UNIVERSIDADE ABERTA

UM MÉTODO ESTATÍSTICO PARA IDENTIFICAÇÃO DE FRAUDE EM EXAMES DE LARGA ESCALA COM QUESTÕES DO TIPO MÚLTIPLA ESCOLHA

EDUARDO AUGUSTO COMENDA COTRIM

Mestrado em Estatística, Matemática e Computação

Na área de especialização de Estatística Computacional

Dissertação de Mestrado orientada por

Professora Doutora Catarina S. Nunes

Professora Doutora Maria João Oliveira

Setembro/2022

UNIVERSIDADE ABERTA

UM MÉTODO ESTATÍSTICO PARA IDENTIFICAÇÃO DE FRAUDE EM EXAMES DE LARGA ESCALA COM QUESTÕES DO TIPO MÚLTIPLA ESCOLHA

EDUARDO AUGUSTO COMENDA COTRIM

Mestrado em Estatística, Matemática e Computação

Na área de especialização de Estatística Computacional

Dissertação de Mestrado orientada por

Professora Doutora Catarina S. Nunes

Professora Doutora Maria João Oliveira

Setembro/2022

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho.

Confirmo que em não recorri à prática de plágio ou a qualquer forma de falsificação de resultados.

Universidade Aberta, 28 de março de 2022.

Assinatura: _____

UM MÉTODO ESTATÍSTICO PARA IDENTIFICAÇÃO DE FRAUDE EM EXAMES DE LARGA ESCALA COM QUESTÕES DO TIPO MÚLTIPLA ESCOLHA

RESUMO

No Brasil, exames de larga escala são utilizados em concursos para admissão a cargos públicos ou para ingresso em universidades. Existem organizações criminosas especializadas em fraudar tais exames, causando enormes danos para a sociedade, possibilitando que pessoas não qualificadas e desonestas ingressem nas universidades e funções públicas, em detrimento de pessoas qualificadas e honestas.

Em busca de uma forma de provar cientificamente a ocorrência de fraude em exames compostos por questões tipo múltipla-escolha, desenvolveu-se um método de análise estatística da similaridade das respostas dos candidatos.

O método se baseia no fato de que as respostas dadas por uma população de candidatos em determinado exame seguem uma distribuição probabilística, cujos parâmetros podem ser estimados a partir de dados intrínsecos do exame. Compara-se o número de respostas coincidentes obtidas entre cada par de candidatos com o número que seria esperado, e calcula-se a probabilidade associada a essa ocorrência. Destacam-se os casos cuja probabilidade de ocorrência é muito pequena, menor que um nível de significância pré-estabelecido.

O método foi desenvolvido de maneira a preservar a segurança, de modo a garantir que todos os candidatos sinalizados como fraudadores tenham probabilidade elevada de terem cometido a fraude, mesmo correndo o risco de eventualmente deixar de indicar algum candidato culpado. Isso é feito através da escolha adequada do nível de significância para os testes de hipóteses.

As limitações de aplicabilidade do método são analisadas através de simulação de dados, determinando-se os limites dentro dos quais o método pode ser aplicado de forma eficaz e fiável.

Palavras-chave: Distribuição Bernoulli com probabilidades variáveis; Teorema de Liapounov; Teorema do Limite Central para variáveis não identicamente distribuídas; identificação de fraude em exames tipo múltipla escolha.

A STATISTICAL METHOD FOR THE IDENTIFICATION OF FRAUD IN LARGE SCALE EXAMS WITH MULTIPLE CHOICE QUESTIONS

SUMMARY

In Brazil, large-scale exams are used in selection processes for admission to public positions or universities. There are criminal organizations specialized in defrauding such exams, causing enormous damage to society, allowing unqualified and dishonest people to enter universities and public functions, instead of qualified and honest people.

In search of a way to scientifically prove the occurrence of fraud in exams composed of multiple-choice questions, a statistical analysis method to determine the similarity of the candidates' answers was developed.

The method is based on the fact that the answers given by a population of candidates in a given exam follow a probability distribution, whose parameters can be estimated from the intrinsic data of the exam. The number of coincident responses between each pair of candidates is compared with what would be expected, and the probability associated with this occurrence is calculated. Cases whose probability of occurrence is very small, less than a pre-established level of significance, stand out.

The method was developed to preserve security, in a way that it guarantees that all candidates indicated as fraudsters have a high probability of having committed the fraud, even at the risk of eventually failing to nominate a guilty candidate. This is done by choosing the appropriate level of significance for the hypothesis tests.

The limitations of applicability of the method is analyzed through data simulation, determining the limits within which the method can be applied effectively and reliably.

Keywords: Bernoulli Distribution with variable probabilities; Liapounov Theorem; Central Limit Theorem for non-identical variables; fraud identification in multiple-choice exams.

ÍNDICE

ÍNDICE DE TABELAS

LISTA DE ABREVIATURAS, SIGLAS E ACRÔNIMOS

CESPE/CEBRASPE: instituição vinculada à Universidade de Brasília - Brasil, responsável por exames vestibulares e pela realização de diversos concursos para cargos públicos.

ENEM: Exame Nacional do Ensino Médio, um exame de larga aplicação utilizado no Brasil como uma forma de avaliar candidatos para ingresso em instituições de ensino superior.

FUVEST: Fundação para o Vestibular, instituição responsável pelos exames vestibulares para ingresso nos cursos da Universidade de São Paulo – Brasil.

GNU Awk: linguagem de programação informática utilizada na rotina computacional para o cruzamento das respostas dos candidatos e cálculos preliminares.

S.A.T.: exame de larga aplicação utilizado nos Estados Unidos como uma forma de avaliar candidatos para ingresso em instituições de ensino superior.

LISTA DE NOTAÇÃO DAS VARIÁVEIS

α : nível de significância adotado para os testes de hipóteses.

c_{jk} : estatística proporcional ao grau de similaridade de todas as respostas entre j,k.

E_1 e E_2 : erros tipo I ou falso positivo; e tipo II ou falso negativo no teste de significância.

fr_j^{iA}: frequência relativa de respostas A à questão i no estrato de habilidade do candidato j.

H_0 e H_1 : hipóteses nula e alternativa para o teste de significância.

I : número de questões no exame.

i : uma questão qualquer variando de 1 a I.

j,k : um candidato qualquer variando de 1 a N.

M : um valor inteiro no intervalo entre 0 e I.

m_{jk} : o número total de respostas coincidentes observado entre os candidatos j,k.

N : número total de candidatos.

ν_i : conjunto de possíveis alternativas de respostas à questão i.

ω : índice de similaridade utilizado por WOLLACK (1997, 2004).

$p^{iA}{}_j$: probabilidade do candidato j responder alternativa A à questão i.

$p^i{}_{jk}$: probabilidade das respostas à questão i dos candidatos j,k coincidirem.

$q^i{}_{jk}$: probabilidade das respostas à questão i dos candidatos j,k não coincidirem.

$r^i{}_j$ é a "nota" do candidato j na questão i.

r_j : número de questões respondidas corretamente pelo candidato j ("nota").

$u^i{}_{jk}$: variável aleatória binária para a questão i entre os candidatos j e k.

$u_{jk} = (u_{jk1}, u_{jk2}, ..., u_{jkI})$: sequência de I variáveis aleatórias binárias entre os candidatos j,k.

valor-p: probabilidade de significância, utilizada para avaliar o teste de hipóteses.

$x^i{}_j$: resposta do candidato j para a questão i .

1 – INTRODUÇÃO

1.1 – Exames de larga escala

Exames de larga escala são provas ou avaliações aplicadas a um grande número de pessoas, desde algumas centenas até centenas de milhares. Atualmente exames como esses são aplicados em muitos países tais como os Estados Unidos da América, o Brasil, o Reino Unido, Canadá, Holanda, além de outros. As finalidades mais comuns são para a seleção e classificação de alunos para admissão em universidades; e para a seleção e contratação de servidores públicos, através de concursos públicos.

No primeiro caso, exames de larga escala para selecionar e classificar candidatos para as vagas em universidades ou instituições de ensino superior, podemos citar como exemplos: o *S.A.T.*, nos Estados Unidos; o ENEM – Exame Nacional do Ensino Médio, de aplicação nacional no Brasil; ou ainda, os vestibulares realizados de forma independente por algumas instituições universitárias, dentre as quais exemplificamos o exame vestibular da FUVEST, para acesso aos cursos de primeiro ciclo da Universidade de São Paulo, além de várias outras, tais como Universidade Estadual Paulista – UNESP, Fundação Getúlio Vargas – FGV São Paulo, *etc*.

O principal exame para acesso às universidades públicas no Brasil é o ENEM - Exame Nacional para o Ensino Médio, e em seu sítio da Internet encontramos:

> O exame aperfeiçoou sua metodologia e, em 2009, passou a ser utilizado como mecanismo de acesso à educação superior, por meio do Sistema de Seleção Unificada (SISU), do Programa Universidade para Todos (PROUNI) e de convênios com instituições portuguesas. Os participantes do ENEM também podem pleitear financiamento estudantil em programas do governo, como o Fundo de Financiamento Estudantil (FIES).
> Pela primeira vez, o INEP realizará o ENEM Digital, com aplicação em janeiro e fevereiro de 2021. A prova em computador está prevista para mais de 96 mil participantes.
>
> Enem — Inep (www.gov.br) [15/02/2021]

O ENEM também é aceito como critério de seleção para admissão em 51 Universidades e Institutos de Nível Superior em Portugal, como pode ser visto em:

> Os resultados individuais do Exame Nacional do Ensino Médio (ENEM) podem ser usados nos **processos seletivos de instituições de educação portuguesas. Mais de 50 universidades, institutos politécnicos e escolas superiores** têm acordo interinstitucional com o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), que garante acesso facilitado às notas dos estudantes brasileiros interessados em cursos de graduação em Portugal. Cada instituição define as regras e os pesos para uso das notas.
>
> Enem Portugal — Inep (www.gov.br) [15/02/2021]

O ENEM em 2020 possuía 2 exames com 90 questões de múltipla escolha, cada qual com 4 alternativas de respostas possíveis. Em 2021 foi aplicado para aproximadamente 96 mil candidatos, totalmente à distância, por computador. Os exames e grelhas de respostas do ENEM podem ser acedidos em https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem/provas-e-gabaritos .

Além desse exame, diversas Universidades brasileiras, principalmente as maiores, realizam seus próprios concursos para acesso, chamados de vestibulares. Um bom exemplo é a USP - Universidade de São Paulo, a maior do país, cujo vestibular é realizado pela sua fundação FUVEST, que teve em 2020, por exemplo, 117.019 candidatos a disputar 8.317 vagas de primeiro ciclo (dados acedidos em https://acervo.fuvest.br/fuvest/2020/) [15/02/2021].

O concurso vestibular de 2020 possuía 90 questões objetivas com 5 alternativas de resposta cada. Essa é a primeira fase e somente uma pequena proporção é aprovada para a segunda fase, que será realizada com questões escritas. Os exames e grelhas de respostas de todos os exames vestibulares realizados pela FUVEST podem ser acedidos em https://acervo.fuvest.br/fuvest/ .

Uma segunda grande vertente dos exames de larga escala são os concursos para cargos públicos no Brasil, que são bastante disputados pois apresentam uma faixa de remuneração bastante atraente no contexto da sociedade brasileira e certa estabilidade no emprego, além de reforma com mais benefícios. São dezenas de milhares de servidores públicos contratados anualmente, selecionados por concursos que têm em sua primeira fase (às vezes a única) exames com questões do tipo múltipla-escolha.

Para alguns cargos mais baixos, o concurso se resume simplesmente ao exame com questões múltipla-escolha, para selecionar e classificar os aprovados. Outros cargos mais altos, têm várias fases no concurso, sendo que a primeira, eliminatória, que filtra grande número de candidatos, é o exame com questões múltipla-escolha.

Como exemplos, citamos dois concursos realizados pelo antigo CESPE, vinculado à Universidade de Brasília, atualmente chamado CEBRASPE:

- Um concurso para a carreira de Diplomata realizado em 2018 (26 vagas, 5294 candidatos), cujo exame foram 73 questões, cada uma com 4 subitens, cujas possibilidades de respostas são “Certo”, “Errado”, ou “Sem Resposta”. Salário inicial na

época era o equivalente a 4.097€ mensais, enquanto o salário-mínimo nacional era equivalente a 241€, portanto, equivalente a cerca de 17 vezes o salário-mínimo;

- Um concurso para a carreira de Agente de Polícia Federal realizado em 2009 (200 vagas, 63294 candidatos), cujo exame possuía 120 questões, cujas possibilidades de respostas são "Certo", "Errado", ou "Sem Resposta". Salário inicial na época era de 2.789€ mensais, enquanto o salário-mínimo nacional era equivalente a 155€, portanto, equivalente a cerca de 18 vezes o salário-mínimo.

Os exames e as grelhas de respostas de ambos os concursos, bem como todas as informações relacionadas a esses e demais concursos públicos realizados por essa entidade foram acedidos em 15/02/2021 nos sítios:

http://www.cespe.unb.br/concursos/_antigos/default.asp
http://www.cespe.unb.br/concursos/DPFAGENTE2009/
http://www.cespe.unb.br/concursos/IRBR_18_DIPLOMACIA/

1.2 – Fraudes nos exames e concursos

A sociedade brasileira tem vivido um movimento crescente de procura por cargos públicos, principalmente em busca de salários considerados atraentes e relativa sensação de segurança e estabilidade no emprego público. Isso faz com que os concursos para provimento de cargos se tornem cada vez mais concorridos, e, portanto, de difícil aprovação. Essa crescente dificuldade e concorrência aumentam o interesse por fraudar os exames, em busca de aprovação.

Essas fraudes podem ocorrer de várias maneiras, principalmente:

(I) de forma individual, por um único candidato, através de algum tipo de meio inidôneo de acesso à informação, vulgarmente conhecido como "cola", ou através da comunicação remota de informações ("ponto eletrônico");

(II) no envolvimento de dois candidatos que realizem o exame, sentados fisicamente próximos e que, de alguma maneira venham a copiar ou repassar suas respostas entre si, ou que possuam algum meio de comunicação entre si;

(III) de forma coletiva, onde um grupo de candidatos tem acesso prévio às respostas ou às questões, respondendo à prova de forma igual ou bastante semelhante às respostas ou às questões que tiveram acesso prévio, independentemente do local onde estejam realizando o exame.

Esta última forma pode atingir um número considerável de candidatos, causando vários efeitos indesejáveis, tais como:

- candidatos que não conseguiriam aprovação, acabam por obtê-las, vindo a conquistar um cargo público indevidamente;

- candidatos que obteriam aprovação, caso o concurso fosse idôneo, acabam por ficar com pior classificação, sendo injustamente reprovados;

- a entidade que promove o concurso, despende recursos e não consegue realmente selecionar os candidatos mais preparados, em decorrência da fraude, selecionando assim, indivíduos que se utilizam de meio criminoso para ingressar no serviço público.

Percebendo-se da imensa oportunidade de auferir ganhos de grande monta, dada a já mencionada crescente procura por concursos públicos, criminosos organizam-se de maneira estruturada para cometer reiteradas fraudes, operando um verdadeiro negócio, ilícito, comercializando provas, questões e temas, prejudicando não só os candidatos idôneos e as entidades públicas promotoras dos concursos, mas em última análise, toda a nossa sociedade, que teria que conviver, quando do sucesso da fraude, com servidores públicos inidôneos, capazes de cometer crime e que, provavelmente iriam reiterar esse tipo de conduta, quando no exercício de suas funções.

A Polícia Federal do Brasil já realizou diversas investigações e algumas operações com o fim de acabar com organizações criminosas desse tipo e prender seus integrantes, tais como a Operação Tormenta em 2009 e a Operação Afronta em 2016.

O objetivo deste trabalho é desenvolver um método de análise de similaridade de respostas, que sinalize uma provável ocorrência ou não de fraude em exames de larga aplicação, compostos somente por questões com respostas do tipo múltipla escolha, com o mesmo número de alternativas de respostas para todas as questões, e que possuem somente uma resposta correta por questão.

1.3 – Breve comentário sobre os métodos existentes

Há várias décadas pesquisadores vêm estudando ou desenvolvendo métodos para a identificação de fraude em exames com questões objetivas do tipo múltipla escolha. Em 1972, W. Angoff publicou *The Development of Statistical Indices for Detecting Cheaters*, onde propõe oito índices estatísticos que buscam detectar fraudes (ANGOFF, 1972).

Desde então, o tema despertou interesse de muitos pesquisadores e muitos estudos foram desenvolvidos, propondo índices e métodos, bem como revisando-os e comparando-os entre si, quanto à efetividade e facilidade de aplicação, dentre os quais destacamos os trabalhos de: ANGOFF (1972), ARGENAL *et al* (2004), BAY (1995), CIZEK e WOLLACK (2017), CODY (1985), FRARY (1992), FRARY, TIDEMAN & WATTS (1977), HANSON, HARRIS & BRENNAN (1987), HOLLAND (1996), LEWIS e THAYER (1998), McMANUS, LISSAUER & WILLIAMS (2005), SOTARIDONA (2003),SOTARIDONA e MEIJER (2001, 2003), VAN DER LINDEN e SOTARIDONA (2006), WESOLOWSKY (2000) e WOLLACK (2004), referidos na Bibliografia.

Deve-se ressaltar que a maior parte desses trabalhos centraram foco na análise da fraude individual, isto é, um indivíduo copiar respostas de outro, fenômeno diferente da fraude múltipla, onde vários indivíduos têm acesso às questões antes do momento do exame, mesmo estando em localidades distintas, que é o foco deste trabalho.

Muitos dos conceitos utilizados neste trabalho também foram aplicados em algum ou em vários dos trabalhos citados, e serão referenciados de acordo com a sequência utilizada para o desenvolvimento da metodologia.

1.4 – Metodologia de investigação e estrutura da dissertação

Quanto à metodologia de investigação, seguiremos o seguinte caminho:

(I) entenderemos o problema a ser analisado como sendo modelável sob o ponto de vista estatístico;

(II) definiremos algumas variáveis e enunciaremos os pressupostos considerados, citando, quando necessário, os resultados teóricos que nos permitem assumir tais pressupostos;

(III) para demonstrar a aplicação do método, ao invés de utilizarmos dados de exames e concursos reais, que implicariam em questões de confidencialidade e sigilo, utilizaremos simulação matemática, uma técnica de simulação de números pseudoaleatórios, para gerar dados de exames fictícios, equivalentes a exames reais. Isto nos traz a vantagem de poder gerar quaisquer conjunto de dados, controlando os parâmetros intrínsecos da forma mais conveniente para estudá-los, tais como número total de candidatos, número de questões, número de alternativas de resposta, número de fraudadores, percentagem de cópia;

(IV) com a aplicação do método aos dados simulados, poderemos concluir sobre a eficácia e robustez do método, bem como seus limites de aplicação no que concerne os parâmetros dos exames a analisar.

Esta dissertação está estruturada da seguinte forma:

- no capítulo 2, faremos algumas considerações preliminares importantes, inclusive desmistificando alguns dos argumentos mais comuns usados por fraudadores, quando tentam explicar a razão de terem tido tantas respostas coincidentes com outros candidatos;
- no capítulo 3, faremos o desenvolvimento do método propriamente dito, apresentando seus fundamentos teóricos;
- no capítulo 4, explicamos como criamos nossas bases de dados simuladas e aplicamos o método aos dados, variando diversas situações, com intuito de analisar a influência que essas variações implicam nos resultados, a fim de mostrar tanto a eficácia e robustez do método, como também, determinar os seus limites de aplicação.
- no capítulo 5 apresentamos as conclusões a que chegamos neste trabalho, de acordo com os resultados obtidos na aplicação do método.

Ao final, trazemos as referências bibliográficas, dois anexos contendo as rotinas computacionais utilizadas para a simulação de dados e aplicação do método, e um anexo com exemplos de exames de larga escala.

2 – CONSIDERAÇÕES PRELIMINARES

2.1 – Independência local entre os eventos

Não obstante haja várias diferenças entre cada um dos métodos existentes, todos parecem concordar que em um exame idôneo, onde não tenha havido fraude, as respostas dadas por cada candidato a cada questão, são eventos independentes entre si. Essa hipótese se baseia no fato de que as respostas escolhidas por um candidato para uma determinada questão, não devem influenciar em nada as respostas escolhidas por outro candidato àquela mesma questão, uma vez que os exames devem ser resolvidos individualmente, sem qualquer tipo de comunicação entre os candidatos durante o exame.

