nas sete tabelas a seguir.

Tabela 3.2.1: Resultado ao final do ano letivo do 6°Ano do Ensino Fundamental

Resultado na AD	Aprovado	Reprovado	Total
AP/AP	35	0	35
AP/AP c R	24	1	25
AP/IN	9	0	9
AP c R/AP	29	0	29
AP c R/AP c R	53	4	57
AP c R/IN	29	7	36
IN/AP	9	0	9
IN/AP c R	20	5	25
IN/IN	42	26	68
Total	250	43	293

Tabela 3.2.2: Resultado ao final do ano letivo do 7°Ano do Ensino Fundamental

Resultado na AD	Aprovado	Reprovado	Total
AP/AP	18	0	18
AP/AP c R	15	0	15
AP/IN	23	3	26
AP c R/AP	9	0	9
AP c R/AP c R	14	1	15
AP c R/IN	22	3	25
IN/AP	0	1	1
IN/AP c R	7	0	7
IN/IN	21	9	30
Total	129	17	146

Tabela 3.2.3: Resultado ao final do ano letivo do 8°Ano do Ensino Fundamental

Resultado na AD	Aprovado	Reprovado	Total
AP/AP	2	0	2
AP/AP c R	9	0	9
AP/IN	22	6	28
AP c R/AP	1	0	1
AP c R/AP c R	14	1	15
AP c R/IN	21	11	32
IN/AP	0	0	0
IN/AP c R	4	1	5
IN/IN	15	22	37
Total	88	41	129

Tabela 3.2.4: Resultado ao final do ano letivo do 9°Ano do Ensino Fundamental

Resultado na AD	Aprovado	Reprovado	Total
AP/AP	7	0	7
AP/AP c R	4	0	4
AP/IN	14	3	17
AP c R/AP	3	1	4
AP c R/AP c R	11	0	11
AP c R/IN	14	8	22
IN/AP	1	0	1
IN/AP c R	2	0	2
IN/IN	11	7	18
Total	67	19	86

Tabela 3.2.5: Resultado ao final do ano letivo do 1°Ano

Resultado na AD	Aprovado	Reprovado	Total
AP/AP	2	1	3
AP/AP c R	8	0	8
AP/IN	6	2	8
AP c R/AP	7	0	7
AP c R/AP c R	2	1	3
AP c R/IN	8	17	25
IN/AP	1	0	1
IN/AP c R	1	2	3
IN/IN	2	13	15
Total	37	36	73

Tabela 3.2.6: Resultado ao final do ano letivo do 2°Ano do Ensino Médio

Resultado na AD	Aprovado	Reprovado	Total
AP/AP	10	0	10
AP/AP c R	4	1	5
AP/IN	1	3	4
AP c R/AP	13	0	13
AP c R/AP c R	8	3	11
AP c R/IN	6	7	13
IN/AP	0	0	0
IN/AP c R	3	0	3
IN/IN	2	10	12
Total	47	24	71

Tabela 3.2.7: Resultado ao final do ano letivo do 3°Ano do Ensino Médio

Resultado na AD	Aprovado	Reprovado	Total
AP/AP	7	0	7
AP/AP c R	3	1	4
AP/IN	4	1	5
AP c R/AP	1	0	1
AP c R/AP c R	3	0	3
AP c R/IN	1	2	3
IN/AP	0	0	0
IN/AP c R	0	0	0
IN/IN	1	1	2
Total	20	5	25

Capítulo 4

Construção dos intervalos de confiança *bootstrap* através do R

Com base na teoria do método de reamostragem *bootstrap* e utilizando comandos do programa R, este desenvolvido no apêndice A, e aquele no segundo capítulo desse trabalho, construímos três diferentes intervalos de confiança *bootstrap* chamados de *bootstrap* padrão, percentis *bootstrap* e *bias-corrected and acelerated* que tem como abreviação padrão BC_a.

Em todos os casos o que denominamos parâmetro de interesse θ foi a média do percentual de alunos aprovados, fizemos 10000 replicações *bootstrap* e construímos o intervalo de confiança com probabilidade de cobertura igual a 95%

Para facilitar nossos cálculos, atribuímos o número 0 para o aluno que foi reprovado e 1 para o aluno aprovado, assim, por exemplo, o conjunto de dez dados 0, 0, 1, 0, 1, 0, 1, 0, 0, 1 , têm quatro aprovados e seis reprovados, note que, a média desse conjunto de dez dados é 0,4 que representa a porcentagem dos alunos aprovados, nosso parâmetro de interesse.

Escolhemos, para construirmos os intervalos de confiança utilizando o R de forma detalhada, a amostra relativa aos alunos do 6°Ano do Ensino Fundamental que foram considerados inaptos em Português e em Matemática, a qual pode ser vista na tabela 3.2.1. Ali,

podemos verificar que foram 42 alunos aprovados e 26 reprovados.

4.1 Construção do intervalo de confiança *bootstrap* padrão

Construímos o seguinte programa:

```
#intervalo de confiança bootstrap padrão
# amostras bootstrap
medias<-numeric(0)
dp<-numeric(0)
for(i in 1:10000) {
sextoinin<-c(rep(1,42),rep(0,26))
x<-sample(sextoinin,68, replace=TRUE)
medias[i]<-mean(x)
dp[i]<-sd(x)
}
mediaoriginal<-mean(sextoinin)
dpgeral<-sd(medias)
hist(medias, main="Intervalo de confiança Bootstrap Padrão", xlab=
"Médias bootstrap", ylab="frequência relativa", density=15, freq=FALSE)

curve(dnorm(x,mean=mediaoriginal,sd=dpgeral),col=2,lwd=2,add=TRUE)
#intervalo de confiança
sup<- mediaoriginal + dpgeral* qnorm(1-0.025)
inf<- mediaoriginal + dpgeral* qnorm(0.025)
sup
[1] 0.7329991
```

```
inf
[1] 0.502295
mediaoriginal
[1] 0.6176471
dpgeral
[1] 0.05885415
```

Posteriormente colocamos mais detalhes no gráfico.

```
legend('topright', legend = c( expression(mu= =0.618), expression
(sigma= =0.059)), text.col=c(2,2), cex=1.)
x<-rep(inf,9)
y<-0:8
lines(x,y,col= 'red',lty=2,lwd=2)
text(0.42,5," 0.502295", col= 'blue')
x<-rep(sup,9)
y<-0:8
lines(x,y,col= 'red',lty=2,lwd=2)
text(0.8,5," 0.7329991", col= ' blue ')
```

O resultado final desse programa pode ser visto na figura a seguir.

