Universidade de Brasília - UnB
Instituto de Ciências Exatas - IE
Departamento de Estatística - EST

Teoria de Resposta ao Item para avaliação de desempenho na disciplina Probabilidade e Estatística da UnB

Tonny Barbosa Pereira

Orientador: Profº. Guilherme Souza Rodrigues

Brasília
1 de julho de 2019

Tonny Barbosa Pereira

Teoria de Resposta ao Item para avaliação de desempenho na disciplina Probabilidade e Estatística da UnB

Relatório final apresentado à disciplina de Trabalho de Conclusão de Curso II de graduação em Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Orientador: Profº. Guilherme Souza Rodrigues

Brasília
1 de julho de 2019

Resumo

Esse estudo tem como objetivo analisar os escores dos alunos que cursam a disciplina de Probabilidade e Estatística na Universidade de Brasília no segundo semestre de 2018, bem como as características dos itens que compõem os testes realizados por esses alunos, de acordo com os métodos de Teoria de Resposta ao Item (TRI), que também é aplicado no ENEM e no vestibular da UnB.
É introduzido um modelo de TRI, bem como suas características. Além disso, apresentamos formas de estimar os parâmetros dos itens e as habilidades dos alunos, por meio de um modelo de TRI. É feita uma análise descritiva dos dados e uma análise das estimativas. Verificamos se o modelo é apropriado e quais itens estão apresentando problemas na análise. Por fim é feita uma comparação entre diferentes métodos de estimação dos parâmetros.

Palavras Chave : Teoria da Resposta ao Item, Modelo dicotômico de 3 parâmetros, Probabilidade e estatística, Avaliação, Método Bayesiano.

Abstract

This study aims to analyze the scores of the students who study Probability and Statistics at the University of Brasília (UnB) in the second half of 2018, as well as the characteristics of the items that make up the tests performed by these students, according to the Item Response Theory (IRT), which is also applied in the ENEM (National High School Exam) and in the UnB's admission test.
A model of IRT is introduced as well as its characteristics. In addition, we present ways of estimating the parameters of the items and the students' abilities, through a IRT model. A descriptive analysis of the data and an analysis of the estimates are made. We verify that the model is appropriate and which items are presenting problems in the analysis. Finally, a comparison is made between different methods of parameter estimation.

Key Words: Item Response Theory, Dichotomous Model of 3 parameters, Probability and statistics, Evaluation, Bayesian method.

Sumário

Lista de ilustrações

Lista de tabelas

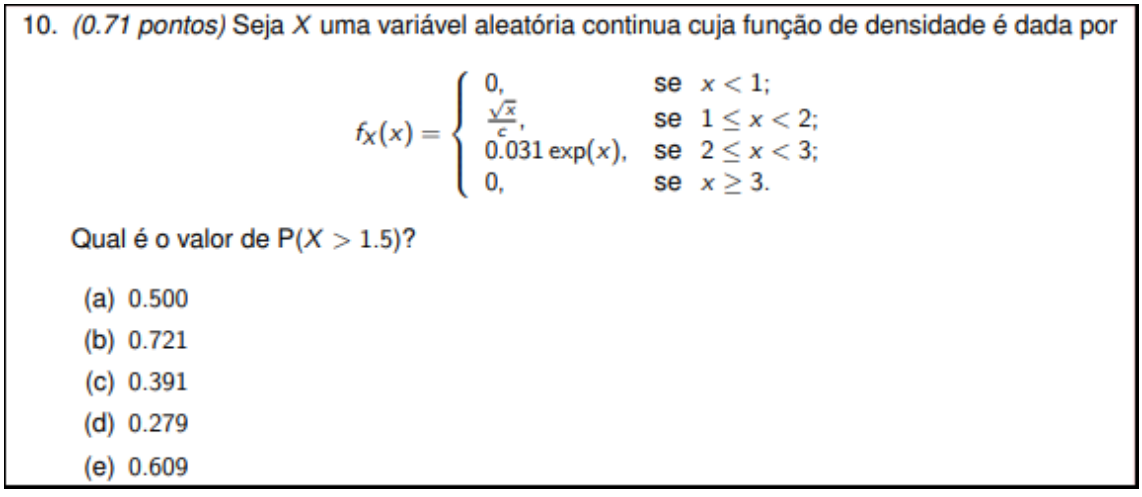
Introdução

O Departamento de Estatística (EST) oferta, em média, 10 turmas de Probabilidade e Estatística por semestre. A disciplina é obrigatória para diversos cursos da área de Ciências Exatas, como Matemática, Computação, Química e algumas Engenharias.

No primeiro semestre de 2018, o Departamento de Estatística unificou a disciplina de Probabilidade e Estatística. A unificação permitiu que os 500 alunos que cursam semestralmente a disciplina, pudessem passar pelo mesmo processo de avaliação.

As provas dos alunos de PE são realizadas 3 vezes por semestre. Cada prova possui 10 questões (itens, na terminologia da Teoria de Resposta ao Item) de múltipla escolha, cada qual com 5 alternativas distintas e apenas uma correta. Um banco de questões foi criado para gerar provas aleatórias para as diversas turmas, sendo que para cada item da prova, o banco conta com 3 diferentes modelos de questões. Atualmente o banco possui 90 questões, mas a previsão para o próximo semestre é que esse número chegue a 300.

Um exemplo de questão gerada pelo banco pode ser visto na figura 1. Nesse exemplo, a probabilidade requerida e a constante normalizadora variam de aluno para aluno, de modo a reduzir o risco de fraudes nas provas. Vale ressaltar que os valores são escolhidos de forma que se mantenha, na medida do possível, o mesmo nível de dificuldade para todos os alunos.



10. *(0.71 pontos)* Seja X uma variável aleatória continua cuja função de densidade é dada por

$$f_X(x) = \begin{cases} 0, & \text{se } x < 1; \\ \frac{\sqrt{x}}{c}, & \text{se } 1 \leq x < 2; \\ 0.031 \exp(x), & \text{se } 2 \leq x < 3; \\ 0, & \text{se } x \geq 3. \end{cases}$$

Qual é o valor de $P(X > 1.5)$?

(a) 0.500
(b) 0.721
(c) 0.391
(d) 0.279
(e) 0.609

Figura 1 – *Exemplo de questão aplicada nas provas de Probabilidade e Estatística*

A partir dos resultados obtidos com a realização das provas, é gerado um banco de dados que contém informações sobre o aluno[1] (curso, semestre, turma, etc) bem como se

[1] Por questões de privacidade, tivemos acesso aos dados anonimizados

sua resposta para cada item está correta ou errada. Um exemplo desse banco de dados pode ser encontrado na seção 2.1.

Com base nesse banco de dados, utilizaremos a Teoria de Resposta ao Item (TRI) para construir modelos estatísticos que verifiquem a qualidade dos itens e calcule o nível de habilidade de cada aluno. Apesar de haver 3 provas, o modelo é ajustado ao final do semestre, combinando as respostas dos alunos nas 3 avaliações. Dessa forma, assumiremos que a habilidade latente do aluno é a mesma ao longo do semestre.

1 Objetivos

1.1 Objetivos Gerais

Utilizar a Teoria de Resposta ao Item para propor melhorias ao sistema de avaliação dos alunos que cursam a disciplina de Probabilidade e Estatística da UnB. Este trabalho não tem pretensão de substituir a avaliação da disciplina pelo processo de TRI, entretanto os professores poderão optar por utilizar as habilidades estimadas da maneira que lhes for conveniente. É importante ressaltar que, embora estejamos avaliando a disciplina de PE, os procedimentos de análise desenvolvidos podem ser aplicados em outros cursos, desde que tenham um sistema de avaliação semelhante.

1.2 Objetivos Específicos

- Analisar numericamente as características das questões utilizadas nas avaliações, incluindo seu nível de dificuldade e seu poder discriminante.

- Identificar questões que não estejam bem calibradas (isto é, são demasiadamente fáceis ou difíceis) ou que não possui relação direta com a habilidade dos alunos (tem baixo poder de discriminação).

- Obter o nível de habilidade latente dos alunos.