Assim como refere BAY (1995), a hipótese de independência local, feita na *Item Response Theory (IRT)* também será feita aqui. Segundo HAMBLETHON, SWAMINATHAN, & ROGERS (1991), *apud* BAY (1995:2), a propriedade de independência local significa que para um dado candidato (ou todos os candidatos de um dado estrato de habilidade) a probabilidade de um padrão de resposta é igual ao produto das probabilidades associado com as respostas do candidato aos itens individuais.

Já o conceito de independência entre cada item para um determinado candidato é, até certo ponto, intuitivo, visto que, de fato, na imensa maioria dos exames, duas questões quaisquer não têm entre si uma relação de dependência. Mesmo que duas questões versem sobre um mesmo assunto, cada uma avaliará um detalhe de conhecimento, de forma que o fato de um indivíduo saber a resposta correta para uma, não implica que saberá a correta para a outra; nem tampouco que não saberá. O que pode ocorrer, e de fato ocorre, é que, se um indivíduo apresenta maior conhecimento no assunto em questão, maior será a probabilidade de acertar o item. Do contrário, com pouco conhecimento, menor será a probabilidade de acertá-la. Portanto, o método estatístico, para estimar a probabilidade de resposta ao item, deve se condicionar a esse maior ou menor conhecimento, ou habilidade de cada candidato.

Sendo assim, nossa questão passa a ser: calcular o nível de respostas coincidentes esperadas probabilisticamente e comparar com as coincidências de respostas efetivamente obtidas. Essa comparação permitirá concluir se as respostas aos testes foram independentes e, portanto, idôneas, ou se existe uma grande probabilidade de um tipo de fraude.

Conforme SOTARIDONA (2003), a observação principal sobre a qual repousa seu método é aquela que se as respostas de dois candidatos são probabilísticas e se um deles não teve acesso às respostas do outro, as respostas são estatisticamente independentes. Se um teve acesso a algumas das respostas do outro e copiou-as, as respostas de ambos nesses itens vão ser não somente dependentes, mas até perfeitamente coincidentes. Entretanto, se um não copiou nenhuma resposta do outro, como as suas respostas são probabilísticas, é ainda possível que algumas delas coincidam. O problema estatístico que nós consequentemente enfrentamos é decidir quanta coincidência nós devemos aceitar antes de rejeitar a hipótese nula que um candidato não copiou a resposta de nenhum dos itens no teste.

Segundo WOLLACK (2004), quaisquer dois candidatos trabalhando independentemente irão produzir padrões de respostas dos itens com alguma quantidade de sobreposição. Índices estatísticos funcionam essencialmente comparando a quantidade de superposição entre dois candidatos à quantidade normal que seria esperada se fosse sabido que os dois candidatos responderam independentemente um do outro. A estatística proposta por WOLLACK (2004) compara o número de respostas coincidentes entre um par de candidatos com o número de coincidências esperadas aleatoriamente.

2.2 - O mito do "estudamos juntos"

Sem dúvida o argumento de defesa mais comumente utilizado por dois ou mais candidatos que são acusados de fraude devido a uma grande quantidade de respostas coincidentes é o de que ambos estudaram e se prepararam juntos para o exame, e por isso teria ocorrido um alto índice de coincidências.

Ora, tal argumento não prospera, uma vez que em exames realizados por imenso número de candidatos, muitos são os pares ou grupos de candidatos que estudam e se preparam juntos, mas cujo nível de respostas coincidentes respeita os limites trazidos pela estatística, sendo que apenas alguns poucos atingem um padrão de coincidências inaceitável do ponto de vista probabilístico.

Nesse sentido referem HOLLAND (1996) e WESOLOWSKY (2000). Se assim fosse, como muitos estudam juntos, têm os mesmos professores, seguem os mesmos cursos e utilizam o mesmo material preparatório, era de se esperar que ocorressem muitos casos com altos índices de coincidência, o que efetivamente não ocorre. Apenas uns poucos

se destacam de forma aberrante. Isso acontece porque as questões dos exames são elaboradas justamente para diferenciar, separar os diversos graus de conhecimento e assimilação de determinado conteúdo, fazendo com que as coincidências entre respostas de dois indivíduos que tenham tido preparação idêntica, devido a diferentes graus de assimilação, tenham um nível de respostas coincidentes dentro de certos limites, que podem ser previstos estatisticamente.

Segundo HOLLAND (1996), enquanto há outras maneiras possíveis de que pequenos valores da estatística proposta por si possam ocorrer sem indicar cópia ou comunicação, há uma explicação comumente proposta para respostas incorretas em comum – que dois candidatos estudaram juntos e aprenderam tudo errado da mesma maneira – o que ele nunca encontrara evidências em casos reais do Escritório de Segurança dos Exames (EUA). Esta explicação ignora a natureza discreta da maioria das questões de teste de múltipla escolha. Raramente os elementos causadores de distração de uma questão têm alguma relação com os de outra. Se muitas questões erradas são respondidas de forma idêntica, a explicação "estudamos juntos" implica em um grande desentendimento de vários assuntos que coloca em questão a verdadeira noção de estudo.

Já WESOLOWSKY (2000) refere que o argumento mais comum de estudantes acusados de cópia (de longe) é o "que nossas respostas são semelhantes porque nós estudamos juntos". O modelo assume independência nas respostas dos estudantes e, consequentemente, esse argumento aparentemente plausível tem que ser considerado. Primeiro, devemos reconhecer que programas como este, avaliam um número muito grande de pares de estudantes. Como discutido anteriormente, há geralmente oportunidades muito limitadas para sentar-se adjacentemente, e um grande número de estudantes estudam juntos. O efeito "estudamos juntos", ou outras imperfeições do modelo, deveriam produzir grande número de pares fortemente semelhantes que não estavam sentados adjacentemente. Parece bastante implausível que as violações do modelo estão somente restritas a pares sentados adjacentemente. Consequentemente parece implausível que o efeito "estudamos juntos" é forte o suficiente para causar falsas acusações, isto indica que o modelo é muito robusto com respeito a violações às suas hipóteses. Isto também indica que a defesa "estudamos juntos" é plausível apenas superficialmente.

2.3 - Comparação de todas as respostas *versus* somente respostas incorretas

Os diversos métodos existentes podem ser separados em dois grandes grupos:

(I) os que consideram todas as respostas coincidentes, dentre os quais destacamos alguns: FRARY, TIDEMAN & WATTS (1977), BAY (1995), WOLLACK, (1997, 2004) e WESOLOWSKY (2000);

(II) os que consideram apenas as respostas erradas coincidentes, dentre os quais destacamos alguns: ANGOFF (1972), CODY (1985), HANSON, HARRIS & BRENNAN (1987) e HOLLAND (1996).

Os defensores deste segundo grupo apresentam como argumento principal o fato de que não seria justo acusar algum candidato de fraude, baseando-se em um grande número de respostas certas coincidentes, visto que o candidato se prepara justamente para responder as questões de forma correta, e não poderia ser penalizado por isso; e ainda, que candidatos bem preparados tenderão a acertar mais questões, gerando um grande número de respostas certas coincidentes. Num primeiro momento, tal argumento parece válido, levando-se em conta que, em princípio, parece ser mais raro dois candidatos coincidirem em uma resposta errada, do que em uma correta. No entanto, imagine uma questão em que a maioria dos candidatos errou, poucos tendo respondido a alternativa correta. Ora, neste caso, será bem mais provável que dois candidatos coincidam em uma alternativa errada, do que na correta. Embora esses casos sejam minoria em um exame equilibrado, eles existem e trazem consigo um conteúdo de informação que não deve ser desprezado.

Segundo CIZEK (2001), dois desses métodos, o desenvolvido por FRARY, TIDEMAN & WATTS (1977), e o desenvolvido por WOLLACK (2004) são tecnicamente superiores. Esses procedimentos oferecem mais força para identificar fraude, enquanto prevenindo contra identificação errônea de fraude (erros tipo I ou falso positivos), e podem ser usados em amostras relativamente pequenas (em torno de 200 candidatos). Diferentemente de outros métodos que se apoiam somente em número de erros em comum, o que pode distorcer os resultados quando a habilidade como um todo não é levada em conta, aqueles dois incorporam informações das respostas certas em comum e probabilidades diferenciais de seleção de opções incorretas.

No mesmo sentido refere SOTARIDONA (2003), quando menciona que WOLLACK (1997, 2004) destacou que a potência de uma estatística que não leva em conta a informação de questões respondidas corretamente, parece decrescer devido à redução no

número de questões operacionais usadas. Ele propõe uma estatística para incorporar a informação sobre cópia que está contida nas respostas corretas coincidentes, em adição à informação nas respostas incorretas coincidentes.

Também ARGENAL *et al.* (2004) refere que ANGOFF (1972) desenvolveu 8 índices estatísticos usando uma variedade de variáveis para identificar fraudadores e descobriu que as estatísticas envolvendo a contagem de respostas certas e a contagem de respostas erradas era mais eficiente para identificar casos de cópia. E que a maioria dos métodos usaram erros correlacionados ou respostas incorretas coincidentes, tais como por exemplo ANGOFF (1972); CODY (1985); HOLLAND (1996); SOTARIDONA e MEIJER (2001); porque a fraude é sugerida se dois estudantes consistentemente escolhem alternativas erradas idênticas para os mesmos itens. E menciona ainda, que a presença de respostas incorretas coincidentes ou compartilhadas produz evidência de cópia, mas o peso dado para esta evidência deve ser proporcional à probabilidade de o suspeito copiador escolher aquela alternativa independentemente. Se quase todos os candidatos selecionaram a mesma resposta errada então uma coincidência na incorreta não seria e não deveria ser considerada uma evidência forte de conspiração. Mas se a escolha/alternativa específica é altamente incomum, a evidência de cópia daquele item é mais alta. No entanto, ressalta que outros métodos, tais como por exemplo FRARY, TIDEMAN & WATTS (1977); WOLLACK (1997, 2004); SOTARIDONA e MEIJER (2003) incorporaram respostas corretas coincidentes em seus modelos, em adição às respostas incorretas coincidentes e que SOTARIDONA e MEIJER (2003) concordaram que a evidência de cópia de respostas é mais forte se baseada em coincidência de respostas incorretas, do que se baseada em coincidência de respostas corretas, mas uma perda de informação é evidente se as respostas corretas coincidentes são descartadas.

Este trabalho irá focar em obter parâmetros ou estimadores para a quantidade de respostas coincidentes totais, incluindo respostas certas e erradas, que se poderia esperar em um exame cujas respostas dos candidatos fossem, de fato, independentes entre si e, portanto, não houvesse ocorrido qualquer tipo de fraude.

3 – DESENVOLVIMENTO DO MÉTODO

Assim como fez WESOLOWSKY (2000), o modelo que passaremos a desenvolver simplesmente analisa o número de respostas coincidentes e ignora outros padrões suspeitos, tais como grupos ou sequências de respostas coincidentes. Isto para se fazer o mínimo de hipóteses possíveis, especialmente sobre o comportamento de candidatos fraudadores, algo difícil de descrever sem se fazer muitas suposições subjetivas. Outra razão é para manter o método compreensível e facilmente aplicável.

3.1 - Definição do espaço amostral

O primeiro aspeto importante reside na definição do espaço amostral a ser avaliado. Dada uma população com um número total de candidatos N que realizou determinado exame, para analisar-se a existência ou não de fraude, será necessário confrontar-se as respostas de cada um dos N candidatos com os N-1 demais candidatos, contando o número de respostas coincidentes entre cada par de candidatos j e k , onde j e k representam quaisquer candidatos variando de 1 a N, de forma a compor uma lista com o número total de respostas coincidentes m_{jk} para cada par de candidatos j,k.

Portanto, o número de comparações de respostas feitas cresce com o quadrado do número de candidatos, pela expressão N.(N-1)/2. Esse número de comparações também se refletirá no número de cálculos a ser realizado pela rotina computacional, afetando diretamente o tempo de processamento e, às vezes até, inviabilizando que a mesma seja executada em determinado equipamento.

O cruzamento das respostas de 12.000 candidatos entre si, por exemplo, em um exame de 100 questões, implica em 7.199.400.000 de comparações de resultados. Imagine-se um concurso com 200.000 candidatos, realizando um exame com 100 questões. A quantidade de comparações necessárias seria 1.999.990.000.000, o que é bastante grande e pode ser inviável, dependendo dos recursos informáticos disponíveis.

Por outro lado, deve-se ter em mente que o objetivo deste trabalho é propor um método para se sinalizar prováveis fraudes nos exames. Considerando que, na grande maioria das vezes, os eventuais fraudadores estariam entre os aprovados ou entre os melhores classificados; e que se os mesmos estiverem muito mal classificados, menos importante seria a sua identificação, então é interessante para este método, no caso de a

população do concurso ser grande, restringir o espaço amostral de modo a escolher o grupo de candidatos melhor classificado, a fim de trabalhar com uma base de dados de tamanho adequado.

Nesse sentido, costuma-se considerar necessário que, no mínimo, todos os candidatos aprovados sejam avaliados. Em algumas situações, de acordo com peculiaridades do concurso em foco, esse grupo pode ser ampliado ou restringido, dependendo de onde podem se situar os suspeitos da fraude, quando houver.

Sempre que for computacionalmente viável, é preferível adotar o espaço amostral como sendo todos os candidatos que realizaram o exame, pois aumentará a precisão do método. Na seção 4.9 serão discutidos os limites considerados mínimos para o tamanho do espaço amostral.

3.2 – O evento "respostas coincidentes" como uma variável aleatória

Quando se compara as respostas dadas por dois candidatos a uma determinada questão, apenas duas situações podem ocorrer: as respostas serem coincidentes; ou as respostas não serem coincidentes (respostas diferentes). O evento "respostas coincidentes" pode ocorrer ou não, de forma independente, para cada questão *i* do exame ($i=1,\ldots,I$), onde *I* é o número total de questões do exame.

Introduz-se a seguinte variável aleatória:

$$u_{jk}^{i} = \begin{cases} 1, \text{se os candidatos } j,k \text{ selecionaram respostas iguais para a questão } i; \\ 0, \text{se os candidatos } j,k \text{ não selecionaram respostas iguais para a questão } i. \end{cases} \quad (1)$$

Então, u^i_{jk} tem a seguinte distribuição Bernoulli:

$$\begin{cases} \mathrm{P}(u_{jk}^{i} = 1) = p_{jk}^{i} \ ; \\ \mathrm{P}\left(u_{jk}^{i} = 0\right) = q_{jk}^{i} = 1 - p_{jk}^{i} \ . \end{cases} \quad (2)$$

com valor esperado p^i_{jk} e variância $p^i_{jk}.q^i_{jk}$, onde

p^i_{jk}: probabilidade das respostas à questão *i* dos candidatos *j*,*k* coincidirem;

q^i_{jk} : probabilidade das respostas à questão *i* dos candidatos *j*,*k* não coincidirem.

Cada u^i_{jk} pode ser interpretado como um resultado de um total de *I* ensaios com a probabilidade de sucesso p^i_{jk} ou fracasso q^i_{jk}, variando para cada ensaio *i* para cada par de candidatos *j*,*k*.

O valor esperado p^i_{jk} está contido no intervalo [0, 1]; e a variância $p^i_{jk}.q^i_{jk}$ está contida no intervalo [0; 0,25].

Seja uma sequência com I termos:

$$u_{jk} = (u^1_{jk}, u^2_{jk}, ..., u^I_{jk}) \quad (3)$$

Para um par de candidatos quaisquer j,k, para as questões i variando de 1 a I, a sequência $u_{jk} = (u^1_{jk}, u^2_{jk}, ..., u^I_{jk})$, será uma sequência de variáveis aleatórias binárias, pois podem assumir valores 0 ou 1, dependendo da coincidência ou não das respostas à cada questão i dadas pelos candidatos j,k, também pode ser percebida como uma sequência de ensaios de Bernoulli.

Define-se a variável aleatória m_{jk} como sendo o número total de respostas coincidentes entre os candidatos j,k, somando-se u_{jk} para todas as questões i de 1 a I:

$$m_{jk} = \sum_{i=1}^{I} u^i_{jk} \quad (4)$$

Os valores possíveis de m_{jk} são inteiros positivos no intervalo $[0, I]$, $I \in \mathbb{N}$.

O valor esperado de m_{jk} será:

$$E(m_{jk}) = E\left(\sum_{i=1}^{I} u^i_{jk}\right) = \sum_{i=1}^{I} E(u^i_{jk}) = \sum_{i=1}^{I} p^i_{jk} \quad (5)$$

A variância de m_{jk} será, usando-se a hipótese de independência entre respostas das diferentes questões i:

$$Var(m_{jk}) = Var\left(\sum_{i=1}^{I} u^i_{jk}\right) = \sum_{i=1}^{I} Var(u^i_{jk}) = \sum_{i=1}^{I} p^i_{jk}.(1 - p^i_{jk}) \quad (6)$$

Define-se então uma estatística que será proporcional ao grau de similaridade de todas as respostas entre j,k:

$$c_{jk} = \frac{m_{jk} - E(m_{jk})}{\sqrt{Var(m_{jk})}} \quad (7)$$

Note que, caso as respostas de todos os candidatos a todas as questões forem iguais entre si, então Var(m_{jk}) seria igual a 0, e nosso método seria inválido, pois c_{jk} é indeterminada para Var(m_{jk}) igual a 0. Tal situação é altamente improvável de se verificar em um exame real, mas caso aconteça, o método não poderá ser aplicado. Portanto, cada $p^i_{jk}.q^i_{jk}$ deve estar contido no intervalo]0; 0,25], para que Var(m_{jk}) seja diferente de 0.

Nas seções 4.3 e 4.4 do capítulo seguinte explicamos passo-a-passo como realizamos os cálculos para estimar estes parâmetros em um conjunto de dados.

3.3 - Distribuição binomial composta ou generalizada

Se a probabilidade p^i_{jk} de respostas coincidentes entre dois candidatos quaisquer fosse constante para toda questão i variando de 1 a I; e considerando que as

respostas entre as diversas questões são independentes entre si, então a distribuição de probabilidades de u_{jk} seria a distribuição binomial, com o número de ensaios *I*. Assim também observa BAY (1995).

Mas não se pode supor que a probabilidade p^i_{jk} seja constante, mas sim varie para toda questão *i* variando de 1 a *I*, de modo que a distribuição de probabilidades de u_{jk} não segue a distribuição binomial. Este tem sido chamado de modelo de ensaios independentes repetidos (SVESHNIKOV, 1968, *apud* WESOLOWSKY, 2000, p. 914).

WESOLOWSKY (2000) apresenta a seguinte função geradora de probabilidade (adequando-se à nossa notação):

$$P(m_{jk} = M) = \frac{1}{M!}\left(\frac{\partial^M G(u)}{\partial u^M}\right)_{u=0} \quad (8)$$

onde $M \in (1,\ldots,I)$ e:

$$G(u) = \prod_{i=1}^{I}(p^i_{jk}.u + (1 - p^i_{jk})) \quad (9)$$

A distribuição de probabilidades que uma sequência de ensaios de Bernoulli, com probabilidades variáveis entre os ensaios, segue, é muitas vezes conhecida como a distribuição binomial composta ou binomial generalizada (*compound binomial*).

De fato, como também referem VAN DER LINDEN e SOTARIDONA (2006), o número de alternativas coincidentes é o resultado de uma série de ensaios de Bernoulli independentes, cada um com uma probabilidade diferente de uma coincidência aleatória. Consequentemente, a distribuição de u_{jk} pertence à família da binomial generalizada, às vezes também chamada de binomial composta.

Regista-se que a distribuição binomial composta (ou generalizada) foi o modelo adotado na proposição dos métodos de FRARY, TIDEMAN & WATTS (1977), CODY (1985), HANSON, HARRIS & BRENNAN (1987), BAY (1995), WOLLACK (1997, 2004), WESOLOWSKY (2000) e VAN DER LINDEN e SOTARIDONA (2006).

No entanto, o cálculo da distribuição binomial composta não é simples e requer recursos computacionais poderosos, principalmente quando é grande o número de questões do exame e de candidatos. Conforme WESOLOWSKY (2000), a distribuição de probabilidade conhecida como a binomial composta não é um método prático para calcular a distribuição de probabilidade de u_{jk}. Isto pode ser feito por um método recursivo computacional que é algumas vezes usado na literatura. Entretanto, este recurso, apesar de fácil de descrever, é computacionalmente lento. Deve-se lembrar que nós desejamos procurar por todos os possíveis pares de candidatos e este número é a combinação de *N* dois

a dois, igual a $N.(N\text{-}1)/2$. Portanto, se o número de comparações de respostas cresce com o quadrado do número de candidatos *N* que realizaram o exame, quanto maior o *N*, maior será o número de comparações e consequentemente, o tempo computacional para se realizar os cálculos de todos os pares de candidatos, aumentará proporcionalmente ao quadrado de *N*.

