Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Camila Neves Souto

Uso de Regressão Isotônica Suavizada na Detecção de Funcionamento Diferencial do Item

Brasília
2019

Camila Neves Souto

Uso de Regressão Isotônica Suavizada na Detecção de Funcionamento Diferencial do Item

Dissertação submetida ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília como requisito para a obtenção do título de Mestre em Estatística.

Área de Concentração: Probabilidade e Estatística

Orientador: Prof. Dr. Antônio Eduardo Gomes

Coorientador: Prof. Dr. Dalton Francisco de Andrade

Brasília

2019

aos meus pais, Carlos e Olga, e minhas irmãs, Tamara e Maria Fernanda, grandes causadores de DIF em minha vida

Agradecimentos

A Deus, que me ampara nos momentos mais difíceis, que torna tudo possível. Agradeço pelo dom da vida e pelas possibilidades que tive ao longo dela.

À minha família amada. Aos meus queridos pais, Carlos e Olga, que sempre batalharam para que tivéssemos a oportunidade de estudar e sempre nos incentivaram a buscar o melhor. À minha irmã Tamara, minha maior e melhor conselheira. À minha irmã Maria Fernanda, que com sua doçura esteve sempre para me apoiar.

À minha amiga Laís, que esteve comigo desde o primeiro dia de aula da graduação e trilhamos juntas o caminho até chegar ao mestrado. Sem nossa parceria, essa jornada teria sido outra, juntas fomos mais longe. À minha amiga Marília, que sempre incentiva e apoia desde as minhas mais tradicionais decisões aos meus mais inesperados caminhos escolhidos. Ao meu compadre Edgard, que sempre me encorajou na realização desse trabalho.

Aos colegas do Inep, ao amigo Lucas, que sempre me ajudou nos desafios de programar em R, sempre com muita paciência e inteligência. À amiga Thaysa, minha fiel companheira nos forrós, essenciais para refrescar a mente e renovar as energias tão necessárias nesse processo. À amiga Isabella, que com seu alto astral, é capaz de melhorar o humor de qualquer um. À equipe com quem tenho o prazer de dividir o trabalho, aos colegas, Carlos, Christyne, Daniel, Juliana, Margarete e Rachel. Ao coordenador Fábio Bravin e ao diretor Carlos Moreno, os quais colaboraram com a redução da carga de trabalho.

Aos colegas de mestrado, com quem tive o prazer de partilhar as angústias e alegrias dessa fase, em especial, Alessandra, Helen e Rayany.

À Universidade Federal de Brasília, aos colegas da secretaria de curso, em especial Karen e André. Aos docentes do programa de mestrado da UnB, que com seu conhecimento e dedicação nos enriquecem, em especial à professora Cira, aos professores Raul, Bernardo

e Nakano. À BCE, sempre de portas abertas para nos acolher e nos trazer a paz num ambiente agradável e inspirador.

E um agradecimento especial ao meu orientador Prof. Doutor Antônio Eduardo pelas conversas e trocas de experiências semanais, pela dedicação em me auxiliar nesse projeto e pelas ajudas incontáveis. Ao Prof. Dr. Dalton pelo auxílio na coorientação deste trabalho.

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

"*You can choose courage or you can choose comfort. You cannot have both.*"

Brené Brown

"*Só o sofrimento constrói.*"

Autor desconhecido

Resumo

A presente dissertação tem como objetivo enriquecer o estudo de Funcionamento Diferencial do Item (DIF, do inglês, *Differencial Item Functioning*). Para isso, foi feita revisão bibliográfica de alguns dos métodos já existentes para detecção de itens que apresentam DIF e se propôs um método não paramétrico, que estima as curvas característica dos itens dos dois grupos comparados via regressão isotônica. O método proposto foi comparado com outros métodos já existentes para detecção de itens com DIF, Mantel-Haenzel e Regressão Logística. O estudo foi feito a partir de simulações com 100 replicações para cada configuração do teste. As configurações variam de acordo com a dimensão do DIF, o tipo de DIF — uniforme ou não-uniforme — o tamanho da amostra e a proporção de respondentes em cada grupo. A comparação consistiu em observar, com os resultados das simulações, a proporção de detecção de DIF em itens que de fato apresentavam DIF e a proporção de falso DIFs apontados pelos métodos. Além disso, foi feita uma análise do impacto desses itens que apresentam DIF em um teste na estimação da proficiência dos respondentes.

Palavras-chave: Funcionamento Diferencial do Item (DIF), Teoria de Resposta ao Item (TRI), Regressão Isotônica, Mantel-Haenzel, Regressão Logística.

Abstract

This project aims to enrich the Differential Item Functioning (DIF) study. In this way, a bibliographical review of some of the existing methods for detection of items with DIF was done and a non-parametric method, which estimates the item characteristic curves of the two compared groups through isotonic regression, was proposed. The presented method was compared with other methods already developed to detect the DIF items, Mantel-Haenzel and logistic regression. I made a simulation study with 100 replications for each test configuration. The settings vary according to the size of DIF, the type of DIF — uniform or non-uniform — the sample size and the proportion of examinees in each group. Considering the results of the simulations, the comparison consisted in observing the proportion of DIF detection in items that in fact presented DIF and the proportion of false DIFs indicated by the methods. In addition, an analysis was made of the impact of these items that presented DIF in a test in the estimation of the proficiency of the examinees.

Key words: Differential Item Functioning (DIF), Item Response Theory (IRT), Isotonic Regression, Mantel-Haenzel, Logistic Regression.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A Teoria de Resposta ao Item (TRI), introduzida por Lord (1952), foi mundialmente difundida somente nos últimos 30 anos. A demora em se tornar popular, deve-se, principalmente, ao fato de a TRI exigir uma capacidade computacional consideravelmente maior que a Teoria Clássica na construção do teste e na estimação das proficiências. Além disso, quando do seu desenvolvimento, não havia softwares eficientes e computadores acessíveis para se fazer análises pela TRI. Consequentemente, com o desenvolvimento computacional, veio a aplicação em massa da TRI. No Brasil, modelos de TRI são aplicados em grandes avaliações como no Exame Nacional do Ensino Médio (ENEM) e no Sistema de Avaliação da Educação Básica (SAEB).

A TRI é uma teoria estatística que relaciona a habilidade do candidato e a resposta dele aos itens do teste. Esta teoria apresenta diferenças consideráveis — e, de certa forma, vantagens também — em relação a Teoria Clássica do Teste (TCT) que se baseia no escore bruto do respondente no teste. Testes baseados na TRI apresentam parâmetros invariantes dentre as amostras de respondentes dentro da população de interesse; o parâmetro da habilidade é invariante dentre as amostras dos itens do teste dentro da população de itens que medem a habilidade de interesse; tem-se uma estimativa do erro para cada habilidade estimada; tem-se a probabilidade de sucesso para cada item por respondente para todas proficiências; e tanto o parâmetro de dificuldade do item quanto a habilidade dos respondentes são medidas na mesma escala. Assim sendo, a TRI facilita a solução de possíveis erros na construção do teste, a comparação de proficiências entre respondentes, a estimação da proficiência dos indivíduos, a estimação do erro e, ainda, facilita análises do Funcionamento Diferencial do Item (do inglês, *Differencial Item Functioning* - DIF), que é a hipótese de um item, no que diz respeito a suas características, comportar-se

diferentemente entre dois ou mais grupos de indivíduos.

Com o avanço da TRI em avaliações de larga escala e de sua aplicação em geral, cresce o interesse pelo estudo de fatores a ela relacionados. O estudo de DIF é um deles. Essa dissertação visa desenvolver um método não paramétrico para detecção de DIF através da estimação das Curvas Características dos Itens (CCIs) via Regressão Isotônica, além de estudar os efeitos do DIF na estimação da proficiência.

Capítulo 2

Teoria de Resposta ao Item - TRI

2.1 Introdução

Quando se deseja medir alguma característica de indivíduos na área da educação ou da psicometria, muitas vezes não é possível medi-la diretamente, como inteligência, por exemplo, apesar de todos terem uma ideia de o que se está dizendo quando uma pessoa é descrita como inteligente. Essa característica é comumente denominada de traço latente. O traço latente é uma variável intuitivamente fácil de se entender, mas não tão simples de se medir; é uma variável que não pode ser observada diretamente, assim como "nível de depressão", "nível de vida saudável que uma pessoa leva", "proficiência em leitura", "inteligência", como foi dito, etc. Para estimar o traço latente de interesse é necessária a observação de variáveis secundárias relacionadas a ele. São duas as teorias utilizadas para estimar o quanto uma pessoa possui de determinado traço latente: a Teoria Clássica dos Testes - TCT, que utiliza o escore no teste como referência de medida, e a Teoria de Resposta ao Item - TRI, a qual tem como foco os itens e não o teste como um todo. Muitas vezes, a variável a ser estimada é a proficiência em escrita, leitura, matemática etc, então o termo genérico"proficiência"(ou habilidade) é usado quando se referir a um traço latente como esses.

A TRI permite comparar grupos diferentes que responderam, não necessariamente, a mesma prova. É necessário apenas que as provas tenham alguns itens em comum para se comparar populações que responderam a provas diferentes. No Brasil, o primeiro teste com a utilização da técnica derivada da TRI foi em 1995, na análise de dados do Sistema Nacional de Ensino Básico. Com a TRI, é possível comparar proficiência de estudantes do 5º e 9º ano do ensino fundamental e 3ª série do ensino médio que responderam a itens

diferentes em diferentes anos.

Diferentemente de características facilmente mensuráveis como, por exemplo, o peso ou a altura, que já possuem uma escala definida e popularizada, é necessário estabelecer uma escala de medida para definir quanto de um traço latente uma pessoa possui, pode-se dizer também definir uma régua. Por motivos técnicos, definir essa régua e o que representa cada número dessa régua é uma tarefa difícil. O padrão é que esta tenha como ponto central o valor zero. Ela teoricamente abrange proficiências no intervalo contínuo $(-\infty, +\infty)$ mas, na prática, um limite (-5, +5) é suficiente para descrever a proficiência de praticamente toda a população, que é o correspondente a proficiências distantes a 3 desvios padrão da média.

Usualmente, quando se deseja medir a proficiência de um grupo de indivíduos, elabora-se um teste, que contém itens (questões). Nesta dissertação se estudará DIF em itens de testes que medem apenas um traço latente, ou seja, cada teste mede uma característica de interesse. Além disso, os itens do teste são dicotômicos, pois para cada item (i), um indivíduo (j) pode respondê-lo corretamente ou não.

2.2 Os modelos

Esta dissertação se propõe a desenvolver um novo método para identificação de DIF. Assumindo que cada indivíduo que responda a um item do teste possua uma proficiência, cada examinado terá uma estimativa da proficiência que o posicionará em algum lugar da escala. Essa proficiência será denotada pela letra grega theta, θ. Para cada nível de proficiência, o indivíduo terá uma certa probabilidade de acerto para cada item. Essa probabilidade será denotada por P(θ). Para um item com comportamento padrão, essa probabilidade será baixa para pessoas com níveis baixos de proficiência e será alta para pessoas com níveis altos de proficiência. Em um gráfico da proficiência pela probabilidade de acerto dada a proficiência, o resultado será uma curva com a forma de um S suavizado, como é mostrado na figura 2.1. A probabilidade de acerto é próxima de zero para níveis baixos de proficiência quando não há possibilidade de acerto ao acaso, já para um item de múltipla escolha com 5 alternativas, por exemplo, essa probabilidade seria de 0,20. E cresce até os níveis mais altos da escala, onde a probabilidade de acerto é próxima de 1. Essa curva descreve a relação entre a probabilidade de um indivíduo responder positivamente

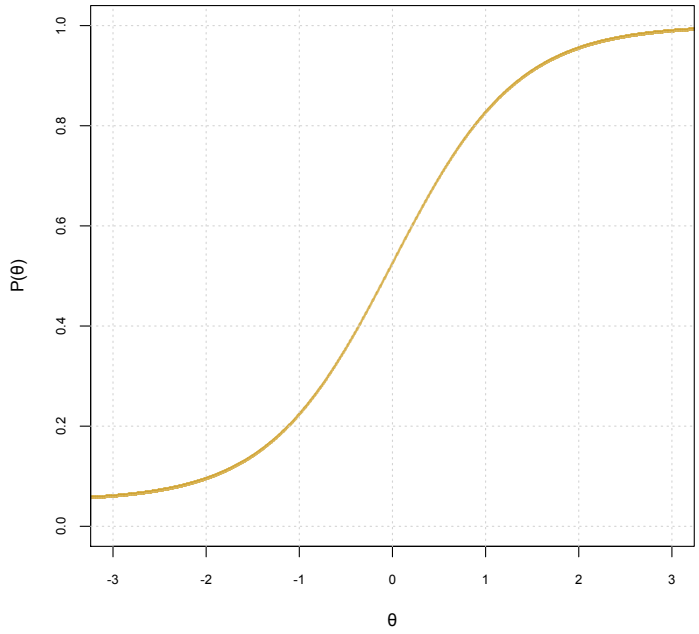


Figura 2.1: Representação de uma típica Curva Característica do Item.

um item e a escala de proficiência. Na Teoria de Resposta ao Item, é conhecida como Curva Característica do Item — CCI. A CCI é a base da TRI; qualquer aspecto que se vá estudar da teoria depende dessa curva. Portanto, para se estudar detecção de DIF, tema desta dissertação, é necessário esclarecer alguns pontos essenciais para determinação da CCI. Cada item em um teste terá sua CCI.

A TRI consiste em um grupo de modelos matemáticos que essencialmente têm o objetivo de representar a probabilidade que um respondente, j, com proficiência, θ, tem de responder positivamente um item, i. Os diversos modelos encontrados na literatura são determinados basicamente por três aspectos:

1. pela natureza do item — dicotômicos, politômicos ou resposta contínua;
2. pelo número de populações envolvidas — apenas uma ou mais de uma; e
3. pela quantidade de traços latentes que está sendo medida — apenas um ou mais de um.

Nesta dissertação serão apresentados modelos apenas para testes com itens dicotômicos, que envolvem apenas uma população e que medem apenas um traço latente, pois estas serão

as características dos testes os quais conterão os itens a serem testados sob a presença ou não de DIF. De acordo com Andrade et al. (2000), precursores no estudo da TRI no Brasil, os modelos mais utilizados são os modelos logísticos, os quais se diferenciam basicamente pela quantidade de parâmetros que levam em consideração para caracterizar o item. Os modelos são conhecidos como modelos logísticos de 1, 2 e 3 parâmetros e consideram respectivamente:

1. somente a dificuldade do item;
2. a dificuldade e a discriminação;
3. a dificuldade, a discriminação e a probabilidade de resposta correta dada por indivíduos de baixa proficiência, que seria a probabilidade de acerto ao acaso.

2.2.1 O Modelo Logístico de Um Parâmetro (ML1)

O Modelo Logístico de Um Parâmetro — ML1 — da TRI é frequentemente chamado de Modelo de Rasch, pois foi o matemático George Rasch (1960) que propôs o modelo unidimensional de um parâmetro. No ML1, apenas os parâmetros de dificuldade dos itens do teste são estimados (o parâmetro de discriminação dos itens é tratado como uma constante, considerado, portanto, igual para todos os itens). Embora teoricamente definido, na prática é difícil a utilização do ML1, pois presume que a função depende apenas da dificuldade do item e não de outras características. O ML1 é definido como

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

com $i = 1, 2, ..., I$ e $j = 1, 2, ..., n$, em que

U_{ij} é uma variável dicotômica que assume valor 1, quando o indivíduo j responde corretamente o item i e assume valor 0 quando o indivíduo j não responde corretamente o item i.