- Comparar os escores obtidos pelo modelo com os escores reais e observar a correlação entre essas variáveis.

2 Metodologia

2.1 Descrição do banco de dados

O banco de dados que será utilizado nesse trabalho foi obtido no decorrer do semestre passado (2º/2018), à medida em que os alunos iam realizando as provas. O banco contém informações acadêmicas gerais sobre o aluno, por exemplo: curso, semestre letivo, turma e período da disciplina. O banco também nos informa se o aluno acertou ou não cada item. Itens não marcados são considerados como respostas erradas.

O banco de dados que iremos obter ao final do semestre, é exemplificado pela figura 2:

Questão	Aluno	Curso	Turma	Semestre	Período	Resposta
1	123	Matemática	B	4	Diurno	certo
28	067	Matemática	B	2	Diurno	errado
60	453	Matemática	C	2	Noturno	certo
1	568	Engenharia Civil	D	3	Diurno	errado
45	498	Engenharia Civil	D	5	Diurno	certo
88	075	Engenharia Civil	D	3	Diurno	errado
5	211	Engenharia Civil	D	2	Diurno	certo
24	531	Engenharia Civil	F	3	Diurno	certo
52	210	Engenharia Civil	E	2	Diurno	certo
11	314	Computação	C	3	Noturno	errado
24	178	Computação	C	4	Noturno	certo
70	298	Computação	C	3	Noturno	errado
14	024	Química	A	4	Diurno	certo
33	013	Química	A	2	Diurno	certo
61	523	Química	B	2	Diurno	certo

Figura 2 – *Exemplo do banco de dados*

2.2 Teoria de Resposta ao Item

Para avaliar o aprendizado de crianças, costuma-se aplicar uma prova e, de acordo com o número de acertos, tem-se uma ideia do quanto a criança absorveu nas aulas. Essa é uma forma de quantificar a habilidade latente do aluno, ainda que a mesma não seja diretamente observável. Em outras palavras, a habilidade de uma criança em matemática, por exemplo, é uma característica complexa e imensurável. Entretanto, se a nota é muito baixa, temos uma indicação de que aquela criança necessita de uma atenção especial de

pais e educadores. Essa estratégia nos permite não somente identificar os alunos com aprendizagem deficitária, mas também os alunos medianos e os de destaques.

A Teoria Clássica das Medidas (TCM) surgiu para propor um método capaz de medir o nível de conhecimento de um indivíduo. O princípio básico da teoria é de que, quanto mais acertos em uma determinada avaliação, maior o domínio sobre o assunto. Na TCM, o instrumento de medida criado para avaliação dos indivíduos depende do particular conjunto de questões que compõem a prova. Alunos submetidos a uma mesma prova ou a provas semelhantes possuem seus resultados comparáveis, o que possibilita também comparações entre grupos. Entretanto, nessa teoria, torna-se inviável a comparação entre indivíduos que não foram submetidos às mesmas provas.

A fim de sanar esse problema, foi desenvolvido a Teoria de Resposta ao Item (TRI), que possibilita a comparação de trações latentes de indivíduos de populações iguais submetidos a testes diferentes e de indivíduos de população diferentes quando submetidos a teste iguais. A TRI começou a ser utilizada em larga escala, chegando no Brasil inicialmente em 1995, na análise dos resultados do Sistema Nacional de Ensino médio (SAEB), que permitiu a comparação do desempenho dos alunos de diferentes séries.

Nos dias de hoje, o maior exemplo que temos de aplicação da TRI é no Exame Nacional do Ensino Médio (ENEM). As questões são divididas previamente em fáceis, médias e difíceis, e são misturadas ao longo da prova de forma que o estudante não saiba qual questão pertence a qual grupo. Através da aplicação do TRI, se constatado que ele errou muitas perguntas fáceis e acertou muitas perguntas difíceis, é deduzido que ele "chutou", assim a média do estudante que chutou cai.

Para o ENEM, o objetivo é evitar que o candidato consiga se valer do fator sorte na hora de responder as questões. Dessa forma, o processo seleciona os alunos de uma forma mais justa, o que traz um impacto positivo para as faculdades e reforça a cultura de que o importante é uma boa preparação do aluno.

2.2.1 Modelos de teoria de resposta ao item

A TRI é um conjunto de modelos estatísticos que estimam a probabilidade de um indivíduo acertar um item em função das características dos itens (nível de dificuldade e poder de discriminação) e dos indivíduos (habilidade latente, curso, entre outros).

Os modelos de TRI dependem dos seguintes atributos: natureza do item, número de populações e quantidade de traços latentes. A natureza do item pode ser dividia em dicotômicos (2 categorias de resposta) e politômicos (mais de 2 categorias). As quantidades de traços latentes determinam se o modelo é unidimensional, quando há apenas um traço a ser calculado, ou multidimensional, quando há mais de um traço.

Nesse estudo iremos trabalhar com itens dicotômicos, ou seja, o item possui apenas uma alternativa correta, que irá determinar que o aluno acertou o não. Em outras palavras, caso a resposta certa para um dado item seja A, o modelo trata igualmente os casos em que o aluno marca as opções B, C, D ou E, já que nesses casos o aluno errou o item. Iremos considerar os discentes matriculados na disciplina de Probabilidade e Estatística como uma única população e será estimada uma única habilidade latente dos indivíduos que, por sua vez, pode possuir covariáveis associadas a este eles.

Há casos em que é possível estimar várias habilidades latentes, por exemplo poderíamos supor que os alunos não têm simplesmente uma habilidade em "estatística", mas sim uma habilidade em inferência e outra distinta em probabilidade. Nesse caso, as habilidades de um mesmo aluno serão correlacionadas.

2.2.2 Modelos para itens dicotômicos

Os modelos de resposta ao item mais utilizados para itens dicotômicos são os modelos logísticos[1], sendo que há basicamente três tipos, e se diferenciam pelo número de parâmetros utilizados para descrever o item. O mais utilizado é o modelo logístico unidimensional de 3 parâmetros (ML3), por ser mais completo. Os parâmetros a serem considerados no ML3 são: dificuldade do item, poder de discriminação do item e probabilidade do item ser acertado ao acaso. Os outros dois modelos podem ser obtidos facilmente a partir deste.

Nesse estudo, o modelo inicial que iremos utilizar é definida pela normal acumulada e não pela logística. A estrutura desse modelo é dada por:

$$
\begin{aligned}
&Y_{ij} \sim Ber(p_{ij}) \\
&p_{ij} = c_j + (1 - c_j)\Phi(a_j\theta_i - b_j) \\
&b_j \sim N(0, 1) \\
&c_j \sim Beta(5, 17) \\
&\theta_i \sim N(0, 1)
\end{aligned}
$$

Sendo i alunos e j questões, onde:

- Y_{ij} assume valor 1 se o aluno i acertar a questão j e 0 caso contrário.
- p_{ij} indica a probabilidade do aluno i acertar a questão j.
- Φ é a função de distribuição acumulada da normal padrão.

[1] O termo *logistico* se refere à forma com que as habilidades se relacionam com a probabilidade de acertar.

- c_j é a probabilidade de um aluno com habilidade extremamente baixa ($\theta_i \to -\infty$) acertar a questão.
- a_j representa o poder discriminante da questão j.
- b_j representa a dificuldade do item j.
- θ_i representa a habilidade latente do aluno i.

Mesmo após extensa pesquisa, não foi possível encontrar a distribuição do parâmetro a_j, entretanto suspeitamos que ele segue uma distribuição log-normal.

Nesse modelo, supomos que respostas de indivíduos diferentes são independentes e que os itens respondidos por cada indivíduo são condicionalmente independentes (isso é, dada a habilidade θ_i).

2.2.3 Função Característica do Item

O modelo assume que quanto maior a habilidade do indivíduo, maior a probabilidade de ele acertar um item, embora essa relação não seja linear.

A Curva Característica do item (CCI) é dada pela relação da probabilidade de o indivíduo acertar o item com os parâmetros latentes do item. A inclinação e o deslocamento representados pelos parâmetros a_j e b_j, respectivamente, faz com que a CCI assuma um formato de "S", na mesma escala da habilidade.