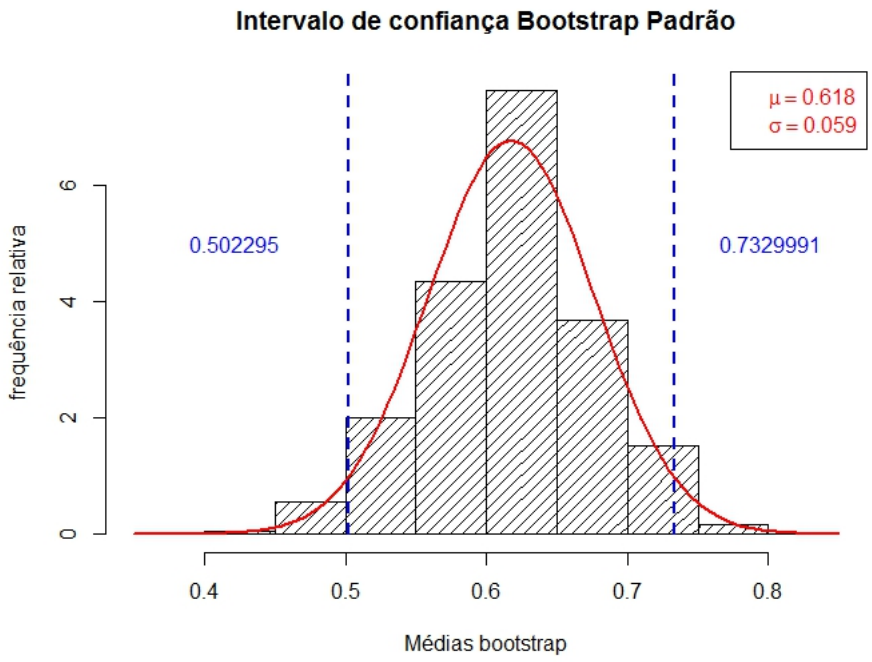

Figura 4.1.1: Intervalo de confiança *bootstrap* padrão

A partir desse resultado, podemos dizer que um aluno, considerado IN/IN no 6°Ano do Ensino Fundamental, tem chance de ser aprovado de $61,76\%$, com uma margem de erro de $\pm 11,54\%$.

Observação 2. O cálculo da margem de erro é feito do seguinte modo:$0,6176471 - 0,502295 = 0,1153521$, isto é, $11,54\%$.

4.2 Construção do intervalo de confiança baseado nos percentis *bootstrap*

Construímos o seguinte programa:

```
#Intervalo de confiança baseado nos percentis bootstrap

# amostras bootstrap
```

```
medias<-numeric(0)
dp<-numeric(0)
for(i in 1:10000) {
sextoinin<-c(rep(1,42),rep(0,26))
x<-sample(sextoinin,68, replace=TRUE)
medias[i]<-mean(x)
dp[i]<-sd(x)
}
hist(medias, main="Intervalo de confiança baseado nos percentis
bootstrap",xlab="Médias bootstrap",ylab="frequência relativa",
density=15, freq=FALSE)
#Intervalo de confiança baseado nos percentis bootstrap
mediasordenadas<- sort(medias)
inf <- mediasordenadas [10000*0.05]
sup<- mediasordenadas [10000*(1-0.05)]
sup
[1] 0.7352941
inf
[1] 0.5
```

Posteriormente colocamos mais detalhes no gráfico.

```
x<-rep(inf,9)
y<-0:8
lines(x,y,col= 'blue',lty=2,lwd=2)
text(0.45,7," 0.5", col= 'blue')
x<-rep(sup,9)
y<-0:8
lines(x,y,col= 'blue',lty=2,lwd=2)
```

```
text(0.77,7," 0.7352941", col= 'blue')
```

O resultado final desse programa pode ser visto na figura a seguir.

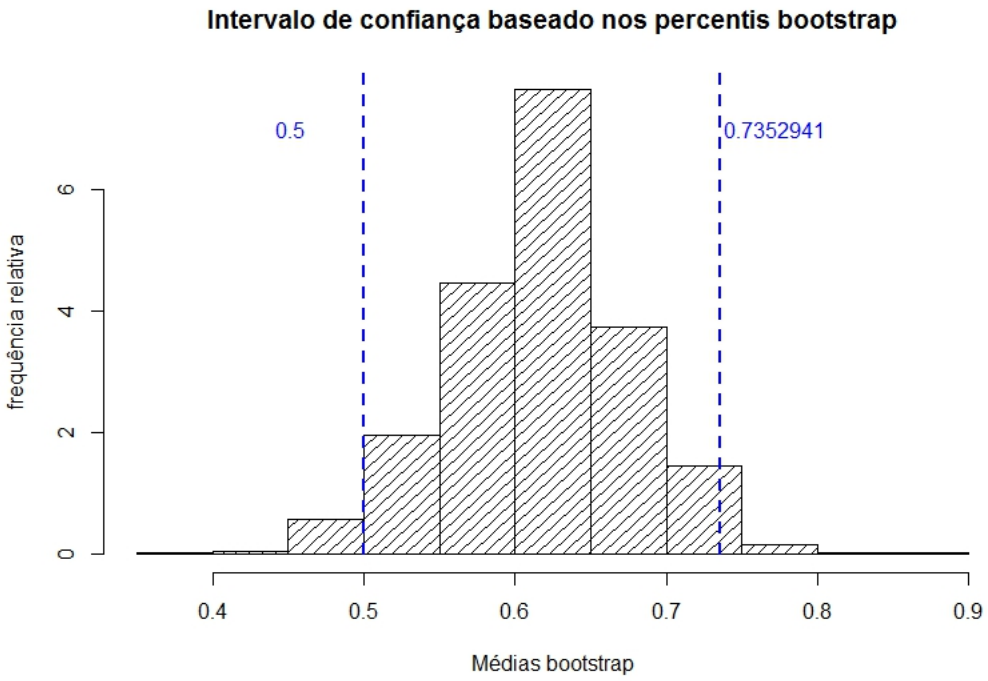

Figura 4.2.1: Intervalo de confiança baseado nos percentis *bootstrap*

A partir desse resultado, podemos dizer que um aluno, considerado IN/IN no 6°Ano do Ensino Fundamental, tem chance de ser aprovado de $\frac{0,5+0,7352941}{2} = 0,61764705$, ou seja, $61,76\%$ com uma margem de erro de $\pm 11,76\%$.

Observação 3. O cálculo da margem de erro é feito do seguinte modo:$0,61764705 - 0,5 = 0,11764705$, isto é, $11,76\%$.

4.3 Construção do Intervalo de confiança percentis BC_a

Construímos o seguinte programa:

```
#Intervalo de confiança BCa
```

```
# amostras bootstrap
medias<-numeric(0)
dp<-numeric(0)
for(i in 1:10000) {
sextoinin<-c(rep(1,42),rep(0,26))
x<-sample(sextoinin,68, replace=TRUE)
medias[i]<-mean(x)
dp[i]<-sd(x)
}
mediaoriginal <-mean(sextoinin)
mediageral <-mean(medias)
hist(medias, main=" Intervalo de confiança BCa ",xlab="Médias
bootstrap",ylab="frequência relativa", density=15, freq=FALSE)
#Intervalo de confiança BCa
mediasordenadas<- sort(medias)
quantidademenor<-sum(ifelse(medias< mediaoriginal,1,0))
zo<-qnorm(quantidademenor/10000)
n<-numeric(0) d<-numeric(0)
for(i in 1:10000) {
n[i]<-( mediageral - medias[i])^3
d[i]<-( mediageral - medias[i])^2
}
numerador<-sum(n)
denominador<-sum(d)
a<- numerador/(6*( denominador)^(3/2))
alfaum<-pnorm(zo+(zo+qnorm(0.025))/(1-a*(zo+qnorm(0.025))))
alfadois<-pnorm(zo+(zo+qnorm(1-0.025))/(1-a*(zo+qnorm(1-0.025))))
```

```
inf <- mediasordenadas [10000* alfaum]

sup<- mediasordenadas [10000* alfadois]

sup

[1] 0.7205882

inf

[1] 0.4852941
```

Posteriormente colocamos mais detalhes no gráfico.

```
x<-rep(inf,9)

y<-0:8

lines(x,y,col= 'blue',lty=2,lwd=2)

x<-rep(sup,9)

y<-0:8

lines(x,y,col= 'blue',lty=2,lwd=2)

text(0.45,7," 0.4852941", col= 'blue') #inf

text(0.77,7," 0.7205882", col= 'blue') #sup
```

O resultado final desse programa pode ser visto na figura a seguir.

