UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA E ESTATÍSTICA

SELEÇÃO QUANTÍLICA NO PACOTE *TestFraud* PARA DETECÇÃO DE FRAUDES EM TESTES

Robinson Ortega Meza

Orientação: **Prof. Dr. Héliton Ribeiro Tavares**
Coorientação: **Profa. Dra. Maria Regina Madruga Tavares**

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001

Belém
2020

Robinson Ortega Meza

SELEÇÃO QUANTÍLICA NO PACOTE *TestFraud* PARA DETECÇÃO DE FRAUDES EM TESTES

Dissertação apresentada ao Curso de Mestrado em Matemática e Estatística da Universidade Federal do Pará, como pré-requisito para a obtenção do título de Mestre em Estatística.

Orientação: **Prof. Dr. Héliton Ribeiro Tavares**

Coorientação: **Profa. Dra. Maria Regina Madruga Tavares**

Belém

2020

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

M617s Meza, Robinson Ortega
Seleção Quantílica no pacote TestFraud para detecção de fraudes em testes / Robinson Ortega Meza. — 2020.
xv, 70 f. : il. color.

Orientador(a): Prof. Dr. Héliton Ribeiro Tavares
Coorientação: Profª. Dra. Maria Regina Madruga Tavares
Dissertação (Mestrado) - Programa de Pós-Graduação em Matemática e Estatística, Instituto de Ciências Exatas e Naturais, Universidade Federal do Pará, Belém, 2020.

1. Métodos para detecção de fraude em testes. 2. Seleção Quantílica. 3. Avaliação em larga escala. 4. Teoria da Resposta ao Item. I. Título.

CDD 310

Robinson Ortega Meza

SELEÇÃO QUANTÍLICA NO PACOTE *TestFraud* PARA DETECÇÃO DE FRAUDES EM TESTES

Esta Dissertação foi julgada e aprovada para a obtenção do grau de Mestre em Estatística, no Programa de Pós-Graduação em Matemática e Estatística da Universidade Federal do Pará.

Belém, 14 de Fevereiro de 2020

João Marcelo B Protazio

Prof. Dr. João Marcelo Brazão Protázio
(Coordenador do Programa de Pós-Graduação em Matemática e Estatística – UFPA)

Banca Examinadora

Prof. Dr. Héliton Ribeiro Tavares
PPGME/UFPA
Orientador

Prof. Dr. Marcus Pinto da Costa da Rocha
PPGME/UFPA
Examinador Interno

Fabrício Martins da Costa

Prof. Dr. Fabrício Martins da Costa
Depto. Mat., Estat. & Infom./UEPA
Examinador Externo

Aos meus amados e inspiradores pais e irmãos.

Agradecimentos

Agradeço a Deus e seu Espirito Santo, por ter me dado uma vida abençoada usando sua imerecida providência, por tudo sua guia e amor incondicional, com os quais obtive um grande aprendizagem intelectual e espiritual.

Aos meus pais Robinson Ortega e Doris Meza, e meus irmãos Edinson, Fernando, Mercedes, Luisa, Eduardo, Encho e a toda o resto da família por sempre acreditarem em mim e por serem o principal motivo para eu nunca desistir dos meus sonhos.

Uma profunda gratidão a Sandra Salgado minha futura esposa pela paciência, amor, ajuda e compreensão neste árduo caminho. Também, agradeço a Mile Salgado minha irmã especial que sempre acreditou em mim e me apoiou incondicionalmente neste projeto. Quero agradecer especialmente a Beatriz e Fernando Sousa, minha família brasileira, além deles, todos os que me aceitaram aqui no Brasil com amor e que infelizmente não posso citar-lhes, mas saibam que estão em meu coração.

Agradeço a todos professores do PPGME e, particularmente, aos professores Dr. Héliton Tavares, Dra. Regina Tavares, Dr. João Marcelo Pratázio e Valcir João da Cunha Farias que acreditaram em mim e que contribuíram imensamente para meu crescimento acadêmico e profissional, pois, orientando-me com paciência durante o curso. Obrigada pelo voto de confiança, pelo apoio e pelos conselhos que me foram dados, estes foram muito importantes pra mim.

A todos os amigos do PPGME, por fornecerem um ambiente de aprendizagem agradável. Agradeço, em especial, a Paulo Germano, meu irmão brasileiro e meus amigos Miguel Monteiro, Charles De Albuquerque, Rafael Moraes, Paulo Emilio, Thamara Medeiros, Alice Moraes, Lizandra Farias, Fernando Campos, Andrey Nascimento, Armando Paiva, pelos preciosos momentos de compartilhação de conhecimento e experiências de vida. Infelizmente não posso citar todos, mas saibam que estão em meu coração.

Finalmente, gostaria de agradecer à UFPA pelo ensino gratuito de qualidade, ao LAM, ao PPGME e à CAPES, sem os quais essa dissertação dificilmente poderia ter sido realizada e a todos mais que eu não tenha citado nesta lista de agradecimentos, mas que de uma forma ou de outra contribuíram não apenas para a minha dissertação, mas também para eu ser quem eu sou.

"Eu não sou um homem perfeito, mas para ti eu serei minha melhor versão possível"

RψS

"Ainda que a minha mente e o meu corpo enfraqueçam, Deus é a minha força, Ele é tudo o que eu sempre preciso"

O rei David

Resumo

Em exames de larga escala é fundamental que os testes sejam aplicados com confiança, a fim de produzir inferências consistências sobre a proficiência dos examinados. Os testes estatísticos de detecção de fraudes aplicados a esses exames geram maior credibilidade aos mesmos. Por outro lado, essas avaliações envolvem um grande número de candidatos, e a aplicação de alguns métodos estatísticos para detecção de fraudes se tornam enviáveis em tempo hábil. Neste estudo propõe-se uma alternativa para esses testes, consistindo na otimização quantílica do pacote *TestFraud*. Essa seleção quantílica diminui a quantidade de pares analisados com base no limiar de alto escore, pois, a fonte (s) e copiador (c) estão acima dessa limiar. Outra contribuição deste trabalho é a criação da estatística T^* ponderada, que minimiza significativamente a taxa de *falso positivo*. Foram realizados estudos de simulações para avaliar a eficiência do método proposto e obtenção de valores críticos para T^*. O trabalho finaliza com uma aplicação a dados reais do ENEM-2018, avaliando as quatro áreas do exame e selecionando-se o quantil 0,95 e um nível de significância de 0,01 para indicação de fraude.

PALAVRAS-CHAVE: Métodos para detecção de fraude em testes, Seleção Quantílica, Avaliação em larga escala, Teoria da Resposta ao Item.

Abstract

In large-scale examinations it is critical that the tests should be applied with confidence in order to produce consistent inferences about the proficiency of the examiners. Statistical fraud detection tests applied to these exams give them greater credibility. On the other hand, these evaluations involve a large number of applicants, and the application of some statistical methods for fraud detection becomes feasible in a timely manner. This study proposes an alternative to these tests, consisting of the quantile optimization of the TestFraud package. This quantile selection decreases the number of pairs analyzed based on the high score threshold because the source (s) and copier (c) are above this threshold. Another contribution of this work is the creation of the weighted T^* statistic, which significantly minimizes the false positive rate. Simulation studies were carried out to evaluate the efficiency of the proposed method and to obtain critical values for T^*. The work ends with an application to real data from ENEM-2018, evaluating the four areas of the exam and selecting the 0.95 quantile and a significance level of 0.01 for indication of fraud.

KEYWORDS: Methods for test fraud detection, Quantile Selection, Large-scale evaluation, Item Response Theory.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

1.1 Aspectos gerais

A popularização de testes em larga escala impõe a necessidade de cuidados adicionais com possíveis fraudes, particularmente quando há diversas vantagens associadas aos desempenhos nos testes, tal como a conquista de vagas em universidades públicas. Uma ferramenta de extrema importância para esse objetivo é a aplicação de métodos estatísticos para detecção de fraudes, o que torna o processo mais transparente. Esses métodos evoluíram muito nos últimos anos devido ao forte avanço computacional [5]. A detecção de fraudes pelos métodos estatísticos consiste em avaliar pares de respostas de parte dos candidatos. Assim, a detecção se concentra na fraude denominada *cola*, física ou eletrônica, que consiste no repasse de gabarito entre indivíduos, e cuja fonte é de alta proficiência.

Apesar do avanço computacional, esses métodos estatísticos de detecção de fraudes não são exequíveis em alguns exames de larga escala, como por exemplo o ENEM, devido ao demasiado tempo de processamento. Isso se dá pelo elevado número de pares de respostas a serem analisadas. Por exemplo, em uma avaliação com 200.000 candidatos teria-se um total de pares a serem analisados de 19.999.900.000. Essa quantidade de combinações é inviável para análise em tempo hábil, com exceção de processamento em supercomputadoras. Nesse sentido, Souza [13] desenvolveu métodos que reduziram significativamente o tempo computacional, através de um pacote que trabalha em paralelo, denominado *Test-Fraud*. Este pacote avalia conjuntamente os sete índices de similaridade (ω, GBT, K, K_1, K_2, S_1 e S_2), por medio da função *Fraud.Indices*. Em seu estudo utilizou a metodologia de filtros de proficiências, pontuação mínima, e uma melhoria nas funções computacionais dos índices.

Na computação dos pares de indivíduos a serem avaliados, considera-se um deles como *fonte* e outro como *copiador*, e como antecipado, a fonte é considerada de alta proficiência. De forma geral, a fraude por *cola* é extremamente prejudicial às avaliações educacionais, pois em geral ocorrem por comunicação eletrônica e envolve grande números de indivíduos. Nesse tipo de fraude também existe a possibilidade de interação entre candidatos próximos na sala de aplicação de teste. Em avaliações educacionais, esse tipo de fraude é o objetivo dos métodos de detecção de fraude.

Nessas avaliações de larga escala os instrumentos de testes são elaborados de acordo com as diretrizes educacionais. Estas determinam as competências e habilidades de acordo com o nível de escolaridade dos examinados [4]. Os testes avaliativos são calibrados de acordo com a Teoria da Resposta ao Item (TRI), que leva em consideração a dificuldade, discriminação e acerto ao acaso dos índices. A partir dos resultados obtidos tem-se uma percepção acerca da qualidade do ensino. Essa percepção direciona as ações dos gestores públicos para as possíveis intervenções educacionais. Assim, os testes avaliativos devem ser aplicados com extrema confiança para não prejudicar as inferências acerca do conhecimento dos avaliados.

Um exemplo dessas avaliações de grande escala, no Brasil, é o Exame Nacional do Ensino Médio (ENEM). Esse descreve em termos quantitativos a qualidade do ensino médio brasileiro. O ENEM também é utilizado como forma parcial ou integral de ingresso das universidades públicas nacionais, além de algumas instituições internacionais que usam a nota desse exame. Ainda é utilizado como critério para programas sociais de educação, como por exemplo, o Programa Universidade para Todos (ProUni). Em outros países tem-se avaliações semelhante, um caso particular é a Colômbia, onde existe o Exame do Estado Saber 11 criado pelo Instituto Colombiano de Fomento da Educação Superior (ICFES), que avalia a qualidade da educação superior neste país.

Portanto, devido a importância dessas avaliações educacionais é indispensável que os exames sejam aplicados com lisura, a fim de garantir estimativas confiáveis acerca das proficiências dos examinados e uma avaliação correta sobre o sistema de ensino. Dessa forma é fundamental a identificação de fraudes, especialmente por *cola*. A identificação desse tipo de fraude é possível através dos métodos estatísticos de detecção de fraudes.

Por outro lado, essa detecção deve ocorrer em tempo compatível com as datas dos exames aplicados, para não prejudicar a organização dos mesmos e fazer a exclusão dos examinados fraudadores. Logo, é imprescindível a otimização desses métodos e o desenvolvimento de novos algoritmos, a fim de minimizar o tempo de processamento em exames de grandes escala.

1.2 Justificativa

A adequação de alguns métodos para detectar possíveis indícios de fraudes em avaliações em larga escala já foram desenvolvidos notavelmente, com base em implementações computacionais paralelas por Souza [13]. No entanto, mesmo com alguns avanços realizados na implementação, os processos ainda levam tempos significativamente altos e há necessidade de metodologias alternativas de seleção de pares. Ainda, os métodos implementados não consideram o controle das taxas de pessoas injustamente acusadas (*falsos positivos*), isso se deve em parte aos grandes volumes de dados que precisam ser processados. Portanto, é necessário otimizar essas implementações e reduzir significativamente as taxas de *falsos positivos* (FP).

1.3 Objetivos

1.3.1 Objetivo geral

Propor e implementar a seleção quantílica para detecção de potenciais fraudes em testes e otimizá-los no pacote *TestFraud* no ambiente R.

1.3.2 Objetivos específicos

i) Propor e desenvolver a metodologia de seleção quantílica de pares.

ii) Criar uma opção no pacote *TestFraud* para selecionar limiares quantílicos dos indivíduos de maiores proficiências.

iii) Verificar a eficiência da seleção quantílica através da simulação de dados.

iv) Avaliar a taxa de *falso positivo* (FP).

v) Aplicar a seleção quantílica nas provas no ENEM.

1.4 Organização da dissertação

Este trabalho encontra-se dividido em 6 capítulos, a saber:

- No Capítulo 1 é feita uma introdução ao conceito de avaliação em larga escala, em particular o ENEM, e a importância dos métodos de detecção para este exame, são abordados os aspectos gerais, justificativa e importância do trabalho, os objetivos geral e específicos, e o sumário da dissertação.
- No Capítulo 2 são apresentados os métodos implementados neste trabalho na área de detecção de fraude.
- No Capítulo 3 são apresentados os contextos dos conjuntos de dados e alguns métodos para controlar o tempo e a quantidade de candidatos em testes de larga escala.
- No Capítulo 4 são apresentados os métodos de redução de pares, de controle das taxas de *falsos positivos* e as novas alternativas do pacote *TestFraud*.
- No Capítulo 5 são apresentados resultados de simulação para analisar o desempenho das mudanças realizadas para a melhoria do pacote *TestFraud*, os resultados da estatística T^* ponderada e a análise da aplicação dos dados de ENEM.
- No Capítulo 6 são apresentadas as considerações finais e recomendações para trabalhos futuros.

No próximo capítulo será apresentada a Teoria da Resposta ao Item (TRI), metodologia utilizada para elaboração de testes educacionais mais válidos, fidedignos e precisos na aferição da proficiência de um aluno em determinada área do conhecimento em avaliações em larga escala. Além disso, também serão apresentados os principais métodos de detecção de fraudes por *cola* e o problema das taxas de *falsos positivos*.

Capítulo 2

Síntese dos principais métodos da área

2.1 Teoria da Resposta ao Item

É um conjunto de modelos matemáticos para traços latentes (variáveis não observados diretamente), onde se utilizam variáveis secundarias que estejam relacionadas com as latentes para poder inferir sobre os valores desses traços, é dizer que a Teoria da Resposta ao Item (TRI) tem como objetivo representar a probabilidade de um indivíduo dar uma certa resposta a um item como função dos parâmetros do item e de seus traços latentes, habilidades ou proficiências na área de conhecimento avaliada. Essa relação é diretamente proporcional, pois quanto maior valor de traço latente maior é a probabilidade de responder acertadamente ao item [1].

Os diferentes modelos propostos por a TRI dependem principalmente de três fatores:

i) do tipo de item (dicotômicos ou politômicos);

ii) da quantidade de populações envolvidas (uma ou mais de uma);

iii) do número das variáveis latentes que estão sendo medida (uma ou mais de uma).

Neste trabalho usaremos os modelos unidimensionais, isto é, os modelos que só estudam um traço latente, exemplos dessas variáveis latentes são a habilidade/proficiência em Português, grau de maturidade de uma empresa em Gestão pela qualidade, entre outros [1].

2.1.1 Modelo Logístico de 3 Parâmetros (ML3)

Na análise de itens dicotomizados ou dicotômicos (considerados como certo ou errado) é utilizados basicamente 3 (três) tipos de modelos logísticos que são determinados pelo número de parâmetros utilizados [1]:

1. Modelo Logístico de 1 Parâmetro (ML1) ou Modelo de Rasch: a dificuldade do item;
2. Modelo Logístico de 2 Parâmetros (ML2), sendo estes: a dificuldade e discriminação;
3. Modelo Logístico de 3 Parâmetros (ML3), sendo estes: a dificuldade, a discriminação e o acerto casual.

Os modelos mencionados anteriormente possuem unidimensionalidade (uma habilidade ou fator dominante) e a suposição da independência local ou independência condicional, isto é, que fixada a proficiência ou habilidade de um indivíduo os itens são respondidos de maneira independente. Segundo Hamblento & Swaminathan [6], na realidade é necessário só a unidimensionalidade dos modelos para cumprir com a suposição de independência local.

Dos modelos anteriores, o Modelo Logístico de 3 Parâmetros (ML3) é o mais utilizado e é dado por:

$$P(U_{ij} = 1|\theta_j) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta_j - b_i)}}, \tag{2.1}$$

com $i = 1, 2, \cdots, I$, e $j = 1, 2, \cdots, n$, em que:

- U_{ij} uma variável dicotômica que assume valor 1 quando o indivíduo j responde corretamente o item i, ou 0 caso contrário;
- θ_j é a habilidade do respondente (traço latente) do j-ésimo indivíduo;
- $P(U_{ij} = 1|\theta_j)$ é a probabilidade do indivíduo j com traço latente θ_j acertar o item i;
- b_i é o parâmetro de dificuldade (ou de posição) do item i, medido na mesma escala de θ_j;
- a_i é o parâmetro de discriminação (ou inclinação) do item i, com valor proporcional à inclinação da Curva Característica do Item no ponto b_i;
- c_i é o parâmetro de acerto casual do item i;
- D é um fator de escala, constante e igual a 1. Utiliza-se o valor $1,702$ quando deseja-se que a função logística forneça resultados semelhantes ao da função ogiva normal.

Sendo as estimações dos parâmetros a_i, b_i, c_i dos itens e da variável latente θ_j dos indivíduos uns dos principais objetivos na TRI.

A partir do ML3 se podem obter os modelos ML2 e ML1 (Modelo de Rasch) só basta utilizar valores específicos para a_i e c_i, isto é, para obter o Modelo Logístico de 2 Parâmetros não tem que existir a possibilidade de acerto ao acaso, ou seja, $c_i = 0$. Mas para obter o Modelo Logístico de 1 Parâmetro (ML1) ou de Rasch, além, de não ter resposta ao acaso precisa que os itens tenham o mesmo poder de discriminação, ou seja, $c_i = 0$ e $a_i = 1$. Note-se que o ENEM utiliza o ML3 para inferir os valores (estimar) as proficiências dos alunos nas quatro áreas de conhecimento existente no exame.

2.1.2 Modelo de Resposta Nominal (MRN)

É um modelo logístico geral aplicável a todas as categorias de respostas escolhidas em um teste com itens de múltipla escolha [1]. De acordo com esses mesmos autores, o Modelo de Resposta Nominal (MRN), foi elaborado por Bock [3], com o objetivo de maximar a precisão da habilidade estimada usando toda a informação contida nas respostas dos indivíduos, e não apenas se o item foi respondido corretamente ou incorretamente. No modelo MRN a probabilidade de um indivíduo j selecionar uma particular opção v (de V opções avaliáveis) do item i é dada por:

$$P_{iv}(\theta_j) = \frac{e^{(\zeta_{iv}+\lambda_{iv}\theta_j)}}{\sum_{v=1}^{V} e^{(\zeta_{iv}+\lambda_{iv}\theta_j)}}, \tag{2.2}$$

com $i = 1, 2, \cdots, I$, $j = 1, 2, \cdots, n$, e $v = 1, 2, \cdots, V$. Em cada θ_j, a soma das probabilidades sobre as V opções, $\sum_{v=1}^{V} P_{iv}(\theta_j)$ é 1. As quantidades ζ_{iv} e λ_{iv} são parâmetros denominados, respectivamente, de intercepto e inclinação do item para alternativa v do item i. Um dos métodos usados para estimar esses parâmetros e a habilidade é o de máxima verossimilhança [2], [3].

2.2 Métodos de detecção de fraudes

Os sete métodos estatísticos para detecção de possíveis copiadores que serão apresentados, foram elaborados de acordo com a literatura na categoria dos índices de similaridade de respostas, os quais, tem como objetivo analisar o grau de concordância entre dois ve-

tores de respostas. Esses índices de similaridade de resposta podem ainda ser classificados com base em dois atributos:

- a distribuição estatística de referência em que se baseiam,
- e a evidência de que a cópia de resposta está sendo usada ao calcular a probabilidade de concordância entre dois vetores de resposta.

Na subseção atual se descreverá e fornecerá uma visão geral de alguns desses índices [16]. Além disso, na seguinte Tabela 2.1 é apresentada uma comparação das principais características dos sete índices que compõem o pacote *TestFraud* utilizado neste trabalho.