A escala da habilidade é arbitraria, onde o importante não é a magnitude e sim as relações de ordem existente entre os pontos. O parâmetro b é medido na mesma escala do parâmetro θ e o parâmetro c não depende de escala. O parâmetro a é obtido pela derivada da tangente no ponto de inflexão.

Um exemplo de curva característica do item está representado na figura 3 :

A probabilidade de acerto aumenta de maneira não-linear em função da proficiência do aluno. A marcação das letras indica a interpretação geométrica dos parâmetros a, b e c, que nesse caso assumem os valores 270, 550 e $0,1$, respectivamente. Com uma breve análise dessa CCI, podemos verificar que um indivíduo com habilidade latente igual a 550 possui 55% de probabilidade de acertar o item, já um indivíduo com habilidade igual a 300 possui 10% de probabilidade de acertar o item, o que corresponde a probabilidade de acertar o item ao acaso.

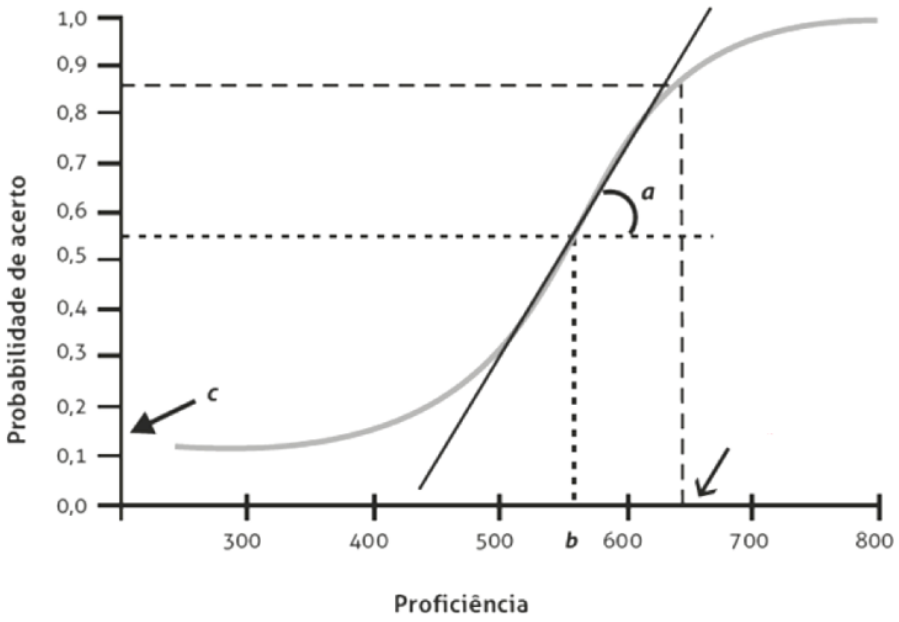


Figura 3 – *Exemplo de Curva Característica do Item (CCI)*

Fonte: Internet

2.2.4 Função de Informação do Item

A função de informação do item é bastante utilizada com a CCI, e permite analisar o quanto o item contém de informação em relação a habilidade latente de um indivíduo. A função é dada por:

$$I_j(\theta) = a_j^2 \frac{Q_j(\theta)}{P_j(\theta)} \left[\frac{P_j(\theta) - c}{1 - c} \right]^2$$

Onde,

- $I_j(\theta)$ é a "informação" fornecida pela questão j no nível de habilidade θ.
- $P_j(\theta) = p_{ij}$.
- $Q_j(\theta) = 1 - P_j(\theta)$.

Essa função nos mostra a importância dos parâmetros sobre o valor da informação do item. A informação é maior:

- Quando b_j se aproxima de θ.
- Quanto maior for a_j.
- Quanto mais c aproximar de 0.

A função de informação do teste é obtida pelo somatório das informações dos itens que compõem o teste. Entretanto, como os alunos realizam provas distintas, a função de informação depende da turma.

$$I(\theta) = \sum_{j=1}^{I} I_j(\theta).$$

A curva de informação é normalmente usada junto com a CCI, e pode ser representada pela figura 4.

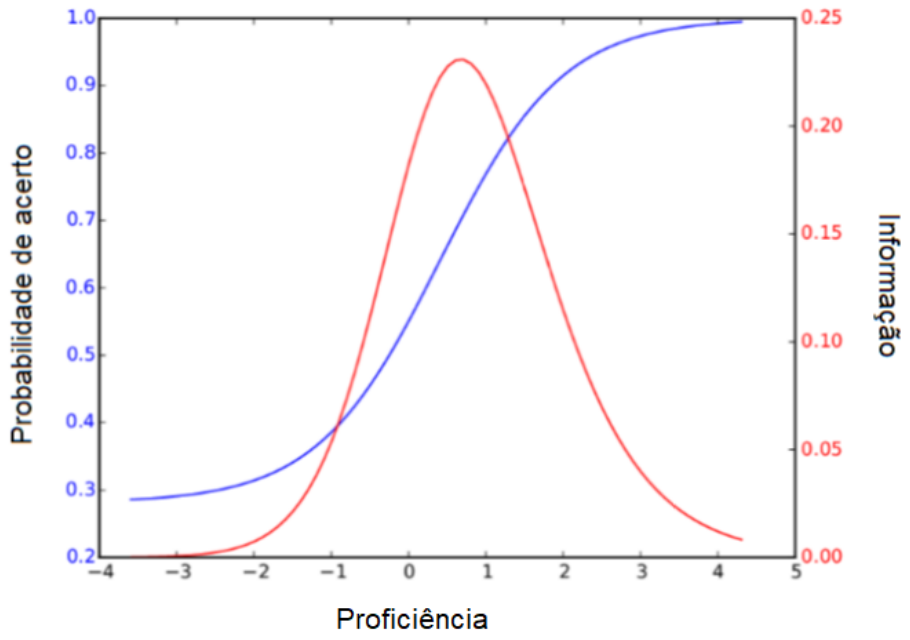


Figura 4 – *Exemplo de Curva de Informação do Item*

Fonte: Internet

2.3 Estimação dos Parâmetros

Um dos pontos mais importantes da Teoria de Resposta ao Item, é a estimação das habilidades dos indivíduos e dos parâmetros dos itens.

Como dito anteriormente, serão criados modelos para determinar a probabilidade de um indivíduo acertar um item (ver seção 2.2.2), e esses modelos dependem apenas das habilidades latentes e dos parâmetros que caracterizam o item. Em geral, esses parâmetros são desconhecidos e conhecemos somente a resposta do aluno aos itens.

Do ponto de vista teórico, a estimação dos parâmetros pode ser feita em três situações diferentes, quando já se conhece os parâmetros dos itens e precisa estimar as habilidades,

quando já se conhece as habilidades dos indivíduos e deseja estimar os parâmetros dos itens ou então quando se deseja estimar as habilidades latentes e parâmetros dos itens simultaneamente, aqui estamos interessados no último caso. A estimação geralmente é feita pelo Método da Máxima Verossimilhança ou por Métodos Bayesianos.

2.3.1 Estimação por Máxima Verossimilhança

A estimação por máxima verossimilhança geralmente é feita através da aplicação de algum método numérico iterativo, como Newton-Raphson. Há duas abordagens comuns para a estimação desses parâmetros, a estimação conjunta e a estimação em duas etapas, onde se estima primeiro o parâmetro dos itens e depois as habilidades.

A estimação conjunta tem uma exigência computacional muito grande quando o número de parâmetros a serem estimados simultaneamente também é grande. Neyman & Scott (1948) notaram que quando o número de indivíduos cresce, o número de parâmetros θ_i também cresce, e o EMV dos parâmetros dos itens pode ser assintoticamente viesado. Lord (1968) diz que os EMV dos parâmetros dos itens e habilidades podem ser não-viesados quando crescem o número de itens e de indivíduos.

Bock & Lieberman (1970) a fim de resolver o problema da possível inconsistência dos estimadores, desenvolveram o método de estimação em duas etapas. Na primeira, é estimado o parâmetro dos itens através da Máxima Verossimilhança Marginal, considerando uma distribuição para as habilidades dos indivíduos e integrando a função de verossimilhança em relação a θ. Em seguida, as habilidades são estimadas individualmente pela máxima verossimilhança, a partir dos parâmetros dos itens estimados anteriormente.

