Universidade de Brasília

Instituto de Ciências Exatas

Departamento de Estatística

Não-Respostas Intencionais na Teoria da Resposta ao Item

Helen Indianara Seabra Gomes

Orientador: Raul Yukihiro Matsushita

Brasília

2018

Não-Respostas Intencionais na Teoria da Resposta ao Item

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidadede de Brasília como parte dos requisitos necessários à obtenção do título de Mestre em Estatística.

Comissão Julgadora:

- Prof. Dr. Raul Yukihiro Matsushita (orientador) - EST/UnB
- Profª. Drª. Cibele Queiroz da Silva (coorientadora) - EST/UnB
- Prof. Dr. Heliton Ribeiro Tavares (membro externo) - EST/UFPA
- Prof. Dr. Antonio Eduardo Gomes (membro interno) - EST/UnB
- Prof. Dr. Eduardo Yoshio Nakano (suplente) - EST/UnB

Aos meus pais!

Agradecimentos

Esta dissertação é o que é hoje por conta dos muitos indivíduos que me apoiaram durante esta jornada. Gostaria de aproveitar esta oportunidade para expressar meus agradecimentos àqueles que me ajudaram de forma direta e indiretamente em vários aspectos na realização deste trabalho. Em especial, eu gostaria de agradecer:

A Deus, pois foi Ele quem colocou pessoas tão especiais ao meu lado, sem as quais certamente eu não teria ido tão longe.

Ao meu professor e orientador, Prof. Dr. Raul Matsushita, por seus ensinamentos, correções, paciência e dedicação na minha dissertação. Sem a sua orientação, sugestões e conselhos, este trabalho certamente não teria sido concluído. Sou muito grata por ter enriquecido meu conhecimento com um exemplo de profissional durante este período. Obrigada por acreditar em meu potencial e que eu seria capaz.

Aos meus pais, Maricelis e Valdemar, esta dissertação não teria sido possível sem o amor, carinho, apoio, incentivo e confiança que depositam em mim para a realização de meus sonhos. Obrigada pelo amor incondicional e por compreenderem a minha ausência nos últimos anos por conta da vida acadêmica. Vocês são meus exemplos de pessoa na vida.

Aos meus irmãos, Suzi, Herberth e Evelyn, e meus sobrinhos amados Victor, Jordan, Adrian, Olívia e Emanuel que são a minha fonte de alegria, alívio e de certa forma, um indicador bom e confiável de incentivo, sempre fazendo meus dias melhores seja perto ou longe. Amo vocês!

À minha prima/irmã amada, Gabriele. Obrigada pela força, incentivo, e por sempre estar me apoiando e acreditando que eu seria capaz de chegar até aqui e ir além do que imaginava ir. Sonhávamos com isso desde pequeninas. Amo você, prima!

A todos os meus colegas de classe do programa de pós-graduação em Estatística da UnB, em especial ao Alex Luiz (nerd), pelo seu grande potencial, pela sua inteligência e o dom de repassar seu incrível conhecimento para seus colegas. Obrigada por não medir esforços em nos ajudar e sempre se disponibilizar em sanar minhas dúvidas quando mais precisava. Você é sensacional!

Aos amigos, Leandro e Damião por todos os conhecimentos compartilhados, vocês com certeza marcaram esse mestrado pela alegria e censo de humor sem igual, não esquecerei dos papos descontraídos, da zoação quanto ao meu sotaque e de ser imitada pelo Dami (risos). Vocês são demais!

À minha amiga Luana, minha best do mestrado, aquela que compartilhou comigo momentos incríveis, difíceis e de superação, que no final sempre deu tudo certo, uma irmã que o mestrado me deu e que pretendo levar para a vida toda. Não caberia aqui tudo o que a gente viveu durante esses dois anos, ficam as lembranças de muitas aventuras memoráveis. Thanks, best!

À minha amiga, Cecília, a japa que sei que posso contar sempre, que foi excepcionalmente favorável comigo em todos os momentos em relação a qualquer tipo de assunto, uma pessoa incrível que compartilhou muito conhecimento comigo e tornou meus dias especiais em Brasília. Obrigada por me tirar de casa para descontrair sempre que eu precisava dar uma pausa. Obrigada pelos dias e noites de estudos em sua casa. E obrigada principalmente, por ter um coração tão generoso.

À minha amiga Elisangela, por termos compartilhado muitos momentos bons (por ser mais ansiosa que eu) e juntas superarmos isso. Obrigada pelo companheirismo nos últimos semestres, foram essenciais para que tudo corresse bem. E obrigada por não levar a sério minhas zoeiras (risos)! Gratidão!

Às minhas amigas de Brasília, Camila e Laís, as alunas especiais mais alto níveis que conheci, obrigada pela paciência de vocês em repassar seus incríveis conhecimentos comigo. Vocês têm toda minha admiração. Sou muito grata, meninas!

Ao Brunno Thadeu, por toda parceria e grandíssima contribuição, da graduação até o mestrado, pela parceria no projeto do CEBRASPE. Obrigada pelo apoio tanto emocional quanto intelectual ao longo deste processo, principalmente nessa fase final. Sou muito grata por poder compartilhar esta jornada com você. Obrigada por tudo!

A todos os professores do PPGEST-UnB, em especial ao Antonio Eduardo, Bernardo

Borba, Eduardo Nakano, Cibeli Queiroz, Cira Guevara, Gustavo Gilardoni, e Juliana Bettini. Fica aqui minha eterna gratidão a vocês!

Ao professor Luiz Pasquali e ao Instituto de psicologia social e do trabalho da UnB, no qual tive a honra de poder cursar uma disciplina. Fui muito bem recebida e vivi uma experiência incrível de ampliar meus conhecimentos em outra área.

Aos meus professores da UFPA por terem me apoiado e acreditado em mim, sem vocês certamente não teria chegado até aqui. Agradeço em especial ao professor Heliton Tavares, Regina Tavares e Marinalva Maciel. Minha gratidão por vocês é imensurável. Vocês são muito incríveis! Obrigada por tudo!

Aos meus amigos de graduação da UFPA, que juntos vivemos as experiências mais incríveis. Obrigada pelo companheirismo, pelos momentos de estudos e descontrações, foi essencial para que eu desse um passo a mais e chegasse até aqui. Agradeço em especial a Alice e Thamara que no último semestre de faculdade, estávamos mais juntas do que nunca, vivendo o sonho de alcançar o mestrado. Obrigada, meninas!

Ao meu amigo, Miguel Souza (o senhor humilde), uma pessoa sensacional, que tenho muito orgulho de ter conhecido e poder compartilhar minha vitória. Obrigada por todo apoio e confiança que tem em mim, sempre me fazendo acreditar que eu poderia ir além. Você é demais, Miguel!

À minha ex-chefe e amiga, Vanessa Pamplona, por me apoiar, incentivar e se preocupar comigo desde o primeiro semestre de faculdade, quando ainda tinha dúvidas do que realmente queria para minha vida profissional. Você me ensinou muito e contribuiu para que eu chegasse até aqui! Eterna gratidão!

Aos meus amigos de Tracuateua, minha cidade natal, que sempre fizeram que minhas férias fossem as melhores possíveis. Compartilharam comigo boas conversas, diversões e me proporcionaram a quantidade certa de risos nos momentos certos. Em especial à Lara, Miria, Samuel e Wanessa. Vocês são demais!

À CAPES pelo apoio financeiro.

Ao CEBRASPE por ter me concedido a participação e ampliado meu conhecimento no projeto, cujo o título: Avaliações seriadas para acesso ao ensino superior: um ensaio sobre não-respostas intencionais na teoria da resposta ao item.

Finalmente, gostaria de agradecer ao PPGEST-UnB por ter me acolhido tão bem durante estes dois anos em que fiz parte.

"*Até aqui nos ajudou o Senhor.* "

1 Samuel 7:12

"*Nothing is impossible. Some things are just less likely than others.*"

Jonathan Winters

Resumo

O presente trabalho apresenta um modelo bidimensional não-compensatório da teoria da resposta ao item para lidar com não-respostas intencionais em testes com itens dicotômicos. Uma dimensão fornece informações sobre o comportamento de omissão, chamado de propensão a responder, enquanto a outra dimensão está relacionada à habilidade do indivíduo. O modelo é ajustado aos dados de um exame do tipo *high stake* (alto risco) feito por 10.822 estudantes do ensino médio que participaram do programa de avaliação seriada da Universidade de Brasília em 2008. Nesse tipo de exame há grande incidência de não-respostas devido a particular forma de correção, em que uma resposta errada anula uma resposta correta. A estimação dos parâmetros dos itens (dificuldade e discriminação) foi feita via Máxima Verossimilhança Marginal. A proficiência e a propensão do candidato foram estimadas pelo método da Esperança a Posteriori. Na análise de ajuste dos dados ao modelo foi utilizada a medida de distância de Bhattacharyya como uma alternativa à medida quiquadrado. Em geral, as frequências observadas de acerto foram inferiores às suas respectivas frequências esperadas. Mesmo assim, 40 itens se mostraram aderentes ao modelo ajustado. Observou-se que os candidatos menos proficientes são menos propensos a responder de forma errônea, pois tendem a deixar a resposta em branco. Isso sugere que a decisão de responder ou não seja mais importante do que a decisão de responder corretamente ou não. Dessa forma, este trabalho mostra que a resposta em branco deve ser tratada como uma informação não ignorável, e que não tem relação apenas com a proficiência do candidato, mas também com as características dos itens e o traço latente propensão a responder.

Palavras-chave: Teoria da Resposta ao Item; Não-Resposta; distância de Bhattacharyya; teste high-stake

Abstract

The present work introduces a two-dimensional non-compensatory model of Item Response Theory to deal with intentional non-responses in tests with dichotomous items. One dimension provides information about the behavior of omission, called the propensity to respond, while the other dimension is related to the ability of the individual. The model is adjusted to the data of an examination of the type *high stake* made by 10.822 students of high school who participated in the program of Evaluation of the University of Brasília in 2008. In this type of examination there is a large incidence of non-responses by its particular form of correction, in which a wrong answer negates a correct answer. The estimation of the items parameters (difficulty and discrimination) was made via maximum Marginal likelihood. The candidate's proficiency and propensity were estimated by the expected a posteriori method. In the analysis of the adjustment of the data to the model was used the measure of distance of Bhattacharyya as an alternative to the chi-squared measure. In general, the observed frequencies of the hit were lower than their expected frequencies. Even so, 40 items have shown themselves adhering to the adjusted model. It has been observed that less proficient candidates are less likely to respond erroneously because they tend to leave the answer blank. That suggests that the decision to respond or not to respond is more important than the decision to respond correctly or not. In this way, this work shows that the blank answer should be treated as non-ignorable information, and that it is not only related to the candidate's proficiency, but also to the characteristics of the items and the latent trace propensity to respond.

keywords: Item response theory; Non-response; Distance from Bhattacharyya; High-stake test

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Identificar as habilidades cognitivas de um indivíduo, representado-as quantitativamente na forma de uma pontuação numérica confiável constitui um dos objetivos das avaliações educacionais e psicológicas (Erguven, 2013). Para isso, instrumentos psicométricos têm sido propostos para a construção de escalas métricas relativas a características individuais intangíveis como a inteligência, os traços de personalidade, os estados emocionais, as aptidões e —em avaliações educacionais — a proficiência e a habilidade na mobilização de conhecimentos para solucionar problemas. Tais características pessoais que não podem ser observadas diretamente denominam-se variáveis latentes (Baker, 2001).

Entre os métodos estatísticos existentes para a modelagem de variáveis latentes encontram-se a análise fatorial exploratória e a confirmatória, os modelos de equação estrutural e a teoria da resposta ao item (Beaujean, 2014; Finch e French, 2015). Em particular, o interesse deste trabalho diz respeito aos modelos da teoria da resposta ao item (TRI) aplicados em avaliações em larga escala no Brasil.

A TRI é uma abordagem para a análise de variáveis latentes, na qual as propriedades dessas variáveis são descritas a partir das respostas de um grupo de examinandos a um conjunto de itens (teste). Nessa perspectiva, considera-se que cada examinando possua um ou mais traços latentes, de modo que esses traços se relacionem com a probabilidade de acerto a determinado item do teste. Portanto a TRI descreve a relação entre a probabilidade de um examinando dar uma resposta correta a um item, as características desse item (como sua dificuldade), os traços latentes (habilidades ou proficiência) desse sujeito (Andrade et al., 2000).

A teoria do traço latente foi introduzida por Lord (1952) em sua tese de doutorado. Em seguida, Birnbaum (1957) escreveu uma série de relatórios técnicos sobre modelos

logísticos e estimativas de parâmetros dos modelos (Birnbaum, 1957, 1958). Rasch (1960) publicou seu livro propondo vários modelos de resposta ao item. Baker (1961) tratou comparação empírica entre as funções logística padrão e ogiva, enquanto Lord e Novick (1968), e Wright (1968) trabalharam em modelos para dados dicotômicos. Samejima (1968) propôs modelo para resposta graduada politômica. Andrade et al. (2000) foram os pioneiros na apresentação da TRI para a comunidade estatística nacional.

De fato, durante as últimas décadas, a avaliação educacional tem sido enriquecida com a aplicação da TRI para o desenvolvimento de testes educacionais. A combinação de avanços metodológicos e softwares cada vez mais poderosos têm aumentado a aplicabilidade e o interesse de pesquisadores para tal metodologia. A TRI oferece uma alternativa à Teoria Clássica dos Testes (TCT), que por muitos anos foi a metodologia dominante no desenvolvimento de testes educacionais. Mas devido a suas limitações, como a de não permitir comparar o desempenho de estudantes em testes diferentes (Pasquali e Primi, 2003), a TRI obteve mais espaço ao lado da TCT. Na TRI é possível separar os aspectos característicos do item (como sua dificuldade) da proficiência de cada examinando, além de possibilitar o acompanhamento do desempenho de indivíduos ao longo do tempo.

Para as avaliações educacionais em larga escala no Brasil, em particular, a TRI tem sido usada pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) desde 1995 no Sistema de Avaliação da Educação Básica (SAEB). Mas sua aplicação para essa finalidade ganhou destaque no Brasil a partir de 2009 quando ela foi adotada pelo INEP para o ENEM, Exame Nacional do Ensino Médio (INEP, 2012). Atualmente, outros programas avaliativos a utilizam, como o Sistema de Avaliação do Rendimento Escolar de Estado de São Paulo (SARESP) e a Avaliação Nacional do Rendimento Escolar (Prova Brasil). Entre outros exemplos internacionais, a TRI é utilizada pela avaliação Pisa (Programme for International Student Assessment, desde 2000) e o Toefl (Test of English as a Foreign Language, desde 1978).

Na prática, na maioria dos casos são considerados modelos unidimensionais, aqueles que se restringem a um único traço latente. Nesse caso, postula-se que todos os itens de teste avaliam uma única habilidade ou característica do examinando. Segundo Reckase (2009) esta suposição é atendida quando todos os itens de teste medem a mesma capacidade subjacente composta e os examinandos usam somente essa habilidade para responder ao teste. Esta suposição pode ser questionável, dado que alguns itens exigem a mobilização de

múltiplas habilidades para se alcançar uma resposta correta. Assim, como extensões dos modelos univariados, os modelos multidimensionais de Teoria de Resposta ao Item (TRIM) pressupõem que os itens sejam construídos em diferentes domínios, e que os instrumentos sejam projetados para medir várias dimensões dessas construções.

Uma dessas dimensões diz respeito a não-respostas (*missigness*) causadas pelas estratégias adotadas pelos examinandos, e os propósitos para os quais se aplicam o teste. Por isso é importante fazer a distinção entre avaliação educacional e processo seletivo. No segundo caso, o interesse maior do examinador consiste em discriminar os candidatos mais proficientes. E a competição proporciona, provavelmente, efeitos devidos a estratégias diversas utilizadas pelos candidatos para a apresentação das respostas. Por exemplo, enquanto o chute é uma recomendação para o caso de um candidato não saber responder a determinado item de uma prova do ENEM (INEP, 2012), para a prova do PAS/UnB a recomendação seria deixá-lo em branco (Cebraspe, 2016).

Nesta dissertação, o foco se concentra no segundo caso, em que se considera a resposta em branco como uma informação não ignorável, em que se discute que uma avaliação educacional que remete a um processo seletivo constitui situação especial. Dependendo da percepção do examinando sobre a sua capacidade de responder ao item, ele pode deixar de respondê-lo intencionalmente. Em parte, sem dúvida, a não-resposta se relaciona com a baixa proficiência do respondente sobre o conteúdo em questão. Mas também é possível que outro respondente, com proficiência mais elevada, seja impelido a não responder por motivos diversos. Embora não seja objeto deste trabalho, a existência de um traço latente relacionado a esse fenêmeno abriria portas, por exemplo, para futuras investigações comportamentais (Evans, 2008; Da Silva et al., 2015).

Partindo-se da premissa de que o respondente possua um traço latente que represente sua propensão a responder ou não a um item que verse sobre determinado assunto, e que ele adote uma estratégia para responder ou não com base nesse traço latente, este trabalho propõe a desenvolver um modelo que permita extrair informações proporcionadas pelo silêncio dos examinandos. Assim, por exemplo, seria possível avaliar a qualidade de um item não apenas com base na probabilidade de acerto, mas também pela probabilidade de ele não ter sido respondido.

Uma possível abordagem consiste em utilizar um modelo que seja função de um traço latente bivariado $(\theta_1; \theta_2)$, em que θ_2 representa a proficiência do indivíduo e θ_1 diz respeito

à sua propensão em apresentar uma resposta. Esse tipo de modelagem permite que as não-respostas sejam incorporadas à análise, proporcionando informações relevantes acerca da avaliação educacional. Essa linha foi introduzida por Knott et al. (1990) e Albanese e Knott (1992), e seguida por diversos autores (*e.g*, Knott e Tzamourani, 1997; Bartholomew et al., 1997; O'Muircheartaigh e Moustaki,1996, 1999; Moustaki e Knott, 2000; Moustaki e O'Muircheartaigh, 2000, 2002). Outra perspectiva de modelagem proposta por Holman e Glas (2005), que se baseia na modelagem da distribuição conjunta dos traços latentes e das características dos itens, foi considerada por outros autores (*e.g.*, Glas e Pimentel, 2006; Pimentel, 2005; Rose et al., 2010; Bertoli-Barsotti e Punzo, 2013). Santos et al. (2016) apresentam uma extensão do modelo de Holman e Glas (2005) para o caso multidimensional sob a perspectiva bayesiana.

Em nosso caso particular, a avaliação educacional também constitui processo seletivo, de modo que uma não-resposta deva ser computada como uma resposta errada (INEP, 2012). Embora autores como Rose et al. (2010) questionem essa prática por conta dos vícios na estimação da proficiência dos examinandos, tem-se na literatura que o ambiente competitivo (de processo seletivo, por exemplo), por si só, proporcionaria vício na produção dos próprios dados (e.g., Abdelfattah, 2007; Lievens et al., 2009). Por isso, aqui, em vez de vício, presume-se haver distinção entre uma proficiência do examinando em um ambiente avaliativo não competitivo com aquela observada em um processo seletivo. Este trabalho propõe um modelo alternativo dependente de traços latentes e das características do item, que considere a possibilidade de haver dois tipos de respostas erradas: a que foi apresentada de forma divergente do gabarito oficial e a que foi omitida.

O restante desta dissertação se organiza da seguinte maneira. O Capítulo 2 apresenta uma breve revisão da literatura acerca dos modelos logísticos unidimensionais 2.2 e multidimensionais 2.3 da TRI, restringindo-nos ao caso das respostas dicotômicas.

O Capítulo 3 descreve o modelo para a não-resposta intencional, em que se considera uma estrutura hierárquica: primeiro o candidato decide se responde ou não responde; e depois, caso ele responda, o resultado é classificado como certa ou divergente do gabarito oficial. Esse modelo baseia-se em probabilidades condicionais descritas por uma combinação de modelos logísticos de dois parâmetros. Como resultado, obtém-se uma curva característica bidimensional, em que a dimensão 1 representa a propensão de um indivíduo apresentar resposta divergente contra a possibilidade de ele deixar a resposta em branco; e

a dimensão 2 denota a proficiência do indivíduo sobre determinado assunto, sob um ambiente competitivo. Anstes, porém, uma breve discussão sobre a tipologia da não-resposta se encontra na Seção 3.2, e a Seção 3.3 aborda sobre as possíveis dinâmicas relacionadas com o objeto do estudo, em que se distingue uma avaliação de baixo grau competitivo (baixo risco, *low-stakes*) com outra de alto risco (alto grau competitivo, *high-stakes*).

Para se avaliar a aderência de um item ao modelo proposto, o Capítulo 4 sugere a utilização da distância de Bhattacharyya como uma alternativa para o coeficiente χ^2 e apresenta um método sequencial para a estimação das proficiências a partir de itens já calibrados. Um exemplo ilustrativo acerca da aplicação dos métodos propostos se encontra no Capítulo 5, considerando parte dos dados da prova da 3^a etapa do Programa de Avaliação Seriada da UnB (PAS/UnB) realizada em 2008. Finalmente, o Capítulo 6 apresenta uma conclusão deste trabalho.

Capítulo 2

Alguns modelos e questões pertinentes à teoria da resposta ao item

2.1 Introdução

Evidentemente, a elaboração de um bom instrumento de avaliação educacional requer itens de boa qualidade. Para essa avaliação, os instrumentos psicométricos são fundamentais e este capítulo se propõe a fazer uma breve revisão sobre a teoria da resposta ao item (TRI). Antes, porém, como um contraponto, iniciaremos com a teoria clássica dos testes (TCT), também conhecida como teoria clássica da medida, introduzida por Spearman (1904).

Essencialmente, a análise clássica se fundamenta em propriedades psicométricas dos itens obtidas com base no escore total (T), na contagem das respostas corretas dadas por cada examinando a um conjunto de itens. Com base no escore T obtém-se parâmetros descritivos dos itens que ajudam a explicar a distribuição das respostas apresentadas pelos respondentes (Borgatto e de Andrade, 2012).

Entre os principais parâmetros destacam-se o índice de dificuldade (fração dos examinandos que apresentaram respostas corretamente); e o índice de discriminação (capacidade de a cobrança proposta no item diferenciar os respondentes de maior habilidade dos demais, sendo medida pela diferença entre a proporção de acertos do grupo de examinandos que apresenta os maiores escores e a do grupo com os menores escores). Segundo Pasquali (2009), a TCT tem como principal interesse produzir testes de qualidade, e seu foco não é o traço latente do indivíduo e sim o seu comportamento. Em geral, o modelo é especificado, simplesmente, na forma linear

$$T_j = V_j + E_j, \tag{2.1}$$

no qual T_j representa o escore bruto do indivíduo j; V_j denota uma função dos escores esperados do indivíduo j; e E_j representa o erro aleatório, tendo supostamente média nula.

A estimação de parâmetros dos modelos na TCT é conceitualmente simples, requerendo geralmente suposições mínimas. No entanto, Klein (2013) lista algumas limitações dessa metodologia, como, por exemplo, (i) a dependência entre os parâmetros descritivos dos itens e o perfil do grupo de examinandos submetidos ao teste; (ii) dependência entre os escores do teste e as características dos itens apresentados aos examinandos; (iii) impossibilidade de se fazer comparações entre proficiências de indivíduos que realizam provas distintas; (iv) inexistência de resultados referentes ao desempenho de um examinando relativo a certo item.

Devido a essas limitações da TCT, a TRI preencheu o vácuo no meio educacional, e hoje, a TCT e a TRI são considerados métodos que se complementam (Borgatto e de Andrade, 2012).

Enquanto os modelos de TCT consideram os resultados com base na relação linear entre o escore esperado e observado, a TRI modela a probabilidade de acerto de um examinando a certo item do teste, levando-se em consideração as características do item (isto é, os seus parâmetros) e os traços latentes desse examinando que representem suas habilidades (ou proficiências) acerca da cobrança proposta nesse item. Um aspecto marcante dessa metodologia é a possibilidade de se construir uma escala avaliativa comum, o que permite comparar os resultados dos examinandos mesmo que eles respondam a itens diferentes. Com respeito a taxonomia dos modelos da TRI, eles dependem basicamente de três fatores (Andrade et al., 2000): (i) da natureza do item – podendo ser dicotômico ou politômico; (ii) do número de populações envolvidas – uma população ou mais; (iii) e da quantidade de traços latentes que está sendo avaliado – um traço latente ou mais.

Em relação ao primeiro fator, a TRI permite especificar diferentes modelos dependendo dos níveis de resposta em consideração. Por exemplo, os itens dicotômicos são aqueles que possuem dois estados possíveis (1 = certo ou 0 = errado). Já os itens politômicos possuem mais de dois níveis de resposta, podendo ser ordenáveis ou não ordenáveis. Por exemplo, as opções de resposta, 1 = discordo completamente, 2 = discordo um pouco,..., 5 = totalmente de acordo constituem um exemplo de escala ordinal.

No que diz respeito ao segundo fator, Andrade et al. (2000) ressaltam que, em avaliação educacional, as características de uma população variam dependendo do objetivo do estudo, sendo definidas pelo próprio desenho da amostragem em que se identifica quantas (e quais) populações devem ser consideradas.

O terceiro ponto, quando apenas uma habilidade afeta as respostas assume-se unidimensionalidade do traço latente. Por outro lado, caso seja necessária a mobilização de dois ou mais traços latentes para se responder corretamente aos itens propostos, tem-se um espaço multidimensional de traços latentes.

Nesta dissertação nos restringiremos aos modelos para testes constituídos por itens dicotômicos, aplicados para dois traços latentes relacionados com a não-resposta.

A seguir, na Seção 2.2, apresentaremos o modelo (unidimensional) logístico de três parâmetros (ML3P). A Seção 2.3 aborda sobre os modelos logísticos multidimensionais na perspectiva de Reckase (2009).

2.2 Modelo Logístico Unidimensional de Três Parâmetros -ML3P

Considere um grupo de n examinandos submetidos a um teste constituído por I itens de natureza dicotômica. Para cada examinando j com traço latente (habilidade ou proficiência) θ_j que responda ao item i, define-se o ML3P como uma função de probabilidade representada como (Andrade et al., 2000)

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

com $i = 1, 2..., I$, e $j = 1, 2...n$, em que U_{ij} é uma variável indicadora que assume valor igual a 1 caso o indivíduo j responda corretamente ao item i, ou 0 caso contrário; a_i é o parâmetro de discriminação ou de inclinação do item i; b_i é o parâmetro de dificuldade (ou posição) do item i; c_i denota o parâmetro de acerto casual (probabilidade do indivíduo j com baixa habilidade θ_j responder corretamente ao item i); e D é um fator de escala, que, ao longo do restante desta dissertação, assume valor unitário.

A função de probabilidade definida em 2.2 é chamada de curva característica do item, cujas formas são exemplificadas na Figura 2.1. O parâmetro b_i ajusta a locação da curva de crescimento. Para que a probabilidade de acerto seja superior a 0,5, basta que o examinando j possua θ_j superior a b_i. Assim, à medida que b_i aumenta, maior será a exigência

sobre o examinando. Por isso b_i representa a dificuldade do item. Como $b_i - \theta$ representa um desvio, tanto b_i como θ_j se definem na mesma escala.

O parâmetro a_i reflete o grau em que um item pode discriminar entre examinandos com altas e baixas habilidades θ. Valores positivos altos indicam maior poder de discriminação, ou seja, proporciona uma curva que permite separar melhor os indivíduos com θ abaixo do parâmetro b_i daqueles que possuem habilidades acima do parâmetro b_i. Em modelos unidimensionais não se espera que a_i assuma valores negativos, já que a probabilidade de o indivíduo j responder corretamente ao item i não pode diminuir à medida que sua habilidade aumenta (Andrade et al., 2000).

Teoricamente, por ser uma probabilidade, o parâmetro c pode assumir valores entre 0 e 1, embora não seja razoável haver item com $c > 0,5$. Por representar o limite assintótico para $\theta \to -\infty$, esse parâmetro reflete as chances de um indivíduo de proficiência muito baixa responder corretamente ao item i.

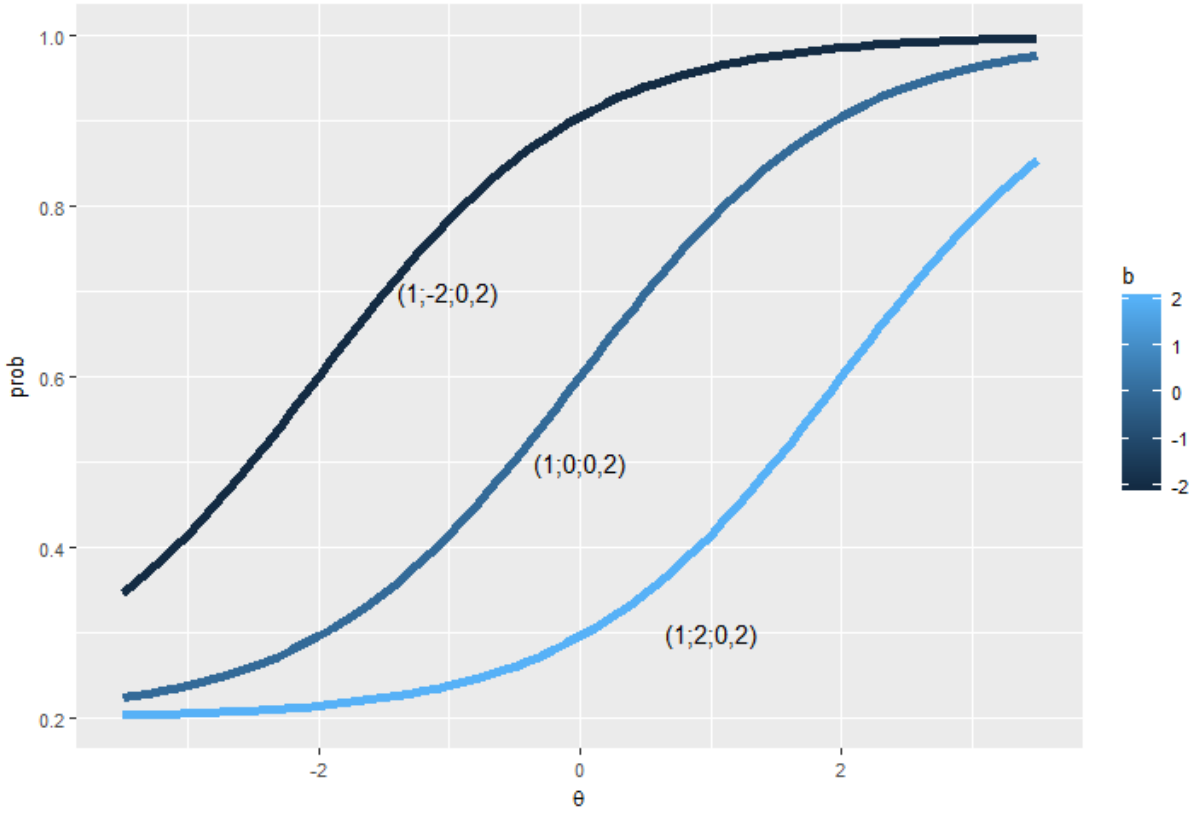


Figura 2.1: Exemplos de curvas características do item (CCI) com parâmetros (a; b; c).