3.4 - Teorema do Limite Central para a soma de variáveis independentes não identicamente distribuídas

O Teorema do Limite Central é considerado um dos mais importantes no estudo de probabilidades e versa sobre a convergência de somas de variáveis aleatórias para uma distribuição normal. Segundo ROSS (2004), a distribuição normal foi introduzida pelo matemático francês Abraham de Moivre em 1733 e foi usada por ele para aproximar probabilidades associadas com variáveis aleatórias binomiais quando o parâmetro binomial *n* é grande. Este resultado foi mais tarde estendido por Laplace e outros pesquisadores, e agora é englobado em um teorema de probabilidade conhecido como o Teorema do Limite Central (ou Teorema Central do Limite), que fornece uma base teórica à observação empírica muitas vezes notada na prática, que muitos fenômenos aleatórios obedecem, pelo menos aproximadamente, uma distribuição de probabilidade normal.

Há várias versões deste teorema, mas talvez sua forma mais conhecida seja a que é chamada por vários pesquisadores como Teorema do Limite Central de *De Moivre* e *Laplace*, enunciado a seguir, conforme refere ROSS (2004).

Teorema. Seja $u_1, u_2, u_3, \ldots, u_I$ uma sequência de variáveis aleatórias independentes e identicamente distribuídas cada qual tendo valor esperado μ e variância σ^2. Então para *I* grande, a distribuição de

$$u_1 + u_2 + u_3 + \ldots + u_I \tag{10}$$

é aproximadamente normal, com valor esperado $I\mu$ e variância $I\sigma^2$. Segue do Teorema do Limite Central que

$$\frac{u1 + u2 + u3 + \ldots + uI - I\mu}{\sigma\sqrt{I}} \tag{11}$$

é aproximadamente uma variável aleatória normal padrão; logo, para *I* grande,

$$P\left\{\frac{u1 + u2 + u3 + \ldots + uI - I\mu}{\sigma\sqrt{I}} < x\right\} \approx P\{Z < x\} \tag{12}$$

onde Z é uma variável aleatória normal padrão ($\mathcal{N}(0,1)$).

Não obstante esta forma do teorema seja a mais comumente utilizada, este resultado não é útil para nosso trabalho, pois restringe-se a variáveis identicamente distribuídas, que não é o nosso caso.

No problema que abordamos, estamos diante de uma soma de variáveis independentes duas a duas, mas não identicamente distribuídas.

Quando comparamos as respostas do exame de um candidato *j* com outro *k*, cada questão *i* é uma variável aleatória de Bernoulli, sendo que as respostas dos candidatos podem coincidir ($u^i_{jk} = 1$), ou não coincidir ($u^i_{jk} = 0$). Para cada questão *i*, a probabilidade de ocorrer esse evento "respostas coincidentes" varia, de forma que as variáveis aleatórias u_i, que iremos somar para termos o número total de coincidências, não são identicamente distribuídas.

O resultado que precisamos é o Teorema do Limite Central aplicável à soma de variáveis aleatórias independentes não identicamente distribuídas. Os resultados seguintes podem ser encontrados em LEHMANN (1999).

Teorema Liapounov. Sejam X_i, ($i = 1, \ldots, n$) variáveis aleatórias independentemente distribuídas, com valor esperado $E(X_i)=\xi_i$ e variâncias σ_i^2 e com terceiros momentos finitos. Denotemos $X := \sum_{i=1}^{n} X_i$; $X/n := \bar{X}$; $\xi := \sum_{i=1}^{n} \xi i$; $\xi/n := \bar{\xi}$. Se

$$Y_n = \frac{\bar{X}-E(\bar{X})}{\sqrt{Var(\bar{X})}} = \frac{\sqrt{n}(\bar{X}-\bar{\xi})}{\sqrt{\frac{(\sigma_1^2+\cdots+\sigma_n^2)}{n}}} \tag{13}$$

então

$$Y_n \xrightarrow{D} \mathcal{N}(0,1) \tag{14}$$

desde que

$$[E(\sum|X_i - \xi_i|^3)]^2 = o[(\sum \sigma_i^2)^3] \tag{15}$$

Isto é provado em, por exemplo, FELLER (volume II) (1970).

Corolário. Sejam X_i, ($i = 1, \ldots, n$) variáveis aleatórias independentemente distribuídas, com valor esperado ξ_i e variância σ_i^2 e suponha que $X_1,\ldots,X_n$ são uniformemente limitadas, isto é, existe uma constante A tal que

$$|X_i| \leq A \text{ para todo } i \tag{16}$$

Então

$$Y_n \xrightarrow{D} \mathcal{N}(0,1)$$

desde que

$$s_n^2 = \sum_{i=1}^{n} \sigma_i^2 \to \infty \tag{17}$$

Prova. Temos que

$$\sum|X_i - \xi_i|^3 \leq 2A\sum(X_i - \xi_i)^2 \quad (18)$$

e logo

$$E\sum|X_i - \xi_i|^3 \leq 2As_n^2 \quad (19)$$

O lado esquerdo de (15) é consequentemente $\leq 4A^2 s_n^4$, que é em $o[(s_n^2)^3]$ quando $s_n^2 \to \infty$. Isto prova (15) e consequentemente, (14).

Binomial de Poisson. Considere uma sequência dos chamados ensaios binomiais de Poisson, isto é, ensaios binomiais com probabilidade variável, com probabilidades de sucesso $p_i = p_1, p_2, \ldots p_n$. Seja $X_i = 1$ ou 0 se o i-ésimo ensaio é um sucesso ou fracasso e denotemos: $p := \sum_{i=1}^{n} p_i$ e $p/n := \bar{p}$. Então os $X_1,\ldots,X_n$ são uniformemente limitados e logo

$$\frac{\sqrt{n}\left(\frac{X}{n} - \bar{p}\right)}{\sqrt{\sum_{i=1}^{n}\frac{p_i q_i}{n}}} \xrightarrow{D} \mathcal{N}(0,1) \quad (20)$$

desde que

$$s_n^2 = \sum_{i=1}^{n} p_i q_i \to \infty \text{ quando } n \to \infty \quad (21)$$

A condição (21) é satisfeita sempre que existe uma constante $0 < a < 1$ tal que

$$a < p_i < 1 - a \text{ para todo } i. \quad (22)$$

Como ambos p_i e q_i são $> a$, temos que $p_i q_i > a^2$ e logo $s_n^2 > na^2 \to \infty$.

Concluímos destes resultados que, dada uma sequência de variáveis aleatórias independentemente distribuídas, e suponha que essa sequência seja uniformemente limitada, então a soma das variáveis aleatórias da sequência converge em distribuição para uma distribuição normal padrão, desde que a soma das variâncias das variáveis da sequência tenda a infinito, quando o número de termos tende a infinito.

A sequência u_{jk} (3), representa a sequência de comparação de respostas entre os candidatos j e k, sob o pressuposto de as respostas a cada questão i serem independentes duas a duas, e trata-se de uma sequência com I termos, cujos valores são somente 0 ou 1. É, portanto, uma sequência uniformemente limitada e satisfaz (16). E a soma das suas variâncias, Var(m_{jk}) (6) satisfaz (17), uma vez que cada $p^i_{jk}.q^i_{jk} \in$]0; 0,25]. Logo, (14) será válida.

Interessante notar que o lado esquerdo de (20) é exatamente análoga à definição de c_{jk} (7), pois:

$$\frac{\sqrt{n}\left(\frac{X}{n}-\bar{p}\right)}{\sqrt{\sum_{i=1}^{n}\frac{p_i q_i}{n}}}=\frac{\sqrt{n}\left(\frac{X}{n}-\frac{p}{n}\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}p_i q_i}}=\frac{n\left(\frac{X}{n}-\frac{p}{n}\right)}{\sqrt{\sum_{i=1}^{n}p_i q_i}}=\frac{X-p}{\sqrt{\sum_{i=1}^{n}p_i q_i}}$$

que equivale, ressalvada a diferença da notação das variáveis, à definição de

$$c_{jk}=\frac{m_{jk}-E(m_{jk})}{\sqrt{Var(m_{jk})}}$$

Assim, com base nos resultados apresentados por LEHMANN (1999), podemos assumir que quando o número de termos *I* da sequência u_{jk} (3) é suficientemente grande, a distribuição da soma dos seus termos m_{jk} (4) converge em distribuição para uma distribuição normal e a estatística c_{jk} (7) converge em distribuição para uma distribuição normal padrão.

$$c_{jk} \xrightarrow{D} \mathcal{N}(0,1) \qquad (23)$$

Quanto ao número de termos necessário da sequência u_{jk}, para que esta seja suficientemente grande, de forma que a convergência em (23) seja válida, esse número será estimado através múltiplas simulações de dados, conforme explicado no capítulo seguinte, na seção 4.8.

3.5 - Aproximação pela distribuição normal

A distribuição normal já foi usada por ANGOFF (1972) para comparar os valores de nível de significância com algumas das suas estatísticas. Mais além, FRARY, TIDEMAN & WATTS (1977), WOLLACK (2004) e WESOLOWSKY (2000) usaram a modelação pela binomial composta aproximada pela distribuição normal.

Conforme FRARY, TIDEMAN e WATTS (1977), *apud* HANSON, HARRIS e BRENNAN (1987:6;23), ao invés de usar a distribuição binomial composta para calcular a probabilidade de que o número de respostas que o suspeito copiador respondeu identicamente à fonte seja no mínimo tão grande quanto a observada, usam o valor esperado e o desvio padrão desta distribuição para calcular uma estatística padronizada, e afirmam que deveria ser aproximadamente distribuída pela normal.

O índice "ω" utilizado nos trabalhos de Wollack é definido como (WOLLACK, 2004):

$$\omega = \frac{\text{(número de coincidências)} - \text{(número de coincidências esperadas)}}{\text{(desvio padrão do número de coincidências)}} \qquad (24)$$

SOTARIDONA (2003) refere que para todo par de candidatos, a distribuição de "ω" (WOLLACK, 1997) aproxima a normal padrão quando o número de questões do teste se torna infinitamente grande. Consequentemente, os valores de "ω" podem ser avaliados para a significância estatística usando a distribuição normal padrão.

E WOLLACK (2004) explica, a respeito do seu índice "ω", que independentemente da habilidade do alegado copiador e da resposta dada pela alegada fonte, haverá alguma probabilidade de uma coincidência de resposta para cada item. Para alguns itens, pode ser alta, para alguns itens, pode ser baixa. Antes de calcular "ω", é necessário determinar a probabilidade de o alegado copiador selecionar a resposta da alegada fonte para todos os itens no exame. A soma dessas probabilidades ao longo de todos os itens iguala o número esperado de respostas coincidentes. Para calcular "ω", o número esperado de coincidências é subtraído do número real de respostas coincidentes entre os dois candidatos, e a diferença é dividida pelo desvio padrão do número de coincidências, o que proporciona uma medida da quantidade de variabilidade esperada no número observado de respostas coincidentes. O índice "ω" entre dois indivíduos é calculado por um programa de computador, baseado nas sequências de respostas de todos os candidatos em um dado teste. Para indivíduos que respondem independentemente, "ω" conduz a um valor que é aproximadamente distribuído pela Normal com média de 0 e desvio padrão de 1.

Importante observar o que nos trazem VAN DER LINDEN e SOTARIDONA (2006), que referem que os testes baseados na estatística "ω" têm uma distribuição nula que é postulada como Normal. Para a estatística "ω", o postulado é baseado no Teorema do Limite Central. Eles lembram que geralmente temos que ser cuidadosos ao invocar tais teoremas quando nós temos um caso de variáveis independentes, mas não identicamente distribuídas, como nessa aplicação, pois os pressupostos para aplicação do teorema poderiam não estar satisfeitos. Entretanto, para tais variáveis binárias, as condições no Teorema do Limite Central para variáveis não identicamente distribuídas, conhecidas como o teorema de Liapounov (LEHMANN, 1999), subsistem, e a normalidade em grandes amostras, para o número de alternativas coincidentes é garantida.

3.6 - Estimação da probabilidade de resposta coincidente

Para calcularmos c_{jk} (7) precisamos conhecer m_{jk} (4), E(m_{jk}) (5) e Var(m_{jk}) (6). Comparando as respostas entre os candidatos *j*,*k* e contando as respostas coincidentes, obtemos m_{jk}.

Para estimarmos E(m_{jk}) e Var(m_{jk}), precisamos estimar p^i_{jk}. Para toda questão *i* variando de 1 a *I*, a p^i_{jk}, é a probabilidade de a resposta de um candidato qualquer *j* coincidir com a resposta de outro candidato qualquer *k*, ou, em outras palavras, é a probabilidade de o candidato *j* e o candidato *k* terem respondido a mesma alternativa na dada questão *i*.

É na maneira de estimar essa probabilidade p^i_{jk}, que cada método apresenta sua própria proposta. E justamente aqui reside um ponto fundamental que pode tornar o método mais ou menos preciso e eficaz no que se propõe. Por isso vamos analisar como alguns autores desenvolveram este tema.

FRARY, TIDEMAN & WATTS (1977), *apud* HANSON, HARRIS & BRENNAN (1987:5), usam funções da pontuação total no teste para obter as probabilidades de respostas nos itens. Os parâmetros das funções que dão as probabilidades de resposta para um item particular variam com as "dificuldades" de resposta do item marginal sobre o grupo de candidatos de interesse (*i.e.*, a proporção de escolha de uma alternativa de resposta particular do item), número de itens no exame, e a média total da pontuação no exame.

Já CODY (1985), *apud* HANSON, HARRIS & BRENNAN (1987:6-7), usa as "dificuldades" de resposta dos itens para o grupo de candidatos de interesse como as probabilidades de resposta dos itens necessárias no cálculo da binomial composta e sugere uma aproximação binomial baseada nas probabilidades médias dos itens. E alternativamente, também usa "dificuldades" de respostas dos itens para pessoas com pontuação no teste similares ao suspeito copiador. Isto requer a partição da faixa de pontuação do teste em intervalos e calcular-se as "dificuldades" de respostas dos itens para todos os intervalos de pontuação. As probabilidades da binomial composta para uma dada pessoa suspeita de cópia são então calculadas com base nas "dificuldades" de respostas dos itens no intervalo de pontuação em que esteja a pontuação do suspeito copiador.

Portanto, tanto FRARY, TIDEMAN & WATTS (1977) como CODY (1985), para estimar a probabilidade de resposta em um dado item, fazem suposições e propõem funções que servem para estimar essas probabilidades.

Segundo SOTARIDONA (2003), a estatística de WOLLACK (1997, 2004) é muito similar à proposta por FRARY, TIDEMAN & WATTS (1977). A diferença principal é na maneira que o valor esperado de coincidências é estimado; enquanto WOLLACK usa o modelo de resposta nominal condicional pelo nível de habilidade do copiador, pelo vetor de resposta das questões da fonte, e pelos parâmetros das questões; já FRARY, TIDEMAN & WATTS usam as distrações e dificuldades da teoria dos testes clássica e a razão entre o número de pontos corretos do copiador e a média do número de pontos corretos para todos os candidatos. Como mostrado naquele estudo e em SOTARIDONA e MEIJER (2001), se os parâmetros do item no modelo de resposta nominal podem ser estimados confiavelmente, "ω" parece ser a melhor escolha para detecção de cópia porque é sensível ao longo de todos os níveis de habilidade do copiador e pode também ser usado para detectar cópia de respostas para tamanhos de amostra pequenos.

Logo, pelo que vimos, um aspecto fundamental que não se pode olvidar é a questão de condicionar a probabilidade de resposta aos itens de acordo com o nível de habilidade ou conhecimento dos candidatos.

Segundo HANSON, HARRIS & BRENNAN (1987), *apud* BAY (1995) a probabilidade de coincidência da resposta da suposta fonte no item é estimada pela "dificuldade de resposta", isto é, dividindo-se o número de estudantes que deu a mesma resposta que a fonte para aquele item , dividido pelo número de estudantes que respondeu ao item.

Já BAY (1995) refere que, a estatística que propõe e aquela proposta por FRARY, TIDEMAN & WATTS (1977), são ambas baseadas na mesma distribuição, a binomial composta, portanto era esperado que elas teriam desempenho com semelhante efetividade. Além disso, era esperado que a de BAY (1995) teria desempenho melhor, já que não emprega uma aproximação pela distribuição normal padrão. Consequentemente, há necessidade de entender por que esta teve um desempenho tão ruim com respeito às taxas observadas de falso positivo. A única razão poderia ser a estimação de probabilidade de selecionar uma resposta particular, a qual é constante para todos os copiadores para a de BAY (1995), mas é função da habilidade estimada para a de FRARY, TIDEMAN & WATTS (1977).

Vemos que BAY (1995) atribui o pior desempenho do índice que propôs em relação ao de FRARY, TIDEMAN & WATTS (1977), a não ter usado uma função da

habilidade do candidato para estimar a probabilidade de selecionar uma resposta particular. Convém lembrar que ambas as estatísticas se baseiam na mesma distribuição, no entanto o índice de BAY (1995) calcula a distribuição binomial composta, enquanto o de FRARY, TIDEMAN & WATTS (1977), a aproxima pela distribuição normal; e mesmo assim, apresenta melhor desempenho.

Parece-nos evidente que quanto menos suposições fizermos acerca de como ocorre a probabilidade de resposta a um determinado item, mais robusto será o método. Isto porque quando fazemos mais suposições, estas podem ser, ou não, verdadeiras no caso real. Assim, menos suposições implicam em menor chance de o modelo proposto diferir do caso real.

Por essa razão, neste trabalho não supomos que um candidato é a fonte e outro copiador, nem tampouco propomos uma função específica ou modelo para estimar a probabilidade de resposta ao item. Vamos nos limitar a estimar a probabilidade de as respostas de ambos os candidatos coincidirem, se elas forem dadas de forma independente entre si.

Segundo WESOLOWSKY (2000), a maioria dos métodos na literatura adotam modelos basicamente similares. Probabilidades de respostas corretas e incorretas são estimadas de respostas reais de classes. Um ponto importante é como aproximar a probabilidade de que um candidato responderá corretamente uma dada questão. Parece razoável assumir que isso depende da pontuação total do candidato no exame, como também da dificuldade da questão em particular. Alguns trabalhos antigos, e mesmo alguns muito recentes, no entanto, não incorporam a habilidade do candidato no modelo. Uma abordagem para incorporar essa consideração é dividir a classe em estratos, tal que pode ser assumido que os candidatos em cada estrato são de habilidade aproximadamente igual (HARPP e HOGAN, 1993, 1996).

Assim, entendemos que a forma adotada por HARPP e HOGAN (1993, 1996) *apud* WESOLOWSKY (2000:3), assume menos hipóteses por não propor nenhuma função para calcular a probabilidade de resposta ao item, mas sim, apenas dividindo os candidatos em estratos de acordo com sua habilidade.

Pelo que vimos, parece bastante razoável supor que p^{i}_{jk} dependa do nível de habilidade dos candidatos j,k.

Para avaliar e classificar o nível de habilidade de cada candidato *j*, conta-se o número de questões que cada um acertou, de acordo com a grelha de respostas oficiais para toda a população de candidatos avaliada. Seja:

$$r_j^i = \begin{cases} 1, \text{se } x_j^i \text{ coincide com a grelha de respostas;} \\ 0, \text{se } x_j^i \text{ não coincide com a grelha de respostas.} \end{cases} \quad (25)$$

Então:

$$r_j = \sum_{i=1}^{I} r_j^i \quad (26)$$

onde:

x^i_j é a resposta do candidato *j* à questão *i;*

r^i_j é a nota do candidato *j* na questão *i;*

r_j : número de questões respondidas corretamente pelo candidato *j* (“nota”).

Agrupa-se a população de candidatos em estratos, de modo a reunir os que acertaram um número de questões semelhante no mesmo estrato, pois desta forma terão aproximadamente o mesmo nível de habilidade. A análise do tamanho ideal dos estratos, simulando-se várias possibilidades e comparando-se os resultados, é apresentada na seção 4.7.