Esse método, entretanto, leva a resultados questionáveis, uma vez que, na segunda etapa, não é devidamente considerada a incerteza em relação aos parâmetros das questões, o que leva a uma subestimação da variabilidade do estimador das habilidades.

O processo de integração pelo EMV fica muito complexo quando o número de itens e indivíduos aumenta, pois o número de integrais a serem resolvidas é muito grande. Uma alternativa para o método de estimação é a aplicação de técnicas Bayesianas.

2.3.2 Estimação Bayesiana

O método de estimação bayesiana consiste em estabelecer distribuições a priori para os parâmetros de interesse e utilizar o Teorema de Bayes para calcular a distribuição a posteriori, a partir da qual se obtêm as estimativas desses parâmetros.

A estimação Bayesiana apresenta solução para problemas gerados pela estimação de máxima verossimilhança, que se dá principalmente quando todos os itens são respondidos

corretamente ou incorretamente por um indivíduo ou quando todos os indivíduos acertam ou erram determinado item. O problema também pode ser causado por alguma estimativa dos parâmetros cair fora do intervalo esperado.

Considerando as limitações dos parâmetros, em que a_j deve ser positivo, b_j assume qualquer valor real e c_j está no intervalo [0,1], as distribuições a priori devem ser determinadas de acordo com essas limitações, exigindo um tratamento diferente para cada um dos parâmetros.

A técnica para realizar o procedimento de estimação, surge a partir do Teorema de Bayes, em que a distribuição à posteriori de Θ é dada por:

$$P(\Theta|Y) \propto P(Y|\Theta)P(\Theta)$$

Onde:

- $P(\Theta|Y)$ é a distribuição à posteriori de Θ.
- $P(Y|\Theta)$ é a função de verossimilhança assumido para as observações.
- $P(\Theta)$ é a distribuição à priori de Θ.
- $\Theta = (a, b, c, \theta)$ é o conjunto de parâmetros, com

  $a = (a_1, ..., a_n)$, n itens.

  $b = (b_1, ..., b_n)$, n itens.

  $c = (c_1, ..., c_n)$, n itens.

  $\theta = (\theta_1, ..., \theta_m)$, m alunos.

  $Y = (Y_{11}, ..., Y_{ij}, ..., Y_{mn})$.

Como os parâmetros são independentes, por suposição, temos a forma estendida:

$$P(\Theta|Y) = P(a, b, c, \theta|Y) \propto P(Y|a, b, \theta)P(a)P(b)P(c)P(\theta)$$

A função de verossimilhança $P(Y|\Theta)$ é dada pelo produto das distribuições Y_{ij}:

$$P(Y|a, b, c, \theta) = \left[\prod_{j=1}^{n}\prod_{i=1}^{m} p_{ij}^{Y_{ij}}(1 - p_{ij})^{1-Y_{ij}}\right]$$

Os parâmetros a serem estimados seguem distribuições normais, com médias e variâncias definidas anteriormente (2.2.2), e são dadas pelas seguintes expressões:

$$P(b) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(b_j\right)^2\right]$$

$$P(\theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\theta_i\right)^2\right]$$

$$P(c) = \prod_{j=1}^{n} \frac{\Gamma(22)}{\Gamma(5)\Gamma(17)} c_j^4 (1 - c_j)^{16}$$

2.3.3 Método MCMC

A distribuição à posteriori é definida a partir de uma integral (a constante normalizadora) que não pode ser calculada diretamente, nem mesmo usando as técnicas de integração numéricas tradicionais. Uma alternativa para podermos calcular a distribuição à posteriori é utilizando o método de simulação Monte Carlo via Cadeia de Markov (MCMC).

O método MCMC é bastante utilizado em problemas de inferência Bayesiana. Ele permite obter amostras de uma distribuição alvo (nesse caso a posteriori), desde que se saiba o núcleo da expressão da função de densidade ou probabilidade dessa distribuição.

Deve-se escolher um núcleo de transição de uma Cadeia de Markov que satisfaça certas condições de regularidade. O método então constrói uma cadeia de Markov a partir dela, de modo que a cadeia terá a distribuição a posteriori como distribuição limite. A partir daí, simula-se a realização da cadeia e ela irá convergir para a distribuição à posteriori.

Os resultados obtidos são assintóticos, de forma que, quando o tamanho de amostras simuladas tende a infinito, a cadeia converge para a distribuição de interesse. Entretanto, é possivel que o número de amostras necessarias para uma boa aproximação seja grande. Por isso, é preciso investigar se a cadeia convergiu e descartar as observações iniciais (processo conhecido como *burning*). Como as amostras simuladas por esse processo são dependentes umas das outras, costuma-se aproveitar apenas uma amostra a cada tantas amostras, de forma a minimizar a dependência entre elas (thining).

2.3.4 Extensões do modelo

Os modelos de 3 parâmetros podem ser muito diferentes entre si a depender das distribuições dos parâmetros. Nesse trabalhos iremos utilizar um modelo mais simples, em que estimamos apenas uma habilidade latente e as distribuições dos parâmetros θ_i, b_j e a_j seguem uma distribuição normal com médias centradas no zero. Para o parâmetro c_j foi definido uma priori com distribuição Beta(5,17) já que os itens apresentam 5 opções de

resposta. Como c_j é limitado por 0 e 1, uma distribuição a priori Beta(α, β) foi proposta por Swaminathan e Gifford (1986). Estes parâmetros são definidos como α = mp + 1 e β = m (p - 1) +1, onde p = 1/n com n = número de alternativas para cada item (Harwell & Baker, 1991).

Para modelos mais complexos, poderíamos utilizar variáveis regressoras para definirmos a distribuição a priori, com base nas características dos alunos (semestre, curso, turma, ano de ingresso na universidade, etc). Também seria possível estimar mais de uma habilidade, por exemplo uma habilidade diferente para cada prova realizada, já que são feitas 3 provas ao longo do semestre (sem considerar a prova substitutiva).

Para esses modelos é mais recomendável utilizar o RStan (Stan integrado ao R). O Stan é um software voltado para inferência Bayesiana, que consiste em realizar métodos computacionais de simulação, sendo muito útil para resolver problemas na TRI. Todavia para os modelos mais simples, é mais prático utilizar funções do R próprias para TRI, por exemplo as funções *bairt* e *sirt*, que consistem em realizar simulações de amostras a partir dos dados e de uma distribuição fornecida a priori. Através do resultado das amostras simuladas, é obtida a estimativa dos parâmetros de interesse.

3 Resultados

Neste capítulo iremos apresentar os resultados obtidos a partir das análises do banco de dados. O banco conta com 487 respondentes, em que cada um respondeu a 30 itens durante o semestre (há casos em que o aluno deixa de fazer uma das provas, dessa forma respondendo menos itens). Os itens que não foram respondidos são considerados como errados.

O banco também conta com 82 itens que foram sorteados aleatoriamente entre as turmas, de forma a compor as provas dos alunos. Os resultados foram obtidos através do software R bem como a elaboração dos gráficos.

3.1 Análise Descritiva

A estatística descritiva é a etapa inicial da análise e extremamente importante para descrever e mostrar um panorama mais completo dos dados. Será abordado alguns gráficos em que poderemos identificar algumas medidas referentes aos alunos que cursaram Probabilidade e Estatística no 2º semestre de 2018, dessa forma gerando mais informação sobre o banco de dados.

A disciplina de Probabilidade e Estatística está presente em cursos de diferentes áreas de conhecimento, seja exatas, humanas e até mesmo ciências biológicas. Aproximadamente 16 cursos tem a disciplina de PE em sua grade curricular. Na figura 5, podemos ver a relação da nota média dos alunos em cada curso.