θ_j representa a proficiência (traço latente) do j-ésimo indivíduo.

$P(U_{ij} = 1|\theta_j)$ é a probabilidade de um indivíduo j, com proficiência θ_j responder corretamente ou concordar ou satisfazer as condições do item i, chamada de curva característica do item — CCI.

$\bar{a}$ representa a discriminação dos itens, é constante, fixa para todos os itens.

b_i é o parâmetro de dificuldade do item i, medido na mesma escala da habilidade.

D é um fator de escala, constante e igual a 1. Utiliza-se o valor 1,7 quando deseja-se que a função logística forneça resultados semelhantes ao da função ogiva normal.

Para o ML1, são necessárias duas premissas: a primeira, que já foi dita anteriormente, que todos os itens devem discriminar igualmente, ou seja, o parâmetro de discriminação, $\bar{a}$, é tratado como uma constante e, a segunda, o acerto casual é definido como zero.

O parâmetro b, único parâmetro estimado no ML1, é medido na mesma unidade da habilidade (θ). Para todos os modelos, o parâmetro b é a proficiência necessária para que uma pessoa tenha probabilidade de responder positivamente o item igual a $(1 + c)/2$. No ML1: $c = 0$, portanto, é a habilidade necessária para que a probabilidade de responder positivamente ao item seja igual a $1/2$. Consequentemente, quanto maior o valor de b maior deve ser a proficiência do respondente para ter alta probabilidade de responder positivamente o item e vice-versa.

Além disso, o ML1 implica que respondentes com o mesmo número de itens respondidos positivamente terão o mesmo valor estimado para θ (proficiência), independentemente de terem apresentado padrão de resposta diferente.

2.2.2 O Modelo Logístico de Dois Parâmetros (ML2)

Birnbaum (1968) inovou os modelos não lineares de 2 e 3 parâmetros sugeridos por Lord (1952) e propôs um modelo logístico não-linear para descrever dois parâmetros distintos do item, o ML2. Os dois parâmetros são (1) discriminação do item e (2) dificuldade do item. O ML2 é amplamente utilizado em instrumentos de medida nos quais não têm a possibilidade de o respondente "chutar"a resposta, portanto, não leva em consideração o parâmetro de acerto ao acaso, um exemplo seriam questionários de satisfação.

Com as representações gráficas dos parâmetros apresentados nas figuras 2.2 e 2.3 fica mais claro como os parâmetros impactam nessas duas características dos itens (discriminação e dificuldade).

A Figura 2.2 apresenta duas CCIs com aspectos distintos para dois itens, denominados de Item 1 e Item 2. As curvas têm inclinações diferentes, apesar de o ponto de inflexão

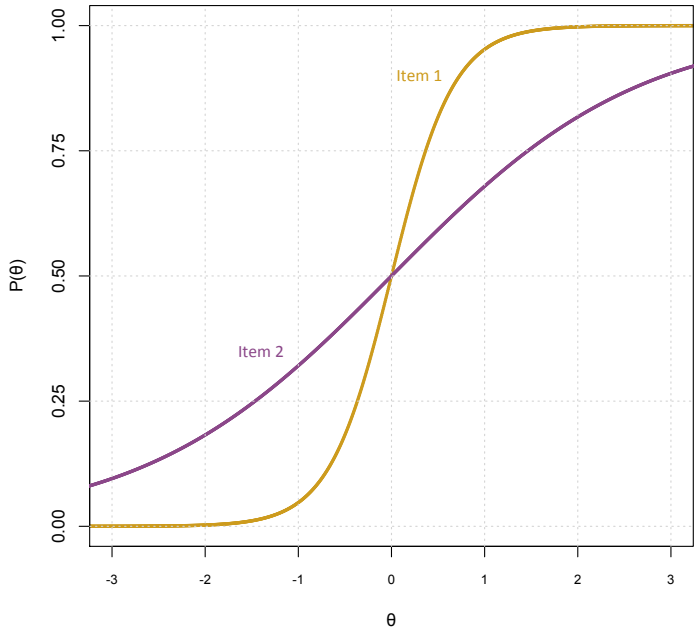


Figura 2.2: CCIs com mesmo ponto de inflexão (mesmo parâmetro b) mas com diferentes inclinações (diferentes parâmetros a).

das duas ser o mesmo, em $\theta = 0$ para o qual $P(\theta)$=0,5. Pelo gráfico, fica evidente que a inclinação da curva reflete uma característica do item. Ela reflete o poder de discriminação do item, que é a diferença de probabilidade de acerto do item para pessoas com proficiências diferentes. Um item que apresenta baixa inclinação é um item que não apresenta grande diferença na probabilidade de responder positivamente ao item independentemente da proficiência dos respondentes. A diferença entre as curvas revela a disparidade entre os dois itens no poder de discriminação dos respondentes com proficiências variadas. Para o Item 1, a discriminação é forte entre pessoas com proficiência acima de zero e abaixo de zero, para θ de -1 a 1 a curva é bastante íngreme, mostrando grande poder de discriminação nesse intervalo. Ao se comparar o Item 2 com o Item 1, percebe-se uma inclinação menor para todos os valores de θ, ou seja, um poder de discriminação menor. Essa característica do item refere-se ao parâmetro de discriminação (ou parâmetro a). No Item 2, indivíduos que apresentam proficiência de -1,5 têm probabilidade de acerto cerca de 0,5 menor que de indivíduos que apresentam proficiência igual a 1,5. Já para o Item 1, essa diferença de probabilidade é de quase 1.

Outros dois itens têm suas CCIs plotadas na figura 2.3. Os itens, denominados de Item

3 e Item 4, apresentam CCIs com inclinação idêntica no centro, no entanto, seus pontos de inflexão são distantes na escala do θ. Ou seja, o que diferencia as duas curvas é o deslocamento da esquerda para a direita e não a inclinação. Essa diferença no posicionamento das curvas mostra uma diferença na dificuldade do item. Esse parâmetro é chamado de parâmetro de dificuldade (ou parâmetro b). Para respondentes com mesma proficiência, o Item 4 é mais difícil pois está localizado mais à direita na escala do θ. Uma pessoa com habilidade $\theta = 0$ tem probabilidade de cerca de 0,85 de acertar o item 3 *versus* uma probabilidade de aproximadamente 0,12 de acertar o item 4.

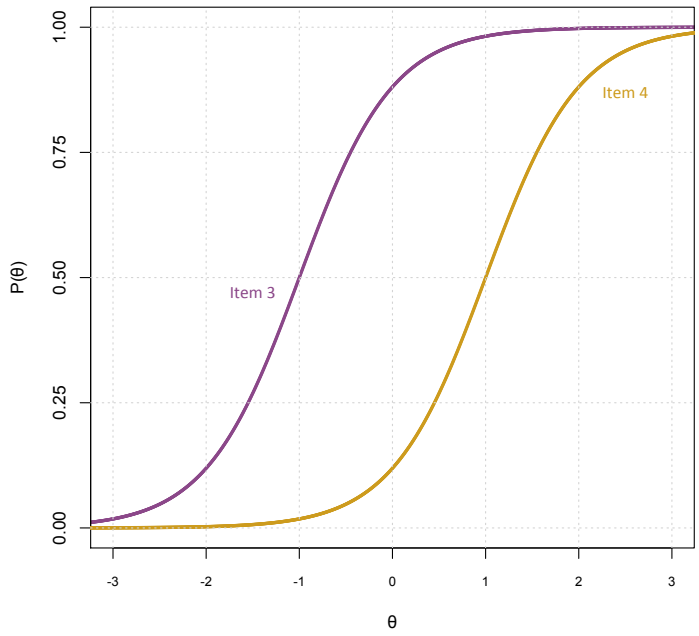


Figura 2.3: CCIs com mesma inclinação (mesmo parâmetro a) mas com diferentes pontos de inflexão (diferentes parâmetros b).

O Modelo Logístico de Dois Parâmetros - ML2 - é representado matematicamente como

$$P(U_{ij} = 1|\theta_j) = \frac{1}{1 + e^{-Da_i(\theta_j - b_i)}}. \tag{2.2}$$

Todos os termos da equação são os mesmos descritos para a equação (2.1) com a diferença de que para o ML2, cada item tem um parâmetro de discriminação e não é constante como para o ML1. No ML2, a_i é o parâmetro de discriminação do item i, com valor proporcional à inclinação da CCI no ponto b_i (que é seu ponto de inflexão). Aqui não é

esperado a_i negativo, pois indicaria que indivíduos com alta proficiência teriam probabilidade menor de acertar um item do que indivíduos com baixa proficiência. Se encontrado um item com esta característica, é um indicativo de que o item precisa de revisão. Baixos valores de a_i indicam itens com baixo poder de discriminação, os quais apresentam probabilidades de acerto semelhantes para indivíduos com habilidades bastante diferentes. Valores muito altos de a_i indicam itens que discriminam respondentes em praticamente dois grupos. No limite, o grupo com proficiência acima de b_i terá probabilidade de acerto igual a 1 e o grupo com proficiência abaixo de b_i terá probabilidade de acerto igual a 0. Portanto, para comparar indivíduos dentro do mesmo grupo, o item não é útil quando a_i apresentar valores muito altos.

2.2.3 O Modelo Logístico de Três Parâmetros (ML3)

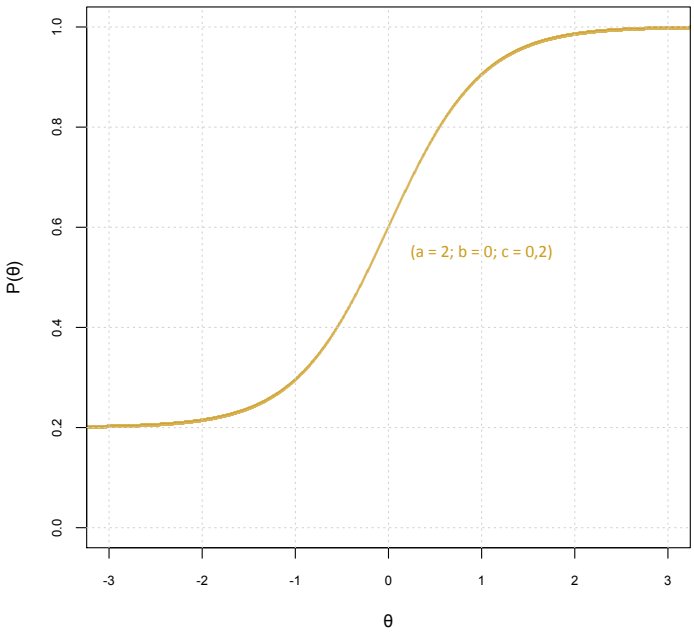


Figura 2.4: Curva Característica do Item para um Modelo Logístico de Três Parâmetros

Dentre os três modelos logísticos da TRI, o Modelo Logístico de Três Parâmetros — ML3 — é o mais completo e, atualmente, o mais utilizado (Andrade et al., 2000). Além de incluir em seu modelo os dois parâmetros considerados no ML2, inclui também o parâmetro de acerto ao acaso (parâmetro c). Esse terceiro parâmetro representa a probabilidade de

acerto para respondentes com proficiência muito baixa. O modelo completo para o ML3 é dado por

$$P(U_{ij} = 1|\theta_j) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta_j - b_i)}}. \tag{2.3}$$

Todos os termos são os mesmos definidos para a equação (2.2) com o acréscimo do parâmetro c_i, que é definido como o parâmetro de acerto ao acaso do item i, ou seja, é a probabilidade de um indivíduo com proficiência assintoticamente baixa, acertar o item i.

A representação gráfica de um modelo completo ML3, é mostrada na figura 2.4. Para esse modelo, $c = 0{,}2$, ou seja, a probabilidade de acerto do item para proficiências muito baixas é 0,2 e o ponto de inflexão, com $b = 0$, é em $\theta = 0$. Seria um exemplo de item de múltipla escolha com 5 alternativas.

2.2.4 Função de Informação do Item

Uma medida poderosa utilizada, em conjunto com a CCI, para descrever itens e testes, para selecionar itens dos testes e para comparar testes é a *Função de Informação do Item* - FII, denotada por $I_i(\theta)$. Ela retorna quanto um item (ou teste) contribui para a medida do traço latente. $I_i(\theta)$ é definida por

$$I_i(\theta) = \frac{\left[\frac{d}{d\theta}P_i(\theta)\right]^2}{P_i(\theta)Q_i(\theta)}, \tag{2.4}$$

em que

$I_i(\theta)$ é a "informação"fornecida pelo item i no ponto θ;

$P_i(\theta)$ é a função de resposta ao item, igual a $P(U_{ij} = 1|\theta_j)$; e

$Q_i(\theta) = 1 - P_i(\theta)$

A equação (2.4) é válida para o ML1 e ML2, descritos nas equações (2.1) e (2.2). Para o Modelo Logístico de Três Parâmetros, descrito na equação (2.3), $I_i(\theta)$ é dada por

$$I_i(\theta) = D^2a_i^2\frac{Q_i(\theta)}{P_i(\theta)}\left[\frac{P_i(\theta) - c_i}{1 - c_i}\right]^2. \tag{2.5}$$

Fica explícito da equação (2.5) e da figura 2.5 como cada parâmetro contribui para informação do item. Isto é, $I_i(\theta)$ é maior:

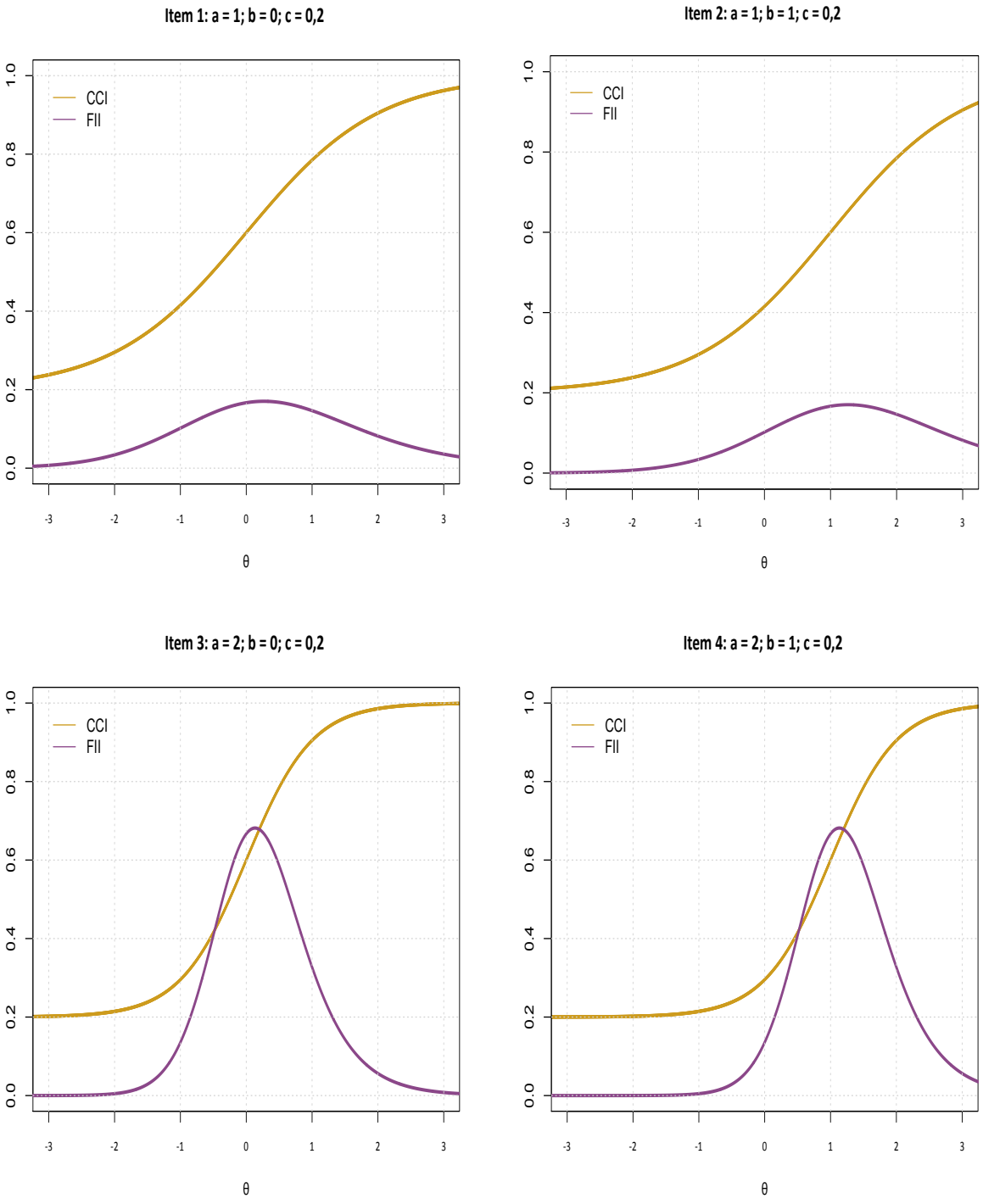


Figura 2.5: Representação da Curva Característica e de Informação de 4 Itens.

i. quanto mais perto θ estiver de b_i ;

ii. quanto maior for o parâmetro a_i; e

iii. quanto mais próximo de 0 o parâmetro c_i estiver

A função de informação do item tem um papel importante no desenvolvimento do teste e avaliação dos itens, pois ela indica qual a contribuição de cada item para todos os valores de estimativa da habilidade. Essa contribuição depende do poder de discriminação do item (quanto maior, mais inclinada a curva e maior o salto em $P(\theta)$) e essa contribuição também depende da dificuldade do item).