Para a utilização desses modelos, certas suposições precisam ser atendidas para que os resultados sejam precisos e úteis. Ao contrário da TCT que faz suposições ao nível

do teste, a TRI faz suposições a nível do item. Elas referem-se à unidimensionalidade, a independência local ou condicional e a adequação do item ao modelo proposto. A suposição de unidimensionalidade estabelece que as respostas ao item sejam conduzidas por um único traço subjacente (Ayala, 2009). Isso implica que os itens que constituem um teste meçam a mesma habilidade latente. Se essa suposição for violada e se os itens de teste realmente medirem traços múltiplos, mas se forem modelados unidimensionalmente, haverá imprecisão na estimação dos parâmetros dos itens e na obtenção dos traços θ_j. Por exemplo, Walker e Beretvas (2003) verificaram diminuição da precisão dos resultados com a utilização de um modelo TRI unidimensional quando na verdade os itens deveriam ser considerados como bidimensionais. Nesse caso, se dois examinandos possuírem níveis semelhantes de um traço principal, mas diferentes níveis de habilidades secundária, caso apenas o primeiro traço seja observado, os examinandos seriam classificados como semelhantes quando na realidade não são.

A suposição de independência local acompanha naturalmente a suposição de unidimensionalidade (Lord, 1980), e implica que as respostas a um item sejam independentes das respostas aos outros itens, dependendo apenas do nível de habilidade θ do indivíduo. Em outras palavras, se os itens forem localmente independentes, não serão correlacionados após o condicionamento em θ_j (DeMars, 2010). Seja $U_{ij} = (u_{1j}, u_{2j}, ..., u_{ij})$, com $i = 1...I$ e $J = 1...n$, o vetor aleatório de respostas observadas do sujeito j com habilidade θ_j e seja $\zeta = (\zeta_1, ..., \zeta_I)$ o conjunto de parâmetros dos itens. A suposição de independência local pode ser expressa como

$$P(\boldsymbol{u_j}|\theta_j, \boldsymbol{\zeta}) = P(u_{1j}|\theta_j, \zeta_1)P(u_{2j}|\theta_j, \zeta_2)...P(u_{ij}|\theta_j, \zeta_I) = \prod_{i=1}^{I} P(u_{ij}|\theta_j, \zeta_i), \qquad (2.3)$$

em que $P(\boldsymbol{u_j}|\theta_j, \boldsymbol{\zeta})$ é a probabilidade de que o vetor de respostas observadas dos itens para uma pessoa com nível de traço θ_j tenha o padrão $\boldsymbol{u_i}$. Expressando a equação 2.3 em termos do θ_j, obtemos a função de verossimilhança,

$$L(\theta_j|\boldsymbol{u_i}, \boldsymbol{\zeta}) = \prod_{j=1}^{n} P(u_{ij}|\theta_j, \zeta_i).$$

Generalizando para a probabilidade $P(\boldsymbol{u}|\boldsymbol{\theta})$ de um conjunto completo de respostas de n pessoas para I itens em um instrumento, em que $\boldsymbol{\theta} = (\theta_1, \theta_2..., \theta_n)$ representa o vetor de habilidades para todos os examinandos, tem-se a equação 2.4:

$$L(u_{i.}|\theta_j,\zeta) = \prod_{j=1}^{n} P(u_{.j}|\theta_j,\zeta) = \prod_{j=1}^{n}\prod_{i=1}^{I} P(U_{ij} = u_{ij}|\theta_j,\zeta_i), \tag{2.4}$$

em que na última igualdade assume-se que a distribuição de U_{ij} só depende de ζ através de ζ_i.

Vale ressaltar que o modelo logístico de dois parâmetros (ML2P) e um parâmetro (ML1P) são casos particulares do ML3P. O ML2 especifica que o parâmetro c de acerto casual seja igual a 0, enquanto o ML1P, também conhecido como o modelo de Rasch (1960), faz a restrição adicional de discriminações iguais para todos os itens, ou seja $a_i = 1, i = 1, ..., I$.

Estimação dos parâmetros

Em relação à estimativa dos parâmetros, existem diversos métodos para estimar as habilidades dos indivíduos e os parâmetros dos itens. No âmbito da estimativa de máxima verossimilhança, tem-se se três métodos principais: a Máxima Verossimilhança Conjunta (MVConj), a Máxima Verossimilhança Condicional (MVC) e a Máxima Verossimilhança Marginal (MVM).

As aplicabilidades das MVConj e MVC são bastante limitadas. A MVConj pode ser usada para modelos TRI de um, dois e três parâmetros. O método MVConj funciona ao estimar simultaneamente o parâmetro do item e do indivíduo através de um procedimento iterativo. A estimativa da MVConj sofre de inconsistências dos parâmetros, uma vez que o número de estimativas aumenta com o número de observações (Little e Rubin, 1984; Baker e Kim, 2004)

O problema da inconsistência pode ser contornado usando o método de máxima verossimilhança condicional (MVC). A MVC foi um método sugerido por Andersen (1970) em que se baseia na propriedade de que o resultado da soma é uma estatística suficiente em relação à variável latente θ do indivíduo em modelos de um parâmetro da família Rasch. A limitação desse método é que infelizmente a MVC não é aplicável para modelos de dois e três parâmetros.

O método de estimação MVM é o mais aplicado, um estimador de MV alternativo foi desenvolvido de modo que é aplicável a modelos de um, dois e três parâmetros (Bock e Lieberman, 1970; Bock e Aitkin, 1981, Baker e Kim, 2004). O problema da inconsistência

pôde ser resolvido assumindo uma distribuição da variável latente θ que pode ser descrita de forma paramétrica. Em vez de estimar todos os parâmetros do indivíduo, apenas os parâmetros da distribuição $g(\theta)$ precisam ser estimados conjuntamente com os parâmetros do item. O número de estimativas é independente do tamanho da amostra. Para obter as estimativas de parâmetros, o método faz uso do algoritmo EM (Ayala, 2009).

Geralmente, assume-se a distribuição normal (multivariada), que é suficientemente especificada pelo vetor de valores esperados $E(\theta)$ e pela matriz de variância-covariância (θ). Em um primeiro estágio, estima-se os parâmetros dos itens. Em seguida, as estimativas de habilidade individuais são estimadas em um procedimento de estimativa subsequente, levando os parâmetros dos itens previamente estimados como conhecidos. Diferentes estimadores de habilidade foram desenvolvidos, como a estimação por máxima verossimilhança (MV), estimação pelo máximo da posteriori (MAP) e estimação pela esperança ou média a posteriori (EAP). O método EAP é recomendado por alguns autores (e.g. Mislevy e Stocking, 1989) por possuir vantagem da estimação ser calculada diretamente e não havendo necessidade de fazer uso de métodos iterativos.

Nesta dissertação será utilizado o método MVM e EAP para obter as estimativas dos parâmetros do itens e habilidades, respectivamente.

Análise de ajuste do modelo

Em qualquer aplicação de um modelo de TRI é importante avaliar em que medida os pressupostos do modelo são válidos para os dados fornecidos e a forma como os itens de teste se adequam ao modelo selecionado. A violação dos pressupostos do modelo de TRI, ou inadimplência entre o modelo usado e os dados de teste, podem levar a estimativas erradas ou instáveis dos parâmetros do modelo (Fan, 1998). Para detectar itens problemáticos é preciso avaliar em termos de sua qualidade de ajuste usando um teste estatístico ou uma análise de resíduos. É importante enfatizar que um ajuste adequado de um modelo para os dados é essencial para uma análise de itens bem sucedida. Os itens com problemas geralmente são identificados através dos valores de seus índices de discriminação (o valor de a_i será um baixo valor positivo ou mesmo negativo) e os índices de dificuldade (os itens não devem ser nem muito fáceis nem muito difíceis de avaliar para um determinado grupo de examinandos).

Conforme Hambleton et al. (1991) e Stone (2000), uma estratégia comum para avaliar

o ajuste de itens de um modelo de TRI pode ser resumida da seguinte forma: (1) estimar o item e os parâmetros de habilidade sob o modelo escolhido, (2) classificar os examinandos em k grupos homogêneos em termos de suas estimativas de habilidade, (3) calcular proporções de respostas observadas em cada grupo para o item sob investigação, (4) obter as proporções de respostas previstas em cada grupo usando as estimativas de parâmetros de item e habilidade sob o modelo de interesse e (5) calcular as estatísticas baseadas no teste quiquadrado, comparando os valores observados e previstos.

Por exemplo, para se avaliar a aderência de um item i a determinado modelo ajustado, a estatística de teste Q_1 pode ser obtida da forma (Hambleton et al., 1991):

$$Q_{1i} = \sum_{j=1}^{k} \frac{N_j[P_{ij} - E(P_{ij})]^2}{E(P_{ij})[1 - E(P_{ij})]}, \tag{2.5}$$

em que os examinandos são divididos em k categorias com base em suas estimativas de habilidade θ (usa-se, por exemplo, o ponto médio das habilidades do subgrupo); i denota o item; j é a categoria de habilidade (subgrupo); N_j é o número de sujeitos dentro da categoria; P_{ij} é a proporção observada de respostas corretas no item i na categoria j; $E(P_{ij})$ é a proporção de respostas esperadas dos sujeitos que acertaram o item i na categoria j.

A estatística Q_1 tem distribuição quiquadrado com $k - m$ graus de liberdade, m é o número de parâmetros do modelo selecionado. Se o valor observado da estatística for maior que o valor crítico (obtido a partir da tabela do quiquadrado), a hipótese nula de que a ICC se ajusta aos dados é rejeitada e então deve ser encontrado um modelo que melhor se ajuste aos dados.

A análise de resíduos também é um procedimento importante para a qualidade de ajuste do modelo na TRI. Neste método, depois de escolhido um modelo supostamente válido, os parâmetros dos itens e habilidades são estimados e pode-se se fazer previsões sobre o desempenho de vários grupos de habilidade. O objetivo desta análise, consiste em verificar se a diferença entre o desempenho real dos sujeitos e o desempenho predito pelo modelo difere estatisticamente de 0. Esaa diferença pode ser expressa da forma

$$r_{ij} = P_{ij} - E(P_{ij}),$$

em que r_{ij} é a diferença entre o desempenho do item observado para um subgrupo de examinandos e o desempenho esperado do item do subgrupo, também conhecido como

resíduo bruto.

Existe uma limitação desse resíduo, pois ele não leva em consideração o erro de amostragem associado à pontuação correta da proporção esperada dentro de uma categoria de habilidade. Para considerar este erro de amostragem, calcula-se o z_{ij} padronizado dividindo o resíduo bruto pelo erro padrão da proporção esperada correta,

$$z_{ij} = \frac{P_{ij} - E(P_{ij})}{\sqrt{E(P_{ij})[1 - E(P_{ij})]/N_j}}.$$

2.3 Modelos Multidimensionais

A abordagem multidimensional da teoria da resposta ao item (TRIM) ganhou maior destaque no início das décadas de 70 e 80, quando os modelos unidimensionais foram estendidos para combinar diversos conhecimentos em avaliação educacional, em que, supostamente, a resposta de uma pessoa a um item é influenciada por vários traços latentes (Yeh, 2007,Reckase, 2009). Os conceitos usados nos modelos TRIM compartilham muitas semelhanças com aqueles usados nos modelos unidimensionais. Por exemplo, a probabilidade de um indivíduo responder corretamente a determinado item ainda é expressa em função das características do item e do indivíduo. No entanto, essas características individuais são representadas em um espaço com dimensão p, na forma de um vetor de parâmetros $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ (Reckase, 2009). Semelhantemente aos modelos unidimensionais, a suposição de independência local também se aplica no contexto da TRIM. Essa suposição estabelece que as respostas dada por um indivíduo particular são reciprocamente independentes, sendo modeladas somente com base em $\boldsymbol{\theta}$ e o vetor dos parâmetros dos itens ($\boldsymbol{\gamma}$). As análises de dimensionalidade também devem remeter ao número de traços latentes especificados no modelo TRIM.

Reckase (1972) propôs uma extensão do modelo de Rasch ao caso multidimensional. Em 1982, McKinley e Reckase desenvolveram modelos logísticos TRIM de dois e três parâmetros. Esses modelos são caracterizados como modelos compensatórios, ou seja, os p traços latentes se combinam linearmente, de modo que uma habilidade baixa pode ser compensada por outra mais elevada. Simultaneamente, nos anos setenta e oitenta, outro conjunto de modelos foram desenvolvidos e receberam o nome de modelos TRIM não compensatórios. Nesse tipo de modelo, os traços latentes se combinam de forma

multiplicativa, de forma que não é possível compensar uma habilidade baixa por outra (Sympson, 1978; Whitely, 1980).

2.3.1 Modelo Logístico Multidimensional Compensatório

Os modelos compensatórios remetem a uma combinação linear de proficiências associadas a uma forma logística ou ogiva normal para se especificar a probabilidade da ocorrência de uma resposta correta (Reckase, 2009). Se um item requer duas proficiências diferentes, que se definem em um espaço bidimensional, a alta proficiência de uma pessoa no primeiro traço latente pode compensar a baixa proficiência no segundo (ou vice-versa). Por exemplo, suponha que em um problema de matemática haja duas dimensões subjacentes, cuja solução pode ser encontrada mediante conhecimentos matemáticos ou pela própria leitura do texto. Nesse caso, a primeira dimensão pode refletir a proficiência em matemática, enquanto a segunda dimensão remete à proficiência em leitura. Se um indivíduo possuir alta proficiência de leitura, ele poderia ser capaz de compensar, em certa medida, uma eventual proficiência baixa em matemática (Svetina, 2011).

Considerando um teste constituído por I itens e n examinandos, a forma do Modelo Logístico Multidimensional Compensatório de três parâmetros proposto por McKinley e Reckase (1980), que representa a probabilidade de um indivíduo j com traços latentes $\theta_{1_j}, \theta_{2_j}, ..., \theta_{k_j}$, responder corretamente ao item i é expressa por:

$$P(U_{ij} = 1|\boldsymbol{\theta_j}, \boldsymbol{a_i}, d_i, c_i) = c_i + (1 - c_i)\frac{1}{1 + \exp\{-\sum_{k=1}^{p} a_{ki}\theta_k + d_i\}}, \tag{2.6}$$

que U_{ij} é a resposta do indivíduo j ao item i, com $i = 1, 2, ..., I$, I consiste no número de itens de teste; p consiste na dimensão do vetor de traços latentes (habilidades); d_i está associado a dificuldade do i-ésimo item, e pode ser representado como:

$$d_i = -b_i\sqrt{(a_{1i}^2 + a_{2i}^2 + ... + a_{ki}^2)},$$

$\boldsymbol{a_i} = (a_{1i}, ..., a_{pi})$ representa o vetor de discriminação do i-ésimo item; c_i é a probabilidade de acerto casual do item para indivíduos com baixa habilidade.

Note que as proficiências que constituem o expoente do modelo 2.6 têm uma relação aditiva. Cada parâmetro de discriminação do modelo multidimensional pode ser interpretado da mesma forma que o parâmetro de discriminação do modelo unidimensional dado em 2.2

Reckase (2009), ou seja, ele está associado à inclinação da superfície de resposta do item na dimensão do traço latente correspondente. O poder de discriminação multidimensional do item i ($MDISC_i$) é definido como

$$MDISC_i = \sqrt{\sum_{k=1}^{p} a_{ki}^2},$$

de modo que, $MDISC_i$ representa o tamanho do vetor do item i.

Já o parâmetro de dificuldade multidimensional para o item i , d_i, não deve ser interpretado da mesma maneira que o parâmetro b_i dos modelos unidimensionais. Mas existe uma forma equivalente ($MDIFF_i$) que representa a distância da origem até o ponto do vetor de maior inclinação, definida como:

$$MDIFF_i = \frac{-d_i}{MDISC_i}.$$

No caso multidimensional, a função 2.6 é chamada de superfície característica do item (SCI). A Figura 2.2 exemplifica a forma do modelo compensatório para o caso bidimensional através de um gráfico de superfície tridimensional e ilustra a probabilidade de uma resposta correta a um item como uma função da SCI. O gráfico mostra claramente que um θ alto em uma dimensão pode compensar um θ baixo na outra dimensão. Por exemplo, uma pessoa com $\theta = -2$ na segunda dimensão do item ainda pode ter uma probabilidade de acerto tão alto quanto 0,9 se o valor da habilidade (θ) do examinando na primeira dimensão for próximo de 3.

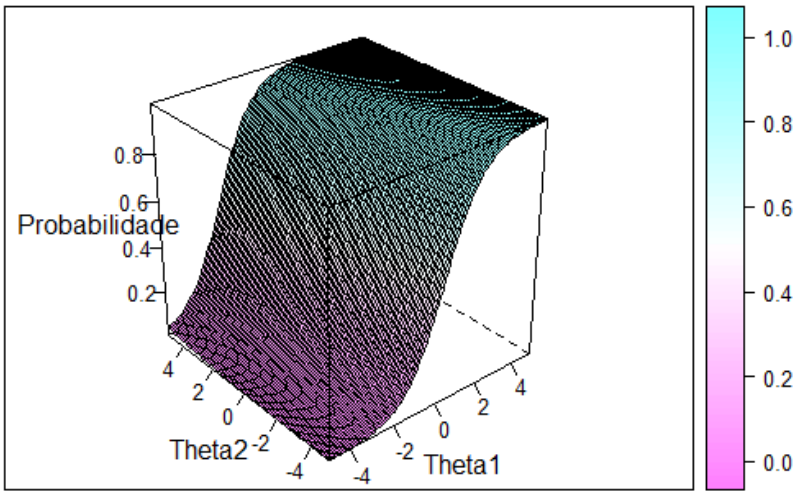


Figura 2.2: Gráfico de superfície para a probabilidade de resposta correta a um item compensatório de duas dimensões com $a_1 = 1,2, a_2 = 0,3, d = 1.0$ (adaptado de Reckase, 2009)

2.3.2 Modelo Logístico Multidimensional Não-Compensatório

Já nos modelos não compensatórios, há separação das tarefas cognitivas do item de teste em partes, de modo que cada uma delas seja modelada de maneira unidimensional. Nesse caso, a probabilidade de ocorrência de uma resposta correta para o item será o produto das probabilidades relativas a cada parte, o que proporciona características não-lineares para essa classe de modelos (Reckase, 2009). Se um item em um teste exige duas competências diferentes, o domínio em uma delas pode não ser suficiente para compensar a falta da outra. Em outras palavras, todas as proficiências subjacentes envolvidas na cobrança devem ser mobilizadas para que haja sucesso na resposta do examinando. Por exemplo, em um item que verse sobre analogia verbal, o domínio em duas proficiências, como a construção de regras e a avaliação de regras pode ser necessário para que o item seja respondido com sucesso. Nesse caso, se um indivíduo tem alta habilidade na construção de regras, mas possuindo baixa habilidade na avaliação de regras, a probabilidade de ele responder corretamente pode não ser alta. Esse tipo de relacionamento é a razão pela qual esses modelos são frequentemente chamados como não-aditivos ou multiplicativos (Svetina, 2011).

O modelo logístico multidimensional não compensatório de três parâmetros (Sympson, 1978; Whitely, 1980) pode ser representado pela seguinte expressão:

$$P(U_{ij} = \boldsymbol{\theta}, \boldsymbol{a_i}, \boldsymbol{b_i}, c_i) = c_i + (1 - c_i) \prod_{k=1}^{p} \frac{1}{1 + \exp\{(-D_{a_{ki}}\theta_{jk} + b_{ki})\}},$$

em que $\boldsymbol{a_i}$, c_i e $\boldsymbol{\theta}$ são definidos como em 2.6, e o vetor $\boldsymbol{b_i} = (b_{i1}, ..., b_{pi})$ representa a dificuldade do item associado a cada traço latente.

A Figura 2.3 mostra um item não-compensatório bidimensional. Uma pessoa com um θ_{jk} baixo não terá uma alta probabilidade de acertar o item, não importa o quanto mais alto seja o outro valor de uma outra proficiência θ_{jk}. A probabilidade de resposta correta para valores baixos de θ_1 ou θ_2, é próxima de zero. Somente quando ambos os valores θ forem altos, o modelo produziria maior probabilidade de resposta correta.

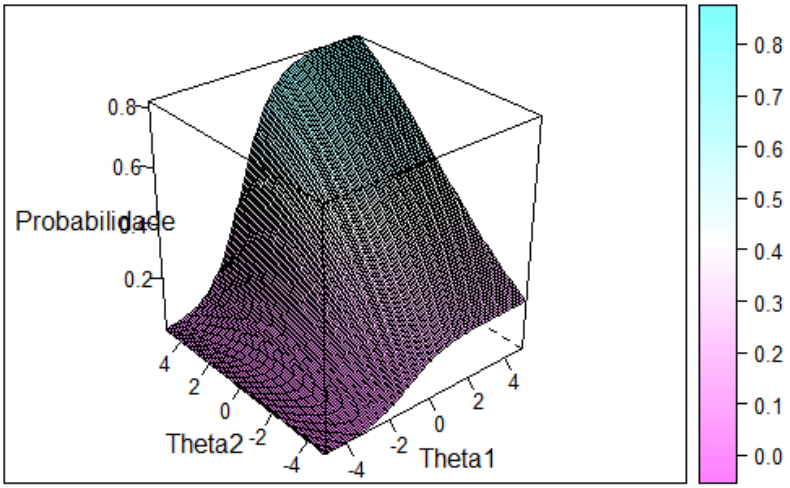


Figura 2.3: Gráfico de superfície para a probabilidade de resposta correta para um item não-compensatório de duas dimensões com $a_1 = 1, 2, a_2 = 0, 5, b_1 = 1, b_2 = 0$ (adaptado de Reckase, 2009)

Smith (2009) relata que esse modelo geralmente exige alto custo computacional para seu ajuste devido à sua natureza multiplicativa. No entanto, esta dissertação tratará de um um caso de modelo bidimensional não-compensatório, cujo processo de estimação remete a dois modelos unidimensionais (Capítulo 3). Por isso, aspectos pertinentes à estimação de parâmetros não serão abordados aqui.

Capítulo 3

Não-Respostas Itencionais

3.1 introdução

Na literatura, encontram-se algumas abordagens para o problema de não-respostas de itens do tipo *missing not at random* (MNAR). O'muircheartaigh e Moustaki (1999) dentro do contexto de modelo padrão simétrico, introduziram uma abordagem para variável latente para lidar com não-respostas. Holman e Glas (2005), assim como Korobko et al. (2008) utilizaram a mesma ideia para tratar as não-respostas não ignoráveis. Eles usaram a matriz $\mathbf{D}$ (ver 3.2) das variáveis indicadoras d_i a fim de estabelecer um modelo para "propensão à resposta". As estimativas de proficiência θ podem ser usadas para calcular pesos para cada observação (Moustaki e Knott, 2000). De maneira alternativa, o modelo de medição para essa propensão de resposta foi inserido ao modelo de estimação da variável de interesse (habilidade), θ. Dessa forma, o modelo resulta em um modelo multidimensional de TRI e as informações das respostas omitidas em relação a habilidade do indivíduo (θ) são levadas em consideração na estimativa de máxima verossimilhança. O modelo TRIM pode ser especificado de várias maneiras para ilustrar essa relação entre habilidade e propensão à resposta. Uma dessas abordagens é considerar a não-resposta como uma nova categoria de resposta em um modelo de resposta nominal, evitando a modelagem de uma segunda variável latente (Moustaki e Knott, 2000). Em um contexto geral, a ideia básica desses modelos é especificar um modelo de seleção (Heckman, 1979) que explique a possível dependência estocástica entre a habilidade θ e a ocorrência dos dados omitidos.

A seguir, apresentaremos duas outras discussões pertinentes ao nosso trabalho, concernentes à tipologia das não-respostas (Seção 3.2) e a necessidade de se fazer distinção entre os processos avaliativos e os seletivos (Seção 3.3), além de um modelo para não-respostas

intencionais na Seção 3.4.

3.2 Tipos de Não-Respostas

Aplicações que envolvem preenchimentos de formulários estão sujeitas ao problema de dados faltantes, ou não respostas. Em avaliação educacional, os dados faltantes ou não-respostas são um problema frequente, o que requer uma atenção especial dos pesquisadores. Se o mecanismo desses tipos de dados for tratado de forma indevida, a questão pode se tornar um problema particularmente crítico. Por isso, Rubin (1976), define uma taxonomia para os mecanismos de dados faltantes, classificando-os como *missing completely at random* (MCAR), *missing at random* (MAR) ou *missing not at random* (MNAR). Se a motivação para a não-resposta for independente de qualquer outra variável, seja observável ou não-observável, então tal mecanismo gerador de não-respostas é MCAR. Já as não-respostas que não dependem de variáveis não-observáveis, embora elas possam depender de variáveis observáveis, denominam-se MAR. E por fim, as não-respostas a um item são consideradas MNAR quando seu mecanismo gerador depende das variáveis não-observáveis como, por exemplo, uma variável latente (Little e Schenker, 1995; Little e Rubin, 2002).

Para os casos MCAR e MAR, no contexto da estimação de parâmetros pelo método da máxima verossimilhança, Little e Rubin (2002) discutem que as não-respostas podem ser consideradas ignoráveis por não haver relação entre seu processo gerador e os parâmetros objetos da estimação. Nesse caso, a presença desses dados faltantes não proporcionariam vícios sistemáticos nas estimativas dos parâmetros. Em contraste, no caso em que os dados são MNAR, os dados faltantes são não ignoráveis, no qual a própria lacuna de informação fornece informações adicionais relativos ao objeto de estudo. Essa informação silenciosa precisa ser de algum modo considerada no processo de estimação dos parâmetros para que os resultados estatísticos não sejam distorcidos pela omissão dos dados.

Em um sentido amplo, podemos distinguir dois casos principais de dados faltantes. Tais se referem ao desenho do processo de coleta de dados. O primeiro tipo remete ao caso em que há um desenho incompleto estabelecido pelo próprio observador (*planned missingness*). Nesse caso de não-respostas planejadas encontram-se, por exemplo, o delineamento aleatório incompleto (*random incomplete design*, teste de múltiplos estágios (*multistage testing design*) e o teste direcionado (*targeted testing design*). Em situações como essas, a

aleatorização é um meio possível para que os dados não observados sejam do tipo MAR (Little e Schenker,1995; Schafer,1997; Mislevy e Wu,1996; Eggen e Verhelst, 2011).

O segundo tipo refere-se à situação em que a não-resposta é consequência de uma escolha do respondente (*unplanned missingness*). Esse é o caso em tela nesta dissertação, em que o item do formulário é apresentado ao respondente, ele tem tempo suficiente para avaliá-lo (ou seja, não se considera o caso em que o item não foi apresentado ao respondente) e, por alguma razão, o candidato decide não responder (Mislevy e Wu, 1996; Bertoli-Barsotti e Punzo, 2013). Embora a literatura exemplifique outras situações (*e.g*, Holman e Glas, 2005), nessa dissertação o interesse se volta para a não-resposta intencional associada não apenas à baixa proficiência do respondente. Aqui, para o caso particular em que uma resposta errada penaliza o escore bruto de um candidato em um processo seletivo (Cebraspe, 2016), considera-se que a não-resposta seja motivada, supostamente, por uma razão estratégica, e que seu processo gerador dependa das características do item e de um traço latente do examinando (propensão a responder).

A questão de como lidar com essas não-respostas não ignoráveis em avaliações educacionais tem sido discutida desde o final dos anos 90, quando foram desenvolvidos modelos para não-respostas não ignoráveis (ver, Glas e Pimentel, 2008; Holman e Glas, 2005; Korobko et al., 2008; Moustaki e Knott, 2000; O'muircheartaigh e Moustaki, 1999; Rose, 2013; Rose et al., 2010).

Há evidências fortes de que as não-respostas aos itens em testes estão relacionadas às características dos indivíduos que não são observadas pelos examinadores. Por exemplo, Rose et al. (2010) mostraram que, nos dados da avaliação PISA 2006, a proporção de itens respondidos corretamente está substancialmente correlacionada com a proporção de respostas omitidas. Os participantes do teste com proporções mais baixas de respostas corretas tiveram, em média, mais respostas omitidas. Da mesma forma, Culbertson (2011) encontrou em avaliações educacionais de larga escala que a probabilidade de omissões de itens aumenta com a diminuição dos níveis de proficiência do candidato. Essas descobertas indicam uma relação entre a não-resposta e o desempenho do teste, o que sugere uma dependência entre a ocorrência de não-respostas e a proficiência dos indivíduos. Portanto, as não-respostas dependem de variáveis não-observáveis, o que é característica para um mecanismo de dados faltantes não-ignoráveis, do tipo MNAR. Esse fato ressalta a necessidade de métodos apropriados para lidar com não-respostas não ignoráveis em medidas

psicológicas e educacionais.

Para contextualizar os três tipos de dados ausentes, no ambiente da TRI segundo Rubin (1976), considere a matriz de dados completa $\boldsymbol{U} = (\boldsymbol{U}_{\text{obs}}, \boldsymbol{U}_{\text{miss}})$ que seguindo uma estrutura para variável latente, consiste nas respostas dos itens que foram observadas $\boldsymbol{U}_{\text{obs}}$ e as respostas que foram omitidas $\boldsymbol{U}_{\text{miss}}$ pelos n examinandos aos I itens do teste. Considerando-se uma variável latente θ e um conjunto de covariáveis como gênero, status socioeconômico, entre outras que constituem uma matriz $\boldsymbol{Z}$, o caso MCAR representaria o caso em que a distribuição de dados missings independe de $\boldsymbol{U}_{\text{obs}}$, $\boldsymbol{U}_{\text{miss}}$, $\boldsymbol{Z}$ e θ, ou seja,

$$\boldsymbol{P}(\boldsymbol{D}|\boldsymbol{U}_{\text{obs}}, \boldsymbol{U}_{\text{miss}}, \boldsymbol{Z}, \theta) = P(\boldsymbol{D}),$$

em que, a matriz $\boldsymbol{D}$ constituída por variáveis indicadoras d_{ij} representa a ocorrência dos valores de U_i, ou seja, $d_{ij} = 1$ se U_{ij} é observado, e $d_{ij} = 0$, se caso contrário.

Se a distribuição dos dados faltantes depender apenas das variáveis observáveis $\boldsymbol{U}_{\text{obs}}$ e $\boldsymbol{Z}$, mas não depender dos valores das variáveis não-observáveis $\boldsymbol{U}_{\text{miss}}$ e θ, mantendo-se o mecanismo de dados do tipo MAR. Tem-se

$$\boldsymbol{P}(\boldsymbol{D}|\boldsymbol{U}_{\text{obs}}, \boldsymbol{U}_{\text{miss}}, \boldsymbol{Z}, \theta) = \boldsymbol{P}(\boldsymbol{D}|\boldsymbol{U}_{\text{obs}}, \boldsymbol{Z}).$$

Finalmente, O terceiro tipo de mecanismo de dados faltantes, descrito como

$$\boldsymbol{P}(\boldsymbol{D}|\boldsymbol{U}_{\text{obs}}, \boldsymbol{U}_{\text{miss}}, \boldsymbol{Z}, \theta) \neq \boldsymbol{P}(\boldsymbol{D}|\boldsymbol{U}_{\text{obs}}, \boldsymbol{Z}),$$

denomina-se MNAR, sendo o oposto do MAR, isto é, a distribuição condicional dos dados faltantes e $\boldsymbol{U}_{\text{obs}}$ e $\boldsymbol{Z}$ depende dos dados não-observáveis $\boldsymbol{U}_{\text{miss}}$ e θ.