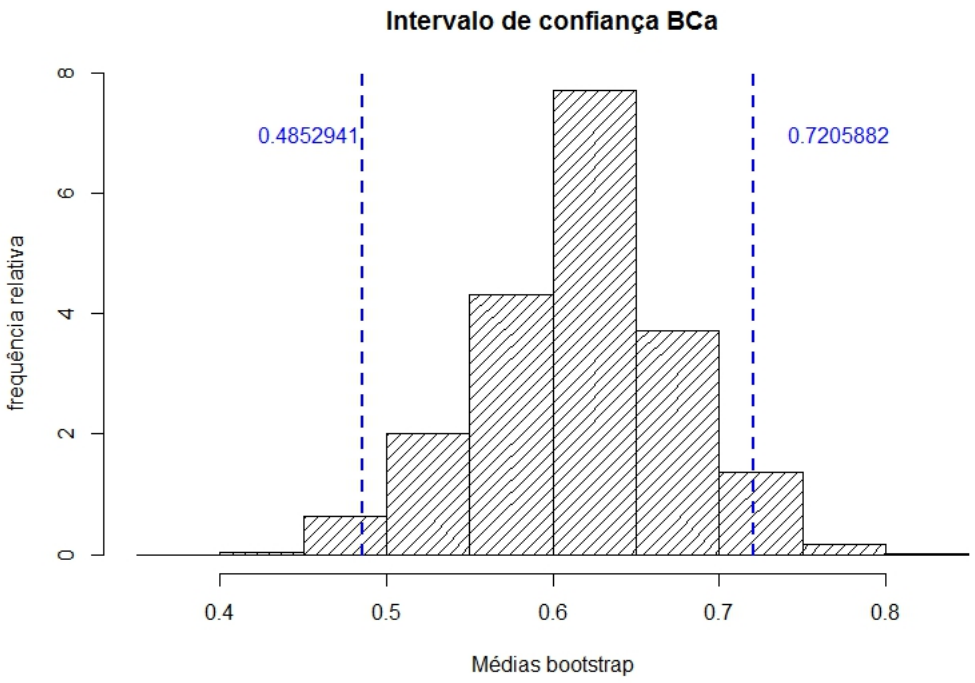

Figura 4.3.1: Intervalo de confiança percentis BC_a

A partir desse resultado, podemos dizer que um aluno, considerado IN/IN no 6°Ano do Ensino Fundamental, tem chance de ser aprovado de $\frac{0,4852941+0,7205882}{2} = 0,60294115$, ou seja, $60,29\%$ com uma margem de erro de $\pm 11,76\%$.

Observação 4. O cálculo da margem de erro é feito do seguinte modo:$0,60294115 - 0,4852941 = 0,11764705$, isto é, $11,76\%$.

4.4 Resultados dos demais dados

Aqui nessa seção, apresentaremos de forma mais resumida os intervalos de confiança construídos, sem detalharmos os programas utilizados, pois todos foram construídos de modo análogo à seção anterior.

Nos casos em que amostra apresentou 100% de aprovação ou reprovação não apresentaremos os intervalos de confiança, pois, quando calculamos, o limite inferior ficou

igual ao superior que por sua vez era igual a 1, o que é equivale dizer 100%, exceto nos percentis BC_a, que não foi possível de calculá-lo pois o denominador era igual a 0.

Ressalta-se que os valores encontrados se referem aos intervalos *bootstrap* a 95% confiança.

Inicialmente apresentaremos os resultados encontrados em cada ano escolar sem considerarmos os resultados na AD.

Tabela 4.4.1: Estimativa da média de aprovados a 95% de confiança referente ao ano escolar sem considerarmos os resultados na AD

Estimativa da média de aprovados a 95% de confiança via intervalos de confiança bootstrap						
Ano escolar	bootstrap padrão		percentis bootstrap		percentis BC_a	
	Média	Margem de erro	Média	Margem de erro	Média	Margem de erro
6°Ano	0,85324235	0,04022505	0,85153585	0,03924915	0,8464164	0,0409556
7°Ano	0,88356165	0,05146315	0,88013695	0,05136985	0,87328765	0,05136985
8°Ano	0,68217055	0,08018545	0,68217055	0,07751935	0,67054265	0,08139535
9°Ano	0,77906975	0,08806785	0,7732558	0,0872093	0,7616279	0,0872093
1°Ano	0,5068493	0,1167266	0,50684935	0,10958905	0,49315065	0,10958905
2°Ano	0,66197185	0,10955055	0,66197185	0,11267605	0,64788735	0,11267605
3°Ano	0,8	0,1564101	0,8	0,16	0,76	0,16

A seguir apresentaremos os resultados encontrados em cada ano escolar de acordo com os resultados na AD.

Tabela 4.4.2: Resultados referentes ao 6°Ano do Ensino Fundamental

Estimativa da média de aprovados a 95% de confiança via intervalos de confiança bootstrap						
Resultado na AD	bootstrap padrão		percentis bootstrap		percentis BC_a	
	Média	Margem de erro	Média	Margem de erro	Média	Margem de erro
AP/AP c R	0,94159565	0,05840435	0,94	0,06	0,88	0,12
AP c R/AP c R	0,92982455	0,06630955	0,9210526	0,0614035	0,89	0,07
AP c R/IN	0,80555555	0,12905065	0,7916667	0,125	0,76	0,13
IN/AP c R	0,8	0,1569264	0,8	0,16	0,76	0,16
IN/IN	0,61764705	0,11535205	0,61764705	0,11764705	0,60	0,12

O resultado esperado ao final do ano letivo para os alunos desse ano escolar com base nos resultados obtidos na AD, tabela 4.4.2, evidenciam uma porcentagem mínima de aprovação de 48%, no caso de IN/IN pelo método percentis BC_a, e a porcentagem chega a 100% de aprovação no caso de AP/AP em todos os métodos. A menor margem de erro ocorre no caso dos AP c R/AP c R que é no máximo 7% para mais ou para menos .