Birnbaum (1968) mostrou que o item fornece a maior informação em $\theta_{\max}$ dado por

$$\theta_{\max} = b_i + \frac{1}{Da_i}\ln\left[0,5\left(1+\sqrt{1+8c_i}\right)\right]. \tag{2.6}$$

Se o acerto ao acaso é mínimo, ou seja, $c_i = 0$, então $\theta_{\max} = b_i$. Em geral, quando $c_i > 0$, a maior informação do item é em θ um pouco maior que o nível de dificuldade b_i do item, como está representado pelas curvas na figura 2.5.

Quatro itens estão representados na figura 2.5. Ao se comparar os itens 1 e 3 (e os itens 2 e 4), visualiza-se o que a equação (2.4) indica; quanto mais inclinada a curva (maior o parâmetro a_i), maiores são os valores da informação nos pontos de máximo apresentados pela função de informação do item. Quando comparados os itens 1 e 2 (e os itens 3 e 4) nota-se também que a FII acompanha a dificuldade do item; o seu ponto de máximo é um pouco à frente do parâmetro b_i pois o parâmetro c_i é maior que 0 (quando $c_i = 0$, o ponto de máximo da FII coincide com o parâmetro b_i).

A utilidade da função de informação do item no desenvolvimento e avaliação do teste depende de quão ajustada está a CCI aos dados do teste. Se está mal ajustada, as correspondentes estatísticas do item e a função de informação do item, consequentemente, não serão confiáveis. Mesmo se a CCI estiver bem ajustada, um item pode ter interpretação limitada se seu parâmetro a_i é baixo e seu parâmetro b_i é alto. Além disso, o uso dos itens (questões) vai depender da necessidade de cada teste. Isso porque um mesmo item pode retornar bastante informação em um intervalo de habilidade, mas apresentar baixa $I_i(\theta)$ para outro intervalo de habilidade (Hambleton et al., 1991).

2.2.5 Função de Informação do Teste

A informação de um teste em θ é simplesmente a soma das funções de informação dos itens em θ. Ela é denotada por $I(\theta)$ e é dada por

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta). \tag{2.7}$$

Pela equação (2.7) fica claro que a contribuição de cada item para a informação do teste é independente. Portanto, as funções de informação dos itens podem ser determinadas sem o conhecimento dos outros itens do teste. Isso não é possível na Teoria Clássica do Teste. Na TCT, a contribuição dos itens do teste para a confiabilidade do teste e os índices de discriminação dos itens não podem ser determinados independentemente das características dos outros itens do teste. Isso porque o escore do teste, que é usado nesses cálculos, é dependente da escolha dos itens. Ao mudar um item do teste, o escore do teste e os índices do teste mudarão também.

A informação de um teste na habilidade θ está relacionada inversamente com a precisão que a habilidade é estimada naquele ponto. Essa relação é dada por

$$\mathrm{SE}(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}. \tag{2.8}$$

$\mathrm{SE}(\hat{\theta})$ é o erro padrão de estimação. Como $I(\theta)$ depende de θ, o erro padrão da estimação da proficiência também depende, diferentemente do erro padrão de estimação da teoria clássica.

Capítulo 3

Funcionamento Diferencial do Item

3.1 Introdução

Resultados provenientes de avaliações educacionais são amplamente utilizados no Brasil e em todo o mundo. Notas obtidas a partir de provas elaboradas pela TRI ou TCT são utilizadas desde as primeiras etapas educacionais, passando pela disputada entrada em uma instituição de ensino superior até uma prova para se alcançar o tão sonhado emprego. Usuários do teste (seja quem está respondendo o teste, quem elaborou o teste ou quem solicitou o teste — uma instituição de ensino ou quem provém a vaga de emprego, por exemplo) acreditam que o resultado do teste é comparável entre vários grupos e que esta é uma comparação justa.

Ao se construir um instrumento de avaliação, o que se deseja é que a nota alcançada por indivíduo seja uma medida relacionada apenas com a habilidade do indivíduo na variável medida. O maior desafio, portanto, ao se elaborar um teste e, provavelmente, o que mais importa para o público, é que este seja justo, não favorecendo assim um grupo em detrimento de outro, principalmente quando o teste a que se refere tem poder de seleção.

De acordo com Holland e Wainer (2012), os primeiros estudos do que mais tarde ficou conhecido como *viés do item* (do inglês, *item bias*) aparentemente foram realizados no começo dos anos 1960. Os estudos foram idealizados para desenvolver métodos que buscavam analisar diferenças culturais e para investigar a afirmação de que a principal razão para a grande disparidade entre resultados, em testes de habilidade cognitiva nos EUA, de estudantes negros e hispânicos e de estudantes brancos é que os testes contém itens que estão fora do domínio das culturas minoritárias. O pressuposto era de que esses itens tratavam de conteúdos que estudantes de culturas minoritárias tinham menos oportunidade

de aprender. O objetivo principal desses estudos era, e continua sendo, a identificação de qualquer item que seja enviesado e remoção desse item do teste.

O termo *bias* em inglês é definido no dicionário como divergência da verdade ou uma informação que não é correta devido ao método utilizado na coleta ou na forma de apresentá-la. No âmbito da estatística, isso se traduz na tendência de uma estimativa desviar do valor verdadeiro. No contexto social, seria um conceito próximo ao preconceito, uma motivação irracional de comportamento.

Com o intuito de verificar a validade da hipótese de que grupos minoritários apresentavam um desempenho inferior em relação a grupos majoritários por captar injustamente o conhecimento e habilidades que faziam parte da cultura de brancos de classe-média, por exemplo, e não de outras culturas, técnicas foram desenvolvidas para que se pudesse identificar esses itens enviesados. No entanto, isso era apenas uma descoberta estatística, sem uma interpretação ou julgamento. Ao interpretar os itens, alguns deles realmente eram enviesados, no sentido de que as minorias se deparavam com uma injusta desvantagem ao respondê-lo. Outros itens foram julgados justos, mesmo apresentando diferença entre os grupos, por se tratar de um conhecimento educacional importante para todos os estudantes, que no entanto era conhecido de forma desigual entre os grupos. E ainda havia outros itens para os quais a razão para a diferença encontrada entre os grupos era desconhecida.

Ficou claro que estava havendo um problema semântico, no qual a mesma palavra, *viés*, estava sendo usada com, pelo menos, dois diferentes significados: estatístico e social. Sugestões foram dadas para que outro termo fosse usado diferentemente de *viés* quando se tratasse apenas de uma constatação estatística, livre de uma análise e avaliação do viés em um contexto social. Finalmente a expressão *funcionamento diferencial do item* (do inglês, *differential item funcioning* — DIF) passou a ser utilizada.

Hambleton et al. (1991) caracteriza um item que apresenta DIF como um item o qual indivíduos que, apesar de apresentarem a mesma habilidade que está sendo medida pelo teste, mas que pertençam a grupos diferentes, não apresentam a mesma probabilidade de acerto do item. De um modo geral, DIF refere-se a diferenças em propriedades psicométricas dos itens entre os grupos [Aliste (1996)]. Ao conduzir uma análise de DIF, é comum que se tenha, pelo menos, dois grupos de interesse: o grupo focal e o grupo de referência. O primeiro, geralmente refere-se a uma minoria ou a um grupo tradicionalmente considerado em desvantagem, enquanto que o segundo, refere-se a uma maioria ou a um

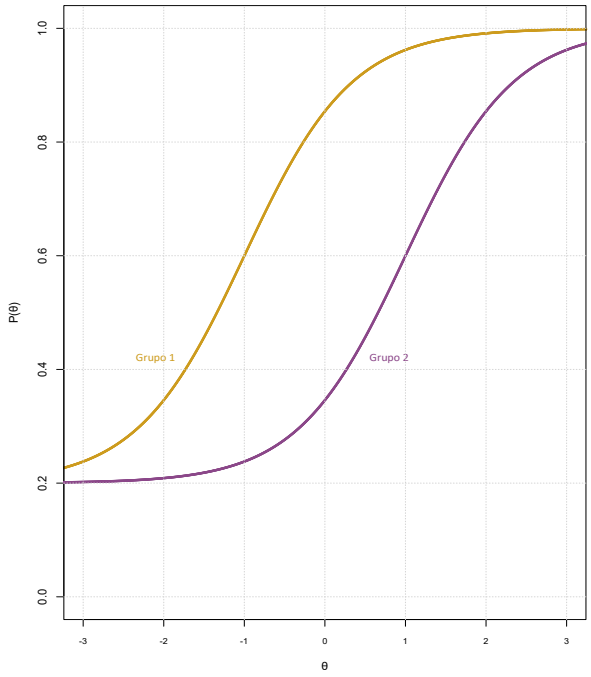
grupo privilegiado [Cuevas e Cervantes (2012)].

Em 1988, Holland e Thayer colaboraram para definir a diferença entre os dois conceitos [Herrera et al. (2005)]. Hoje em dia, viés é utilizado para se referir a uma opinião que leva em consideração o objetivo do teste assim como a informação contextual dos grupos, que pode explicar o DIF em um item específico. No geral, o estudo do DIF é o primeiro passo, a análise estatística, a fim de decidir se um item estaria enviesado em relação a um grupo particular [Cuevas e Cervantes (2012)].

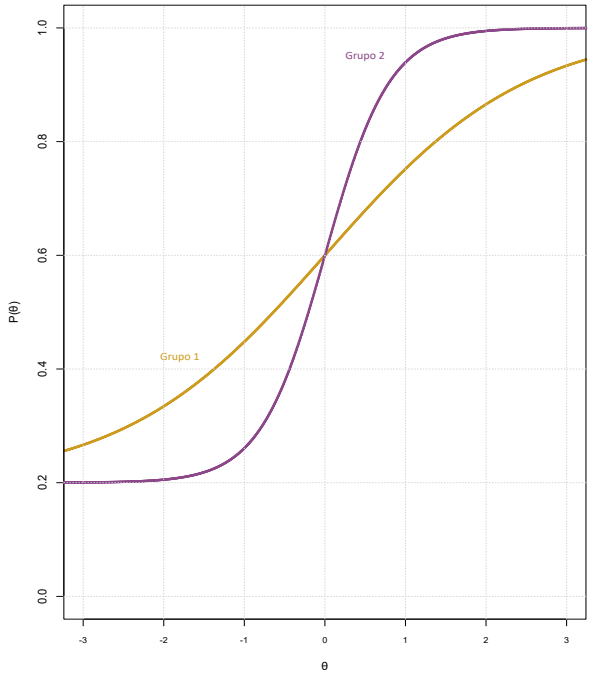
3.2 DIF: Uniforme ou Não-uniforme

Angoff (1993), um dos pioneiros no estudo de viés do teste no *Educational Testing Service (ETS)* — maior organização privada de avaliação e teste educacional sem fins lucrativos do mundo, responsável pela elaboração de inúmeros testes nos Estados Unidos — comentou que elaboradores de testes afirmaram que eles são frequentemente confrontados com resultados de DIF que eles não entendem; mesmo estudando o item que apresenta DIF, não conseguem chegar a uma explicação do motivo de alguns itens que parecem ser perfeitamente construídos apresentarem valores elevados de DIF. A ambiguidade e confusão vivida por desenvolvedores de testes talvez tenha alguma relação com a suposição, que nem sempre fica clara, de que o teste é unidimensional, ou seja, mede apenas um traço latente. De fato, a presença de DIF pode indicar que o teste está medindo um fator secundário (Osterlind e Everson, 2009) ou, muitas vezes, apenas não se sabe a causa raiz do DIF encontrado.

Os tipos de DIF são divididos em dois mais comuns, *uniforme* e *não-uniforme*. Como foi visto, um item apresenta DIF quando existe uma dependência condicional entre a probabilidade de acerto e a qual grupo o examinado pertence. *DIF uniforme* é a forma mais simples de DIF e ocorre quando a dependência condicional é constante para todos os valores de θ. Isto é, um grupo tem uma vantagem constante em relação ao outro para todos os valores de θ, ou seja, as CCIs tem o mesmo parâmetro de discriminação, mas o parâmetro de dificuldade muda [Camilli e Shepard (1994)]. Isso pode ser visto como evidência de que uma outra dimensão está sendo abordada pelo item e que os grupos diferem na distribuição dessa dimensão [Herrera Rojas (2006)]. Quando um item é classificado como apresentando DIF não-uniforme, o parâmetro de difculdade é o mesmo, mas o de discriminação, não



(a) DIF uniforme



(b) DIF não-uniforme

Figura 3.1: Representação de DIF uniforme (primeiro gráfico) e DIF não-uniforme (segundo gráfico).

[Cuevas e Cervantes (2012)]. Nesse caso, a interpretação do DIF implicaria que a variância dos grupos na dimensão irrelevante não é a mesma ou que a correlação entre as duas

dimensões é diferente entre os grupos [Herrera Rojas (2006)]. Por último, quando um item apresenta tanto DIF uniforme quanto não-uniforme, tanto a dificuldade quanto a discriminição são diferentes entre os grupos [Herrera Rojas (2006)].

As diferenças entre DIF uniforme e não-uniforme estão ilustradas na figura 3.1, na qual foram plotadas as curvas características de dois itens para dois grupos.

A diferença entre as CCIs sugere que respondentes dos dois grupos, com mesma proficiência, não têm a mesma probabilidade de acertar o item. DIF uniforme é evidente quando as CCIs dos dois grupos são diferentes, não se cruzam e um grupo apresenta uma vantagem constante em relação ao outro. DIF não-uniforme, diferentemente, ocorre quando as CCIs dos dois grupos são diferentes e a diferença entre as probabilidades de acerto para os dois grupos varia dependendo do θ. Elas podem se cruzar ou não em algum valor de θ. Ao analisar os gráficos, fica evidente que a área entre as CCIs indica o grau de DIF.