3.3 Processos avaliativos x Seletivos

Um dos principais objetivos da avaliação em contextos educacionais é medir o desempenho dos examinandos para fins de inferências estatísticas e tomada de decisões. A precisão das decisões pode depender das consequências educacionais para os examinandos ou para a escola na qual estudam. Os formuladores de políticas educacionais muitas vezes utilizam testes para finalidades como monitoraramento do sistema educacional e o acompanhamento do desempenho dos estudantes, com vistas à melhoria da qualidade educacional das escolas

e dos educadores. Esses testes também podem servir para certificar os indivíduos de acordo com níveis específicos de desempenho (Klein e Hamilton, 1999; Hamilton et al., 2002).

Conforme Abdelfattah (2007), esses tipos de testes diferem em suas consequências e podem ser responsáveis por alguma variação no desempenho do examinando. De um lado, os testes denominados *high-stakes* (de alto risco) refletem um conjunto de políticas que incluem procedimentos que proporcionam recompensas como consequência dos resultados dos exames em avaliações. Nesse tipo de teste o examinando deve alcançar uma pontuação elevada, para obter um benefício desejado. Isso ocorre, por exemplo, em processos seletivos que visam a admissão em faculdades e universidades ou a obtenção de uma bolsa de estudos. Isso propicia um ambiente de competição, uma vez que tais testes acarretam consequências pessoais que dizem respeito a oportunidades e perspectivas de vida dos examinandos. Os examinandos que se submetem a esse tipo de teste têm maior motivação na busca por melhores desempenhos. E além disso, lançam mão de estratégias para maximizar seus desempenhos.

Em contraste, os testes *low-stakes* se aplicam tipicamente para avaliar a eficácia das escolas, fornecer informações sobre o desempenho dos alunos para os professores, formuladores de políticas, pais e outros. Em nível individual, as consequências dos testes *low-stakes* (de baixo risco) são consideradas relativamente menores tanto para professores quanto para os estudantes. No entanto, em nível institucional, as escolas de alto desempenho recebem recompensas de desempenho acadêmico, que podem incluir o reconhecimento público ou recursos a utilizar para a melhoria da escola (Greene et al., 2003; Hamilton et al., 2002; Thomas, 2005). No Brasil, o ENADE seria um exemplo de teste *low-stake*, pois não há competição direta entre os examinandos, nem tampouco haveria consequências resultantes dos seus desempenhos nesse teste.

No que diz respeito às instruções comportamentais dos candidatos a responderem os itens de testes, Whelpley (2014) ressalta que as diferenças na instrução dessas respostas podem, no entanto, proporcionar algumas consequências quando se trata de testes *high-stakes* e *low-stakes*. Os pesquisadores têm a hipótese de que os indivíduos que recebem instruções comportamentais em testes de alto risco são motivados para maximizar suas pontuações respondendo de uma maneira estratégica que eles percebem ser a mais apropriada. Lievens et al. (2009) confirmou essa hipótese e concluiu que, em avaliações de alto risco, os examinandos que recebem instruções de tendência comportamental, escolhem respostas que

acreditam ser estratégicamente melhores.

Tanto os testes *low-stake* como *high-stake*, estão sucetíveis a não-resposta. Há várias razões pelas quais os candidatos optam por omitir as respostas em um teste, como por exemplo, baixo nível de habilidade, baixa motivação, falta de atenção ou tempo e a apenação a respostas erradas. Weeks et al. (2015) destacam que a tendência para omitir respostas também pode ser associada ao tipo de item, formato de resposta, e se a natureza do teste é de *low-stakes* ou *high-stakes* para os participantes. Em um contexto de alto risco, as omissões de respostas podem ser mais consequentes (na qual, por exemplo, os candidatos recebem uma pontuação menor em um vestibular (Yamamoto e Everson, 1997). Para testes *low-stakes* (por exemplo, em levantamentos em grande escala, onde nenhum índice individual é retornado/relatado) como no ENADE, a questão maior é provavelmente a baixa motivação dos candidatos (DeMars, 2000; Wise et al., 2006; Wise e DeMars, 2005).

Evidentemente, dependendo do número de respostas omitidas, a codificação dessas respostas como incorretas provavelmente introduzirá um viés negativo nas estimativas da dificuldade do item (por exemplo, Rose et al., 2010). E a dimensão desse viés aumentará à medida que o número de não-respostas codificadas como incorreta aumenta. Weeks et al. (2015) também ressaltam que, simultaneamente, ocorre diminuição nas estimativas da habilidade do examinando. Por outro lado, esses autores discutem que se as respostas omitidas forem codificadas como não alcançadas e excluídas da estimativa, as estimativas dos parâmetros do item e da habilidade tendem a apresentar viés menor, embora isso pressuponha que não haja razões sistemáticas para as omissões. Em compensação, os erros padrão associados serão maiores em relação às estimativas com respostas omitidas codificadas como incorretas, simplesmente porque o número de respostas incluídas na determinação da estimativa será menor ao ignorar as não-respostas.

Em nossa perspectiva, preferimos não considerar que haja vício nas estimativas, mas que, simplesmente, os parâmetros se diferenciam em função do ambiente. Os parâmetros em um ambiente competitivo simplesmente devem ser diferentes daqueles existentes em outro não competitivo. Desse modo, o próprio desenho do processo proporciona alterações nos objetos de inferência estatística. Por exemplo: em um processo seletivo, o examinador geralmente se restringe a discriminar os mais proficientes dos demais e, portanto, não haveria qualquer interesse em obter estimativas precisas para todo o grupo de examinandos. Além disso, a própria competição, e as estratégias adotadas pessoalmente por cada candi-

dato, poderia proporcionar um efeito de confundimento com a proficiência do indivíduo em um ambiente não competitivo. Na próxima seção apresentaremos um modelo alternativo para não-respostas intencionais, em que a não-resposta é codificada como resposta errada. Nesse caso particular, uma resposta errada penaliza o escore bruto de um candidato e, por isso, considera-se que a não-resposta seja motivada, supostamente, por uma razão estratégica, e que seu processo gerador dependa das características do item e de um traço latente do examinando (propensão).

3.4 Uma modelagem da não-resposta intencional

Considere a situação em que há um grupo de n respondentes submetidos a um teste constituído por I itens. Seja R_{ij} uma variável aleatória indicadora que assume valor 1, se o indivíduo j $(1 \leq j \leq n)$ responde ao item i $(1 \leq i \leq I)$, ou valor 0, se ele não responde a esse item. Defina agora outra variável aleatória indicadora U_{ij}, tal que $U_{ij} = 1$, caso o item i seja respondido corretamente, ou $U_{ij} = 0$, caso o item i não seja respondido corretamente ou tenha sido deixado em branco pelo indivíduo j. A Figura 3.1 evidencia a estrutura hierárquica entre essas variáveis, mostrando os possíveis caminhos para se observar o resultado U_{ij}. Primeiramente, o indivíduo j decide se responde ou não ao item i. Caso ele não responda, atribui-se o resultado $U_{ij} = 0$. Caso contrário, em momento posterior, sua resposta é classificada como certa ou errada. A Tabela 3.1 esboça a forma da distribuição conjunta das variáveis R_{ij} e U_{ij}, cujas probabilidades π_{00}, π_{01} e π_{11} podem ser representadas em termos de probabilidades condicionais como:

$$\pi_{00} = \mathrm{P}(R_{ij} = 0, U_{ij} = 0) = \mathrm{P}(R_{ij} = 0|U_{ij} = 0) \cdot \mathrm{P}(U_{ij} = 0); \tag{3.1}$$

$$\pi_{01} = \mathrm{P}(R_{ij} = 1, U_{ij} = 0) = \mathrm{P}(R_{ij} = 1|U_{ij} = 0) \cdot \mathrm{P}(U_{ij} = 0); \tag{3.2}$$

$$\pi_{11} = \mathrm{P}(R_{ij} = 1, U_{ij} = 1) = \mathrm{P}(R_{ij} = 1|U_{ij} = 1) \cdot \mathrm{P}(U_{ij} = 1) = \mathrm{P}(U_{ij} = 1). \tag{3.3}$$

Observa-se que as probabilidades (3.1), (3.2) e (3.3) dependem dos índices i, j, que foram omitidos para simplificar a notação. Assim, definindo-se $0^0 \equiv 0$, a distribuição de probabilidade conjunta pode ser escrita como

$$\mathrm{P}(R_{ij} = r, U_{ij} = u) = [\mathrm{P}(R_{ij} = r|U_{ij} = u)]^{1-u} \cdot \mathrm{P}(U_{ij} = u), \tag{3.4}$$

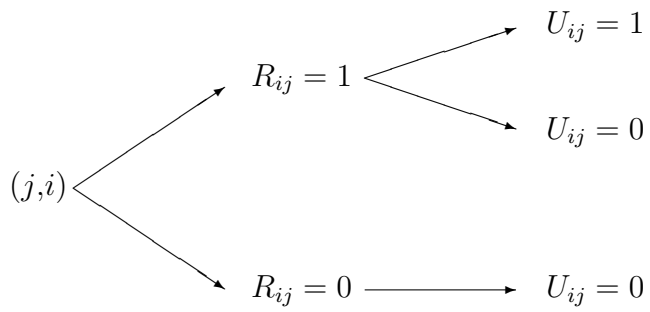


Figura 3.1: Possíveis configurações de respostas

Tabela 3.1 - Distribuição conjunta das variáveis R_{ij} (responde = 1, não responde = 0) e U_{ij} (acerta = 1, não acerta = 0)

		R_{ij}		
		0	1	total
U_{ij}	0	π_{00}	π_{01}	$\pi_{0\cdot}$
	1	0	π_{11}	π_{11}
	total	π_{00}	$\pi_{\cdot 1}$	1

em que $r, u \in \{0, 1\}$.

Seguindo uma prescrição similar àquela feita por Knott et al. (1990) e Albanese e Knott (1992), considere que os termos da Eq. (3.4) sejam expressos por modelos logísticos de dois parâmetros nas formas

$$\mathrm{P}(R_{ij} = r|U_{ij} = 0) = \frac{\exp\left[ra_{1,i}(\theta_{1,j} - b_{1,i})\right]}{1 + \exp\left[a_{1,i}(\theta_{1,j} - b_{1,i})\right]}, \tag{3.5}$$

e

$$\mathrm{P}(U_{ij} = u) = \frac{\exp\left[ua_{2,i}(\theta_{2,j} - b_{2,i})\right]}{1 + \exp\left[a_{2,i}(\theta_{2,j} - b_{2,i})\right]}, \tag{3.6}$$

em que $r, u \in \{0, 1\}$, $a_{1,i}$ e $a_{2,i}$ são parâmetros de discriminação do item i, e $b_{1,i}$ e $b_{2,i}$ são seus parâmetros de dificuldade (Andrade et al., 2000, Reckase; 2009). O traço $\theta_{2,j}$ denota a proficiência do indivíduo j, e $\theta_{1,j}$ representa a propensão de esse indivíduo em responder incorretamente. No modelo unidimensional (3.6), respostas omitidas são tratadas como incorretas. Considerando as Eqs. (3.1), (3.2) e (3.3), as variáveis latentes $\theta_{1,j}$ e $\theta_{2,j}$ se relacionam com as transformações logito

$$\eta_{1,ij} = \ln \frac{\pi_{01}}{\pi_{00}} = a_{1,i}(\theta_{1,j} - b_{1,i}),$$

e

$$\eta_{2,ij} = \ln \frac{\pi_{11}}{\pi_{00} + \pi_{01}} = a_{2,i}(\theta_{2,j} - b_{2,i}).$$

Substituindo-se os modelos logísticos unidimensionais (3.5) e (3.6) em (3.4), obtém-se um modelo bidimensional na forma

$$\begin{aligned}
\mathrm{P}(R_{ij} = r, U_{ij} = u) &= [\mathrm{P}(R_{ij} = r|U_{ij} = u)]^{1-u} \cdot \mathrm{P}(U_{ij} = u) \\
&= \left[\frac{\mathrm{e}^{r\eta_{1,ij}}}{1 + \mathrm{e}^{\eta_{1,ij}}}\right]^{1-u} \cdot \frac{\mathrm{e}^{u\eta_{2,ij}}}{1 + \mathrm{e}^{\eta_{2,ij}}} \\
&= \frac{\mathrm{e}^{r(1-u)(\eta_{1,ij}+u\eta_{2,ij})}}{\left[1 + \mathrm{e}^{\eta_{1,ij}}\right]^{1-u} \cdot \left[1 + \mathrm{e}^{\eta_{2,ij}}\right]} \\
&= \frac{\mathrm{e}^{r(1-u)(\eta_{1,ij}+u\eta_{2,ij})}}{1 + \mathrm{e}^{\eta_{2,ij}} + (1-u)\left[\mathrm{e}^{\eta_{1,ij}} + \mathrm{e}^{\eta_{1,ij}+\eta_{2,ij}}\right]} \\
&= \frac{\mathrm{e}^{r(1-u)(\eta_{1,ij}+u\eta_{2,ij})}}{1 + \mathrm{e}^{\eta_{2,ij}} + (1-u)\mathrm{e}^{\eta_{1,ij}}\left[1 + \mathrm{e}^{\eta_{2,ij}}\right]}.
\end{aligned} \tag{3.7}$$

Esse modelo para $(\theta_{1,j}, \theta_{2,j})$ é do tipo não compensatório (Reckase, 2009), uma vez que uma baixa proficiência $\theta_{2,j}$ não pode ser compensada pela propensão $\theta_{1,j}$. Neste trabalho, para a calibração dos itens da aplicação feita no Capítulo 5, utiliza-se o método de estimação por máxima verossimilhança marginal (Andrade et al., 2000; Reckase, 2009). em que $f(\theta_{1,j}, \theta_{2,j}) = f(\theta_{2,j}|\theta_{1,j}).f(\theta_{1,j})$.

Capítulo 4

A distância de Bhattacharyya

4.1 Introdução

Hambleton et al. (1991) discutem que os modelos da TRI, ao contrário da TCT, são *falsifiable models*. Isso significa que a relação entre as respostas dos examinandos a certo item em função dos parâmetros desse item e das proficiências individuais, pode não ocorrer de acordo com o modelo escolhido. Por isso, é preciso avaliar a aderência de cada item ao modelo proposto, para que os desempenhos dos examinandos ao final do teste sejam adequadamente relacionados com suas respectivas proficiências. Porém, este ainda constitui um dos pontos fracos da TRI. Por exemplo, Neel (2004) mostrou que o uso da estatística quiquadrado na análise de ajuste do item é enganosa na medida em que mostra itens que não se ajustam bem quando se pode considerar que os itens se ajustam bem e vice-versa. Os testes utilizados para avaliar a qualidade de ajuste do modelo em TRI, assim como muitas técnicas estatísticas, são sensíveis ao tamanho da amostra. À medida que o tamanho da amostra aumenta, os testes tornam-se cada vez mais poderosos e mais e mais itens são rejeitados (Neel, 2004). A estatística quiquadrado apresenta alguns problemas, Moore (1986) enumera os motivos: a "arbitrariedade introduzida pela necessidade de escolher células" e "descartar informações dentro das células".

A arbitrariedade das células é um dos principais problemas no uso da estatística quiquadrado. Conforme usado em estatísticas como Q_1 (Eq. 2.5), intervalos iguais são criados ao longo da escala de habilidade e um valor de P_{ij} é selecionado para representar a probabilidade de sucesso relativa a esse intervalo. Porém, esses intervalos são arbitrários. Dessa maneira, a estatística quiquadrado depende dessa arbitrariedade, o que a torna, eventualmente, ineficiente.

Como esses intervalos criados são arbitrários, por exemplo, o seu comprimento. Os intervalos que dão um valor particular do quiquadrado podem dar um valor diferente, se os intervalos fossem de um comprimento diferente.

Um segundo problema é que Q_1 usa o valor do ponto médio de P_{ij} do intervalo na escala θ (outros valores, como o máximo e o mínimo podem ser usados). Ao usar esse valor único para representar todos os pontos no intervalo, as probabilidades possivelmente diferentes ao longo do intervalo são ignoradas. Tratar todos os pontos no intervalo como tendo o mesmo P_{ij} descarta a informação do P_{ij} desigual que existe durante o intervalo devido aos diferentes valores de θ_j. Isso só piora quando os intervalos são combinados, devido ao baixo tamanho da amostra, como é frequentemente feito nos testes de qualidade de ajuste do quiquadrado, porque um único valor de P_{ij} deve então representar um intervalo ainda maior na escala P_{ij}.

Além disso, as diferenças nas proporções observadas podem ser mascaradas pela seleção de intervalos. Isso pode acontecer se a primeira de duas regiões adjacentes na escala de habilidade mostrar uma proporção baixa, enquanto a segunda mostra uma proporção elevada. Se essas duas regiões sucessivas estiverem incluídas no mesmo intervalo, a proporção total pode ser muito próxima do valor apropriado e correto.

Para a avaliação das predições proporcionadas pelo modelo ajustado, por haver diversos casos em que há incidência de baixas frequências esperadas, em lugar da tradicional estatística χ^2, sugere-se a utilização da distância de Bhattacharyya (Schweppe, 1967; Ray, 1989). Inicialmente, introduziremos o coeficiente de Bhattacharyya na Seção 4.2, que representa a similaridade (proximidade) entre duas distribuições como, por exemplo, a esperada e a observada. Com base nesse coeficiente, medidas de distâncias podem ser definidas, como a de Bhattacharyya, a de Hellinger e a de Matusita (4.3). A Seção 4.2 apresenta critérios para a avaliação da proximidade entre duas distribuições discretas, com base em medidas de informação. A Seção 4.3 trata das medidas de divergências entre distribuições, em que se discute a relação entre a distância χ^2 e a de Bhattacharyya. Essa seção também apresenta quatro exemplos de aplicações em testes de hipóteses. Finalmente, a Seção 4.4 apresenta algumas considerações acerca deste capítulo.

4.2 Proximidade entre duas distribuições discretas

Considere que $\boldsymbol{P} = (p_1, \ldots, p_m)$ e $\boldsymbol{Q} = (q_1, \ldots, q_m)$ representem as distribuições discretas de probabilidades, nas quais $p_k \geq 0$ e $q_k \geq 0$ para $k = 1, 2, \ldots, m$, com $\sum p_k = \sum q_k = 1$. Define-se o *coeficiente de Bhattacharyya* entre essas distribuições como

$$\begin{aligned} A(\boldsymbol{P}||\boldsymbol{Q}) &= A(p_1, \ldots, p_m; q_1, \ldots, q_m) \\ &= \sum_{k=1}^{m} (p_k q_k)^{\frac{1}{2}}. \end{aligned} \tag{4.1}$$

Nota-se que a definição (4.1) constitui uma combinação linear das médias geométricas entre p_k e q_k, de modo que $A(\boldsymbol{P}||\boldsymbol{Q})$ representa uma medida de proximidade entre duas populações (ou amostras ou grupos). Em outras palavras, trata-se de uma medida da quantidade de sobreposição entre dois conjuntos de dados.

Do ponto de vista da álgebra linear, $A(\boldsymbol{P}||\boldsymbol{Q})$ é o produto interno entre os vetores $\sqrt{\boldsymbol{P}} = (\sqrt{p_1}, \ldots, \sqrt{p_m})$ e $\sqrt{\boldsymbol{Q}} = (\sqrt{q_1}, \ldots, \sqrt{q_m})$. Como $\sqrt{\boldsymbol{P}}$ e $\sqrt{\boldsymbol{Q}}$ são vetores unitários, já que $||\sqrt{\boldsymbol{P}}||^2 = ||\sqrt{\boldsymbol{Q}}||^2 = \sum p_k = \sum q_k = 1$, tem-se que

$$A(\boldsymbol{P}||\boldsymbol{Q}) = \sqrt{\boldsymbol{P}} \cdot \sqrt{\boldsymbol{Q}} \tag{4.2}$$

$$= ||\sqrt{\boldsymbol{P}}|| \cdot ||\sqrt{\boldsymbol{Q}}|| \cos\varphi \tag{4.3}$$

$$= \cos\varphi, \tag{4.4}$$

ou seja, $A(\boldsymbol{P}||\boldsymbol{Q})$ representa o coseno do ângulo φ formado entre $\sqrt{\boldsymbol{P}}$ e $\sqrt{\boldsymbol{Q}}$. Logo, $A(\boldsymbol{P}||\boldsymbol{Q}) = 1$ significa que $\sqrt{\boldsymbol{P}}$ e $\sqrt{\boldsymbol{Q}}$ são vetores paralelos e, consequentemente, as distribuições $\boldsymbol{P}$ e $\boldsymbol{Q}$ se sobrepõem mutuamente (similares). Por outro lado, se $A(\boldsymbol{P}||\boldsymbol{Q}) = 0$, tais vetores são perpendiculares. Assim, $A(\boldsymbol{P}||\boldsymbol{Q})$ representa uma medida de similaridade entre duas distribuições.

Entre as propriedades do coeficiente de Bhattacharyya, tem-se a não-negatividade, pois $A(\boldsymbol{P}||\boldsymbol{Q}) \geq 0$. A igualdade ocorre se $p_k q_k = 0$ para todo k. Além disso, trata-se de uma medida simétrica com respeito aos pares $\{(p_k, q_k)\}$, ou seja, $A(p_1, \ldots, p_m; q_1, \ldots, q_m) = A(p_{k_1}, \ldots, p_{k_m}; q_{k_1}, \ldots, q_{k_m})$, em que $\{k_1, \ldots, k_m\}$ representa uma permutação abritrária do conjunto de índices $\{1, 2, \ldots, m\}$. Por construção, ele é simétrico, $A(\boldsymbol{P}||\boldsymbol{Q}) = A(\boldsymbol{Q}||\boldsymbol{P})$ e, finalmente, $A(\boldsymbol{P}||\boldsymbol{Q}) \leq A(\boldsymbol{P}||\boldsymbol{P}) = A(\boldsymbol{Q}||\boldsymbol{Q}) = 1$.

4.3 Distâncias

A divergência quiquadrado da distribuição $\boldsymbol{P}$ relativamente a $\boldsymbol{Q}$ é definida como

$$\begin{aligned}\chi^2(\boldsymbol{P}||\boldsymbol{Q}) &= \sum_{k=1}^{m} \frac{(p_k - q_k)^2}{q_k} \\ &= \sum_{k=1}^{m} q_k \left(\frac{p_k}{q_k} - 1\right)^2.\end{aligned}$$

Quanto maior for a proximidade entre $\boldsymbol{P}$ e $\boldsymbol{Q}$, menor será o valor χ^2, e vice-versa. Como alternativas a essa medida outras podem ser sugeridas. Em particular, enumeramos aqui algumas distâncias encontradas na literatura que se relacionam com a medida $A(\boldsymbol{P}||\boldsymbol{Q})$. Por exemplo, define-se a distância de Bhattacharyya entre $\boldsymbol{P}$ e $\boldsymbol{Q}$ como

$$D_B(\boldsymbol{P}||\boldsymbol{Q}) = -\ln A(\boldsymbol{P}||\boldsymbol{Q}). \tag{4.5}$$

Essa distância mede a similaridade entre duas distribuições. Enquanto o coeficiente de Bhattacharyya mede a proximidade relativa entre dois conjuntos de dados, a distância de Bhattacharyya mede a separabilidade entre eles. Como $0 \leq A(\boldsymbol{P}||\boldsymbol{Q}) \leq 1$, tem-se que $0 \leq D(\boldsymbol{P}||\boldsymbol{Q}) < +\infty$. Entre as outras distâncias, a de Hellinger se define como

$$D_H(\boldsymbol{P}||\boldsymbol{Q}) = \sqrt{1 - A(\boldsymbol{P}||\boldsymbol{Q})}, \tag{4.6}$$

e a de Matusita é dada por

$$\begin{aligned}D_M(\boldsymbol{P}||\boldsymbol{Q}) &= \sum_{k=1}^{m} \left(\sqrt{p_k} - \sqrt{q_k}\right)^2 \\ &= 2 - 2A(\boldsymbol{P}||\boldsymbol{Q}).\end{aligned}$$

Como as distâncias de Hellinger e a de Matusita se relacionam linearmente com o coeficiente $A(\boldsymbol{P}||\boldsymbol{Q})$, elas também são medidas de "proximidade" entre duas distribuições, possuindo propriedades análogas. Por isso, neste trabalho restringimos nossa atenção à distância de Bhattacharyya.

4.3.1 Relação entre χ^2 e D_B

Embora as medidas χ^2 e D_B sejam funcionalmente distintas, encontramos a seguinte equivalência entre elas para a situação em que $\boldsymbol{P} \approx \boldsymbol{Q}$.

Inicialmente, considere o caso $m = 2$, tal que $p_1 = q_1 + \varepsilon$, em que ε corresponde a uma variação suficientemente pequena. Havendo apenas duas classes, a divergência quiquadrado pode ser escrita como

$$\begin{aligned}
\chi^2(\boldsymbol{P}||\boldsymbol{Q}) &= \sum_{k=1}^{2} \frac{(p_k - q_k)^2}{q_k} \qquad (4.7)\\
&= \frac{(p_1 - q_1)^2}{q_1} + \frac{(1 - p_1 - 1 + q_1)^2}{1 - q_1}\\
&= \frac{(p_1 - q_1)^2}{q_1} + \frac{(p_1 - q_1)^2}{1 - q_1}\\
&= (p_1 - q_1)^2 \left(\frac{1}{q_1} + \frac{1}{1 - q_1}\right)\\
&= \frac{(p_1 - q_1)^2}{q_1(1 - q_1)}\\
&= \frac{\varepsilon^2}{q_1(1 - q_1)}. \qquad (4.8)
\end{aligned}$$

O *coeficiente de Bhattacharyya* pode ser escrito como

$$\begin{aligned}
A(\boldsymbol{P}||\boldsymbol{Q}) &= \sum_{k=1}^{2} (p_k q_k)^{\frac{1}{2}} \qquad (4.9)\\
&= \sqrt{p_1 q_1} + \sqrt{(1 - p_1)(1 - q_1)}\\
&= \sqrt{(q_1 + \varepsilon) q_1} + \sqrt{(1 - q_1 - \varepsilon)(1 - q_1)}\\
&= \sqrt{\left(1 + \frac{\varepsilon}{q_1}\right) q_1^2} + \sqrt{\left(1 - \frac{\varepsilon}{1 - q_1}\right) (1 - q_1)^2}\\
&= q_1 \sqrt{\left(1 + \frac{\varepsilon}{q_1}\right)} + (1 - q_1) \sqrt{\left(1 - \frac{\varepsilon}{1 - q_1}\right)}. \qquad (4.10)
\end{aligned}$$

Considerando a aproximação de Taylor de terceira ordem $\sqrt{1 + x} \approx 1 + \frac{x}{2} - \frac{x^2}{8} + \frac{x^3}{16}$ em torno de $x = 0$ temos

$$\begin{aligned}
A(\boldsymbol{P}||\boldsymbol{Q}) &\approx q_1 \left(1 + \frac{\varepsilon}{2q_1} - \frac{\varepsilon^2}{8q_1^2} + \frac{\varepsilon^3}{16q_1^3}\right) + (1 - q_1) \left(1 - \frac{\varepsilon}{2(1 - q_1)} - \frac{\varepsilon^2}{8(1 - q_1)^2} - \frac{\varepsilon^3}{16(1 - q_1)^3}\right)\\
&\approx 1 - \frac{\varepsilon^2}{8q_1(1 - q_1)} + \frac{\varepsilon^3(1 - 2q_1)}{16q_1^2(1 - q_1)^2} \qquad (4.11)\\
&\approx 1 - \frac{1}{8}\chi^2(\boldsymbol{P}||\boldsymbol{Q}) + \frac{\varepsilon(1 - 2q_1)}{16q_1(1 - q_1)}\chi^2(\boldsymbol{P}||\boldsymbol{Q})\\
&\approx 1 - \frac{1}{8}\left[1 - \frac{\varepsilon(1 - 2q_1)}{2q_1(1 - q_1)}\right]\chi^2(\boldsymbol{P}||\boldsymbol{Q}). \qquad (4.12)
\end{aligned}$$

Assim, para $q_1 = 1/2$ ou $|\varepsilon| \downarrow 0$ haveria uma aproximação linear na forma $A(\boldsymbol{P}||\boldsymbol{Q}) \approx 1 - \chi^2(\boldsymbol{P}||\boldsymbol{Q})/8$. A Seção 4.3.2.1 exemplifica uma situação amostral em que o erro ε pode ser controlado com base no tamanho da amostra.

De modo similar, considere agora o caso geral, em que $p_k = q_k + \varepsilon_k$ para $k = 1, \ldots, m$, de modo que $\sum_{k=1}^{m} \varepsilon_k = 0$. Agora, a divergência quiquadrado da distribuição pode ser escrita como

$$\chi^2(\boldsymbol{P}||\boldsymbol{Q}) = \sum_{k=1}^{m} \frac{\varepsilon_k^2}{q_k},$$

e o *coeficiente de Bhattacharyya* pode ser escrito como

$$\begin{aligned} A(\boldsymbol{P}||\boldsymbol{Q}) &= \sum_{k=1}^{m} (p_k q_k)^{\frac{1}{2}} \\ &= \sum_{k=1}^{m} \sqrt{(q_k + \varepsilon_k) q_k} \\ &= \sum_{k=1}^{m} \sqrt{\left(1 + \frac{\varepsilon_k}{q_k}\right) q_k^2} \\ &= \sum_{k=1}^{m} q_k \sqrt{\left(1 + \frac{\varepsilon_k}{q_k}\right)}. \end{aligned}$$

Tomando-se novamente as aproximações de terceira ordem para cada termo, teremos

$$\begin{aligned} A(\boldsymbol{P}||\boldsymbol{Q}) &\approx \sum_{k=1}^{m} q_k \left(1 + \frac{\varepsilon_k}{2q_k} - \frac{\varepsilon_k^2}{8q_k^2} + \frac{\varepsilon_k^3}{16q_k^3}\right) \\ &= \sum_{k=1}^{m} q_k + \sum_{k=1}^{m} \frac{\varepsilon_k}{2} - \sum_{k=1}^{m} \frac{\varepsilon_k^2}{8q_k} + \sum_{k=1}^{m} \frac{\varepsilon_k^3}{16q_k^2} \\ &= 1 - \frac{1}{8}\chi^2(\boldsymbol{P}||\boldsymbol{Q}) + \frac{1}{16} \sum_{k=1}^{m} \frac{\varepsilon_k^3}{q_k^2}. \end{aligned} \tag{4.13}$$

Como $A(\boldsymbol{P}||\boldsymbol{Q}) > 0$, observe que sua aproximação de terceira ordem mostrada em (4.13) requer $\chi^2(\boldsymbol{P}||\boldsymbol{Q}) < 8 + \frac{1}{2}\sum_{k=1}^{m} \frac{\varepsilon_k^3}{q_k^2}$. Se, por exemplo, $q_k = 1/m$ (distribuição uniforme), e se os erros ε_k forem simétricos, de modo que $\sum_{k=1}^{m} \varepsilon_k^3 = 0$, então o último termo em (4.13) seria nulo, o que proporcionaria uma aproximação linear na forma $A(\boldsymbol{P}||\boldsymbol{Q}) \approx 1 - \chi^2(\boldsymbol{P}||\boldsymbol{Q})/8$. Situações amostrais, em que $|\varepsilon_k| \downarrow 0$ serão ilustradas na Seção 4.3.2.