Tabela 2.1 *Análise comparativa dos sete índices do Pacote TestFraude*

Características	Índices						
	ω	GBT	K	K_1	K_2	S_1	S_2
Examina só Respostas Incorretas Idênticas	Não	Não	Sim	Sim	Sim	Sim	Não
Examina Respostas Idênticas	Sim	Sim	Não	Não	Não	Não	Sim
Modelo de Distribuição	N(0,1)	BC	$B(w_s, p)$	$B(w_s, \hat{p}_1^*)$	$B(w_s, \hat{p}_2^*)$	$P(\hat{\mu}_{c'}^w)$	$P(\hat{\mu}_{c'}^w)$

2.2.1 Índice ω

Este índice analisa o número de respostas idênticas entre dois indivíduos, tanto corretas quanto incorretas. Assim, segundo Wollack [15] se considera que exista um indivíduo c suspeito de copiar as respostas do indivíduo s, e h_{cs} o número de itens respondidos de forma idênticas pelos indivíduos c e s em um teste de múltipla escolha com opções $v = 1, \cdots, V$. Logo, condicionando às respostas de s se pode definir h_{cs} como

$$h_{cs} = \sum_{i=1}^{I} 1[u_{ic} = u_{is}], \tag{2.3}$$

sendo $i = 1, 2, \cdots, I$ e representa o i-ésimo item, u_{ic} e u_{is} são as alternativas do item i escolhidas pelos examinados c e s respectivamente. Além, tem-se a função indicadora,

$$1[u_{ic} = u_{is}] = \begin{cases} 1, & \text{se } c \text{ e } s \text{ selecionaram a mesma alternativa } v, \\ 0, & \text{c. c.} \end{cases} \tag{2.4}$$

Para a obtenção da distribuição de h_{cs}, calcula-se a probabilidade de c selecionar as respostas providas por s dada a habilidade do examinado c (θ_c), o vetor de respostas do examinado s (U_s) e a matriz de parâmetros dos itens (ξ). Logo, o valor esperado dessa distribuição é igual a

$$\begin{aligned} E(h_{cs}|\theta_c, U_s, \xi) &= E\left[\sum_{i=1}^{I} 1(u_{ic} = u_{is}|\theta_c, U_s, \xi)\right] \\ &= \sum_{i=1}^{I} E\left[1(u_{ic} = u_{is}|\theta_c, U_s, \xi)\right] \\ &= \sum_{i=1}^{I} \left[P(u_{ic} = u_{is}|\theta_c, U_s, \xi)\right]. \end{aligned} \tag{2.5}$$

Assumindo que as respostas dos indivíduos aos itens são localmente independentes e a partir das Equações (2.4) e (2.5) condicionando as respostas em s e os parâmetros dos itens, h_{cs} é a soma de variáveis Bernoulli independentes cada uma com probabilidade de sucesso, é dizer, com média igual a

$$P(u_{ic} = u_{is}|\theta_c, U_s, \xi), \tag{2.6}$$

sendo assim, o desvio-padrão de h_{cs} é dado por

$$\sigma_{h_{cs}} = \sqrt{\sum_{i=1}^{I} [P(u_{ic} = u_{is}|\theta_c, U_s, \xi)][1 - P(u_{ic} = u_{is}|\theta_c, U_s, \xi)]}. \tag{2.7}$$

Para obter $P(u_{ic} = u_{is}|\theta_c, U_s, \xi)$ usa-se o MRN, descrito na Seção 2.1.2.

Utilizando o Teorema Central do Limite (TCL) tem-se que

$$\omega = \frac{h_{cs} - \sum_{i=1}^{I} [P(u_{ic} = u_{is}|\theta_c, U_s, \xi)}{\sqrt{\sum_{i=1}^{I} [P(u_{ic} = u_{is}|\theta_c, U_s, \xi)][1 - P(u_{ic} = u_{is}|\theta_c, U_s, \xi)]}} = \frac{h_{cs} - E(h_{cs}|\theta_c, U_s, \xi)}{\sigma_{h_{cs}}}, \tag{2.8}$$

tem distribuição aproximadamente normal com média 0 e variância 1, ($\omega \sim N(0, 1)$). Logo, segundo Sotaridona [12] e Wollack [15] considerando o valor observado ou p-valor

de ω é possível obter evidências que o indivíduo c cometeu fraude. Para isto basta verificar se o valor observado de ω é maior ou igual que o valor crítico para um nível de significância (α) assumido, pois, quanto maior o valor de ω mais forte é a evidência de que c copiou de s.

2.2.2 Teste da Binomial Generalizada (GBT)

O método GBT analisa o número de respostas idênticas, tanto corretas quanto incorretas, entre dois indivíduos, através da família de distribuição binomial generalizada ou binomial composta (esse último nome se utilizará porque é o mais apropriado para o caso de amostragem aleatório de itens) [14]. Então, seja P_{M_i} a probabilidade das respostas dos examinados c e s ao item i coincidirem, logo se pode calcular P_{M_i} como

$$P_{M_i} = \sum_{v=1}^{V} P_{civ} \cdot P_{siv}, \tag{2.9}$$

sendo P_{civ} e P_{siv} as probabilidades de c e s selecionar a alternativa v do mesmo item i, respectivamente. Neste trabalho, essas probabilidades foram computadas com o Modelo de Resposta Nominal (MRN), mas se pode utilizar algum outro modelo de resposta.

Portanto, a probabilidade de serem observadas exatamente n correspondências (ou respostas iguais) em I itens entre dois examinados é igual a

$$f_I(n) = \sum \left(\prod_{i=1}^{I} P_{M_i}^{u_i} Q_{M_i}^{1-u_i} \right) = \sum \left(\prod_{i=1}^{I} P_{M_i}^{u_i} (1 - P_{M_i})^{1-u_i} \right), \tag{2.10}$$

sendo Q_{M_i} o complemento de P_{M_i}, isto é,

$$Q_{M_i} = 1 - P_{M_i} \tag{2.11}$$

e

$$u_i = \begin{cases} 1, & \text{se } c \text{ e } s \text{ respondem identicamente ao item } i, \\ 0, & \text{c.c.} \end{cases} \tag{2.12}$$

em que o somatório da Equação (2.10) refere-se a todas as combinações possíveis de n respostas coincidentes em I itens. Então, ilustrando a equação (2.10), suponha o caso onde se tem duas respostas iguais em três itens, como segue

$$f_3(2) = \sum \left(\prod_{i=1}^{3} P_{M_i}^{u_i} Q_{M_i}^{1-u_i} \right) = Q_{M_1} P_{M_2} P_{M_3} + P_{M_1} Q_{M_2} P_{M_3} + P_{M_1} P_{M_2} Q_{M_3} \tag{2.13}$$

Logo, o índice GBT é definido como a cauda superior da distribuição binomial composta e, assim, a probabilidade de observar o número de respostas incorretas idênticas (w_{cs}) mais o número de respostas corretas coincidentes (R_{cs}) ou mais em I itens é igual a

$$\sum_{n=w_{cs}+R_{cs}}^{I} f_I(n). \tag{2.14}$$

Para caracterização se houve fraude entre os examinados c e s, basta verificar se P_{M_i} é menor o igual ao nível de significância α, em caso contrario, ($P_{M_i} > \alpha$) não houve fraude [16].

2.2.3 Índice K

A diferencia dos índices anteriores (ω e GBT) que analisavam o número de respostas idênticas (corretas ou incorretas), o índice K, proposto por Holland [7], avalia só as coincidências de respostas incorretas num teste de multiplica escolha entre dois examinandos. Para cada par de examinados, considere-se as mesmas notações anteriores: o indivíduo suspeito de copiar as respostas é c, e o examinado fonte dessas respostas é s.

Para esta subseção serão utilizadas as seguintes notações:

- j, com ($j = 1, \cdots, J$), denotando os examinados;
- i, com ($i = 1, \cdots, I$), denotando os itens;
- v, com ($v = 1, \cdots, V$), denotando as alternativas de um item;
- w_j sendo o número de respostas "erradas"do examinado j;
- r, com $r = 1, \cdots, c', \cdots, R$, denotando os subgrupos de examinados, em que cada subgrupo tem um número distinto de respostas incorretas, R é o número total de subgrupos ($R = I + 1$, salvo se houver algum subgrupo vazio), além disso, cada subgrupo possui no mínimo um examinado e que $\sum_{r=1}^{R} n_r = J - 1$, denota-se aqui c' como o subgrupo ao qual o examinado c pertence e n_r é o número total de examinados de cada subgrupo r;
- j', com $j' = 1, \cdots, n_r$, denotando os examinados dentro de um subgrupo r específico.
- $\mathbf{M_r} = (M_{r1}, \cdots, M_{rj'}, \cdots, M_{rn_r})$ sendo um vetor dos números de respostas incorretas idênticas às da fonte em um particular subgrupo r;

- $\mathbf{M_{c'}} = (M_{c'1}, \cdots, M_{c'n_r})$ denotando o vetor do número de respostas incorretas idênticas às da fonte de $n_{c'}$ examinados do subgrupo c', sendo este o subgrupo que possui o mesmo número de respostas incorretas do copiador.

- $m_{rj'}$ sendo o valor observado do número de respostas incorretas idênticas entre o examinado rj' e s;

- $Q_r = \frac{w_r}{I}$ como a proporção de respostas incorretas de um subgrupo r, sendo w_r o número de respostas erradas do subgrupo r e I é o número total de itens do teste.

Esse índice possui duas formulações para serem obtidas, estas são: a construção por uma distribuição amostral empírica (dados empíricos) e a construção fundamentada em um modelo aproximado (distribuição teórica).

2.2.3.1 Índice K Baseado na Distribuição Empírica

A construção do índice K de forma empírica utiliza os dados empíricos de J examinados respondendo a I itens. Para essa finalidade, sugeriu-se adotar os seguintes passos:

- definir o grupo de examinados com o mesmo número de respostas incorretas de c (subgrupo c');

- para cada examinado do subgrupo c', definir o número de itens incorretos idênticos ao examinado s, obtendo-se assim o vetor $\mathbf{M_{c'}}$.

Assim, dadas as premissas anteriores, pode-se definir o índice K como a proporção de examinados com o mesmo número de respostas incorretas que o copiador e cujo número de respostas incorretas correspondentes com a fonte é maior ou igual $m_{c'c}$, onde $m_{c'c}$ é o número de respostas incorretas idênticas entre c e s. Portanto, o índice K é dado por:

$$K = \frac{\sum_{j'=1}^{n_{c'}} I_{c'j'}}{n_{c'}}, \tag{2.15}$$

com a função indicadora a seguir,

$$I_{c'j'} = \begin{cases} 1, & se \quad m_{c'j'} \geq m_{c'c}, \\ 0, & c.c. \end{cases} \tag{2.16}$$

Portanto, para a análise de fraude tem-se que quando K é pequeno, há evidência estatística que o examinado c copiou as respostas de s. No entanto, note-se que as habilidades dos examinandos irão influenciar diretamente na contagem de correspondência

de respostas incorretas, ou seja, o número de respostas incorretas idênticas é necessariamente pequena quando s ou c ou ambos têm muitas respostas corretas (altas proficiências). Devido esta situação, obtêm-se o índice K condicionalmente sobre o número de contagens incorretas da copiadora suspeita. Com isso o índice é obtido através apenas de uma pequena quantidade de examinandos do subgrupo c' (considere-se como pequenas amostras $J \leq 100$ e indica-se utilizar uma aproximação teórica para a distribuição de correspondência empírica), o qual influencia na precisão do valor do índice K, além disso, se tem o impedimento da obtenção do erro Tipo I pré-especificado de 0,01 [11].

2.2.3.2 Índice K Baseado na Aproximação Teórica

Com o propósito de evitar ao máximo o problema presente na implementação empírica do índice K e não apontar um examinado injustamente de fraude (*Falso Positivo*). Holland [7] propôs obter o índice a partir de uma distribuição teórica do número de respostas incorretas iguais entre o examinado c' e s, sendo esta variável simplesmente denotada por M. Segundo esse mesmo autor a distribuição de M pode ser aproximada por uma distribuição binomial para calcular a probabilidade do examinado c possuir o número de respostas incorretas iguais as do examinado s maior que os outros examinados pertencentes ao subgrupo c'. Portanto, tem-se que

$$M \overset{aprox.}{\sim} Bin(w_s, p), \tag{2.17}$$

em que w_s o número de respostas incorretas de s, o qual é conhecido, e p é a probabilidade esperada de M, no entanto, p é desconhecido [11]. Assim, Holland [7] sugeriu dois formas de aproximar p:

- p é computado para que a distribuição binomial e a distribuição empírica de M tenham as mesmas médias.

Seja $\bar{m}'_c$ a média da distribuição empírica de concordância tem-se que

$$\bar{m}_{c'} = \frac{\sum_{j'=1}^{n_{c'}} m_{c'j'}}{n_{c'}}. \tag{2.18}$$

Assim, tem-se uma estimativa de p através de $p^*_{c'}$ definida como

$$p^*_{c'} = \frac{\bar{m}_{c'}}{w_s}. \tag{2.19}$$

Logo, seja K^* o índice K baseado em $p^*_{c'}$, então K^* é obtido por

$$K^* = P(M \geq m_{c'c}) = \sum_{w=m_{c'c}}^{w_s} \binom{w_s}{w} (p^*_{c'})^w (1 - p^*_{c'})^{w_s - w}. \tag{2.20}$$

Observar-se que o cálculo $p^*_{c'}$ dependente dos vetores de respostas dos examinados no subgrupo c', por tal motivo, devem estar disponíveis. Além disso, o valor de $p^*_{c'}$ é sensível ao tamanho da amostra tornando-se menos confiável quando a amostra é pequena [11], [12].

A segunda sugestão de Holland [7] para obter uma estimativa de p foi através do método da regressão linear, na qual essa regressão fosse calculada utilizando a proporção de respostas incorretas (Q_r) de cada subgrupo como a variável preditora.

- O mesmo autor mostrou empiricamente com grandes bases de dados que p^*_r é linearmente relacionado a Q_r e p^*_r é definido de modo análogo como em (2.19).

Assim, seja $\hat{p}_r$ a estimativa da probabilidade binomial de p^*_r utilizando Q_r. A expressão para $\hat{p}_r$ usando regressão linear é:

$$\hat{p}_r = \begin{cases} a + bQ_r, & \text{se} \quad 0 < Q_r \leq 0.3; \\ [a + 0.3b] + 0.4b[Q_r - 0.3], & \text{se} \quad 0.3 < Q_r \leq 1. \end{cases} \tag{2.21}$$

Note-se que os valores de a e b são os parâmetros de intercepto e de inclinação, respectivamente. Além disso, segundo Sotaridona & Meijer [11] observa-se que a e b devem ser especificados para estimar $\hat{p}_r$ do modelo de regressão anterior, sendo estas condicionadas ao valor Q_r. Holland [7] usou $a = 0,085$ e diferentes valores para b baseado na configuração do teste específico utilizado. Porém, não está claro como esses valores foram obtidos no estudo feito por Holland. Para verificar se houve fraude entre os examinados compara-se o valor observado de K com o nível de significância adotado e se o valor de $K < \alpha$, então eles fizeram fraude [7], [11].

2.2.4 Índices K_1 e K_2

Com o propósito de minimizar erros Sotaridona & Meijer [11] propuseram $\hat{p}^*_1$ e $\hat{p}^*_2$ como estimavas de p^*_r, sendo essas duas alternativas baseados, respectivamente, a partir de uma regressão linear e uma quadrática utilizando Q_r como variável preditora. Isto é,

$$\hat{p}^*_1 = \beta_0 + \beta_1 Q_r + \epsilon_r, \tag{2.22}$$

e

$$\hat{p}_2^* = \beta_0 + \beta_1 Q_r + \beta_2 Q_r^2 + \epsilon_r, \tag{2.23}$$

sendo, β_0 e β_1 os parâmetros de intercepto e inclinação, respectivamente, β_2 é um parâmetro de regressão e $\epsilon_r \sim N(0, \sigma^2)$ é o erro. Então, usando as estimativas de p_r^*, tem-se duas versões do índice K, K_1 e K_2, no qual cada uns é dado pelas seguintes probabilidades:

$$K_1 = P(M \geq m_{c'c}) = \sum_{w=m_{c'c}}^{w_s} \binom{w_s}{w} (\hat{p}_1^*)^w (1 - \hat{p}_1^*)^{w_s - w} \tag{2.24}$$

e

$$K_2 = P(M \geq m_{c'c}) = \sum_{w=m_{c'c}}^{w_s} \binom{w_s}{w} (\hat{p}_2^*)^w (1 - \hat{p}_2^*)^{w_s - w}. \tag{2.25}$$

A diferença de $p_{c'}^*$ com $\hat{p}_1^*$ e $\hat{p}_2^*$, é que $p_{c'}^*$ só utiliza as informações do subgrupo c' para estimar p, mas $\hat{p}_1^*$ e $\hat{p}_2^*$ utilizam os dados de todos os R subgrupos nesta estimativa. Além disso, Sotaridona & Meijer [11] mostraram que $\hat{p}_2^*$ gerou melhores estimativas para p que $\hat{p}_1^*$ e $p_{c'}^*$. Para verificar se houve fraude entre os examinados compara-se os valores observados de K_1 e K_2 com o nível de significância adotado e se os valores de $K_1, K_2 < \alpha$, então eles fizeram fraude.

2.2.5 Índices S_1 e S_2

2.2.5.1 Índices S_1

O índice S_1, foi proposto por Sotaridona & Meijer [12] e é similar ao índice K_2, porque, também é baseado no número de respostas incorretas idênticas entre os examinados c' e s (ou variável aleatória M). As diferenças entre esses dois índices se dá porque

- Para o índice S_1, a variável aleatória M segue uma distribuição de Poisson, enquanto que ao índice K_2 foi atribuída uma distribuição binomial para M.
- A estimação do parâmetro p, em K_2 é feita por um modelo de regressão quadrática (ver seção 2.2.4), mas para o índice S_1, a estimação do valor esperado μ (ou média de M) da distribuição de Poisson é feita a partir do modelo log-linear.

Então, o índice S_1 é representado, assim:

$$S_1 = P(M \geq m_{c'c}) = \sum_{w=m_{c'c}}^{w_s} \frac{e^{-\hat{\mu}_{c'}} \hat{\mu}_{c'}^w}{w!}, \tag{2.26}$$

sendo $\hat{\mu_{c'}}$ a estimativa para μ utilizando o modelo log-linear e este é dado por:

$$\log(\mu_r) = \beta_0 + \beta_1 w_r, \forall\, r, \tag{2.27}$$

onde β_0 é o intercepto, β_1 a inclinação, μ_r o valor esperado da variável Poisson $M_{rj'}$ e w_r é o número de respostas incorretas do subgrupo r. Assim, a média ajustada para o subgrupo c' ($\hat{\mu_{c'}}$) é:

$$\hat{\mu}_{c'} = e^{\beta_0 + \beta_1 w_{c'}}. \tag{2.28}$$

Portanto, quanto menor o valor de S_1, mais forte é a evidência das respostas terem sido copiadas [12].

2.2.5.2 Índices S_2

Como os índices K, K_1, K_2 e S_1 são baseados em examinar só as respostas incorretas idênticas, então Sotaridona & Meijer [12] propuseram o índice S_2, o qual incorpora em seus cálculos informações não somente de correspondências de respostas incorretas, mas também das respostas corretas iguais apresentadas pelos examinandos para a sua computação. A justificativa é de que os índices que só consideram as respostas incorretas idênticas se tornam "insensíveis", pois, não inclui os casos onde os indivíduos c e s são amigos, e s compartilha suas respostas com c nos itens onde tem quase certeza que está correta ou onde c subornar a s para compartilhar seus itens corretamente respondidos.

Então, seja $M^*_{rj'}$ a soma entre o número de respostas coincidentes incorretas e o número de respostas coincidentes corretas ponderadas entre s e o examinado rj' pertencente a um subgrupo r específico. A expressão $M^*_{rj'}$ é dada por

$$M^*_{rj'} = M_{rj'} + \sum_{i^*} \delta_{i^* rj'}, \tag{2.29}$$

onde $\delta_{i^* rj'}$ é a estimativa da informação de cópia do item i^* pelo examinado rj' e i^* representado os itens respondidos corretamente pela fonte. O termo $\delta_{i^* rj'}$ é definida por:

$$\delta_{i^* rj'} = f(P_{i^* rj'}) = d_1 e^{d_2 P_{i^* rj'}}, \tag{2.30}$$

com $0 \leq \delta_{i^* rj'} \leq 1$, e $P_{i^* rj'}$ a probabilidade do examinado rj' responder corretamente ao

item i^*. Em virtude do método da máxima verossimilhança $P_{i^*rj'}$ é estimado por

$$\hat{P}_{i^*rj'} = \frac{\sum_{j'=1}^{n_r} I_{(u_{i^*rj'}=u_{i^*s})}}{n_r}, \tag{2.31}$$

onde

$$I_{(u_{i^*rj'}=u_{i^*s})} = \begin{cases} 1, & se \quad j' \quad responder \quad corretamente \quad ao \quad item \quad i^*, \\ 0, & c.c. \end{cases} \tag{2.32}$$

Logo, os valores d_2 e d_1 são dados por

$$d_2 = -\left(\frac{1+g}{g}\right), \tag{2.33}$$

$$d_1 = -\left(\frac{1+g}{1-g}\right)^{d_2 P_{i^*c}}, \tag{2.34}$$

no qual g a probabilidade de individuo que desconhece o item acertá-lo ao acaso, é dizer que se um item tem V alternativas, então $g = 1/V$ (para mais detalhes vide Sotaridona & Meijer (2003), pág. 36) [12].

Quando não há respostas corretas coincidentes entres rj' e s, tem-se que o segundo termo da Equação (2.29) zera e portanto, $M_{rj'}$ se torna um caso especial de $M^*_{rj'}$. Em contrapartida, quando não há respostas incorretas coincidentes entres rj' e s o primeiro termo da Equação (2.29) zera e $M^*_{rj'} = \sum_{i^*} \delta_{i^*rj'}$, tornando-se uma variável sensível para todo conjunto de respostas. Para a aplicação o valor de $M^*_{rj'}$ é tratado como um número inteiro [12]. Assim, o índice S_2 é definido sobre a distribuição de Poisson de $M^*_{c'c}$ (ou simplesmente M^*) e usa-se o modelo log-linear para estimar a média de M^*, como foi realizado no índice S_1. Portanto, o índice S_2 é definido como

$$S_2 = P(M^* \geq m^*_{c'c}) = \sum_{w=m^*_{c'c}}^{I} \frac{e^{-\hat{\mu}_{c'}} \hat{\mu}^w_{c'}}{w!}, \tag{2.35}$$

em que $m^*_{c'c}$ é o número observado de coincidências incorretas e corretas ponderada entre os indivíduos c e s. Assim, para valores de S_2 menores ao α adotado ou pequenos valores de S_2, tem-se maior evidência que a cópia ocorreu [12].