É importante notar que, quando as CCIs se cruzam, parte da área pode ser tratada como DIF positivo e parte como DIF negativo (Camilli e Shepard, 1994). Essas distinções são fundamentais para se entender e interpretar análises de DIF. Métodos diferentes podem ser necessários dependendo se há suspeita de DIF uniforme ou não-uniforme.

Um dos objetivos desta dissertação é o desenvolvimento de um método para detecção de DIF. Nas subseções a seguir, serão descritos alguns métodos já apresentados na literatura.

3.3 Método Mantel-Haenszel para detecção de DIF

Com o intuito de contornar dificuldades na utilização da simples estatística de qui-quadrado, já que esta é inapropriada quando a frequência observada é pequena (circunstância comum ao se estudar DIF), na década de 50, dois pesquisadores da área da medicina - o bioestatístico Nathan Mantel e o epidemiologista Willian Haenszel - desenvolveram um procedimento baseado na distribuição qui-quadrado para comparação de grupos, mas adaptado para amostras estratificadas (Mantel e Haenszel, 1959). Mais de duas décadas depois, esse procedimento de qui-quadrado modificado foi retomado e adaptado por Holland et al. (1985) e Holland e Thayer (1988) para o uso de detecção de DIF. Essa metodologia desenvolvida é chamada de Método de Mantel-Haenszel (M-H) e é hoje, possivelmente, o método mais usado de detecção de DIF (Osterlind e Everson, 2009).

3.3.1 Tabela de Contingência 2 × 2 × M

Para a utilização do método M-H, os dois grupos a serem analisados são divididos em M estratos de proficiência (ou habilidade) baseados no escore total do teste, definindo assim grupos que serão comparados para detecção de DIF. Usualmente, os indivíduos são divididos em 4 ou 5 níveis de habilidades. Em seguida, para cada nível de habilidade, é construída uma tabela de contingência *2* x *2* ou uma grande tabela de contingência tridimensional *2* × *2* × *M*. Na tabela, é apresentada a frequência de respostas certas e erradas para cada grupo de habilidade.

Tabela 3.1 - Tabela de contingência *2* (grupos) × *2* (escores do item) × *M* (nível de habilidade) apresentada em partes

	Escore do Item		
	Certo (C)	Errado (E)	Total(N)
Grupo 1 (1)	C_{1m}	E_{1m}	N_{1m}
Grupo 2 (2)	C_{2m}	E_{2m}	N_{2m}
Total (t)	C_{tm}	E_{tm}	N_{tm}

Para cada item, cada respondente tem o escore de acerto ou erro (aqui estão incluídas as não-respostas). A soma de acertos e erros pode ser organizada em uma tabela de contingência *2* × *2* × *M* para cada item a ser estudado. São dois níveis de grupos, grupo 1 e grupo 2, que também pode ser chamado de grupo *focal* que é o foco da análise e *de referência*, que serve como base de comparação do grupo focal; são dois níveis de resposta do item, *certo* ou *errado*; e são M níveis de habilidade na variável de interesse (que são os próprios itens). A tabela de contingência *2* (grupos) × *2* (escores do item) × *M* (nível de habilidade) para cada item está representada em partes de *2* x *2* (são M partes por item) na Tabela 3.1.

A hipótese nula de DIF para o método M-H pode ser expressa como

$$H_0 : \frac{\left[\frac{C_{1m}}{E_{1m}}\right]}{\left[\frac{C_{2m}}{E_{2m}}\right]} = 1 \tag{3.1}$$

em que $m = 1, ..., M$ ou, alternativamente, como

$$H_0 : \left[\frac{C_{1m}}{E_{1m}}\right] = \left[\frac{C_{2m}}{E_{2m}}\right] \tag{3.2}$$

com $m = 1, ..., M$.

Em outras palavras, a hipótese nula estipula que a chance de acertar um dado item é a mesma tanto para o grupo 1 como para o grupo 2, para todos os M níveis de habilidade. Se essa igualdade não é satisfeita para qualquer um dos M níveis, rejeita-se a hipótese nula e conclui-se que o item analisado apresenta DIF.

3.3.2 Razão de Chance de M-H

Em seu trabalho original, Mantel e Haenszel (1959) desenvolveram um teste de qui-quadrado da hipótese nula de DIF contra a hipótese alternativa conhecida com a hipótese da razão de chance constante,

$$H_1 : \frac{C_{2m}}{E_{2m}} = \alpha \frac{C_{1m}}{E_{1m}},\ m = \ 1,\ ...,\ M \text{ e } \alpha \neq 1. \tag{3.3}$$

Quando $\alpha = 1$, a hipótese alternativa se reduz à hipótese nula de DIF. O parâmetro α é chamado de *razão de chance comum* na M-ésima tabela 2×2, porque sob H_1, o valor de α é a mesma razão de chance para todo m,

$$\alpha_m = \frac{\dfrac{C_{2m}}{E_{2m}}}{\dfrac{C_{1m}}{E_{1m}}} = \frac{C_{2m}E_{1m}}{C_{1m}E_{2m}} \tag{3.4}$$

Se não houver diferença entre os grupos, então a razão de chance é igual a 1 (isto é, $\alpha_m = 1$), indicando equilíbrio entre o grupo 1 e grupo 2 e interpretada como ausência de DIF. No entanto, quando $\alpha_m > 1$, o grupo 2 apresenta um percentual de respostas corretas significativamente maior que o grupo 1, pelo menos para aquela faixa de proficiência. De modo contrário, quando $\alpha_m < 1$, o grupo 1 apresenta um percentual de respostas corretas significativamente maior que o grupo 2 para aquele nível de proficiência.

Mantel e Haenszel (1959) também propuseram uma estimativa da razão de chance comum, $\hat{\alpha}_{\text{MH}}$, dada por

$$\hat{\alpha}_{\text{MH}} = \frac{\sum_m \dfrac{C_{1m}E_{2m}}{N_{tm}}}{\sum_m \dfrac{C_{2m}E_{1m}}{N_{tm}}}. \tag{3.5}$$

Esta é uma estimativa do tamanho do efeito do DIF, que vai de 0 a ∞, em que $\alpha_{\text{MH}} = 1$ indica ausência de DIF. Devido a dificuldade de interpretação, a transformação logarítmica

apresentada na equação (3.6), conhecida por $MH_{\text{D-DIF}}$, mais utilizada por ser simétrica em torno do zero e mais fácil de interpretar,

$$MH_{\text{D-DIF}} = -2,35\ln\left(\hat{\alpha}_{\text{MH}}\right). \tag{3.6}$$

A nova escala traz o centro do índice para $MH_{\text{D-DIF}} = 0$, onde 0 é interpretado como ausência de DIF. O sinal de menos resulta na interpretação de que quando $MH_{\text{D-DIF}}$ é positivo, o DIF é em favor do grupo 1 e quando é negativo, o DIF é em favor do grupo 2.

3.3.3 Estatística do Teste Qui-Quadrado

A estatística do teste associada ao método M-H é dada por

$$MH - \chi^2 = \frac{\left[\left|\sum_m C_{2m} - \sum_m \text{E}\left(C_{2m}\right)\right| - 0,5\right]^2}{\sum_m \text{Var}\left(C_{2m}\right)} \tag{3.7}$$

em que

$$\text{E}\left(C_{2m}\right) = \text{E}\left(C_{2m}|\alpha = 1\right) = \frac{N_{2m}C_{tm}}{N_{tm}},$$
$$\text{Var}\left(C_{2m}\right) = \text{Var}\left(C_{2m}|\alpha = 1\right) = \frac{N_{2m}C_{tm}N_{1m}E_{tm}}{N_{tm}^2\left(N_{tm} - 1\right)},$$

e $MH - \chi^2$ tem distribuição aproximada de uma qui-quadrado com um grau de liberdade.

Explicação completa e mais detalhada do método M-H pode ser encontrada em Osterlind e Everson (2009) e Holland e Wainer (2012).

3.4 Regressão logística para detecção de DIF

Regressão logística tem sido utilizada como método para detecção de DIF em diferentes contextos, como saúde, educação e psicologia. O método consiste em ajustar os modelos nas equações (3.8), (3.9) e (3.10).

$$logit(P(U = 1)) = \beta_0 + \beta_1\theta + \beta_2 g + \beta_3\theta g \tag{3.8}$$

$$logit(P(U = 1)) = \beta_0 + \beta_1\theta + \beta_2 g \tag{3.9}$$

$$logit(P(U = 1)) = \beta_0 + \beta_1\theta \tag{3.10}$$

em que

- $P(U = 1)$ é a probabilidade de responder corretamente, concordar ou satisfazer as condições de um determinado item;

- θ é a proficiência do examinado no teste ou o escore total; e

- g é o grupo ao qual o examinado pertence.

A comparação entre esses modelos pela estatística G^2 (estatística do teste de razão de verossimilhança com distribuição χ^2 com número de graus de liberdade igual à diferença entre o número de parâmetros dos modelos comparados) permite identificar se o item apresenta DIF (THISSEN (1988)) assim como o tipo de DIF: uniforme, não uniforme ou ambos. Cuevas e Cervantes (2012) apontam que, além disso, o teste deve ser complementado pela medida do efeito do tamanho da amostra a fim de adicionar informação para o desenvolvedor do teste sobre a magnitude das diferenças entre os dois grupos. Segundo Kirk (1996), pequenas amostras não apresentam efeitos estatísticos interessantes e grandes amostras podem se deparar com resultados estatisticamente significativos em que o efeito é muito pequeno e não há significância prática. Nesse contexto, o $R^2\Delta$, definido como a diferença dos R^2 entre os modelos, é proposto como a medida de efeito natural do tamanho da amostra (Zumbo (1999)).

De acordo com ESPITIA (2009), um item apresenta DIF uniforme, em termos estatísticos, se a estatística G^2 é significante entre os modelos (3.9) e (3.10). Um item apresenta DIF não-uniforme se a estatística G^2 é significante entre os modelos (3.8) e (3.9). Se um item apresentar tanto DIF uniforme quanto DIF não uniforme, ele é classificado como apresentando DIF misto.

3.5 DIF e a TRI

Progresso significativo, marcado como um dos maiores avanços na história da teoria do teste, foi dado pela publicação da dissertação de F. M. Lord (Lord, 1952), na qual Lord explica o modelo da Teoria de Resposta ao Item (TRI). Não demorou muito para que esse modelo passasse a ser amplamente utilizado no estudo de DIF. Assumindo que a habilidade que está sendo avaliada é unidimensional e que o item mede esta mesma habilidade,

a curva característica do item é completamente determinada por seus parâmetros, independentemente da natureza do grupo para qual se está definindo a curva. A curva, como já foi explicado no capítulo 2, na maioria das vezes, é definida em três parâmetros: a, que é proporcional à inclinação da curva e representa o poder de discriminação do item; b, o nível de proficiência, θ, no ponto de inflexão, representando a dificuldade do item; e c, o parâmetro de acerto ao acaso, correspondente à probabilidade de pessoas com proficiências muito baixas acertarem o item.

Pela natureza única da curva característica do item, o fato das curvas de dois grupos não serem as mesmas é evidência de que pelo menos um dos três parâmetros é diferente entre os grupos. Testes estatísticos estão disponíveis para testar a significância dessa diferença em relação aos três parâmetros (THISSEN, 1988). Analogamente, um item não apresenta DIF se as curvas características do item para todos os grupos forem idênticas (Hambleton et al., 1991). A diferença entre as curvas também pode ser testada pela análise da área gerada entre as curvas — isto é, quanto as curvas estão distantes entre si.

A vantagem ao se utilizar as curvas geradas pela TRI para estudar a presença de DIF no item é que, diferentemente de outros métodos, não apenas a diferença entre grupos que diz respeito à dificuldade como também diferenças em relação ao poder de discriminação ou até mesmo ao parâmetro de acerto ao acaso são levadas em consideração. Outros métodos ignoram diferenças entre itens no que diz respeito ao poder de discriminação e acerto ao acaso ou assumem que elas não existem. O fato de existirem essas diferenças de comportamento de resposta dos itens entre grupos, especialmente no que diz respeito ao poder de discriminação, fez da TRI o método usado para estudar DIF e para identificar itens com DIF.

Na seção seguinte será descrito um método que é baseado na TRI para estudar DIF, comparando os parâmetros dos itens.

3.5.1 Comparação dos parâmetros dos itens

Se os parâmetros de duas funções características do item forem idênticos, então as curvas serão idênticas em todos os seus pontos e a probabilidade de uma resposta correta será a mesma para os dois grupos para qualquer θ. A hipótese nula de que a função característica do item é a mesma para o grupo 1 e 2 é definida como

$$H_0 : a_1 = a_2; b_1 = b_2; c_1 = c_2 \tag{3.11}$$

na qual, os números subscritos denotam o grupo que gerou a estimativa do respectivo parâmetro. Se a hipótese for rejeitada para algum dos três parâmetros, pode-se dizer que este item possui DIF.

Para testar a hipótese nula, as estimativas dos parâmetros do item e as matrizes de variância e covariância das estimativas são necessárias. Ao se estimarem os parâmetros do item e da habilidade em cada grupo, a escala dos parâmetros deve ser especificada. Na maioria das vezes, isto é feito normalizando a estimativa da habilidade ou da dificuldade em cada grupo. Padronizar a estimativa da habilidade normalmente resulta em escalas diferentes para cada grupo e as estimativas dos parâmetros do item não estarão na mesma escala. No entanto, padronizar o parâmetro de dificuldade, resultará em estimativas dos parâmetros do item que estarão na mesma escala.

Depois que as estimativas dos parâmetros estão em uma mesma escala, a matriz de variância e covariância das estimativas dos parâmetros é computada para cada grupo. Primeiro, a matriz de informação é computada para cada grupo e é invertida. As matrizes de variância e covariância dos dois grupos são adicionadas para produzir a matriz de variância e covariância da diferença entre as estimativas. A estatística para testar a hipótese nula é dada por

$$\chi^2 = (a_{\text{diff}} b_{\text{diff}} c_{\text{diff}})' \sum{}^{-1} (a_{\text{diff}} b_{\text{diff}} c_{\text{diff}}), \tag{3.12}$$

em que

$$a_{\text{diff}} = a_2 - a_1; b_{\text{diff}} = b_2 - b_1; c_{\text{diff}} = c_2 - c_1$$

e $\sum$ é a matriz de variância e covariância das diferenças entre as estimativas dos parâmetros. O teste estatístico (para grandes amostras), assintoticamente, segue uma distribuição qui-quadrado com p graus de liberdade, o qual p é o número de parâmetros comparados. Para modelos de três parâmetros, quando a, b e c são comparados para cada item, p=3; para modelos de dois parâmetros, p=2; para modelos de um parâmetro, p=1. No caso do modelo de um parâmetro, a expressão para a estatística de qui-quadrado é consideravelmente mais simples; a estatística do teste nesse caso é

$$\chi^2 = \frac{b^2_{\text{diff}}}{\nu(b_1) + \nu(b_2)}, \tag{3.13}$$

na qual $\nu(b_1)$ e $\nu(b_2)$ são as recíprocas das funções de informação para a estimativa do parâmetro de dificuldade. O parâmetro c, na maioria das vezes, é mal estimado e, portanto, apresenta elevados erros padrão. A inclusão do parâmetro no teste estatístico pode produzir um teste bastante conservador, ou seja, um teste que não é poderoso na detecção de DIF. Uma alternativa a isso é comparar apenas os parâmetros a e b e ignorar o parâmetro c. Essa se mostra uma opção razoável porque se existir diferença nos parâmetros a e b entre os grupos, então as funções característica dos grupos serão diferentes, independentemente dos valores do parâmetro c; se não houver diferença estatisticamente significativa entre os parâmetros a e b, diferença entre os parâmetros c dos grupos seria muito irreal para se chegar à conclusão de que a função característica do item é diferente (Lord, 1980).