Tabela 4.4.3: Resultados referentes ao 7°Ano do Ensino Fundamental

Estimativa da média de aprovados a 95% de confiança via intervalos de confiança bootstrap						
Resultado na AD	bootstrap padrão		percentis bootstrap		percentis BC_a	
	Média	Margem de erro	Média	Margem de erro	Média	Margem de erro
AP/IN	0,88073195	0,11926805	0,8653846	0,1346154	0,8269231	0,1346154
AP c R/AP c R	0,9035578	0,0964422	0,9	0,1	0,83333335	0,16666665
AP c R/IN	0,87688135	0,12311865	0,88	0,12	0,82	0,14
IN/IN	0,7	0,1636508	0,7	0,1666667	0,66666665	0,16666665

O resultado esperado para esse alunos, tabela 4.4.3, denota novamente uma porcentagem mínima de aprovação de $66,67\%$, no caso de IN/IN, com margem de erro $\pm 16,67\%$ pelo método percentis BC_a, e a porcentagem chega a 100% de aprovação no caso de AP/AP, AP/AP c R, AP c C/AP, IN/AP c C. Cabe citar que no caso IN/AP tivemos apenas um aluno em todo o período em questão, conforme tabela 3.2.2, o que prejudica, ou até mesmo impossibilita uma análise.

Tabela 4.4.4: Resultados referentes ao 8°Ano do Ensino Fundamental

Estimativa da média de aprovados a 95% de confiança via intervalos de confiança bootstrap						
Resultado na AD	bootstrap padrão		percentis bootstrap		percentis BC_a	
	Média	Margem de erro	Média	Margem de erro	Média	Margem de erro
AP/IN	0,78571425	0,15314285	0,76785715	0,16071425	0,73214285	0,16071425
AP c R/AP c R	0,90251935	0,09748065	0,9	0,1	0,83333335	0,16666665
AP c R/IN	0,65625	0,1646709	0,65625	0,15625	0,609375	0,171875
IN/AP c R	0,72539715	0,27460285	0,7	0,3	0,6	0,4
IN/IN	0,4054054	0,1583627	0,4054054	0,1621622	0,37837835	0,1621622

Em relação à porcentagem mínima de aprovação esperada tem-se apenas $21,62\%$, no caso de IN/IN, pelo método percentis BC_a, e varia de 40% a 100% nos casos de IN/AP c R, pelo método percentis *bootstrap*. Note que a margem de erro é muito grande, fato que se deve ao número muito pequeno da amostra. Percebe-se também que a porcentagem de aprovação é de até 100% para os alunos considerados AP c R/AP c R, isto em todos os intervalos *bootstrap* construídos.

Tabela 4.4.5: Resultados referentes ao 9°Ano do Ensino Fundamental

Estimativa da média de aprovados a 95% de confiança via intervalos de confiança bootstrap						
Resultado na AD	bootstrap padrão		percentis bootstrap		percentis BC_a	
	Média	Margem de erro	Média	Margem de erro	Média	Margem de erro
AP/IN	0,82233645	0,17766355	0,8235294	0,1764706	0,73529415	0,20588235
AP c R/AP	0,66020995	0,33979005	0,625	0,375	0,5	0,5
AP c R/IN	0,63636365	0,20172995	0,63636365	0,18181815	0,5909091	0,1818182
IN/IN	0,6111111	0,2271364	0,6111111	0,2222222	0,55555555	0,22222225

Notamos que a aprovação esperada, conforme a tabela 4.4.5, no caso dos AP c R/AP, pelo método percentis BC_a, é de 50% com margem de erro $\pm 50\%$, isto é, varia de 0% a 100%, isto se deve ao número muito pequeno da amostra, o que prejudica uma análise mais pormenorizada. Com relação aos demais casos, verificamos que a menor aprovação esperada é de $55,56\%$ no caso dos IN/IN pelo método percentis BC_a, e a maior é de $82,35\%$,

com margem de erro $\pm 17,65\%$, isto por meio do método percentis *bootstrap*, em relação aos considerados AP/IN.

Tabela 4.4.6: Resultados referentes ao 1°Ano do Ensino Médio

Estimativa da média de aprovados a 95% de confiança via intervalos de confiança bootstrap						
Resultado na AD	bootstrap padrão		percentis bootstrap		percentis BC_a	
	Média	Margem de erro	Média	Margem de erro	Média	Margem de erro
AP/AP	0,56773675	0,43226325	0,5	0,5	0,5	0,5
AP/IN	0,72631195	0,27368805	0,75	0,25	0,5625	0,3125
AP c R/AP c R	0,56372925	0,43627075	0,5	0,5	0,5	0,5
AP c R/IN	0,32	0,1825786	0,34	0,18	0,3	0,18
IN/AP c R	0,42932235	0,42932235	0,5	0,5	0,33333335	0,33333335
IN/IN	0,1530255	0,1530255	0,16666665	0,16666665	0,13333335	0,13333335

Por meio da tabela 4.4.6, podemos notar que para a maior parte dos alunos, os considerados AP c R/IN, a expectativa de aprovação é de 32% com margem de erro $\pm 18,26\%$ através do método *bootstrap* padrão. Com excessão aos casos que a amostra era de 100% de aprovação, a saber, AP/AP c R, AP c R/AP, IN/AP, a maior expectativa de aprovação ocorre no caso dos AP/IN que também foi de até 100% em todos intervalos construídos. Destacamos que esse ano escolar foi o único em que ocorreu um caso no qual um aluno considerado AP/AP reprovou ao término do ano letivo, como a amostra era muito pequena margem de erro chegou a $\pm 50\%$. Ainda, a expectativa máxima de aprovação é de apenas $16,67\%$ com margem de erro $\pm 16,67\%$, isto por meio do método percentis *bootstrap*, no caso dos alunos IN/IN, sendo a menor dentre todos os anos escolares.

Tabela 4.4.7: Resultados referentes ao 2°Ano do Ensino Médio

Estimativa da média de aprovados a 95% de confiança via intervalos de confiança bootstrap						
Resultado na AD	bootstrap padrão		percentis bootstrap		percentis BC_a	
	Média	Margem de erro	Média	Margem de erro	Média	Margem de erro
AP/AP c R	0,72350805	0,27649195	0,7	0,3	0,6	0,4
AP/IN	0,33541715	0,33541715	0,375	0,375	0,25	0,25
AP c R/AP c R	0,7272727	0,2656412	0,72727275	0,27272725	0,63636365	0,27272725
AP c R/IN	0,46153845	0,27035015	0,5	0,2692308	0,42307695	0,26923075
IN/IN	0,18748595	0,18748595	0,20833335	0,20833335	0,16666665	0,16666665

A aprovação esperada, conforme tabela 4.4.7, é de $72,35\%$ com margem de erro $\pm 27,65\%$ através do método *bootstrap* padrão para o caso dos AP/AP c R. Em relação aos AP c R/AP ocorreu em 100% dos casos. A menor expectativa de aprovação se dá com os alunos IN/IN na qual é de apenas $16,67\%$ com margem de erro $\pm 16,67\%$, isto por meio do método percentis BC_a.