Finalmente, nessas mesmas condições, considerando a aproximação de primeira ordem $\ln(1 + x) \approx x$, temos

$$D_B(\boldsymbol{P}||\boldsymbol{Q}) = -\ln A(\boldsymbol{P}||\boldsymbol{Q}) \approx \frac{1}{8}\chi^2(\boldsymbol{P}||\boldsymbol{Q}) - \frac{1}{16} \sum_{k=1}^{m} \frac{\varepsilon_k^3}{q_k^2}, \tag{4.14}$$

ou, dependendo da situação,

$$D_B(\boldsymbol{P}||\boldsymbol{Q}) = -\ln A(\boldsymbol{P}||\boldsymbol{Q}) \approx \frac{1}{8}\chi^2(\boldsymbol{P}||\boldsymbol{Q}). \tag{4.15}$$

Para fins de inferência estatística, investigaremos a seguir como as aproximações (4.13), (4.14) e (4.15) poderiam ser aplicadas a testes de hipóteses que dizem respeito a comparações entre duas distribuições.

4.3.2 Aplicações em testes de hipóteses

4.3.2.1 Caso 1

Seja $U_1, \ldots, U_n$ uma amostra aleatória simples retirada de uma população Bernoulli com probabilidade de sucesso p_1. Considerando o estimador de máxima verossimilhança de p_1 dado por $\hat{p_1} = (U_1 + \ldots + U_n)/n$, deseja-se testar a hipótese nula $H_0 : p_1 = q_1$ contra $H_1 : p_1 \neq q_1$, na qual $0 < q_1 < 1$ é um valor hipotético conhecido. Como alternativa às tradicionais estatísticas Z e χ^2 aplicáveis para esse caso, estudaremos as propriedades da estatística $D_B(\boldsymbol{P}||\boldsymbol{Q})$, em que, neste caso particular, $\boldsymbol{Q} = (q_1, 1 - q_1)$ representa a distribuição hipotética H_0, e $\boldsymbol{P} = (\hat{p}_1, 1 - \hat{p}_1)$ é a distribuição empírica.

Primeiramente efetuaremos o estudo considerando uma distribuição empírica $\boldsymbol{P}$ obtida sob H_0, e em seguida, para efetuarmos comparações entre as estatísticas, trataremos $\boldsymbol{P}$ sob H_1.

Sob H_0, para n suficientemente grande sabe-se que $Z = \frac{\hat{p_1} - q_1}{\sqrt{q_1(1-q1)/n}} \sim \mathrm{N}(0, 1)$, ou seja, tem-se imediatamente

$$\hat{p_1} = q_1 + \varepsilon_1,$$

em que $\varepsilon_1 \sim \mathrm{N}(0, q_1(1 - q1)/n)$. Sob a hipótese nula, a equação (4.8) multiplicada por n assume a seguinte forma particular:

$$\begin{aligned} n\chi^2(\boldsymbol{P}||\boldsymbol{Q}) &= \frac{\varepsilon^2}{q_1(1-q_1)/n} = \frac{n(\hat{p_1} - q_1)^2}{q_1(1-q_1)} = \frac{n \cdot n^2(\hat{p_1} - q_1)^2}{nq_1n(1-q_1)} \qquad (4.16) \\ &= n \cdot \frac{(n\hat{p_1} - nq_1)^2}{nq_1n(1-q_1)} = n \cdot \frac{(o_1 - e_1)^2}{e_1e_2} \\ &= \frac{n}{n}\left[\frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2}\right] \\ &= \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2}, \qquad (4.17) \end{aligned}$$

na qual $e_1 = nq_1$ e $e_2 = n(1 - q_1)$ representam as frequências esperadas, e $o_1 = n\hat{p}_1$ e $o_2 = n(1 - \hat{p}_1)$ são as frequências observadas correspondentes. Essa forma particular é a estatística χ^2 de Pearson para uma tabela 2×1. Para n suficientemente grande ela segue distribuição χ^2 com 1 grau de liberdade, pois $n\varepsilon_1^2/\{q_1(1 - q_1)\} = Z^2$.

De (4.9), o coeficiente de Bhattacharyya pode ser expresso como

$$A(\boldsymbol{P}||\boldsymbol{Q}) = \sum_{k=1}^{2} (\hat{p_k} q_k)^{\frac{1}{2}}. \tag{4.18}$$

Agora, com respeito a aproximações, a partir de (4.11), (4.15) e considerando $\varepsilon_1 = Z\sqrt{q_1(1-q_1)/n}$, podemos escrever

$$D_B(\boldsymbol{P}||\boldsymbol{Q}) = -\ln A(\boldsymbol{P}||\boldsymbol{Q}) \approx \frac{\varepsilon_1^2}{8q_1(1-q_1)} - \frac{\varepsilon_1^3(1-2q_1)}{16q_1^2(1-q_1)^2} \tag{4.19}$$

$$\approx \frac{Z^2}{8n} - \frac{Z^3(1-2q_1)}{16n\sqrt{nq_1(1-q_1)}}. \tag{4.20}$$

Multiplicando-se (4.20) por n, podemos escrever

$$8nD_B(\boldsymbol{P}||\boldsymbol{Q}) \sim Z^2 - \frac{1-2q_1}{2\sqrt{nq_1(1-q_1)}} Z^3. \tag{4.21}$$

Verifica-se, portanto, que a contribuição do termo cúbico diminui à medida que $q_1 \to 0,5$ ou n aumenta. Nessa situação, podemos considerar a aproximação $8nD_B(\boldsymbol{P}||\boldsymbol{Q}) \sim \chi^2_{(1)}$. Por outro lado, a presença do termo cúbico tende a se destacar, por exemplo, à medida que $q_1 \downarrow 0$ ou $q_1 \uparrow 1$.

Os valores críticos da estatística $8nD_B(\boldsymbol{P}||\boldsymbol{Q})$ podem ser obtidos com base na sua distribuição amostral empírica gerada pelo método de Monte de Carlo. Para exemplificar numericamente, considere três testes com hipóteses nulas $q_1 = 0,3$, 0,5 e 0,99, sob tamanhos amostrais iguais a $n = 500$ e 5000.

Com esses valores de q_1 e n, tomando-se 10^7 realizações da variável Z, e aplicando-as na (4.21), observamos os valores críticos empíricos relativos aos níveis de significância do teste $\alpha = 0,1\%$, 1% e 5% que se encontram na Tabela 4.1. Para o caso assintótico, os valores críticos foram obtidos diretamente da distribuição $\chi^2_{(1)}$.

Em outra simulação, desta vez com base em 5 mil replicações de uma distribuição binomial com parâmetros n e q_1, utilizando agora (4.17) e (4.18), a Figura 4.1 mostra as dispersões entre as estatísticas $n\chi^2(\boldsymbol{P}||\boldsymbol{Q})$ e $8nD_B(\boldsymbol{P}||\boldsymbol{Q})$ obtidas sob H_0, na qual, como se espera, a aproximação linear $8nD_B(\boldsymbol{P}||\boldsymbol{Q}) \approx n\chi^2(\boldsymbol{P}||\boldsymbol{Q})$ tende a melhorar à medida que n aumenta. Assintoticamente, sob H_0, ambas as estatísticas seguem distribuição χ^2 com 1 grau de liberdade. As linhas pontilhadas na Figura 4.1 indicam o valor crítico $6,63$ com base na distribuição $\chi^2_{(1)}$ relativo ao nível $\alpha = 1\%$ (Tabela 4.1). Observe que, para os casos $q_1 = 0,3$ e 0,5, a aproximação linear tende a piorar na região de rejeição de H_0.

Tabela 4.1 - Caso 1. Valores críticos empíricos c_α correspondentes ao teste $H_0 : p_1 = q_1$ *versus* $H_1 : p_1 \neq q_1$, para $q_1 = 0,3$, 0,5 e 0,99, com tamanhos amostrais iguais a $n = 500$, 5000, e níveis de significância $\alpha = 0,1\%$, 1% e 5% obtidos com base em 10^7 realizações da variável Z aplicadas em (4.21). Para o caso assintótico, $8nD_B(\boldsymbol{P}||\boldsymbol{Q}) \sim \chi^2_{(1)}$.

n	q_1								
	0,3			0,5			0,99		
	0,1%	1%	5%	0,1%	1%	5%	0,1%	1%	5%
500	10,89	6,65	3,84	10,78	6,63	3,84	16,05	8,17	3,69
5000	10,84	6,63	3,84	10,80	6,63	3,84	11,65	6,72	3,82
$+\infty$	10,83	6,63	3,84	10,83	6,63	3,84	10,83	6,63	3,84

Com base nos resultados empíricos mostrados na Tabela 4.1, H_0 é rejeitada caso a estatística em questão for superior a c_α (Tabela 4.1). Por exemplo, para o nível $\alpha = 1\%$, $q_1 = 0,3$ e $n = 500$, a hipótese nula é rejeitada se $8nD_B(\boldsymbol{P}||\boldsymbol{Q}) > 6,65$. A Figura 4.2 ilustra as funções de poder do teste, B(p), obtidas com base nesse critério de decisão. À medida que a amostra aumenta,ambas as estatísticas tendem a apresentar poderes equivalentes. Dependendo da região, porém, uma estatística proporciona poder do teste superior a outra. Por exemplo, para $n = 500$ e $p > 0,99$, a utilização da estatística $8nD_B(\boldsymbol{P}||\boldsymbol{Q})$ oferece um teste com maior poder, mas para $p < 0,99$, observa-se o contrário.

4.3.2.2 Caso 2

Seja $U_1, \ldots, U_n$ uma amostra aleatória simples retirada de uma distribuição discreta tal que $p_1 = \mathrm{P}(U_i = 1) > 0$, $p_2 = \mathrm{P}(U_i = 2) > 0$ e $p_3 = \mathrm{P}(U_i = 3) = 1 - p_1 - p_2$. Considerando a variável indicadora $I_{ki} = 1$, se $U_i = k$, e $I_{ki} = 0$, se $U_i \neq k$, e que $I_{1i} + I_{2i} + I_{3i} = 1$, os estimadores de máxima verossimilhança para as probabilidades p_1, p_2 e p_3 são, respectivamente, $\hat{p_1} = (I_{11} + \ldots + I_{1n})/n$, $\hat{p_2} = (I_{21} + \ldots + I_{2n})/n$ e $\hat{p_3} = 1 - \hat{p_1} - \hat{p_2}$. Deseja-se testar a hipótese nula $H_0 : \boldsymbol{P} = \boldsymbol{Q}$ contra a alternativa $H_1 : \boldsymbol{P} \neq \boldsymbol{Q}$, em que $\boldsymbol{P} = (p_1, p_2, p_3)$ e $\boldsymbol{Q} = (q_1, q_2, q_3)$ representa uma distribuição hipotética objeto de comparação.

Considere agora a estatística $D_B(\hat{\boldsymbol{P}}||\boldsymbol{Q})$. Sua distribuição amostral pode ser obtida empiricamente com base em (4.14), para $k = 3$, em que $\varepsilon_k/\sqrt{q_k} \sim \mathrm{N}(0, (1 - q_k)/n)$. A covariância entre as variáveis aleatórias ε_1 e ε_2 é dada por

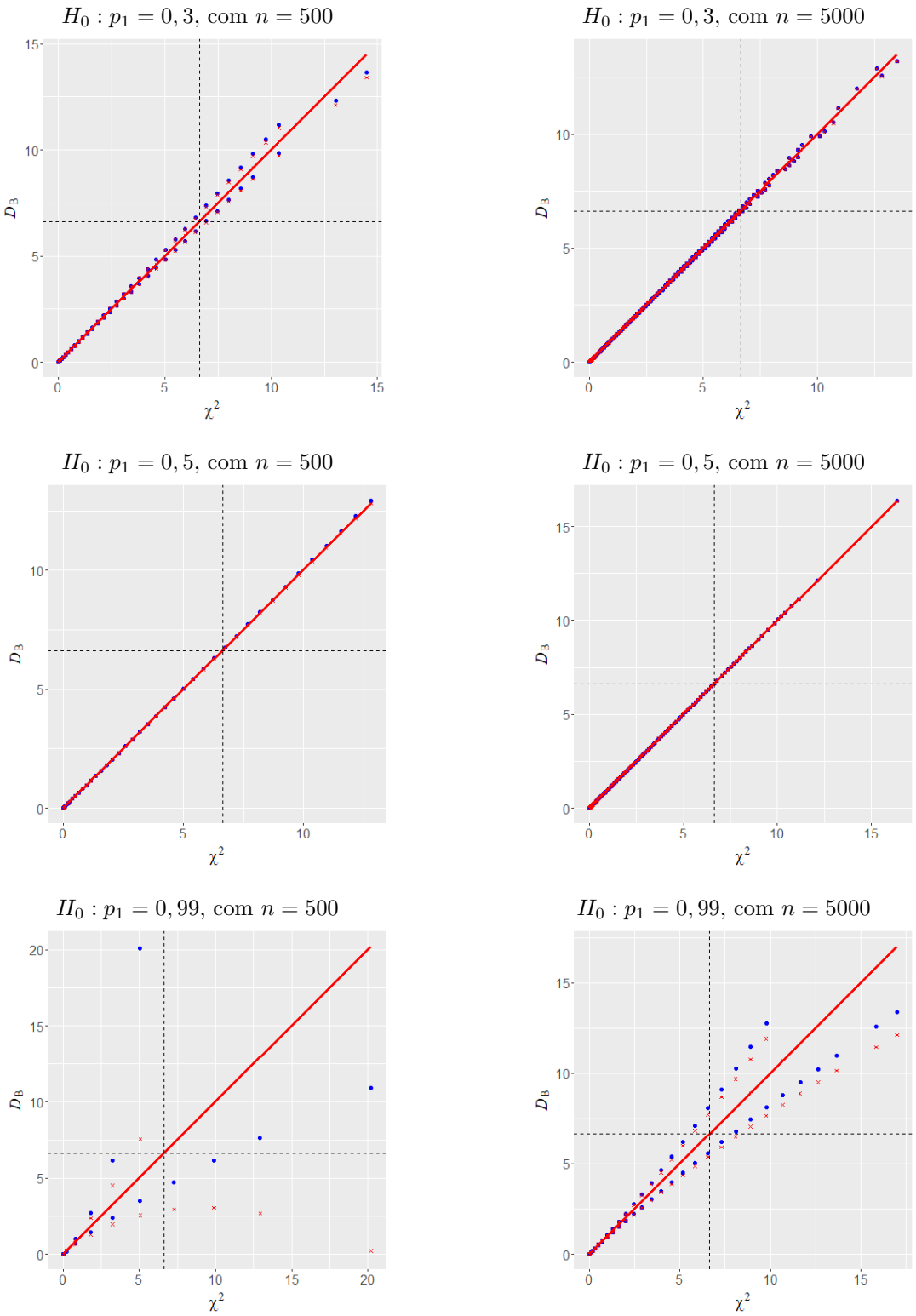


Figura 4.1: Dispersões entre $n\chi^2(\boldsymbol{P}||\boldsymbol{Q})$ e $8nD_B(\boldsymbol{P}||\boldsymbol{Q})$ (pontos). Cinco mil estatísticas obtidas sob a hipótese nula do teste correspondente. A linha sólida representa a reta $8nD_B(\boldsymbol{P}||\boldsymbol{Q}) = n\chi^2(\boldsymbol{P}||\boldsymbol{Q})$, e $\times$ representa a aproximação $8nD_B(\boldsymbol{P}||\boldsymbol{Q}) \approx \left[1 - 0,5Z(1-2q_1)/\sqrt{nq_1(1-q_1)}\right]\chi^2_{(1)}$. As linhas pontilhadas indicam o valor crítico 6,63 relativo ao nível $\alpha = 1\%$ com base na distribuição $\chi^2_{(1)}$.

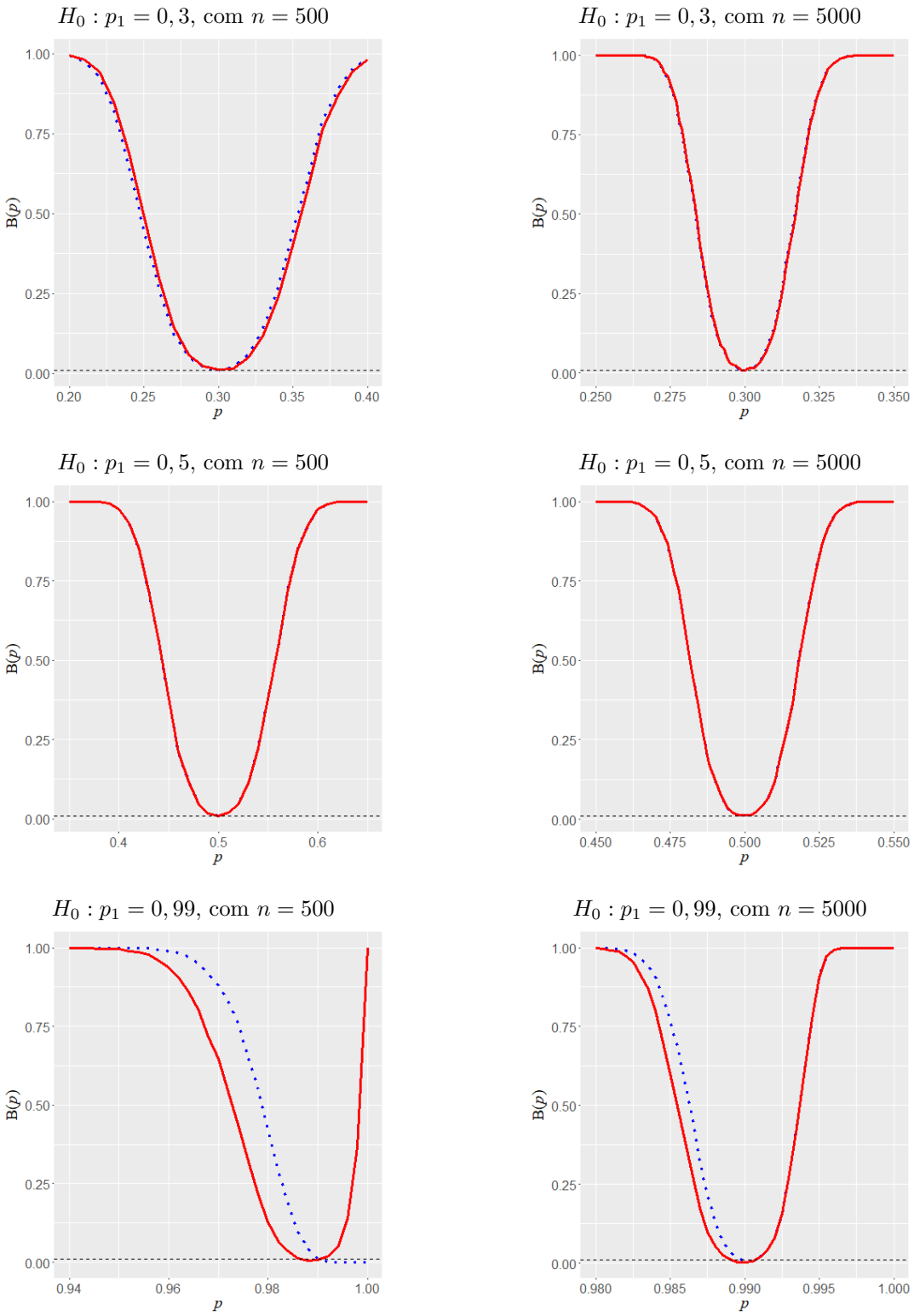


Figura 4.2: Poderes dos testes, $B(p)$, com nível de significância $\alpha = 1\%$, mediante 5 mil replicações das estatísticas $n\chi^2(\boldsymbol{P}||\boldsymbol{Q})$ (linha pontilhada) e $8nD_B(\boldsymbol{P}||\boldsymbol{Q})$ (linha contínua).

$$\begin{aligned}
\mathrm{Cov}[\varepsilon_1, \varepsilon_2] = \mathrm{E}[\varepsilon_1, \varepsilon_2] &= \mathrm{E}[(\hat{p_1} - q_1)(\hat{p_2} - q_2)] = \mathrm{E}[\hat{p_1}\hat{p_2}] - q_1 q_2 \\
&= \frac{1}{n^2}\mathrm{E}\left[\sum_{i=1}^{n}\sum_{i'=1}^{n} I_{1i} I_{2i'}\right] - q_1 q_2 \\
&= \frac{1}{n^2}\left\{\sum_{i=i'=1}^{n} \mathrm{E}\left[I_{1i} I_{2i}\right] + \sum\sum_{i \neq i'} \mathrm{E}\left[I_{1i} I_{2i'}\right]\right\} - q_1 q_2 \\
&= 0 + \frac{1}{n^2}\sum\sum_{i \neq i'} \mathrm{E}[I_{1i}]\mathrm{E}[I_{2i'}] - q_1 q_2 \\
&= \frac{1}{n^2}\sum\sum_{i \neq i'} q_1 q_2 - q_1 q_2 \\
&= \frac{n(n-1)}{n^2} q_1 q_2 - q_1 q_2 = -\frac{q_1 q_2}{n}
\end{aligned} \tag{4.22}$$

Consequentemente,

$$\mathrm{Cov}\left[\frac{\varepsilon_1}{\sqrt{q_1}}, \frac{\varepsilon_2}{\sqrt{q_2}}\right] = \frac{\mathrm{Cov}[\varepsilon_1, \varepsilon_2]}{\sqrt{q_1 q_2}} = -\frac{\sqrt{q_1 q_2}}{n},$$

de modo que

$$\rho = \mathrm{Corr}\left[\frac{\varepsilon_1}{\sqrt{q_1}}, \frac{\varepsilon_2}{\sqrt{q_2}}\right] = -\sqrt{\frac{q_1 q_2}{(1-q_1)(1-q_2)}}.$$

Logo, conjuntamente, $(\varepsilon_1/\sqrt{q_1}, \varepsilon_2/\sqrt{q_2})$ segue uma distribuição normal bivariada cuja função de densidade é dada por

$$f(\varepsilon_1/\sqrt{q_1}, \varepsilon_2/\sqrt{q_2}) = \frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)}\mathrm{e}^{-Z^2/2},$$

em que $\sigma_k = \sqrt{(1-q_k)/n}$, e

$$
\begin{aligned}
Z^2 &= \frac{1}{1-\rho^2}\left\{\frac{\varepsilon_1^2}{q_1\sigma_1^2} - 2\frac{\rho\varepsilon_1\varepsilon_2}{\sigma_1\sigma_2\sqrt{q_1q_2}} + \frac{\varepsilon_2^2}{q_2\sigma_2^2}\right\} \\
&= \frac{(1-q_1)(1-q_2)}{(1-q_1)(1-q_2)-q_1q_2}\left\{\frac{n\varepsilon_1^2}{q_1(1-q_1)} + \frac{2n\sqrt{q_1q_2}\varepsilon_1\varepsilon_2}{(1-q_1)(1-q_2)\sqrt{q_1q_2}} + \frac{n\varepsilon_2^2}{q_2(1-q_2)}\right\} \\
&= \frac{(1-q_1)(1-q_2)}{1-q_1-q_2+q_1q_2-q_1q_2}\left\{\frac{n\varepsilon_1^2}{q_1(1-q_1)} + \frac{2n\varepsilon_1\varepsilon_2}{(1-q_1)(1-q_2)} + \frac{n\varepsilon_2^2}{q_2(1-q_2)}\right\} \\
&= \frac{n(1-q_1)(1-q_2)}{1-q_1-q_2}\left\{\frac{\varepsilon_1^2}{q_1(1-q_1)} + \frac{2\varepsilon_1\varepsilon_2}{(1-q_1)(1-q_2)} + \frac{\varepsilon_2^2}{q_2(1-q_2)}\right\} \\
&= n\left\{\frac{\varepsilon_1^2(1-q_2)}{q_1q_3} + \frac{2\varepsilon_1\varepsilon_2}{q_3} + \frac{\varepsilon_2^2(1-q_1)}{q_2q_3}\right\} \\
&= n\left\{\frac{\varepsilon_1^2(q_3+q_1)}{q_1q_3} + \frac{2\varepsilon_1\varepsilon_2}{q_3} + \frac{\varepsilon_2^2(q_3+q_2)}{q_2q_3}\right\} \\
&= n\left\{\frac{\varepsilon_1^2}{q_1} + \frac{\varepsilon_1^2}{q_3} + \frac{2\varepsilon_1\varepsilon_2}{q_3} + \frac{\varepsilon_2^2}{q_2} + \frac{\varepsilon_2^2}{q_3}\right\} \\
&= n\left\{\frac{\varepsilon_1^2}{q_1} + \frac{\varepsilon_2^2}{q_2} + \frac{\varepsilon_1^2+2\varepsilon_1\varepsilon_2+\varepsilon_2^2}{q_3}\right\} \\
&= n\left\{\frac{\varepsilon_1^2}{q_1} + \frac{\varepsilon_2^2}{q_2} + \frac{\varepsilon_3^2}{q_3}\right\} \\
&= n\chi^2(\boldsymbol{P}||\boldsymbol{Q}).
\end{aligned}
$$

Assim, com base em (4.14), a distribuição amostral de $8nD_B(\hat{\boldsymbol{P}}||\boldsymbol{Q})$ pode ser gerada aproximadamente como

$$
8nD_B(\hat{\boldsymbol{P}}||\boldsymbol{Q}) \approx n\sum_{k=1}^{3}\frac{\varepsilon_k^2}{q_k} - \frac{n}{2}\sum_{k=1}^{3}\frac{\varepsilon_k^3}{q_k^2}, \tag{4.23}
$$

em que $(\varepsilon_1, \varepsilon_2)$ segue uma distribuição normal bivariada com vetor de médias nulo e matriz de covariâncias

$$
\Sigma = \frac{1}{n}\begin{bmatrix} q_1(1-q_1) & -q_1q_2 \\ -q_1q_2 & q_2(1-q_2) \end{bmatrix},
$$

com $q_3 = 1-q_1-q_2$ e $\varepsilon_3 = -(\varepsilon_1+\varepsilon_2)$.

Novamente, utilizaremos o método de Monte de Carlo para obter alguns valores críticos para a estatística $8nD_B(\boldsymbol{P}||\boldsymbol{Q})$. Considere o teste $H_0 : \boldsymbol{P} = \boldsymbol{Q}$ contra $H_1 : \boldsymbol{P} \neq \boldsymbol{Q}$, para os casos $\boldsymbol{Q} = (0,4; 0,3; 0,3)$, $(0,7; 0,2; 0,1)$ e $(0,8; 0,15; 0,05)$ com tamanhos amostrais iguais a $n = 500$ e 5000. Com esses valores de q_1 e n, tomando-se 10^7 realizações do vetor aleatório $(\varepsilon_1, \varepsilon_2)$, aplicando-as em (4.23), registramos os valores críticos empíricos para os

Tabela 4.2 - Caso 2. Valores críticos empíricos c_α correspondentes ao teste $H_0 : \boldsymbol{P} = \boldsymbol{Q}$ contra $H_1 : \boldsymbol{P} \neq \boldsymbol{Q}$, com tamanhos amostrais iguais a $n = 500$, 5000, e níveis de significância $\alpha = 0,1\%$, 1% e 5% obtidos com base em 10^7 realizações do vetor aleatório $(\varepsilon_1, \varepsilon_2)$ aplicadas em (4.21). Para o caso assintótico, $8nD_B(\boldsymbol{P}||\boldsymbol{Q}) \sim \chi^2_{(2)}$.

n	$\boldsymbol{Q}$								
	$(0,4; 0,3; 0,3)$			$(0,7; 0,2; 0,1)$			$(0,8; 0,15; 0,05)$		
	0,1%	1%	5%	0,1%	1%	5%	0,1%	1%	5%
500	13,87	9,22	5,99	14,33	9,30	5,99	14,95	9,43	5,99
5000	13,83	9,21	5,99	13,82	9,21	5,99	13,95	9,22	5,99
$+\infty$	13,82	9,21	5,99	13,82	9,21	5,99	13,82	9,21	5,99

níveis de significância $\alpha = 0,1\%$, 1% e 5% (Tabela 4.2). Para o caso assintótico, os valores críticos foram obtidos diretamente da distribuição $\chi^2_{(2)}$.

Como no caso anterior, estatísticas $n\chi^2(\boldsymbol{P}||\boldsymbol{Q})$ e $8nD_B(\boldsymbol{P}||\boldsymbol{Q})$ foram obtidas sob H_0 com base em 5 mil realizações de uma variável aleatória multinomial com parâmetros n e $\boldsymbol{Q} = (q_1, q_2, q_3)$. A Figura 4.1 mostra suas dispersões para alguns casos. As linhas pontilhadas na Figura 4.3 indicam o valor crítico $9,21$ com base na distribuição assintótica $\chi^2_{(2)}$ relativo ao nível $\alpha = 1\%$ (Tabela 4.1). Observe que a aproximação linear tende a melhorar à medida que n aumenta, sendo que, em geral, ela se mostra pior na região de rejeição de H_0.

Considerando as probabilidades $p = p_1 + \delta$ e $r = p_2 - \delta$, para $|\delta| < 0,15$, a Figura 4.4 ilustra as funções de poder do teste, B(p), considerando-se os valores críticos empíricos apresentados na Tabela 4.2. À medida que a amostra aumenta, ambas as estatísticas tendem a apresentar poderes equivalentes. Novamente, dependendo da região, uma estatística proporciona poder do teste superior a outra. Por exemplo, para $n = 500$ e $p < 0,8$, a estatística $8nD_B(\boldsymbol{P}||\boldsymbol{Q})$ oferece poder levemente superior à estatística $\chi^2_{(2)}$, mas o contrário ocorre para $p > 0,08$.