Para uma dada questão *i*, que tem como possíveis respostas, digamos, cinco alternativas nominadas de *A, B, C, D* ou *E*, a probabilidade da resposta à questão *i* de dois candidatos *j* e *k* coincidirem, resume-se à soma das probabilidades de ambos responderem simultaneamente a mesma alternativa. Desta forma,

$$p_{jk}^i = P\left(x_j^i = x_k^i\right) = P\left(x_j^i = \mathrm{A} \wedge x_k^i = \mathrm{A}\right) + P\left(x_j^i = \mathrm{B} \wedge x_k^i = \mathrm{B}\right) + P\left(x_j^i = \mathrm{C} \wedge x_k^i = \mathrm{C}\right) + P\left(x_j^i = \mathrm{D} \wedge x_k^i = \mathrm{D}\right) + P\left(x_j^i = \mathrm{E} \wedge x_k^i = \mathrm{E}\right) \quad (27)$$

Assim, se conseguirmos estimar a probabilidade de cada candidato responder cada alternativa para cada questão *i*, podemos estimar a probabilidade de a resposta de cada par de candidatos *j*,*k* para cada questão *i* coincidir.

Para cada questão *i* entre 1 e *I* e para cada estrato de candidatos, classificados de acordo com seus respetivos números de respostas certas r_j , estima-se a probabilidade de cada candidato pertencente àquele estrato responder cada uma das alternativas possíveis, tomando-se a respetiva frequência relativa de cada alternativa, no estrato em que se encontra o candidato (de acordo com seu nível de habilidade); frequência, esta, que naturalmente variará de acordo com o nível de dificuldade da questão e das alternativas possíveis e com o nível de habilidade daquele estrato de candidatos. Assim, tomamos

$$p_j^{iA} = P\left(x_j^i = A\right) \approx fr_j^{iA} \quad (28)$$

Onde:

$p_j^{iA} = P\left(x_j^i = A\right)$ é a probabilidade do candidato *j* responder alternativa *A* à questão *i*;

fr_j^{iA} é a frequência relativa de respostas *A* à questão *i* no estrato de habilidade do candidato *j*.

3.7 – Cálculo das frequências relativas de alternativas de respostas

Para cada estrato de candidatos (classificados pelo nível de habilidade), para cada questão *i*, separamos as respostas dadas por todos os candidatos por cada alternativa possível, contamos o número de respostas em cada alternativa e dividimos pelo número total de respostas, calculando assim, a frequência relativa de cada alternativa de resposta, para cada uma das questões *i*, para cada estrato de candidatos.

Supondo que cada questão *i* possa receber as alternativas de respostas:

$$v_i = \{A, B, C, D, E, sem\ resposta, nula\} \quad (29)$$

Calculam-se as respetivas frequências relativas:

$$fr_j^{iA} = \frac{nº\ respostas\ A\ dadas\ à\ questão\ i\ no\ estrato\ do\ candidato\ j}{n}$$

$$fr_j^{iB} = \frac{nº\ respostas\ B\ dadas\ à\ questão\ i\ no\ estrato\ do\ candidato\ j}{n}$$

$$fr_j^{iC} = \frac{nº\ respostas\ C\ dadas\ à\ questão\ i\ no\ estrato\ do\ candidato\ j}{n}$$

$$fr_j^{iD} = \frac{nº\ respostas\ D\ dadas\ à\ questão\ i\ no\ estrato\ do\ candidato\ j}{n}$$

$$fr_j^{iE} = \frac{nº\ respostas\ E\ dadas\ à\ questão\ i\ no\ estrato\ do\ candidato\ j}{n}$$

$$fr_j^{i\ sem\ resposta} = \frac{nº\ respostas\ "sem\ resposta"\ dadas\ à\ questão\ i\ no\ estrato\ do\ candidato\ j}{n}$$

$$fr_j^{i\ nula}, = \frac{nº\ respostas\ nulas\ dadas\ à\ questão\ i\ no\ estrato\ do\ candidato\ j}{n} \quad (30)$$

Repetem-se os cálculos:

- para toda questão *i* variando de 1 a *I*;
- para todos os estratos de habilidades de candidatos existentes.

3.8 - Estimação da probabilidade de certo número de respostas coincidentes

Com todas as frequências relativas (30) calculadas, armazenadas em uma tabela, poder-se-ão estimar todos os p^i_{jk}, aplicando-se (28) em (27).

Com todos os p^i_{jk}, estimam-se E(m_{jk}) (5) e Var(m_{jk}) (6).

Como já se conhece m_{jk} (4), através da contagem das respostas coincidentes entre os candidatos j,k , pode-se então estimar a probabilidade de ocorrência de um dado c_{jk}(7), utilizando-se a distribuição normal como aproximação, pois $c_{jk} \xrightarrow{D} \mathcal{N}(0,1)$ (23):

$$c_{jk} = \left[\frac{m_{jk} - E(m_{jk})}{\sqrt{Var(m_{jk})}}\right] \approx \mathcal{N}(0,1) \tag{31}$$

onde $\mathcal{N}(0,1)$ é a distribuição de probabilidade Normal, em que a função de distribuição é dada por

$$\Phi: \mathbb{R} \to \mathbb{R}_+ : \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt \, , \forall x \in \mathbb{R} \tag{32}$$

3.9 – Correção de continuidade

Deve-se notar que a sequência $\underline{u_{jk}} = (u^1_{jk}, u^2_{jk}, \ldots, u^I_{jk})$ (3), é uma sequência de variáveis aleatórias binárias, que podem assumir valores 0 ou 1, portanto, trata-se de uma sequência de variáveis aleatórias discretas.

Pelo Teorema do Limite Central aplicado à soma de variáveis não identicamente distribuídas, pudemos aproximar m_{jk} (4) que é uma soma de variáveis discretas, pela distribuição normal, e aproximar c_{jk}, (7) por uma distribuição normal padrão $\mathcal{N}(0,1)$, que é uma distribuição de variável aleatória contínua.

Estamos interessados em calcular a probabilidade de ocorrência de um certo número de ocorrências m_{jk}, que é uma variável aleatória discreta, utilizando, para tal, a função de distribuição Normal. Como esta é uma distribuição de probabilidades absolutamente contínua, significa que quase certamente uma variável aleatória com distribuição Normal nunca é igual a um valor fixo.

Devido à aproximação de uma variável discreta ser estimada por uma distribuição de probabilidades contínua, é conveniente aplicar a seguinte correção de continuidade em c_{jk}, a fim de melhorar a precisão da aproximação utilizada, conforme também refere ROSS (2004):

$$c_{jk} \approx \Phi\left[\frac{m_{jk}+0{,}5-E(m_{jk})}{\sqrt{Var(m_{jk})}}\right] - \Phi\left[\frac{m_{jk}-0{,}5-E(m_{jk})}{\sqrt{Var(m_{jk})}}\right] \quad (33)$$

3.10 – Pressupostos de validade e hipóteses

O objetivo deste trabalho é apresentar um método que seja aplicável a exames de larga escala com questões do tipo múltipla-escolha, que permita identificar se o exame foi idôneo ou se ocorreu algum tipo de fraude e, neste caso, identificar os fraudadores.

Até a seção anterior, desenvolvemos uma forma de estimar c_{jk} (33), que corresponde à ocorrência de um determinado número de coincidências m_{jk} (4) entre um par de candidatos quaisquer *j*,*k*.

Para chegarmos até esse ponto, foram necessários assumir os seguintes pressupostos:

I. Independência local entre as respostas de todos os pares de candidatos *j*,*k*.
II. Sequência de variáveis u_{jk} (3) é uniformemente limitada.
III. A soma das variâncias Var(m_{jk}) (6) das variáveis da sequência u_{jk} tende para infinito, quando o número de termos tende para infinito.

Se esses três pressupostos forem atendidos, aplica-se o Teorema do Limite Central à soma das variáveis m_{jk} e qualquer que seja a sua distribuição, convergirá para uma distribuição Normal, desde que o número de itens da sequência seja suficientemente grande.

Quanto aos pressupostos I e II, estão satisfeitos, pois u_{jk} é uma sequência com *I* termos, cujos valores são somente 0 ou 1. Portanto, é uniformemente limitada; e a soma das suas variâncias, Var(m_{jk}), é finita e tende a infinito quando o número de termos *I* tende a infinito.

Quanto ao pressuposto I, será avaliado através de testes de hipóteses. Considerando as seguintes hipóteses:

H_0 : as respostas de todos os candidatos *j*,*k* foram realizadas de forma independente entre si.

H_1 : as respostas de alguns candidatos *j*,*k* (ao menos um par) não foram realizadas de forma independente entre si.

Se a hipótese nula não for rejeitada, então o pressuposto I está satisfeito e podemos aproximar a distribuição de m_{jk} pela Normal.

Definimos α como sendo o nível de significância adotado para o teste de hipóteses. Ao avaliarmos aplicando a distribuição Normal, se a conclusão do teste for não rejeitar a hipótese nula, então concluiremos que não foi encontrada nenhuma evidência que as respostas entre todos os candidatos *j,k* não foram dadas de forma independente e logo, não foi encontrada nenhuma evidência de fraude no exame.

No sentido contrário, após avaliarmos o teste de hipóteses, se a conclusão for a de rejeitar a hipótese nula, aceitando a hipótese alternativa, então concluiremos que há evidências estatisticamente significativas de que respostas de alguns candidatos *j,k* (no mínimo um par), não foram dadas de forma independente e logo, há probabilidade estatisticamente significativa para um nível de significância α de ter havido fraude no exame.

3.11 – Testes de hipóteses através do valor-p

Para avaliarmos a rejeição ou não da hipótese-nula, utilizaremos o teste de hipóteses por valor-p , chamado como probabilidade de significância, que corresponde a, considerando a hipótese nula como verdadeira, a probabilidade de obter–se uma estatística de teste igual ou mais extrema que a estatística observada a partir de uma amostra da população. Adotaremos um nível de significância α : se o valor-p for menor que o nível de significância α adotado, rejeita-se a hipótese nula.

Quanto ao valor de nível de significância α a ser adotado e consequentemente, de valor-p que será utilizado como limite para se rejeitar ou não a hipótese-nula, este tema será desenvolvido e explicado no capítulo seguinte deste trabalho, na seção 4.5.

3.12 – Erros associados ao teste de hipóteses

Nesta altura, é oportuno definirmos os erros associados aos testes de hipóteses que aplicaremos. Seja o erro tipo I ou falso positivo:

E_1 : o método considerar um candidato idôneo como fraudador.

E seja o erro tipo II ou falso negativo:

E_2 : o método considerar um candidato fraudador como idôneo.

Como pretendemos identificar candidatos que tenham cometido algum tipo de fraude, que possivelmente serão acusados de tal feito ou que talvez sofrerão algum tipo de consequência desse suposto ato ilícito, nossa melhor escolha certamente é configurar e calibrar o método de modo a minimizar, com a maior segurança possível, a ocorrência de

erros tipo I (falsos positivos), para não sinalizar um inocente como sendo fraudador; mesmo pagando o preço de cometer possíveis erros tipo II (falsos negativos), ou seja, deixando de sinalizar algum eventual fraudador.

No mesmo sentido, BAY (1995) refere que para decidir se um par de candidatos têm ou não uma similaridade de respostas altamente incomum, um valor-limite de número de coincidências tem que ser estabelecido. Para fazer isso, estes dois tipos de erros potenciais têm que ser considerados. Falsos positivos (também referidos como erros tipo I) ocorrem quando um par de candidatos é acusado de cópia quando a cópia não aconteceu. Falsos negativos (também referidos como erros tipo II) ocorrem quando um par de candidatos que copiou não é acusado de fazê-lo. Um valor-limite de número de coincidências mais alto conduz a uma taxa de falsos negativos mais alta e falsos positivos mais baixa. A estratégia em definir o valor-limite de número de coincidências envolve especificar uma taxa de falso positivo; e a escolha óbvia é reduzir a probabilidade de falsamente acusar candidatos inocentes por esse tipo de erro, pois tem consequências mais sérias que o caso contrário.

3.13 - Fator de correção de Bonferroni

O método de análise desenvolvido até aqui, consiste em compararmos as respostas de cada candidato com a de todos os outros, a fim de identificar, caso exista, uma quantidade de coincidências que não seriam esperadas sob o ponto de vista estatístico. Isso faz com que, para N candidatos, cada candidato j será comparado com N-1 outros candidatos, contando-se suas respostas coincidentes dois a dois.

Quando se avalia a significância de um evento raro, não através de um único teste, mas através de múltiplos testes, produz-se um efeito que é por vezes conhecido como *data dredging* , ou "inflação de significância" (α *inflation*).

Em outras palavras, pode-se entender o efeito da seguinte maneira: quando se analisa um concurso com 20000 candidatos, por exemplo, faz-se quase 200 milhões de comparações entre pares de candidatos. Desse modo, um evento tão raro quanto o de acertar o prêmio máximo do concurso da lotaria "Euromilhões", jogando-se a aposta mínima, que ocorre com a probabilidade aproximada de uma em 140 milhões, seria esperado que ocorresse uma vez quando se compara as provas dos 20000 candidatos do concurso entre si. Portanto, um evento tão raro como o de ganhar no Euromilhões, seria esperado que

aparecesse uma vez nos cruzamentos de respostas dos candidatos desse concurso, sem que isso significasse que teria havido qualquer tipo de fraude.

Assim, quando se fazem múltiplos testes de significância, temos que corrigir o cálculo da probabilidade do evento, de modo que a ocorrência de eventos raros não seja incorretamente interpretada.

Existem vários métodos para se minimizar esse problema das comparações múltiplas. Um dos métodos de ajuste usado para esse tipo de situação, por diversos pesquisadores do assunto é conhecido como o fator, ou a correção de Bonferroni, e consiste em dividir o valor da significância α, pelo número de testes realizados, conforme referido, por exemplo, em CIZEK e WOLLACK (2017).

Vários estudos já foram desenvolvidos usando esse método de ajuste, demonstrando que é conservador e diminui a ocorrência de erros tipo I, conforme os trabalhos mencionados a seguir.

A esse respeito, LEWIS e THAYER (1998) mencionam que uma modificação envolve a multiplicação do índice pelo número de comparações sendo feitas. Esta modificação pretende controlar a taxa de erro tipo I (um erro tipo I neste caso consiste em rejeitar a hipótese nula que dois candidatos estavam trabalhando independentemente, quando a hipótese é realmente verdadeira).

Também WESOLOWSKY (2000) refere que a principal preocupação é a prevenção de falsas acusações. O modelo é adequado para pesquisar grandes classes e os resultados são simples de interpretar. Simulação e a correção de Bonferroni são usados para prevenir falsas acusações devido ao *data dredging*. A ênfase é em prevenir falsas acusações (controlando o erro tipo I) e não em aumentar o número de detecções. Há duas maneiras básicas de usar um programa de detecção estatística. Uma é buscar evidência se há alguma razão anterior para suspeitar de candidatos em particular, por exemplo, um relatório de um fiscal sobre comportamento suspeito durante um exame. Este é um teste de hipótese padrão. O outro uso é mapeamento; que é a comparação das respostas de todos os pares de candidatos em uma tentativa de detectar cópia, da qual não havia evidência prévia. Os níveis de corte de Bonferroni usados são muito conservadores.

Ainda, MCMANUS, LISSAUER & WILLIAMS (2005) mencionam que também utilizam a correção de Bonferroni como uma correção para inflação de significância (teste de significância múltipla). Dado um grande número de pares analisados, então se

argumenta que com certeza é o caso de que alguns pares irão atingir um nível arbitrário de significância. Certamente se está lidando com números muito grandes de pares de candidatos e níveis muito pequenos de probabilidade. O ajuste de Bonferroni está, portanto, fazendo seu trabalho propriamente e evitando erros tipo I.

Entre os diversos métodos que utilizam a correção de Bonferroni, citamos como exemplos, o de WOLLACK (2004) e o de WESOLOWSKY (2000).

Nas múltiplas comparações de respostas, analisa-se e conta-se o número de coincidências m_{jk} (4) de cada candidato *j* com cada um dos outros *N*-1 candidatos, e estimam-se as respetivas probabilidades de ocorrência de c_{jk} (33) através da distribuição normal padrão $\mathcal{N}(0,1)$). Mas para cada candidato *j*, realizam-se (*N*-1) comparações, de modo que em relação aos valores-p absolutos que consideraremos nos testes de hipóteses, realizaremos a seguinte correção de Bonferroni:

$$valor - p_c(c_{jk}) = \left(\Phi \left[\frac{m_{jk}+0{,}5-E(m_{jk})}{\sqrt{Var(m_{jk})}} \right] - \Phi \left[\frac{m_{jk}-0{,}5-E(m_{jk})}{\sqrt{Var(m_{jk})}} \right] \right).(N - 1) \qquad (34)$$

onde $valor\text{-}p_c(c_{jk})$ é o valor-p corrigido de c_{jk} .

Nos testes de hipóteses, esse valor-p corrigido por Bonferroni será comparado com o nível de significância α adotado para decidir-se rejeitar, ou não, a hipótese-nula, de modo a minimizar os erros tipo I, mesmo correndo o risco de se aumentar os erros de tipo II, pois consideramos que seja menos grave deixar de identificar um candidato que tenha fraudado o exame (erro tipo II), do que identificar erroneamente como fraudador, um candidato que não tenha fraudado (erro tipo I).

3.14 - Cálculo da probabilidade conjunta de respostas coincidentes

Supondo que determinado candidato *j* apresente simultaneamente eventos de grande coincidência de respostas com mais de um candidato, por exemplo, candidatos *k* e *l* com os respetivos c_{jk} e c_{jl}. Então, sob a validade da hipótese nula de independência entre as respostas de todos os candidatos, podemos dizer que a probabilidade conjunta de ocorrência simultânea c_{jk} e c_{jl} será:

$$c_j = c_{jk} \,.\, c_{jl} \qquad (35)$$

Onde c_j representa a probabilidade conjunta de ocorrência simultânea de m_{jk} e $m_{jl,}$ respetivamente entre os pares de candidatos *j*,*k*, e *j*,*l*.

A expressão anterior pode ser estendida para quaisquer quantidades de eventos raros ocorridos simultaneamente, e representa a probabilidade conjunta associada à ocorrência simultânea de mais de um evento raro entre o mesmo candidato *j* e outros candidatos distintos. Consideramos como eventos raros, aqueles que ocorrem entre dois candidatos *j,k,* cujo *valor-p*$_c$ (c_{jk}) < α, lembrando que a determinação do nível de significância α que adotaremos, será realizada através da simulação de dados, conforme apresentado no capítulo seguinte, na seção 4.5.