2.2.6 Taxa de *Falso Positivo* (FP)

Segundo Zopluoglu, Cizek & Wollack [16], a taxa de *falso positivo* (FP) é a probabilidade de um método identificar examinados que fizeram fraudes, quando na realidade isso não

aconteceu. De forma equivalente, é a proporção de pares honestos detectados falsamente como copiadoras. Nesse contexto, a grande preocupação é obter uma estatística conservadora, ou seja, uma estatística para a qual as taxas de FP sejam no máximo possível equivalentes ao valor esperado ou adotado, pois uma má classificação de indivíduos honestos como copiadores pode ser muito grave a nível individual. Nesse sentido, segundo a literatura tem-se os chamados tipos de erros, classificados da seguinte forma:

Taxa de Erro Tipo I : ocorre quando um índice caracteriza erroneamente um fato como fraude quando não é;

Taxa de Erro Tipo II : ocorre quando um índice não consegue detectar um caso de fraude que aconteceu.

Segundo Satoridona [10], os índices de cópia que não conseguem manter a taxa de erro Tipo I nominal deve ser considerado inaceitáveis. Por outro lado, um índice de cópia não deve ser excessivamente conservador; caso contrário, seu poder de detectar verdadeiros examinandos como copiadores será muito baixo. Neste trabalho centra-se no controle das taxas de erro Tipo I, e para isso é proposta uma estatística T^* ponderada descrita no Capitulo 4, especificamente na Seção 4.2.

No próximo capítulo serão apresentados os contextos dos conjuntos de dados e alguns métodos para controlar o tempo e a quantidade de candidatos em testes de larga escala.

Capítulo 3

Distribuições dos escores e notas finais

Na TRI a construção das estatísticas de itens e indivíduos são consistentes, e as proficiências são equalizadas na métrica iniciada em 2009, quando o ENEM foi reformulado, em que as quatro áreas tiveram as proficiências estimadas e transformadas para média 500 e desvio-padrão 100, no conjunto dos candidatos concluintes da modalidade regular. A partir de 2009, as médias e desvios-padrão poderiam variar, mantendo o ano 2009 como referência na escala denotada por (500; 100).

No entanto, as distribuições de cada área apresentam assimetria. Isso ocorre devido as misturas de populações diferentes, como escola públicas e particulares, regiões do país e outras populações. O fator tempo, também, pode alterar a forma da distribuição, pois, ao longo dos anos a média das notas pode subir ou descer.

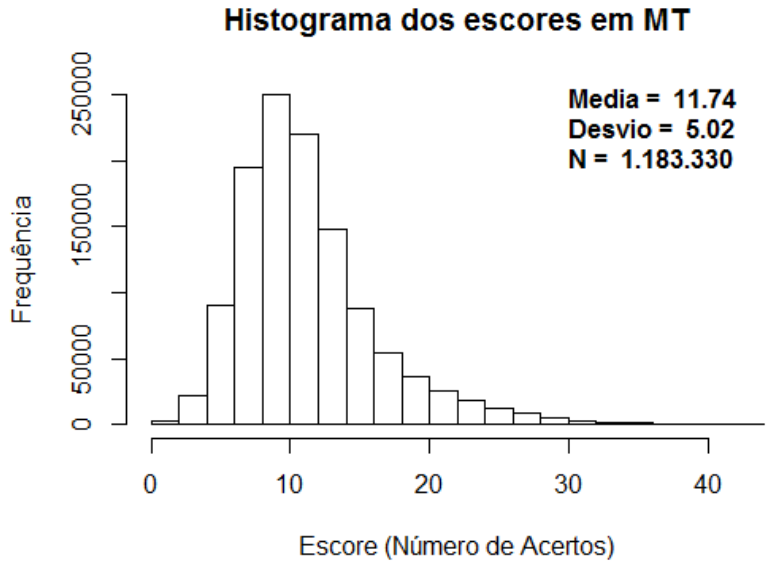
Portanto, as diferenças entre a formas das distribuições de cada área para um valor particular de escore impactam no número de pares de candidatos a serem analisados. Dessar forma, é necessário o estudo da distribuição de cada área para seleção do número de candidatos analisados.

3.1 Distribuição dos escores observados

Na Teoria Clássica dos Testes (TCT) o escore corresponde a somar todas as questões em um teste, sendo comumente adotado o valor 0 (zero) em caso de erro e 1 para respostas corretas, mas também pode haver ponderações para os itens. Essa teoria é muito útil para avaliar as provas como um todo e principalmente a curva característica de cada item. Na Figura 3.1, a média de acertos dos alunos em Matemática foi de 11,7400 com desvio-padrão de 5,0200 e coeficiente percentílico de curtose de 5,7670, o qual indica que é uma distribuição leptocúrtica. Nesta prova, a distribuição das notas é assimétrica a direita,

tendo um coeficiente de assimetria de Pearson de 1,3640 pois, poucos examinados têm notas altas, acima de 20 questões. Essa distribuição é particular para esse caderno.

Figura 3.1 *Histograma do número de acertos da prova azul de Matemática e suas tecnologias, ENEM 2018*



Conforme as Figuras 3.2 a 3.4 observa-se o comportamento diferenciado da distribuição dos escores. As notas na prova de Linguagens e Códigos apresenta menos assimetria (0,4700) em comparação com as provas de Ciências Humanas e Ciências da Natureza, 0,8680 e 1,4410 respectivamente. Dessa forma, essas distribuições têm características próprias de cada área, ou seja, as distribuições das provas de ciências Humanas e Ciências da Natureza apresentam distribuição leptocúrtica, tendo os coeficientes de curtoses iguais a 3,3220 e 6,4980, respectivamente e, a prova de Linguagens e Códigos tem distribuição platicúrtica, pois seu coeficiente de curtose foi 2,6060. Em particular no ENEM, pode haver variação da distribuição na mesma área ao longo dos anos de aplicação.

Figura 3.2 *Histograma do número de acertos da prova azul de Ciências Humanas e suas tecnologias, ENEM 2018*

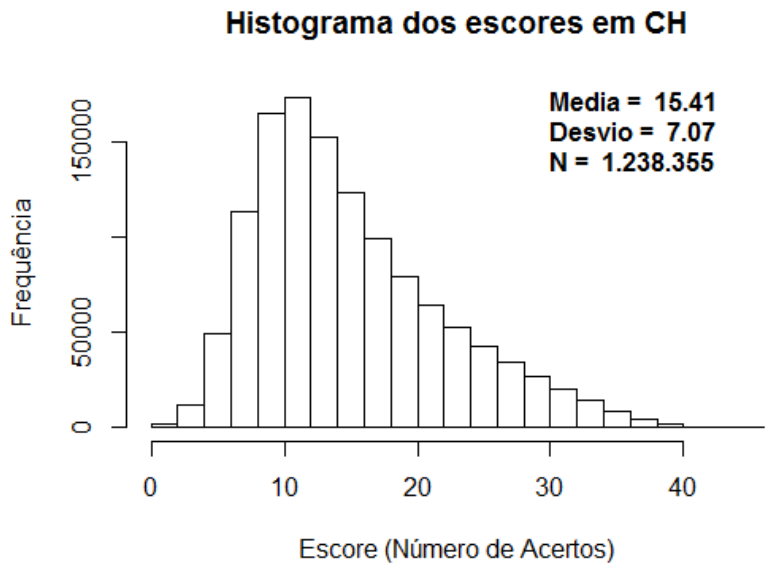


Figura 3.3 *Histograma do número de acertos da prova azul de Ciências da Natureza e suas tecnologias, ENEM 2018*

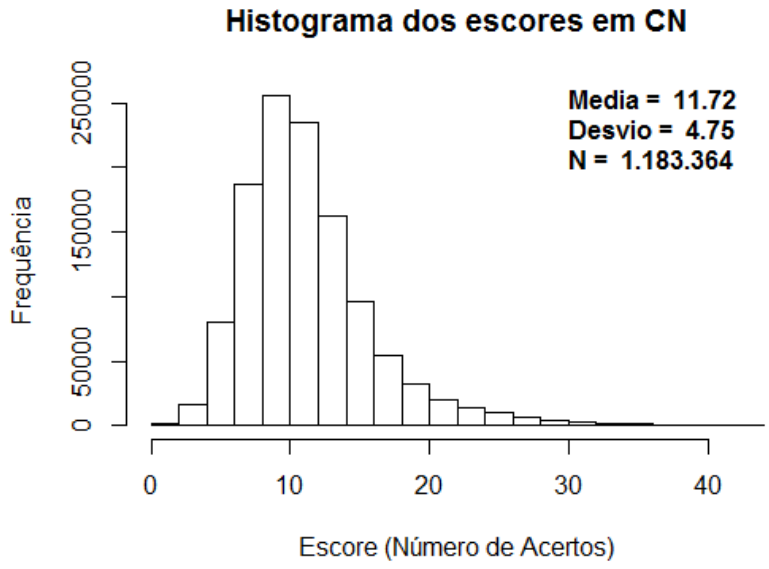
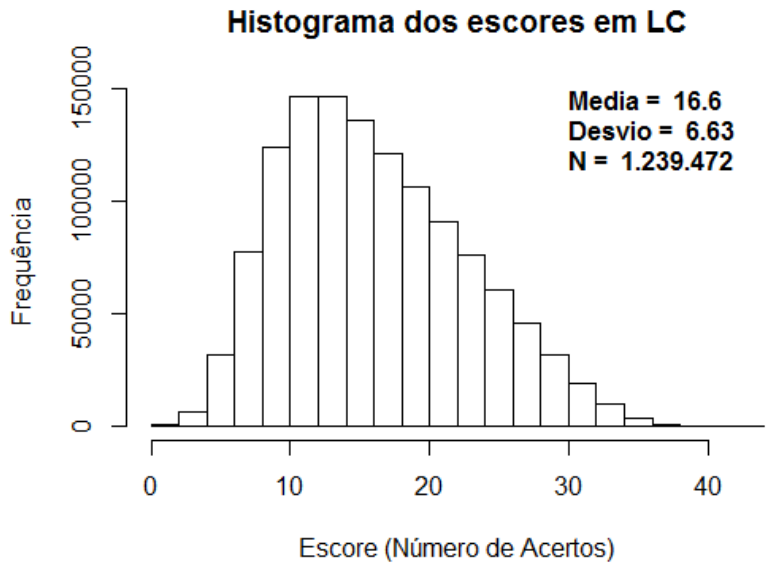


Figura 3.4 *Histograma do número de acertos da prova azul de Linguagens, Códigos e suas tecnologias, ENEM 2018*



Caso a opção fosse a fixação de um escore mínimo para a seleção dos indivíduos a entrarem na análise, poderia haver conjuntos bem distintos em cada área do exame. Portanto, essas distribuições demostram que deve ser realizado um estudo do quantil a ser analisado para cada distribuição. Como já se discutiu antes, vários fatores afetam a forma da curva de cada prova, e dentro da mesma área existe a mistura de várias populações.

3.2 Distribuição dos escores verdadeiros

Dado um conjunto de parâmetros de itens $\boldsymbol{\zeta}_i = (a_i, b_i, c_i), i = 1, \cdots, I$, o escore esperado (ou verdadeiro, ou *true score*, em inglês) de um indivíduo com habilidade θ_j é dado por:

$$T_j = \sum_{i=1}^{I} P(U_{ij} = 1|\theta_j, \boldsymbol{\zeta}_i). \tag{3.1}$$

Nota-se que para indivíduos de habilidade consideravelmente baixa, teremos que cada probabilidade de acerto é baixa, e com isso teremos T_j também baixo, aproximadamente $\overline{c} = \sum_{i=1}^{I} c_i$. Por outro lado, para indivíduos de alta proficiência, cada probabilidade de acerto será alta, e assim teremos T_j próxima de I. Com isso, vemos que o escore verdadeiro é uma versão contínua do escore observado, mas que considera as características dos itens obtidas via TRI.

3.3 Distribuição das proficiências ou habilidades

Pelas Figuras de 3.5 a 3.8, observa-se que a distribuição de cada área apresenta assimetria diferente. Sendo a prova de Matemática a mais assimetria a direita com 0,6920 e tendo uma distribuição platicúrtica, pois o coeficiente de curtose foi de 2,9050, a prova de Linguagem e Códigos tem uma distribuição platicúrtica (curtose de 2,5190) e assimetria a esquerda (-0,1760), as provas de Ciências Humanas e Ciências da Natureza presentaram distribuição platicúrtica com coeficiente de curtose de 2,1710 e 2,8600, respectivamente. Além disso, Ciências Humanas apresentou assimetria a esquerda (-0,2490) e Ciências da Natureza assimetria a direita (0,6080).

Figura 3.5 *Histograma das proficiências da prova azul de Matemática e suas tecnologias, ENEM 2018*

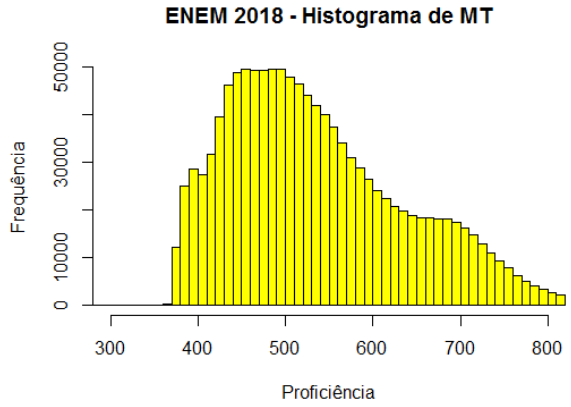


Figura 3.6 *Histograma das proficiências da prova azul de Ciências Humanas e suas tecnologias, ENEM 2018*

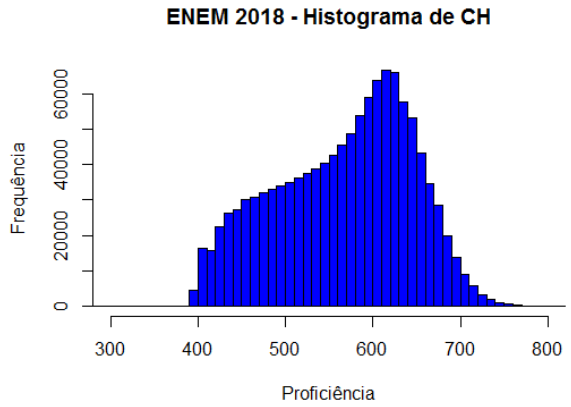


Figura 3.7 *Histograma das proficiências da prova azul de Ciências da Natureza e suas tecnologias, ENEM 2018*

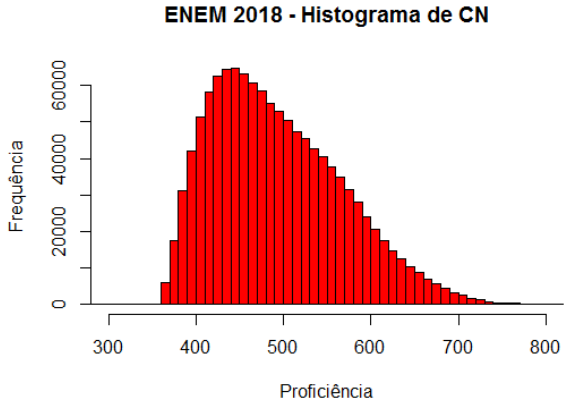
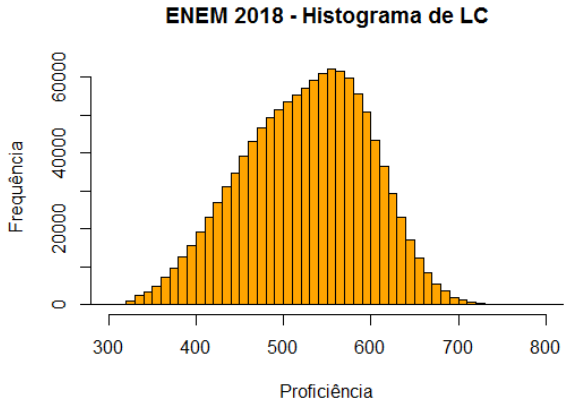


Figura 3.8 *Histograma das proficiências da prova azul de Linguagens, Códigos e suas tecnologias, ENEM 2018*



Tem-se que a distribuição das proficiências difere bastante segundo a área. Logo, a valor cada quantil a ser definido depende da particular distribuição em análise. É basicamente a mesma interpretação que houve para distribuição dos escores. Logo, para seleção quantílica existe a necessidade de analisar cada área de forma particular, pois, um quantil q para área de matemática difere totalmente das demais.

3.4 Número de indivíduos

Se o objetivo é identificar fraudes entre candidatos que realizam uma prova, então, deve-se comparar o padrão de resposta de dois indivíduos (c e s) usando os métodos da Seção 2.2. Certamente é melhor analisar os indivíduos com maiores habilidades do que

fazê-lo para todos os indivíduos da prova, porque se eles copiaram é esperado que seus resultados sejam altos. Logo, com base nessa premissa pode-se dizer que essa quantidade de indivíduos a ser analisado é muito menor em relação ao total, e é de suma importância quando se trata de testes de larga escala.

Por exemplo, suponha que a análise deva ser aplicada para todos os indivíduos que realizam o teste, o que implica que se deve comparar todas as combinações possíveis de pares dos respondentes. Assim, se há J indivíduos que participam desta prova, tem-se $\frac{J(J-1)}{2}$ pares a serem analisados. Exemplificando ainda mais isto, considere-se o exame (de larga escala) ENEM e seja o $J = 500.000$ estudantes, portanto, ao realizar a combinação $\binom{J}{2}$, tem-se 124.999.750.000 pares a serem considerados na análise. Observe-se que é praticamente impossível realizar a análise de um J grande como essa e menos ainda de todos pares possíveis em tempo hábil, salvo em supercomputadores, pois vale lembrar que o ENEM apresentou milhões de participantes inscritos nos últimos anos. Assim, a pergunta é, qual é o número total de indivíduos para serem analisados no tempo hábil de uma prova de larga escala? Para responder essa pergunta é necessário conhecer as seguintes variáveis: (i) o tempo que leva um determinado computador para analisar um par de candidatos, e (ii) o tempo limite proporcionado pela entidade avaliadora. Portanto, determina-se o número total de pares $(p^{'})$ através da equação,

$$p^{'} = \frac{t_l}{t_0}, \tag{3.2}$$

em que t_l é o tempo limite para apresentar os resultados e t_0 é o tempo que leva um determinado computador em realizar a análise de um par suspeito de fraude, na mesma unidade de medida.

Assim, com o resultado da quantidade de pares da Equação 3.2 se pode determinar o número de indivíduos (n_i) que se pode analisar para um certo tempo limite, através da seguinte equação:

$$n_i = \frac{1 + \sqrt{1 + 4p^{'}}}{2} \simeq \sqrt{p'}, \tag{3.3}$$

nota-se que os conjuntos J e n_i cumprem com a relação $J \geq n_i$.

As próximas seções e o capítulo seguinte deste trabalho abordam propostas com o objetivo de diminuir a quantidade de pares a serem analisados. Essas propostas são baseadas

nas proficiências, também chamada habilidades, ao invés dos escores observados usado por Souza [13].

3.4.1 Escore mínimo

Se não há condições de analisar todos os candidatos de um teste de larga escala como dito na Seção 3.4, seja por motivos do tempo limite ou por não ter os equipamentos necessários ou por outros motivos, é recomendado utilizar pontos de cortes para dizer quais indivíduos serão analisados. Neste trabalho, adotar-se-á limiares sobre uma escala de habilidade, denotado, H_m para ter subconjuntos de participantes que possam ser analisados num tempo hábil.

Tomando o exemplo do ENEM da Seção 3.4, ele é utilizado para obter vagas nas melhores universidades do Brasil, o qual implica que há um desejo grande de alguns indivíduos para usar meios ilícitos e, assim, conseguir o objetivo de acessar a essas. Mas, para obter excelentes resultados desta prova que permita seu acesso os candidatos pagam uma determinada quantia a quadrilhas especializadas em esquemas de fraude para receber as respostas de cada itens, sendo estas geralmente fornecidas por comunicação eletrônica através de um indivíduo (ou mais) com alta proficiência e pertencente a quadrilha. Dada essa premissa é esperado que as habilidades estimadas para esses candidatos sejam altas, pois como foi dito as respostas provém de um individuo de alta proficiência. Portanto, é intuitivo que indícios de fraudes recaiam sobre examinados com altas habilidades. Por isso, foi proposto neste trabalho limiares nas altas habilidades de tal forma que abaixo desses pontos de cortes o examinado não entra no processo de detecção de fraude, gerando um número de examinados menor e, por consequente, o número de pares também diminui.

Estimativas para H_m depende fortemente da distribuição das proficiências que aqui serão usado, e, portanto, do número de itens e do comportamento geral da distribuição. Estimativas preliminares podem ser obtidas com base em resultados de simulação, ressaltando que em aplicações reais é fundamental que a etapa preliminar dedique-se a estimar tal distribuição para haver estimativas de H_m mais apropriadas.