4.3.2.3 Caso 3

Considere agora uma variação do Caso 1, em que a amostra não seja identicamente distribuída. Seja $U_1, \ldots, U_n$ uma sequência de variáveis aleatórias independentes retiradas de uma população Bernoulli com probabilidade de sucesso $\hat{p}_{1,j}$, para $j = 1, \ldots, n$. A frequência relativa de sucessos na amostra é $\tilde{\pi}_1 = (U_1 + \ldots + U_n)/n$. Na prática, $\tilde{\pi}$ poderia representar, por exemplo, a frequência de acertos a determinado item de um teste em

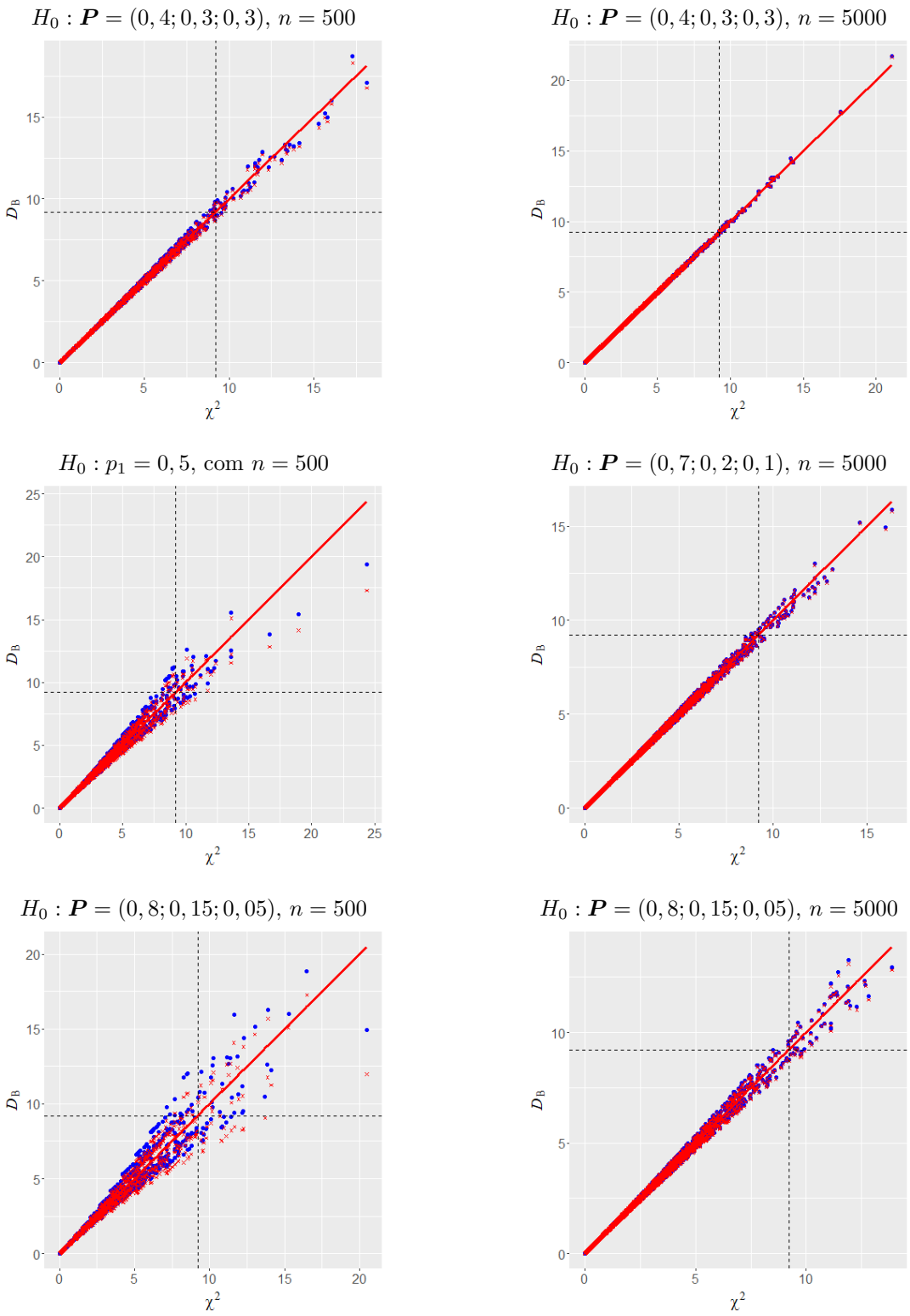


Figura 4.3: Dispersões entre $n\chi^2(\boldsymbol{P}||\boldsymbol{Q})$ e $8nD_B(\boldsymbol{P}||\boldsymbol{Q})$ (pontos). Cinco mil estatísticas obtidas sob a hipótese nula do teste correspondente. A linha sólida representa a reta (4.15), enquanto $\times$ representa a aproximação (4.14). As linhas pontilhadas indicam o valor crítico 9,21 relativo ao nível $\alpha = 1\%$ com base na distribuição $\chi^2_{(2)}$.

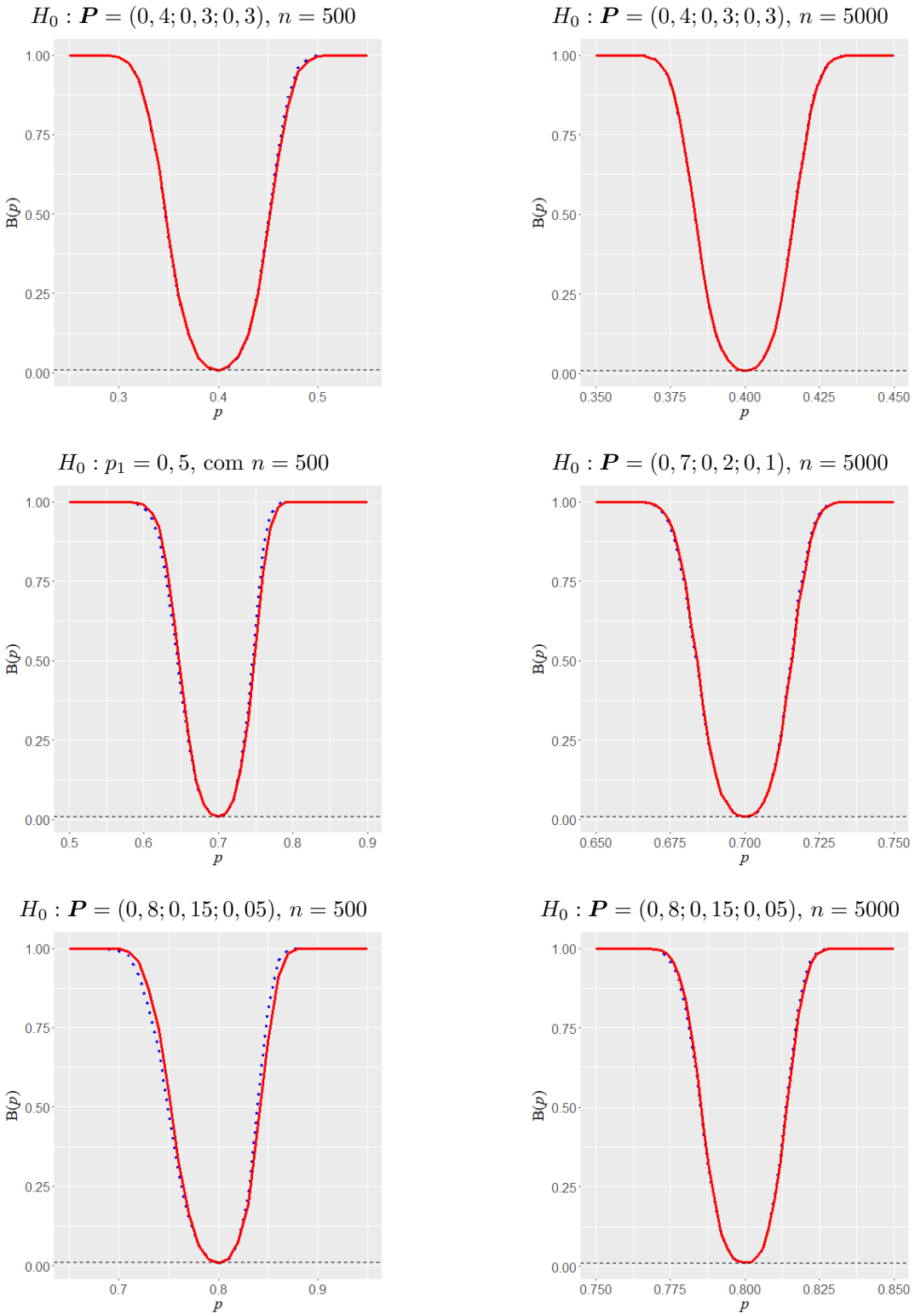


Figura 4.4: Poderes dos testes, $B(p)$, com nível de significância $\alpha = 1\%$, mediante 5 mil replicações das estatísticas $n\chi^2(\boldsymbol{P}||\boldsymbol{Q})$ (linha pontilhada) e $8nD_B(\boldsymbol{P}||\boldsymbol{Q})$ (linha contínua).

que os examinandos possuem habilidades distintas. Caso as probabilidades individuais $\{p_{1,j} : 1 \leq j \leq n\}$ sejam hipoteticamente conhecidas (por exemplo, mediante ajuste de um modelo da TRI), define-se a frequência esperada de sucessos como $\hat{\pi}_1 = (\hat{p}_{1,1} + \ldots + \hat{p}_{1,n})/n$.

Assim, deseja-se testar a hipótese nula $H_0 : \tilde{\pi}_1 = \hat{\pi}_1$, dado um conjunto hipotético de probabilidades individuais, $\{\hat{p}_{1,j} : 1 \leq j \leq n\}$, contra $H_1 : \tilde{\pi}_1 \neq \hat{\pi}_1$. A Tabela 4.3 descreve resumidamente esta situação.

Tabela 4.3 - Representação das distribuições $\boldsymbol{Q}$ e $\boldsymbol{P}$ referentes ao Caso 3, na qual $\tilde{\pi}_1$ e $\tilde{\pi}_2 = 1 - \tilde{\pi}_1$ representam as frequências empíricas, e $\hat{\pi}_1$ e $\hat{\pi}_2 = 1 - \hat{\pi}_1$ são os valores esperados correspondentes.

	sucessos	fracassos
$\boldsymbol{Q}$	$\hat{\pi}_1$	$\hat{\pi}_2$
$\boldsymbol{P}$	$\tilde{\pi}_1$	$\tilde{\pi}_2$

Para exemplificar, suponha que, para certo amostral n, uma coleção particular de probabilidades $\{\hat{p}_{1,j} : 1 \leq j \leq n\}$ seja obtida mediante realizações de uma distribuição Uniforme no intervalo $[0,25; 0,95]$. Mantendo-se tal coleção fixada, consideramos 5 mil replicações da Tabela 4.3 para $n = 1000$, 5000 e 10000. A Figura 4.5 mostra as dispersões entre as estatísticas $n\chi^2(\boldsymbol{P}||\boldsymbol{Q})$ e $8nD_B(\boldsymbol{P}||\boldsymbol{Q})$ obtidas sob H_0 para $n = 1000$ e 5000. Para essa situação é possível observar a relação $8nD_B(\boldsymbol{P}||\boldsymbol{Q}) \approx n\chi^2(\boldsymbol{P}||\boldsymbol{Q})$.

Sob H_0, porém, essas estatísticas não seguem distribuição χ^2 com 1 grau de liberdade. A soma $S_n = U_1 + \ldots + U_n$, que segue a distribuição Poisson-Binomial (Butler e Stephens, 2017), possui média $\mu_S = \sum \hat{p}_i = n\hat{\pi}_1$ e variância $\sigma_S^2 = \sum \hat{p}_i(1 - \hat{p}_i) = \mu_S - \mu_S^{(2)} = n\left(\hat{\pi}_1 - \hat{\pi}^{(2)}\right)$, na qual $\mu_S^{(2)} = \sum \hat{p}_i^2$ e $\hat{\pi}^{(2)} = \sum \hat{p}_i^2/n$. Nesse caso, tem-se o resultado limite de Liapunov na qual

$$Z = \frac{\sum_{i=1}^{n}(U_i - \hat{p}_i)}{\sqrt{\sum_{i=1}^{n} \hat{p}_i(1 - \hat{p}_i)}} \xrightarrow{D} \mathrm{N}(0, 1) \tag{4.24}$$

à medida que n aumenta. A padronização (4.24) pode ser reescrita como

$$\begin{aligned} Z &= \frac{\sum_{i=1}^{n}(U_i - \hat{p}_i)}{\sqrt{\sum_{i=1}^{n}\hat{p}_i(1-\hat{p}_i)}} = \frac{\frac{n}{n}\sum_{i=1}^{n}(U_i - \hat{p}_i)}{\sqrt{\frac{n}{n}\sum_{i=1}^{n}\hat{p}_i(1-\hat{p}_i)}} \\ &= \frac{n(\tilde{\pi}_1 - \hat{\pi}_1)}{\sqrt{n\left(\hat{\pi}_1 - \hat{\pi}_1^{(2)}\right)}} \\ &= \frac{\tilde{\pi}_1 - \hat{\pi}_1}{\sqrt{\left(\hat{\pi}_1 - \hat{\pi}_1^{(2)}\right)/n}} \\ &= \frac{\varepsilon_1}{\sqrt{\left(\hat{\pi}_1 - \hat{\pi}_1^{(2)}\right)/n}}, \end{aligned} \tag{4.25}$$

em que $\hat{\pi}_1^{(2)} = \sum_{i=1}^{n}\hat{p}_{1,i}^2/n$.

Assim, para grandes amostras, considere

$$\hat{\kappa} n \chi^2(\boldsymbol{P}||\boldsymbol{Q}) \sim \chi^2_{(1)},$$

na qual

$$\hat{\kappa} = \frac{\hat{\pi}_1(1-\hat{\pi}_1)}{\hat{\pi}_1 - \hat{\pi}_1^{(2)}} \tag{4.26}$$

denota o fator de correção devido à não homogeneidade das probabilidades de sucesso. Note que $\hat{\kappa} = 1$ se $\hat{p}_i = \hat{\pi}_1$ para todo $i = 1, \ldots, n$.

Alguns autores, como Eisinga et al. (2013), discutem que, dependendo do tamanho amostral, a aproximação (4.24) pode não ser muito boa nas caudas por causa da assimetria da soma S_n. Mas para as nossas aplicações, veremos a seguir que ela se mostra razoável (Tabela 4.5).

Para obtermos uma expressão aproximada para a geração da distribuição amostral mediante simulações de Monte Carlo, a partir de (4.11), (4.15), (4.25) e (4.26), considerando $\varepsilon_1 = Z\sqrt{\left(\hat{\pi}_1 - \hat{\pi}_1^{(2)}\right)/n}$, e efetuando-se as devidas adaptações notacionais, podemos escrever

$$\begin{aligned} D_B(\boldsymbol{P}||\boldsymbol{Q}) = -\ln A(\boldsymbol{P}||\boldsymbol{Q}) &\approx \frac{\varepsilon_1^2}{8\hat{\pi}_1(1-\hat{\pi}_1)} - \frac{\varepsilon_1^3(1-2\hat{\pi}_1)}{16\hat{\pi}_1^2(1-\hat{\pi}_1)^2} \\ &\approx \frac{Z^2\left(\hat{\pi}_1 - \hat{\pi}_1^{(2)}\right)}{8n\hat{\pi}_1(1-\hat{\pi}_1)} - \frac{Z^3\left(\hat{\pi}_1 - \hat{\pi}_1^{(2)}\right)^{3/2}(1-2\hat{\pi}_1)}{16n^{3/2}\hat{\pi}_1^2(1-\hat{\pi}_1)^2} \\ &\approx \frac{Z^2}{8n\hat{\kappa}} - \frac{Z^3(1-2\hat{\pi}_1)}{16n\hat{\kappa}\sqrt{n\hat{\kappa}\hat{\pi}_1(1-\hat{\pi}_1)}} \end{aligned}$$

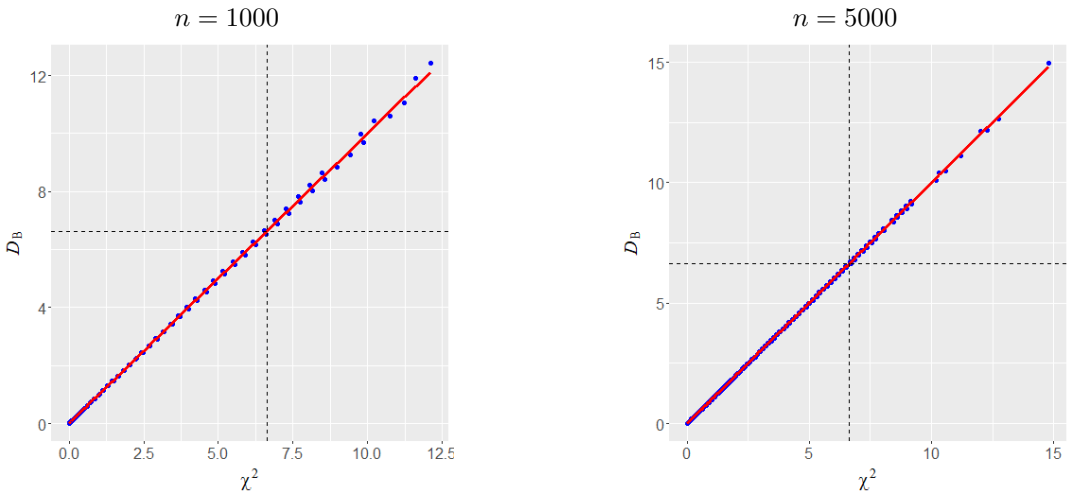


Figura 4.5: Dispersões entre $n\hat{\kappa}\chi^2(\boldsymbol{P}||\boldsymbol{Q})$ e $8n\hat{\kappa}D_B(\boldsymbol{P}||\boldsymbol{Q})$ com base em cinco mil replicações. A linha sólida representa a reta $8n\hat{\kappa}D_B(\boldsymbol{P}||\boldsymbol{Q}) = n\hat{\kappa}\chi^2(\boldsymbol{P}||\boldsymbol{Q})$. As linhas pontilhadas indicam o valor crítico 6,63 relativo ao nível $\alpha = 1\%$ com base na distribuição $\chi^2_{(1)}$.

Logo, temos a aproximação

$$8n\hat{\kappa}D_B(\boldsymbol{P}||\boldsymbol{Q}) \sim Z^2 - \frac{1 - 2\hat{\pi}_1}{2\sqrt{n\hat{\kappa}\hat{\pi}_1(1-\hat{\pi}_1)}}Z^3. \tag{4.27}$$

Com base nesse resultado, efetuando-se 10^7 replicações da distribuição normal padrão Z, tomando-se também um conjunto fixo de probabilidades $\{\hat{p}_{1,j} : 1 \leq j \leq n\}$ como hipótese nula, realizadas de uma distribuição Uniforme no intervalo $[0,25; 0,95]$ (para $n = 1000$ e 5000), obtivemos as distribuições amostrais empíricas das estatísticas $n\hat{\kappa}\chi^2(\boldsymbol{P}||\boldsymbol{Q})$ e $8n\hat{\kappa}D_B(\boldsymbol{P}||\boldsymbol{Q})$. A Tabela 4.4 mostra seus valores críticos empíricos c_α para os níveis de significância $\alpha = 0,1\%$, 1% e 5%.

Tomando-se agora os níveis de significância nominais de 0,1%, 1% e 5% e seus valores críticos correspondentes a partir da distribuição assintótica $\chi^2_{(1)}$, efetuamos 10^5 realizações da Tabela 4.3, para um dado conjunto fixo de probabilidades $\{\hat{p}_{1,j} : 1 \leq j \leq n\}$ (retiradas de uma distribuição Uniforme no intervalo $[0,25; 0,95]$). A Tabela 4.5 mostra que os percentuais de ocorrência do erro do tipo I se encontram bastante próximos dos seus respectivos valores nominais.

Para uma ilustração do poder do teste, consideramos a transformação logito das probabilidades individuais, $\text{logit} = \ln[\hat{p}_{1,j}/(1-\hat{p}_{1,j})]$, e depois consideramos probabilidades sob a hipótese alternativa na forma $\hat{p}^*_{1,j} = \exp(\text{logit} + \delta)/(1 + \exp(\text{logit} + \delta))$, para $-3 < \delta < 3$. A Figura 4.6 ilustra as funções de poder do teste, $\text{B}(p)$, considerando-se os valores críticos empíricos apresentados na Tabela 4.4.

Tabela 4.4 - Caso 3. Valores críticos empíricos c_α correspondentes ao teste $H_0 : \tilde{\pi}_1 = \hat{\pi}_1$, dado um conjunto $\{\hat{p}_{1,j} : 1 \leq j \leq n\}$, contra $H_1 : \tilde{\pi}_1 \neq \hat{\pi}_1$, com tamanhos amostrais iguais a $n = 1000$ e 5000 e níveis de significância $\alpha = 0,1\%$, 1% e 5% obtidos com base em 10^7 realizações de (4.27). Para o caso assintótico, $n\hat{\kappa}\chi^2(\boldsymbol{P}||\boldsymbol{Q}) \sim \chi^2_{(1)}$.

n	estatística					
	$n\hat{\kappa}\chi^2(\boldsymbol{P}\|\|\boldsymbol{Q})$			$8n\hat{\kappa}D_B(\boldsymbol{P}\|\|\boldsymbol{Q})$		
	0,1%	1%	5%	0,1%	1%	5%
1000	10,80	6,64	3,84	10,81	6,64	3,84
5000	10,80	6,64	3,84	10,80	6,64	3,84
$+\infty$	10,83	6,63	3,84	10,83	6,63	3,84

Tabela 4.5 - Níveis empíricos de significância (%) para a situação do Caso 3, considerando 10^5 realizações da Tabela 4.3, tomando-se os valores críticos assintóticos 3,84; 6,63 e 10,83 da distribuição $\chi^2_{(1)}$, relativos aos níveis nominais de significância $\alpha = 0,1\%$, 1% e 5%.

n		estatísticas	
	α nominal	$n\hat{\kappa}\chi^2(\boldsymbol{P}\|\|\boldsymbol{Q})$	$8n\hat{\kappa}D_B(\boldsymbol{P}\|\|\boldsymbol{Q})$
1000	0,1%	0,11	0,12
	1,0%	1,03	1,06
	5,0%	4,84	4,84
5000	0,1%	0,08	0,09
	1,0%	1,00	1,00
	5,0%	4,94	5,09

4.3.3 Caso 4

Agora consideraremos a aplicação para fins de comparação entre duas tabelas 2×2, no contexto do modelo apresentado no capítulo anterior. Considerando as variáveis binárias R_{ij} — que assume valor 1 se o indivíduo j responde ao item i ou valor 0 caso ele deixe de respondê-lo —, e U_{ij}, que é unitária se o indivíduo j acerta ao item i e nula se ele não acerta, o modelo 3.7 propicia estimativas das probabilidades individuais, de maneira que as frequências absolutas esperadas para o item i do grupo de n examinandos podem ser escritas como agregações das estimativas das probabilidades proporcionadas pelo modelo,

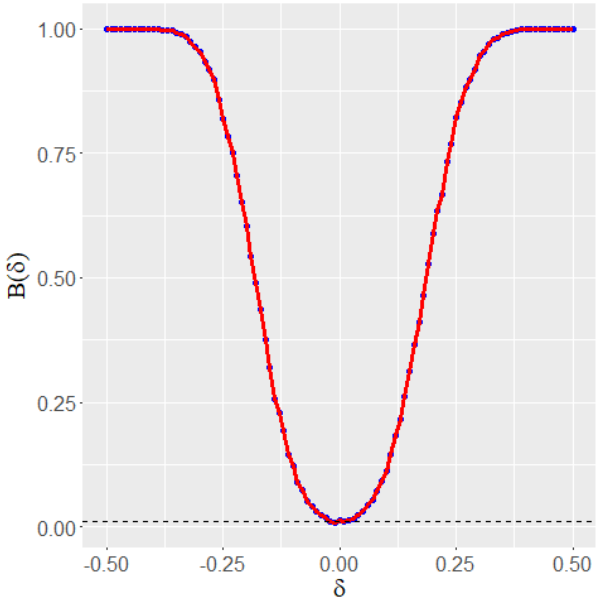


Figura 4.6: Poderes dos testes, $B(\delta)$, com nível de significância $\alpha = 1\%$, relativos ao teste $H_0 : \tilde{\pi}_1 = \hat{\pi}_1$ contra $H_1 : \tilde{\pi}_1 \neq \hat{\pi}_1$, mediante 5 mil replicações das estatísticas $n\hat{\kappa}\chi^2(\boldsymbol{P}||\boldsymbol{Q})$ (linha pontilhada) e $8n\hat{\kappa}D_B(\boldsymbol{P}||\boldsymbol{Q})$ (linha contínua), com $n = 1000$.

ou seja,

$$\hat{n}_{01} = \sum_{j=1}^{n} \hat{\mathrm{P}}(R_{ij} = 1|U_{ij} = 0) \cdot \hat{\mathrm{P}}(U_{ij} = 0);$$

$$\hat{n}_{11} = \sum_{j=1}^{n} \hat{\mathrm{P}}(U_{ij} = 1);$$

$$\hat{n}_{00} = n - \hat{n}_{01} - \hat{n}_{11}.$$

Assim, as frequências esperadas, segundo o modelo ajustado para o item i, podem ser definidas como

$$\hat{\pi}_{00} = \frac{\hat{n}_{00}}{n};$$

$$\hat{\pi}_{01} = \frac{\hat{n}_{01}}{n};$$

$$\hat{\pi}_{11} = \frac{\hat{n}_{11}}{n}.$$

Agora, considere as frequências observadas (empiricamente) para cada item i

$$\tilde{\pi}_{00} = \frac{1}{n}\sum_{j=1}^{n} \mathrm{I}(R_{ij} = 0, U_{ij} = 0);$$
$$\tilde{\pi}_{01} = \frac{1}{n}\sum_{j=1}^{n} \mathrm{I}(R_{ij} = 1, U_{ij} = 0);$$
$$\tilde{\pi}_{11} = \frac{1}{n}\sum_{j=1}^{n} \mathrm{I}(R_{ij} = 1, U_{ij} = 1),$$

que I é uma função indicadora tal que $\mathrm{I}(R_{ij} = r, U_{ij} = u) = 1$, se $R_{ij} = r$ e $U_{ij} = u$, e $\mathrm{I}(R_{ij} = r, U_{ij} = u) = 0$, se $R_{ij} \neq r$ ou $U_{ij} \neq u$, para $r, u = 0$ ou 1. Desse modo, essas frequências empíricas correspondem à razão entre o número de casos relativos ao evento de interesse e o total n de examinandos.

Assim, as frequências esperadas, segundo o modelo ajustado para o item i, podem ser definidas como

$$\hat{\pi}_{00} = \frac{\hat{n}_{00}}{n};$$
$$\hat{\pi}_{01} = \frac{\hat{n}_{01}}{n};$$
$$\hat{\pi}_{11} = \frac{\hat{n}_{11}}{n}.$$

Agora, considere as frequências observadas (empiricamente) para cada item i

$$\tilde{\pi}_{00} = \frac{1}{n}\sum_{j=1}^{n} \mathrm{I}(R_{ij} = 0, U_{ij} = 0);$$
$$\tilde{\pi}_{01} = \frac{1}{n}\sum_{j=1}^{n} \mathrm{I}(R_{ij} = 1, U_{ij} = 0);$$
$$\tilde{\pi}_{11} = \frac{1}{n}\sum_{j=1}^{n} \mathrm{I}(R_{ij} = 1, U_{ij} = 1),$$

em que I é uma função indicadora tal que $\mathrm{I}(R_{ij} = r, U_{ij} = u) = 1$, se $R_{ij} = r$ e $U_{ij} = u$, e $\mathrm{I}(R_{ij} = r, U_{ij} = u) = 0$, se $R_{ij} \neq r$ ou $U_{ij} \neq u$, para $r, u = 0$ ou 1. Desse modo, essas frequências empíricas correspondem à razão entre o número de casos relativos ao evento de interesse e o total n de examinandos.

Se o propósito for obter, para cada item, a distância entre a Tabela 4.6(a) e a 4.6(b), podemos considerar a distância χ^2 dada por

$$\chi^2(\boldsymbol{P}||\boldsymbol{Q}) = \left[\frac{(\hat{\pi}_{11} - \tilde{\pi}_{11})^2}{\hat{\pi}_{11}} + \frac{(\hat{\pi}_{01} - \tilde{\pi}_{01})^2}{\hat{\pi}_{01}} + \frac{(\hat{\pi}_{00} - \tilde{\pi}_{00})^2}{\hat{\pi}_{00}}\right],$$

Tabela 4.6 - Representação da distribuição conjunta observada (empiricamente) e a esperada para um item i com base no modelo 3.7, relativas às variáveis R (respondeu = 1, não respondeu = 0) e U (acertou = 1, não acertou = 0).

(a) observada

		R		
		0	1	total
U	0	$\tilde{\pi}_{00}$	$\tilde{\pi}_{01}$	$\tilde{\pi}_{0\cdot}$
	1	0	$\tilde{\pi}_{11}$	$\tilde{\pi}_{11}$
	total	$\tilde{\pi}_{00}$	$\tilde{\pi}_{\cdot 1}$	1

(a) esperada

		R		
		0	1	total
U	0	$\hat{\pi}_{00}$	$\hat{\pi}_{01}$	$\hat{\pi}_{0\cdot}$
	1	0	$\hat{\pi}_{11}$	$\hat{\pi}_{11}$
	total	$\hat{\pi}_{00}$	$\hat{\pi}_{\cdot 1}$	1

e a distância de Bhattacharyya dada por

$$D_B(\boldsymbol{P}||\boldsymbol{Q}) = -\ln\left(\sqrt{\hat{\pi}_{11}\tilde{\pi}_{11}} + \sqrt{\hat{\pi}_{01}\tilde{\pi}_{01}} + \sqrt{\hat{\pi}_{00}\tilde{\pi}_{00}}\right).$$

Caso o objetivo seja avaliar a distribuição marginal correspondente à variável R, ou seja, no que se refere à distância entre a fração esperada e a observada das não-respostas, pode-se definir as medidas

$$D_{B,0}(\boldsymbol{P}_0||\boldsymbol{Q}_0) = -\ln\left(\sqrt{\hat{\pi}_{00}\tilde{\pi}_{00}} + \sqrt{(1-\hat{\pi}_{00})(1-\tilde{\pi}_{00})}\right)$$

e

$$\chi^2_0(\boldsymbol{P}_0||\boldsymbol{Q}_0) = \left[\frac{(\hat{\pi}_{00}-\tilde{\pi}_{00})^2}{\hat{\pi}_{00}} + \frac{(1-\hat{\pi}_{00}-(1-\tilde{\pi}_{00}))^2}{1-\hat{\pi}_{00}}\right].$$

Essas duas últimas medidas remetem ao Caso 3, em que $n\hat{\kappa}\chi^2_0(\boldsymbol{P}||\boldsymbol{Q})$ e $8n\hat{\kappa}D_{B,0}(\boldsymbol{P}||\boldsymbol{Q})$, com $\hat{\kappa} = \frac{\hat{\pi}_1(1-\hat{\pi}_1)}{\hat{\pi}_1-\hat{\pi}_1^{(2)}}$ e $\hat{\pi}_1^{(2)} = \sum_{j=1}^n \hat{\mathrm{P}}^2(R_{ij}=0)/n$, seguem assintoticamente uma distribuição $\chi^2_{(1)}$.