3.15 – Síntese passo-a-passo das etapas e fórmulas do método

A fim de auxiliar na compreensão do método desenvolvido, apresentamos a seguir a síntese dos cálculos que devem ser realizados, na ordem em que devem ser efetuados, com suas respetivas fórmulas e suas identificações numéricas.

a) Para cada par de candidatos *j,k* (*j* variando de 1 a *N*-1; e *k* variando de 2 a *N*), comparam-se as respostas de ambos a cada questão variando de 1 a *I*, e calculam-se:

$$u_{jk}^{i} = \begin{cases} 1, \text{se } x_j^i = x_k^i \ ; \\ 0, \ \text{se } x_j^i \neq x_k^i \ . \end{cases} \quad (1)$$

$$m_{jk} = \sum_{i=1}^{I} u_{jk}^{i} \quad (4)$$

b) Para cada candidato *j* de 1 a *N*, para cada questão *i* de 1 a *I*, calculam-se:

$$r_j^i = \begin{cases} 1, \text{se } x_j^i \text{ coincide com a grelha de respostas;} \\ 0, \text{se } x_j^i \text{ não coincide com a grelha de respostas;} \end{cases} \quad (25)$$

$$r_j = \sum_{i=1}^{I} r_j^i \quad (26)$$

c) Divide-se a população de candidatos em estratos de habilidade de acordo com as respetivas notas r_j, conforme explicado na seção 4.7 a seguir.

d) Para cada estrato de candidatos, para cada questão *i*, supondo que cada questão *i* possa receber as alternativas de respostas:

$$v_i = \{A, B, C, D, E, sem\ resposta, nula\} \quad (29)$$

Calculam-se as respetivas frequências relativas:

$$fr_j^{iA} = \frac{n^o \text{ respostas A dadas à questão i no estrato do candidato j}}{n}$$

$$fr_j^{iB} = \frac{n^o \text{ respostas B dadas à questão i no estrato do candidato j}}{n}$$

$$fr_j^{iC} = \frac{\text{nº respostas C dadas à questão i no estrato do candidato j}}{n}$$

$$fr_j^{iD} = \frac{\text{nº respostas D dadas à questão i no estrato do candidato j}}{n}$$

$$fr_j^{iE} = \frac{\text{nº respostas E dadas à questão i no estrato do candidato j}}{n}$$

$$fr_j^{i\ sem\ resposta} = \frac{\text{nº respostas "sem resposta" dadas à questão i no estrato do candidato j}}{n}$$

$$fr_j^{i\ nula}, = \frac{\text{nº respostas nulas dadas à questão i no estrato do candidato j}}{n} \quad (30)$$

Repetem-se os cálculos:

- para toda questão *i* variando de 1 a *I*;

- para todos os estratos de habilidades de candidatos existentes.

e) Para cada par de candidatos *j*,*k* (*j* variando de 1 a *N*-1; e *k* variando de 2 a *N*), calculam-se:

$$p_{jk}^i = P\left(x_j^i = x_k^i\right) = P\left(x_j^i = \mathrm{A} \wedge x_k^i = \mathrm{A}\right) + P\left(x_j^i = \mathrm{B} \wedge x_k^i = \mathrm{B}\right) + P\left(x_j^i = \mathrm{C} \wedge x_k^i = \mathrm{C}\right) + P\left(x_j^i = \mathrm{D} \wedge x_k^i = \mathrm{D}\right) + P\left(x_j^i = \mathrm{E} \wedge x_k^i = \mathrm{E}\right) \quad (27)$$

Onde, para todas $x_j^i \in v_i$:

$$p_j^{iA} = P\left(x_j^i = A\right) \approx fr_j^{iA} \quad (28)$$

f) Com os p_{jk}^i para todo par de candidatos *j*,*k*, estimam-se:

$$\hat{E}\left(m_{jk}\right) = \sum_{i=1}^{I} p_{jk}^i \quad (5)$$

$$\widehat{Var}\left(m_{jk}\right) = \sum_{i=1}^{I} p_{jk}^i . \left(1 - p_{jk}^i\right) \quad (6)$$

$$c_{jk} = \frac{m_{jk} - \hat{E}\left(m_{jk}\right)}{\sqrt{\widehat{Var}\left(m_{jk}\right)}} \quad (7)$$

$$valor - p_c(c_{jk}) = \left(\Phi\left[\frac{m_{jk}+0{,}5-E\left(m_{jk}\right)}{\sqrt{Var\left(m_{jk}\right)}}\right] - \Phi\left[\frac{m_{jk}-0{,}5-E\left(m_{jk}\right)}{\sqrt{Var\left(m_{jk}\right)}}\right] \right).(N - 1) \quad (34)$$

g) Com a lista dos valores-p_c (c_{jk}) ordenada em ordem crescente, compara-se o menor valor-p_c (c_{jk}) com o nível de significância α definido para o teste de hipóteses, conforme explicado no capítulo seguinte, na seção 4.5, para avaliar a rejeição ou não da hipótese-nula H_0.

4 – APLICAÇÃO DO MÉTODO

4.1 – Escolha dos parâmetros a estudar

Todo concurso ou exame de larga aplicação possui alguns parâmetros intrínsecos que influenciam na possibilidade de este método de identificação de fraudes ser ou não aplicável, bem como na sua maior ou menor eficácia.

Selecionamos alguns desses parâmetros para analisar, relevantes para determinar quais os limites de aplicação do método desenvolvido, para que, quando for aplicado em uma situação real na análise e na sinalização se houve ou não fraude nas respostas de um determinado exame, os resultados sejam eficazes e confiáveis. Os parâmetros analisados e cuja influência avaliamos são os descritos na tabela 4.1.

Tabela 4.1 – Parâmetros estudados para avaliar a eficácia e robustez do método

Tamanho dos estratos de candidatos por nível de habilidade: 50, 100, 200 e 1 desvio padrão*
Número de questões do exame (*I*): 25, 50, 75, 100 e 200
Número de candidatos que realizaram o exame (*N*): 50, 100, 1000, 10000 e 100000
Número de candidatos que fraudaram: 2, 5, 10 e 50
Percentagem de respostas copiadas nas fraudes simuladas: 50%, 60% e 75%
Quantidade de alternativas possíveis em cada questão: 2 (certo, errado) e 5 (A, B, C, D e E)

*Observação: o tamanho de estrato correspondente a "1 desvio padrão" será explicado na seção 4.7 a seguir.

A metodologia utilizada para se avaliar a influência de cada um desses parâmetros foi variar apenas um parâmetro de cada vez, mantendo-se todos os demais fixados, e analisar como a variação daquele parâmetro influenciava nos resultados e na eficácia do método.

As combinações de parâmetros que foram simuladas e avaliadas são as que estão descritas nas seções 4.7 a 4.12.

4.2 – Geração de bancos de dados fictícios por simulação de números pseudoaleatórios

Neste trabalho, a aplicação do método desenvolvido para identificação de fraudes foi realizada utilizando-se bancos de dados fictícios, criados por simulação matemática, correspondentes a respostas de candidatos a exames de larga escala. Isso traz várias vantagens, pois além de não tratarmos com dados reais de pessoas, que estariam revestidos de proteção de confidencialidade e sigilo, podemos controlar o espaço amostral

de acordo com cada parâmetro que queremos analisar isoladamente e testar a eficácia do método em várias situações diferentes (Tabela 4.1).

Os bancos de dados fictícios foram criados por simulação matemática, utilizando uma técnica para a geração de números pseudoaleatórios, implementada através de um algoritmo que criamos no software estatístico "R", de uso livre e disponível em https://cran.r-project.org, da forma que explicamos a seguir.

Para um dado conjunto de parâmetros (número de questões do exame, número de candidatos totais, quantidade de alternativas possíveis em cada questão), criamos um primeiro banco de dados de respostas através da geração de números pseudoaleatórios, que chamaremos de "independentes", o qual corresponde ao banco de dados onde a hipótese nula de independência entre as respostas é respeitada, uma vez que as respostas foram obtidas por um processo de geração de números pseudoaleatórios, sem qualquer dependência entre as respostas.

Para cada questão, a probabilidade que um determinado candidato escolha uma determinada alternativa, varia para cada alternativa dentro de uma mesma questão, e varia também, de questão para questão. Essas variações devem ser levadas em conta na simulação dos dados. Assim, geramos um primeiro ficheiro tipo "csv", somente com respostas geradas aleatoriamente, de forma independente, que simula um exame onde não houve fraude entre os candidatos.

Em seguida simulamos uma fraude, criando outro banco de dados com os mesmos parâmetros do banco inicial, que só continha respostas aleatórias e independentes, apenas acrescentando ao fim, um certo número de candidatos cujas respostas são fortemente dependentes entre si. Desta forma o banco terá uma grande maioria de respostas independentes, e algumas respostas geradas por forte dependência, o que simula exatamente o que acontece nos casos de fraude.

A percentagem de respostas copiadas nas fraudes simuladas dentro do algoritmo implementado determinará quão dependentes serão as respostas entre si, e poderemos alterá-la para mais ou para menos, a fim de testar a sensibilidade, e até calibrar o nosso sistema de identificação de fraudes.

Assim, com os mesmos parâmetros fixados, criamos um segundo banco de dados, chamado de "dependentes", através da inserção no primeiro banco de dados "independentes", de um certo número de candidatos com uma certa quantidade de respostas

necessariamente iguais entre si, simulando assim, artificialmente, uma "fraude no exame", com um número de "fraudadores" que escolhemos e com uma quantidade de respostas copiadas que também escolhemos. Dessa forma, necessariamente quebramos a hipótese-nula de independência entre as respostas. Além disso, sabemos quem são os candidatos inseridos, que correspondem aos "fraudadores", bem como a percentagem de cópia das respostas, escolhida por nós na geração dos dados. Assim, geramos um segundo ficheiro tipo "csv", composto pelo mesmo ficheiro anterior que continha somente candidatos com respostas independentes, mas "contaminado" com candidatos com respostas dependentes.

Para que possamos classificar os candidatos em estratos de habilidade, e necessário determinar o número de respostas corretas de cada candidato, e para isso, é necessário em determinado passo, que conheçamos a grelha de respostas oficial do exame, para que seja comparado com as respostas dos candidatos. Utilizamos as respostas simuladas mais comuns entre os candidatos, para atribuir-lhes como sendo a grelha de respostas oficial, gerando um terceiro ficheiro tipo "csv".

Para cada uma das situações contidas na tabela 4.1, geramos três bancos de dados, que correspondem às situações apresentadas na tabela 4.2.

Tabela 4.2 – Correspondência entre os bancos de dados simulados e situação real

Banco de dados simulado	Correspondência à situação real
respostas_ind.csv : todos os candidatos com respostas independentes	Exame sem fraude
respostas_dep.csv : alguns candidatos com respostas dependentes	Exame com fraude
gabarito.csv : grelha de respostas simulada	Grelha de respostas oficial do exame

Os dados de saída do algoritmo são três ficheiros: *respostas_ind.csv*, *respostas_dep.csv* e *gabarito.csv*, que serão então, analisados pela metodologia até aqui apresentada, da maneira explicada na seção a seguir.

No anexo I apresentamos o algoritmo (em *R*) que desenvolvemos e utilizamos na geração dos bancos de dados de respostas fictícias, exemplificando através de um suposto exame com os seguintes parâmetros: 100 questões, com 5 alternativas de respostas (letras "A" a "E"), realizado por 1009 candidatos, com 10 candidatos fraudadores, com 75% (em média) de respostas compartilhadas entre si (dependentes).

4.3 – Comparação das respostas dos pares de candidatos e cálculos preliminares

Para iniciarmos a aplicação do método aos conjuntos de dados, que contém uma lista com *N* candidatos, cada um com uma sequência de *I* respostas às questões, percorremos as seguintes etapas:

a) Comparar a sequência de respostas a todas as questões de 1 a *I*, de cada candidato *j*, com a sequência de respostas do ficheiro "grelha de respostas oficiais", e contar o número de coincidências, obtendo-se assim, r_j (26); que corresponderá à "nota" do candidato *j*.

b) Com todos os r_j, calcular a "nota" média dos candidatos e o desvio padrão da "nota". Formar então, estratos para agrupar os candidatos por seu nível de habilidade, conforme explicado na seção 4.7 a seguir.

c) Para cada par de candidatos *j,k* (*j* variando de 1 a *N*-1; *k* variando de 2 a *N*), comparar as respostas de ambos a cada questão de 1 a *I*, obtendo-se os valores de u^i_{jk} (1) e consequentemente, da sequência u_{jk} (3).

d) Somar os termos de u_{jk}, obtemos m_{jk} (4) para todo par de candidatos *j,k*.

e) Para cada estrato de habilidade de candidatos, para cada questão *i*, separar as respostas dadas por todos os candidatos daquele estrato, por cada alternativa de resposta possível contida em v_i (29), contar o número de respostas em cada alternativa e dividir pelo número total de respostas do estrato, calculando assim, as frequências relativas de cada alternativa de resposta, para cada uma das questões *i*, para cada estrato de candidatos (30).

f) Estimar os p^i_{jk} (27), para toda questão *i* de 1 a *I*, para todo par de candidatos *j,k*, através de $p_j^{iA} = P\left(x_j^i = A\right) \approx fr_j^{iA}$ (28), utilizando-se as frequências relativas já calculadas.

g) Com todos os p^i_{jk}, estimar as $E(m_{jk})$ (5) e $Var(m_{jk})$ (6) e a estatística c_{jk} (7) para todo par de candidatos *j,k*.

h) Criar uma lista em ordem decrescente de c_{jk} para todos os pares de candidatos *j,k*.

Para realizar esse processo repetitivo milhares ou milhões de vezes, dependendo do número de candidatos *N* e do número de questões *I*, desenvolvemos um algoritmo lógico sequencial com todas essas etapas e recorremos a uma rotina computacional criada especificamente para este fim, que se encontra apresentada no Anexo II.

4.3.1 – Erro associado aos cálculos computacionais

Essa linguagem de programação informática utiliza nessa rotina variáveis do tipo "*float*", que efetua os cálculos com 48 casas decimais. Portanto, podemos admitir que o erro existente nos cálculos computacionais da rotina será da ordem de 10^{-48}. Após os cálculos, na apresentação dos resultados, o valor de c_{jk} (7) é arredondado para 6 casas decimais.

Com os valores de c_{jk}, com 6 casas decimais, estimamos os valores-p corrigidos (34), utilizando a função de distribuição Normal através do software *Excel* (da *Microsoft*), que possui precisão de cálculo de 15 algarismos significativos.

Mas como os valores de c_{jk} são arredondados para 6 casas decimais nos dados de saída da rotina computacional feita em GNU Awk versão 4.1.4, o erro associado a todos os cálculos computacionais será da ordem de 10^{-6}.

4.4 – Método de análise dos bancos de dados de respostas

No primeiro banco de dados, que chamamos de "independentes" aplicamos o método descrito por todo o capítulo 3, conforme explicado na seção anterior, e determinamos, para aquele conjunto de parâmetros fixados, qual o maior nível de respostas coincidentes possível de se obter entre os candidatos, uma vez que a hipótese de independência entre as respostas foi respeitada, que equivale a um exame não fraudado. Isto é, calcula-se os c_{jk} (7) para todos os pares de candidatos j,k. O menor c_{jk} será considerado na estimação do valor-p corrigido (34) para aquele conjunto de parâmetros fixados, correspondendo ao menor nível de significância que de fato ocorreu, sob a validade da hipótese-nula de independência.

Novamente aplicamos o método para o segundo banco de dados, que chamamos de "dependentes", determinando qual o nível de respostas coincidentes obtidos agora, que a "fraude" simulada faz parte do conjunto. Isto é, calcula-se os c_{jk} para todos os pares de candidatos j,k. Os menores c_{jk} agora obtidos, e seus correspondentes valores-p corrigidos, serão comparados com os do correspondente banco de dados "independentes". O que se espera deste método de análise, é que, de preferência todos, mas senão pelo menos a maioria, os candidatos com respostas dependentes devem estar entre os menores c_{jk} e seus correspondentes valores-p corrigidos, com valores bastante inferiores ao mínimo do banco de dados "independentes", que respeita a hipótese-nula de independência.

Por fim, comparamos os níveis de significância das respostas coincidentes ocorridas, através dos menores c_{jk} e seus correspondentes valores-p corrigidos, entre o primeiro banco de dados, gerado com respostas totalmente independentes, e o segundo banco de dados, gerado com os mesmos parâmetros de número de candidatos, número de questões e de número de alternativas em cada questão, mas que foi criado com a inserção forçada de um certo número de candidatos com um certo número de repostas copiadas entre si. O método deverá destacar os candidatos que foram inseridos com cópia de respostas através de menores c_{jk} e seus correspondentes valores-p corrigidos, que correspondem aos níveis de significância de respostas coincidentes entre si, bastante inferiores que os menores c_{jk} e seus correspondentes valores-p corrigidos dos candidatos com respostas independentes obtêm entre si.

Isto nos levaria a concluir pela rejeição da hipótese-nula neste banco de dados onde foram inseridos candidatos com respostas dependentes, mostrando que o método é capaz de distinguir quando a hipótese-nula de independência é violada, que corresponderia em uma situação real de um exame, de algum tipo de fraude. E também que o método mostra, através dos menores c_{jk} obtidos, e seus correspondentes valores-p corrigidos, inferiores ao nível de significância mínimo adotado α, quais foram os candidatos que violaram a independência entre as respostas, identificando os fraudadores.

Repetimos esse processo para várias combinações de parâmetros descritos na Tabela 4.1, para analisar como que cada um dos parâmetros influencia na eficácia do método, ou seja, na sua capacidade de destacar os candidatos com respostas copiadas, dos demais, com respostas independentes. Analisamos cada um dos parâmetros separadamente e descrevemos os resultados nas seções 4.7 a 4.12.

4.5 - Determinação do nível de significância adotado para os testes de hipóteses

Como vimos no capítulo 3, a variável aleatória m_{jk} (4) é uma soma de variáveis aleatórias binárias de Bernoulli com probabilidades variáveis, que segue uma distribuição conhecida como Binomial Composta ou Binomial Generalizada e que, sob determinadas hipóteses, pode-se aplicar o Teorema do Limite Central para variáveis não identicamente distribuídas e pode-se aproximá-la por uma distribuição Normal. Entre essas hipóteses está a de independência local entre as variáveis binárias, que no nosso caso são as coincidências ou não-coincidências das respostas dos candidatos, dois a dois. Assim, se o

exame ou concurso não foi fraudado, haverá independência entre as respostas de todos os candidatos e a distribuição da soma de respostas coincidentes m_{jk} poderá ser aproximada pela Normal. Por outro lado, se houve fraude através do compartilhamento de respostas entre candidatos, a independência entre as respostas foi quebrada, a soma m_{jk} não seguirá mais a distribuição Normal.

Quando analisamos um determinado exame ou concurso real, não sabemos se houve ou não fraude, isto é, se a hipótese da independência foi respeitada ou não. É justamente isso que queremos descobrir com este método, quando tal hipótese é quebrada, o que configura dependência entre respostas de alguns candidatos (pelo menos um par), portanto, a ocorrência de fraude. Analisamos isso através do teste de hipóteses, determinando um valor-p, corrigido por Bonferroni, devido à multiplicidade de testes, abaixo do qual, rejeitaremos a hipótese-nula H_0 de independência entre as respostas dos candidatos, concluindo pela hipótese alternativa H_1, de não-independência entre algumas respostas e, portanto, indício estatisticamente significativo de violação da regra do exame que pressupõe que cada candidato responda às questões individualmente, de forma independente.

Mas qual deverá ser esse valor de significância α, abaixo do qual rejeitamos a hipótese de independência? Poderíamos utilizar como referência valores comumente utilizados na literatura em testes de significância; ou ainda, valores de referência que outros autores utilizaram em seus trabalhos nesta área de investigação por índices de similaridade de resposta. Mas as realidades daqueles estudos são diferentes dos casos que estamos estudando, quanto a vários aspectos, tais como: tamanhos de bases de candidatos, fraudes realizadas por grupos múltiplos de candidatos ao invés de apenas pares, entre outros.

Por essa razão, precisamos encontrar quais os limites de significância mínimos que de fato ocorrem dentro do nosso modelo de índice de similaridade c_{jk}, quando a hipótese-nula de independência entre as respostas é respeitada, o que equivale a dizer que o exame foi idôneo, sem fraude.

Para chegar a esses valores, servimo-nos das ferramentas de simulação, variando os parâmetros dos exames simulados, e verificando os valores-p mínimos que ocorrem. Assim conseguiremos definir quais os níveis de significância que deveriam ser considerados como limites, até os quais, a quantidade de respostas coincidentes seria considerada como resultado idôneo, fruto de respostas independentes dos candidatos, sem

que houvesse qualquer tipo de fraude. Esses níveis de significância limites separarão os casos de soma de coincidências suspeitas das não-suspeitas.

Através de simulação, com a geração de números pseudoaleatórios, utilizando o algoritmo listado no Anexo I, criamos diversos bancos de dados, de forma que as respostas de cada candidato são localmente independentes.

Em seguida aplicamos o método, conforme descrito nas seções anteriores 4.3 e 4.4, aos bancos de dados com respostas independentes, para todas as combinações de parâmetros apresentadas na tabela 4.3 e identificamos em cada caso, qual o menor valor-p corrigido (34) que de fato ocorreu.

Lembramos que, como todos esses bancos de dados contêm respostas independentes entre si, estes equivalem a um exame idôneo onde não houve qualquer compartilhamento de respostas entre os candidatos. Portanto, os menores valores-p corrigidos encontrados, ocorreram respeitando-se a hipótese-nula H_0 de independência entre as respostas.

O menor valor encontrado entre todos os casos simulados para o valor-p corrigido mínimo foi 0,002. Assim, decidimos adotar como nível de significância crítico para os testes de hipóteses, $\boldsymbol{\alpha = 0{,}001}$, metade do valor mínimo obtido em todas as simulações.

Portanto, quando obtivermos na comparação entre dois candidatos quaisquer *j,k*, um valor-p corrigido inferior ao valor crítico $\alpha = 0{,}001$, rejeitaremos a hipótese-nula H_0 de independência entre as respostas de todos os candidatos, aceitando hipótese-alternativa H_1 de que as respostas de alguns candidatos *j,k* (ao menos um par) não foram realizadas de forma independente entre si. Dito de outra forma, haverá indício estatisticamente significativo de que a regra do exame, que pressupõe que cada candidato responda às questões individualmente, de forma independente dos demais candidatos, teria sido, de alguma forma, violada.