É esperado que alunos de Engenharias venham a ter as maiores notas, isso se comprova através do gráfico, em que os cursos de maiores notas foram Engenharia Civil, Engenharia Elétrica e Engenharia de Controle e Automação, com notas médias equivalentes a $6,35$, $5,93$ e $5,92$ respectivamente. Já as menores notas médias pertencem aos cursos de Ciências Biológicas$(3,96)$ e Engenharia Ambiental$(4,06)$.

De forma parecida, também podemos ver a nota dos alunos por turma, dessa vez por meio de boxplox (Figura 6). A vantagem desse gráfico é que permite observar um pouco da dispersão dos dados. Notamos que a turma CA obteve maiores notas que as demais turmas, isso se confirma pois a turma é composta principalmente por alunos de Engenharia Civil e Engenharia de Controle e Automação. É interessante ressaltar que nas turmas AA e DB não houve um aluno sequer que obteve 10 na nota final.

Das 3 provas que os alunos fazem durante o semestre, é observado uma pequena queda na nota dos alunos da primeira para a segunda prova e um grande aumento das notas

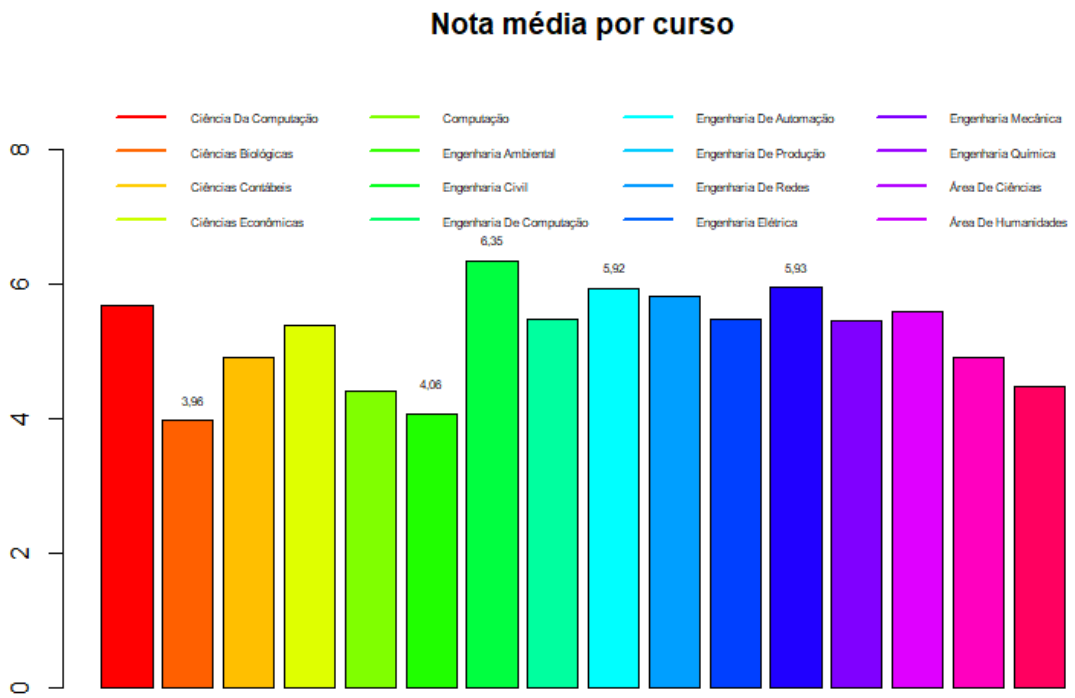


Figura 5 – *Nota média dos alunos por curso*

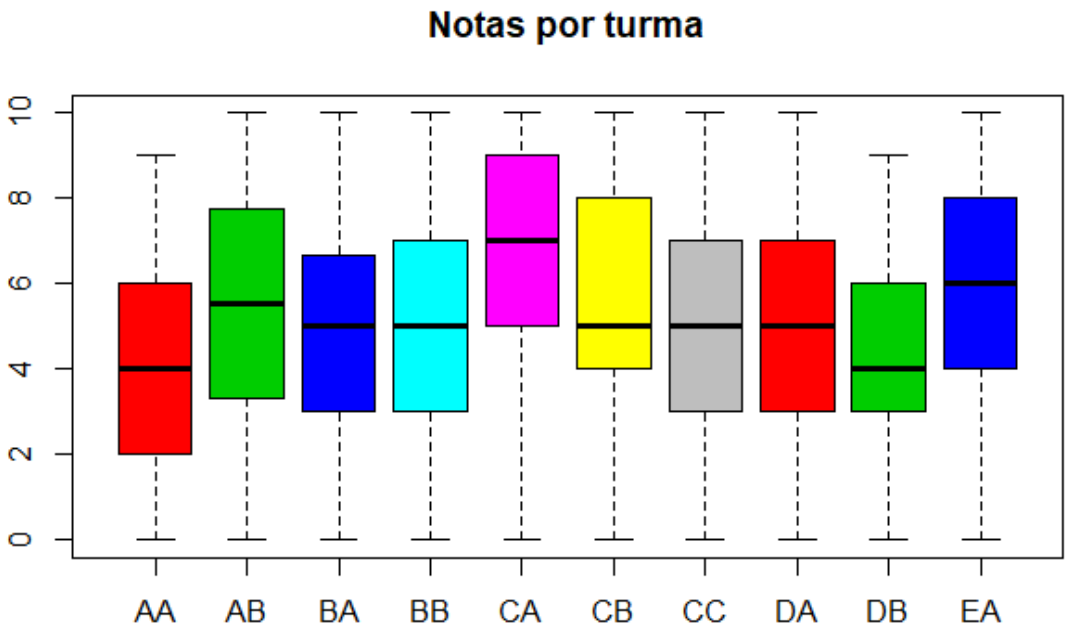
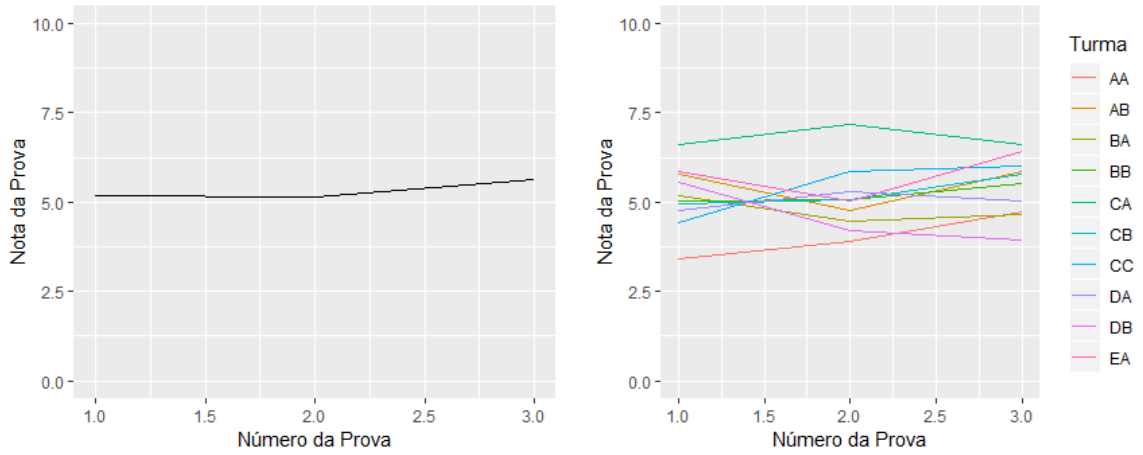


Figura 6 – *Nota dos alunos por turma*

na terceira prova. Já observando por turma, algumas delas não apresentam o mesmo comportamento. Os alunos da turma CA foram melhores na prova 2, já os alunos da turma DB obtiveram os piores resultados na prova 3.



(a) Nota média (b) Nota média por turma

Figura 7 – *Nota média dos alunos em cada prova*

Cerca de 52,77% dos alunos obtiveram aprovação nessa disciplina (desconsiderando-se a prova substitutiva), o que parece ser um valor relativamente baixo. A porcentagem de alunos que obtiveram cada menção ao final do semestre pode ser vista no gráfico 8. Apenas 3 alunos reprovaram com SR e 20 alunos foram aprovados com menção máxima. O gráfico apresenta uma distribuição parecida com uma normal, com grande concentração de alunos que obtiveram MI e MM.