3.4.2 Quantis

Para a seleção dos limiares serão utilizados quantis de uma distribuição de X. Segundo Milone [9] o quantil é uma medida de posição que divide X em partes iguais. Também,

diz que os quantis fomentam cortes nos dados. Essas medidas podem ser utilizadas para converter variáveis quantitativas em intervalares, estabelecer quantidades ou porcentagens de itens menores e maiores que certo referencial, calcular assimetria das distribuições e definir o centro de dados com média não definível, entre outras. Por suas características, o quantil só é definível para X ordenado crescente ou decrescente, além disso, não é afetado pelos extremos do conjunto, como acontece com a média ou variância (portanto, pode ser calculada para séries abertas), nem é algebricamente manipulável, o qual implica que não se pode obter o quantil de X dos quantis de seus subconjuntos. O total de partições é um número arbitrário definido pelo usuário, em função de objetivos e interesses específicos. Formalmente a definição dos quantis e dada por:
Se q é o número de partições de no conjunto X, de tamanho N, então $q \in \mathbb{Z}_+^*$ e $1 < q < N$. Com isso, se S_r é o quantil de ordem r de X, então $1 \leq r \leq q-1$.
Isso implica que o número de partições é um inteiro positivo maior que a unidade e menor que o total de dados, e que os quantis só são totalmente determináveis para conjuntos com mais de q elementos (se $q \geq N$, as primeiras e os últimos quantis fazem referência a valores fora do campo de definição de X). S_r é perfeitamente determinável para conjuntos constituídos de variáveis quantitativas ou por um número conveniente de elementos. Se X está agrupado em uma distribuição de frequências acumuladas, S_r é tal que:

Tabela 3.1 *Quantil de ordem r de X acumulado em uma distribuição*

Absoluta	Relativa (%)
$\sum_{i=1}^{S_r} F_i = N\frac{r}{q}$	$\sum_{i=1}^{S_r} P_i = 100\frac{r}{q}$

sendo r é a ordem do quantil, q é o número de partições e N é o total de pontos de X.

Se X é um vetor ordenado, então

$$S_r = x_{[pos_r]} = x_{[Zpos_r]} + Fpos_r(x_{[Zpos_{r+1}]} - x_{[Zpos_r]}), \tag{3.4}$$

no qual $x_{[i]}$ representa o i-ésimo elemento de X, $pos_r = \frac{r}{q}(N+1)$ define a posição de S_r em X, $Zpos_r$ e $Fpos_r$ as partes inteira e fracionária de pos_r, respectivamente.

No caso contrario, se for uma variável contínua, é dizer, $X = \{x| \in \mathbb{R}\}$, é dado por

$$\int_{L_i}^{S_r} p(x)dx = \frac{r}{q}, \tag{3.5}$$

onde L_i é o limite inferior de X e $p(x)$ é a função de frequência (ou probabilidade). Dada a definição anterior, utilizar-se-á para determinar limiares em conjuntos contínuos como é o caso das habilidades.

3.5 Tempo de processamento

Qualquer entidade pública ou privada (como bancos, universidades, etc) que faz processos seletivos através de provas para ocupar determinadas vagas (acadêmicas, emprego, entre outras) têm em comum um tempo limite para apresentar os resultados dos candidatos nesse processo. Por isso, é fundamental ter um controle sobre o mesmo para cumprir com os cronogramas estabelecidos e não gerar dúvidas do processo seletivo. Além disso, com esse controle do tempo pode-se eliminar do processo os indivíduos fraudulentos, se for o caso para não prejudicar aos honestos. A fraude aqui depende da importância das vagas que determinada entidade oferece, pois no caso do ENEM, onde os estudantes desejam ingressar às melhores ou importantes universidades do Brasil é um lugar adequado para numerosos fraudes, como já verificado em anos anteriores.

O tempo de processamento está associado ao número de indivíduos. Por tal motivo, a questão é, se tem-se um número J de candidatos em uma prova de larga escala, como determinar o tempo de processamento para analisar a quantidade de pares suspeitos de fraudes? E aí pode-se dizer que depende de variáveis como o tempo limite da entidade, dos sistemas computacionais aos quais se tem acesso, ao tempo que leva examinar um par suspeito de fraude e a quantidade de indivíduos desejados para examinar em relação se fizeram copia.

Assim, se tem-se t_0, o tempo de processamento (em segundos) que leva um determinado computador para executar a análise de um par suspeito de fraude, então precisa-se conhecer o tempo total de processamento ao analisar J indivíduos que fizeram uma prova, com o objetivo de determinar os possíveis candidatos fraudulentos, e pode-se utilizar a seguinte equação,

$$t_t = \frac{n_i(n_i - 1)}{2} \times t_0 \tag{3.6}$$

Essa equação permite dizer se a análise pode ser realizada no tempo limite que tem a entidade. No entanto, no caso que não seja possível, pode-se diminuir o número de indivíduos através de filtros e cumprir com o cronograma de publicação dos resultados sem deixar de analisar os possíveis copiadores para excluí-os do processo.

No próximo capítulo são apresentados os procedimentos para a diminuição de pares a serem analisados, o controle das taxas de *falsos positivos* e as novas alternativas do pacote *TestFraud*.

Capítulo 4

Seleção quantílica e estatística de teste

4.1 Seleção quantílica e quantitativos

Neste trabalho utilizou-se dentro do pacote *TestFraud* a função *quantile* do programa R para escolher os quantis ótimos para ser aplicável em tempo hábil os métodos da Seção 2.2 às avaliações de larga escala. Note-se que sem importar quantos candidatos fizeram um teste, essa função *quantile* vai permitir selecionar o número de indivíduos desejados para a análise, pois ela ordena o conjunto de dados e determina a posição a partir de onde inicia o subconjunto de dados escolhidos. Para exemplificar melhor essa situação, suponha que há N indivíduos e adota-se o quantil $q \in [0, 1]$, assim, serão selecionados $n_i = N(1 - q)$ deles para avaliação de fraudes, de forma que $\binom{n_i}{2}$ pares ficaram para avaliação final. Neste estudo serão adotados quantis dependendo do tempo limite e da quantidade de candidatos que se deseja analisar no ENEM, de forma a cumprir o cronogramas de entrega dos resultados da pesquisa para a entidade solicitante.

4.2 Estatística T^* ponderada

Nos sete índices apresentados anteriormente tem-se que nem sempre há concordância nos pares considerados suspeitos, devido a formulação independente de cada um. Dessa forma, propõem-se neste estudo uma estatística T^* ponderada que é função de todos os índices. Ela será obtida para cada par (c, s) de indivíduos, que será suprimindo na notação abaixo. Assim, definimos T^* como

$$T^* = \sum_{t=1}^{\tau} \rho_t \lambda_t, \tag{4.1}$$

sendo $t = 1, 2, ..., \tau$, número de índices considerados e λ_t

$$\lambda_t = \begin{cases} 1, & \textit{par com detecção de fraude no índice } t \\ 0, & \textit{c.c.} \end{cases} \tag{4.2}$$

o ρ_t é o peso do t-ésimo índice, com $\sum_{t=1}^{\tau} \rho_t = 1$ e será denotado t^*_α como o ponto de corte para decidir aos possíveis pares que copiaram.

Para o cálculo dos pesos tem-se a seguinte formulação:

$$\rho_t = \frac{f_t}{\sum_{t^*=1}^{\tau} f_{t^*}}, \tag{4.3}$$

onde $f_t = f_t(\alpha)$ é fator de ajuste em relação ao nível de significância α adotado. Logo, este fator é obtido por

$$f_t = \frac{1}{|\alpha - \hat{\alpha}_t|}, \tag{4.4}$$

em que α é o nível de significância nominal adotado nos índices e $\hat{\alpha}_t$ é a estimativa de detecção dos mesmos, dado por

$$\hat{\alpha}_t = \frac{1}{n} \sum_{i=1}^{n} \lambda_{it}, \tag{4.5}$$

sendo n o total de pares analisados e λ_{it} a indicadora do índice. Além disso, tem-se $\hat{\alpha}_t$ é a estimativa de erro tipo I observado no índice t, ou seja, o nível de *falso positivo* (FP), isto é, considerar dois indivíduos como fraudadores quando na realidade não são. Também é interessante notar que os índices que têm valores de $\hat{\alpha}$ próximo de α gera erros menores e por conseguinte, têm pesos maiores para o cálculo da estatística T^* ponderada.

Ao finalizar o análise de um conjunto de pares com a estatística T^* ponderada, será determinada a qualidade do ajuste das taxas de FP estimados, com relação aos níveis de erro Tipo I adotados. Para isso, será utilizada a soma do quadrado das diferenças (SQD) dos níveis de significância, tanto estimado como adotados. A SQD é dada por:

$$\Delta = \sum_{i=1}^{\gamma^*} (\alpha_i - \hat{\alpha}_i)^2, \tag{4.6}$$

onde γ^* representa o número dos níveis de significância adotados para essa análise, α e $\hat{\alpha}$ são os níveis de erro Tipo I adotado e estimado, respectivamente.

4.3 Adaptação no Pacote *TestFraud*

Este trabalho procura tornar aplicável os métodos da Seção 2.2 sobre provas de larga escala. Por isso, há necessidade de otimizar o pacote *TestFraud* desenvolvido para fazer processamento em paralelo por Souza [13]. A otimização será realizada através da seleção quantílica na distribuição dos candidatos com as maiores habilidades. Neste estudo serão usados os percentis 4% 5%, 7%, 8%, 10%, 12,5%, 15% e 20% nos dados simulados, no entanto, como a base de dados reais é maior à simulada, se utilizaram os percentis 1%, 2%, 3%, 4% e 5%, pois eles permitem controlar (conhecer) o tempo de processamento e taxas de *falso positivo* necessários para analisar um número determinado de indivíduos. O pacote *TestFraud* avalia a similaridade e cópia de respostas entre dois examinados a partir de dados provenientes de respostas de testes de múltipla escolha, os quais podem ser inseridos em sua forma original ou dicotomizada. Para avaliar o erro Tipo I Souza [13] apresentaram uma variável T discreta, representando o número de índices que apontam indícios de fraude, fixado um nível de significância α. Com a variável T, também foi utilizada sua indicadora, definida por:

$$T(t) = \begin{cases} 1, & \textit{par com detecção de fraude no índice } t \\ 0, & \textit{c.c.} \end{cases} \tag{4.7}$$

Souza [13] tomou como critério considerar os candidatos como possíveis fraudulentos quando a variável T apontar 4 ou mais índices de fraude ($T \geq 4$), sem esperar necessariamente um retorno idêntico ao α adotado. A Tabela 4.1 reproduz a probabilidade de não cometer o erro Tipo I obtida pelo autor. No entanto, ressalta-se que neste estudo será adotada a nova proposta da estatística T^* ponderada definida na Seção 4.2.

Tabela 4.1 *Distribuição acumulada da estatística T*

	T						
α	1	2	3	4	5	6	7
0,001	0,99841	0,99958	0,99987	0,99994	0,99996	0,99998	0,99999
0,005	0,99200	0,99714	0,99895	0,99932	0,99961	0,99981	0,99992
0,010	0,98413	0,99347	0,99732	0,99815	0,99883	0,99942	0,99977
0,020	0,96841	0,98501	0,99312	0,99498	0,99659	0,99822	0,99920
0,050	0,92146	0,95489	0,97646	0,98162	0,98596	0,99218	0,99585

A estatística T assume apenas valores discretos: 0, 1, 2, 3, 4, 5, 6, 7. A estatística T^* assumirá valores decimais no intervalo $[0, 1]$, dependendo dos pesos de cada índices.

No próximo capítulo são apresentados resultados de simulação para analisar o desempenho das mudanças realizadas para a melhoria do pacote *TestFraud*.

Capítulo 5

Resultados

Quando há varias alternativas para realizar um determinado processo, deve-se selecionar um ou mais de acordo com algum critério que otimiza os resultados. Neste estudo considerou-se três alternativas de análise para a seleção dos indivíduos que serão avaliados, referidas como *Variáveis de Seleção*: (*i*) escore observado, (*ii*) escore verdadeiro, e (*iii*) proficiência ou habilidades. Além disso, deseja-se otimizar e controlar o tempo de processamento e quantidades de indivíduos a analisar num teste deste mesmo tipo. Este último deve-se à utilização dos quantis ou limiares sobre a variável de seleção escolhida. Também, pode selecionar os candidatos que se acredita haver indícios de fraude, visando eliminá-los do processo ao qual estão participando.

A taxa de *falso positivo* (FP) apresentada neste trabalho retorna melhores resultados que os determinados por Souza [13], devido à proposta da estatística T^* ponderada apresentada na Seção 4.2. Isto contribui enormemente para evitar de acusar de fraude algum candidatos honestos, controlando de forma mais eficiente o erro Tipo I.

Portanto, várias modificações foram realizadas a fim de ter o controle do processo, sendo a seleção quantílica, complementada pela estatística T^* ponderada e a seleção do conjunto de dados a analisar, as principal mudança propostas neste trabalho. As seções a seguir mostram com maiores detalhes as principais modificações ou amplificações desses três aportes no pacote *TestFraud* desenvolvido por Souza [13].

Máquina de teste

Para todos os resultados obtidos na próxima seção utilizou-se um computador com processador *AMD Ryzen 7 2700*, que possui 8 núcleos físicos com capacidade de executar 16 *threads*, ou seja, possui capacidade de emular 16 núcleos (físicos e lógicos), e opera à frequência de 3.2 Ghz (Max Turbo 4.1 GHz), com 32 GB de memória RAM, Cache L3: 16MB, Cache L2: 4MB, Potência: 65 W. Utilizou-se o sistema operacional Windows 10 Pro 64 bits. Para medir o tempo de execução dos processos utilizou-se o pacote *microbenchmark*-[8].

5.1 Seleção quantílica no pacote *TestFraud*

A seleção quantílica no pacote *TestFraud* é muito útil quando o objetivo é controlar o tempo e saber a quantidade de indivíduos a serem analisados. Sobre a variável de seleção, adotou-se as proficiências, diferentemente do adotado por Souza [13], que utilizou o escore observado, com o uso de um escore mínimo para seleção de indivíduos. Uma vantagem em trabalhar com as habilidades é que esta assume valores contínuos, enquanto a diferença de escores observados assume valores discretos. Assim, as principais atribuições realizadas ao pacote *TestFraud* foram criar uma função chamada *Simula_deteccao* onde o Escore mínimo (E_{min}) de Souza [13] mudou para quantis. O tipo de escore a ser utilizado pode ser: os escores observados, esperados ou as habilidades.

A Figura 5.1 apresenta a implementação da função *Simula_deteccao* com os valores que seriam atribuídos se não se preenche nada nela. Ela só acrescentou as opções para determinados argumentos, por exemplo, no argumento de *type_sco*, tem-se a oportunidade de selecionar as três variáveis de seleção citadas anteriormente, denotadas por *true_sco* (escore verdadeiro), *score_TCT* (escore observado) e *score_theta* (proficiências), respectivamente. No argumento *sco_min* (linha 27) no lugar de escolher um valor das notas como foi realizado por Souza [13] é permitido selecionar um quantil específico. Seguidamente estão as funções de gerar os dados simulados, a serem abordadas com maior detalhes nas próximas seções.

Figura 5.1 *Funções Default no pacote TestFraud para a seleção quantílica*

```
27  Simula_deteccao <- function(J = 500, I = 45, s = 5, dif_sco = 45, sco_min = 0.95,
28                              sig = 0.001, semente = 2, typ="A", Numero_de_pares = "ALL", type_sco="true_score") {
29
30    #Definindo o numero de cores que serao usados (numero de cores total -1)
31    c1 <- makeCluster(detectCores()-1)
32    #Registrando que ira usar processamento paralelo conforme o numero de cores em c1
33    registerDoParallel(c1)
34    #%%%%%%%%%%%%%%%%%% Funcao para gerar parametros de itens (MRN) %%%%%%%%%%%%%%%
35    GeraItens=function(I,s){...}
45    #%%%%%%%%%%%%%%%%%% Funcao para gerar habilidades N(0,1) %%%%%%%%%%%%%%%%
46    GeraTheta <- function(J){...}
50    #%%%%%%%%%%%%%%%%%% Funcao para gerar respostas politomicas (MRN) %%%%%%%%%%%%%%%%
51    #ipar sera o objeto retornado pela funcao GeraItens
52    #theta sera o objeto retornando pela funcao GeraTheta
53    #typ defince se a saida das respostas sera Numerica ou Alfabetica (typ="A")
54    #configurado apenas para s=5<-------------
55    Gera_Dados_Nominal <- function(ipar, theta, typ="A"){...}
77    #%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
78    #%%%%%%%%%%%%%%%%%%%%%%%% Finalizacao das funcoes de geração de dados %%%%%%%%%%%%%%%%%%
79    #%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
80
81
82    #%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
83    ##################### Função para deteccao de fraudes (Fraud.Indices) ##################
84    #%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
85
86    #data: Matriz de repostas (nao dicotomizada)
87    #item.par: Matriz dos parametros do MRN
88    #pair: vetor de tamanho 2, (copiador, fonte) [NESTA ORDEM]<...
265   ################################### Gerando dados ##################################
266   #%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
267
268
269   #fixando a semente
270   set.seed(semente)
```

Na linha 298 da Figura 5.2 apresentam-se as estimativas das habilidades segundo Modelo Logístico de 3 parâmetros e é utilizado o pacote *mirt*. Além, na linha 301 se calcula o escore observado para cada indivíduo que realizou o teste. Da linha 304 até 314 se encontra o calculo do escore esperado ou verdadeiro, usando a equação dada por [1]. Para as linhas 318 até 324 estão as condições das opções do argumento *type_sco* com que se deseja trabalhar e as seleções quantílicas para cada uma dessas. Na linha 326 só está chamando *score_maiores* ao quantil escolhido do tipo de escala usado e por último A Figura mostra na linha 328 a criação das combinações entre os candidatos selecionados.

Figura 5.2 *As Três opções do argumento type_sco da função Simula_deteccao*

```
295  cat("\nEstimativas dos parametros dos itens (3PL): \n");   print(round(ipar.dic,4))
296
297  #habilidades
298  theta_ml3 = fscores(mirt.3PL)# theta estimado (ML3)
299
300  #calculando escores dos individuos
301  score.indiv <- rowSums(scored.data, na.rm = T) #(TCT)
302  #calculando escore esperado (true escore)
303  #ml3
304  p_1 <- function (z_i,th){
305    z_i[3]+((1-z_i[3])/(1+exp(-z_i[1]*(th-z_i[2]))))#ml3 c=z_i[3], b=z_i[2], a=z_i[1]
306  }
307  true_score = rep(NA,J)
308  for (j in 1:J){
309    s=0
310    for (i in 1:I){
311      s=s+p_1(z_i=ipar.dic[i,],th=theta_ml3[j])
312    }
313    true_score[j]=s
314  }
315  cat("Concluida a geracao do gab(key), scored.data e theta\n")
316  #selecionando os maiores escores (sco_min)
317  #type_sco="true_score" ou escore (TCT) type_score="score_TCT"
318  if (type_sco=="true_score"){
319    aux_sco=true_score >= quantile(true_score,sco_min)
320  } else if (type_sco=="score_TCT"){
321    aux_sco=score.indiv >= quantile(score.indiv,sco_min)
322  } else if (type_sco=="score_theta") {
323    aux_sco=theta_ml3 >= quantile(theta_ml3,sco_min)#<----------------------------
324  }
325  #
326  score_maiores=(1:J)[aux_sco]
327  #Gerando os pares
328  pairs <-  as.data.frame(t(combn(score_maiores,2)))
329  #pairs <-  as.data.frame(t(combn(J,2)))  <----------- para gerar todos os pares
```

A Figura 5.3 apresenta nas linhas 341 até 347 a forma de fazer os pares, de acordo com a variável de seleção escolhida (type_sco). As linhas 348 até 380 não serão comentadas porque são as linhas da função *Fraud.Indices*, desenvolvida por Souza [13] (para mais detalhe ver [13]). Na linha 382 se apresenta o exemplo de simulação de teste de múltipla escolha (s = 5 alternativas) com 45 itens para 20.000 indivíduos que se encaixa no modelo do ENEM. Ainda, se selecionam os 5% (seleção quantílica de 0,95) dos candidatos com maior habilidades (note-se que o *type_sco* é *score_theta*), se adota um nível de significância de 0,001 para identificar os pares suspeitos de fraudes e se fixa a semente igual a 2, arbitrariamente.