Tabela 4.4.8: Resultados referentes ao 3°Ano do Ensino Médio

Estimativa da média de aprovados a 95% de confiança via intervalos de confiança bootstrap						
Resultado na AD	bootstrap padrão		percentis bootstrap		percentis BC_a	
	Média	Margem de erro	Média	Margem de erro	Média	Margem de erro
AP/AP c R	0,662514	0,337486	0,625	0,375	0,5	0,5
AP/IN	0,72681435	0,27318565	0,7	0,3	0,6	0,4
AP c R/IN	0,56925035	0,43074965	0,5	0,5	0,5	0,5
IN/IN	0,5	0,5	0,5	0,5	0,25	0,25

A aprovação esperada é de 100% nos considerados AP/AP, pois a amostra era de 100% de aprovado. De modo geral as margens de erro eram grandes, isto devido ao pequeno tamanho das amostras. As menores margens de erro ocorrem no caso dos AP/IN, dentre os quais destacamos a margem de erro $\pm 27,32\%$, cuja aprovação esperada é de $72,68\%$, valores estes referentes ao método *bootstrap* padrão, conforme tabela 4.4.8.

4.5 Respostas às questões de estudo

A fim de responder às questões de estudo, que são o objeto da pesquisa desse trabalho, padronizamos como referência, a média do percentual de aprovados de cada um dos intervalos de confiança *bootstrap* contruídos, e responderemos de acordo com cada tipo de intervalo contruído, isto é, poderemos ter até três respostas distintas cada uma fundamentada num intervalo de confiança. Ressalta-se que esses intervalos são de 95% confiança.

Na primeira questão de estudo desejamos encontrar em qual ano, do fundamental ou médio, um aluno mesmo considerado IN/IN na AD tem maior probabilidade de ser aprovado. A partir dos resultados verificamos, por meio dos três intervalos de confiança contruídos, que isto ocorre no 7° ano Ensino Fundamental, cuja expectativa de aprovação é de 70%, por meio do intervalo de confiança *bootstrap* padrão e percentis *bootstrap*, sendo que no primeiro a margem de erro é de $\pm 16,37\%$ e no por meio do segundo a margem de erro é de $\pm 16,67\%$. Em relação ao intervalo de confiança percentis BC_a a expectativa de aprovação é de $66,67\%$, cuja margem de erro é de $\pm 16,67\%$.

Na segunda questão de estudo desejamos encontrar o oposto da primeira, ou seja, mesmo considerado AP/AP na AD, em qual ano tem maior probabilidade de ser reprovado. Após analisar os resultados notamos que isto ocorre no 1° ano do Ensino Médio, aliás, foi o único ano que ocorreu um caso de reprovação dentre os alunos que foram considerados AP/AP, em todos os demais anos escolares os alunos obtiveram 100% de aprovação. Nesse ano escolar destacamos que todos os intervalos de confiança construídos tiveram uma margem de erro muito grande, fato que se deve ao pequeno tamanho da amostra.

Já na terceira questão de estudo almejamos de modo geral, de acordo com o resultado na AD, em cada ano escolar, qual a probabilidade de aprovação. Os dados inerentes a essa resposta podem ser vistos na seção 4.4 desse trabalho. Ali, notamos, conforme a tabela 4.4.2, que os resultados do 6° Ano do Ensino Fundamental cuja menor chance de aprovação é de 60% no caso IN/IN com margem de erro de $\pm 12\%$ através do método percentis BC_a. Com relação ao 7° Ano do Ensino Fundamental, tabela 4.4.3, também a menor expectativa

de aprovação é no caso dos IN/IN, com $66,67\%$,com margem de erro de $\pm 16,67\%$ através do método percentis BC_a. Os resultados do 8° Ano do Ensino Fundamental, os quais podem ser vistos na tabela 4.4.4, mostram que a expectativa mínima de aprovação é de $37,84\%$ com margem de erro de $\pm 16,22\%$ através do método percentis BC_a que ocorre, também, com os alunos considerados IN/IN. Em relação ao 9° Ano do Ensino Fundamental, tabela 4.4.5, a menor probabilidade de aprovação é de 50% com margem de erro de $\pm 50\%$, porém, ao conttrário dos anos anteriores, ocorre no caso dos AP c R/AP, novamente através do método percentis BC_a. Destacamos, em relação ao 1° Ano do Ensino Médio, cujos resultados estão exibidos na tabela 4.4.6, que a expectativa máxima de aprovação, no caso dos IN/IN, é de $16,67\%$ com margem de erro de $\pm 16,67\%$, isto através do método percentis *bootstrap*. Em relação ao 2° Ano do Ensino Médio, conforme tabela 4.4.7, novamente é baixa a expectativa de aprovação do alunos considerados IN/IN, que é de no máximo $20,83\%$ com margem de erro de $\pm 20,83\%$, isto através do método percentis *bootstrap*. Por fim, os resultados referentes ao 3° Ano do Ensino Médio, conforme tabela 4.4.8, em relação ao caso dos alunos considerados IN/IN a expectativa de aprovação é de 50% a margem de erro é de $\pm 50\%$, isto através dos métodos percentis *bootstrap* e *bootstrap* padrão.

Finalmente, na quarta e última questão de estudo desejamos conhecer de modo geral a probabilidade de aprovação de um aluno novo em cada ano escolar, sem considerarmos o resultado na AD. A tabela 4.4.1 exibe todos esses resultados. Ali, destacamos que a menor expectativa de aprovação ocorre no 1° ano do Ensino Médio, com apenas $49,32\%$ com margem de erro de $\pm 10,96\%$ através do método percentis BC_a , e o ano que tem a maior expectativa de aprovação é o 7° ano do Ensino Fundamental, essa expectativa é de $88,36$ com margem de erro de $\pm 5,15\%$ através do método *bootstrap* padrão. A maior expectativa de aprovação no Ensino Médio ocorre no 3° ano, a qual é de 80% a margem de erro é de $\pm 15,64\%$, isto através do método *bootstrap* padrão e, por fim, em relação ao Ensino Fundamental, a menor expectativa de aprovação ocorre no 8° ano, com $67,05\%$ com margem de erro de $\pm 8,14\%$ através do método percentis BC_a.

Capítulo 5

Conclusão

Esse trabalho teve como foco principal responder questões inerentes à realidade do CMCG, propiciando dados estatísticos que subsidiem estratégias para selecionar qual público alvo deve receber apoio pedagógico, ou seja, identificar qual grupo de alunos apresenta uma maior probabilidade de reprovação, baseados na primeira avaliação que estes realizam ao ingressar no SCMB, a Avaliação Diagnóstica.

Ao realizarmos a coleta dos dados referentes à AD realizada no CMCG, notamos que a quantidade desses dados era pequena. Então, após uma análise no conteúdo básico de Estatística Descritiva e também de Estatística Inferencial, verificamos que, devido a esse pequeno tamanho da amostra, era necessário utilizarmos uma técnica de reamostragem, para tanto, escolhemos o método *bootstrap*.

O método *bootstrap* é particularmente apropriado quando o cálculo de estimadores por métodos analíticos é difícil ou não se conhece a função de distribuição de probabilidades, pois não necessita de muitas suposições para a estimação dos parâmetros das distribuições de interesse.

Entretanto, ainda que esse método de reamostragem seja muito útil, necessitamos de um tamanho razoável da amostra para encontrarmos intervalos de confiança mais estreitos, obtendo, assim, uma maior precisão nos cálculos. Em diversos casos, cuja amostra era muito

pequena, encontramos intervalos de confiança com limite inferior igual a zero e o superior, a um, ou seja, não poderiamos concluir nada sobre esses casos, pois a margem de erro foi de $\pm 50\%$.