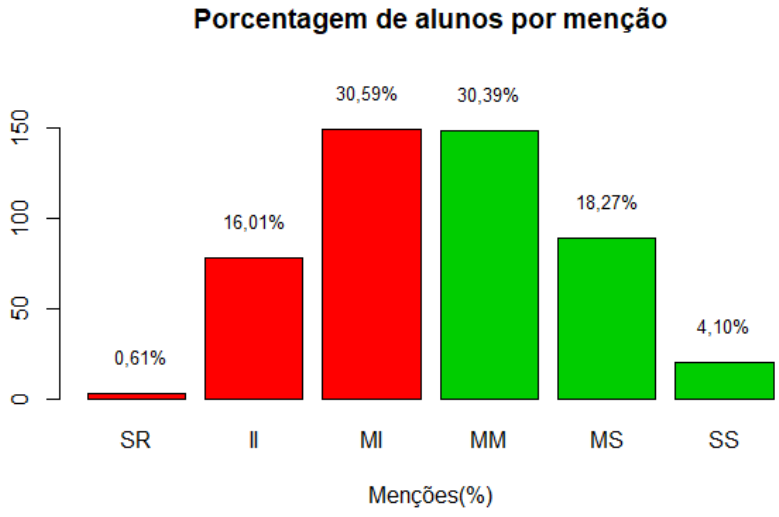


Figura 8 – *Menção dos alunos*

Quanto ao ano de matrícula, os alunos que ingressaram na Universidade no ano de 2018 obtiveram melhores resultados. O gráfico apresenta um comportamento bem curioso, visto que as notas decrescem dos alunos com matrícula em 2013 a 2015, e tornam a subir até 2018. Houve apenas 1 aluno com matrícula em 2010, que não foi aprovado na disciplina

e sequer fez a terceira prova.

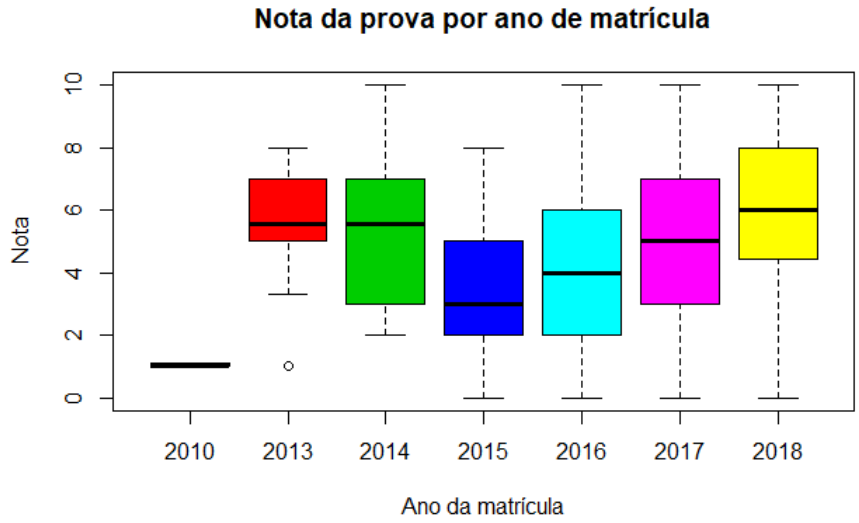


Figura 9 – *Notas dos alunos por ano de ingresso na Universidade*

Por fim, foram observados os itens das provas que os alunos têm mais dificuldade e mais facilidade. Na prova 1 os alunos tiveram mais acerto no item 3 e mais erros no item 4. Na segunda prova, os alunos acertaram mais o item 1 e erraram o 9 e na terceira prova acertaram mais o 2 e erraram o 9. Dos assuntos abordados na disciplina, os alunos apresentaram mais facilidade em Regra da Probabilidade Total, referente ao item 3 da prova 1, e apresentaram mais dificuldade em obter os p-valores, referente ao item 9 da prova 3.

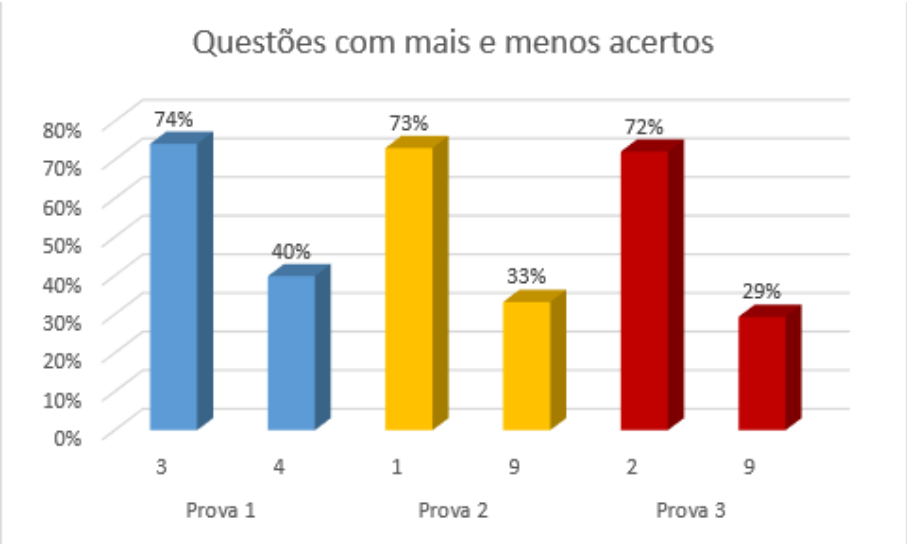


Figura 10 – *Itens com mais e menos acertos em cada prova*

Os gráficos apresentados até aqui contribuem para uma identificação mais detalhada

do banco de dados. As características dos itens, de dificuldade, discernimento e acerto ao acaso, bem como as proficiências dos alunos, serão abordados daqui pra frente.

3.2 Análise via TRI

A estimação dos parâmetros, como vimos anteriormente, foi feita pelo método Bayesiano, utilizando o pacote do R *bairt* e a função *mcmc.3pnob*. O banco de dados foi ajustado de forma que cada aluno represente as linhas, os itens representem as colunas e os valores dessa matriz assumem valor 1 para caso o aluno acertou o item e 0 caso o aluno errou o item. Os itens não respondidos pelos alunos tiveram valores NA.

Para executarmos a função, definimos o número de iterações em 500.000, o que é um valor consideravelmente alto para checarmos a convergência da cadeia de Markov. Valores maiores que esse impediam a execução da função pois exigia um trabalho computacional muito grande. Também definimos o valor de thining como 10, para eliminarmos a correlação entre os valores subsequentes da cadeia bem como o valor de burning igual a zero para que seja analisado todo o comportamento da Cadeia de Markov.

Também definimos a distribuição a priori do parâmetro c_j como sendo uma Beta de parâmetros $\alpha = 5$ e $\beta = 17$.

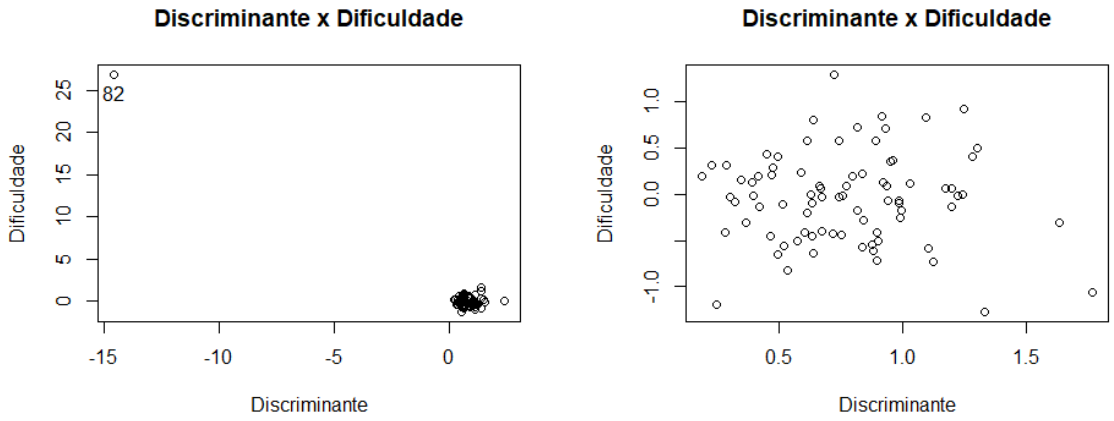
3.2.1 Parâmetros dos itens

Obtidas as estimativas dos parâmetros é possível identificar itens que apresentem irregularidades. Valores negativos do discriminante indicam que a probabilidade de acerto do item diminui a medida que aumenta a habilidade do aluno, o que significa que o item apresenta algum tipo de "problema". Valores muito grandes da dificuldade do item nos mostram que o item está muito além da habilidade dos alunos e que deveria ser substituído ou eliminado.