Figura 5.3 *As Três opções do argumento type_sco da função Simula_deteccao*

```
#aux2=score.indiv[pairs[,2]]<score.indiv[pairs[,1]]
if (type_sco=="true_score"){
  aux2=true_score[pairs[,2]]<true_score[pairs[,1]]
} else if (type_sco=="score_TCT"){
  aux2=score.indiv[pairs[,2]]<score.indiv[pairs[,1]]
} else if (type_sco=="score_theta") {
  aux2=theta_ml3[pairs[,2]]<theta_ml3[pairs[,1]]#<--------------------------Perguntar qual theta é para usar
}
pairs[aux2,1:2]=pairs[aux2,2:1]
if(Numero_de_pares!="ALL") {aux3=sample(1:nrow(pairs),Numero_de_pares); pairs=pairs[aux3,]}
#**************************Finalizando filtros*********************************#
wc  <- ncol(scored.data)-rowSums(scored.data)
subgroups <- vector("list",(ncol(scored.data)+1))
lengths <- rep(0,(ncol(scored.data)+1))
for(j in 1:(ncol(scored.data)+1)){}
#pairs=pairs[1:1000,]#<------------------- (1000 primeiros apenas)
npairs=nrow(pairs);cat(npairs)
resultsCD22 <- rep(NA,7)
if(typ==1){} else {}
ti=Sys.time()
#utilizacao da funcao Fraud.Indices (com uso dos parametros verdadeiros do MRN e habilidades verdadeiras)
#colocar os estimados ou usar os verdadeiros mesmo <--------(Perguntar para o Prof. Heliton)
pairs[,3:10]=foreach(i=1:nrow(pairs),.combine = rbind)%dopar%{}
stopCluster(c1)
tf=Sys.time()-ti
print(tf)
return(list(pairs=pairs,responses=responses,key=gab,score.indiv=score.indiv,true_score=true_score,score_maiores=s
            scored.data=scored.data,theta=theta,theta_ml3=theta_ml3,theta_mrn=theta_mrn,wc=wc,
            subgroups=subgroups,lengths=lengths,ipar=ipar,ipar.dic=ipar.dic,parametros=parametros,tf=tf))
}

resultado <- Simula_deteccao(J = 20000, I = 45, s = 5, dif_sco = 45, sco_min = 0.95,
                             sig = 0.001, semente = 2, typ="A", Numero_de_pares = "ALL",type_sco = "score_theta")
```

Tabela 5.1 *Descrição do algoritmo da Estatística T^* ponderada*

Linha	Descrição
420	Adotou-se o mesmo nível de significância que foi colocado no argumento *sig* da função *Simula_deteccao*.
423	Separou-se as 8 colunas dos índices e da função indicadora de fraude (variável λ_t) das duas colunas dos pares (que são as duas primeiras).
425-432	Gerou-se a matriz indicadora de 0 e 1, dependendo se os índices detectaram fraude ou não, ou seja, se seus p-valores foram menores que o α adotado 0,001.
433-436	Somou-se a quantidade total de índices que detectaram fraude.
437	Determinou-se os pesos para cada índice segundo a Equação 4.3 da Seção 4.2
439	Calculou-se o fator de correção descrita na Equação 4.4 da mesma seção dita anteriormente.
441	Foi realizado a padronização dos pesos de acordo à Equação 4.5.
443-446	Realizou-se a soma dos produto dos pesos com a matriz indicadora de fraude como o representa a Equação 4.1.
447-450	Tomou-se os valores de t^*_α maiores que $quantile(T^*, 1-\alpha)+0,0001$ para imprimir o resultado do nível de significância estimado ($\hat{\alpha}$).

Figura 5.4 *Algoritmo da Estatística T^* ponderada*

```
#pares=read.table("pairs.csv")
##### Definiar alpha
alpha=0.001

#### Taxa de falso-positivo (FP)
pares_indices=paresF[,3:9]

matrix_ind=matrix(0,nrow(pares_indices),7) ### matriz de indicadores "0" ou "1"
for (i in 1:nrow(pares_indices)) {
  for (j in 1:7) {
    if (pares_indices[i,j] < alpha) {
      matrix_ind[i,j] = 1} else {
        matrix_ind[i,j] = 0}
  }
}
vetor=matrix(0,1,7) ### soma das colunas
for (j in 1:7){
  vetor[j]=sum(matrix_ind[,j])
}
tfp=vetor/nrow(pares_indices)  ### taxa de falso positivo (alpha chapeu)
#### Fator de correção
fc=1/abs(alpha-tfp)
### Padronização (soma do peso igual a 1)
p=fc/sum(fc)
##### Estatistica T
T=matrix(0,nrow(matrix_ind),1)
for (i in 1:nrow(matrix_ind)){
  T[i]=sum(matrix_ind[i,]*p)
}
hist(T)
T1=T[which(T>quantile(T,1-alpha)+0.0001),]
Tp=length(T1)/length(T)
Tp
```

5.2 Estudo das habilidades (H_m)

Com o objetivo de analisar as propostas apresentadas nos Capítulo 3 e 4 fez-se o uso de dados simulados tendo como cenário 20.000 indivíduos respondendo a 45 itens com 5 alternativas cada item, sendo as habilidades dos indivíduos simuladas a partir de uma $N(0,1)$ e os parâmetros dos itens do MRN a partir das seguintes distribuições: $\lambda_{iv} \sim N(0,1)$ e $\zeta_{iv} = 2\lambda_{iv} + X$, sendo $X \sim N(0, 10^{-2})$. As Tabelas 5.2 e 5.3 contêm os parâmetros dos 5 primeiros itens dos 45 simulados.

Tabela 5.2 *Parâmetros λ de inclinação para MRN*

Item (i)	λ_1	λ_2	λ_3	λ_4	λ_5
1	-1,0290	0,9930	0,7190	0,4390	-1,1220
2	-0,0570	-0,4000	0.9910	0,5360	-1,0700
3	0,9030	-0,2720	-1,0370	0,3420	0,0660
4	-0,5540	0,1080	-0,7140	1,7140	-0,5550
5	0,4620	-0,3210	-0,5410	-0,4990	0,8990

Tabela 5.3 *Parâmetros ζ de intercepto para MRN*

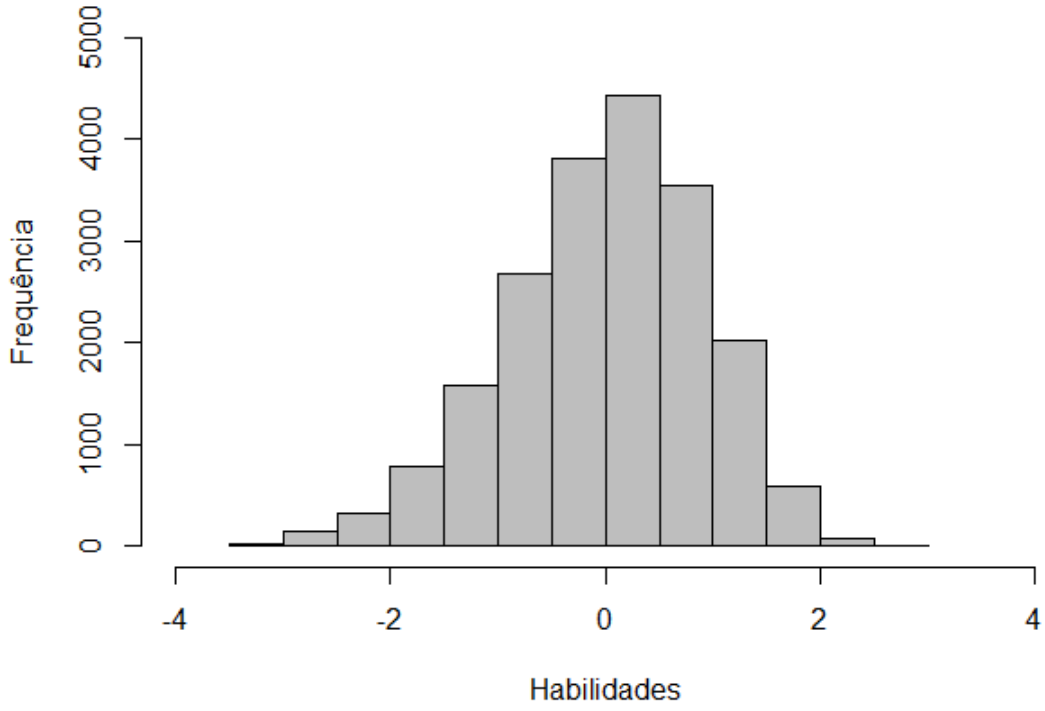
Item (i)	ζ_1	ζ_2	ζ_3	ζ_4	ζ_5
1	-2,0020	1,7220	1,4950	1,0430	-2,2590
2	-0,1590	-0,8410	2,0670	1,1790	-2,2460
3	1,9310	-0,5090	-2,0270	0,6830	-0,0780
4	-1,1280	0,1110	-1,4010	3,3960	-0,9780
5	1,1410	-0,6250	-1,1100	-1,1960	1,7900

As propostas dos Capítulos 3 e 4 foram baseados nas habilidades estimadas dos indivíduos através do modelo ML3. Nota-se que a escala utilizada neste estudo foi de -4 até 4. Na Figura 5.5 se apresenta a distribuição das habilidades dos candidatos, e verifica-se que a distribuição apresentou uma leve assimetria negativa, com valor de -0,4229. E portanto, a média que foi 0,0780 difere levemente da mediana que tem valor de 0,000001. Além disso, a Tabela 5.4 verifica-se que a menor proficiência foi -3,2957, a maior foi 2,7276, seu desvio-padrão foi de 0,9162 e seu curtose foi de 3,0491.

Tabela 5.4 *Estatística descritiva das habilidades dos indivíduos*

Mín	Q_1	Média	Mediana	Q_3	Máx	SD	Simetria	Curtose
-3,2957	-0,5791	0,0780	0,000001	0,6520	2,7276	0,9162	-0,4229	3,0491

Figura 5.5 *Histograma das Habilidades*



Dada a simulação da população de 20.000 candidatos, aplicou-se os limiares para reduzir o número de indivíduos notavelmente e controlar o tempo de processamento. Portanto, na Tabela 5.5, apresentam-se os tempos que a máquina usada levou nessa simulação e que foi descrita ao inicio deste capitulo. Utilizou-se a Equação 3.6 para obter o tempo por par em segundos.

5.3 Análise das Taxas de *Falsos Positivos* (FP)

Para controle do erro Tipo I (probabilidade de indicar fraude a determinados candidatos quando na verdade não ocorreu) nos conjuntos definidos por quantis, aplicou-se a estatística T^* ponderada definida na Seção 4.2 sobre a matriz identificadora de fraude e foi atribuído um peso a cada índice, tendo os maiores pesos os índices que retornaram os p-valores mais próximos ao α adotado e sempre que foram menores que este nível de significância (α). Neste trabalho se adotou os valores de significância (α) iguais a 0,001, 0,005, 0,01, 0,02, 0,03, 0,04 e 0,05 tanto para os dados simulados como os reais do ENEM.

Tabela 5.5 *Tempo de processamento usando a seleção quantílica para 20.000 indivíduos*

Quantis	J Analisar	Pares totais	Tempo total (s)	Tempo total (h)	Tempo por par
0.96	800	139.600	6.568,4560	1,8245	0,0205
0.95	1000	499.500	10.376,0900	2,8822	0,0208
0.93	1400	979.300	21.202,6800	5,8896	0,0216
0.92	1600	1.279.200	27.827,6100	7,7298	0,0217
0.90	2000	1.999.000	43.719,9800	12,1444	0,0218
0.875	2500	3.123.750	72.118,4800	20,0329	0,0230
0.85	3000	4.498.500	109.867,3000	30,5186	0,0244
0.80	4000	7.998.000	211,80000	58,8333	0,0264

Depois de introduzir a função *quantile* no pacote *TestFraud* se selecionaram oito quantis, 0,96, 0,95, 0,93, 0,92, 0,90, 0,875, 0,85 e 0,80 da população dos 20.000 candidatos e adotou-se diferentes níveis de significância. Com eles, foi possível calcular os α estimados ($\hat{\alpha}$) e os valores de t^*_α para cada uns deles. Também, determinou-se o quantil que melhor retornar as taxas de FP através da medida da Soma do Quadrado da Diferença (SQD ou Δ). Além disso, obteve-se a quantidade total de pares, o subconjuntos de indivíduos que ficaram para analisar.

Na Tabela 5.6 observe-se que depois de aplicar os quatro quantis 0,94, 0,95, 0,93 e 0,92 ficaram 800, 1.000, 1.400 e 1.600 indivíduos, respectivamente. Com essas quantidades de candidatos se realizou as combinações, $\binom{800}{2}$, $\binom{1.000}{2}$, $\binom{1.400}{2}$, $\binom{1.600}{2}$, obtendo os números totais de pares ($p^{'}$) iguais a 139.600, 499.500, 979.300 e 1.279.200 respectivamente. Além disso, os valores dos níveis de significância estimados ($\hat{\alpha}$) foram próximos aos esperado (ou adotados), tendo diferenças na sexta casa decimal entre cada conjuntos analisados e por conseguinte, geraram valores pequenos na SQD. Portanto, em virtude dos valores de Δ, pode-se afirmar que o quantil 0,92, o qual gerou $\Delta = 7 \times 10^{-12}$ retorna melhor as taxas de FP em comparação com os outros três primeiros quantis.

Na Tabela 5.7 observe-se que depois de aplicar os quatro quantis 0,90, 0,875, 0,85 e 0,80 ficaram 2.000, 2.500, 3.000 e 4.000 indivíduos, respectivamente. Com essas quantidades de candidatos se realizou as combinações, $\binom{2.000}{2}$, $\binom{2.500}{2}$, $\binom{3.000}{2}$, $\binom{4.000}{2}$, obtendo os números totais de pares ($p^{'}$) iguais a 1.999.000, 3.123.750, 4.498.500 e 7.998.000 respectivamente. Além disso, os valores dos níveis de significância estimado ($\hat{\alpha}$) foram próximos aos esperado

(ou adotados), tendo diferenças na sexta casa decimal entre os conjuntos analisados e, por conseguinte, geraram valores pequenos na SQD, com $\Delta = 7 \times 10^{-12}$. Portanto, os quatro quantis retornam as mesmas taxas de FP.

Tabela 5.6 *Resultados das taxas do FP nos quantis 0,96, 0,95, 0,93 e 0,92 para 20.000 candidatos*

J Analisar	Pares (p')	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
		0,001	0,0006	0,0009	
		0,005	0,0033	0,0049	
		0,01	0,0064	0,0099	
800	139.600	0,02	0,0130	0,0199	1×10^{-10}
		0,03	0,0195	0,0299	
		0,04	0,0253	0,0399	
		0,05	0,0308	0,0499	
J Analisar	Pares (p')	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
		0,001	0,0007	0,0009	
		0,005	0,0040	0,0049	
		0,01	0,0068	0,0099	
1000	499.500	0,02	0,0132	0,0199	$4,5 \times 10^{-11}$
		0,03	0,0203	0,0299	
		0,04	0,0260	0,0399	
		0,05	0,0309	0,0499	
J Analisar	Pares (p')	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
		0,001	0,0006	0,0009	
		0,005	0,0031	0,0049	
		0,01	0,0063	0,0099	
1400	799.300	0,02	0,0126	0,0199	$2,2 \times 10^{-11}$
		0,03	0,0191	0,0299	
		0,04	0,0262	0,0399	
		0,05	0,0349	0,0499	
J Analisar	Pares (p')	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
		0,001	0,0006	0,0009	
		0,005	0,0030	0,0049	
		0,01	0,0062	0,0099	
1600	1.279.200	0,02	0,0128	0,0199	7×10^{-12}
		0,03	0,0196	0,0299	
		0,04	0,0267	0,0399	
		0,05	0,0336	0,0499	

Tabela 5.7 *Resultados das taxas do FP nos quantis 0,90, 0,875, 0,85 e 0,80 para 20.000 candidatos*

J Analisar	Pares ($p^{'}$)	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
		0,001	0,0006	0,0009	
		0,005	0,0032	0,0049	
		0,01	0,0062	0,0099	
2000	1.999.000	0,02	0,0127	0,0199	7×10^{-12}
		0,03	0,0197	0,0299	
		0,04	0,0274	0,0399	
		0,05	0,0345	0,0499	
J Analisar	Pares ($p^{'}$)	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
		0,001	0,0006	0,0009	
		0,005	0,0033	0,0049	
		0,01	0,0062	0,0099	
2500	3.123.750	0,02	0,0128	0,0199	7×10^{-12}
		0,03	0,0202	0,0299	
		0,04	0,0275	0,0399	
		0,05	0,0339	0,0499	
J Analisar	Pares ($p^{'}$)	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
		0,001	0,0007	0,0009	
		0,005	0,0034	0,0049	
		0,01	0,0066	0,0099	
3000	4.498.500	0,02	0,0135	0,0199	7×10^{-12}
		0,03	0,0209	0,0299	
		0,04	0,0279	0,0399	
		0,05	0,0353	0,0499	
J Analisar	Pares ($p^{'}$)	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
		0,001	0,0007	0,0009	
		0,005	0,0052	0,0049	
		0,01	0,0061	0,0099	
4000	7.998.000	0,02	0,0163	0,0199	7×10^{-12}
		0,03	0,0223	0,0299	
		0,04	0,0262	0,0399	
		0,05	0,0320	0,0499	

Das Figuras 5.6 até 5.9, apresentam-se os valores esperados ou os níveis de significância adotado (α) das taxas de FP, representada pela linha de cor verde versus os valores observados ou estimados ($\hat{\alpha}$) das taxas de FP determinados pela estatística T^* ponderada e é representada pela linha de cor vermelha. Nota-se que graficamente não se pode afirmar que quantil retornar melhor as taxas de *falsos positivos*, porque todas apresentam as duas linhas sobrepostas e é impossível identificar diferenças ao observar-las. Por tal motivo, foi incluído nos gráficos as SQD, as quais permitem determinar o gráfico com o menor valor, e portanto, o melhor ajustado entre as taxas de FP adotadas e as estimadas. Isto permite ter um excelente conhecimento do erro Tipo I, porque com nesta estatística se pode determinar (aproximadamente) a quantidade de indivíduos acusados como *falsos positivos*, utilizando a Equação 3.3 e sabendo os níveis de α adotados, e o quantil selecionado.