As respostas às questões norteadoras, através dos intervalos *bootstrap* com 95% confiança, destacam que o 1° ano do Ensino Médio é o ano escolar que os alunos novos têm mais dificuldade de cursá-lo e, como aqui mostramos, é o único ano no qual um aluno mesmo considerado apto em Português e em Matemática não podemos inferir sobre sua expectativa de aprovação ao final do ano letivo, pois a margem de erro é de $\pm 50\%$, isto através dos métodos percentis *bootstrap* e percentis BC_a. Outro fator que observamos é que, em relação ao Ensino Fundamental, é no 8° ano que os alunos ingressantes têm mais chances de reprovar.

Através da análise dos dados, outro fato interessante que cabe ressaltar, é que foi no 7° ano Ensino Fundamental que os alunos ingressantes, mesmo inaptos, têm mais chances de reverterem esse quadro e alcancerem a aprovação, isto através de todos os intervalos de confiança construídos, isto é, *bootstrap* padrão, percentis *bootstrap* e percentis BC_a.

É evidente que apenas o resultado obtido na AD não irá condicionar o desempenho durante o ano letivo do aluno, certamente existem outros fatores, os quais poderão ser pesquisados em trabalhos futuros. Também é possível, a partir dos resultados obtidos nessa pesquisa, criar estratégias de apoio pedagógico aos grupos de alunos com maior probabilidade de reprovação e, posteriormente, analisar a eficácia dessa estratégia criada, comparando com os resultados aqui obtidos.

Apêndice A

O programa R

O programa R é um software livre para computação estatística e construção de gráficos que pode ser baixado e distribuído gratuitamente. Foi criado originalmente por Ross Ihaka e por Robert Gentleman, daí o nome R que provêm em parte das iniciais dos criadores, no departamento de Estatística da universidade de Auckland, Nova Zelândia, e foi desenvolvido por um esforço colaborativo de pessoas em vários locais do mundo.

Nesse apêndice vamos resumir os camandos utilizados no programa R visando a construção de intervalos de confiança *bootstrap*, maiores detalhes sobre o programa R pode ser visto em [14] ou na própria ajuda interna do programa e ainda utilizamos [15] como referência para construção de gráficos.

A.1 Comandos utilizados nesse trabalho

Antes de qualquer comando sempre aparecerá o sinal de maior > siginificando que o R está pronto para receber comandos. Para inserir comentários, os quais não serão reconhecidos como comandos, basta uitilzar o sinal de sustenido (jogo da velha) # .

A.1.1 Operações básicas

Os operadores matemáticos mais básicos do R são: `+` para soma, `-` subtração, `*` multiplicação, `/` divisão e `^` exponenciação.

A.1.2 Vetores com valores numéricos

```
> notas<-c(5,8,7,3,10)
```

O comando `<-` (sinal de menor e sinal de menos) significa dizer "salve os dados a seguir com o nome de notas". A letra `c` significa colocar junto. Entenda como "agrupe os dados entre parênteses dentro do objeto que será criado" neste caso no objeto notas.

Para ver os valores (o conteúdo de um objeto), basta digitar o nome do objeto na linha de comandos.

```
> notas
[1] 5 8 7 3 10 # resultado exibido pelo programa.
```

A.1.3 Algumas funções

A função `sqrt()` é a função para calcular a raiz quadrada.

```
> sqrt(notas)
[1] 2.236068 2.828427 2.645751 1.732051 3.162278 # raiz quadrada de cada
```

uma das notas

A função `sum()` soma todos os valores das notas.

```
> sum(notas)
[1] 33
```

A função `length()` fornece o número de observações `(n)` dentro do objeto.

```
> length(notas)
[1] 5
```

O R têm funções prontas para calcular a média, variância e desvio padrão.

A função `mean()`calcula a média dos valores de um objeto.

```
> mean(notas)
[1] 6.6
```

A função `var()`calcula a variâcia dos valores de um objeto.

```
> var(notas)
[1] 7.3
```

A função `sd()`calcula o desvio padrão dos valores de um objeto.

```
> sd(notas)
[1] 2.701851
```

Note que a raiz quadrada da variância é o desvio padrão.

```
> sqrt(var(notas))
[1] 2.701851
```

A função `sort()` coloca os valores de um objeto em ordem crescente ou em ordem decrescente.

```
> sort(notas) # para colocar em ordem crescente, isto é, construir o rol.
[1] 3 5 7 8 10
> sort(notas, decreasing=TRUE) # para colocar em ordem decrescente
[1] 10 8 7 5 3
```

A função `sample()` é utilizada para realizar amostras aleatórias e funciona do seguinte modo: `sample(x, size=1, replace = FALSE)`, sendo que x é o conjunto de dados do qual as amostras serão retiradas, `size` é o o número de amostras retiradas e em `replace` você indica se a amostra deve ser feita com reposição `(TRUE)` ou sem reposição `(FALSE)`.

```
> sample(notas, 3, replace=TRUE)
[1] 10 5 3 # note que é uma amostra aleatória
```

Para gerar repetições temos a função `rep`.

```
> rep(2,5) # repete o valor 2 cinco vezes
[1] 2 2 2 2 2
```

O R também têm funções que nos fornecem informações, isto é, valores referentes a distribuição normal os quais são implementados por argumentos que combinam letras com o termo `norm`. Vamos ver dois exemplos com a distribuição normal padrão.

```
> qnorm(0.975) # Calcula o valor de Z^α, no caso, Z^0,975.
[1] 1.959964
> pnorm(1.959964) # Calcula o valor de Φ que é a função distribuição acumulada
```

de uma normal padrão. Note que $Z^{\alpha} = \Phi^{-1}$.

```
[1] 0.975
```

A.1.4 Operações com vetores

Caso queira acessar apenas um valor do conjunto de dados use colchetes `[]`. Isto é possível porque o R salva os objetos como vetores, assim, a sequência na qual você incluiu os dados é preservada. Por exemplo, vamos acessar o quarto valor do objeto notas.

```
> notas[4] # Qual o quarto valor de notas?
[1] 3
```

Note que se desejarmos, por exemplo a quarta maior nota basta criar um novo objeto dos dados ordenados e utilizar o comando acima.

```
> notasordenados<-c(sort(notas))
> notasordenados[4]
[1] 8
```

A.2 Gráficos

O R têm diversas funções para construção gráfica como por exemplo a construção de linhas (retas), histogramas, e a curva normal de Gauss. Vejamos exemplos de cada uma dessas curvas:

Para construirmos um histograma precisamos ter um conjunto de dados, por exemplo,

`temperaturas<-c(27, 22, 34, 32, 16, 26, 25, 19, 21, 29, 34, 29, 26, 28, 18, 24, 26, 17, 28, 22, 19, 33, 21, 21, 17, 18, 17, 27, 41, 39, 36)`

`hist(temperaturas)#` Construa o histograma do conjunto de dados denominado temperaturas. Cujo gráfico resultante pode ser visto na figura a seguir.