É possível perceber, com certa facilidade, itens que destoam dos demais por meio de um gráfico de dispersão entre os parâmetros de discriminação e dificuldade do item. No modelo inicial, o gráfico de dispersão nos mostra que há um item que apresenta valores completamente diferentes dos demais. O item 82 apresentou ambos problemas citados acima: discriminante negativo e dificuldade muito alta. É possível que o item, de fato, apresente problemas e portanto deveria ser corrigido do banco de questões.

Eliminando o item 82 e ajustando um novo modelo, todos os demais itens apresentaram um comportamento regular, como podemos ver no gráfico abaixo.

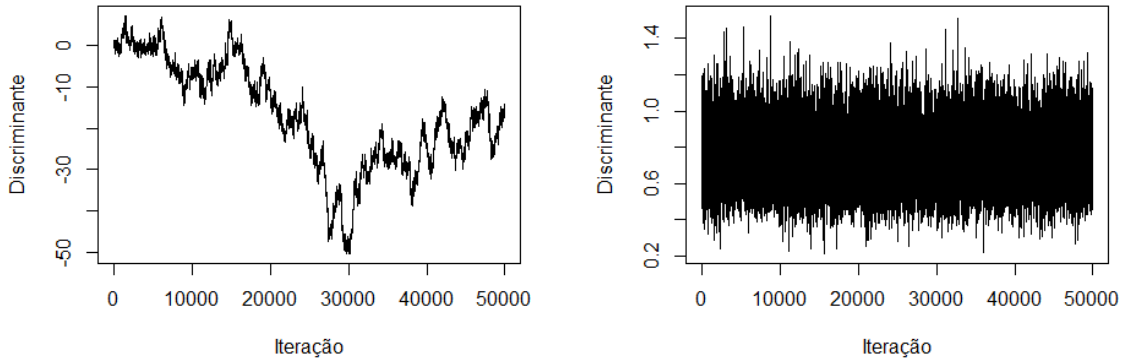
Outro indicativo que temos da irregularidade do item 82 é através da Cadeia de Markov. O comportamento da cadeia para esse item é totalmente irregular e não foi possivel observar



(a) Modelo com o item 82 (b) Modelo sem o item 82

Figura 11 – *Gráfico de dispersão dos parâmetros de Discriminação e Dificuldade do item*

a convergência da cadeia mesmo após meio milhão de iterações. Para os demais itens do modelo, as Cadeias de Markov apresentaram um comportamento dentro do esperado. Como comparativo, utilizamos a cadeia do item 2, em que a cadeia converge.



(a) Cadeia do item 82 (b) Cadeia do item 2

Figura 12 – *Comportamento da Cadeia de Markov ao longo das iterações*

Dadas essas circunstancias, foi preferível trabalhar com o modelo sem o item 82.

A curva característica é uma importante ferramenta gráfica para observarmos a probabilidade do indivíduo acertar o item de acordo com sua proficiência. Como seria inviável analisar as CCI das 81 questões, pegamos 4 itens que demonstrassem bastante a diferença entre elas: os itens de maior e menor dificuldade (itens 47 e 59, respectivamente) e os itens de maior e menor poder discriminante (itens 65 e 55).

A curva do item 47 é deslocada mais para a direita, indicando um maior nível de

dificuldade, ou seja, a probabilidade de acerto do item se dá a níveis de proficiência mais elevados. Enquanto isso as curvas verde e vermelha são deslocadas mais para a esquerda, indicando que esses itens são mais fáceis. Quanto ao discriminante, a curva do item 65 foi a mais ingrime, o que indica o maior poder de discriminação. É a curva que mais se aproxima do formato de "S"esperado por esse modelo. A curva azul, entretanto, aparenta uma relação quase linear, o que nos mostra um poder de discriminação muito baixo. Em outros termos, o item 55 não discrimina bem alunos com proficiência alta dos alunos com proficiência baixa.

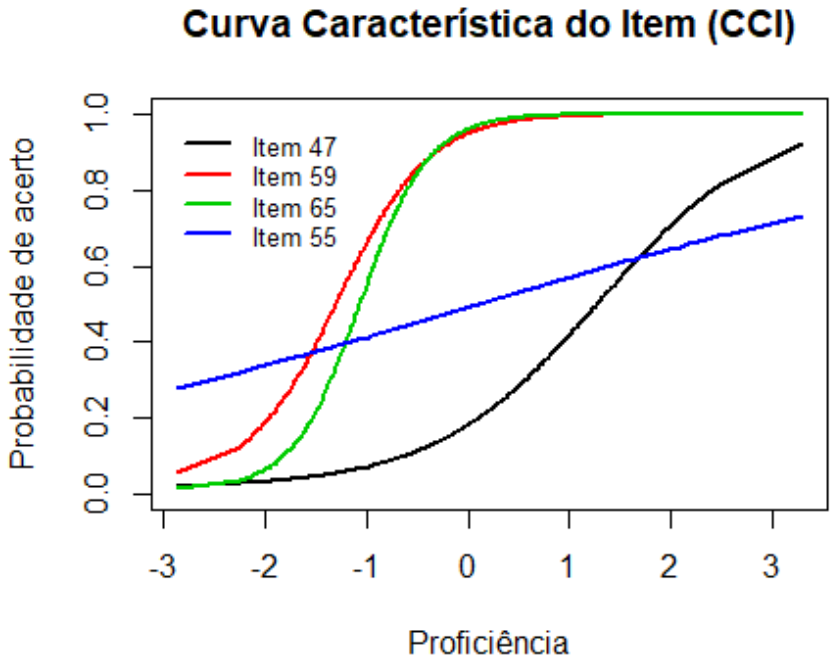


Figura 13 – *Curvas Característica dos Itens*

Outro gráfico importante a ser analisado na TRI é a função de informação. Assim como na CCI, a curva preta também é deslocada mais para a direita, indicando que o item 47 contém mais informação para habilidades mais elevadas. A curva verde é mais acentuada que as demais, isso ocorre devido ao seu discriminante ser elevado, e atinge seu máximo em níveis de proficiência próximos a -1. A curva do item 55 já não apresenta nem formato de curva, isso indica que o item fornece quase nenhuma informação em todos os níveis de proficiência.

A curva de informação do teste é construída a partir da sobreposição das curvas de informação de todos os itens que compõem o teste, e pode ser apresentada pela figura 15.