Tabela 4.3 – Valor-p corrigido mínimo em bancos de dados simulados de respostas independentes

Nº questões (*I*)	Nº candidatos (*N*)	Valor-p (29) mínimo
200	1000	0,011
200	1000	0,020
200	1000	0,034
200	1000	0,048
100	25	0,003
100	25	> 0,569
100	25	> 0,569
100	25	> 0,569
100	50	0,008
100	50	0,018
100	50	0,036
100	50	0,084
100	100	0,022
100	100	0,028
100	100	0,035
100	100	0,048
100	1000	0,003
100	1000	0,004
100	1000	0,006
100	1000	0,006
100	1000	0,007
100	1000	0,008
100	1000	0,022
100	1000	0,028
100	1000	0,029
100	1000	0,029
100	1000	0,032
100	1000	0,040
100	1000	0,041
100	1000	0,042
100	1000	0,051
100	1000	0,065
100	1000	0,070
100	1000	0,076
100	1000	0,085
100	1000	0,094
100	1000	0,116
100	10000	0,002
100	10000	0,011
100	10000	0,013
100	10000	0,016
100	10000	0,017
100	10000	0,025
100	100000	0,002
100	100000	0,002
100	100000	0,002
100	100000	0,002
100	100000	0,003
100	100000	0,007
75	1000	0,028
75	1000	0,089
50	1000	0,007
50	1000	0,022
25	1000	0,032
25	1000	0,076

4.6 – Exemplo de aplicação do método em um exame simulado

A fim de exemplificar a aplicação do método, seguimos os passos descritos nas seções 4.3 e 4.4 e apresentamos nas tabelas 4.4 e 4.5 os resultados obtidos para um exame simulado com os seguintes parâmetros: 100 questões, com 5 alternativas de respostas (letras "A" a "E"), realizado por 1009 candidatos, com 10 candidatos fraudadores (e 999 independentes), com 75% (em média) de respostas compartilhadas entre os dependentes.

Inicialmente, utilizamos a rotina computacional apresentada no anexo II para compararmos as respostas somente dos 1000 candidatos independentes, efetuando os cálculos descritos na seção anterior; sendo os principais resultados apresentados na tabela 4.4 a seguir, que mostra os 10 pares de candidatos *j*,*k* com os menores c_{jk} (7) e seus correspondentes valores-p corrigidos (34) entre todas as 499500 combinações de pares de candidatos possíveis.

Tabela 4.4 – Exemplo de aplicação em exame com 1000 candidatos independentes

Candidato j	Nota j	Candidato k	Nota k	mjk(4)	E(mjk) (5)	$(Var(mjk))^{1/2}$	cjk (7)	valor-pc (34)
Candidato 689	**45**	**Candidato 755**	**57**	**21**	**41,262608**	**5,085488**	**3,984398**	**0,028653392**
Candidato 379	60	Candidato 859	57	64	44,379045	5,076597	3,864982	0,045799689
Candidato 583	56	Candidato 999	51	60	41,262608	5,085488	3,684482	0,090162461
Candidato 224	53	Candidato 939	53	59	40,193131	5,108937	3,68117	0,090829866
Candidato 93	51	Candidato 95	53	59	40,193131	5,108937	3,68117	0,090829866
Candidato 97	62	Candidato 291	56	63	44,379045	5,076597	3,668	0,095950764
Candidato 193	62	Candidato 462	52	61	42,380534	5,077752	3,666872	0,096324309
Candidato 821	50	Candidato 966	58	61	42,380534	5,077752	3,666872	0,096324309
Candidato 19	55	Candidato 352	56	61	43,159284	5,089353	3,505498	0,171095332
Candidato 217	57	Candidato 753	57	61	43,159284	5,089353	3,505498	0,171095332

Na primeira linha da Tabela 4.4, nota-se que o menor valor-p corrigido encontrado foi aproximadamente 0,029, obtido entre os candidatos de números 689 e 755. Como 0,029 > 0,001, nosso nível de significância α crítico adotado, então não rejeitamos a hipótese-nula H_0 de que todas as respostas entre os pares de candidatos *j*,*k* foram independentes. De fato, geramos todas as respostas dos 1000 candidatos, utilizando a rotina computacional apresentada no anexo I, de forma independente, e os resultados encontrados corroboram isso.

Em seguida, utilizando a rotina computacional apresentada no anexo I, no mesmo banco de dados com 1000 candidatos independentes, inserimos 9 candidatos, de números 1001 a 1009, que copiaram em média, 75% das respostas do candidato número

1000. Assim, sabemos que as respostas dos candidatos números 1 a 999 permanecem independentes, mas as dos candidatos 1000 a 1009 contêm, em média, 75% de respostas dependentes (copiadas).

Então utilizamos a rotina computacional apresentada no anexo II para compararmos as respostas dos 1009 candidatos, 999 independentes e 10 dependentes, efetuando os cálculos descritos na seção anterior; sendo os principais resultados apresentados na Tabela 4.5, que mostra os 50 pares de candidatos *j,k* com os menores c_{jk} e seus correspondentes valores-p corrigidos entre todas as 508536 combinações de pares de candidatos possíveis.

Insta ressaltar, que como há 10 candidatos com respostas dependentes, do número 1000 ao 1009, há 45 combinações de pares *j,k* possíveis entre eles. Assim, o resultado ideal da aplicação do método, seria que esses 45 valores-p corrigidos fossem os menores entre todas as 508536 combinações, todas com valor-p < 0,001, uma vez que violam a hipótese-nula de independência entre as respostas; e que o 46º menor valor-p corrigido, correspondente já a um par de candidatos *j,k* com respostas independentes, fosse por sua vez, > 0,001, uma vez que as respostas foram geradas de forma independente, e portanto, respeita a hipótese-nula H_0.

De fato, notamos que os 45 menores valores-p corrigidos ocorrem entre as 45 combinações de pares de candidatos *j,k* de números 1000 a 1009, justamente os que possuem respostas dependentes, sendo que o menor valor-p corrigido é aproximadamente $1{,}33 \times 10^{-18}$ e pertence ao par de candidatos números 1000,1004; e o 45º menor valor-p corrigido é aproximadamente $9{,}02 \times 10^{-5}$ e pertence ao par de candidatos números 1007,1009. Portanto todas as 45 combinações de pares de candidatos com respostas dependentes, de números 1001 a 1009, obtiveram valor-p corrigido < 0,001, o α que adotamos, de forma que rejeitamos a hipótese-nula de independência entre as respostas desses 10 candidatos, concluindo pela hipótese alternativa de respostas não-independentes, correspondente a algum tipo de fraude.

Já o 46º menor valor-p corrigido é aproximadamente 0,029 e pertence ao par de candidatos números 689,755, o mesmo par de candidatos *j,k* que obteve o valor-p corrigido mínimo quando simulamos somente os 1000 candidatos independentes. Assim como no caso anterior, por seu valor-p corrigido = 0,029 > 0,001, não rejeitamos a hipótese-nula de independência entre suas respostas.

Tabela 4.5 – Exemplo de aplicação em exame com 1009 candidatos, sendo 10 dependentes

Candidato j	Nota j	Candidato k	Nota k	mjk(4)	E(mjk) (5)	Var(mjk))$^{1/2}$	cjk (7)	valor-pc (34)
Candidato 1000	52	**Candidato 1004**	50	89	40,224581	5,111419	9,542442	**1,32676E-18**
Candidato 1000	52	**Candidato 1001**	51	88	40,224581	5,111419	9,346802	**8,41899E-18**
Candidato 1000	52	**Candidato 1003**	52	88	40,224581	5,111419	9,346802	**8,41899E-18**
Candidato 1000	52	**Candidato 1005**	49	88	40,224581	5,111419	9,346802	**8,41899E-18**
Candidato 1000	52	**Candidato 1006**	56	86	41,275252	5,087702	8,790756	**1,31001E-15**
Candidato 1000	52	**Candidato 1002**	49	85	40,224581	5,111419	8,75988	**1,70972E-15**
Candidato 1000	52	**Candidato 1009**	51	85	40,224581	5,111419	8,75988	**1,70972E-15**
Candidato 1000	52	**Candidato 1008**	55	84	41,275252	5,087702	8,397652	**3,84155E-14**
Candidato 1001	51	**Candidato 1004**	50	81	40,224581	5,111419	7,977319	**1,34293E-12**
Candidato 1000	52	**Candidato 1007**	52	80	40,224581	5,111419	7,781678	**6,15508E-12**
Candidato 1004	50	**Candidato 1005**	49	80	40,224581	5,111419	7,781678	**6,15508E-12**
Candidato 1003	52	**Candidato 1004**	50	79	40,224581	5,111419	7,586038	**2,73062E-11**
Candidato 1001	51	**Candidato 1003**	52	78	40,224581	5,111419	7,390398	**1,18066E-10**
Candidato 1002	49	**Candidato 1008**	55	78	41,275252	5,087702	7,218337	**4,15524E-10**
Candidato 1004	50	**Candidato 1008**	55	78	41,275252	5,087702	7,218337	**4,15524E-10**
Candidato 1001	51	**Candidato 1002**	49	77	40,224581	5,111419	7,194757	**4,89608E-10**
Candidato 1002	49	**Candidato 1003**	52	77	40,224581	5,111419	7,194757	**4,89608E-10**
Candidato 1003	52	**Candidato 1005**	49	77	40,224581	5,111419	7,194757	**4,89608E-10**
Candidato 1002	49	**Candidato 1006**	56	77	41,275252	5,087702	7,021785	**1,67698E-09**
Candidato 1003	52	**Candidato 1006**	56	77	41,275252	5,087702	7,021785	**1,67698E-09**
Candidato 1006	56	**Candidato 1009**	51	77	41,275252	5,087702	7,021785	**1,67698E-09**
Candidato 1001	51	**Candidato 1005**	49	76	40,224581	5,111419	6,999117	**1,95429E-09**
Candidato 1004	50	**Candidato 1009**	51	76	40,224581	5,111419	6,999117	**1,95429E-09**
Candidato 1001	51	**Candidato 1008**	55	76	41,275252	5,087702	6,825232	**6,51106E-09**
Candidato 1005	49	**Candidato 1006**	56	76	41,275252	5,087702	6,825232	**6,51106E-09**
Candidato 1001	51	**Candidato 1009**	51	75	40,224581	5,111419	6,803476	**7,5083E-09**
Candidato 1002	49	**Candidato 1004**	50	75	40,224581	5,111419	6,803476	**7,5083E-09**
Candidato 1002	49	**Candidato 1005**	49	75	40,224581	5,111419	6,803476	**7,5083E-09**
Candidato 1004	50	**Candidato 1006**	56	75	41,275252	5,087702	6,62868	**2,4327E-08**
Candidato 1008	55	**Candidato 1009**	51	75	41,275252	5,087702	6,62868	**2,4327E-08**
Candidato 1002	49	**Candidato 1009**	51	74	40,224581	5,111419	6,607836	**2,77666E-08**
Candidato 1003	52	**Candidato 1009**	51	74	40,224581	5,111419	6,607836	**2,77666E-08**
Candidato 1001	51	**Candidato 1006**	56	74	41,275252	5,087702	6,432128	**8,74558E-08**
Candidato 1003	52	**Candidato 1008**	55	74	41,275252	5,087702	6,432128	**8,74558E-08**
Candidato 1005	49	**Candidato 1008**	55	74	41,275252	5,087702	6,432128	**8,74558E-08**
Candidato 1001	51	**Candidato 1007**	52	73	40,224581	5,111419	6,412196	**9,88403E-08**
Candidato 1003	52	**Candidato 1007**	52	73	40,224581	5,111419	6,412196	**9,88403E-08**
Candidato 1004	50	**Candidato 1007**	52	73	40,224581	5,111419	6,412196	**9,88403E-08**
Candidato 1005	49	**Candidato 1009**	51	73	40,224581	5,111419	6,412196	**9,88403E-08**
Candidato 1006	56	**Candidato 1008**	55	73	43,159766	5,090382	5,862082	**2,87464E-06**
Candidato 1006	56	**Candidato 1007**	52	71	41,275252	5,087702	5,84247	**3,22495E-06**
Candidato 1002	49	**Candidato 1007**	52	70	40,224581	5,111419	5,825275	**3,54589E-06**
Candidato 1005	49	**Candidato 1007**	52	68	40,224581	5,111419	5,433994	**3,18704E-05**
Candidato 1007	52	**Candidato 1008**	55	68	41,275252	5,087702	5,252813	**8,40834E-05**
Candidato 1007	52	**Candidato 1009**	51	67	40,224581	5,111419	5,238353	**9,02318E-05**
Candidato 689	45	Candidato 755	57	21	41,275252	5,087702	3,985149	0,02881224
Candidato 379	60	Candidato 859	57	64	44,378914	5,077128	3,864604	0,046274663
Candidato 583	56	Candidato 999	51	60	41,275252	5,087702	3,680394	0,092308579
Candidato 224	53	Candidato 939	53	59	40,224581	5,111419	3,67323	0,094307252
Candidato 93	51	Candidato 95	53	59	40,224581	5,111419	3,67323	0,094307252

Neste caso exemplificado, o método foi capaz de destacar os 10 candidatos com respostas dependentes, de números 1001 a 1009, e também de identificar todas as

possíveis combinações entre eles, com valores-p corrigidos < 0,001, o que equivale dizer que não houve nenhum falso negativo (erro tipo II). Mais importante que isso, todas as combinações de pares de candidatos independentes, de números 1 a 999, obtiveram valor-p corrigido > 0,001, o que equivale dizer que não houve nenhum falso positivo (erro tipo I).

4.7 – Tamanho dos estratos de classificação dos candidatos pelo nível de habilidade

Como vimos na seção 3.6, $p_j^{iA} = P\left(x_j^i = A\right)$ é a probabilidade do candidato *j* responder alternativa *A* à questão *i* e será estimada por fr_j^{iA} que é a frequência relativa de respostas *A* à questão *i* no estrato de habilidade do candidato *j*.

Cabe agora uma consideração quanto a essa estratificação da população de candidatos por estratos de habilidade, que é necessária para se obter uma boa estimação de $p^i{}_{jk}$.(27). Quanto menor o número de candidatos em cada estrato, mais próximos eles estarão quanto ao seu nível de habilidade, mas pior será a aproximação $p_j^{iA} = P\left(x_j^i = A\right) \approx fr_j^{iA}$ (28), pois o tamanho da amostra para o cálculo da frequência relativa será menor. Por outro lado, quanto maior o número de candidatos em cada estrato de habilidade, melhor será essa aproximação da probabilidade de resposta pela frequência relativa, mas candidatos com níveis de habilidade não tão semelhantes pertencerão ao mesmo estrato, sendo suas probabilidades de respostas estimadas pelas mesmas frequências relativas.

Logo, temos que adotar um tamanho de estrato de habilidade, de forma que o nível de habilidade dos candidatos em cada estrato seja semelhante o suficiente, e a aproximação em (28) seja boa o bastante para que as frequências relativas daquele estrato representem com boa precisão as probabilidades de respostas dos candidatos daquele estrato.

Para encontrar o melhor tamanho para os estratos de habilidade, geramos um conjunto de dados conforme explicado na seção anterior, correspondente a um exame com os seguintes parâmetros: 100 questões, com 5 alternativas de respostas (letras “A” a “E”), realizado por 1009 candidatos, com 10 candidatos fraudadores, com 75% (em média) de respostas compartilhadas entre si (dependentes). Com esses parâmetros fixados, criamos os ficheiros “independentes” e o “dependentes”, após a inserção dos candidatos com cópias de respostas.

Aplicamos o método para esses dois ficheiros, seguindo os passos descritos nas seções 4.3 e 4.4, variando apenas o tamanho dos estratos de habilidades em que os candidatos são agrupados: 50, 100, 200 e “1 desvio padrão”.

Este último tamanho de estrato, que chamamos de "1 desvio padrão", é definido como: para cada candidato *j*, comparamos suas respostas a todas as questões de 1 a *I*, com a grelha de respostas oficial, compostas pelas alternativas mais frequentes em cada questão *i*, contando-se o número de coincidências, r_j (26) que corresponderá à "nota" do candidato *j*. Com todos os r_j podemos calcular a "nota" média dos candidatos e o desvio padrão das notas. Formamos então estratos para agrupar os candidatos cujo r_j está compreendido entre o valor da nota média e nota média mais um desvio-padrão, nota média mais dois desvios-padrão, sucessivamente até o candidato de maior nota; e também para nota média menos um desvio-padrão, nota média menos dois desvios-padrão, sucessivamente até o candidato de menor nota. Desta forma, agrupamos todos os candidatos pela sua "nota" r_j que corresponderia em um exame real, ao seu nível de habilidade. Quando um estrato contém menos que 100 candidatos, aglutinamos esse estrato com a anterior, de modo a garantir que nenhum estrato contenha menos que 100 candidatos, de forma que a aproximação $p_j^{iA} = P\left(x_j^i = A\right) \approx fr_j^{iA}$ seja boa o suficiente, pois o tamanho da amostra para o cálculo da frequência relativa será sempre maior ou igual a 100 candidatos.

Em cada conjunto de ficheiros "independentes" e "dependentes", para cada tamanho de estrato, identificamos o menor valor-p corrigido de candidato independente, correspondente ao caso sem fraude, e analisamos os valores-p dos candidatos "dependentes". Assim podemos comparar os resultados do método, quanto a sua capacidade de identificar e separar os candidatos com respostas dependentes, que correspondem aos fraudadores em um exame real com fraude, alterando-se apenas os tamanhos dos estratos de habilidade, mantendo todos os demais parâmetros constantes.

Em seguida, repetimos esse mesmo processo, apenas alterando-se a porcentagem de respostas copiadas, que era de 75% no primeiro caso, para 60% no segundo caso. Por fim, repetimos novamente o processo, agora para 50% de respostas copiadas.

Portanto, geramos 3 pares (respostas independentes e dependentes) de bancos de dados com vários parâmetros iguais (*I*=100 questões; *N*=1000 candidatos independentes e *N*=1009, sendo 999 candidatos independentes + 10 dependentes; com 5 alternativas de respostas, letras "A" a "E"); e apenas um parâmetro distinto entre eles (75% de respostas copiadas entre os 10 candidatos dependentes; 60% e 50%). Esses 3 pares de bancos de dados foram analisados de 4 formas distintas, seguindo o método descrito nas seções 4.3 e 4.4, variando-se o tamanho dos estratos de habilidade em que os candidatos eram agrupados:

estratos com 50 candidatos, 100 candidatos, 200 candidatos e "1 desvio padrão" (conforme explicado anteriormente). Os resultados são apresentados na tabela 4.6.

Tabela 4.6 – Simulações analisadas para determinar o melhor tamanho de estrato

Tamanho estrato	Valor-p mín. independentes	% de respostas copiadas	Valor-p mín. dependentes	Candidatos dependentes	Combinações dependentes
50	0,051	75%	6,4E-19	100%	100%
100	0,040	75%	2,6E-19	100%	100%
200	0,029	75%	1,0E-18	100%	100%
1 desvio padrão	0,029	75%	1,3E-18	100%	100%
50	0,094	60%	8,3E-15	100%	76%
100	0,076	60%	5,9E-15	100%	71%
200	0,085	60%	2,9E-15	100%	71%
1 desvio padrão	0,070	60%	2,1E-14	100%	78%
50	0,006	50%	6,4E-10	70%	13%
100	0,042	50%	4,1E-10	80%	18%
200	0,065	50%	4,4E-09	60%	11%
1 desvio padrão	0,028	50%	1,9E-10	80%	16%

Onde:

Tamanho estrato é o tamanho dos estratos em que os candidatos foram agrupados por suas notas r_j;

Valor-p mín. independentes corresponde ao menor valor-p obtido por um candidato que possui todas as respostas independentes;

% de respostas copiadas é a média de respostas copiadas entre os candidatos com respostas copiadas (dependentes) entre si;

Valor-p mín. dependentes corresponde ao menor valor-p obtido por um candidato que possui uma percentagem de respostas copiadas (dependentes) de outros;

Candidatos dependentes é a percentagem de todos os candidatos com respostas dependentes entre si, que obtiveram ao menos uma vez um valor-p $< \alpha = 0{,}001$; e

Combinações dependentes é a percentagem de todas as combinações possíveis *j,k* entre os candidatos com respostas dependentes entre si, que obtiveram valor-p $< \alpha = 0{,}001$.

Observamos que em todos os casos nenhum dos candidatos com todas as respostas independentes obteve valor-p $< \alpha = 0{,}001$, não ocorrendo, portanto, falsos positivos (erro de Tipo I).

Quando a percentagem de respostas copiadas = 75%, todos os 4 tamanhos de estratos são igualmente eficientes, pois identificam 100% dos candidatos com respostas dependentes e 100% das combinações *j,k* entre eles.