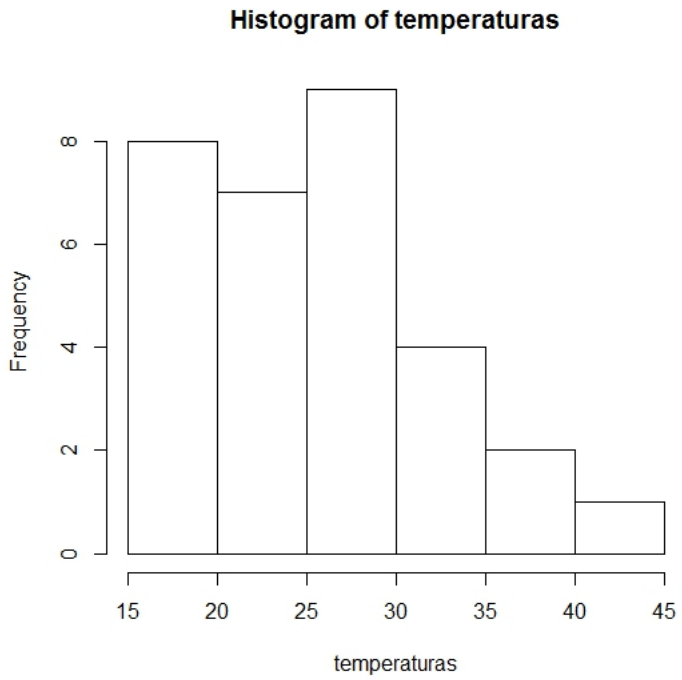

Figura A.2.1: histograma temperaturas

Podemos acrescentar ou alterar informações ao gráfico como título `main="novo título"`, nomenclatura do eixo-x, `xlab="x"`, nomenclatura do eixo-y, `ylab="y"`. Acrescentar linhas de sombreamento, `density=15`. Podemos também construir o histograma através da sua frequência relativa com o comando `freq=FALSE`. Vejamos o resultado na figura a seguir.

`hist(temperaturas, main=" Temperaturas",xlab=" temperaturas", ylab=" frequência relativa", density=15, freq=FALSE)`

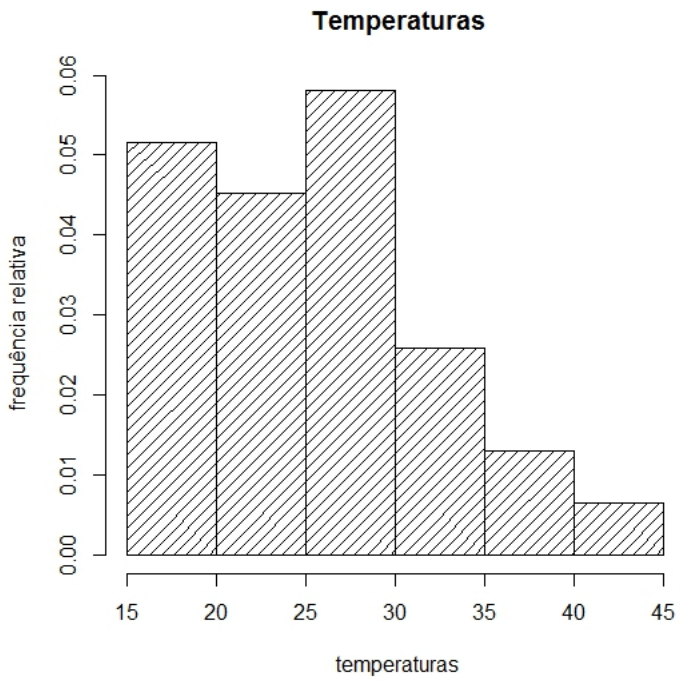

Figura A.2.2: Temperaturas

Podemos traçar a curva da função de densidade da distribuição normal com média, `mean`, e desvio padrão, `sd`, desejados.

`curve(dnorm(x,mean=0,sd=sqrt(1)),lwd=2,col='blue',from=-3,to=3) #` O comando `lwd` altera a espessura da curva, `col`, a cor e `from=,to` a variação dos valores de x.

Para acrescentar legenda e título ao gráfico utilizamos os camandos abaixo descritos:

`legend('topright', legend=c(expression(mu==0), expression(sigma==1)), text.col=c(1,2), cex=1.)#` Os comandos `text.col` e `cex` alteram respectivamente a localização e a espessura da caixa.

`title("Curva Normal")`

O resultado final desses comandos visualizamos na próxima figura.

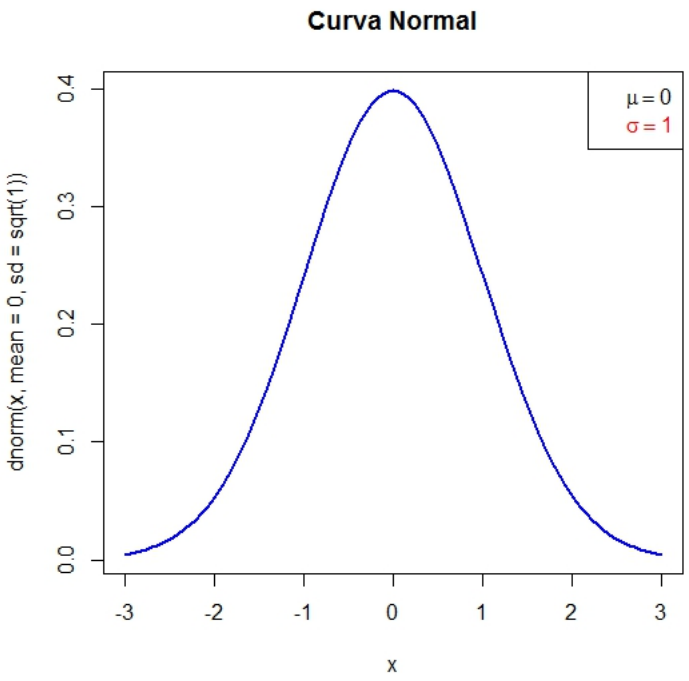

Figura A.2.3: Temperaturas

Podemos também acrescentar uma reta ou linha ao gráfico construído por meio do comando `lines()`.

```
x<-rep(0,2)
y<-0:1
lines(x,y,col= 'red',lty=2,lwd=2) # Cria linha e o comando lty altera a distância do tracejado da linha.
text(0.8,0.1," mediana", col= 'red') # Coloca um texto no gráfico.
```

Vejamos o gráfico resultante da linha com a curva normal.