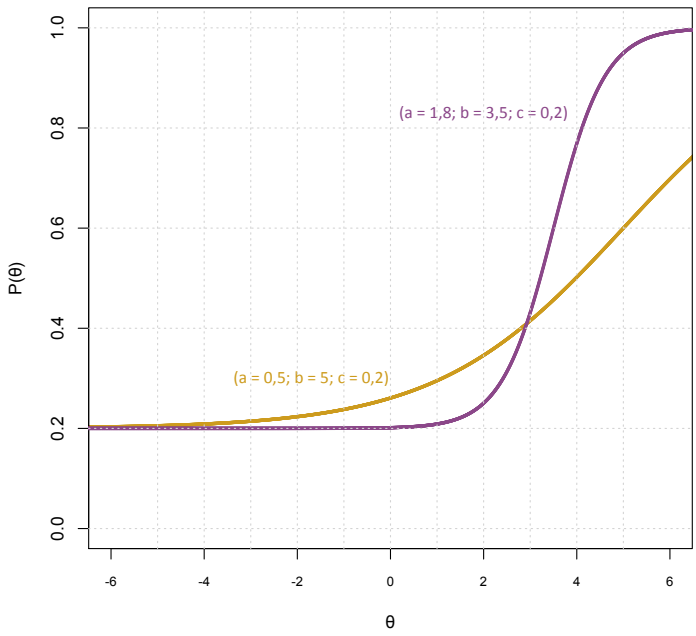


Figura 3.2: Representação das curvas característica do item para dois grupos

A comparação dos parâmetros dos itens como meio de comparar as CCIs para dois grupos — ou seja, como meio de detecção de DIF — já foi criticada com a justificativa de que diferenças significativas entre os parâmetros podem ser encontradas quando, na prática, não existe diferença entre as CCIs em uma faixa de habilidade de interesse específica. No

artigo de Linn et al. (1981) um exemplo foi dado em que os parâmetros dos itens eram significativamente diferentes mas que produziam, para uma certa faixa de proficiência (-3,3), curvas bem próximas, que não excedia em 0,1 a diferença de probabilidade de acerto entre os dois grupos para uma mesma proficiência. No exemplo, os parâmetros dos itens para cada grupo eram os seguintes:

$$\begin{array}{lllll} \text{Grupo 1:} & \text{a} = 1{,}8; & \text{b} = 3{,}5; & \text{c} = 0{,}2 \\ \text{Grupo 2:} & \text{a} = 0{,}5; & \text{b} = 5{,}0; & \text{c} = 0{,}2 \end{array}$$

A representação gráfica do exemplo exposto está ilustrada na figura 3.2, a qual mostra que a maior distância, dentro do intervalo de θ de (-3,3), não excede 0,1. Com o auxílio do gráfico, observa-se, entretanto, que esse item era extramente difícil para ambos os grupos e, portanto, um item inapropriado para grupos com proficiência de -3 a 3. Se as duas CCIs fossem comparadas nas faixas de proficiência em que esse item é apropriado, uma diferença considerável de probabilidade de acerto para indivíduos de mesma proficiência seria observada. Para itens com dificuldade apropriada para pelo menos um dos dois grupos de examinados (itens com parâmetro de dificuldade dentro da faixa de proficiência de interesse), não é possível que se tenham diferenças significativas entre os parâmetros do item para os dois grupos sem uma correspondente diferença entre as curvas.

Capítulo 4

Regressão Isotônica

4.1 Introdução

Em algumas situações práticas, considerando informações *a priori*, é esperado que os dados se comportem de forma ordenada, como é o caso da *proficiência* pela *probabilidade de acerto*. Espera-se que quanto maior a proficiência de um respondente, maior a probabilidade de ele acertar um dado item. Modelos com base na Regressão Isotônica consideram essa característica; o termo *isotônica* refere-se ao fato de a variável resposta aumentar com o aumento da variável independente. Para os casos nos quais a variável resposta decresce com o aumento da variável independente, utiliza-se o termo *antitônica*.

Schlemper (2014) apresenta a regressão isotônica como um bom modelo para estimação não paramétrica da Curva Característica do Item. Uma vez que, na construção dessas curvas, elas apresentam, em geral, forma não decrescente, com o aumento da probabilidade de resposta correta conforme cresce a habilidade do examinado.

A fim de construir um método para detecção de DIF, pretende-se estimar as CCIs de cada grupo pela regressão isotônica. Portanto, para o desenvolvimento das próximas etapas, será feita uma breve apresentação da Regressão Isotônica, além da definição do caso 1 de censura intervalar (também chamado de status corrente), retirada do trabalho de Groeneboom e Wellner (2012).

Definição 4.1. Seja $((T_1, C_1), ..., (T_n, C_n))$ uma amostra aleatória de variáveis aleatórias em $\mathbb{R}^2_+$, tal que T_j e C_j são independentes (não-negativas) com funções de distribuição F e G, respectivamente. Observa-se C_j ("instante da observação") e δ_j, sendo

$$\delta_j = \begin{cases} 1, & \text{se } T_j \leq C_j \\ 0, & \text{c.c .} \end{cases}$$

Um conjunto de dados com esse comportamento é o chamado **caso 1 de censura intervalar** (status corrente). Sem perda de generalidade, assume-se que $C_j \leq C_{j+1}$. Para este caso, o estimador não paramétrico de máxima verossimilhança de F é o valor $\hat{F}(C_j)$, $j = 1, ..., n$, que maximiza a expressão

$$\log L(F) = \sum_{j=1}^{n} \{\delta_j \log F(C_j) + (1 - \delta_j)\log(1 - F(C_j))\} \tag{4.1}$$

sob a restrição de que $F(C_j) \leq F(C_{j+1})$, para $C_j \leq C_{j+1}$.

Definição 4.2. Seja X um conjunto finito $\{x_1, ..., x_k\}$ com uma relação de ordem simples $x_1 \prec x_2 \prec ... \prec x_k$, em que uma relação de ordem simples existe se as seguintes propriedades forem satisfeitas:

i. ser reflexiva: $x \prec x$ para todo $x \in X$;

ii. ser transitiva: $x, y, z \in X$, se $x \prec y$ e $y \prec z$ então $x \prec z$;

iii. ser antissimétrica: $x, y \in X$, $x \prec y$ e $y \prec x$ então $x = y$; e

iv. todo e qualquer elemento de X ser comparável: $x, y \in X$, implica $x \prec y$ ou $y \prec x$.

Uma função f em X é isotônica se $x, y \in X$ e $x \prec y$ implica $f(x) \leq f(y)$.

Seja g uma função em X e w uma função positiva em X, uma função isotônica g^* em X é uma *regressão isotônica de g com pesos w que respeita a ordem simples* $x_1 \prec x_2 \prec ... \prec x_k$ se e somente se g^* minimizar a soma (4.2)

$$\sum_{x \in X} [g(x) - f(x)]^2 w(x) \tag{4.2}$$

em que f varia entre todas as funções isotônicas em X. Portanto, g^* é uma solução de mínimos quadrados restritos para a expressão (4.2)

Entendido isso, g^* é chamada simplesmente de regressão isotônica de g.

Teorema 4.1. Se f é isotônica em X e se a imagem de f está em I então,

$$\sum_x \Delta[g(x), f(x)]w(x) \leq \sum_x \Delta[g(x), g^*(x)]w(x) + \sum_x \Delta[g^*(x), f(x)]w(x) \tag{4.3}$$

consequentemente g^* minimiza

$$\sum_x \Delta[g(x), f(x)]w(x) \tag{4.4}$$

entre todas as funções isotônicas f em I e maximiza

$$H(f) = \sum_x \{\Phi[f(x)] + [g(x) - f(x)]\,\varphi[f(x)]\}\,w(x) \tag{4.5}$$

em que $\varphi(y) = \dfrac{d\Phi(y)}{dy}$.

A função que minimiza (maximiza) é única se Φ é estritamente convexa.

A prova do teorema 4.1 encontra-se em Barlow (1972), página 42.

É possível escrever a expressão (4.1) na forma de (4.5) fazendo $g(C_j) = \delta_j$, $w(C_j) = 1$, $\Phi(x) = x\log(x) + (1-x)\log(1-x)$ e $\varphi(x) = \Phi'(x) = \log(x) - \log(1-x)$.

A regressão isotônica g^* também pode ser obtida pela expressão

$$g^*(C_m) = \max_{i \leq m} \min_{k \geq m} \frac{\sum_{i \leq j \leq k} \delta_{(j)}}{k - i + 1}, \tag{4.6}$$

em que $m = 1, ..., n$.

4.2 Interpretação gráfica — a função minorante convexa máxima

A solução da equação (4.6) também pode ser encontrada plotando os pontos $\left(j, \sum_{i \leq j} \delta_{(j)}\right)$ no plano cartesiano e encontrando a função *minorante convexa máxima* (MCM) - do inglês, *greatest convex minorant* - desses pontos no intervalo [1,k]. A função MCM e a construção do respectivo gráfico são apresentadas a seguir.

Assumindo a ordem simples $x_1 \prec x_2 \prec ... \prec x_k$, deve-se gerar um gráfico da soma acumulada da função g ponderada pelos pesos w:

$$G_j = \sum_{i=1}^{j} g(x_i)w(x_i)$$

pela soma acumulada dos pesos w:

$$W_j = \sum_{i=1}^{j} w(x_i),\ j = 1, 2, ..., k.$$

Ou seja, construir um gráfico dos pontos $P_j = (W_j, G_j)$, $j = 1, 2, ..., k(P_0 = (0, 0))$. Esses pontos constituem um *diagrama de soma acumulada* (DSA) - do inglês, *cumulative sum diagram* - da função g com os pesos w. O declive do segmento que une os pontos P_{j-1} a P_j é exatamente $g(x_j)$, $j = 1, 2, ..., k$.

A regressão isotônica de g é dada pelo declive da função MCM do DSA. A função MCM é a curva da maior de todas as funções convexas que estejam abaixo do DSA. O valor da regressão isotônica g^* no ponto x_j é simplesmente a inclinação a esquerda da MCM no ponto P_j^* com abscissa igual a $\sum_{i=1}^{j} w(x_i)$.

A tabela 4.1 e o gráfico da figura 4.1 apresentam dados de um exemplo ilustrativo retirado do livro de Barlow (1972), página 11.

Tabela 4.1 - Exemplo de DSA e MCM apresentado em Barlow (1972)

j	$w(x_j)$	W_j	$g(x_j)$	G_j	G_j^*	g_j^*
1	1	1	-2	-2	-2	-2
2	2	3	5/2	3	-8/5	1/5
3	3	6	-4/3	-1	-1	1/5
4	2	8	1	1	1	1

Na tabela 4.1,

$$W_j = \sum_{i=1}^{j} w(x_i),\quad G_j = \sum_{i=1}^{j} g(x_i)w(x_i),\quad G_j^* = \sum_{i=1}^{j} g^*(x_i)w(x_i)\quad \text{e } j = 1, 2, 3, 4.$$

A inclinação em P_j do DSA é dada por

$$\frac{G_j - G_{j-1}}{W_j - W_{j-1}} = g(x_j); \tag{4.7}$$

e a inclinação em P_j^* da função MCM é dada por

$$\frac{G_j^* - G_{j-1}^*}{W_j - W_{j-1}} = g^*(x_j),\ j = 1, 2, 3, 4. \tag{4.8}$$

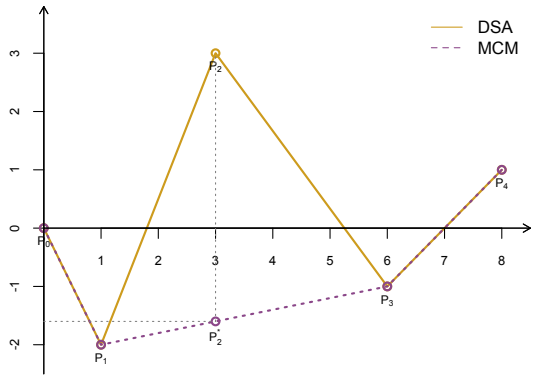


Figura 4.1: Gráficos DSA e MCM para ilustrar o exemplo apresentado em Barlow (1972)

Como a regressão isotônica é a inclinação ($g^*(x_j)$) do ponto na função MCM, pode-se calculá-la para todos os P_j^*. Da equação (4.8), tem-se que

$$g^*(x_1) = -2,\ g^*(x_2) = g^*(x_3) = 1/5 \text{ e } g^*(x_4) = 1.$$

A partir do diagrama da figura 4.1, com as inclinações ($g^*(x_j)$) calculadas, gera-se o gráfico apresentado na figura 4.2.

Algumas propriedades importantes da MCM:

i. o DSA e a MCM coincidem no P_k, isto é:

$$G_k^* = G_k$$

ii. se, para algum i da MCM, P_{i-1}^* está posicionado exatamente abaixo do DSA no ponto P_{i-1}, então o declive da MCM para esquerda e para direita de P_{i-1}^* será o mesmo. Ou seja,

$$G_{i-1}^* < G_{i-1} \Rightarrow g_i^* - g_{i-1}^* = 0,\ j = 1, 2, ..., k.$$

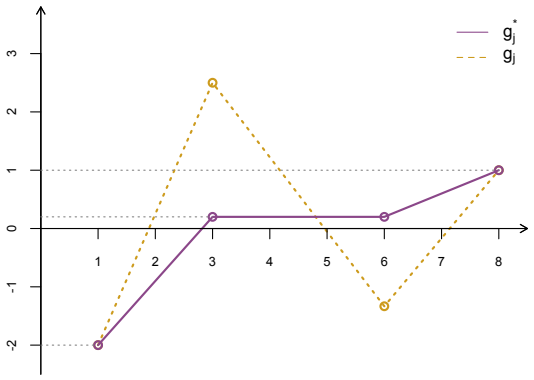


Figura 4.2: Regressão Isotônica para o conjunto de dados apresentados em Barlow (1972).

Capítulo 5

O método

5.1 *Introdução*

A regressão isotônica será utilizada para a obtenção do estimador não paramétrico de máxima verossimilhança (ENPMV) das CCIs. Em particular, para itens nos quais há suspeita de DIF, obtém-se o ENPMV para a CCI separadamente para os dois grupos que supostamente dão origem ao DIF.

Sendo $\hat{\theta}_j$, $j = 1, ..., n$, estimativas de θ_j (obtidas via TRI ou dados pelo escore padronizado), $g_j = g(\hat{\theta}_j) = u_{ij}$, $w_j = w(\hat{\theta}_j) = 1$, a regressão isotônica de g_j com pesos w_j, $j = 1, ..., n$, fornece o ENPMV de $P_i(\theta_j)$. Isso é decorrência da aplicação do Teorema 1.10 de Barlow (1972), apresentado neste trabalho pelo teorema 4.1.