Já as duas primeiras medidas, que se referem ao caso $m = 3$ com heterogeneidade das probabilidades de sucesso, constituem o objeto principal desta ilustração. Para esse caso, devido à dificuldade de se encontrar um fator de correção comum κ que remeta $8n\kappa D_B(\hat{\boldsymbol{P}}||\boldsymbol{Q})$ diretamente a uma distribuição χ^2, restringiremo-nos a estudar empiricamente os comportamentos das estatísticas $n\chi^2(\boldsymbol{P}||\boldsymbol{Q})$ e $8nD_B(\hat{\boldsymbol{P}}||\boldsymbol{Q})$. No caso da estatística χ^2 de Pearson, lembremo-nos que, na sua forma original, ela se baseia em contagens multinomiais. Em nosso caso específico, tais contagens não ocorrem sob a hipótese de haver probabilidade de sucesso constante e, por isso, é preciso obter uma distribuição amostral mais adequada. Sem uma referência assintótica, os valores críticos correspondentes serão ser obtidos com base em distribuições empíricas geradas computacionalmente.

Com respeito à estatística $D_B(\hat{\boldsymbol{P}}||\boldsymbol{Q})$, sua distribuição amostral pode ser obtida empiricamente com base em (4.14), para $m = 3$, em que $\varepsilon_k \sim \mathrm{N}\left(0, \left(\hat{\pi}_k - \hat{\pi}_k^{(2)}\right)/n\right)$ para $k \leq 2$.

A covariância entre as variáveis aleatórias ε_1 e ε_2 é dada por

$$\begin{aligned}
\mathrm{Cov}[\varepsilon_1, \varepsilon_2] &= \mathrm{E}[\varepsilon_1, \varepsilon_2] = \mathrm{E}[(\tilde{\pi}_1 - \hat{\pi}_1)(\tilde{\pi}_2 - \hat{\pi}_2)] = \mathrm{E}[\tilde{\pi}_1\tilde{\pi}_2] - \hat{\pi}_1\hat{\pi}_2 \\
&= \frac{1}{n^2}\mathrm{E}\left[\sum_{i=1}^{n}\sum_{i'=1}^{n} I_{1i}I_{2i'}\right] - \hat{\pi}_1\hat{\pi}_2 \\
&= \frac{1}{n^2}\left\{\sum_{i=i'=1}^{n}\mathrm{E}\left[I_{1i}I_{2i}\right] + \sum\sum_{i\neq i'}\mathrm{E}\left[I_{1i}I_{2i'}\right]\right\} - \hat{\pi}_1\hat{\pi}_2 \\
&= 0 + \frac{1}{n^2}\sum\sum_{i\neq i'}\mathrm{E}[I_{1i}]\mathrm{E}[I_{2i'}] - \hat{\pi}_1\hat{\pi}_2 \\
&= \frac{1}{n^2}\sum\sum_{i\neq i'}\hat{p}_{1,i}\hat{p}_{2,i'} - \hat{\pi}_1\hat{\pi}_2 \\
&= \frac{1}{n^2}\left\{\sum_{i=1}^{n}\sum_{i'=1}^{n}\hat{p}_{1,i}\hat{p}_{2,i'} - \sum_{i=1}^{n}\hat{p}_{1,i}\hat{p}_{2,i}\right\} - \hat{\pi}_1\hat{\pi}_2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\hat{p}_{1,i}\frac{1}{n}\sum_{i'=1}^{n}\hat{p}_{2,i'} - \frac{1}{n^2}\sum_{i=1}^{n}\hat{p}_{1,i}\hat{p}_{2,i} - \hat{\pi}_1\hat{\pi}_2 \\
&= \hat{\pi}_1\hat{\pi}_2 - \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{p}_{1,i}\hat{p}_{2,i}}{n} - \hat{\pi}_1\hat{\pi}_2 \\
&= -\frac{1}{n}\hat{\pi}_{12}, \qquad (4.28)
\end{aligned}$$

em que $\hat{\pi}_{12} = \sum_{i=1}^{n}\hat{p}_{1,i}\hat{p}_{2,i}/n$.

Assim, a distribuição amostral de $8nD_B(\hat{\boldsymbol{P}}||\boldsymbol{Q})$ pode ser gerada aproximadamente com base em (4.23), tomando-se realizações de uma distribuição normal bivariada $(\varepsilon_1, \varepsilon_2)$ cujo vetor de médias é nulo, e cuja matriz de covariâncias é dada por

$$\Sigma = \frac{1}{n}\begin{bmatrix} \left(\hat{\pi}_1 - \hat{\pi}_1^{(2)}\right) & -\hat{\pi}_{12} \\ -\hat{\pi}_{12} & \left(\hat{\pi}_2 - \hat{\pi}_2^{(2)}\right) \end{bmatrix},$$

com $q_3 = 1 - q_1 - q_2$ e $\varepsilon_3 = -(\varepsilon_1 + \varepsilon_2)$.

Para exemplificar, o Programa A.1 apresenta um código em R para simular respostas a um item, conforme o modelo 3.7, com parâmetros $a_1 = 1,9$, $b_1 = 0,1$, $a_2 = 0,9$ e $b_2 = -0,2$. Nesse experimento hipotético, o vetor de proficiências (θ_1, θ_2) segue uma distribuição normal bivariada com médias nulas, variâncias unitárias, e correlação entre as proficiências igual a 0.5, de onde foi retirada uma amostra aleatória de $n = 5.000$ pares

de proficiências. Dessa maneira, cada realização da probabilidade π obtida com base no modelo 3.7 segue uma distribuição *logistica-normal* (Atchison e Shen, 1980).

Como resultado dessa simulação, a Tabela 5.9 apresenta a distribuição das frequências esperadas.

Tabela 4.7 - Frequências esperadas para um item hipotético, relativas às variáveis R (respondeu = 1, não respondeu = 0) e U (acertou = 1, não acertou = 0).

		R	
		0	1
U	0	27,0%	25,4%
	1	0,0%	47,6%

Tomando-se essa coleção $n = 5.000$ probabilidades como uma população hipotética de examinandos com proficiências gaussianas, com base em 10^7 mil realizações de (4.23), foram obtidas as distribuições amostrais empíricas das estatísticas $n\chi^2(\boldsymbol{P}||\boldsymbol{Q})$ e $8nD_B(\boldsymbol{P}||\boldsymbol{Q})$. A Tabela 4.8 mostra seus percentís empíricos de 95%, 99% e 99,9%.

Tabela 4.8 - Valores críticos empíricos para a situação do Caso 4 ($n = 5000$ e proficiências gaussianas).

estatística	95%	99%	99,9%
$n\chi^2(\boldsymbol{P}\|\|\boldsymbol{Q})$	4,615	7,198	11,014
$8nD_B(\boldsymbol{P}\|\|\boldsymbol{Q})$ (forma aproximada)	4,614	7,198	11,018
$8nD_B(\boldsymbol{P}\|\|\boldsymbol{Q})$ (forma exata)	4,617	7,208	11,023

Note que os valores críticos apresentados na Tabela 4.8 diferem dos valores críticos da distribuição $\chi^2_{(2)}$, que seriam respectivamente iguais a 13,815; 9,210 e 5,991.

Tomando-se agora 5 mil réplicas de tabelas observadas, foram obtidas realizações de $n\chi^2(\boldsymbol{P}||\boldsymbol{Q})$ e da estatística $8nD_B(\boldsymbol{P}||\boldsymbol{Q})$ com base na forma (4.5). A Figura 4.7 mostra que há aproximação linear entre as estatísticas χ^2 e $8nD_B$ para essa situação particular. A correlação linear entre essas estatísticas é forte, superior a 0,9998, o que permite sugerir que os poderes oferecidos por elas sejam equivalentes.

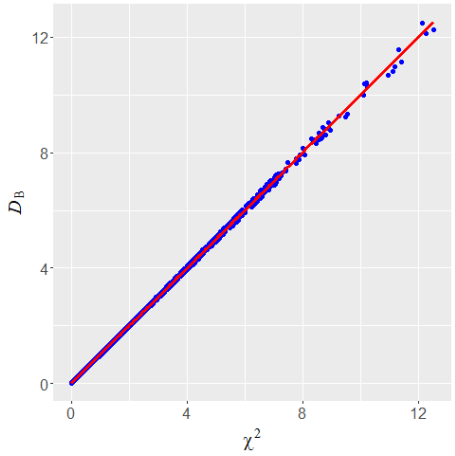


Figura 4.7: Dispersão entre cinco mil estatísticas $n\chi^2(\boldsymbol{P}||\boldsymbol{Q})$ e $8nD_B(\boldsymbol{P}||\boldsymbol{Q})$ (pontos) obtidas sob H_0. A linha sólida representa a reta $8nD_B(\boldsymbol{P}||\boldsymbol{Q}) = n\chi^2(\boldsymbol{P}||\boldsymbol{Q})$.

4.4 Algumas considerações

Este capítulo tratou de um ensaio em que se sugere a distância de Bhattacharyya como uma alternativa à tradicional estatística χ^2. Quando as duas estatísticas se relacionam linearmente, ambas proporcionam poderes equivalentes. Mas para certas situações, nosso estudo empírico constatou poderes distintos. Portanto, ambas as estatísticas poderiam ser aplicadas de modo complementar nessas situações.

Para o caso não IID, a tradicional estatística χ^2, caso seja aplicada sem o fator de correção, não segue a distribuição quiquadrado esperada. Neste capítulo, para o caso $m = 2$, obtivemos o fator de correção e a distribuição amostral assintótica. O estudo do Caso 4 ainda está para ser aperfeiçoado em trabalhos futuros, para que se busque um fator de correção para as medidas de interesse.

Uma vantagem da distância de Bhattacharyya sobre a χ^2 é que ela pode ser efetivamente utilizada na presença frequências esperadas muito baixas. Por isso, no próximo capítulo, ela será utilizada para fins de diagnóstico do modelo. Em contraste, uma desvantagem é que a sua distribuição amostral é obtida mediante aproximações matemáticas. Essa desvantagem, porém, é superada mediante elevação do tamanho amostral.

Capítulo 5

Resultados

Esta seção ilustra uma aplicação do modelo (3.7), considerando a prova de conhecimentos da terceira etapa do subprograma 2006-2008 do PAS/UnB aplicada em 7/12/2008 para 10.822 candidatos. Uma cópia dessa prova e seu gabarito oficial definitivo se encontram disponíveis em `http://www.cespe.unb.br/PAS`. Essa prova possui duas partes. A primeira trata de língua estrangeira (inglês, francês ou espanhol, de acordo com a opção do estudante), e a segunda contempla o restante das disciplinas (artes cênicas, artes visuais, biologia, filosofia, física, geografia, história, língua portuguesa, literatura, matemática, música, química e sociologia).

Após a exclusão de itens anulados, a atenção nesta seção se restringe a 100 itens do tipo Certo (C) ou Errado (E) da segunda parte da prova de conhecimentos. Esses itens foram organizados em grupos de disciplinas, conforme detalha a Tabela 5.1. Cada item desse tipo deve ser julgado de acordo com o comando que se refere, cuja resposta deve ser assinalada na folha de respostas como C ou E. No cálculo do resultado da prova, ao item cuja resposta coincida com o gabarito oficial definitivo atribui-se uma pontuação positiva (acerto); e ao item cuja resposta divirja desse gabarito é atribuído pontuação negativa (discordante). Porém, caso o item seja deixado em branco ou com dupla marcação na folha de respostas, a pontuação é nula (não resposta).

Tabela 5.1 - Distribuição dos itens em grupos de disciplinas

Grupo	disciplinas	quantidade de itens
I	filosofia, geografia, história, língua portuguesa, e sociologia	24
II	física e matemática	24
III	biologia e química	25
IV	artes cênicas, artes visuais e literatura	27

Essa apenação proporciona uma incidência natural de não respostas dependendo das características do item ou do próprio candidato. A Figura 5.0 (a) apresenta a dispersão entre o total de itens não respondidos (T_{nr}) e o total de acertos por candidato (T_a). Seu aspecto é de funil, em que a variabilidade de T_a diminui à medida que T_{nr} aumenta. De modo análogo, a dispersão entre T_{nr} e o total de respostas divergentes por candidato (T_d) apresenta um aspecto triangular (Figura 5.0 (b)). Enquanto as distribuições de T_a e T_d possuem aspecto aproximadamente sinusoidal, a forma da distribuição de T_{nr} evidencia uma concentração de zeros, na qual quase $12,5\%$ dos candidatos responderam todos os itens. A variabilidade da quantidade de acertos ou de respostas divergentes para esses candidatos que não deixaram respostas em branco é elevada, o que permite sugerir, por exemplo, que boa parte desses candidatos apresentem maior propensão θ_1.

A Figura 5.2, que mostra o percentual de respostas em branco para cada item, sugere uma relação entre a incidência de não respostas e os grupos de disciplinas (Tabela 5.1). Com base no teste de Kruskal-Wallis, com 3 graus de liberdade, essa relação é evidenciada com p-valor igual a 0,0005. De fato, não é surpreendente que os itens que versem sobre física, matemática, biologia e química (grupos II e III) apresentem taxas mais elevadas de não-respostas e que os do grupo I tenham taxas menores. Esse fato já permite caracterizar a multidimensionalidade das não-respostas, ou seja, de que se trata de um fenômeno associado a grupos de disciplinas. Por isso, a aplicação do modelo (3.7) será efetuado segundo esses grupos de itens (Tabela 5.1). Outro aspecto, um pouco menos evidente, diz respeito à proporção entre não respostas e respostas divergentes, ou seja, a relação $\pi_{00} : \pi_{01}$ na Tabela 3.1. A Figura 5.3 mostra, para cada item, o percentual de respostas divergentes relativo aos não acertos, ou seja,

$$\mathrm{P}_d = \frac{\pi_{01}}{\pi_{00} + \pi_{01}} \times 100\%. \tag{5.1}$$

Em contraste com a Figura 5.2, a Figura 5.3 não indica claramente se algum grupo de disciplinas apresenta maiores ou menores incidências de respostas divergentes sobre o total de não acertos. O teste de Kruskal-Wallis, com 3 graus de liberdade, também não evidencia uma relação entre os percentuais P_d e os grupos de disciplinas (p-valor = 0,16).

A Figura 5.4 mostra a distribuição dos valores empiricos de π_{00} e π_{11} dos 100 itens do tipo C ou E da segunda parte da prova de conhecimentos (Tabelas 5.3 e 5.4). Os pontos em destaque são referentes aos itens para exemplificação mostrados nas Figuras 5.7, 5.8 e

5.9. Nessa figura, espera-se que a taxa observada de não-resposta ($\tilde{\pi}_{00}$) diminua à medida que a taxa de acerto aumenta ($\tilde{\pi}_{11}$), como os itens 15, 70, 71, 72, 73, 83 e 84. No entanto, alguns itens apresentam anomalias, havendo maior predomínio de respostas divergentes em vez de não-respostas. Por exemplo, o item 35 representa um caso peculiar, com taxa de acerto igual a 18,6%, com baixa taxa de não-resposta (5,1%) e considerável percentual de respostas divergentes (76,3%).

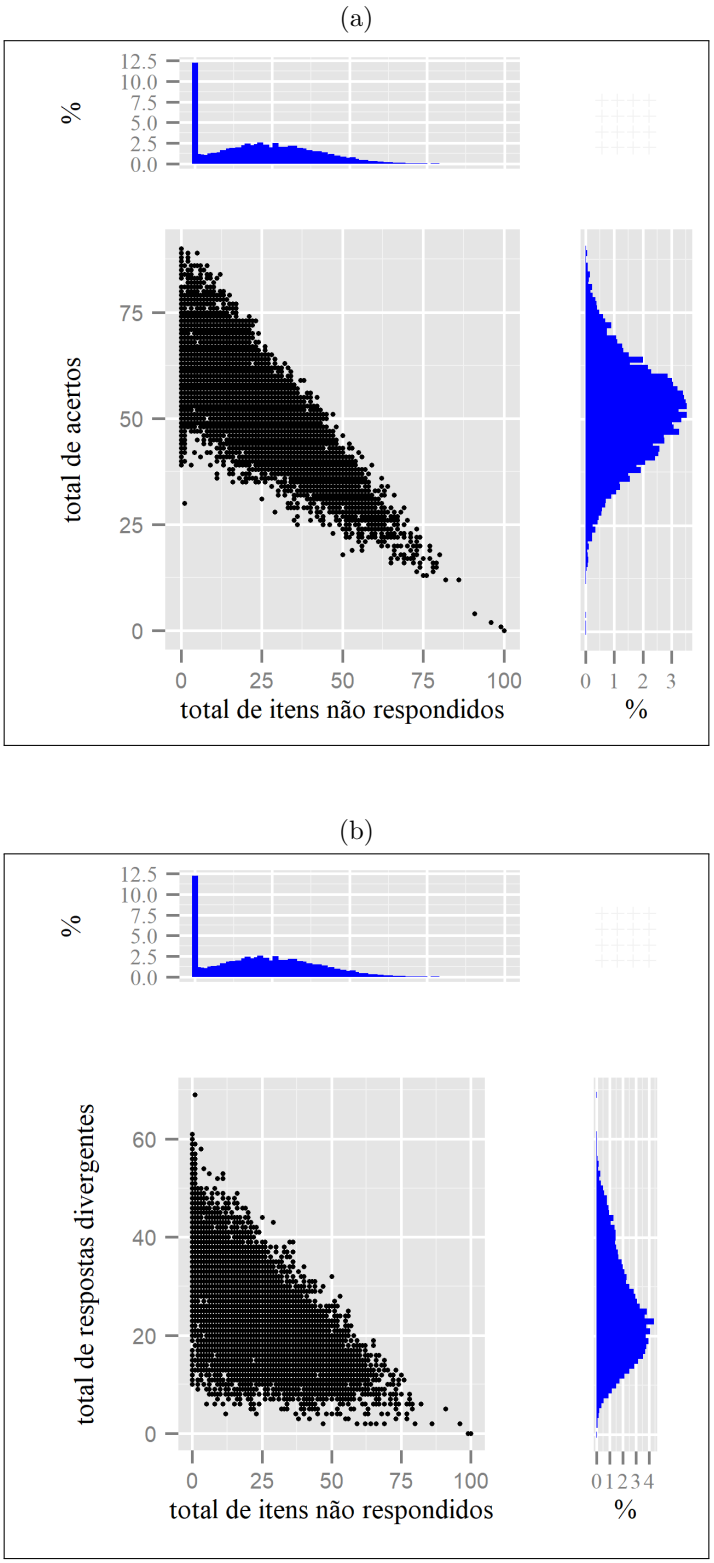


Figura 5.1: Dispersão do total de itens não respondidos *versus* o total de acertos por candidato, 5.0 dispersão entre o total de itens não respondidos e o total de respostas divergentes por candidato, e suas respectivas distribuições marginais.

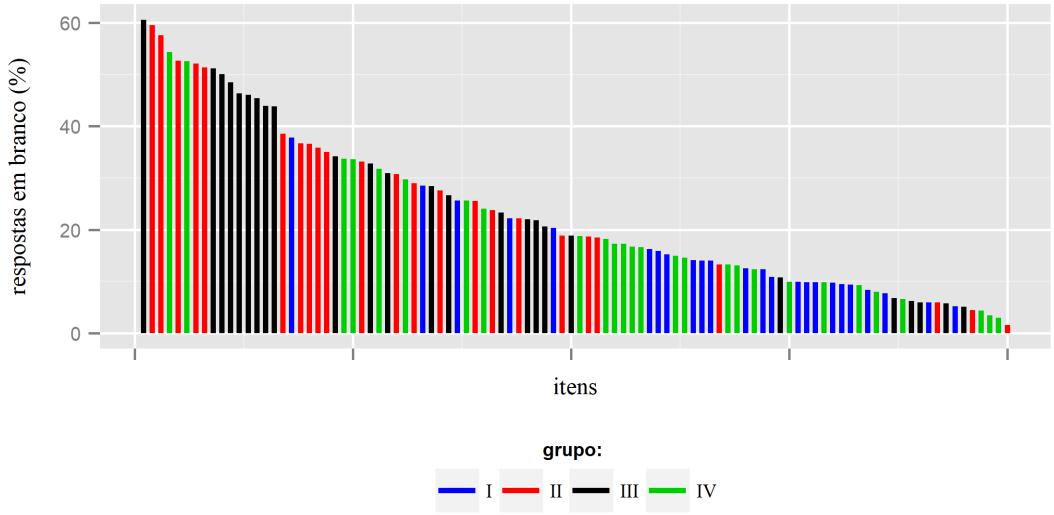


Figura 5.2: Percentual de respostas em branco para cada item, por grupos de disciplinas (I = filosofia, geografia, história, língua portuguesa, e sociologia; II = física e matemática, III = biologia e química, IV = artes cênicas, artes visuais e literatura).

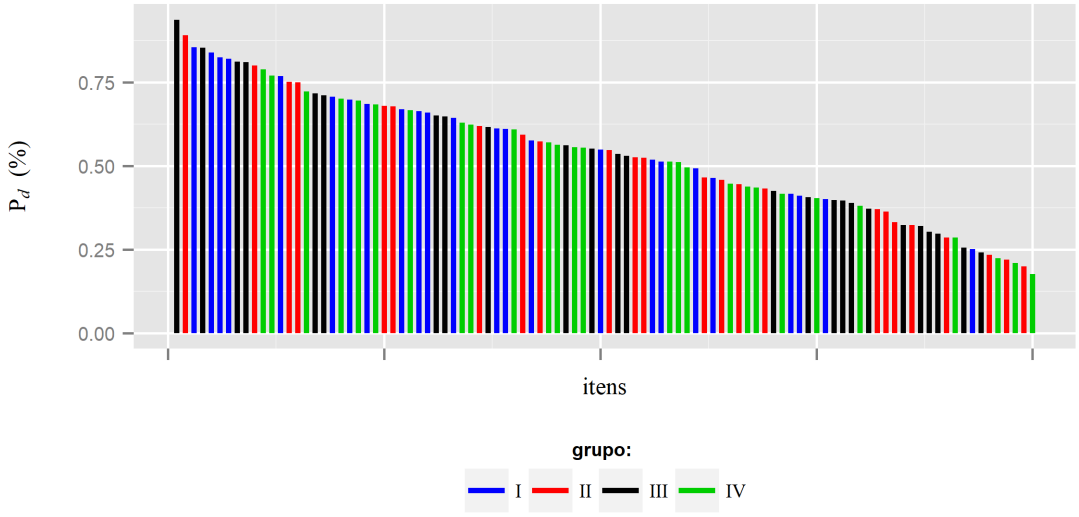


Figura 5.3: Percentual P_d de respostas divergentes relativo aos não acertos para cada item (5.1), por grupos de disciplinas (I = filosofia, geografia, história, língua portuguesa, e sociologia; II = física e matemática, III = biologia e química, IV = artes cênicas, artes visuais e literatura).

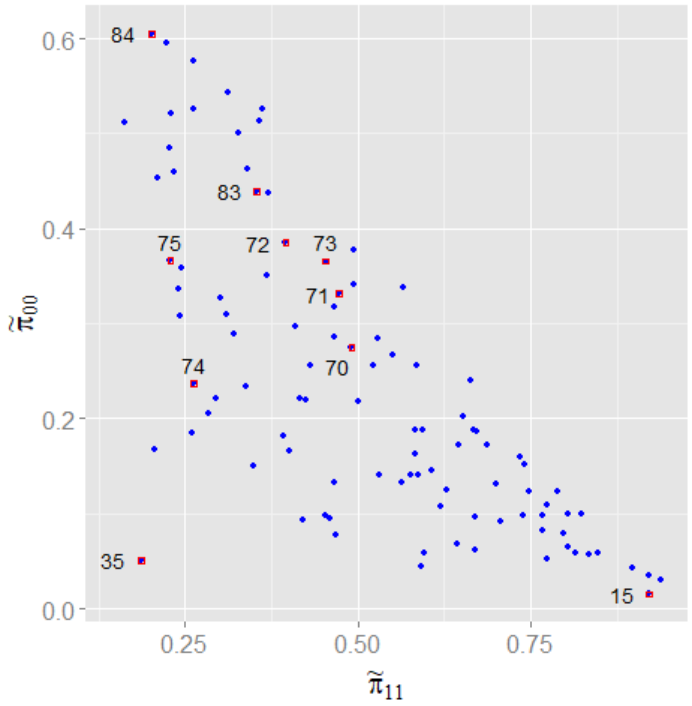


Figura 5.4: Dispersão entre os valores empíricos de π_{00} e π_{11}. Os oito pontos em destaque dizem respeito aos casos para exemplificação (itens 15, 35, 70 a 75, 83 e 84 apresentados nas Figuras 5.7, 5.8 e 5.9.).

As Tabelas 5.2, 5.3, 5.4 e 5.5 mostram as estimativas dos parâmetros dos itens obtidas por máxima verossimilhança marginal, as frequências percentuais observadas e esperadas conforme o modelo ajustado, e as medidas de proximidade D_B e $D_{B,0}$. As Figuras 5.5 e 5.6 apresentam as dispersões entre os parâmetros de discriminação e os de dificuldade, por grupos de disciplinas.

Na dimensão da propensão θ_1, quase a totalidade dos itens apresentam poder razoável de discriminação ($a_1 > 1$), mas com baixo grau de dificuldade $b_1 < 0$. Isso sugere que, de um modo geral, os itens propiciam informação acerca daqueles indivíduos que estão menos propensos a responderem aos itens de modo divergente, porém mediante não-respostas. Em contraste, no que se refere à proficiência θ_2, os itens geralmente apresentam poder de discriminação de menor magnitude, restando alguns itens com algum poder para discriminar os candidatos mais proficientes (e.g., como os itens 20, 77, 78, 82, 83, 95, 96, 118 e 119, que possuem $a_2 > 1$ e $b_2 > 0$).

Para exemplificar, a Figura 5.7 reproduz os itens 83 e 84 (Grupo III), cujas respostas são C, conforme o gabarito oficial definitivo da prova. As estimativas dos parâmetros para o item 83 foram $a_1 = 2,35$, $b_1 = 0,27$, $a_2 = 1,02$ e $b_2 = 0,72$, e para o item 84 foram

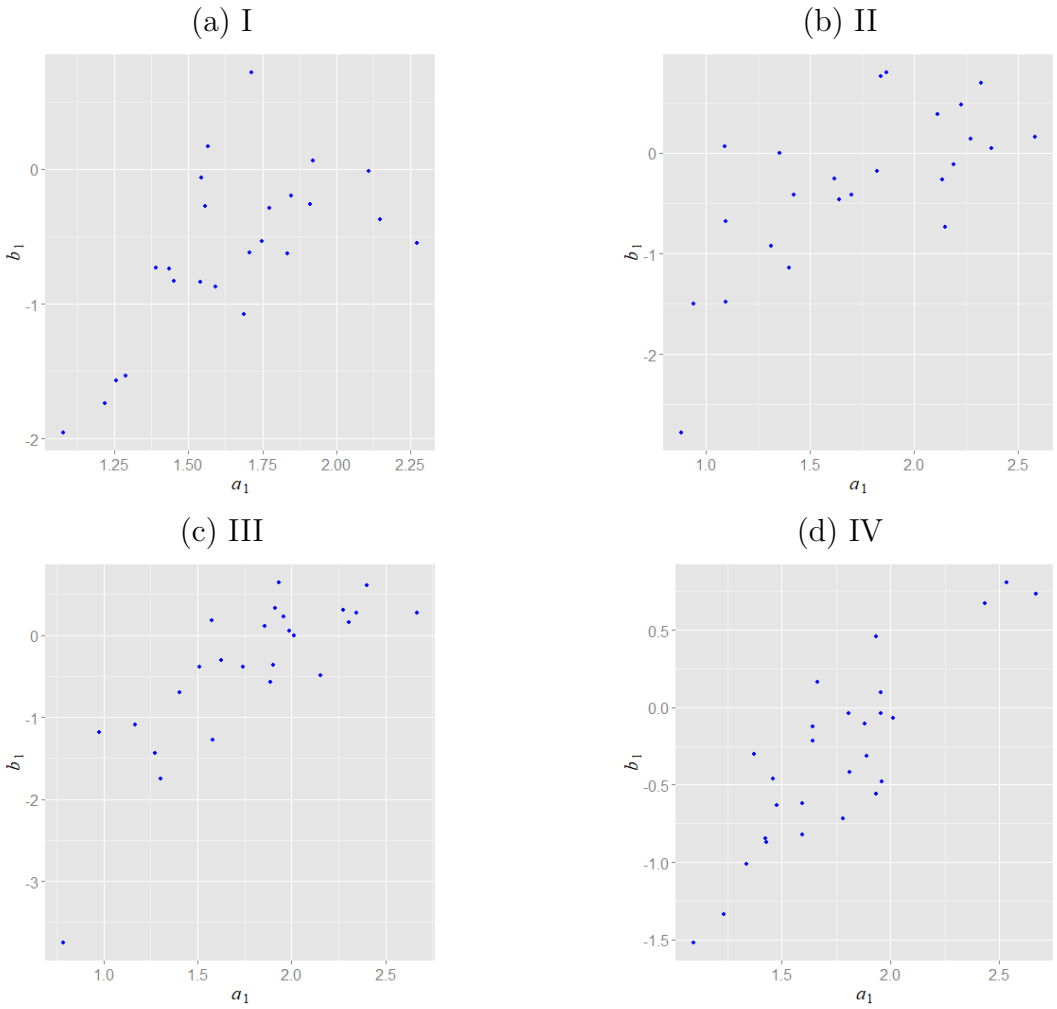


Figura 5.5: Dispersões dos parâmetros dos itens por grupos de disciplinas referentes à propensão θ_1 (I = filosofia, geografia, história, língua portuguesa, e sociologia; II = física e matemática, III = biologia e química, IV = artes cênicas, artes visuais e literatura).