Figura 5.6 *Taxas do Falsos Positivos dos quantis 0,96 e 0,95*

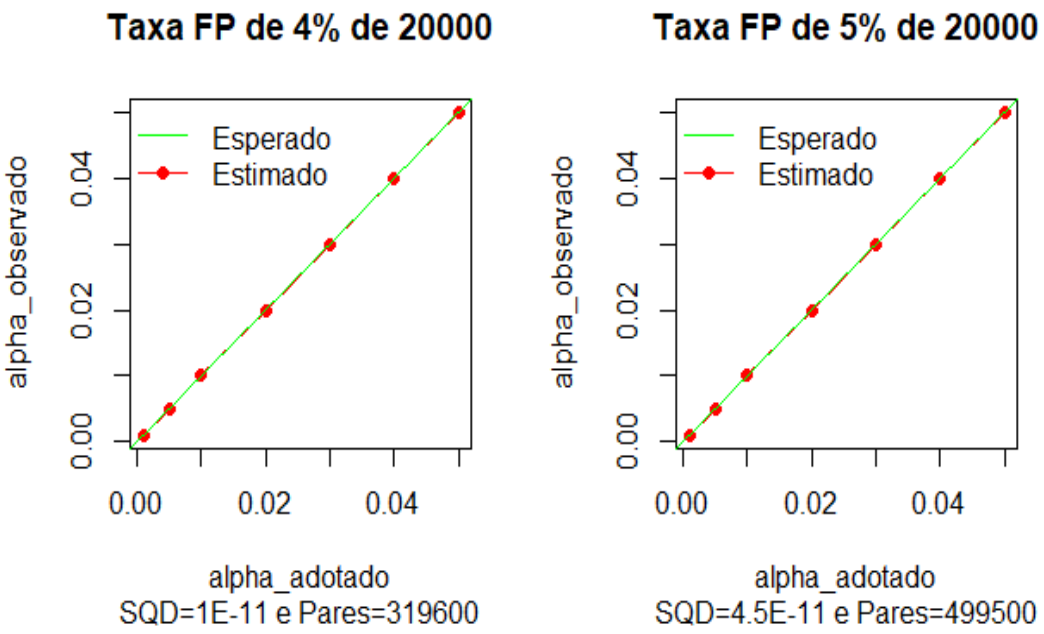


Figura 5.7 *Taxas do Falsos Positivos dos quantis 0,93 e 0,92*

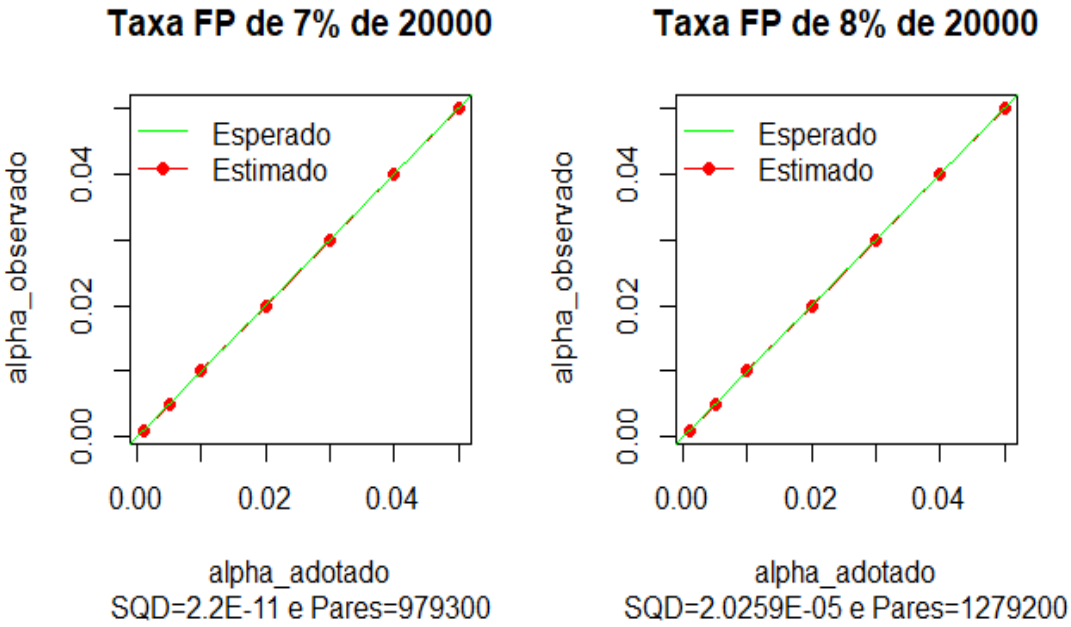


Figura 5.8 *Taxas do Falsos Positivos dos quantis 0,90 e 0,875*

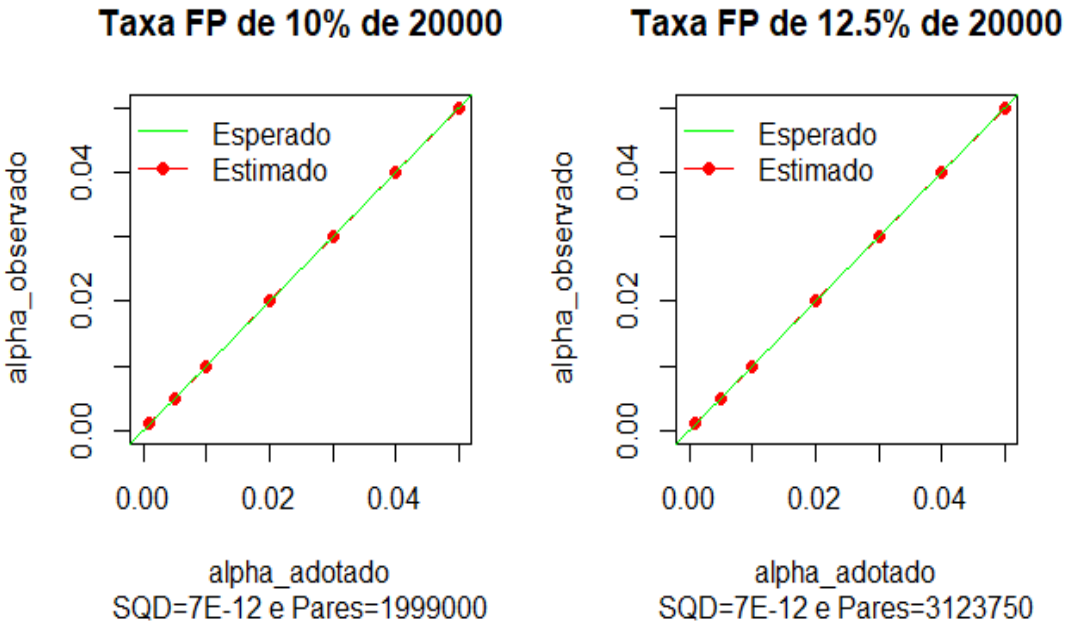
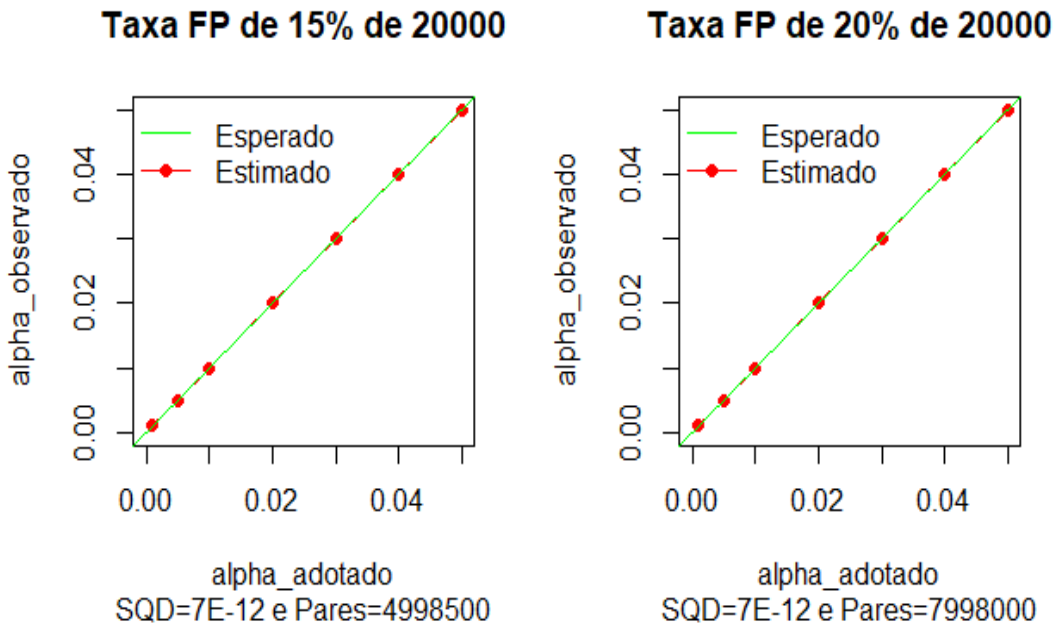


Figura 5.9 *Taxas do Falsos Positivos dos quantis 0,85 e 0,80*

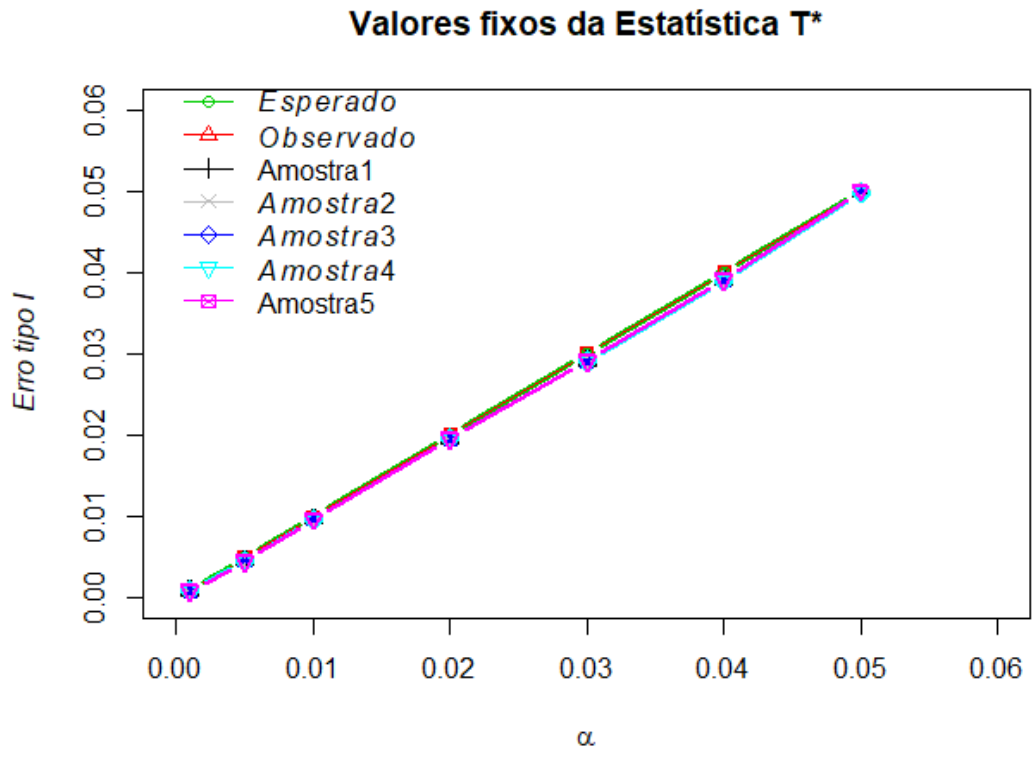


Para analisar melhor a estatística T^* ponderada, foram extraídas cinco amostras do conjunto de 3.123.750 pares (esse conjunto corresponde ao quantil 0,875), todas com tamanho iguais a 1.000.0000. Essas amostras foram selecionada sem substituição de pares e utilizando a função *Sample* de *R*. Assim, fixando os valores de cortes t^*_α do conjunto pertencente ao quantil 0,875 (ou o conjunto de 3.123.750 pares) e os níveis de significância adotados, estimaram-se os valores de $\hat{\alpha}$ e realizou-se a análise dos valores máximos, mínimos, média, desvio-padrão e Δ como se apresentam na Tabela 5.8. Nela se observa quão próximo estiveram as médias, os mínimos e os máximos de $\hat{\alpha}$, e o desvio padrão confirma, tendo medida muito pequena. Também, esses valores foram próximos aos α adotados, o qual gerou o resultado pequeno de Δ. Os valores das taxas de FP estimados para cada amostras (e os do conjunto total, chamados Observados) foram comparadas com os esperados na Figura 5.10, e se observou que existem minímas diferenças dado os valores de corte fixos de t^*_α. Portanto, pode-se afirmar que uma vez determinados os valores de t^*_α de um conjunto inicial se podem seguir utilizando esses mesmos valores nos subconjuntos deste sem prejudicar seus resultados, isto tem sentido quando se precisa estudar um modelo hierárquico, porque na primeira etapa se tem um conjunto maior às anteriores (e esses são subconjuntos do primeiro).

Tabela 5.8 *Estatística descritiva da Estatística T^* ponderada*

α	t^*_α	Mín	Máx	Média	SD	Δ
0,001	0,0006	0,0008	0,0009	0,0008	$1,9078 \times 10^{-5}$	
0,005	0,0033	0,0044	0,0045	0,0045	$3,8746 \times 10^{-5}$	
0,01	0,0062	0,0095	0,0096	0,0096	$5,2719 \times 10^{-5}$	
0,02	0,0128	0,0195	0,0195	0,0195	$1,7742 \times 10^{-5}$	$2,2258 \times 10^{-6}$
0,03	0,0202	0,0290	0,0291	0,0291	$5,9939 \times 10^{-5}$	
0,04	0,0275	0,0388	0,0391	0,0390	$1,0934 \times 10^{-4}$	
0,05	0,0339	0,0497	0,0500	0,0499	$1,4347 \times 10^{-4}$	

Figura 5.10 *Taxas do Falsos Positivos para diferentes amostras do quantil 0,875*



Segundo os resultados anteriores (as tabelas e gráficos), pode-se afirmar que está estatística T^* ponderada retorna as taxas de FP observados praticamente idênticos aos esperados ou adotados, permitindo o controle do erro Tipo I de forma excelente. Até a data não se conhecem resultados de simulação ou de dados reais que permitam controlar de forma eficaz o erro Tipo I como os obtidos neste estudo através desta estatística T^* ponderada.

5.4 Cenário ENEM 2018

A aplicação da metodologia da seleção quantílica em dados reais foi sobre o município de *Teresina do Estado de Piauí*, essa base de dados pertence ao ENEM e se encontram disponíveis no site do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), na página de microdados. A motivação de selecionar essa cidade foi o historial de fraudes que ali se apresentam neste tipo de provas. A quantidade de candidatos que realizaram o ENEM neste município para o ano 2018 ascendeu mais dos 37.000, no entanto, os indivíduos que tiveram presente nas quatro áreas (Matemáticas, Linguagens e Código, Ciências da Natureza e Ciências Humanas, cada uma com 45 itens) foram 37.194, o análise de todos esses candidatos com o pacote *TestFraud* sem utilizar a metodologia que este estudo apresenta (seleção quantílica) levaria uns três meses aproximadamente por área, e como são quatro áreas, o total seriam um ano de processamento, vale ressaltar que é assumindo que essa análise é realizado com o computador que este trabalho utilizou (ver Maquina de teste). Lembre-se que a seleção quantílica foi aplicada sobre a *variável de seleção* habilidades.

Nas Figuras de 5.11 a 5.14 e seus tabelas de estatísticas descritivas (5.9 até 5.12), observa-se que todas as distribuições das habilidades ou proficiências das áreas apresentaram assimetria direita. Sendo a prova de Ciências da Natureza a mais assimetria a direita com o coeficiente de 0,9457 e com a menor média das áreas (-0,1757). Além disso, essa área apresentou uma distribuição leptocúrtica (seu curtose foi de 0,5308) ao igual que a prova de Matemática, pois seu coeficiente de curtose foi de 0,2477, as provas de Linguagem e Ciências Humanas apresentaram distribuição platicúrtica com coeficiente de curtose de -0,6021 e -0,6077, respectivamente.

Figura 5.11 *Histograma das Habilidades na área de Ciências Humanas, ENEM 2018*

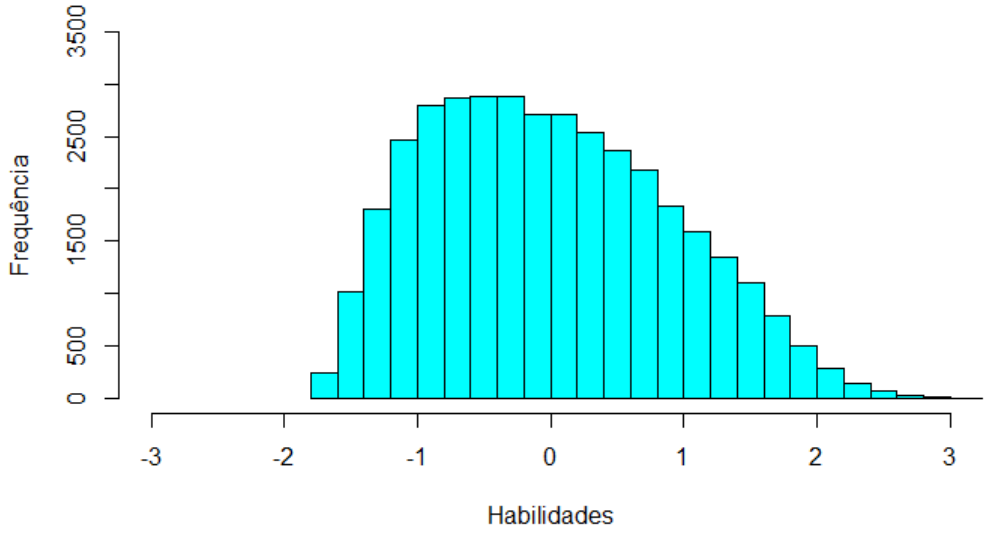


Tabela 5.9 *Estatística descritiva das habilidades dos indivíduos na área de Ciências Humanas e suas tecnologias, ENEM 2018*

Mín	Q_1	Média	Mediana	Q_3	Máx	SD	Simetria	Curtose
-1,7459	-0,7326	-0,0808	-0,000003	0,6503	3,3931	0,9069	0,3569	-0,6077

Figura 5.12 *Histograma das Habilidades na área de Ciências da Natureza, ENEM 2018*

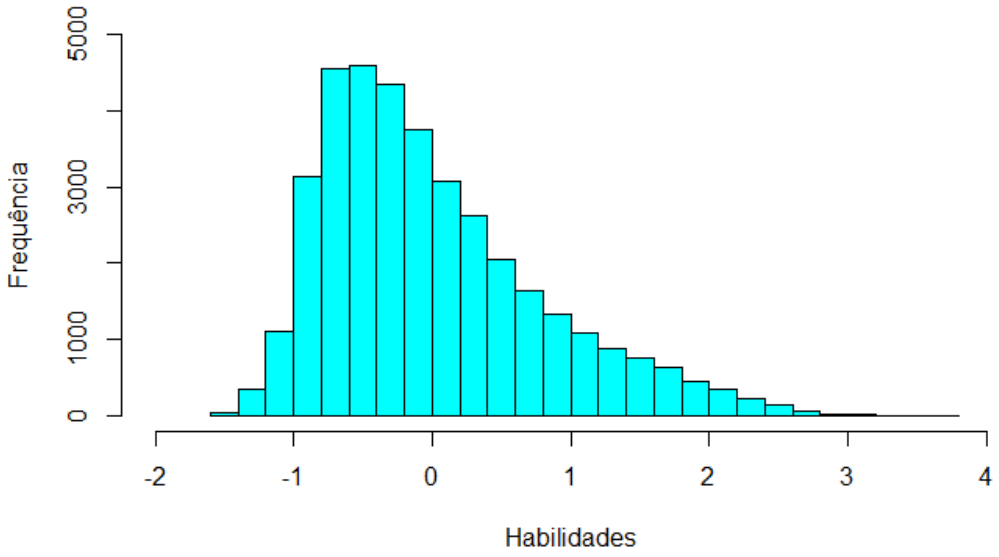


Tabela 5.10 *Estatística descritiva das habilidades dos indivíduos na área de Ciências da Natureza e suas tecnologias, ENEM 2018*

Mín	Q_1	Média	Mediana	Q_3	Máx	SD	Simetria	Curtose
-1,4982	-0,5943	-0,1757	-0,0001	0,4285	3,7368	0,7888	0,9457	0,5308

Figura 5.13 *Histograma das Habilidades na área de Linguagens e Códigos, ENEM 2018*

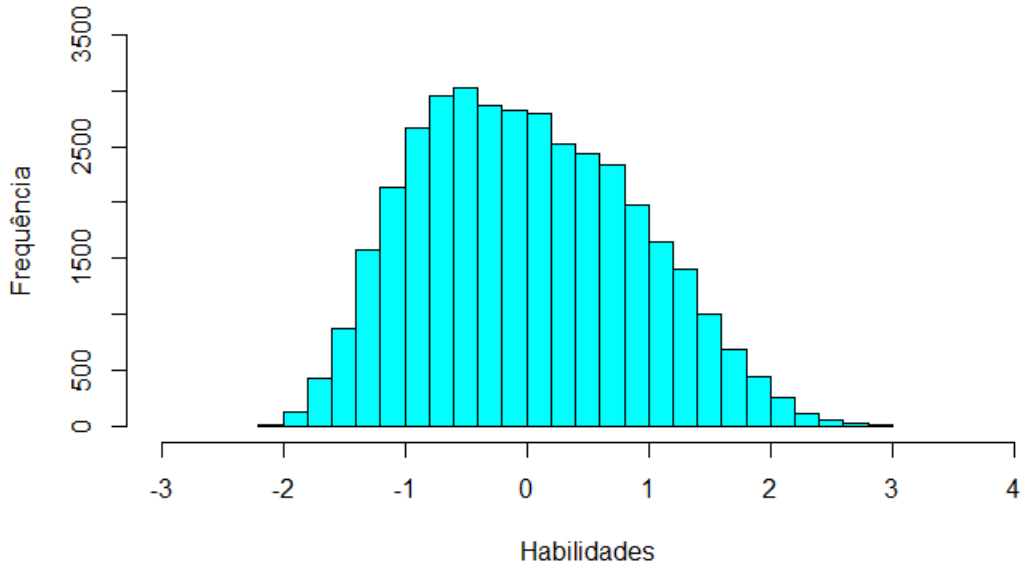


Tabela 5.11 *Estatística descritiva das habilidades dos indivíduos na área de Linguagens, Códigos e suas tecnologias, ENEM 2018*

Mín	Q_1	Média	Mediana	Q_3	Máx	SD	Simetria	Curtose
-2,0576	-0,7019	-0,0642	0,0000002	0,6525	2,9479	0,8911	0,2678	-0,6021

Figura 5.14 *Histograma das Habilidades na área de Matemáticas e suas Tecnologias, ENEM 2018*

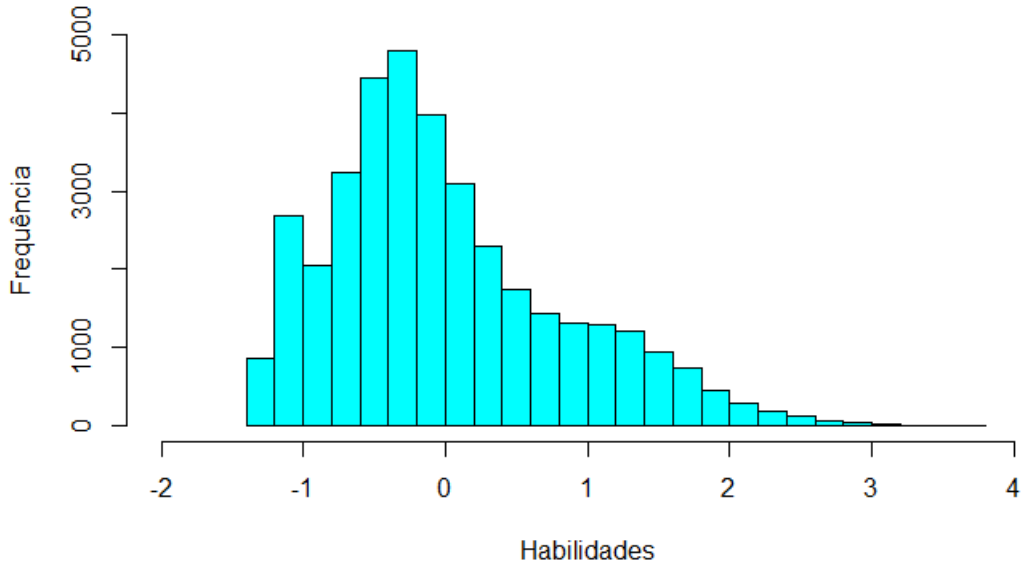


Tabela 5.12 *Estatística descritiva das habilidades dos indivíduos na área de Matemáticas e suas Tecnologias, ENEM 2018*

Mín	Q_1	Média	Mediana	Q_3	Máx	SD	Simetria	Curtose
-1,2817	-0,5775	-0,1739	-0,0001	0,4538	3,6966	0,8230	0,8131	0,2477

Na Tabela 5.13 se apresentam os cinco quantis utilizando sobre a base de dados do ENEM 2018 correspondente ao município de *Teresina-PI*, o tempo por pares, as quantidades de candidatos e os pares analisados. Além disso, o tempo por área em cada quantil e o total de cada uns desses. Portanto, em virtude da Tabela 5.13 se pode afirmar que em todos os quantis a área de Linguagens e Códigos foi a que menos tempo levou no processo de análise de possíveis copiadores. Também observamos que com está metodologia é possível ter o controle sobre o tempo de análise, pois lembremos que realizar a análise sobre toda a base de dados pode levar um tempo de aproximadamente um ano, porém, os quantis o mudam para horas ou semanas, dependendo quantos indivíduos se desejam incluir no processo de detecção de suspeitos de fraude.