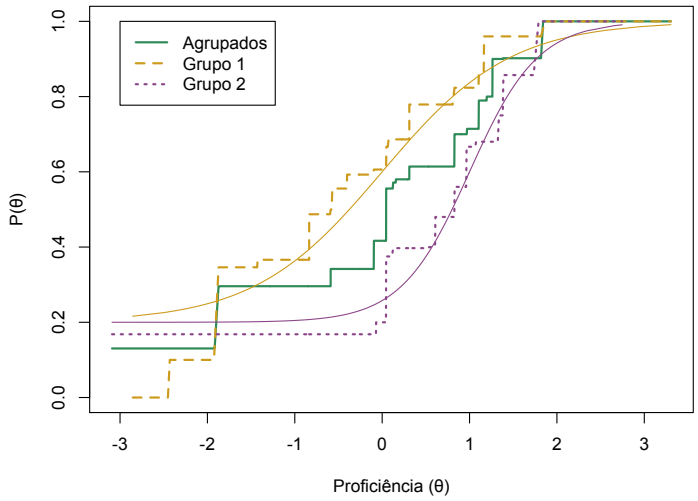


Figura 5.1: Exemplo de CCIs estimadas de dois grupos para um item

A figura 5.1 mostra as CCIs estimadas via regressão isotônica e suas curvas reais. Na cor verde, é a estimativa da CCI para todos os respondentes, sem distinção do grupo ao qual pertencem. Separadamente, para cada grupo, tem-se a CCI real. Em pontilhado, tem-se o ENPMV calculado para o grupo 1 e grupo 2, separadamente.

5.2 Uso de regressão isotônica suavizada na detecção de Funcionamento Diferencial do Item

Para testar a hipótese de existência de DIF em um item, considera-se a hipótese nula H_0 "o item não apresenta DIF"e a hipótese alternativa H_1 "o item apresenta DIF". Sob H_0, obtém-se uma estimativa suavizada do ENPMV via núcleo estimador (Kernel), geram-se respostas artificiais para o item e para cada indivíduo como função de sua proficiência estimada, obtém-se estimativas não-paramétricas da CCI para cada grupo separadamente e calcula-se a distância S entre as curvas.

Utilizou-se três métodos para o cálculo da distância entre as curvas, S_1, S_2 e S_3.

$$S_1 = d_{L_1}(CCI_0, CCI_1) = \int_{-\infty}^{\infty} |F_0(t) - F_1(t)|\, \phi(t)dt \tag{5.1}$$

$$S_2 = d_{L_2}(CCI_0, CCI_1) = \int_{-\infty}^{\infty} (F_0(t) - F_1(t))^2\, \phi(t)dt \tag{5.2}$$

$$S_3 = d_{sup}(CCI_0, CCI_1) = \sup_{t\in[-\infty,\infty]} |F_0(t) - F_1(t)| \tag{5.3}$$

em que F_0 é a estimativa da CCI do grupo 0, F_1 é a estimativa da CCI do grupo 1 e $\phi(t)$ é a densidade da Distribuição Normal (0,1).

Essas são as as estatísticas do teste. Esse procedimento será repetido m vezes, obtendo-se assim m valores para a estatística do teste sob H_0 através de m amostras obtidas via permutação aleatória dos indivíduos nos dois grupos, mantidos os tamanhos originais dos grupos. O valor dessa mesma estatística calculada para os dados reais será comparada com os valores gerados e o p-valor do teste será dado por

$$\text{p-valor} = \frac{\#\{i : S_i > S_{\text{OBS}}, i = 1, ..., n\}}{m}, \tag{5.4}$$

em que S_i é o valor da estatística S para a i-ésima amostra gerada, S_{OBS} é o valor da estatística para os dados reais, calculado suavizando-se o ENPMV da CCI separadamente para cada grupo.

Utilizando as amostras geradas por permutações, obtém-se um p-valor local para cada valor da proficiência, obtido calculando-se, para cada valor de $\hat{\theta}$, a proporção de amostras permutadas que apresentam distância entre as CCIs estimadas não parametricamente para os dois grupos maior do que a distância observada para os dados originais. Com isso, é possível identificar intervalos de valores da proficiência onde a diferença entre as CCIs dos dois grupos é significativa. A figura 5.2 apresenta um exemplo de um item que apresenta DIF e para $\hat{\theta}$ entre -1,9 e 2,8 tem-se p-valor $< 0{,}05$, indicando um distanciamento significativo entre a CCIs dos dois grupos nesse intervalo de $\hat{\theta}$.

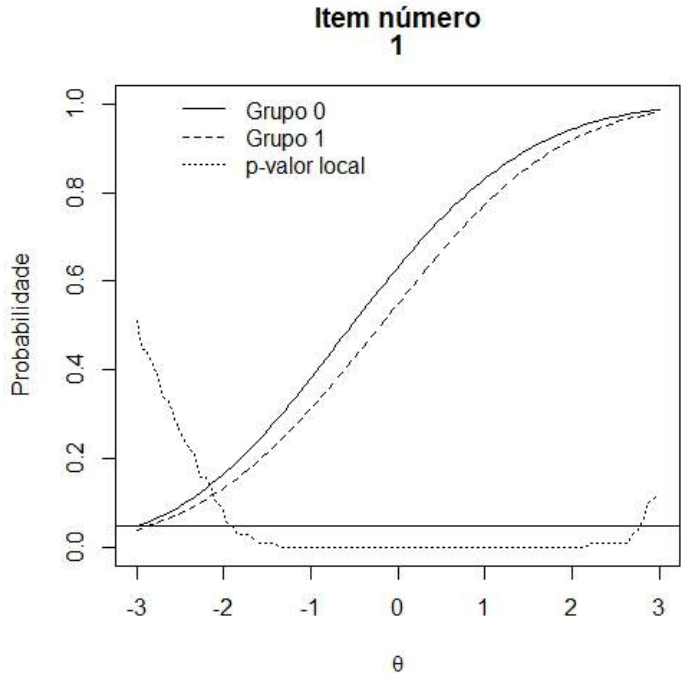


Figura 5.2: Proporção de Itens com DIF Uniforme Identificados

Os p-valores encontrados a partir dos métodos descritos nessa sessão foram comparados com os obtidos a partir do método para detecção de DIF por Mantel-Haenszel e pela Regressão Logística.

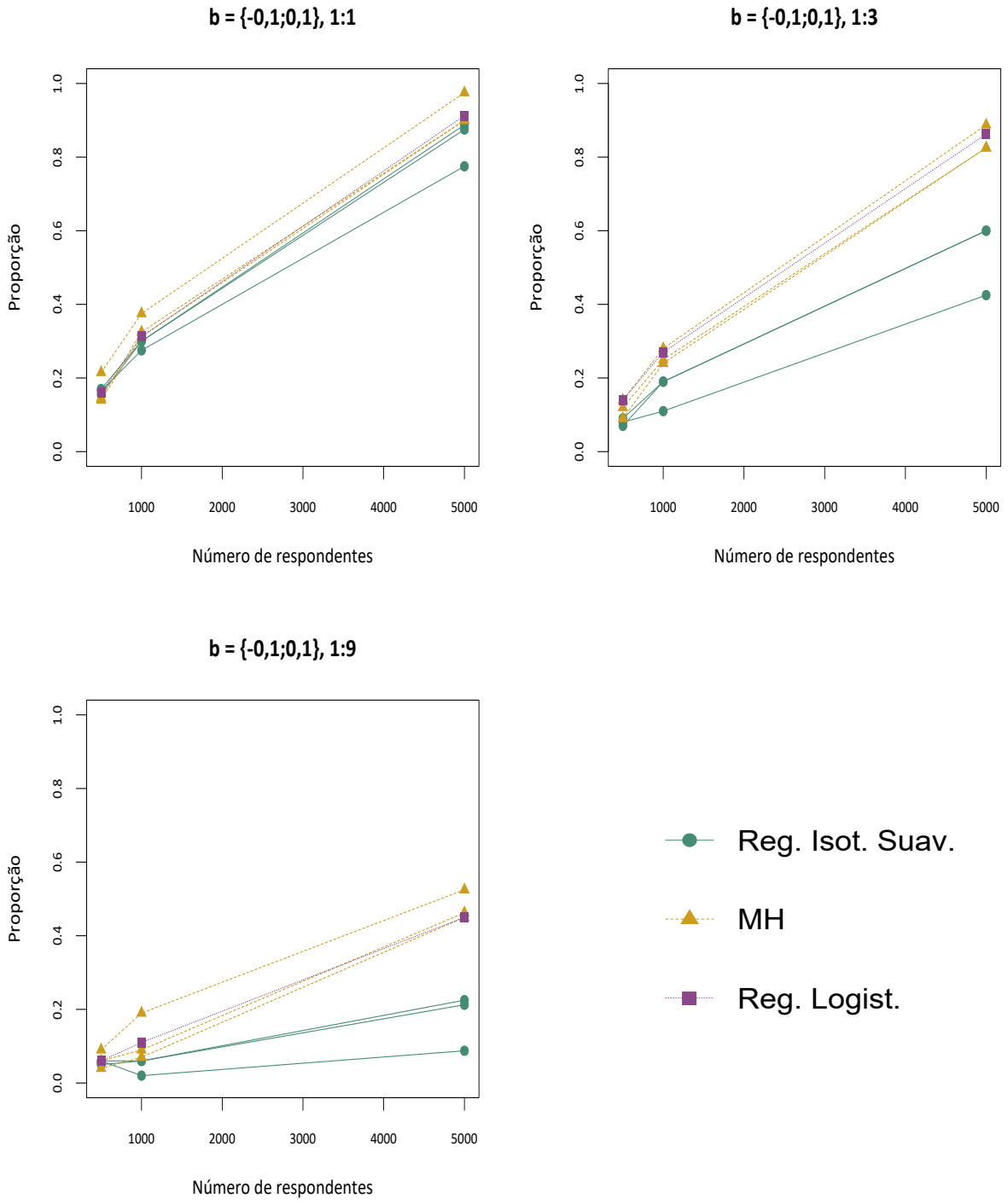


Figura 5.3: Proporção de Itens com DIF Uniforme Identificados para b={-0,1;0,1}.

5.3 Simulação

A comparação entre os métodos foi feita por simulações com as seguintes variáveis independentes:

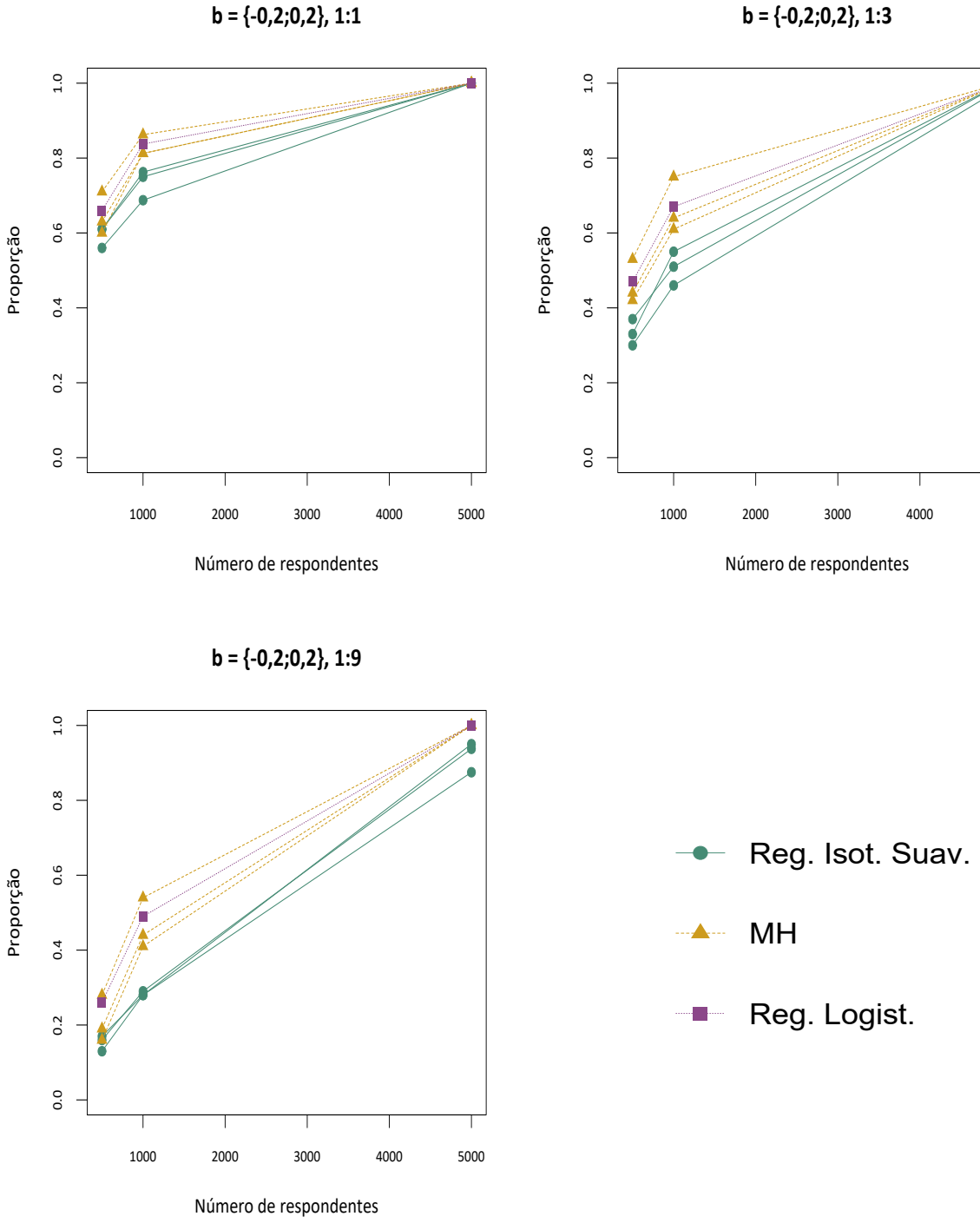


Figura 5.4: Proporção de Itens com DIF Uniforme Identificados para b={-0,2;0,2}.

Tamanho da amostra: 500, 1.000, 5.000, 10.000, 100.000. Esse é o número de respondentes.

Razão do tamanho do grupo: 1, 3, 9. Uma razão de 3 significa que são 3 respondentes no grupo 1 para cada respondente no grupo 0.