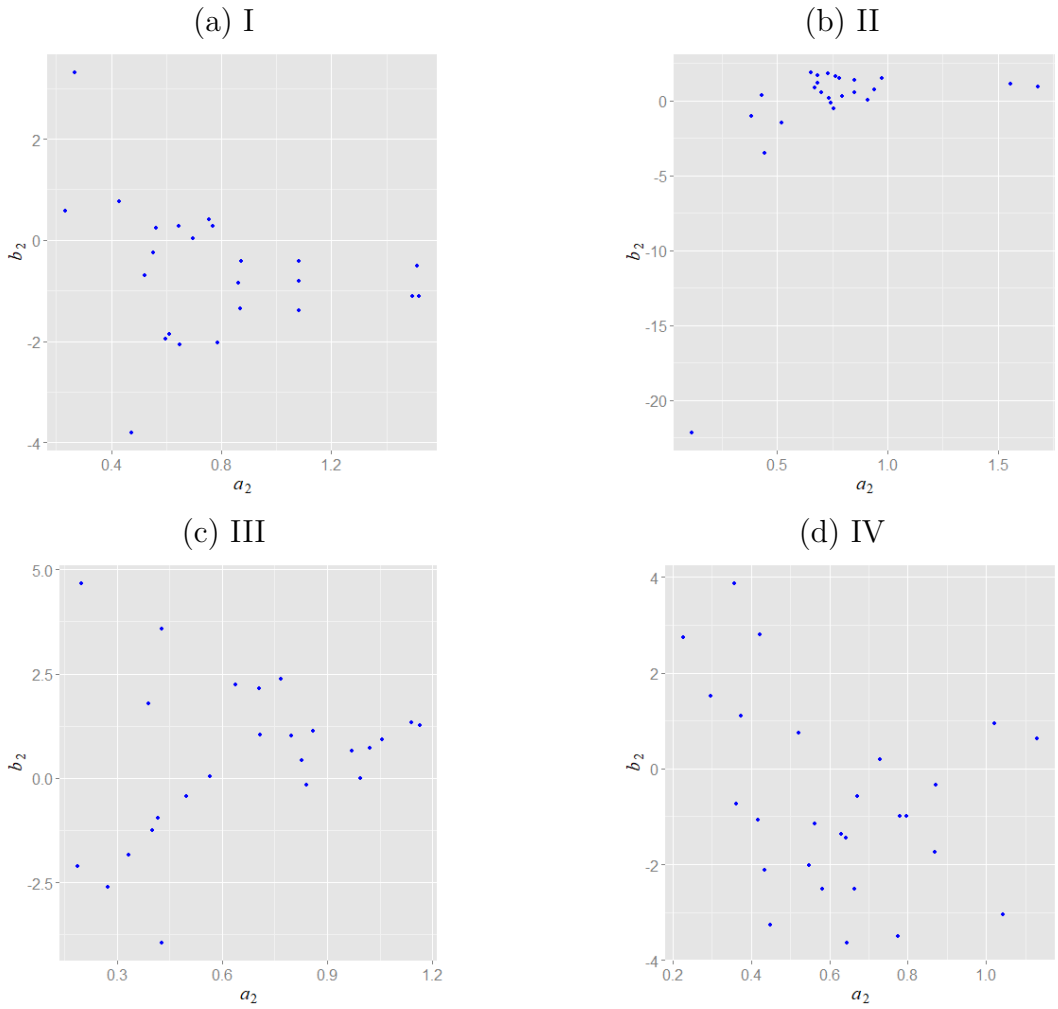


Figura 5.6: Dispersões dos parâmetros dos itens por grupos de disciplinas, referentes à proficiência θ_2 (I = filosofia, geografia, história, língua portuguesa, e sociologia; II = física e matemática, III = biologia e química, IV = artes cênicas, artes visuais e literatura).

As figuras I e II, a seguir, representam, respectivamente, uma substância presente no óleo essencial de rosas e uma substância, a putrescina, exalada por cadáveres em decomposição.

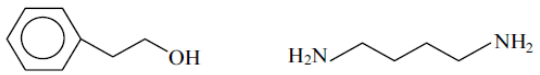


Figura I

Figura II

83 A oxidação da substância representada na figura I, por reação com $KMnO_4$ em meio ácido, permite a obtenção de um aldeído.

84 Em solução aquosa ácida, a putrescina apresenta-se como cátion divalente, produzindo campo elétrico resultante.

Figura 5.7: Itens 83 e 84, terceira etapa do PAS/UnB, subprograma 2006-2008

Tabela 5.2 - Resultados relativos aos itens do grupo I (filosofia, geografia, história, língua portuguesa, e sociologia)

	estimativas dos parâmetros				frequências (%) esperadas			observadas			medidas de proximidade	
item	a_1	b_1	a_2	b_2	$\hat{\pi}_{00}$	$\hat{\pi}_{01}$	$\hat{\pi}_{11}$	$\tilde{\pi}_{00}$	$\tilde{\pi}_{01}$	$\tilde{\pi}_{11}$	D_B	$D_{B,0}$
11	1,43	-0,74	0,87	-0,42	13,6	17,0	69,4	14,0	28,5	57,5	0,0101	2.0×10^{-5}
12	1,54	-0,84	1,08	-0,81	9,7	17,9	72,4	9,7	23,4	66,8	0,0024	2.7×10^{-8}
13	1,29	-1,54	0,77	0,27	9,4	7,0	83,5	9,9	45,0	45,1	0,1183	2.7×10^{-5}
23	1,85	-0,20	0,79	-2,02	8,9	35,2	55,9	9,9	9,6	80,4	0,0525	1.5×10^{-4}
24	1,70	-0,62	0,47	-3,80	5,1	27,0	67,9	5,9	9,4	84,7	0,0282	1.6×10^{-4}
25	1,75	-0,54	0,52	-0,69	14,8	19,5	65,7	16,3	25,5	58,3	0,0033	2.1×10^{-4}
26	1,84	-0,63	1,51	-0,50	12,6	18,8	68,5	12,5	24,8	62,7	0,0027	1.3×10^{-6}
27	2,15	-0,38	1,52	-1,10	9,6	28,5	61,9	9,9	13,4	76,7	0,0181	1.6×10^{-5}
28	2,11	-0,02	0,61	-1,86	13,2	36,4	50,3	15,3	10,6	74,1	0,0516	4.4×10^{-4}
30	1,71	0,72	0,70	0,03	35,8	34,5	29,7	37,9	12,8	49,3	0,0399	2.3×10^{-4}
31	1,69	-1,08	1,08	-1,39	5,2	15,8	79,0	5,2	17,5	77,3	0,0003	5.5×10^{-8}
32	2,27	-0,55	1,49	-1,11	8,3	24,0	67,7	8,3	15,0	76,7	0,0067	6.2×10^{-8}
36	1,39	-0,73	1,08	-0,42	13,5	17,7	68,8	14,0	27,2	58,7	0,0072	3.1×10^{-5}
37	1,91	-0,26	0,65	-2,07	9,9	32,2	58,0	10,9	11,8	77,3	0,0327	1.4×10^{-4}
38	1,77	-0,29	0,60	-1,95	10,8	30,6	58,5	12,3	13,0	74,7	0,0241	2.7×10^{-4}
39	1,92	0,06	0,86	-0,85	18,6	33,4	48,0	20,4	14,5	65,1	0,0264	2.5×10^{-4}
45	1,59	-0,87	0,55	-0,25	12,9	13,5	73,6	14,2	32,8	53,0	0,0300	1.8×10^{-4}
52	1,56	-0,27	0,76	0,41	24,8	17,6	57,6	25,7	31,3	43,0	0,0153	5.7×10^{-5}
62	1,56	0,16	0,87	-1,36	14,7	40,0	45,3	15,9	10,7	73,4	0,0659	1.4×10^{-4}
67	1,22	-1,74	0,43	0,77	8,6	5,9	85,5	9,3	48,6	42,0	0,1529	8.4×10^{-5}
89	1,08	-1,95	0,56	0,23	7,6	6,2	86,2	7,7	45,5	46,8	0,1279	3.7×10^{-6}
90	1,26	-1,57	0,65	0,27	9,2	7,2	83,6	9,5	44,7	45,8	0,1146	1.1×10^{-5}
114	1,45	-0,83	0,27	3,31	20,2	8,1	71,7	22,2	48,4	29,3	0,1404	3.1×10^{-4}
115	1,54	-0,06	0,23	0,57	25,9	22,3	51,8	28,6	24,8	46,6	0,0014	4.5×10^{-4}

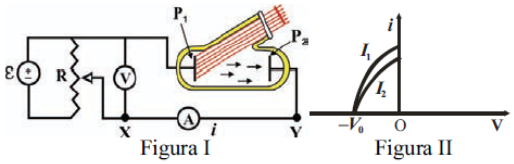
$a_1 = 2,40$, $b_1 = 0,61$, $a_2 = 0,71$ e $b_2 = 2,15$. Na dimensão da propensão, esses itens possuem boa discriminação e parâmetros de dificuldade positivos, o que sugere que eles fornecem alguma informação acerca de candidatos mais propensos a responderem aos itens de modo divergente contra a opção de não responderem. Para esses itens, a Tabela 5.6 mostra uma comparação entre a distribuição percentual conjunta das variáveis R_{ij} e U_{ij} obtida empiricamente (valores entre parênteses) e a distribuição esperada correspondente com base no modelo (3.7). As medidas de proximidade D_B entre a distribuição esperada e a empírica foram, respectivamente, 0,0011 e 0,0087, o que sugere boa aderência desses itens ao modelo ajustado. E, considerando apenas as frequências relativas a π_{00} e seu

Tabela 5.3 - Resultados relativos aos itens do grupo II (física e matemática)

	estimativas dos parâmetros				frequências (%) esperadas			frequências (%) observadas			medidas de proximidade	
item	a_1	b_1	a_2	b_2	$\hat{\pi}_{00}$	$\hat{\pi}_{01}$	$\hat{\pi}_{11}$	$\tilde{\pi}_{00}$	$\tilde{\pi}_{01}$	$\tilde{\pi}_{11}$	D_B	$D_{B,0}$
15	1,10	-1,48	0,11	-22,14	1,6	18,8	79,6	1,5	6,2	92,2	0,0194	8.9×10^{-6}
16	0,88	-2,77	0,38	-1,01	4,4	5,5	90,1	4,4	36,3	59,2	0,0876	1.6×10^{-6}
17	0,94	-1,50	0,43	0,34	13,0	9,8	77,2	13,3	40,2	46,5	0,0734	1.1×10^{-5}
18	1,09	0,07	0,67	0,89	34,7	17,3	48,1	35,0	28,1	36,9	0,0103	7.7×10^{-6}
55	1,70	-0,42	0,75	-0,50	18,7	20,0	61,2	18,8	22,8	58,4	0,0006	1.1×10^{-6}
56	1,87	0,80	0,94	0,74	49,9	23,2	26,8	51,4	12,9	35,8	0,0109	9.9×10^{-5}
57	1,40	-1,14	0,68	1,69	18,1	4,6	77,3	18,5	55,6	25,9	0,2349	1.0×10^{-5}
59	1,64	-0,47	0,77	1,66	30,4	7,1	62,5	30,8	44,9	24,4	0,1339	7.1×10^{-6}
60	1,62	-0,25	0,65	1,89	34,8	8,8	56,4	35,9	39,7	24,4	0,0927	5.7×10^{-5}
64	1,31	-0,92	0,44	-3,49	5,9	22,3	71,8	5,9	12,5	81,6	0,0086	1.2×10^{-8}
65	1,36	0,00	0,52	-1,45	18,1	32,5	49,4	18,6	14,2	67,1	0,0257	2.7×10^{-5}
66	1,09	-0,68	0,70	0,53	22,5	13,0	64,5	22,2	36,1	41,7	0,0422	8.0×10^{-6}
70	2,19	-0,12	0,91	0,05	28,0	19,7	52,3	27,5	23,4	49,1	0,0010	1.4×10^{-5}
71	2,37	0,05	0,74	0,17	31,3	22,0	46,7	33,2	19,5	47,3	0,0005	2.0×10^{-4}
72	2,27	0,14	0,85	0,58	37,5	18,7	43,8	38,5	21,9	39,6	0,0012	5.6×10^{-5}
73	2,58	0,16	0,79	0,28	35,0	22,6	42,4	36,6	18,1	45,3	0,0015	1.3×10^{-4}
74	2,15	-0,74	0,78	1,49	22,6	5,2	72,1	23,7	50,0	26,2	0,1883	8.4×10^{-5}
75	2,14	-0,27	0,73	1,85	35,2	7,6	57,3	36,7	40,4	22,9	0,1098	1.2×10^{-4}
76	1,82	-0,18	0,74	-0,13	24,6	20,9	54,5	25,5	22,3	52,2	0,0003	6.0×10^{-5}
77	2,12	0,38	1,56	1,10	53,1	10,2	36,7	52,1	24,9	22,9	0,0244	4.5×10^{-5}
78	2,32	0,69	1,68	0,91	58,2	14,2	27,6	57,6	16,2	26,2	0,0004	1.5×10^{-5}
79	2,23	0,47	0,85	1,40	51,7	14,5	33,7	52,7	21,1	26,2	0,0055	4.4×10^{-5}
94	1,42	-0,42	0,68	1,21	28,9	10,7	60,4	28,9	38,9	32,2	0,0680	7.4×10^{-8}
102	1,84	0,76	0,97	1,52	58,7	13,5	27,8	59,5	18,2	22,3	0,0033	3.6×10^{-5}

complementar, as distâncias de Bhattacharyya correspondentes foram $D_{B,0} = 4.6 \times 10^{-5}$ e 3.1×10^{-4}, indicando boa aderência das frequências observadas de não-respostas ao modelo ajustado.

Como segundo exemplo, a Figura 5.8 apresenta um conjunto de seis itens (grupo II) propostos sob o mesmo comando (itens 70 a 75). De acordo com o gabarito oficial definitivo, apenas o item 70 deve ser assinalado como E, e os restantes são C. As estimativas dos parâmetros desses itens se encontram na Tabela 5.3. Na dimensão da proficiência θ_2 esses itens apresentam poder de discriminação moderado ($0,73 \leq a_2 \leq 0,91$) e dificuldade positiva ($0 \leq b_2 \leq 1,85$). Com respeito à propensão θ_1 esses itens possuem poder de discriminação elevado ($2,14 \leq a_1 \leq 2,58$) e parâmetros de dificuldade $-0,74 \leq b_1 \leq 0,16$.



É possível transformar energia luminosa em energia elétrica por meio do efeito fotoelétrico. O diagrama mostrado na figura I esquematiza uma célula fotoelétrica. Nesse esquema, um raio luminoso monocromático incide na placa metálica P_1 da célula, retirando elétrons, que se deslocam para a outra placa — P_2 —, o que gera uma corrente elétrica i, que pode ser medida por um amperímetro. Com o objetivo de se determinar a velocidade do elétron mais veloz emitido pela placa iluminada, um potencial reverso — inserido para desacelerar o movimento dos elétrons — é aplicado entre as placas metálicas, com auxílio de uma fonte ε. A figura II apresenta o comportamento da corrente elétrica i, medida como função desse potencial reverso para duas situações de intensidades de iluminação, que permitiram obter os gráficos I_1 e I_2. Com base nessas informações, julgue os itens que se seguem.

70 Uma das exigências para que o amperímetro não interfira na medida é que ele tenha resistência infinita.

71 Mesmo não havendo potencial aplicado entre as placas metálicas (V = 0), há geração de corrente elétrica ($i \neq 0$) devido à luz incidente na placa metálica.

72 No esquema da figura I, se R = 0, ainda é possível que a corrente medida pelo amperímetro seja não-nula.

73 É correto inferir que o gráfico I_1 está associado a uma intensidade de luz maior que a intensidade de luz associada ao gráfico I_2.

74 No circuito ilustrado na figura I, a corrente elétrica i segue no sentido do ponto X para o ponto Y.

75 Infere-se dos gráficos I_1 e I_2 que a velocidade do elétron mais veloz independe da intensidade de luz que incide na placa P_1.

Figura 5.8: Itens de 70 a 75 (Grupo II), terceira etapa do PAS/UnB, subprograma 2006-2008

A Tabela 5.7 mostra a distribuição conjunta esperada de acordo com o modelo (3.7) e a distribuição observada para cada um desses itens. Com exceção dos itens 74 e 75, as medidas de proximidade D_B entre essas distribuições foram inferiores a 0,0015 (Tabela 5.3). Nos itens 74 e 75, que apresentaram $D_B > 0,10$, os percentuais observados de acertos foram bem menores do que os percentuais esperados correspondentes. Porém, os percentuais observados de não-respostas para esses itens foram todos muito próximos aos seus valores esperados ($D_{B,0} \leq 2.0 \times 10^{-4}$).

Alguns itens anômalos podem ser identificados nas Tabelas 5.3 e 5.4. A Figura 5.9 reproduz dois desses casos, os itens 15 e 35 (ambos com respostas E). Na dimensão da proficiência θ_2 o item 15 possui parâmetros com valores muito baixos ($a_2 = 0,11$ e $b_2 = -22,14$), o que o caracteriza como item ruim. Já o item 35 possui parâmetros $a_1 = 0,79$, $b_1 = -3,74$, $a_2 = 0,43$ e $b_2 = 3,58$, o que remete a um caso de baixo poder de discriminação, tanto para na dimensão da propensão como na da proficiência, de grande dificuldade na dimensão θ_2, e com b_1 baixo, o que proporciona maior chance de respondê-lo incorretamente em relação à possibilidade de ele não ser respondido. Porém, esse item apresenta problema de aderência entre o percentual esperado de acerto (93,9%) e o que foi observado (18,6%). (Embora parte da cobrança desse item diga respeito à obra *Operários* de Tarsila do Amaral, a parte crucial refere-se ao conhecimento sobre *classificação biológica*. Por isso, este item foi classificado como parte do Grupo III).

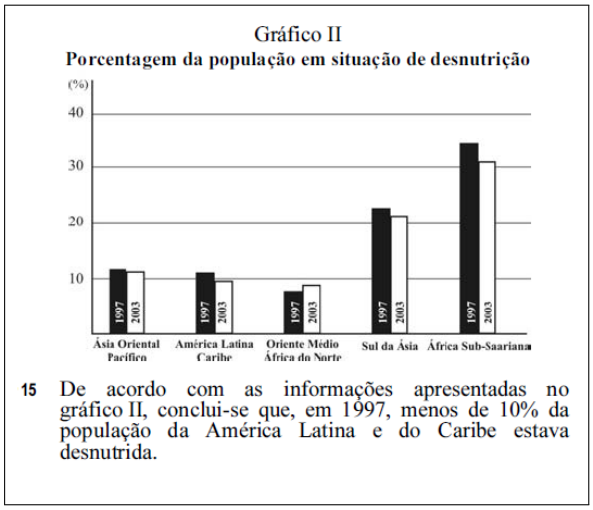


Figura 5.9: Itens 15 (Grupo II) e 35 (Grupo III), terceira etapa do PAS/UnB, subprograma 2006-2008

A Figura 5.10 mostra a distribuição dos valores de D_B, em que a linha vertical separa o grupo de 40 itens considerados aderentes ao modelo ajustado dos demais, sendo 8 do grupo I, 12 do grupo II, 11 do grupo III e 9 do grupo IV. Esses valores são tais que $D_B < 0,015$, sendo seus p-valores obtidos pelo método de Monte Carlo inferiores a 5%. Esses itens apresentam frequências esperadas $\hat{\pi}_{11}$ e $\hat{\pi}_{01}$ mais próximas das frequências correspondentes observadas ($\tilde{\pi}_{11}$ e $\tilde{\pi}_{01}$), tendo valores D_B próximos de zero.

Com respeito aos valores relativos à similaridade entre a fração esperada e a observada das não-respostas, $D_{B,0}$, todos foram menores que 0,0005, sendo estatisticamente nulos com p-valores inferiores a 1%. Observe, na Figura 5.11, que os percentuais de um modo geral se encontram próximos da linha $\tilde{\pi}_{00} = \hat{\pi}_{00}$. Isso sugere que a não-resposta seja um fenômeno fortemente dependente das características do item e de um traço latente do respondente. Mas quanto aos percentuais de acerto, há grande dispersão dos pontos em torno da linha de referência $\tilde{\pi}_{11} = \hat{\pi}_{11}$ (Figura 5.12). Por exemplo, nos itens do Grupo II (física e matemática), há uma quantidade expressiva de itens cujos percentuais observados de acerto ($\tilde{\pi}_{11}$) são inferiores aos valores esperados.

A Figura 5.13 mostra a distribuição conjunta entre θ_1 e θ_2 para cada grupo de disciplinas. Ela evidencia a existência de pelo menos dois grupos de candidatos. A concentração de pontos na parte superior de cada distribuição referem-se àqueles candidatos que não deixaram respostas em branco. A massa restante de pontos corresponde aos demais candidatos. Na dispersão das proficiências associadas ao grupo IV, é possível observar a

existência de outros aglomerados de pontos. Como se espera, os gráficos mostram que os candidatos menos proficientes (θ_2 baixo) tendem a ser menos propensos a responder de forma divergente, preferindo deixar a resposta em branco (θ_1 baixo). Mas à medida que a proficiência aumenta, a propensão tende a se concentrar na região modal da distribuição.

Tabela 5.4 - Resultados relativos aos itens do grupo III (biologia e química)

	estimativas dos parâmetros				frequências (%) esperadas			frequências (%) observadas			medidas de proximidade	
item	a_1	b_1	a_2	b_2	$\hat{\pi}_{00}$	$\hat{\pi}_{01}$	$\hat{\pi}_{11}$	$\tilde{\pi}_{00}$	$\tilde{\pi}_{01}$	$\tilde{\pi}_{11}$	D_B	$D_{B,0}$
19	1,17	-1,09	0,40	-1,26	10,8	15,5	73,7	10,8	27,3	61,9	0,0110	1.1×10^{-6}
20	1,57	0,18	1,16	1,27	45,8	10,1	44,1	46,0	30,4	23,5	0,0443	3.7×10^{-6}
21	1,86	0,11	0,84	-0,15	28,7	25,9	45,3	28,5	18,7	52,8	0,0044	4.9×10^{-6}
22	1,27	-1,43	0,28	-2,61	6,7	12,2	81,1	6,2	26,8	67,0	0,0178	5.2×10^{-5}
35	0,79	-3,74	0,43	3,58	5,1	1,0	93,9	5,1	76,3	18,6	0,5869	2.2×10^{-7}
82	1,74	-0,38	1,00	0,00	23,0	16,7	60,2	21,8	28,0	50,2	0,0097	1.1×10^{-4}
83	2,35	0,27	1,02	0,72	43,0	17,7	39,3	43,9	20,6	35,4	0,0011	4.6×10^{-5}
84	2,40	0,61	0,71	2,15	58,1	12,3	29,6	60,5	19,3	20,2	0,0087	3.1×10^{-4}
86	2,01	0,00	0,50	-0,43	25,1	26,4	48,5	26,7	18,3	55,0	0,0049	1.7×10^{-4}
87	1,99	0,05	0,64	2,25	43,7	9,3	46,9	45,4	33,6	21,0	0,0655	1.4×10^{-4}
88	1,63	-0,31	0,86	1,13	32,0	10,2	57,8	32,8	37,0	30,2	0,0660	3.4×10^{-5}
91	2,27	0,31	0,97	0,66	42,4	19,2	38,4	43,8	19,1	37,1	0,0001	1.0×10^{-4}
92	2,31	0,16	0,77	2,37	49,7	7,4	42,9	51,2	32,6	16,3	0,0797	1.0×10^{-4}
93	2,67	0,28	0,71	1,04	43,4	18,1	38,5	46,3	19,7	34,0	0,0011	4.3×10^{-4}
97	1,58	-1,27	0,33	-1,83	7,3	11,4	81,3	6,7	28,8	64,4	0,0253	6.0×10^{-5}
98	1,51	-0,39	0,42	-0,95	17,6	22,6	59,8	18,8	21,8	59,4	0,0001	1.3×10^{-4}
99	1,94	0,64	0,80	1,03	49,4	20,3	30,4	50,1	17,2	32,8	0,0009	2.6×10^{-5}
100	2,15	-0,48	0,43	-3,93	6,7	29,1	64,2	5,8	10,8	83,5	0,0293	2.1×10^{-4}
101	1,30	-1,75	0,19	-2,11	5,8	7,9	86,3	5,9	34,4	59,6	0,0609	3.1×10^{-6}
103	1,91	0,33	0,57	0,05	32,9	28,1	38,9	34,2	16,4	49,4	0,0112	8.8×10^{-5}
104	1,40	-0,70	0,39	1,78	22,6	10,0	67,4	23,3	42,9	33,8	0,0901	3.2×10^{-5}
111	0,98	-1,19	0,20	4,67	19,8	7,4	72,8	20,6	50,9	28,5	0,1608	5.0×10^{-5}
118	1,90	-0,36	1,06	0,93	31,0	9,3	59,8	30,9	38,0	31,1	0,0745	4.5×10^{-8}
119	1,96	0,23	1,14	1,33	48,3	10,1	41,6	48,4	28,7	22,9	0,0382	1.5×10^{-6}
120	1,89	-0,58	0,83	0,43	22,1	11,7	66,3	22,0	35,5	42,5	0,0465	1.9×10^{-7}

Tabela 5.5 - Resultados relativos aos itens do grupo IV (artes cênicas, artes visuais e literatura)

	estimativas dos parâmetros				frequências (%) esperadas			frequências (%) observadas			medidas de proximidade	
item	a_1	b_1	a_2	b_2	$\hat{\pi}_{00}$	$\hat{\pi}_{01}$	$\hat{\pi}_{11}$	$\tilde{\pi}_{00}$	$\tilde{\pi}_{01}$	$\tilde{\pi}_{11}$	D_B	$D_{B,0}$
33	1,81	-0,42	0,77	-3,51	3,4	34,9	61,7	3,5	4,3	92,2	0,0926	1.8×10^{-6}
34	1,43	-0,87	0,43	-2,12	8,9	19,2	71,9	9,2	20,0	70,8	0,0001	1.8×10^{-5}
40	1,43	-0,85	0,30	1,51	18,1	10,7	71,3	18,2	42,6	39,2	0,0800	1.5×10^{-6}
41	1,60	-0,62	0,58	-2,52	7,5	25,8	66,7	7,9	12,4	79,7	0,0153	3.2×10^{-5}
42	1,34	-1,01	0,37	1,11	16,1	9,7	74,2	16,6	43,3	40,1	0,0902	2.6×10^{-5}
43	1,48	-0,63	0,42	-1,07	14,1	19,6	66,3	14,6	24,8	60,6	0,0022	2.7×10^{-5}
44	1,09	-1,52	0,36	3,87	16,0	3,7	80,3	16,7	62,7	20,6	0,3245	4.6×10^{-5}
46	1,78	-0,72	0,45	-3,27	6,5	23,3	70,2	6,5	13,1	80,4	0,0091	8.5×10^{-7}
47	1,88	-0,11	0,66	-2,52	9,2	38,4	52,4	9,9	7,7	82,3	0,0783	7.3×10^{-5}
48	1,95	-0,04	0,63	-1,36	16,8	33,1	50,1	17,2	13,9	68,8	0,0282	1.6×10^{-5}
49	1,64	-0,12	0,52	0,74	29,1	18,0	52,9	29,8	29,3	41,0	0,0107	2.3×10^{-5}
50	1,24	-1,34	0,23	2,75	13,8	6,9	79,3	15,0	50,0	35,0	0,1550	1.5×10^{-4}
51	1,60	-0,82	0,36	-0,73	13,0	15,1	71,9	13,3	30,4	56,4	0,0181	8.3×10^{-6}
61	1,94	0,46	0,78	-0,98	23,0	41,6	35,4	24,0	9,6	66,3	0,0838	7.5×10^{-5}
68	1,67	0,16	0,67	-0,57	24,4	31,1	44,6	25,7	15,7	58,6	0,0179	1.1×10^{-4}
80	1,38	-0,30	0,42	2,81	33,0	9,5	57,5	33,7	42,2	24,1	0,0986	2.1×10^{-5}
85	1,46	-0,46	0,64	-1,44	12,5	25,7	61,8	13,1	16,9	70,0	0,0059	4.3×10^{-5}
95	2,67	0,73	1,02	0,94	53,4	20,8	25,8	54,4	14,4	31,2	0,0042	5.0×10^{-5}
96	2,54	0,81	1,13	0,63	50,8	25,0	24,2	52,5	11,3	36,2	0,0197	1.6×10^{-4}
105	1,64	-0,22	0,56	-1,15	16,9	27,6	55,5	17,3	18,1	64,6	0,0068	1.3×10^{-5}
106	2,43	0,67	0,87	-0,35	33,0	39,1	27,9	33,8	9,8	56,5	0,0769	3.5×10^{-5}
109	1,96	-0,48	0,64	-3,63	4,2	31,9	64,0	4,4	5,8	89,8	0,0654	1.2×10^{-5}
110	1,81	-0,04	0,87	-1,74	11,6	38,2	50,2	12,3	8,8	78,8	0,0700	6.4×10^{-5}
112	1,94	-0,56	0,55	-2,02	9,8	23,9	66,3	9,8	16,3	73,9	0,0048	1.7×10^{-7}
113	1,89	-0,31	1,04	-3,05	2,9	38,3	58,8	3,0	3,2	93,9	0,1255	3.6×10^{-6}
116	2,01	-0,07	0,80	-0,99	17,9	31,0	51,2	18,8	14,5	66,8	0,0210	6.8×10^{-5}
117	1,96	0,10	0,73	0,20	30,7	23,4	45,9	31,8	21,5	46,7	0,0002	6.3×10^{-5}

Tabela 5.6 - Percentuais esperados para os itens 83 e 84, com base no modelo (3.7), e seus respectivos percentuais empíricos (entre parênteses) relativos às variáveis R_{ij} (respondeu = 1, não respondeu = 0) e U_{ij} (acertou = 1, não acertou = 0).

(a) item 83

U_{ij} \ R_{ij}	0	1
0	43,0% (43,9%)	17,7% (20,6%)
1	0	39,3% (35,4%)

(b) item 84

U_{ij} \ R_{ij}	0	1
0	58,1% (60,5%)	12,3% (19,3%)
1	0	29,6% (20,2%)

Tabela 5.7 - Percentuais esperados para os itens de 70 a 75, com base no modelo (3.7), e seus percentuais empíricos correspondentes (entre parênteses) relativos às variáveis R_{ij} (respondeu = 1, não respondeu = 0) e U_{ij} (acertou = 1, não acertou = 0).

(a) item 70

U_{ij} \ R_{ij}	0	1
0	28,0% (27,5%)	19,7% (23,4%)
1	0	52,3% (49,1%)

(b) item 71

U_{ij} \ R_{ij}	0	1
0	31,3% (33,2%)	22,0% (19,5%)
1	0	46,7% (47,3%)

(c) item 72

U_{ij} \ R_{ij}	0	1
0	37,5% (38,5%)	18,7% (21,9%)
1	0	43,8% (39,6%)

(d) item 73

U_{ij} \ R_{ij}	0	1
0	35,0% (36,6%)	22,6% (18,1%)
1	0	42,4% (45,3%)

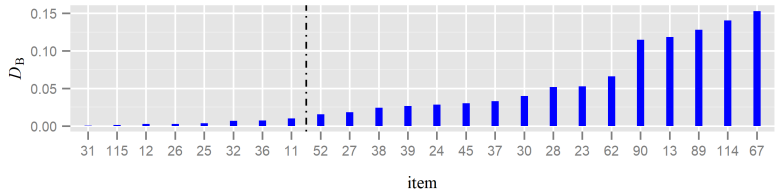
(e) item 74

U_{ij} \ R_{ij}	0	1
0	22,6% (23,7%)	5,2% (50,0%)
1	0	72,1% (26,2%)

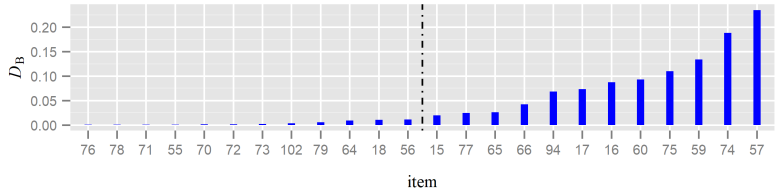
(f) item 75

U_{ij} \ R_{ij}	0	1
0	35,2% (36,7%)	7,6% (40,4%)
1	0	57,3% (22,9%)

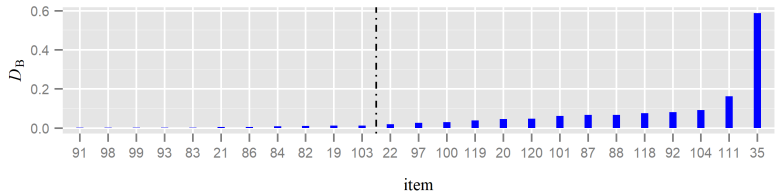
(a) I (filosofia, geografia, história, língua portuguesa, e sociologia)



(b) II (física e matemática)



(c) III (biologia e química)



(d) IV (artes cênicas, artes visuais e literatura)

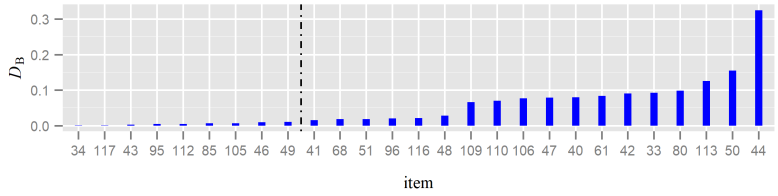


Figura 5.10: Distâncias de Bhattacharyya entre a distribuição observada e a esperada segundo o modelo ajustado, por item e grupos de disciplinas. Valores à esquerda das linhas tracejadas são estatisticamente nulos (p-valores $< 5\%$, por Monte Carlo).