Tabela 5.13 *Resultados de utilizar a seleção quantílica no pacote TestFraud sobre os dados de Teresina-PI nas provas do ENEM 2018*

Quantis	J Analisar	Pares totais	Áreas	Tempo (h)	Tempo total (h)	Tempo por par (s)
0,99	372	69.006	Matemáticas	0,5073	1,9723	0,0264
			Ciências da Natureza	0,5263		0,0274
			Linguagens e Códigos	0,4486		0,0234
			Ciências Humanas	0,4898		0,0255
0,98	744	276.396	Matemáticas	2,0887	8,0029	0,0272
			Ciências da Natureza	2,1136		0,0275
			Linguagens e Códigos	1,8109		0,0235
			Ciências Humanas	1,9896		0,0259
0,97	1116	622.170	Matemáticas	4,8741	18,4845	0,0282
			Ciências da Natureza	4,8791		0,0282
			Linguagens e Códigos	4,1510		0,0240
			Ciências Humanas	4,5801		0,0265
0,96	1488	1.106.328	Matemáticas	8,9105	33,8643	0,0289
			Ciências da Natureza	8,9871		0,0292
			Linguagens e Códigos	7,5833		0,0246
			Ciências Humanas	8,3832		0,0272
0,95	1860	1.728.870	Matemáticas	14,6146	54,3627	0,0304
			Ciências da Natureza	14,2542		0,0296
			Linguagens e Códigos	12,0075		0,0250
			Ciências Humanas	13,4864		0,0280

Lembre-se que a prova do ENEM contém quatro áreas (Linguagens e Código, Ciências da Natureza, Ciências Humanas e Matemáticas), com o objetivo de identificar a possíveis copiadores em três ou quatro áreas se realizou uma combinação de $\binom{4}{3}$ dando como resultados 4, é dizer que temos quatro grupos de três áreas e um só grupo de quatro áreas para analisar que pares foram suspeitos neles. Os grupos das áreas são:

1. Linguagens e Código, Ciências da Natureza e Ciências Humanas.
2. Linguagens e Código, Matemáticas e Ciências Humanas.
3. Linguagens e Código, Matemáticas e Ciências da Natureza.
4. Matemáticas, Ciências da Natureza e Ciências Humanas.
5. Linguagens e Código, Ciências da Natureza, Ciências Humanas e Matemáticas (as 4 áreas).

Para identificar os pares suspeitos de colar em qualquer dos cinco grupos anteriores se tomou arbitrariamente o quantil maior (0,95) e um nível de significância de 0,01. Obtendo os valores da Tabela 5.14, em que se apresentam os 28 pares detectados suspeitos da prova do ENEM para o ano 2018 correspondente ao município de *Teresina-PI*, distribuídos da seguente forma:

i) 17 pares encontrados nas áreas de Linguagens e Código, Ciências da Natureza e Ciências Humanas, o que é o mesmo que o grupo 1.

ii) 11 pares para as áreas de Linguagens e Código, Matemáticas e Ciências Humanas (grupo 2).

Para os outros grupos não foram determinados pares suspeitos (incluído das quatro áreas). Dos 28 pares apresentados na Tabela 5.14 só o indivíduo 99 teve duas repetições (como se pode observar no Gráfico 5.15), o que isso implica é que se têm 55 candidatos suspeitos de fraude, sendo o indivíduo 99 como maior probabilidade de haver realizado cola.

Tabela 5.14 *Pares detectados suspeitos de fraudes nos grupos 1 e 2 das áreas do ENEM 2018 para Teresina-PI*

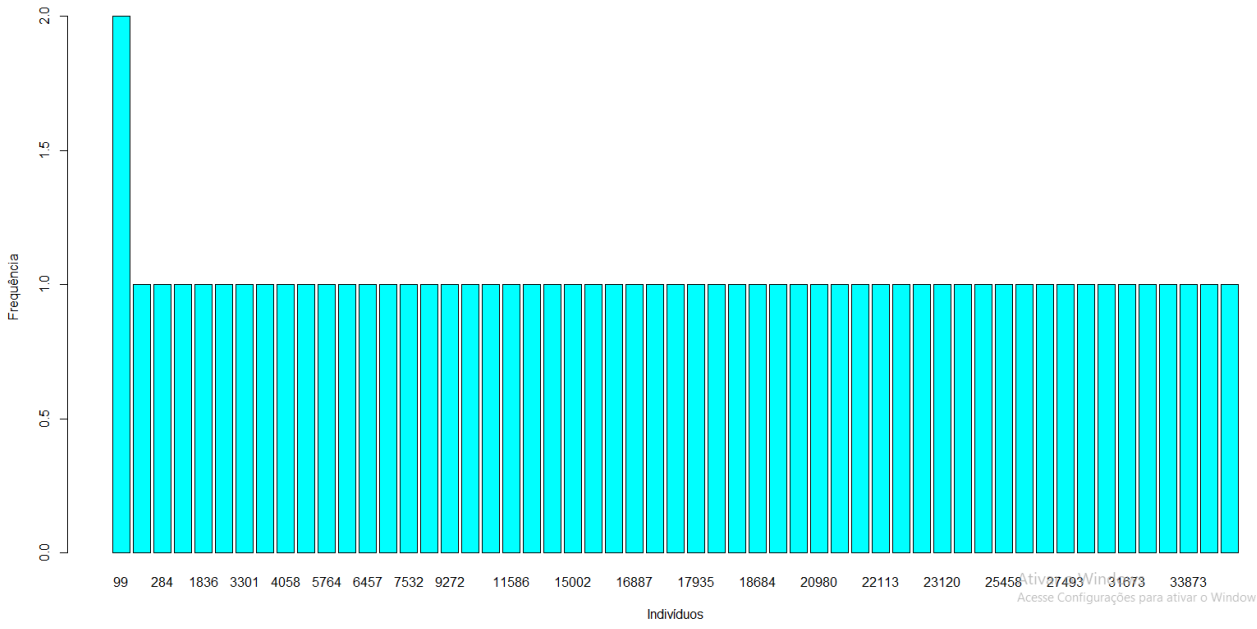
Pares detectados			
222	6084	99	3301
284	16887	99	10497
639	18684	1836	4058
3826	24313	2224	17403
4376	23120	5764	25702
6457	33168	6824	23044
7532	22949	8187	21137
9272	11761	15002	17935
10441	25458	15585	22113
11586	18472	18286	20980
11877	26867	21220	27493
16161	19040	23624	30585
17871	31673	27537	34304
20967	33873	33671	36490

Na Tabela 5.15 se apresentam as frequências dos candidatos identificados como suspeitos de fraude para uma, duas, três ou quatro áreas, vale ressaltar que a quantidade total de pares analisados foram 1.728.870.

Tabela 5.15 *Frequência dos pares detectados suspeitos de fraudes nas áreas do ENEM 2018 para Teresina-PI*

Número de Áreas	1	2	3	4
Frequência	49054	5937	28	0

Figura 5.15 *Frequência dos candidatos suspeitos de fraude nos grupos 1 e 2 das áreas do ENEM 2018 para Teresina-PI*



Nas Tabelas 5.16 até 5.20 observamos as taxas de detecção ($\hat{\alpha}$) de possíveis pares fraudulentos determinados pela estatística T^* ponderada dado seus correspondentes níveis de significância adotados (α), pode-se notar que seus valores diferem notavelmente em comparação como os resultados da simulação para essa estatística (Ver Subseção 5.3). No entanto, segundo os valores de SQD as áreas que estiveram mais próximas aos valores de α adotados foram Matemáticas e Ciências da Natureza. Além disso, se realizou uma média dos SQD para cada quantil, sendo que o quantil de 0,95 presentou o menor valor de todos (0,0002). Os quantis 0,99, 0,98, 0,97 e 0,96 apresentaram media de 0,0009, 0,0006, 0,0004

e 0,0003 respectivamente, isso implica que ele é o melhor ajuste da taxa de detecção em relação à esperada. Também se observam nessas tabelas os pontos de corte da estatística T^*, denotado por t^*_α, assim como a quantidade de candidatos analisados para cada quantil e seus respectivos número de pares.

Tabela 5.16 *Taxas de detecção dos candidatos suspeitos de fraude por área no quantil 0,99*

J Analisar	Pares ($p^{'}$)	Área	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
			0,001	0,0005	0,0007	
			0,005	0,0011	0,0043	
			0,01	0,0004	0,0098	
372	69.006	Matemáticas	0,02	0,0110	0,0199	0,0002
			0,03	0,7634	0,0215	
			0,04	0,6396	0,0306	
			0,05	0,8996	0,0396	
J Analisar	Pares ($p^{'}$)	Área	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
			0,001	0,0127	0,0001	
			0,005	0,0162	0,0014	
			0,01	0,0178	0,0039	
372	69.006	Linguagens e Códigos	0,02	0,0205	0,0096	0,0010
			0,03	0,0231	0,0156	
			0,04	0,0248	0,0226	
			0,05	0,0260	0,0305	
J Analisar	Pares ($p^{'}$)	Área	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
			0,001	0,0010	0,0002	
			0,005	0,0025	0,0023	
			0,01	0,0040	0,0053	
372	69.006	Ciências da Natureza	0,02	0,0064	0,0116	0,0006
			0,03	0,0082	0,0193	
			0,04	0,0102	0,0264	
			0,05	0,0119	0,0339	
J Analisar	Pares ($p^{'}$)	Área	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
			0,001	0,0048	0,00004	
			0,005	0,0070	0,0010	
			0,01	0,0090	0,0025	
372	69.006	Ciências Humanas	0,02	0,0124	0,0062	0,0019
			0,03	0,0151	0,0113	
			0,04	0,0176	0,0165	
			0,05	0,0202	0,0210	

Tabela 5.17 *Taxas de detecção dos candidatos suspeitos de fraude por área no quantil 0,98*

J Analisar	Pares ($p^{'}$)	Área	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
			0,001	0,4646	0,0003	
			0,005	0,3493	0,0027	
			0,01	0,2673	0,0066	
744	276.396	Matemáticas	0,02	0,2210	0,0164	$4,5379 \times 10^{-5}$
			0,03	0,1804	0,0264	
			0,04	0,1221	0,0381	
			0,05	0,0521	0,0497	
J Analisar	Pares ($p^{'}$)	Área	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
			0,001	0,0254	0,0001	
			0,005	0,0287	0,0014	
			0,01	0,0304	0,0039	
744	276.396	Linguagens e Códigos	0,02	0,0332	0,0092	0,0011
			0,03	0,0358	0,0149	
			0,04	0,0376	0,0212	
			0,05	0,0382	0,0296	
J Analisar	Pares ($p^{'}$)	Área	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
			0,001	0,0021	0,0005	
			0,005	0,0041	0,0033	
			0,01	0,0054	0,0072	
744	276.396	Ciências da Natureza	0,02	0,0073	0,0154	0,0001
			0,03	0,0083	0,0245	
			0,04	0,0093	0,0336	
			0,05	0,0103	0,0425	
J Analisar	Pares ($p^{'}$)	Área	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
			0,001	0,0113	0,0001	
			0,005	0,0138	0,0014	
			0,01	0,0160	0,0033	
744	276.396	Ciências Humanas	0,02	0,0193	0,0078	0,0014
			0,03	0,0217	0,0136	
			0,04	0,0236	0,0201	
			0,05	0,0253	0,0265	

Tabela 5.18 *Taxas de detecção dos candidatos suspeitos de fraude por área no quantil 0,97*

J Analisar	Pares ($p^{'}$)	Área	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
			0,001	0,1826	0,0005	
			0,005	0,1618	0,0035	
			0,01	0,1334	0,0082	
1116	622.170	Matemáticas	0,02	0,0676	0,0199	0,0001
			0,03	0,7842	0,0230	
			0,04	0,5464	0,0322	
			0,05	0,4208	0,0426	
J Analisar	Pares ($p^{'}$)	Área	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
			0,001	0,0403	0,0002	
			0,005	0,0419	0,0020	
			0,01	0,0426	0,0048	
1116	622.170	Linguagens e Códigos	0,02	0,0444	0,0110	0,0007
			0,03	0,0459	0,0177	
			0,04	0,0469	0,0248	
			0,05	0,0467	0,0337	
J Analisar	Pares ($p^{'}$)	Área	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
			0,001	0,0036	0,0006	
			0,005	0,0054	0,0040	
			0,01	0,0061	0,0085	
1116	622.170	Ciências da Natureza	0,02	0,0070	0,0179	$1,4481 \times 10^{-5}$
			0,03	0,0067	0,0279	
			0,04	0,0054	0,0385	
			0,05	0,0047	0,0489	
J Analisar	Pares ($p^{'}$)	Área	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
			0,001	0,0195	0,0001	
			0,005	0,0216	0,0018	
			0,01	0,0235	0,0042	
1116	622.170	Ciências Humanas	0,02	0,0262	0,0099	0,0009
			0,03	0,0280	0,0165	
			0,04	0,0293	0,0238	
			0,05	0,0305	0,0311	

Tabela 5.19 *Taxas de detecção dos candidatos suspeitos de fraude por área no quantil 0,96*

J Analisar	Pares ($p^{'}$)	Área	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
			0,001	0,1644	0,0005	
			0,005	0,1391	0,0038	
			0,01	0,1026	0,0089	
1488	1.106.328	Matemáticas	0,02	0,8185	0,0194	0,0001
			0,03	0,5104	0,0234	
			0,04	0,4153	0,0332	
			0,05	0,3560	0,0426	
J Analisar	Pares ($p^{'}$)	Área	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
			0,001	0,0556	0,0002	
			0,005	0,0553	0,0022	
			0,01	0,0551	0,0052	
1488	1.106.328	Linguagens e Códigos	0,02	0,0558	0,0117	0,000645226
			0,03	0,0566	0,0185	
			0,04	0,0567	0,0261	
			0,05	0,0558	0,0350	
J Analisar	Pares ($p^{'}$)	Área	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
			0,001	0,0051	0,0007	
			0,005	0,0057	0,0044	
			0,01	0,0046	0,0095	
1488	1.106.328	Ciências da Natureza	0,02	0,0025	0,0197	$8,39981 \times 10^{-5}$
			0,03	0,9417	0,0298	
			0,04	0,7338	0,0333	
			0,05	0,6259	0,0437	
J Analisar	Pares ($p^{'}$)	Área	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
			0,001	0,0284	0,0002	
			0,005	0,0294	0,0022	
			0,01	0,0305	0,0051	
1488	1.106.328	Ciências Humanas	0,02	0,0319	0,0119	0,0005
			0,03	0,0331	0,0191	
			0,04	0,0332	0,0276	
			0,05	0,0340	0,0355	

Tabela 5.20 *Taxas de detecção dos candidatos suspeitos de fraude por área no quantil 0,95*

J Analisar	Pares (p')	Área	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
			0,001	0,1604	0,0005	
			0,005	0,1323	0,0039	
			0,01	0,0829	0,0094	
1860	1.728.870	Matemáticas	0,02	0,6390	0,0144	0,0001
			0,03	0,4490	0,0236	
			0,04	0,3797	0,0333	
			0,05	0,3394	0,0427	
J Analisar	Pares (p')	Área	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
			0,001	0,0728	0,0003	
			0,005	0,0693	0,0025	
			0,01	0,0677	0,0057	
1860	1.728.870	Linguagens e Códigos	0,02	0,0669	0,0127	0,0004
			0,03	0,0671	0,0199	
			0,04	0,0657	0,028018	
			0,05	0,0638	0,0371	
J Analisar	Pares (p')	Área	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
			0,001	0,0069	0,0007	
			0,005	0,0060	0,0046	
			0,01	0,0027	0,0098	
1860	1.728.870	Ciências da Natureza	0,02	0,8687	0,0196	$3,1555 \times 10^{-5}$
			0,03	0,7052	0,0279	
			0,04	0,5342	0,0351	
			0,05	0,4016	0,0472	
J Analisar	Pares (p')	Área	α	t^*_α	$\hat{\alpha}$	SQD (Δ)
			0,001	0,0381	0,0003	
			0,005	0,0368	0,0027	
			0,01	0,0364	0,0062	
1860	1.728.870	Ciências Humanas	0,02	0,0356	0,0142	0,0002
			0,03	0,0356	0,0224	
			0,04	0,0338	0,0320	
			0,05	0,0340	0,0406	

Nas Figuras 5.16 até 5.20 se apresentam os valores das taxas de detecção dos candidatos suspeitos de fraude, tanto os adotados (α) dada pela linha verde como os observados ($\hat{\alpha}$) representados pela curva vermelha, onde esses últimos foram determinados com a estatística T^* para cada área e em cada uno dos cinco quantis utilizados. Nessas figuras se pode verificar quais áreas estão melhor ajustadas, sendo notável que são as áreas de Matemáticas e Ciências da Natureza e, em média, o quantil 0,95 foi que melhor ajustou as quatro áreas, como foi descrito anteriormente.

Figura 5.16 *Taxa de detecção dos candidatos suspeitos de fraude no quantil 0,99*

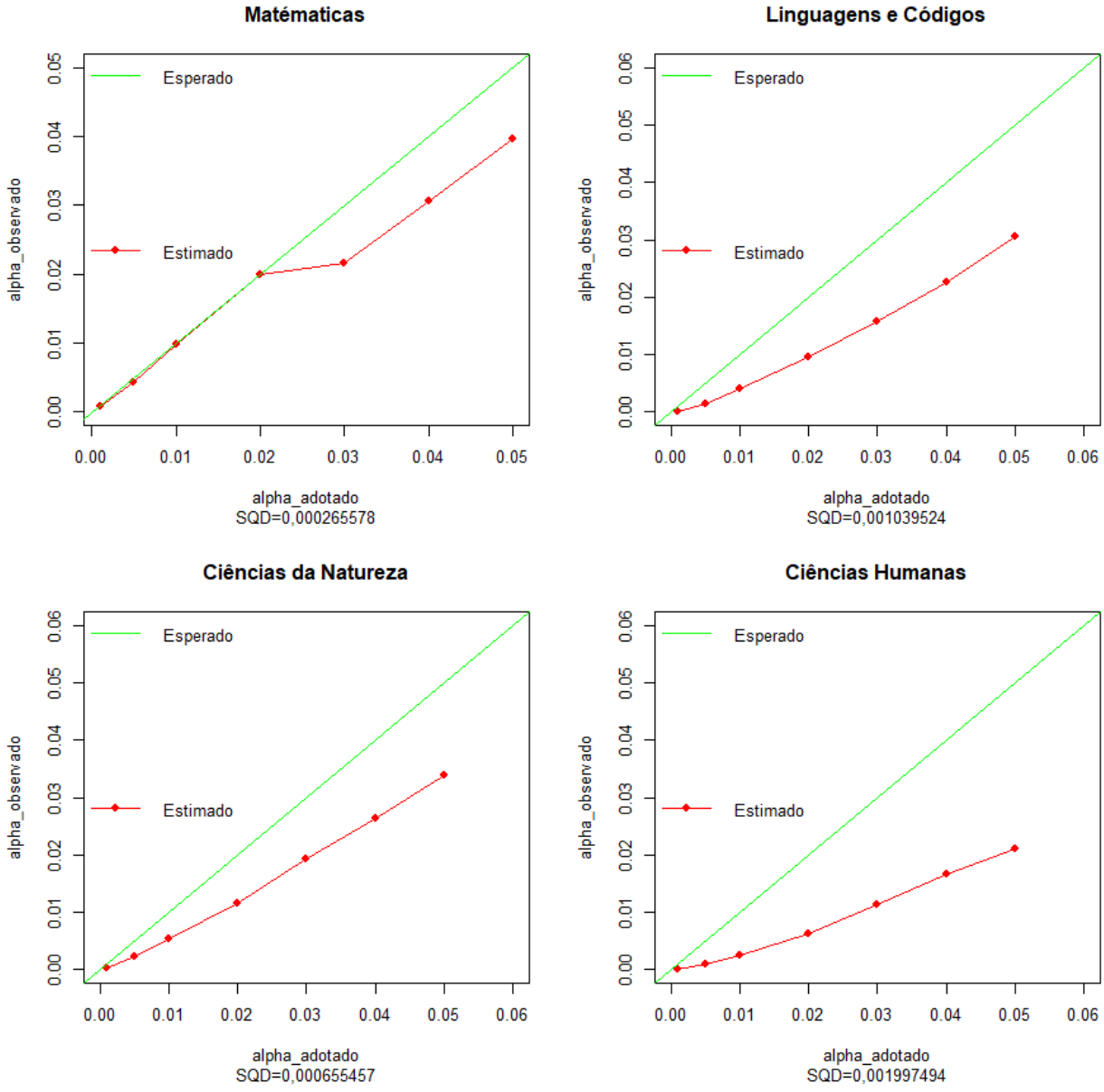


Figura 5.17 *Taxa de detecção dos candidatos suspeitos de fraude no quantil 0,98*