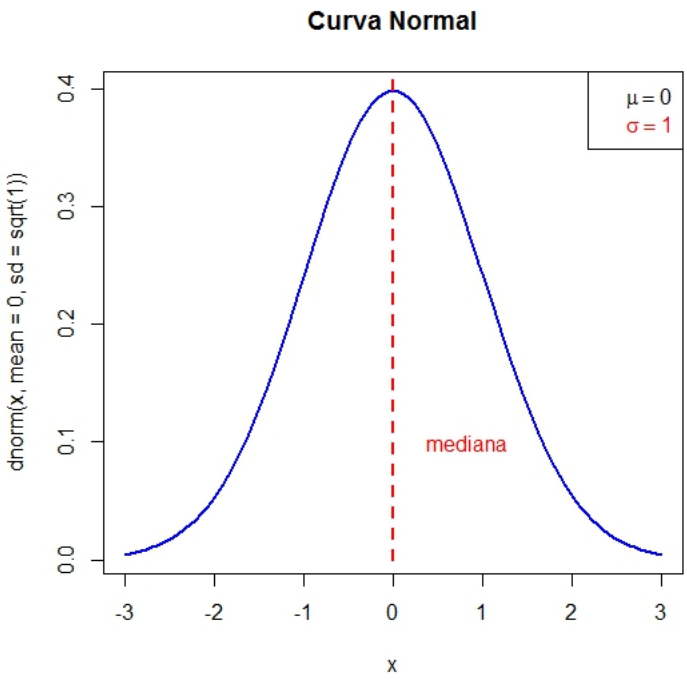

Figura A.2.4: Temperaturas

A.3 Comandos de lógica

O comando `for` é usado para fazer *loopings*, e funciona do seguinte modo: "`for(i in 1:n) {comandos}`" Isso quer dizer que para cada valor i o R vai calcular os comandos que estão entre as chaves `{comandos}`. O "`i in 1:n`" indica que os valores de `i` serão `i = 1` até `i = n`. Vejamos um exemplo:

```
> pot<-numeric(0)
> for(i in 1:4) {
+ pot [i]<-i^i
+ }
> pot
[1] 1 4 27 256
```

Apêndice B

Programa Geral

Aqui nesse apêndice apresentamos o programa geral construído com o intuito de obtermos os intervalos de confiança *bootstrap* padrão, percentis *bootstrap* e BC_a. Analogamente ao programa desenvolvido no quarto capítulo, nosso parâmetro de interesse θ foi a média do percentual de alunos aprovados, fizemos 10000 replicações *bootstrap* e construímos o intervalo de confiança com probabilidade de cobertura igual a 95%.

A seguir exibimos o programa, as únicas alterações necessárias ocorrem na quinta e sexta linhas de acordo com os valores desejados.

```
# PROGRAMA GERAL
# amostras bootstrap
medias<-numeric(0)
dp<-numeric(0)
ap<-(total de aprovados na amostra)
rep<-(total de reprovados na amostra)
totalalunos<- ap+rep
for(i in 1:10000) {
sextoinin<-c(rep(1, ap),rep(0, rep))
x<-sample(sextoinin, totalalunos, replace=TRUE)
```

```
        medias[i]<-mean(x)
        dp[i]<-sd(x)
        }
        mediaoriginal<-mean(sextoinin)
        mediageral <-mean(medias)
        dpgeral<-sd(medias)
        #intervalo de confiança bootstrap padrão
        supbp<- mediaoriginal + dpgeral* qnorm(1-0.025)
        infbp<- mediaoriginal + dpgeral* qnorm(0.025)
        #Intervalo de confiança baseado nos percentis bootstrap
        mediasordenadas<- sort(medias)
        infpb <- mediasordenadas [10000*0.025]
        suppb<- mediasordenadas [10000*(1-0.025)]
        #Intervalo de confiança BCa
        mediasordenadas<- sort(medias)
        quantidademenor<-sum(ifelse(medias< mediaoriginal,1,0))
        zo<-qnorm(quantidademenor/10000)
        n<-numeric(0)
        d<-numeric(0)
        for(i in 1:10000) { n[i]<-( mediageral - medias[i])^3 d[i]<-( mediageral
- medias[i])^2 }
        numerador<-sum(n)
        denominador<-sum(d)
        a<- numerador/(6*( denominador)^(3/2))
        alfaum<-pnorm(zo+(zo+qnorm(0.025))/(1-a*(zo+qnorm(0.025))))
        alfadois<-pnorm(zo+(zo+qnorm(1-0.025))/(1-a*(zo+qnorm(1-0.025))))
        infbca <- mediasordenadas [10000* alfaum]
```

```
supbca<- mediasordenadas [10000* alfadois]


mediabp<-(( supbp + infpb)/2)

errobp<-( supbp - mediabp)

mediapb<-(( suppb + infpb)/2)

errobp<-( suppb - mediapb)

mediabca<-(( supbca + infbca)/2)

errobca<-( supbca - mediabca)

mediabp

errobp

mediapb

errobp

mediabca

errobca
```

Referências Bibliográficas

[1] BRASIL. **Normas Internas de Avaliação Educacional (NIAE)**. Rio de Janeiro, 2011. 1

[2] Costa Neto, P. L. de O. **Estatística**. 3 ed, São Paulo: Edgard Blucher, 2002. 2, 4

[3] PAIVA, Manoel. **Matemática 2**. 1ª ed. São Paulo: Ed. Moderna: 2009 4

[4] IEZZI, G et al, **Fundamentos da Matemática Elementar**, volume 11, São Paulo: Atual editora, 2004. 4

[5] Fonseca, J. S. da, Martins, G. de A. **Curso de estatística**. 6 ed, São Paulo: Atlas, 1996. 320p. 10

[6] Neves, M. **Introdução à Estatístistica e à Probabilidade**. Disponível em: <http://www.isa.utl.pt/dm/estat/estat/seb3.pdf> Acesso em 20/02/2014. 10

[7] Bolfarine, H. **Introdução à Inferência Estatística**. 2 ed, Rio de Janeiro: SBM, 2010. 10

[8] Efron, B. Tibshirani, R. **An Introduction to the Bootstrap**. Chapman and Hall, 1993. Disponível em: <http://staff.ustc.edu.cn/~zwp/teach/Stat-Comp/Efron_Bootstrap_CIs.pdf> Acesso em 20/02/2014. 13, 14

[9] DiCiccio, T. J; Efron, B. **Bootstrap Condence Intervals.** Statistical Science. v.11, n.3, p.189-228. 1996. 13

[10] Cunha, W. J. da; Colosimo, E. A. **Intervalos de confiança bootstrap para modelos de regressão com erros de medida**. Rev. Mat. Estat. São Paulo, v.21, n.2, p.25-41, 2003. 13

[11] Rizzo, A. L. T; Cymrot. R. **Estudo e Aplicações da Técnica Bootstrap**. II Jornada de Iniciação Científica. Universidade Presbiteriana Mackenzie. Disponível em: <http://meusite.mackenzie.com.br/raquelc/ana_lucia.pdf> Acesso em 20/02/2014. 13

[12] Martinez, E. Z.; Louzada-Neto, F. **Estimação intervalar via bootstrap**. Rev. Mat. Estat. São Paulo, v.19, p.217-251, 2001. 13

[13] Kauark, Fabiana; Manhães, F. C.; Medeiros, C. H. **Metodologia da pesquisa : guia prático**. Itabuna: Via Litterarum, 2010. 18

[14] Landeiro, V.L. **Introdução ao uso do programa R**. Disponível em: <http://cran.r-project.org/doc/contrib/Landeiro-Introducao.pdf> Acesso em 20/02/2014. 47

[15] Martins, P.S. **Treinando habilidades de elaboração de gráficos com o software R**. Disponível em: <http://www.professores.uff.br/luciane/images/stories/Arquivos/Rgraficos.pdf> Acesso em 20/02/2014.

47