Quando a percentagem de respostas copiadas = 60%, todos os 4 tamanhos de estratos identificam 100% dos candidatos com respostas dependentes; mas diferem um pouco quanto à identificação das combinações *j,k* entre eles: o melhor caso é para estrato tamanho "1 desvio padrão", com 78% das combinações *j,k* identificadas; e os piores casos são os estratos com 100 e 200 candidatos, com 71% das combinações *j,k* identificadas.

Quando a percentagem de respostas copiadas = 50%, nem todos os candidatos com respostas dependentes são identificados, o que significa a ocorrência de falsos negativos (erro de Tipo II): o melhor caso é para estrato tamanho 100 candidatos e "1 desvio padrão", com 80% dos candidatos com respostas copiadas (dependentes) identificados; e o pior caso é o estrato com 200 candidatos, com 60% dos candidatos com respostas copiadas (dependentes) identificados. O tamanho de estrato com 50 candidatos produziu um par *j*,*k* com valor-p mín ind = 0,006, o que está relativamente próximo do nível de significância α = 0,001.

De um modo geral, para todos os casos, os tamanhos de estratos com 1 desvio padrão de "notas" e fixo com 100 candidatos, apresentaram resultados muito semelhantes, de modo que se pode optar por qualquer um deles. Ambos apresentam bons resultados com relação a erros tipo II (falsos negativos); mantendo-se a segurança de não se cometer erros tipo I (falsos positivos), que é nossa prioridade.

Concluímos que as melhores opções de tamanhos de estrato de candidatos por nível de habilidade, são a de 1 desvio padrão de "notas" e o fixo de 100 candidatos.

4.8 – Número de questões nos exames

Para analisar como o método se comporta em função do número de questões nos exames, simulamos e comparamos 5 conjuntos de dados, variando somente o número de questões: 25, 50, 75, 100 e 200 questões, mantendo-se todos os demais parâmetros fixos: 1000 candidatos, 5 alternativas de respostas, 10 candidatos "fraudadores", com 75% de respostas copiadas (dependentes), tamanho de estrato "1 desvio padrão", conforme mostrado na tabela 4.7.

Para 25 questões, o método não deve ser aplicado, uma vez que no banco de dados com respostas dependentes, o menor valor-p corrigido (34) dos candidatos com respostas copiadas (dependentes) foi 0,003 > α = 0,001, configurando falsos negativos (erros tipo II) para todos os pares de candidatos com respostas dependentes. Isso ocorre porque o método se baseia no Teorema do Limite Central para variáveis não identicamente distribuídas, para aproximar a distribuição real da soma das variáveis u^i_{jk} (1), que é uma binomial composta ou generalizada, pela distribuição Normal. Mas um dos pressupostos necessários para aplicação do teorema, é que o número de parcelas da soma, no caso *I* questões, seja suficientemente grande, o que parece não ocorrer para *I* = 25, tendo em vista

os resultados obtidos de valores-p corrigidos mínimos para os candidatos com respostas dependentes (copiadas).

Tabela 4.7 – Simulações para analisar a influência do número de questões

Nº questões (I)	Valor-p mín. independentes	% de respostas copiadas	Valor-p mín. dependentes	Candidatos dependentes	Combinações dependentes
25	0,032	75%	0,003	0%	0%
50	0,022	75%	3,5E-10	100%	62%
75	0,089	75%	4,3E-10	100%	29%
100	0,029	75%	1,3E-18	100%	100%
200	0,048	75%	1,4E-39	100%	100%

Onde:

Nº questões (I) é o número de questões do exame;

Valor-p mín. independentes corresponde ao menor valor-p obtido por um candidato que possui todas as respostas independentes;

% de respostas copiadas é a média de respostas copiadas entre os candidatos com respostas copiadas (dependentes) entre si;

Valor-p mín. dependentes corresponde ao menor valor-p obtido por um candidato que possui uma percentagem de respostas copiadas (dependentes) de outros;

Candidatos dependentes é a percentagem de todos os candidatos com respostas dependentes entre si, que obtiveram ao menos uma vez um valor-p $< \alpha = 0{,}001$; e

Combinações dependentes é a percentagem de todas as combinações possíveis *j*,*k* entre os candidatos com respostas dependentes entre si, que obtiveram valor-p $< \alpha = 0{,}001$.

Para 50 e 75 questões, o método foi capaz de identificar todos os candidatos com respostas dependentes, em pelo menos um par *j*,*k*; mas não identificou todas as possíveis combinações de pares *j*,*k* dependentes, incorrendo, portanto, em alguns falsos negativos. Não houve nenhum falso positivo (erro tipo I), como era desejável, e foi suficiente para indicar todos os 10 candidatos com respostas copiadas (dependentes).

Para 100 e 200 questões, o método foi capaz de identificar todos os candidatos com respostas copiadas (dependentes), em todas as possíveis combinações de pares *j*,*k* dependentes, sem ocorrerem falsos negativos (erro tipo II). Também não houve nenhum falso positivo (erro tipo I), como era desejável.

Quanto ao número de questões no exame, embora não tenham sido avaliados exames com número de questões maior que 25 e menor que 50, como para 50 questões já houve ocorrência de falsos negativos, concluiu-se que, para um número menor que 50, ocorreriam ainda mais casos de falsos negativos.

Concluímos, portanto, que pelo menos 50 questões são necessárias para a aplicação deste método, sendo desejável 100 questões ou mais, para haver maior eficácia em evitar a ocorrência de falsos positivos, bem como minimizar a ocorrência de falsos negativos (erros tipos I e II, respetivamente).

4.9 – Número de candidatos que realizaram o exame

Para analisar como o método se comporta em função do número de candidatos que realizaram o exame investigado, simulamos e comparamos 19 conjuntos de dados, variando o número de candidatos independentes *N*: 25, 50, 100, 1000, 10000 e 100000; o número de candidatos dependentes: 2, 5, 10 e 50; mantendo-se todos os demais parâmetros fixos: 100 questões, 5 alternativas de respostas, com 75% de respostas dependentes, tamanho de estrato "1 desvio padrão", conforme apresentado na tabela 4.8.

Tabela 4.8 – Simulações para analisar a influência do número de candidatos

Nº candidatos (N)	Nº candidatos "dependentes"	Valor-p mín independentes	Valor-p mín dependentes	Candidatos dependentes	Combinações dependentes
25	2	> 0,569	2,8E-15	100%	100%
25	5	> 0,569	6,2E-14	100%	100%
25	10	0,003	1,1E-15	100%	98%
50	2	0,018	5,0E-15	100%	100%
50	5	0,084	2,1E-18	100%	100%
50	10	0,036	2,4E-18	100%	100%
100	2	0,028	1,7E-13	100%	100%
100	5	0,048	2,8E-16	100%	100%
100	10	0,022	8,1E-17	100%	100%
1000	2	0,032	4,3E-15	100%	100%
1000	5	0,006	2,7E-17	100%	100%
1000	10	0,029	1,3E-18	100%	100%
1000	50	0,116	4,8E-17	100%	87%
10000	2	0,025	4,0E-10	100%	100%
10000	5	0,016	5,1E-19	100%	90%
10000	10	0,013	1,1E-19	100%	96%
10000	50	0,011	5,1E-18	100%	85%
100000	10	0,002	4,2E-17	100%	56%
100000	50	0,002	3,0E-18	100%	84%

Onde:

Nº candidatos (N) é o número de candidatos com respostas independentes que realizaram o exame;

Nº candidatos "dependentes" é o número de candidatos com uma percentagem de respostas copiadas (dependentes) de outro candidato;

Valor-p mín. independentes corresponde ao menor valor-p obtido por um candidato que possui todas as respostas independentes;

Valor-p mín. dependentes corresponde ao menor valor-p obtido por um candidato que possui uma percentagem de respostas copiadas (dependentes) de outros;

Candidatos dependentes é a percentagem de todos os candidatos com respostas dependentes entre si, que obtiveram ao menos uma vez um valor-p $< \alpha = 0{,}001$; e

Combinações dependentes é a percentagem de todas as combinações possíveis *j,k* entre os candidatos com respostas dependentes entre si, que obtiveram valor-p $< \alpha = 0{,}001$.

Em todas as combinações de parâmetros simuladas, não houve nenhum falso positivo (erro tipo I), como era desejável e além disso, o método foi capaz de identificar todos os candidatos com respostas dependentes, em pelo menos um par *j,k,* evitando que algum fraudador não fosse identificado.

Em aproximadamente 63% das combinações de parâmetros simuladas, o método foi capaz de identificar todas as possíveis combinações de pares de candidatos *j,k* com respostas dependentes, sem incorrer, portanto, em nenhum falso negativo (erro tipo II).

Em aproximadamente 37% das combinações de parâmetros simuladas, não identificou todas as possíveis combinações de pares *j,k* dependentes, incorrendo, portanto, em falsos negativos (erros tipo II). De qualquer forma, foi suficiente para indicar todos os candidatos com respostas dependentes em pelo menos um par *j,k.*, evitando que algum fraudador deixasse de ser identificado.

Quanto ao número de candidatos que realizaram o exame, concluiu-se que para $N \leq 100000$, o método funciona com a mesma eficácia, independentemente do tamanho de *N*. Para $N > 100000$, parece aumentar a probabilidade de se ocorrer falsos positivos (erros tipo I), pois os *Valor-p mín. independentes*, entre os candidatos com respostas independentes, estão mais próximos do nível de significância crítico $\alpha = 0{,}001$, alguns sendo da mesma ordem de grandeza. Quando temos 100000 candidatos, fazemos $5x10^9$ comparações entre eles. Neste caso, poder-se-ia aplicar o método com segurança, se diminuíssemos o nível de significância crítico α para os testes de hipóteses, a fim de continuarmos evitando a ocorrência de falsos positivos (erros tipo I).

4.10 – Número de candidatos que fraudaram o exame

Para analisar como o método se comporta em função do número de candidatos que fraudaram o exame, que chamamos de "dependentes", simulamos e comparamos 19 conjuntos de dados, variando o número de candidatos independentes: 50, 100, 1000, 10000 e 100000; o número de candidatos dependentes: 2, 5, 10 e 50; mantendo-se todos os demais parâmetros fixos: 100 questões, 5 alternativas de respostas (letras "A" a "E"), com 75% de respostas dependentes, conforme mostrado na Tabela 4.8.

Cumpre observar que não simulamos todas as combinações de candidatos independentes e dependentes possíveis, por acreditar que algumas não fariam sentido na aplicação prática. Por exemplo, para 50 ou 100 candidatos independentes que realizaram o

exame, simulamos com 2, 5 ou 10 candidatos que fraudaram, mas não com 50. Só simulamos 50 fraudadores, nos concursos com 1000, 10000 e 100000 candidatos independentes que realizaram o exame.

Em todas as combinações de parâmetros simuladas, não houve nenhum falso positivo (erro tipo I), como era desejável e o método foi capaz de identificar todos os candidatos com respostas dependentes, em pelo menos um par *j,k,* evitando que algum fraudador não fosse identificado.

Quanto a falsos negativos (erros tipo II), quanto maior o número de candidatos com respostas dependentes, mais casos de combinações de pares de candidatos *j,k* podem não ser identificadas, mas todos os candidatos dependentes foram identificados em pelo menos um par *j,k.*

Quanto ao número de candidatos que fraudaram o exame, concluiu-se que o método funciona com a mesma eficácia, independentemente da quantidade de candidatos fraudadores, ou "dependentes"; mas quanto maior o seu número, mais falsos negativos podem ocorrer.

4.11 – Quantidade percentual de respostas dependentes entre candidatos fraudadores

Poderíamos simplesmente copiar todas as respostas de um determinado candidato para os outros, gerando respostas 100% idênticas. Esse tipo de fenômeno só ocorre na realidade quando um determinado candidato copia integralmente a folha de respostas de outro candidato. Isso até acontece, mas é raro, não é o tipo de fraude mais comum.

O que as investigações da Polícia Federal do Brasil encontraram, são verdadeiras organizações criminosas que fraudam os resultados, mas que disfarçam essa fraude, repassando boa parte das respostas aos candidatos fraudadores, mas não a totalidade, de forma que as folhas de respostas contêm um grande número de respostas coincidentes, mas que estão longe de ser 100% idênticas.

Para analisar como o método se comporta em função da percentagem de todas as respostas do exame que foram copiadas ou compartilhadas de alguma forma, pelos candidatos fraudadores, que chamamos de "dependentes", simulamos e comparamos 3 conjuntos de dados, variando a percentagem de respostas dependentes: 50%, 60% e 75%, mantendo-se todos os demais parâmetros fixos: 100 questões, 1000 candidatos

independentes, 10 candidatos fraudadores e 5 alternativas de respostas (letras “A” a “E), para quatro tamanhos de estratos: 50, 100, 200 e “1 desvio padrão”.

Os resultados estão apresentados na Tabela 4.6. Em todas as combinações de parâmetros simuladas, não houve nenhum falso positivo (erro tipo I), como era desejável.

Para 75% de respostas dependentes, o método foi capaz de identificar todos os candidatos com respostas dependentes, em todas as possíveis combinações de pares *j*,*k* dependentes, portanto, sem incorrer em nenhum falso negativo (erro tipo II).

Para 60% de respostas dependentes, o método foi capaz de identificar todos os candidatos com respostas dependentes, em pelo menos um par *j*,*k*; mas não identificou todas as possíveis combinações de pares *j*,*k* dependentes, incorrendo, portanto, em alguns falsos negativos nas combinações. Mas foi suficiente para indicar todos os 10 candidatos dependentes.

Para 50% de respostas dependentes, o método foi capaz de identificar 80% dos candidatos com respostas dependentes, em pelo menos um par *j*,*k*; mas não identificou todas as possíveis combinações de pares *j*,*k* dependentes, incorrendo, portanto, em muitas combinações de falsos negativos. Conseguiu indicar apenas 8 dos 10 candidatos dependentes.

Quanto à percentagem de respostas dependentes entre os candidatos fraudadores, concluiu-se que, quando há pelo menos 75% em média de respostas dependentes, ou seja, os candidatos inseridos com respostas copiadas, tiveram em média 75% de respostas copiadas, o método funciona com toda a eficácia, identificando todos os candidatos fraudadores, bem como todas as combinações de candidatos onde há respostas dependentes. Já para 60% de respostas dependentes, o método ainda é bastante eficaz, pois identifica todos os candidatos fraudadores, em pelo menos uma combinação de par de candidatos com respostas dependentes. Para 50% de respostas dependentes, o método é parcialmente eficaz, identificando a maior parte dos fraudadores, mas não todos. Abaixo de 50% de respostas dependentes, a eficácia do método seria bastante variável e não se pode garantir que a maior parte dos candidatos fraudadores seriam identificados.

4.12 – Quantidade de alternativas de respostas em cada questão

Existem diversos tipos de exames com questões cujas respostas são alternativas do tipo múltipla-escolha, e neste trabalho avaliamos dois tipos dos mais comuns:

respostas tipo [A, B, C, D, E] e respostas tipo [CERTO, ERRADO]. O primeiro tipo parece-nos mais comumente aplicado no Brasil, e por isso a maior parte das situações que analisamos foi com esse tipo de respostas.

Para analisar como o método se comporta quando as alternativas de respostas são tipo [CERTO, ERRADO], simulamos e comparamos 16 conjuntos de dados, variando isoladamente cada um dos parâmetros: o número de questões: 25, 50, 75, 100 e 200; o número de candidatos: 50, 100, 1000, 10000 e 100000; e a percentagem de respostas dependentes: 50%, 60% e 75%, sempre com tamanho de estrato "1 desvio padrão".

Tabela 4.9 – Simulações para analisar a influência do número de alternativas de resposta

Nº questões (*I*)	Nº candidatos (*N*)	Nº candidatos dependentes	Valor-p mín ind	% de respostas copiadas	Valor-p mín dep	Candidatos dep	Combinações dep
25	1000	10	0,076	75%	0,007	0%	0%
50	1000	10	0,007	75%	3,3E-06	80%	22%
75	1000	10	0,028	75%	5,6E-12	100%	64%
100	25	10	> 0,569	75%	8,7E-12	100%	87%
100	50	10	0,008	75%	1,4E-13	100%	89%
100	100	10	0,035	75%	9,9E-15	100%	100%
100	1000	2	0,004	75%	1,9E-10	100%	100%
100	1000	5	0,003	75%	2,0E-10	100%	90%
100	1000	10	0,022	50%	1,3E-09	90%	20%
100	1000	10	0,007	60%	1,1E-07	100%	20%
100	1000	10	0,008	75%	7,5E-14	100%	96%
100	1000	50	0,041	75%	1,1E-15	100%	83%
100	10000	10	0,002	75%	3,2E-12	100%	64%
100	100000	10	0,002	75%	7,7E-12	100%	56%
200	1000	10	0,020	50%	1,6E-13	100%	29%
200	1000	10	0,011	75%	1,1E-27	100%	100%

Onde:

Nº questões (I) é o número de questões do exame;

Nº candidatos (N) é o número de candidatos com respostas independentes que realizaram o exame;

Nº candidatos "dependentes" é o número de candidatos com uma percentagem de respostas copiadas (dependentes) de outro candidato;

Valor-p mín ind corresponde ao menor valor-p obtido por um candidato que possui todas as respostas independentes;

% de respostas copiadas é a média de respostas copiadas entre os candidatos com respostas copiadas (dependentes) entre si;

Valor-p mín dep corresponde ao menor valor-p obtido por um candidato que possui uma percentagem de respostas copiadas (dependentes) de outros;

Candidatos dep é a percentagem de todos os candidatos com respostas dependentes entre si, que obtiveram ao menos uma vez um valor-p $< \alpha = 0{,}001$; e

Combinações dep é a percentagem de todas as combinações possíveis *j,k* entre os candidatos com respostas dependentes entre si, que obtiveram valor-p $< \alpha = 0{,}001$.

O comportamento do método nas diferentes combinações de parâmetros com duas alternativas de resposta [CERTO, ERRADO] não diferiu significativamente daquele observado nas combinações de parâmetros com cinco alternativas de resposta [A, B, C, D, E], descrito anteriormente nas seções 4.8 a 4.11, como pode ser observado comparando-se os valores da Tabela 4.9, com as Tabelas 4.7 e 4.8.

Houve em algumas situações, um pequeno aumento de ocorrências de falsos negativos (erros tipo II), deixando-se de identificar algumas das combinações de pares *j*,*k* com respostas dependentes. Todavia, mantém a eficácia de identificar todos os candidatos com respostas dependentes, em pelo menos um par *j*,*k*, de forma que nenhum fraudador deixaria de ser identificado. Também persiste a não-ocorrência de falsos positivos (erros tipo I), o que é ainda mais fundamental.

4.13 – Síntese dos resultados das análises e limites de aplicação do método

Nas seções anteriores deste capítulo, simulamos várias combinações de parâmetros diferentes, verificando-se a eficácia do método em destacar os candidatos com respostas dependentes, dos com respostas independentes. Isso permitiu determinar quais os limites de aplicação do método, de modo eficaz, quanto aos parâmetros intrínsecos dos exames, conforme sintetizado na Tabela 4.10.

Tabela 4.10 – Parâmetros mínimos e ideais para aplicação do método com eficácia

Parâmetro	Valor mínimo	Valor ideal
Número de questões do exame	50	≥100
Número de candidatos que realizaram o exame	25	≥50
Número de candidatos que fraudaram o exame	indiferente	
Percentagem de respostas dependentes nas fraudes	50%	≥75%

5 – CONCLUSÃO

O objetivo deste trabalho é apresentar um método eficaz para analisar exames de larga aplicação, como são muitos dos concursos para provimento de cargos públicos ou exames para admissão em universidades, e identificar se há ou não, indícios de ocorrência de fraudes nas respostas dos candidatos. Em caso positivo, deve sinalizar os candidatos com probabilidade mais elevada de terem respostas não-independentes, que corresponderão a algum tipo de fraude ou cópia, uma vez que as questões devem ser respondidas individualmente.

Insta comentar que o método utiliza a estatística para destacar quais candidatos apresentam respostas que provavelmente não foram obtidas de forma independente, mas não pode, nem pretende, determinar de que maneira se operou o compartilhamento de respostas, de forma que tal questão só pode ser elucidada através de outros esforços de investigação.

Existem vários métodos de análise estatística de similaridade de respostas em exames com testes de múltipla escolha que visam identificar a ocorrência de fraudes. Mas o método apresentado nesta dissertação possui uma diferença importante, ao invés de tentar estimar a probabilidade de coincidência de respostas entre os candidatos utilizando um modelo matemático aproximado; este método estima essa probabilidade utilizando dados reais extraídos das respostas dadas pelos candidatos, através das frequências relativas de respostas. Isso faz com que este método possa ser igualmente eficaz para exames com questões mais fáceis ou mais difíceis, realizados por uma população de candidatos melhor ou pior preparados.