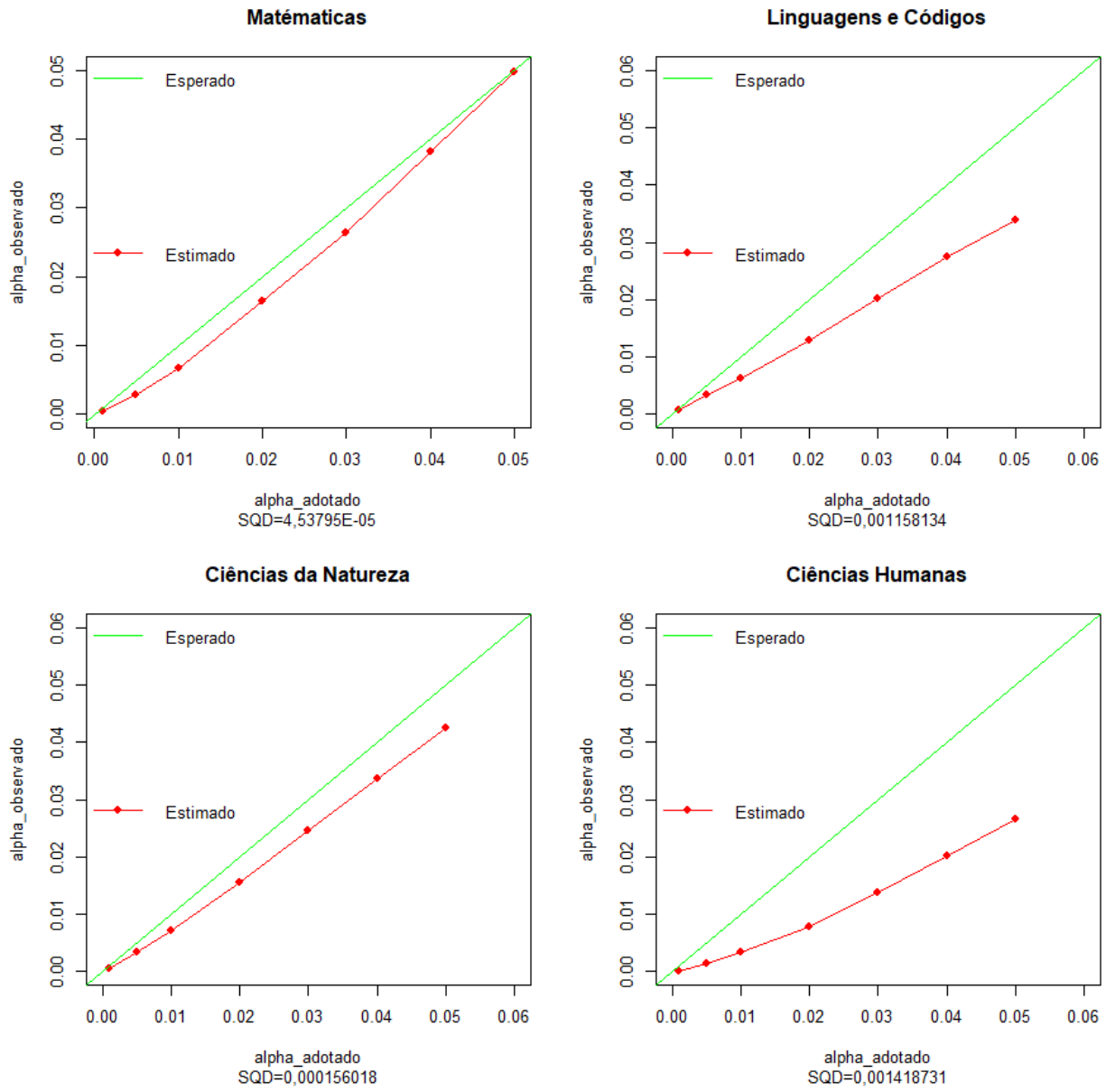


Figura 5.18 *Taxa de detecção dos candidatos suspeitos de fraude no quantil 0,97*

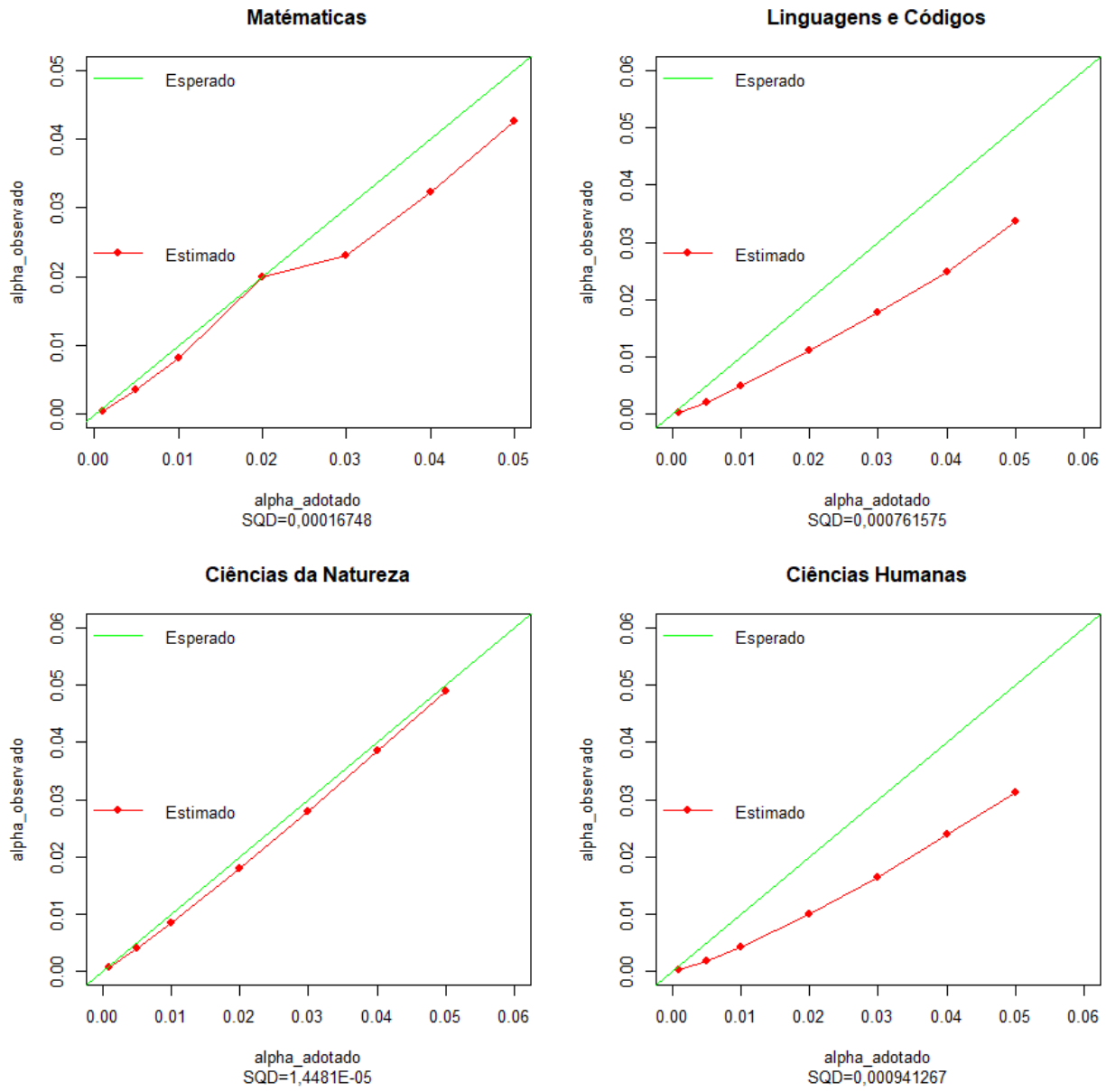


Figura 5.19 *Taxa de detecção dos candidatos suspeitos de fraude no quantil 0,96*

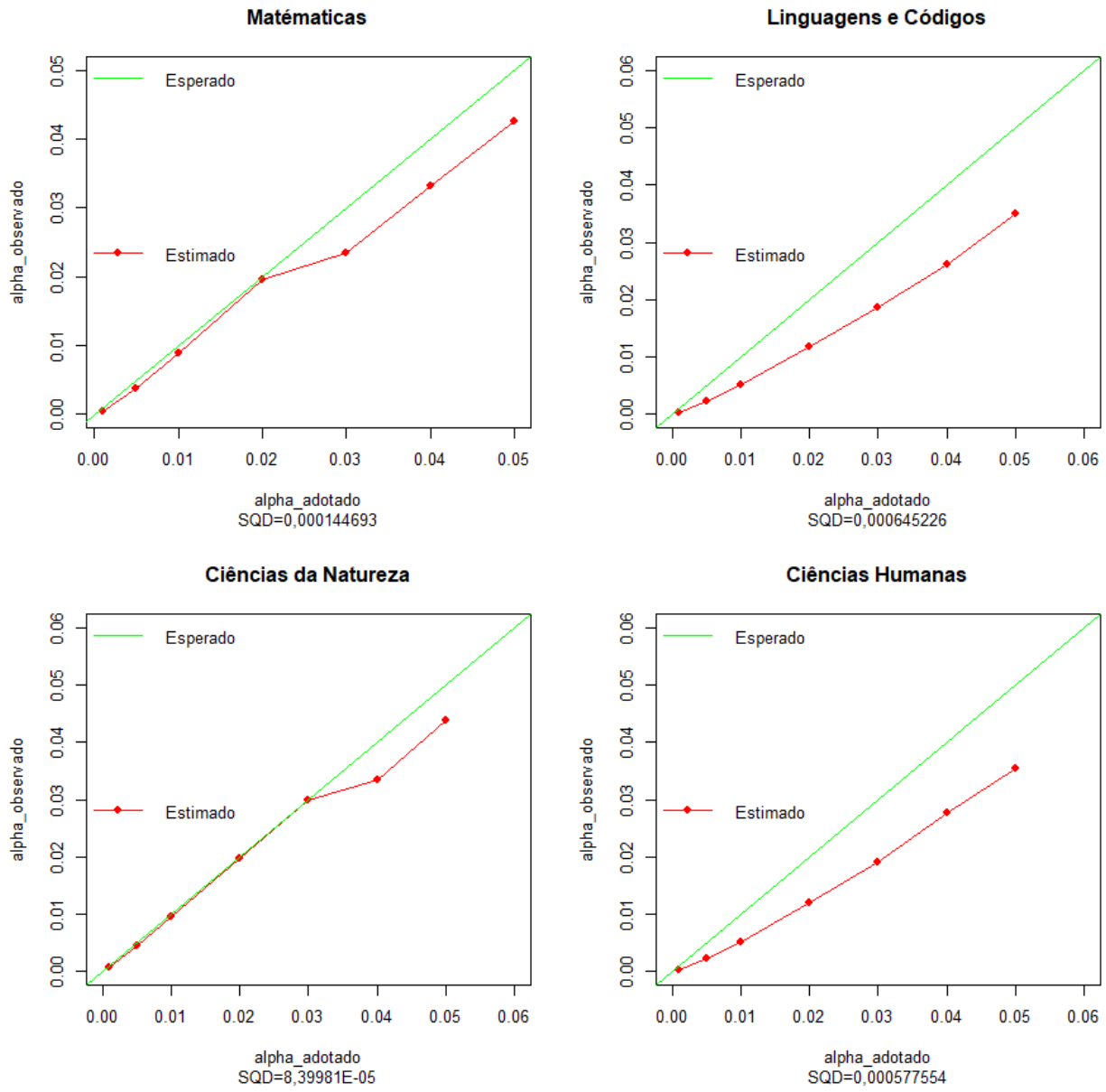
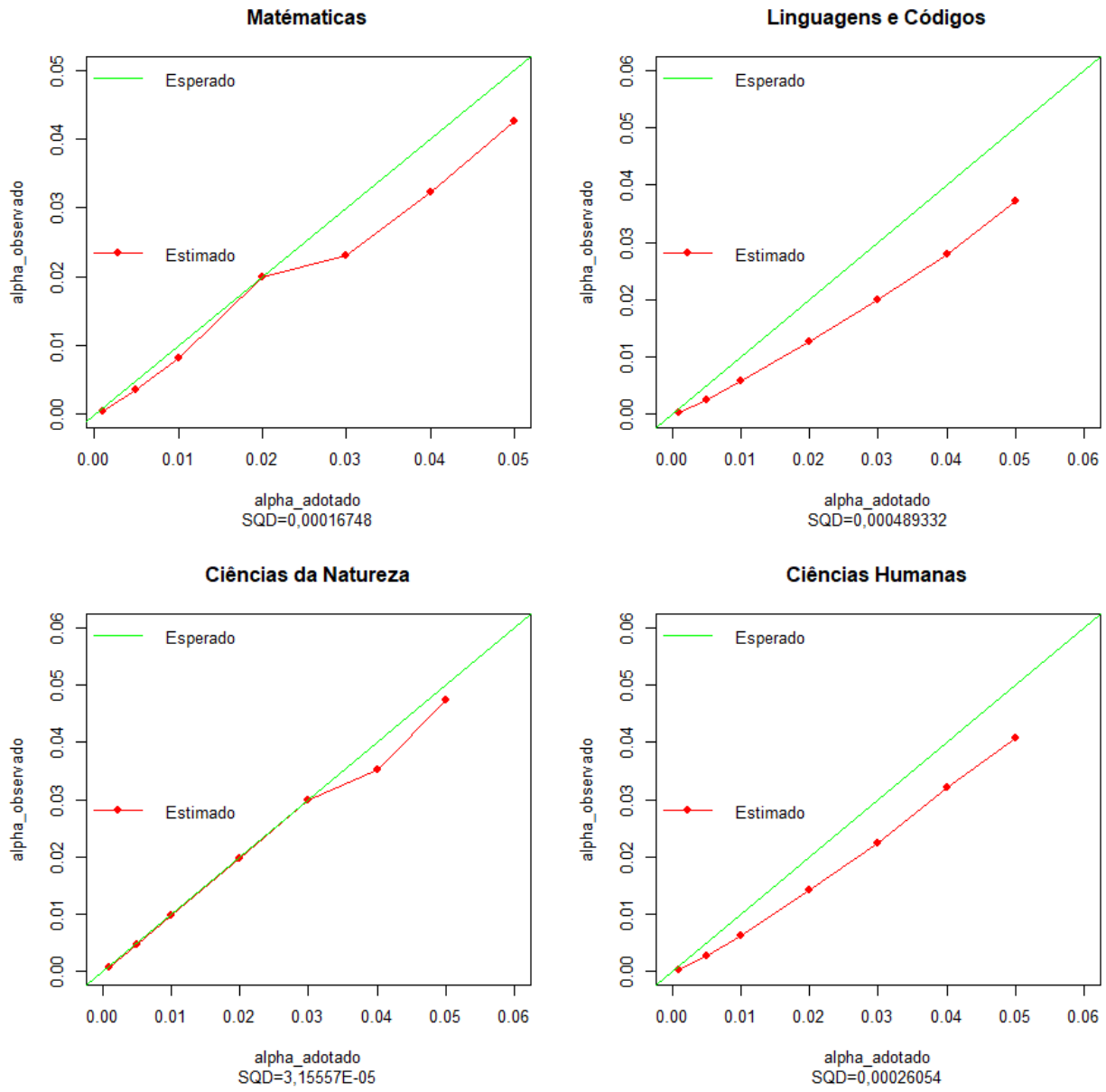


Figura 5.20 *Taxa de detecção dos candidatos suspeitos de fraude no quantil 0,95*



Capítulo 6

Conclusões e Considerações Gerais

Neste trabalho foram apresentados sete métodos estatísticos de detecção fraudes em testes e mostrou-se as dificuldades da aplicações dos mesmo nas avaliações em larga escala. O problema principal é o tempo computacional envolvido no cálculo dos índices desses métodos quando existe um grande número de participantes na avaliação de interesse. Por isso, este estudo apresentou uma proposta para tornar os métodos aplicáveis a esse tipo de avaliações.

Uma nova proposta foi avaliada com o objetivo de diminuir a quantidade de pares a serem analisados: a seleção quantílica aplicada sobre as habilidades ou proficiências dos indivíduos. Em complemento, foi proposta uma metodologia para retornar os valores mais próximos das taxas de *falsos positivos* (FP) que foram adotados. Através de um estudo envolvendo dados simulados mostrou-se que os índices podem detectar indícios de fraude para diferentes níveis de significância.

Foi realizada uma aplicação a dados reais do ENEM 2018 no município de *Teresina-PI*, adotando-se o quantil 0,95 (ou seja, os 5% de maior desempenho) e um nível de significância de 0,01 para indicação de fraude. As taxas de detecção dos possíveis candidatos fraudulentos foram notavelmente menor em comparação com os dados simulados. Mas se determinou dois conjuntos de indivíduos que foram identificados simultaneamente, como possíveis fraudadores, em três áreas diferentes do ENEM. Para as áreas de Linguagens, Matemáticas e Ciências Humanas vinte e um (21) indivíduos (um deles em duas áreas) e para as áreas de Linguagens, Ciências da Natureza e Ciências Humanas trinta e quatro (34) avaliados, todos sem repetições de área. E avaliou-se como se dá a redução no número de pares a serem analisados utilizando diferentes valores da seleção quantílica tanto para os dados simulados como os reais, pois a análise dos diferentes números dos candidatos com maiores habilidades determinou o tempo de processamento computacional. Em particular, essa aplicação dos dados do ENEM para a cidade de *Teresina-PI* permitiram controlar

seu tempo de análise passando de poder levar um ano a só uns dias para determinar quais foram os indivíduos suspeitos de copiar nessa prova realizada nesse lugar. A utilização dessa metodologia é útil quando trata-se de detecção de esquemas de fraude, porque é comum nestes casos que o indivíduo fonte forneça integralmente, ou quase integralmente, as respostas para os copiadores. Além disso, a fonte geralmente possui alta proficiência e, por consequência, vai-se refletir também nas habilidades altas dos indivíduos copiadores. Também para tornar mais transparente o processo da publicação dos resultados e cumprir com os cronogramas pre-estabelecidos.

Em etapas futuras de pesquisa, serão estudados métodos hierárquicos combinados com a seleção quantílica de tal forma que seja o processo mais rápido para avaliações como o ENEM que envolvem milhões de candidatos.

Assim, as próximas etapas trarão algumas propostas para a melhoria do presente trabalho.

6.1 Trabalhos Futuros

Recomenda-se para trabalhos futuros:

- Identificar propriedades das três variáveis de seleção (escore observado, escore verdadeiro ou habilidades), determinando a que gera melhores resultados.

- Comparar o tempo de processamento ao computar os índices separadamente e a versão atual do pacote *TestFraud*.

- Refazer a seleção quantílica utilizando também diferencias de escores (observados ou esperados) ou diferencias de habilidades.

- Corrigir a estatística T^* ponderada através um ajuste utilizando sua distribuição e tomando pontos de cortes baseados nela para melhorar os valores retornados.

- Criar uma estatística nova com pesos pre-estabelecidos para cada índice dados por seu desempenho sem depender dos dados.

- Realizar estudos de simulação em que as distribuições das habilidades simuladas reproduzam as distribuições dos dados reais para cada área.

- Utilizar a Seleção Quantílica e Bases Hierárquicas por área, em que a base na etapa (área do ENEM) 2, utiliza dos pares detectados na etapa 1, e assim em diante.

Referências Bibliográficas

[1] ANDRADE, D. F.; TAVARES, H. R.; VALLE, R. C . *Teoria da Resposta ao Item: conceitos e aplicações. ABE, São Paulo*, 2000.

[2] BAKER, F. B. and KIM, S. H. *Item response theory: Parameter estimation techniques.* CRC Press, 2004.

[3] BOCK, R. D. *Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika*, 37(1):29–51, 1972.

[4] CAED - Centro de Políticas Públicas e Avaliação da Educação, 2008. *O que é avaliação educacional?*. Disponível em: `http://www.portalavaliacao.caedufjf.net/pagina-exemplo/o-que-e-avaliacao-educacional/`. Acesso em: 20 dez. 2018.

[5] CIZEK, G. J.; WOLLACK, J. A. *Handbook of quantitative methods for detecting cheating on tests.* Routledge New York, NY, 2017.

[6] HAMBLETON, R. K. and SWAMINATHAN, H. and ROGERS, H. J. *Fundamentals of item response theory.* Sage, 1991.

[7] HOLLAND, P. W. *Assessing unusual agreement between the incorrect answers of two examinees using the K-index: Statistical theory and empirical support. ETS Research Report Series*, 1996(1):i–41, 1996.

[8] MERSMANN, O. *microbenchmark: Accurate Timing Functions*, 2018. R package version 1.4-6.

[9] MILONE, G. *Estatística: geral e aplicada.* Pioneira Thomson Learning, 2004.

[10] SOTARIDONA, S. L. *Statistical Methods for the Detection of Answer Copying on Achievement Tests.* 2003. 126 f. Tese de Doutorado apresentada ao Departamento de

Metodologia de Pesquisa, Medição e Análise de dados da Universidade de Twente – Holanda.

[11] SOTARIDONA, L. S.; MEIJER, R. R. *Statistical properties of the K-index for detecting answer copying. Journal of Educational Measurement*, 39(2):115–132, 2002.

[12] SOTARIDONA, L. S.; MEIJER, R. R. *Two new statistics to detect answer copying. Journal of Educational Measurement*, 40(1):53–69, 2003.

[13] SOUZA, M. M. *Implementação e otimização do pacote TestFraud para detecção de fraude em testes*. 2019. 42 f. Dissertação (Mestrado em Estatística) – Instituto de Ciências Exatas e Naturais, Universidade Federal de Pará, Belém.

[14] VAN DER LINDEN; WIM J.; SOTARIDONA, L. *Detecting answer copying when the regular response process follows a known response model. Journal of Educational and Behavioral Statistics*, 31(3):283–304, 2006.

[15] WOLLACK, J. A. *A nominal response model approach for detecting answer copying. Applied Psychological Measurement*, 21(4):307–320, 1997.

[16] ZOPLUOGLU, C.; CIZEK, G. J.; WOLLACK, J. A. *Similarity, answer copying, and aberrance: Understanding the status quo. CIZEK, G. J.; WOLLACK, J. A.,"Handbook of quantitative methods for detecting cheating on tests," New York, NY: Routledge*, pages 25–46, 2017.