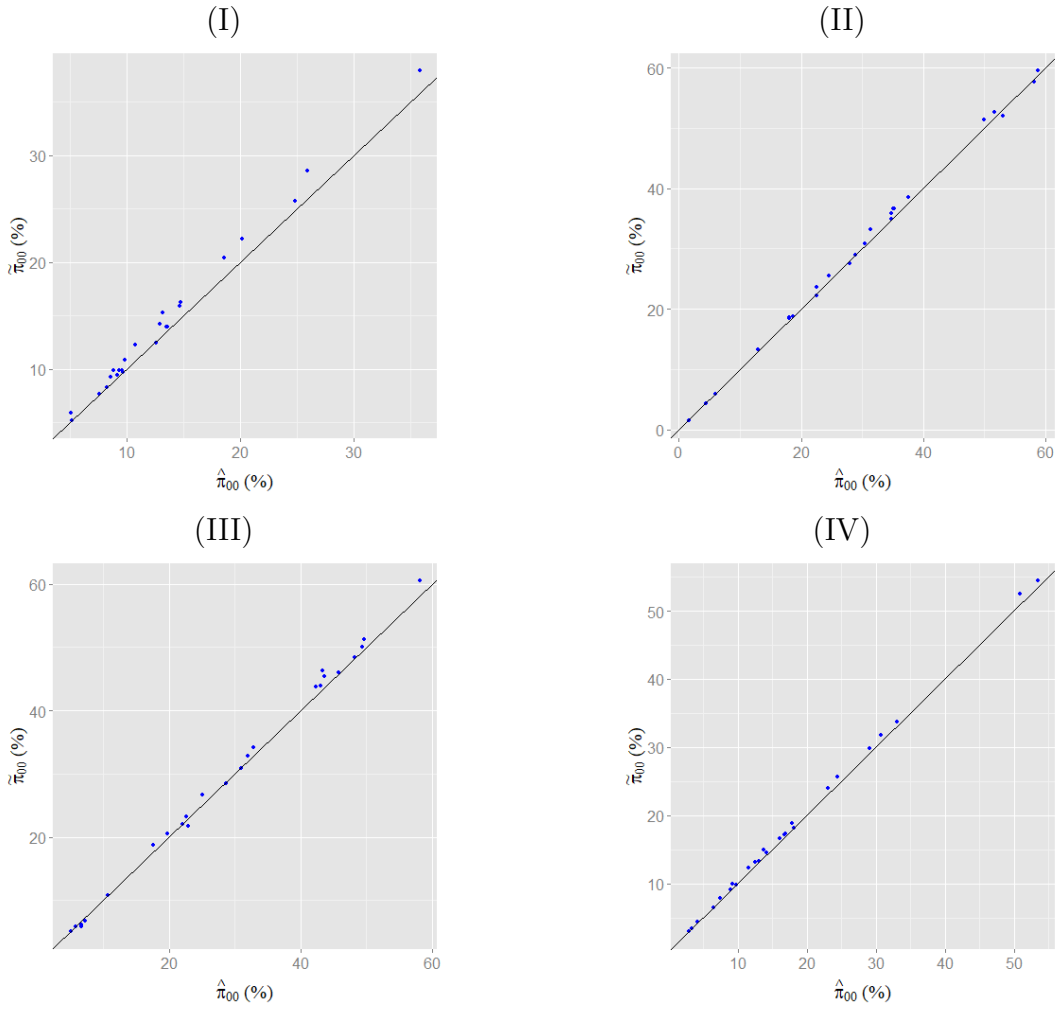


Figura 5.11: Dispersões entre as frequências esperadas de não-respostas ($\hat{\pi}_{00}$) e as observadas ($\tilde{\pi}_{00}$), por grupos de disciplinas (I = filosofia, geografia, história, língua portuguesa, e sociologia; II = física e matemática, III = biologia e química, IV = artes cênicas, artes visuais e literatura). A linha sólida representa o caso $\tilde{\pi}_{00} = \hat{\pi}_{00}$.

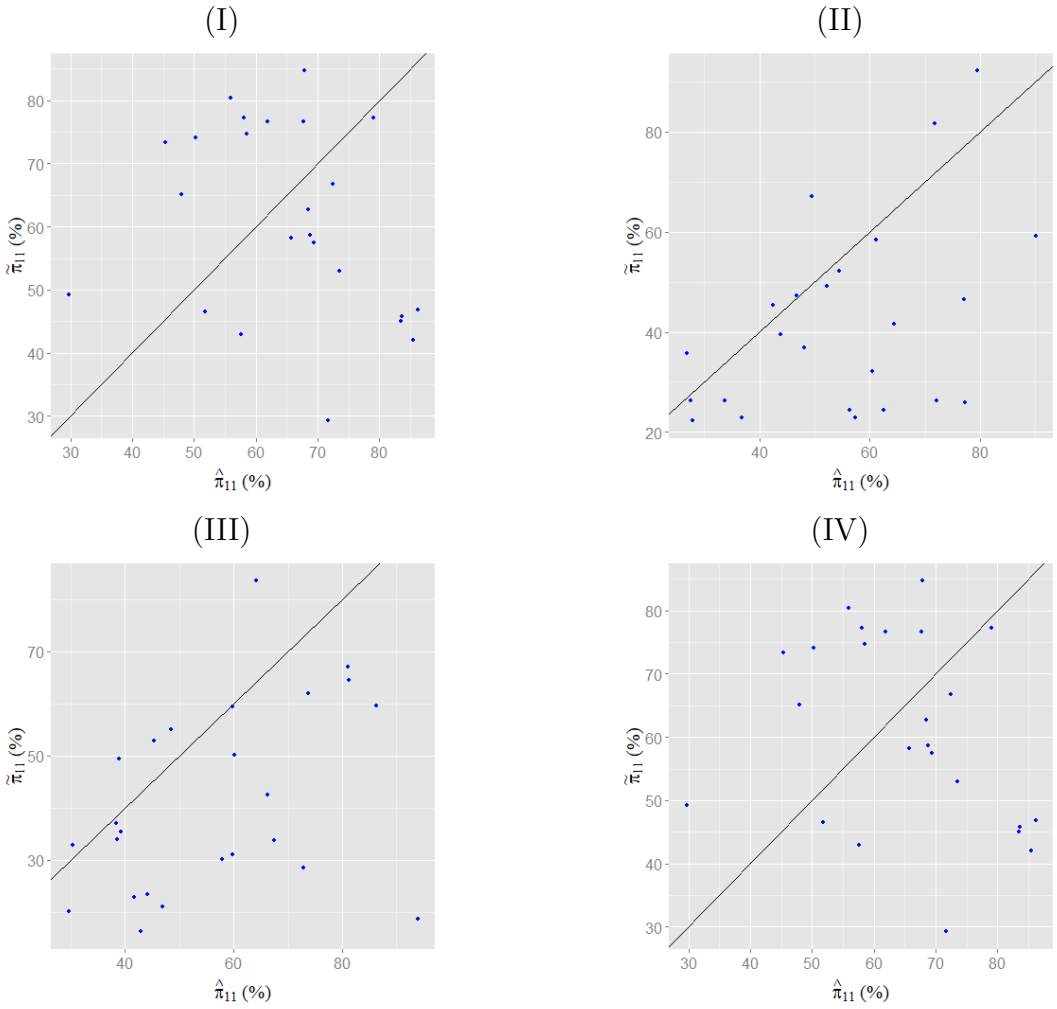


Figura 5.12: Dispersões entre as frequências esperadas de acertos ($\hat{\pi}_{11}$) e as observadas ($\tilde{\pi}_{11}$), por grupos de disciplinas (I = filosofia, geografia, história, língua portuguesa, e sociologia; II = física e matemática, III = biologia e química, IV = artes cênicas, artes visuais e literatura). A linha sólida representa o caso $\tilde{\pi}_{11} = \hat{\pi}_{11}$.

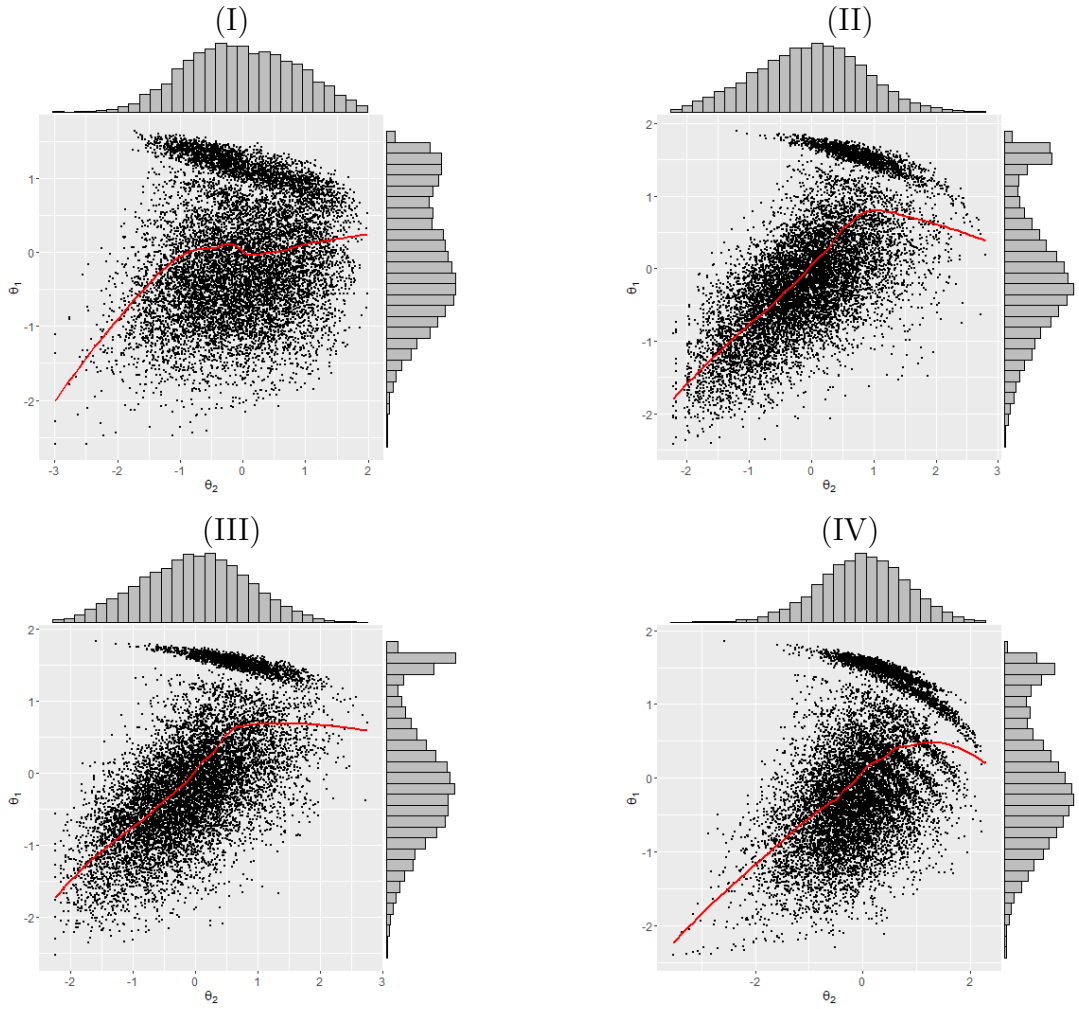


Figura 5.13: Dispersão entre os traços θ_1 e θ_2 por grupos de disciplinas (I = filosofia, geografia, história, língua portuguesa, e sociologia; II = física e matemática, III = biologia e química, IV = artes cênicas, artes visuais e literatura). As linhas sólidas (vermelhas) representam as médias condicionais $\theta_1|\theta_2$ ajustadas não parametricamente pelo método LOESS.

Capítulo 6

Conclusão

6.1 Considerações finais

Este trabalho apresentou um modelo bidimensional que descreve, conjuntamente, as variáveis dicotômicas que representam o acerto (U_{ij}) e a resposta (R_{ij}) referentes a um item i e um examinando j. Esse modelo preescreve uma distribuição para a variável U_{ij} e outra para $R_{ij}|U_{ij} = 0$, em função de traços latentes e de parâmetros do item (dificuldade e discriminação). Essa distribuição condicional diz respeito à propensão de o indivíduo j apresentar resposta ao item i que divirja do gabarito oficial contra a possibilidade de ele deixar o item em branco. Em nosso estudo, essas duas situações remetem ao evento $U_{ij} = 0$.

Como contraponto, Rose et al. (2010) discutem que *treating missing data as wrong appears to be the least desirable way to account for responses MNAR (missing not at random) in large-scale surveys.* Para esses autores, de modo natural, os examinandos tendem a deixar em branco os itens difíceis, além de suas capacidades. Dessa forma, os respondentes administrariam seu próprio teste, escolhendo aqueles itens que sejam compatíveis com suas respectivas proficiências.

Porém, em um processo seletivo, por força da competição, é necessário penalizar tanto aqueles examinandos que propuseram respostas divergentes quanto aqueles que não apresentaram respostas. De fato, quando as não-respostas são consideradas respostas erradas, as estimativas da proficiência são viciadas (Rose et al., 2010; Bertoli-Barsotti e Punzo, 2013). No entanto, para fins de seleção de candidatos, o interesse do examinador se volta para aqueles cujas proficiências sejam mais elevadas. Assim, a proficiência estimada com base em um processo seletivo poderia ser diferente daquela estimada em uma avaliação

educacional.

Em geral, as frequências observadas de acerto foram inferiores às suas respectivas frequências esperadas (Figura 5.12). Mesmo assim, 40 itens se mostraram aderentes ao modelo ajustado (de um total de 100). Esse resultado não é surpreendente, pois a prova do PAS/UnB não foi desenhada sob a perspectiva da TRI, sendo também processo seletivo. De fato, Hambleton et al. (1991) discute que *item responde models, unlike the classical true score model, are 'falsifiable' models* e, por isso, é justificável que haja um conjunto particular de itens que não seja adequadamente descrito por um modelo da TRI.

Este ensaio também sugere que a distribuição conjunta dos traços latentes (θ_1, θ_2) apresenta uma estrutura de dependência não linear, havendo possíveis misturas de distribuições (ou grupos de candidatos). Por exemplo, a presença de examinandos que não deixaram respostas em branco é refletida mediante concentração de pontos na distribuição conjunta. Além disso, observou-se que os candidatos menos proficientes sejam menos propensos a responderem de forma divergente, pois tendem a deixar a resposta em branco. À medida que a proficiência aumenta, porém, a propensão tende a se concentrar na região modal da distribuição.

Por fim, este trabalho mostrou que a resposta em branco deve ser tratada como uma informação não ignorável, sendo relacionada não apenas com a baixa proficiência do respondente, mas também com as características do item e o traço θ_1. Ele representa a propensão de um indivíduo apresentar resposta divergente contra a possibilidade de ele não responder. Com respeito a essa dimensão, todos os itens da prova mostraram-se aderentes ao modelo ajustado, com base na distância de Bhattacharyya. Possivelmente, há uma relação muito forte entre características individuais (além do próprio item) e as não-respostas. Para estudos futuros, sugere-se a execução de investigações comportamentais para se examinar detalhadamente essa possibilidade (Evans, 2008; Da Silva et al., 2015). Os dados utilizados neste trabalho se encontram disponíveis em `https://1drv.ms/f/s!Apx60k7TMXzegYhBVRKeIjqgLcXixA`.

Referências Bibliográficas

Abdelfattah F. A., Response latency effects on classical and item response theory parameters using different scoring procedures, Ohio University, 2007, Tese de Doutorado

Albanese M., Knott M., TWOMISS: a computer program for fitting a one or two-factor logit probit latent model to binary data when observations may be missing, London School of Economics. Technical Report, Statistics Department, 1992

Andersen E. B., Asymptotic properties of conditional maximum-likelihood estimators, Journal of the Royal Statistical Society. Series B (Methodological), 1970, pp 283–301

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

Atchison J., Shen S. M., Logistic-normal distributions: Some properties and uses, Biometrika, 1980, vol. 67, p. 261

Ayala R. J., The theory and practice of item response theory. Guilford Publications, 2009

Baker F. B., Empirical comparison of item parameters based on the logistic and normal functions, Psychometrika, 1961, vol. 26, p. 239

Baker F. B., The basics of item response theory. ERIC, 2001

Baker F. B., Kim S.-H., Item response theory: Parameter estimation techniques. CRC Press, 2004

Bartholomew D. J., De Menezes L. M., Tzamourani P., Latent trait and latent class models applied to survey data, Applications of latent trait and latent class models in the social sciences, 1997, pp 219–232

Beaujean A. A., Latent variable modeling using R: A step-by-step guide. Routledge, 2014

Bertoli-Barsotti L., Punzo A., Rasch analysis for binary data with nonignorable nonresponses, Psicológica, 2013, vol. 34

Birnbaum A., Efficient design and use of tests of a mental ability for various decision-making problems, Randolph Air Force Base, Texas: Air University, School of Aviation Medicine, 1957, vol. 26

Birnbaum A., On the estimation of mental ability, Series Rep, 1958, vol. 15, p. 7755

Bock R. D., Aitkin M., Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm, Psychometrika, 1981, vol. 46, p. 443

Bock R. D., Lieberman M., Fitting a response model forn dichotomously scored items, Psychometrika, 1970, vol. 35, p. 179

Borgatto A. F., de Andrade D. F., Análise Clássica de Testes com diferentes graus de dificuldade, Estudos em Avaliação Educacional, 2012, vol. 23, p. 146

Butler K., Stephens M. A., The Distribution of a Sum of Independent Binomial Random Variables, Methodology and Computing in Applied Probability, 2017, pp 1–15

Culbertson M. J., Is it wrong? Handling missing responses in IRT, 2011

Da Silva S., Matsushita R., De Carvalho M., Prosocial people take better care of their own future well-being, 2015

DeMars C., Item response theory. Oxford University Press, 2010

DeMars C. E., Test stakes and item format interactions, Applied Measurement in Education, 2000, vol. 13, p. 55

Eggen T. J., Verhelst N. D., Item calibration in incomplete testing designs, Psicológica, 2011, vol. 32

Eisinga R., Te Grotenhuis M., Pelzer B., Saddlepoint approximations for the sum of independent non-identically distributed binomial random variables, Statistica Neerlandica, 2013, vol. 67, p. 190

Erguven M., Two approaches to psychometric process: Classical test theory and item response theory, Journal of Education, 2013, vol. 2, p. 23

Evans J. S. B., Dual-processing accounts of reasoning, judgment, and social cognition, Annu. Rev. Psychol., 2008, vol. 59, p. 255

Fan X., Item response theory and classical test theory: An empirical comparison of their item/person statistics, Educational and psychological measurement, 1998, vol. 58, p. 357

Finch W. H., French B. F., Latent variable modeling with R. Routledge, 2015

Glas C. A., Pimentel J. L., Modeling nonignorable missing data processes in item calibration, 2006

Glas C. A., Pimentel J. L., Modeling nonignorable missing data in speeded tests, Educational and Psychological Measurement, 2008, vol. 68, p. 907

Greene J. P., Winters M. A., Forster G., Testing High Stakes Tests: Can We Believe the Results of Accountability Tests? Civic Report., 2003

Hambleton R. K., Swaminathan H., Rogers H. J., Fundamentals of item response theory. vol. 2, Sage, 1991

Hamilton L. S., Stecher B. M., Klein S. P., Making sense of test-based accountability in education. Rand Corporation, 2002

Heckman J. J., Sample Selection Bias as a Specification Error, Econometrica, 1979, vol. 47, p. 153

Holman R., Glas C. A., Modelling non-ignorable missing-data mechanisms with item response theory models, British Journal of Mathematical and Statistical Psychology, 2005, vol. 58, p. 1

Klein R., Alguns aspectos da teoria de resposta ao item relativos à estimação das proficiências, Ensaio: Avaliação e Políticas Públicas em Educação, 2013, vol. 21, p. 35

Klein S. P., Hamilton L., Large-Scale Testing: Current Practices and New Directions., 1999

Knott M., Albanese M. T., Galbraith J., Scoring attitudes to abortion, The Statistician, 1990, pp 217–223

Knott M., Tzamourani P., Fitting a latent trait model for missing observations to racial prejudice data, Applications of latent trait and latent class models in the social sciences, 1997, pp 244–252

Korobko O. B., Glas C. A., Bosker R. J., Luyten J. W., Comparing the difficulty of examination subjects with item response theory, Journal of Educational Measurement, 2008, vol. 45, p. 139

Lievens F., Sackett P. R., Buyse T., The effects of response instructions on situational judgment test performance and validity in a high-stakes context., Journal of Applied Psychology, 2009, vol. 94, p. 1095

Little R., Rubin D., Statistical Analysis with Missing Data, 2002

Little R. J., Rubin D. B., On jointly estimating parameters and missing data by maximizing the complete-data likelihood, The American Statistician, 1984, vol. 37, p. 218

Little R. J., Schenker N., Missing Data, Handbook of Statistical Modeling for the Social and Behavioral Sciences, 1995, p. 39

Lord F. M., The relation of test score to the trait underlying the test, ETS Research Report Series, 1952, vol. 1952, p. 517

Lord F. M., Applications of item response theory to practical testing problems. Routledge, 1980

Lord F. M., Novick M. R., Statistical theories of mental test scores. IAP, 1968

McKinley R. L., Reckase M. D., Computer applications to ability testing, AEDS Journal, 1980, vol. 13, p. 193

Mislevy R. J., Stocking M. L., A consumer's guide to LOGIST and BILOG, 1989

Mislevy R. J., Wu P.-K., Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing, ETS Research Report Series, 1996, vol. 1996

Moore D. S., Tests of chi-squared type, Goodness-of-fit techniques, 1986, vol. 634

Moustaki I., Knott M., Weighting for item non-response in attitude scales by using latent variable models with covariates, Journal of the Royal Statistical Society: Series A (Statistics in Society), 2000, vol. 163, p. 445

Moustaki I., Marcoulides G., et al., Locating"don't know","no answer"and middle alternatives on an attitude scale: a latent variable approach. Lawrence Erlbaum Associates, 2002

Moustaki I., O'Muircheartaigh C., A one dimension latent trait model to infer attitude from nonresponse for nominal data, Statistica, 2000, vol. 60, p. 259

Neel J. H., A new goodness-of-fit test for Item Response Theory, Journal of Modern Applied Statistical Methods, 2004, vol. 3, p. 26

O'Muircheartaigh C., Moustaki I., Item non-response in attitude scales: a latent variable approach. In Proceedings of the American Statistical Association, Section of Survey Research Methods , 1996, p. 938

O'muircheartaigh C., Moustaki I., Symmetric pattern models: a latent variable approach to item non-response in attitude scales, Journal of the Royal Statistical Society: Series A (Statistics in Society), 1999, vol. 162, p. 177

Pasquali L., Psicometria. Revista da Escola de Enfermagem da USP, 2009, 2009, vol. 43

Pasquali L., Primi R., Fundamentos da teoria da resposta ao item: TRI, Avaliação Psicológica, 2003, vol. 2, p. 99

Pimentel J. L., Item Response Theory modeling with nonignorable missing data. University of Twente, 2005

Rasch G., Probabilistic models for some intelligence and attainment tests.. The Danish Institute for Educational Research, 1960

Ray S., On a theoretical property of the bhattacharyya coefficient as a feature evaluation criterion, Pattern recognition letters, 1989, vol. 9, p. 315

Reckase M., Multidimensional item response theory. vol. 150, Springer, 2009

Reckase M. D., Development and application of a multivariate logistic latent trait model, 1972

Rose N., Item nonresponses in educational and psychological measurement, Jena, Friedrich-Schiller-Universität Jena, Diss., 2013, 2013, Tese de Doutorado

Rose N., Davier M., Xu X., Modeling nonignorable missing data with item response theory (IRT), ETS Research Report Series, 2010, vol. 2010

Rubin D. B., Inference and missing data, Biometrika, 1976, vol. 63, p. 581

Samejima F., Estimation of latent ability using a response pattern of graded scores, ETS Research Report Series, 1968, vol. 1968

Santos V. L. F., Moura F. A., Andrade D. F., Gonçalves K. C., Multidimensional and longitudinal item response models for non-ignorable data, Computational Statistics & Data Analysis, 2016, vol. 103, p. 91

Schafer J. L., Analysis of incomplete multivariate data. CRC press, 1997

Schweppe F. C., On the Bhattacharyya distance and the divergence between Gaussian processes, Information and Control, 1967, vol. 11, p. 373

Smith J., Some Issues in Item Response Theory: Dimensionality Assessment and Models for Guessing.. ERIC, 2009

Spearman C., The proof and measurement of association between two things, The American journal of psychology, 1904, vol. 15, p. 72

Stone C. A., Monte Carlo Based Null Distribution for an Alternative Goodness-of-Fit Test Statistic in IRT Models, Journal of Educational Measurement, 2000, vol. 37, p. 58

Svetina D., Assessing dimensionality in complex data structures: A performance comparison of DETECT and NOHARM procedures. Arizona State University, 2011

Sympson J. B., A model for testing with multidimensional items. In Proceedings of the 1977 computerized adaptive testing conference , No. 00014, 1978

Thomas R. M., High-stakes testing: Coping with collateral damage. Routledge, 2005

Walker C. M., Beretvas S. N., Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid, Journal of Educational Measurement, 2003, vol. 40, p. 255

Weeks J. P., von Davier M., Yamamoto K., Using response time data to inform the coding of omitted responses, 2015

Whelpley C. E., How to Score Situational Judgment Tests: A Theoretical Approach and Empirical Test. Virginia Commonwealth University, 2014

Whitely S. E., Multicomponent latent trait models for ability tests, Psychometrika, 1980, vol. 45, p. 479

Wise S. L., Bhola D. S., Yang S.-T., Taking the Time to Improve the Validity of Low-Stakes Tests: The Effort-Monitoring CBT, Educational Measurement: Issues and Practice, 2006, vol. 25, p. 21

Wise S. L., DeMars C. E., Low examinee effort in low-stakes assessment: Problems and potential solutions, Educational assessment, 2005, vol. 10, p. 1

Wright B. D., Sample-free test calibration and person measurement. In Proceedings of the 1967 invitational conference on testing problems , 1968, p. 85

Yamamoto K., Everson H., Modeling the effects of test length and test time on parameter estimation using the HYBRID model, Applications of latent trait and latent class models in the social sciences, 1997, pp 89–98

Yeh C.-C., The effect of guessing on assessing dimensionality in multiple-choice tests: A Monte Carlo study with application, University of Pittsburgh, 2007, Tese de Doutorado

Apêndice

Apêndice A

Simulação

A.1 *Exemplo em R para a obtenção dos valores críticos da estatística D_B, obtidos com base em 1 mil replicações, assumindo-se uma população normal bivariada para as proficiências*

.

```
# ------------------------------------------------------------
# Preparação para a simulação:
# semente, n = número de examinandos, replic = réplicas,
# thetas.s = proficiências gaussianas bivariadas,
# BD, BD.0, Chi.2 = distâncias de Bhattacharyya e chi^2
# ------------------------------------------------------------
set.seed(81348530)
require(MASS)
n.1            <- 5000
r.1            <- 1000
Sigma.thetas <- matrix(c(1,0.5,0.5,1),2,2)
theta.s      <- mvrnorm(n.1, rep(0, 2), Sigma.thetas)
colnames(theta.s) <- c("theta.1","theta.2")
BD     <- NULL
BD.0   <- NULL
Chi.2  <- NULL
Chi0.2 <- NULL
# ------------------------------------------------------------
# parâmetros do item
```

```
# ------------------------------------------------------------
# par.item = (a_1,b_1,a_2,b_2)
par.item <- c(1.9,0.1,0.9,-0.2)
# ------------------------------------------------------------
# probabilidades
# ------------------------------------------------------------
eta.1 <- par.item[1]*(as.numeric(theta.s[,1])-par.item[2])
eta.2 <- par.item[3]*(as.numeric(theta.s[,2])-par.item[4])
P.U   <- 1/(1+exp(-eta.2))
P.R   <- 1/(1+exp(-eta.1))
P.11  <- P.R
P.01  <- (1-P.R)*P.U
P.00  <- 1-P.11-P.01
# ------------------------------------------------------------
# distribuição esperada
# ------------------------------------------------------------
p.esp <- rbind(c(mean(P.00),mean(P.01),mean(P.11)))
# ------------------------------------------------------------
# simulação das estatísticas Chi.2, D.B exato e aproximado
# ------------------------------------------------------------
r.1 <- 10000000
q.1 <-  p.esp[3]
q.2 <-  p.esp[2]
q.3 <-  p.esp[1]

hat.pi.1  <- q.1
hat.pi2.1 <- mean(P.11^2)
hat.pi.2  <- q.2
hat.pi2.2 <- mean(P.01^2)
hat.pi.12 <- mean(P.11*P.01)
v.1 <-  (hat.pi.1 - hat.pi2.1)/n.1
v.2 <-  (hat.pi.2 - hat.pi2.2)/n.1
cov <- -hat.pi.12/n.1
Sigma <- matrix(c(v.1,cov,cov,v.2),2,2)
```

```
varepsilons <- mvrnorm(n = r.1, rep(0, 2), Sigma)
e.1 <- varepsilons[,1]
e.2 <- varepsilons[,2]
cov(e.1,e.2)
e.3 <- -(e.1+e.2)
q.3 <- 1 - q.1 - q.2
Chi.2 <- n.1*((e.1^2)/q.1 + (e.2^2)/q.2 + (e.3^2)/q.3)
D.B.approx   <- Chi.2 - n.1*((e.1^3)/q.1^2 +
(e.2^3)/q.2^2 + (e.3^3)/q.3^2)/2
D.B.exato    <- -8*n.1*log(q.1*sqrt(1+e.1/q.1)+
q.2*sqrt(1+e.2/q.2)+q.3*sqrt(1+e.3/q.3))
quantile(Chi.2, probs = c(0.999, 0.99, 0.95))
quantile(D.B.approx, probs = c(0.999, 0.99, 0.95))
quantile(D.B.exato, probs = c(0.999, 0.99, 0.95))
# ------------------------------------------------------------------
```