Este método é robusto, pois se apoia em somente dois pressupostos. O primeiro é o de independência local entre as respostas, o que é de se esperar em um exame sem fraude, no qual os candidatos respondem as questões individualmente, sem qualquer forma de compartilhamento de respostas. O segundo e último pressuposto é o de aproximar a distribuição da soma de respostas coincidentes entre os candidatos, pela distribuição Normal, valendo-se do Teorema do Limite Central para o caso da soma de variáveis não identicamente distribuídas. Vimos que esse pressuposto é válido, desde que o número de questões do exame seja suficientemente grande. As análises dos casos simulados mostraram que para o número de questões $I \geq 50$, o método é eficaz.

A adoção desses dois únicos pressupostos, e de nenhuma suposição quanto a funções que buscariam prever como as respostas dos candidatos se distribuem, tornam este método bastante robusto para ser aplicado com eficácia em diferentes situações de exames, bem como o diferencia de muitos outros métodos existentes, que tentam, de diferentes maneiras, modelar a distribuição das respostas dos candidatos.

Utilizamos simulação de números pseudoaleatórios para criar diversos exames fictícios com várias combinações de parâmetros diferentes, nos quais aplicamos o método e analisamos a sua eficácia em destacar os candidatos com respostas dependentes, dos com respostas totalmente independentes. Isso permitiu determinar quais os limites mínimos e os ideais para aplicação do método, de modo eficaz.

A análise dos dados simulados permitiu determinar algumas limitações de aplicabilidade, além do número de questões I que deve ser ≥ 50: o método apresenta bons resultados para o número de candidatos $N \geq 25$ e para uma percentagem de respostas copiadas $\geq 50\%$. Abaixo desses limites, o método não deve ser utilizado, pois produzirá resultados incertos, não-fiáveis. Quanto ao número de candidatos com respostas copiadas, e quanto ao número de alternativas de respostas, o método se mostrou estável, não alterando sua eficácia com a variação desses parâmetros.

Por suas características intrínsecas teóricas, como as correções de continuidade e de Bonferroni; e de calibração, mormente na definição do nível de significância α dos testes de hipóteses, o método é extremamente conservador, de forma que prioriza a não-ocorrência de erros tipo I, ou falso positivos, em detrimento de, eventualmente, tolerar a ocorrência de erros tipo II, ou falso negativos. Estes podem ocorrer principalmente quando os parâmetros do exame são próximos dos limites mínimos de aplicação, principalmente quanto ao número de questões ou de percentagem de respostas copiadas.

Todavia, essas mesmas características intrínsecas, nos permitem afirmar que nenhum candidato inocente seja erroneamente acusado de fraude, e que todos os candidatos sinalizados pelo método como sendo candidatos com respostas não-independentes, de fato apresentam evidências estatisticamente significativas de respostas distribuídas de forma não-independente, portanto, violando a regra dos exames que determina que os mesmos devem ser respondidos individualmente por cada candidato.

BIBLIOGRAFIA

ANGOFF, W. The development of statistical indices for detecting cheaters. Educational Testing Service, 1972.

ARGENAL, R.; CO, F.; CRUZ, E.; PATUNGAN, W. A new index for detecting collusion and its statistical properties. National Convention on Statistics, 2004.

BAY, L. Detection of cheating on multiple-choice examinations. American Educational Research Association, ED 421530, TM 028856, 1995.

CIZEK, G. An overview of issues concerning cheating on large-scale tests. National Association of Test Directors 2001 Symposia, 2001.

CIZEK, G.; WOLLACK, J. Handbook of quantitative methods for detecting cheating on tests. Routledge, 2017.

CODY, R. Statistical analysis of examinations to detect cheating. Journal of Medical Education, Vol. 60, 136-137, 1985.

FELLER, W. An introduction to probability theory and its applications. Volume II. John Wiley & Sons, Second Edition, 1970.

FRARY, R.; TIDEMAN, T.; WATTS, T. Indices of cheating on multiple-choice tests. Journal of Educational Statistics, Vol. 2, Number 4, 235-256, 1977.

HANSON, B.; HARRIS, D.; BRENNAN, R. A comparison of several statistical methods for examining allegations of copying. ACT Research Report Series 87-15, 1987.

HOLLAND, P. Assessing unusual agreement between the incorrect answers of two examinees using the K-index: statistical theory and empirical support. Program Statistics Report, Technical Report No. 96-4, Educational Testing Service, 1996.

LEHMANN, E. Elements of large sample theory. Springer, 1999.

LEWIS, C.; THAYER, D. The power of the K-index (or PMIR) to detect copying. Research Report No. 98-49, Educational Testing Service, 1998.

McMANUS, I.; LISSAUER, T.; WILLIAMS, S. Detecting cheating in written medical examinations by statistical analysis of similarity of answers: pilot study. BMJ, volume 330, 1064-1066 + extended version 1-18, 2005.

ROSS, S. Introduction to probability and statistics for engineers and scientists. Elsevier Academic Press, 3rd ed., 2004.

SOTARIDONA, L. Statistical methods for the detection of answer copying on achievement tests. Twente University Press, Netherlands, 2003.

_____ Cheating detection using the S2 copying index. The Philippine Statistician, Vol. 52, Nos. 1-4, 59-67, 2003.

SOTARIDONA, L.; MEIJER, R. Statistical properties of the K-index for detecting answer copying. Faculty of Educational Science and Technology, University of Twente, Report Research 01-06, 2001.

_____ Two new statistics to detect answer copying. Journal of Educational Measurement, Vol. 40, 53-69, 2003.

VAN DER LINDEN, W.; SOTARIDONA, L. Detecting answer copying when the regular response process follows a known response model. Journal of Educational and Behavioral Statistics, Vol. 31, No. 3, 283-304, 2006.

WESOLOWSKY, G. Detecting excessive similarity in answers on multiple choice exams. Journal of Applied Statistics, Vol. 27, No. 7, 909-921, 2000.

WOLLACK, J. Detecting answer copying on high-stakes tests. The Bar Examiner, Vol. 73, Number 2, 35-45, 2004.

ANEXO I

Algoritmo em linguagem “R” utilizado na geração dos bancos de dados fictícios

Autor: Eduardo Augusto Comenda Cotrim

```
# Criando os arquivos utilizados
files <- c("r.csv","s.csv","gabarito.csv","respostas_ind.csv","respostas_dep.csv")
for (f in files) {
  file.create(f)
}
# Definindo os parâmetros do exame a ser simulado
Q <- 100 # Q é o número de questões
NI <- 1000 # NI é o número de candidatos independentes
ND <- 10 # ND é o número de candidatos dependentes
ic <- 0.75 # ic é o índice de coincidências entre os candidatos dependentes
# Gerando as respostas independentes dos candidatos
g <- 1:Q
for (i in 1:Q) {
  pa <- runif(1,0,0.9)
  pb <- runif(1,0,(0.9-pa))
  pc <- runif(1,0,(0.9-pa-pb))
  pd <- runif(1,0,(0.9-pa-pb-pc))
  pe <- (1-pa-pb-pc-pd)
  p <- c(pa, pb, pc, pd, pe)
  r <- sample(letters[1:5], NI, replace = T, prob = p)
  g[i] <- letters[which.max(table(r))]
  m <- matrix (r, nrow = 1, ncol = NI)
  write.table (m, file = "r.csv", append = T, quote = F, sep = ",", row.names = F, col.names = F)
# Gerando as respostas dependentes entre 10 candidatos
  ifelse (runif(1,0,1) <= ic, r2 <- r[NI], r2 <- sample(letters[1:5], 1, replace = T, prob = p))
  ifelse (runif(1,0,1) <= ic, r3 <- r[NI], r3 <- sample(letters[1:5], 1, replace = T, prob = p))
  ifelse (runif(1,0,1) <= ic, r4 <- r[NI], r4 <- sample(letters[1:5], 1, replace = T, prob = p))
  ifelse (runif(1,0,1) <= ic, r5 <- r[NI], r5 <- sample(letters[1:5], 1, replace = T, prob = p))
  ifelse (runif(1,0,1) <= ic, r6 <- r[NI], r6 <- sample(letters[1:5], 1, replace = T, prob = p))
  ifelse (runif(1,0,1) <= ic, r7 <- r[NI], r7 <- sample(letters[1:5], 1, replace = T, prob = p))
  ifelse (runif(1,0,1) <= ic, r8 <- r[NI], r8 <- sample(letters[1:5], 1, replace = T, prob = p))
  ifelse (runif(1,0,1) <= ic, r9 <- r[NI], r9 <- sample(letters[1:5], 1, replace = T, prob = p))
  ifelse (runif(1,0,1) <= ic, r10 <- r[NI], r10 <- sample(letters[1:5], 1, replace = T, prob = p))
  s <- cbind(r2,r3,r4,r5,r6,r7,r8,r9,r10)
  write.table (s, file = "s.csv", append = T, quote = F, sep = ",", row.names = F, col.names = F)
}
m1 <- read.csv("r.csv", header = F, sep = ",")
t1 <- t(m1)
nomes <- paste("CANDIDATO ", 1:NI, ";", sep="")
ind <- cbind(nomes,t1)
write.table (ind, "respostas_ind.csv", col.names = F, quote = F, row.names = F, sep = "")
m2 <- read.csv("s.csv", header = F, sep = ",")
t2 <- t(m2)
nomes <- paste("CANDIDATO ", (NI+1):(NI+ND-1), ";", sep="")
dep <- cbind(nomes,t2)
t3 <- rbind(ind,dep)
write.table (t3, "respostas_dep.csv", col.names = F, quote = F, row.names = F, sep = "")
# Gerando a grelha de respostas oficial
n <- matrix (g, 1, Q, byrow = T)
write.table (n, file = "gabarito.csv", append = T, quote = F, sep = "", row.names = F, col.names = F)
```

ANEXO II

Rotina computacional de comparação das respostas e cálculos preliminares

Autor do algoritmo lógico sequencial: Eduardo Augusto Comenda Cotrim

Autor da rotina implementada em GNU Awk: Robson Alexandre de Araújo Santos

```
#!/usr/bin/awk -f
#/**
# * @link      https://github.com/robsonalexandre
# * @license   https://www.gnu.org/licenses/gpl-3.0.txt GNU GENERAL PUBLIC LICENSE
# * @author    Robson Alexandre <alexandrerobson@gmail.com>
# */

# GNU Awk 4.1.4, API: 1.1 (GNU MPFR 4.0.1, GNU MP 6.1.2)
# Copyright (C) 1989, 1991-2016 Free Software Foundation.
# export LC_NUMERIC=pt_BR.UTF-8
# ENVIROMENTS VARS
#   concurso
#   gabarito
#   metodo
#   str_Rvalidas
#   str_Rinvalidas
#   indice_minimo
# ASSIGNED VARS -v var=val
#   faixafixa
#   decrementar_nota
#   probabilidade_condicional
# Usage: awk -N -F ';' -f analise_combinatoria.awk planilha_respostas.csv

# Média aritmética do vetor a
function avg(a, k, s) { N=length(a); for (k in a) {s+=a[k]}; return s / N; }
# Desvio Padrão Populacional do vetor a com média avg
function desvpad(a, avg, N) { s=0; N=length(a); for (n=1;n<=N;n++) {s+=(a[n] - avg)**2 / N;}; return sqrt(s); }
# Retorna o inteiro maior que x
function menor_inteiro(x) { return (x>0) ? int(x) : int(x-1); }
# Módulo de x
function abs(x) { return ((x < 0.0) ? -x : x); }
# Filtra string return somente alphanum + punct + blank
function str_filter(str) { re="[[:cntrl:]]"; gsub(re, "", str); return str; }

BEGIN {
  # ENV_VARS
  # @required concurso|metodo|gabarito|str_Rvalidas|str_Rinvalidas|indice_minimo
  metodo=ENVIRON["metodo"]
  indice_minimo=ENVIRON["indice_minimo"]
  str_Rvalidas=toupper(ENVIRON["str_Rvalidas"])
  str_Rinvalidas=toupper(ENVIRON["str_Rinvalidas"])
  re_Rvalidas="["str_Rvalidas"]"
  re_Rinvalidas="["str_Rinvalidas"]"
  re="[^" str_Rvalidas str_Rinvalidas "]"
  gabarito=toupper(ENVIRON["gabarito"])
  # Filtro de caracteres esperados
  gsub(re, "", gabarito)
  # Q(Número de questões)
  Q=split(gabarito, g, "")
}
{
  candidato[NR]=$1
  resposta=toupper($2)
```

```
 54    # Filtro de caracteres esperados
 55    gsub(re, "", resposta)
 56    # R(Número de respostas)
 57    R=split(resposta, r, "")
 58
 59    # Cálculo das Notas dos Candidatos
 60    nota=0
 61    for (q=1;q<=Q;q++) {
 62      if (match(g[q], re_Rinvalidas))
 63        # Questão anulada em gabarito oficial
 64        nota++
 65      else if (r[q] == g[q])
 66        # Resposta coincide com gabarito
 67        nota++
 68      else if (decrementar_nota == 1 && ENVIRON["concurso"] ==
    "cespe" && match(r[q], /[CE]/))
 69        # Concurso tipo => cespe, resposta válida não coincidente
    com C ou E do gabarito
 70        nota--
 71    }
 72    notas[NR]=nota
 73    respostas[NR]=resposta
 74  } END {
 75    # N(Número de candidatos)
 76    N=NR
 77
 78    # Média aritmética das notas
 79    media=avg(notas)
 80
 81    # Desvio Padrão Populacional das Notas
 82    dp=desvpad(notas, media)
 83
 84    # Organizando candidatos por faixas
 85    switch (metodo) {
 86      # Faixa Única
 87      case "faixaunica":
 88        f=1
 89        sFaixa[f]=N
 90        for (n=1;n<=N;n++) {
 91          nFaixa[n]=f
 92        }
 93        break
 94
 95      # Tamanho de Faixa Fixa
 96      case "faixafixa":
 97        PROCINFO["sorted_in"]="@val_num_desc"
 98        f=1
 99        nota=0
100        minporFaixa=faixafixa
101        for (n in notas) {
102          if (nota > notas[n] && count >= minporFaixa) {
103            f++
104            count=0
105          }
106          nFaixa[n]=f
107          sFaixa[f]++
108          count++
109          if (count >= minporFaixa) {
110            nota=notas[n]
```

```
111         }
112       }
113       break
114
115     # Tamanho de Faixa Desvio Padrão
116     case "faixadesvpad":
117       for (n=1;n<=N;n++) {
118         # Nota máxima = Q, onde, Q < media + f * desvpad
119         # Faixa f da nota máxima Q é:
120         # f=menor_inteiro((Q - media) / desvpad)
121         f=menor_inteiro((notas[n] - media) / dp)
122
123         sFaixa[f]++     # Quantidade de candidatos em faixa f
124         nFaixa[n]=f     # Candidato n na Faixa f
125       }
126
127       # Ordenar por índice númerico na descendente
128       PROCINFO["sorted_in"]="@ind_num_desc"
129
130       # Se Número de candidatos N >= 200
131       # Reagrupando candidatos se faixa contém menos de 100
132       if (N >= 200) {
133         do {
134           reagrupar=0
135           for (f in sFaixa) {
136             if (sFaixa[f] < 100) {
137               novafaixa=(f>0) ? f-1 : f+1
138               sFaixa[novafaixa]+=sFaixa[f]
139               delete sFaixa[f]
140               for (n=1;n<=N;n++) {
141                 if (nFaixa[n]==f) {
142                   nFaixa[n]=novafaixa
143                 }
144               }
145               reagrupar=1
146             }
147           }
148         } while (reagrupar != 0)
149       } else {
150       # Se o número de candidatos for menor que 200
151       # todos ficarão em uma única faixa
152         f=1
153         sFaixa[f]=N
154         for (n=1;n<=N;n++) {
155           nFaixa[n]=f
156         }
157       }
158       break
159   }
160
161   # Cálculo das frequências relativas
162   for (n=1;n<=N;n++) {
163     split(respostas[n], r, "")
164     f=nFaixa[n]
165     for (q=1;q<=Q;q++) {
166       if (match(r[q], re_Rvalidas)) {
167         # Somatório das repostas válidas por faixa, questão de
    alternativas
168         fr[f,q,r[q]]++
```

```
169       } else {
170         # Somatório das repostas inválidas (respostas em branco ou
    anuladas)
171         sRinvalidas[f,q]--
172       }
173     }
174   }
175   split(str_Rvalidas, r, "")
176   PROCINFO["sorted_in"]="@ind_num_asc"
177   for (f in sFaixa) {
178     for (q=1;q<=Q;q++) {
179       for (alternativa in r) {
180         fr[f,q,r[alternativa]]/=(sFaixa[f] + sRinvalidas[f,q])
181       }
182     }
183   }
184
185   # Cálculo das coincidências efetivas
186   split(str_Rvalidas, r, "")
187   for (j=1;j<=N;j++) {
188     split(respostas[j], r1, "")
189
190     # Coincidências Efetivas Cef[j,k]
191     # Métodos de Probabilidade de Coincidências (pijk)
192     # |-- Probabilidade Condicional de Coincidências (flag:
    probabilidade_condicional)
193     # |-- Probabilidade Total de Coincidências (default)
194     kinicial=(probabilidade_condicional) ? 1 : j
195     for (k=kinicial;k<=N;k++) {
196       if (j != k) {
197         split(respostas[k], r2, "")
198         Cef[j,k]=0
199         for (q=1;q<=Q;q++) {
200           if (r1[q] == r2[q]) {
201             Cef[j,k]++
202           }
203         }
204       }
205     }
206
207   # Cálculo da probabilidade pijk (Probabilidade das respostas dos
    candidatos j e k na questão i serem iguais)
208
209     # Coincidências Esperadas Cesp[j,k]
210     # Cesp[j,k]=Somatório(frj[f,q,r]*frk[f,q,r])
211     #
212     # Método de cálculo da probabilidade de coincidências
    condicional, variância
213     # Este método só é utilizado quando a flag
    coincidencias_esperadas_condicional está habilitada
214     # Cesp+=fr[faixas[i],q,resposta2[q]]
215     # variancia+=fr[faixas[i],q,resposta2[q]]*(1-
    fr[faixas[i],q,resposta2[q]])
216     #
217     # Método de cálculo da probabilidade de coincidências
    esperadas total, variância (default)
218     # Este método é o método de cálculo padrão utilizado,
    probabilidade de coincidências total
219     # fr_i_j=fr[nFaixa[j],q,r[alternativa]]
```

```
220     # fr_i_k=fr[nFaixa[k],q,r[alternativa]]
221     # Cesp+=fr_i_j * fr_i_k
222     # variancia+=(fr_i_j * fr_i_k) * (1 - (fr_i_j * fr_i_k))
223     kinicial=(probabilidade_condicional) ? 1 : j
224     for (k=kinicial;k<=N;k++) {
225       if (j != k) {
226         split(respostas[k], r2, "")
227         Cesp=0                    # Coincidências Esperadas
228         variancia=0
229         for (q=1;q<=Q;q++) {
230           if (probabilidade_condicional) {
231             Cesp+=fr[nFaixa[j],q,r2[q]]
232             variancia+=fr[nFaixa[j],q,r2[q]]*(1-
    fr[nFaixa[j],q,r2[q]])
233           } else {
234             for (alternativa in r) {
235               fr_i_j=fr[nFaixa[j],q,r[alternativa]]
236               fr_i_k=fr[nFaixa[k],q,r[alternativa]]
237               Cesp+=fr_i_j * fr_i_k
238               variancia+=(fr_i_j * fr_i_k) * (1 - (fr_i_j *
    fr_i_k))
239             }
240           }
241         }
242         # Variância = sqrt(Somatório(pijk * (1-pijk)))
243         # DP[j,k] = sqrt(variancia)
244         dp=sqrt(variancia)
245
246         # Índice c j,k c[j,k]=indice
247         indiceC=abs((Cef[j,k]-Cesp))/dp
248
249         # Resultado
250         # CANDIDATO j;NOTA j;CANDIDATO k;NOTA
    k;CEf[j,k];CEsp[j,k],DP(j,k);pijk(c)
251         if (indiceC >= indice_minimo)
252           printf "%s;%d;%s;%d;%d;%f;%f;%f\n", candidato[j],
    notas[j], candidato[k], notas[k], Cef[j,k], Cesp, dp, indiceC
253       }
254     }
255     delete Cef
256   }
257 }
```