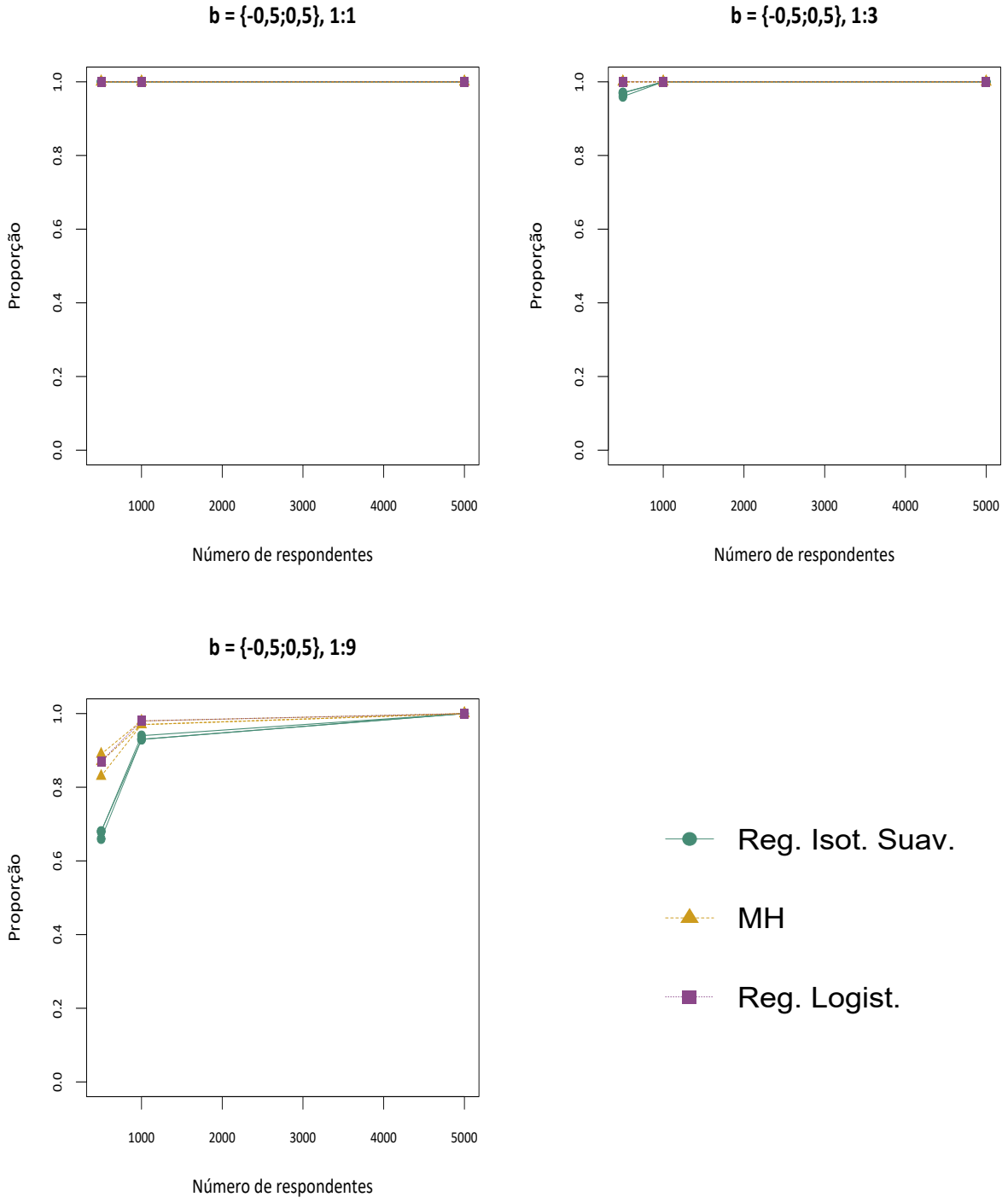


Figura 5.5: Proporção de Itens com DIF Uniforme Identificados b={-0,5;0,5}.

Tamanho do teste: 30. Esse é o número de itens.

Parâmetros do item com DIF: b={-0,1;0,1}, b={-0,2;0,2}, b={-0,5;0,5}, b={-1,0;1,0}, a={1,0;1,5}, a={0,5;1,0}, a={0,5;1,5}. Quando o item tem DIF uniforme, ou seja, a diferença está no parâmetro b, o parâmetro a é $1,5$ para os dois grupos. Quando o item

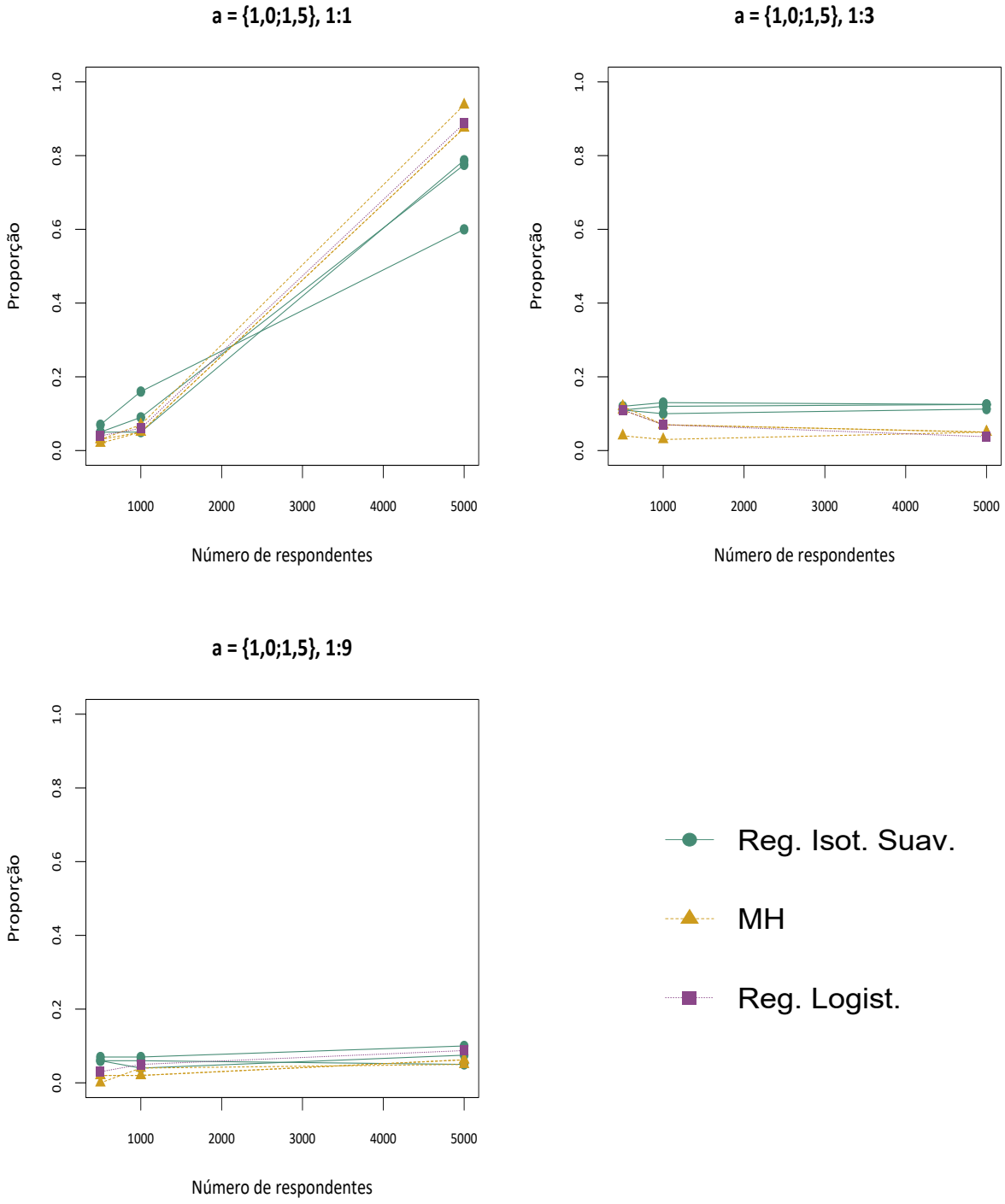


Figura 5.6: Proporção de Itens com DIF Uniforme Identificados.

possui DIF não-uniforme, ou seja, a diferença está no parâmetro a, o parâmetro b é 0 para os dois grupos analisados.

Cada conjunto de dados foi gerado de modo que o primeiro item sempre apresentasse DIF, com seus parâmetros definidos segundo sua configuração e os outros 29 itens tive-

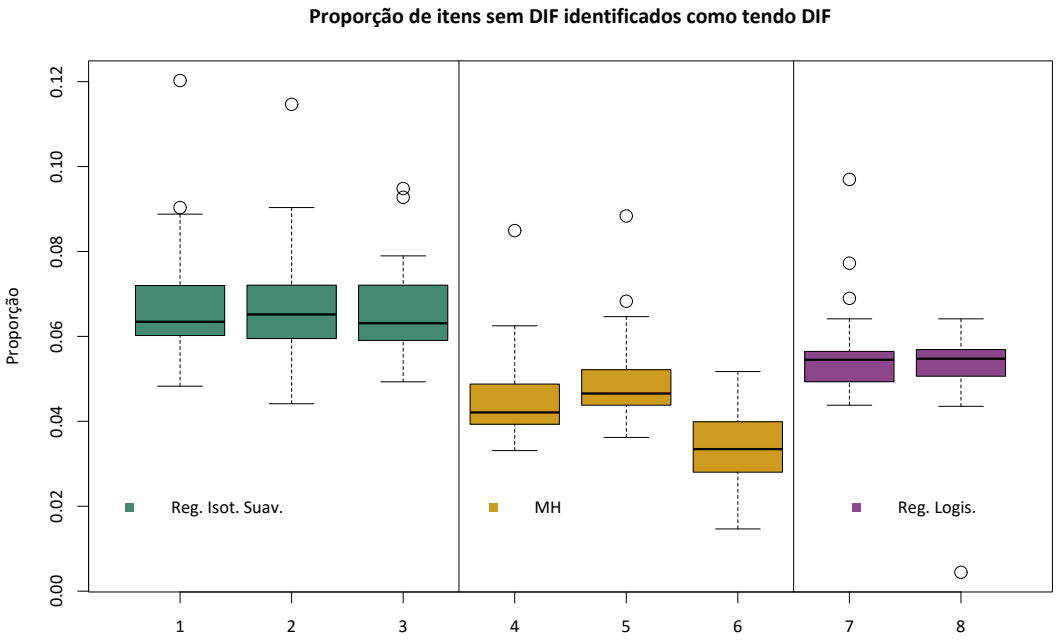


Figura 5.7: Proporção de Itens com falso DIF

ram seus parâmetros gerados aleatoriamente. Os parâmetros a foram gerados por uma distribuição Lognormal média 0 e desvio padrão 0,5 que gerou valores entre 0,3 e 2,5. Os parâmetros b foram gerados por uma distribuição Normal média 0 e desvio padrão 1. E os parâmetros c foram gerados aleatoriamente por uma distribuição Beta (80, 320) que gerou valores entre entre 0,15 e 0,25.

As comparações foram feitas gerando-se 200 amostras para cada configuração e calculada a proporção de vezes que o método detectou DIF no item 1 e a proporção de vezes que detectou DIF nos outros itens, ou seja, detectou um falso DIF. O p-valor utilizado na análise é calculado a partir da comparação da distância entre as curvas dos grupos dos dados reais e a dos dados alocados aleatoriamente entre os grupos, conforme a equação 5.4. Isso foi repetido 100 vezes para cada amostra, m=100.

A eficiência do método descrito na seção 5.1.1 foi comparada com os métodos de Mantel-Haenzel e Regressão Logística para detecção de DIF. Os resultados dessas comparações estão apresentados nas figuras 5.3, 5.4, 5.5, 5.6 e 5.7. As figuras 5.3, 5.4 e 5.5 mostram que tanto a diferença entre os bs estimados para cada grupo, a proporção de respondentes dentro de cada grupo e o tamanho da amostra influenciam na proporção de DIF uniformes

detectados. Pelos gráficos, percebe-se um poder maior de detectar o item que apresenta DIF para os métodos de Mantel-Haenzel e Regressão Logística em comparação com o método desenvolvido neste trabalho, principalmente para os casos de DIFs pequenos, $b = \{-0{,}1;\ 0{,}1\}$, amostras pequenas, menos de 5.000 respondentes ou proporção de respondentes no grupo 0 menor que no grupo 1.

Quando o parâmetro b estimado é de -0,1 para um grupo e 0,1 para o outro grupo (figura 5.3), apenas quando a amostra é de pelo menos 5.000 respondentes e quando a proporção de respondentes entre os grupos é de 1 para 1 que a proporção de itens com DIF detectados é relativamente alta para os três métodos. Quando o parâmetro b estimado é de -0,2 para um grupo e 0,2 para o outro grupo (figura 5.4), essa proporção de itens com DIF detectados aumenta consideravelmente, podendo-se observar altas proporções mesmo em razões de 1 para 3 e até mesmo de 1 para 9 respondentes nos grupos. Quando o parâmetro b estimado é de -0,5 para um grupo e 0,5 para o outro grupo (figura 5.5), as proporções de DIF detectado são próximas de 100% para todos os casos, exceto para a amostra de 1000 respondentes quando a proporção é de 1 para 9.

A figura 5.6 apresenta os casos de DIF não-uniforme, quando a diferença entre os parâmetros estimados para cada grupo está no parâmetro a, para as configurações analisadas, apenas a proporção de respondentes entre os grupos é de 1 para 1 e o número de respondentes é de pelo menos 5.000 é que a proporção de detecção de DIF é relativamente alta. Mesmo assim é consideravelmente mais alta quando comparada a detecção de DIF uniforme.

O boxplot apresentado na figura 5.7 compara a proporção de falsos DIFs detectados por cada método, ou seja, a proporção de itens que foram acusados com presença de DIF mas que eram itens livres de DIF na geração dos parâmetros. As três primeiras figuras correspondem aos três métodos desenvolvidos neste trabalho, com base nos cálculos explicitados em (5.1), (5.2) e (5.3), respectivamente. As três figuras seguintes, geradas com resultados do método para detecção de DIF de Mantel-Haenzel correspondem ao teste clássico, teste exato e teste exato condicional unilateral, respectivamente. As duas últimas figuras, correspondem ao teste da Regressão Logística, uniforme e não-uniforme, respectivamente. A proporção de falto DIF não extrapola os 8% para a grande maioria dos dados. O método de Mantel-Haenzel mostrou-se o melhor, seguido da Regressão Logística.

5.4 Impacto do DIF na estimação da proficiência

Além de investigar a presença de DIF no item, observou-se o impacto da existência de itens com DIF na estimação da proficiência. O impacto foi obtido ordenando os $\hat{\theta}$ (proficiências estimadas) dentro de cada grupo e obtendo as funções empíricas da proficiência para cada grupo. Por serem amostras independentes, utilizou-se o teste de Kolmogorov-Smirnov para testar se as funções de distribuição eram iguais.

A tabela 5.1 apresenta os resultados para cada configuração. Para DIF uniforme, mesmo com amostras de 100 mil respondentes, a proporção de amostras com p-valor menor que 0,05 é pequena para DIF de baixa magnitude, $b = \{$-0,1;0,1$\}$ e $b = \{$-0,2;0,2$\}$. A partir do caso com $b = \{$-0,5;0,5$\}$, percebe-se efeito significativo do DIF na distribuição da proficiência em quase todas as amostras. Para o caso de DIF não-uniforme, a taxa de detecção (p-valor $<$ 0,05) é razoavelmente alta apenas para o caso em que $a = \{$0,5;1,5$\}$, com mil respondentes na amostra. No entanto, a taxa de detecção do efeito do DIF (uniforme e não-uniforme) na distribuição das proficiências estimadas para os dois grupos decresce quando a proporção de respondentes em um dos grupos decresce.

O gráfico da figura 5.8 é a representação do impacto na estimação da proficiência para um teste de 30 itens em que, foram 1000 respondentes, a proporção é de 1:9 respondentes entre os grupos e, para um item, o parâmetro b para o grupo 0 foi de -1,0 e para o grupo 1 foi de 1,0, para os outros itens, os parâmetros dos itens para os dois grupos é o mesmo. O teste K-S verifica se essa distância é significativa, ou seja, se o item com DIF impactou na estimação da proficiência dos grupos comparados. Quanto mais distantes as curvas para cada grupo, maior será o impacto do DIF na estimação da proficiência.

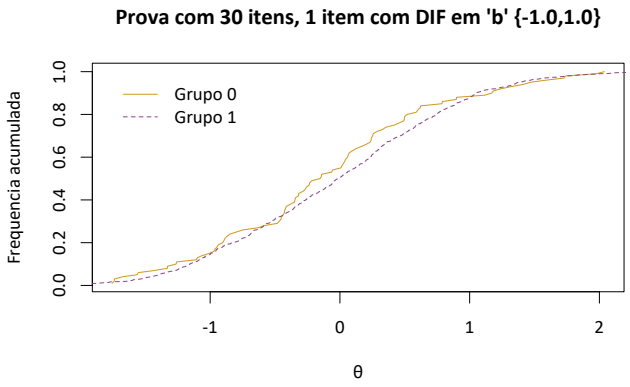


Figura 5.8: Distribuição do $\hat{\theta}$ para um teste com um item com DIF

Tabela 5.1 - Proporção de estimação de proficiências impactadas pela presença de itens com DIF no teste

Itens	DIF	a	b	Número de Respondentes	Proporção entre os grupos	p-valor médio	proporção de p-valores < 0,05
30	1	$\{1,5;1,5\}$	$\{-0,1;0,1\}$	500	1:1	0,4891	0,10
30	1	$\{1,5;1,5\}$	$\{-0,1;0,1\}$	1000	1:1	0,4917	0,04
30	1	$\{1,5;1,5\}$	$\{-0,1;0,1\}$	5000	1:1	0,4664	0,06
30	1	$\{1,5;1,5\}$	$\{-0,1;0,1\}$	10000	1:1	0,4453	0,06
30	1	$\{1,5;1,5\}$	$\{-0,1;0,1\}$	100000	1:1	0,1209	0,46
30	1	$\{1,5;1,5\}$	$\{-0,1;0,1\}$	500	1:3	0,5708	0,03
30	1	$\{1,5;1,5\}$	$\{-0,1;0,1\}$	1000	1:3	0,5552	0,04
30	1	$\{1,5;1,5\}$	$\{-0,1;0,1\}$	5000	1:3	0,4604	0,08
30	1	$\{1,5;1,5\}$	$\{-0,1;0,1\}$	100000	1:3	0,2362	0,28
30	1	$\{1,5;1,5\}$	$\{-0,1;0,1\}$	500	1:9	0,5773	0,06
30	1	$\{1,5;1,5\}$	$\{-0,1;0,1\}$	1000	1:9	0,5555	0,03
30	1	$\{1,5;1,5\}$	$\{-0,1;0,1\}$	5000	1:9	0,4795	0,06
30	1	$\{1,5;1,5\}$	$\{-0,1;0,1\}$	100000	1:9	0,3483	0,14
30	1	$\{1,5;1,5\}$	$\{-0,2;0,2\}$	500	1:1	0,4971	0,08
30	1	$\{1,5;1,5\}$	$\{-0,2;0,2\}$	1000	1:1	0,4828	0,08
30	1	$\{1,5;1,5\}$	$\{-0,2;0,2\}$	5000	1:1	0,4231	0,10
30	1	$\{1,5;1,5\}$	$\{-0,2;0,2\}$	10000	1:1	0,3344	0,18
30	1	$\{1,5;1,5\}$	$\{-0,2;0,2\}$	100000	1:1	0,0224	0,23
30	1	$\{1,5;1,5\}$	$\{-0,2;0,2\}$	500	1:3	0,5543	0,03
30	1	$\{1,5;1,5\}$	$\{-0,2;0,2\}$	1000	1:3	0,5384	0,03
30	1	$\{1,5;1,5\}$	$\{-0,2;0,2\}$	5000	1:3	0,4263	0,11
30	1	$\{1,5;1,5\}$	$\{-0,2;0,2\}$	100000	1:3	0,0341	0,18
30	1	$\{1,5;1,5\}$	$\{-0,2;0,2\}$	500	1:9	0,5605	0,06
30	1	$\{1,5;1,5\}$	$\{-0,2;0,2\}$	1000	1:9	0,5331	0,05
30	1	$\{1,5;1,5\}$	$\{-0,2;0,2\}$	5000	1:9	0,4579	0,07
30	1	$\{1,5;1,5\}$	$\{-0,2;0,2\}$	100000	1:9	0,2200	0,10
30	1	$\{1,5;1,5\}$	$\{-0,5;0,5\}$	500	1:1	0,4392	0,12
30	1	$\{1,5;1,5\}$	$\{-0,5;0,5\}$	1000	1:1	0,3953	0,11
30	1	$\{1,5;1,5\}$	$\{-0,5;0,5\}$	5000	1:1	0,1724	0,47
30	1	$\{1,5;1,5\}$	$\{-0,5;0,5\}$	10000	1:1	0,0964	0,70
30	1	$\{1,5;1,5\}$	$\{-0,5;0,5\}$	100000	1:1	5,32e-12	0,99
30	1	$\{1,5;1,5\}$	$\{-0,5;0,5\}$	500	1:3	0,5239	0,04
30	1	$\{1,5;1,5\}$	$\{-0,5;0,5\}$	1000	1:3	0,4506	0,08
30	1	$\{1,5;1,5\}$	$\{-0,5;0,5\}$	5000	1:3	0,2496	0,36
30	1	$\{1,5;1,5\}$	$\{-0,5;0,5\}$	100000	1:3	8,26e-08	0,95
30	1	$\{1,5;1,5\}$	$\{-0,5;0,5\}$	500	1:9	0,5396	0,06
30	1	$\{1,5;1,5\}$	$\{-0,5;0,5\}$	1000	1:9	0,4960	0,06
30	1	$\{1,5;1,5\}$	$\{-0,5;0,5\}$	5000	1:9	0,3803	0,19
30	1	$\{1,5;1,5\}$	$\{-0,5;0,5\}$	100000	1:9	0,0028	0,98
30	1	$\{1,5;1,5\}$	$\{-1,0;1,0\}$	500	1:1	0,3518	0,22
30	1	$\{1,5;1,5\}$	$\{-1,0;1,0\}$	1000	1:1	0,2192	0,37
30	1	$\{1,5;1,5\}$	$\{-1,0;1,0\}$	5000	1:1	0,0114	0,93
30	1	$\{1,5;1,5\}$	$\{-1,0;1,0\}$	10000	1:1	0,0012	1,00
30	1	$\{1,5;1,5\}$	$\{-1,0;1,0\}$	100000	1:1	0,0000	1,00

Itens	DIF	a	b	Número de Respondentes	Proporção entre os grupos	p-valor médio	proporção de p-valores < 0,05
30	1	$\{1,5;1,5\}$	$\{-1,0;1,0\}$	500	1:3	0,4396	0,06
30	1	$\{1,5;1,5\}$	$\{-1,0;1,0\}$	1000	1:3	0,3208	0,19
30	1	$\{1,5;1,5\}$	$\{-1,0;1,0\}$	5000	1:3	0,0646	0,68
30	1	$\{1,5;1,5\}$	$\{-1,0;1,0\}$	100000	1:3	0,0000	1,00
30	1	$\{1,5;1,5\}$	$\{-1,0;1,0\}$	500	1:9	0,5069	0,05
30	1	$\{1,5;1,5\}$	$\{-1,0;1,0\}$	1000	1:9	0,4368	0,11
30	1	$\{1,5;1,5\}$	$\{-1,0;1,0\}$	5000	1:9	0,2482	0,33
30	1	$\{1,5;1,5\}$	$\{-1,0;1,0\}$	100000	1:9	0,0000	1,00
30	1	$\{1,0;1,5\}$	$\{0,0;0,0\}$	500	1:1	0,47959	0,08
30	1	$\{1,0;1,5\}$	$\{0,0;0,0\}$	1000	1:1	0,50242	0,05
30	1	$\{1,0;1,5\}$	$\{0,0;0,0\}$	5000	1:1	0,46112	0,06
30	1	$\{1,0;1,5\}$	$\{0,0;0,0\}$	10000	1:1	0,43786	0,11
30	1	$\{1,0;1,5\}$	$\{0,0;0,0\}$	100000	1:1	0,25995	0,16
30	1	$\{1,0;1,5\}$	$\{0,0;0,0\}$	500	1:3	0,55994	0,03
30	1	$\{1,0;1,5\}$	$\{0,0;0,0\}$	1000	1:3	0,55711	0,03
30	1	$\{1,0;1,5\}$	$\{0,0;0,0\}$	5000	1:3	0,45485	0,08
30	1	$\{1,0;1,5\}$	$\{0,0;0,0\}$	100000	1:3	0,27377	0,24
30	1	$\{1,0;1,5\}$	$\{0,0;0,0\}$	500	1:9	0,56082	0,04
30	1	$\{1,0;1,5\}$	$\{0,0;0,0\}$	1000	1:9	0,53910	0,03
30	1	$\{1,0;1,5\}$	$\{0,0;0,0\}$	5000	1:9	0,46735	0,04
30	1	$\{1,0;1,5\}$	$\{0,0;0,0\}$	100000	1:9	0,44251	0,10
30	1	$\{0,5;1,0\}$	$\{0,0;0,0\}$	500	1:1	0,49010	0,09
30	1	$\{0,5;1,0\}$	$\{0,0;0,0\}$	1000	1:1	0,52116	0,04
30	1	$\{0,5;1,0\}$	$\{0,0;0,0\}$	5000	1:1	0,47351	0,05
30	1	$\{0,5;1,0\}$	$\{0,0;0,0\}$	100000	1:1	0,32191	0,16
30	1	$\{0,5;1,0\}$	$\{0,0;0,0\}$	500	1:3	0,56932	0,04
30	1	$\{0,5;1,0\}$	$\{0,0;0,0\}$	1000	1:3	0,55388	0,01
30	1	$\{0,5;1,0\}$	$\{0,0;0,0\}$	5000	1:3	0,46002	0,10
30	1	$\{0,5;1,0\}$	$\{0,0;0,0\}$	100000	1:3	0,32624	0,20
30	1	$\{0,5;1,0\}$	$\{0,0;0,0\}$	500	1:9	0,55215	0,05
30	1	$\{0,5;1,0\}$	$\{0,0;0,0\}$	1000	1:9	0,54645	0,04
30	1	$\{0,5;1,0\}$	$\{0,0;0,0\}$	5000	1:9	0,47148	0,04
30	1	$\{0,5;1,0\}$	$\{0,0;0,0\}$	100000	1:9	0,46942	0,04
30	1	$\{0,5;1,5\}$	$\{0,0;0,0\}$	500	1:1	0,47817	0,10
30	1	$\{0,5;1,5\}$	$\{0,0;0,0\}$	1000	1:1	0,49728	0,03
30	1	$\{0,5;1,5\}$	$\{0,0;0,0\}$	5000	1:1	0,41245	0,07
30	1	$\{0,5;1,5\}$	$\{0,0;0,0\}$	100000	1:1	0,04580	0,68
30	1	$\{0,5;1,5\}$	$\{0,0;0,0\}$	500	1:3	0,57102	0,05
30	1	$\{0,5;1,5\}$	$\{0,0;0,0\}$	1000	1:3	0,54251	0,01
30	1	$\{0,5;1,5\}$	$\{0,0;0,0\}$	5000	1:3	0,41823	0,09
30	1	$\{0,5;1,5\}$	$\{0,0;0,0\}$	100000	1:3	0,03436	0,78
30	1	$\{0,5;1,5\}$	$\{0,0;0,0\}$	500	1:9	0,56082	0,04
30	1	$\{0,5;1,5\}$	$\{0,0;0,0\}$	1000	1:9	0,53910	0,03
30	1	$\{0,5;1,5\}$	$\{0,0;0,0\}$	5000	1:9	0,46735	0,04
30	1	$\{0,5;1,5\}$	$\{0,0;0,0\}$	100000	1:9	0,13209	0,32

Capítulo 6

Conclusão

Os resultados das simulações mostraram que o teste proposto tende a detectar a existência de DIF com menos frequência do que os testes de Mantel-Haenszel e via regressão logística. Além disso, o teste proposto tende a indicar a falsa existência de DIF com maior frequência do que os testes de Mantel-Haenszel e via regressão logística. Inclusivamente, é importante considerar, além da presença de DIF em um item, o impacto desse evento na estimação da proficiência. Mesmo que detecte-se DIF, o teste de Kolmogorov-Smirnov balizará a tomada de decisão, pois mesmo um teste com um item que apresente DIF, pode ser que a estimação da proficiência dos respondentes não seja afetada.

6.1 Sugestões de trabalhos futuros

Para enriquecer a análise e a comparação entre os métodos, sugere-se ampliar o estudo de simulação para outros valores do número de respondentes e de itens, além da quantidade do número de itens com DIF em um teste. Também estudar o teste de permutação com as estatísticas que envolvem a distância entre as estimativas das CCIs para cada grupo utilizando os modelos paramétricos da TRI. Outra possibilidade seria utilizar outras estatísticas de distância entre as distribuições empíricas das proficiências nos dois grupos como alternativa ao teste de Kolmogorov-Smironov.

Referências Bibliográficas

Aliste Á. M. F., Funcionamiento diferencial de los ítems. In Psicometría , 1996, p. 371

Andrade D. F., Tavares H. R., da Cunha Valle R., Teoria da Resposta ao Item: conceitos e aplicações, ABE, Sao Paulo, 2000

Angoff W. H., Perspectives on differential item functioning methodology., 1993

Barlow R. E., , 1972 Technical report Statistical inference under order restrictions; the theory and application of isotonic regression

Birnbaum A., Some latent train models and their use in inferring an examinee's ability, Statistical theories of mental test scores, 1968, pp 395–479

Camilli G., Shepard L. A., Methods for identifying biased test items. vol. 4, Sage, 1994

Cuevas M., Cervantes V. H., Differential item functioning detection with logistic regression, Math Sci Hum, 2012, vol. 199, p. 45

ESPITIA A. C. S., Efecto de la Razón de Tamaños sobre la Detección del Funcionamiento Diferencial del ítem mediante Regresión Logística, 2009

Groeneboom P., Wellner J. A., Information bounds and nonparametric maximum likelihood estimation. vol. 19, Birkhäuser, 2012

Hambleton R., Swaminathan H., Rogers H., Fundamentals of Item Response Theory. Measurement Methods for the So, SAGE Publications, 1991

Herrera A.-N., Gómez Benito J., Hidalgo-Montesinos M. D., Detección de sesgo en los ítems mediante análisis de tablas de contingencia, Avances en Medición, 2005, vol. 3, p. 29

Herrera Rojas A. N., Efecto del tamaño de muestra y la razón de tamaños de muestra en la detección de funcionamiento diferencial de los ítems. Universitat de Barcelona, 2006

Holland P., Wainer H., Differential Item Functioning. Taylor & Francis, 2012

Holland P. W., Glymour C., Granger C., Statistics and causal inference, ETS Research Report Series, 1985, vol. 1985

Holland P. W., Thayer D. T., Differential item performance and the Mantel-Haenszel procedure, Test validity, 1988, pp 129–145

Kirk R. E., Practical significance: A concept whose time has come, Educational and psychological measurement, 1996, vol. 56, p. 746

Linn R. L., Levine M. V., Hastings C. N., Wardrop J. L., Item Bias in a Test of Reading Comprehension, Applied Psychological Measurement, 1981, vol. 5, p. 159

Lord F., Applications of Item Response Theory to Practical Testing Problems. Erlbaum Associates, 1980

Lord F. M., The Relation of Test Score to the Trait Underlying the Test, ETS Research Bulletin Series, 1952, vol. 1952, p. 517

Mantel N., Haenszel W., Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease, JNCI: Journal of the National Cancer Institute, 1959, vol. 22, p. 719

Osterlind S. J., Everson H. T., Differential item functioning. vol. 161, Sage Publications, 2009

Rasch G., Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests., 1960

Schlemper B. N., Testes de adequabilidade de ajuste em teoria de resposta ao item, 2014

THISSEN D., Use of item response theory in the study of group differences in trace lines, Test Validity, 1988

Zumbo B. D., A handbook on the theory and methods of differential item functioning (DIF), Ottawa: National Defense Headquarters, 1999