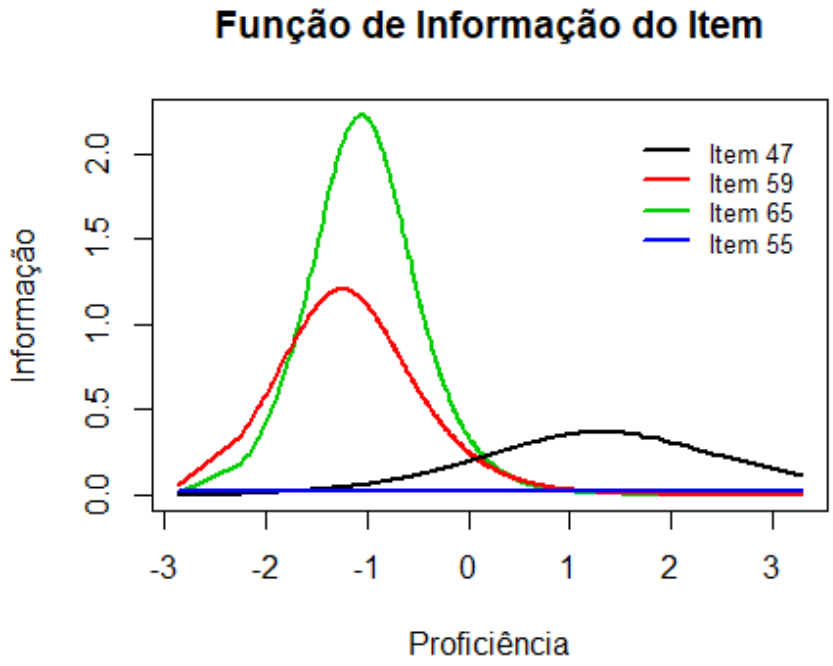


Figura 14 – *Curvas de Informação dos Itens*

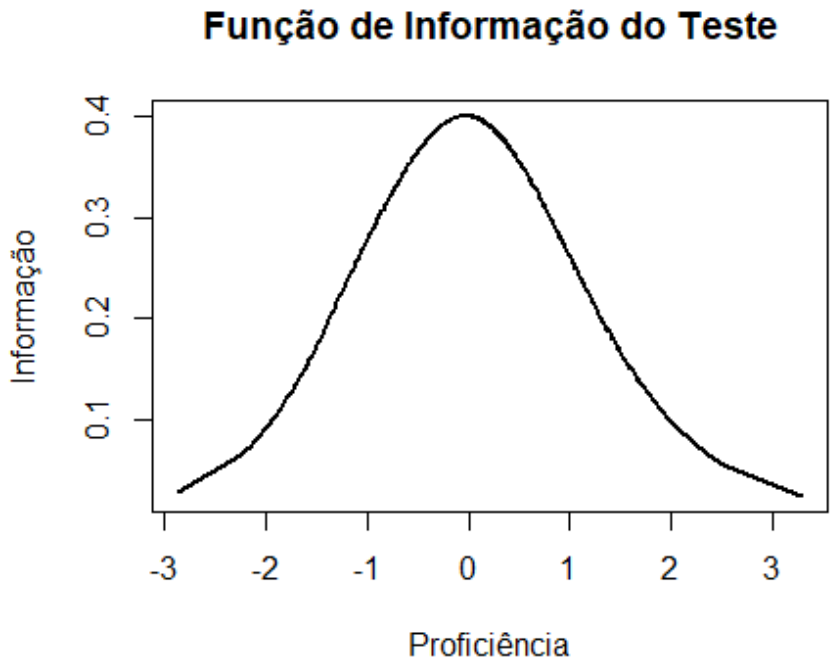


Figura 15 – *Curva de Informação do Teste*

Voltando ao item 82 que havia sido removido do modelo, decidimos apresentá-lo aqui para tentarmos identificar algum possível erro, tanto de enunciado quanto no próprio

resultado. O item pede que se calcule o p-valor em um teste de hipótese, é um assunto que se aborda no final da disciplina de Probabilidade e Estatística e de fato pode ser que os alunos apresentem dificuldade por ser um pouco mais complicado. A questão 82 é equivalente ao 117 do banco de questões.

O enunciado aparentemente não tem nenhum erro. Também não houve erro na resposta. É difícil apontar com precisão qual o motivo das estimativas dos parâmetros desse item apresentarem valores tão irregulares. É possível que a maioria dos alunos tivessem chutado essa questão, por ser um tema mais complicado, e acabaram atraídos por outras opções de resposta. O ideal seria que houvesse um maior número de respondentes de forma a reduzir os problemas nas estimativas.

117. **Questão**
Um mercado na Dinamarca comercializa produtos vencidos que ainda estão próprios para consumo. O mercado acredita que esses produtos devem ser vendidos, em média, num prazo de no máximo 5 dias após a data de vencimento. Um lote com 10 caixas de iogurte, escolhidas ao acaso, obteve tempo médio até a venda de 5.131 dias e desvio padrão de 0.47. Com base nas informações, assumindo que as observações seguem uma distribuição Normal, assinale a altervativa correspondente ao p-valor do teste sobre o tempo médio até a venda desses iogurtes.

(a) 0.200
(b) 0.811
(c) 0.189
(d) 0.806
(e) 0.800

Solução
Primeiramente deve-se observar que o teste é unilateral para a média com variância desconhecida. As hipóteses de teste são

$$H_0 : \mu \leq 5 \ \ versus \ \ H_a : \mu > 5.$$

A estatística de teste normalizada é dada por

$$\frac{\bar{x} - 5}{s/\sqrt{n}} = 0.883$$

e o p-valor

$$\alpha^* = 1 - P(T_9 \leq 0.883)) = 0.2.$$

(a) Verdadeiro
(b) Falso
(c) Falso
(d) Falso
(e) Falso

Figura 16 – *Questão 82*

3.2.2 Parâmetros das habilidades

As habilidades dos alunos, juntamente com os parâmetros dos itens, também foram estimadas pelo do modelo. De acordo com o histogramas das estimativas das proficiências, os valores apresentam um comportamento regular, com grande concentração de alunos com habilidades entre -1 e 1 e poucos alunos com habilidades maior que 2 e menor que -2.

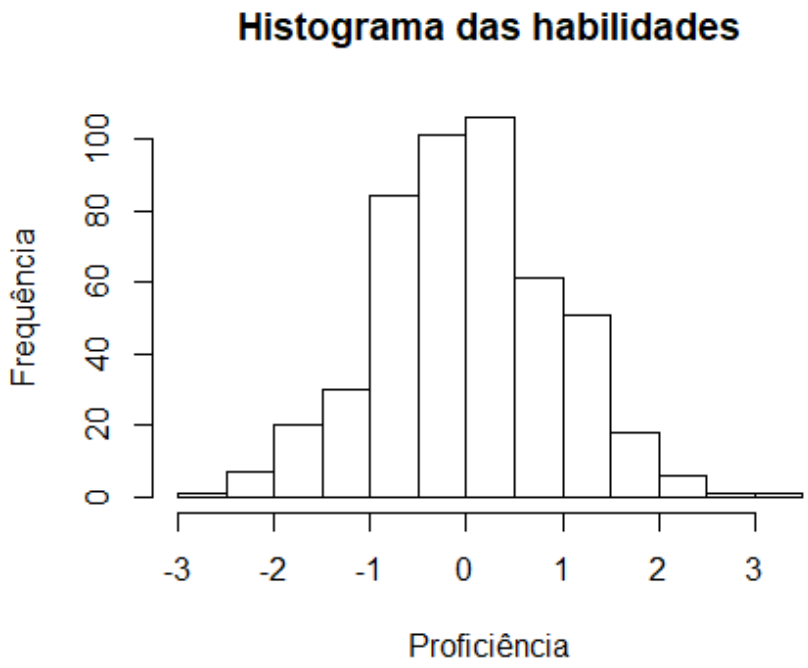


Figura 17 – *Histograma das habilidades dos alunos*

Através das medidas de posição dessas estimativas, também é possível observar um comportamento regular, com valor mínimo estimado de θ de aproximadamente $-2,86$, média e mediana próximo de zero e valor máximo de θ de aproximadamente $3,29$.

Mínimo	-2.863236
1º Quartil	-0.589858
Mediana	0.000026
Média	0.012486
3º Quartil	0.580976
Máximo	3.291880

Tabela 1 – *Medidas de posição das habilidades dos alunos*

Com as habilidades dos alunos estimadas, é possível obter os escores desses alunos caso fossem avaliados pelo modelo de TRI. O gráfico de dispersão da figura 18 nos dá um comparativo entre os escores reais dos alunos e os escores estimados. A correlação entre esses valores foi de aproximadamente 0,877, o que indica uma correlação alta. É interessante observar que os alunos que obtiveram escore 10, teriam seus escores reduzidos pelo modelo de TRI, pelo fato de terem respondido itens mais fáceis que outros alunos, possivelmente devido ao sorteio de itens que é feito para a construção do teste de cada turma.

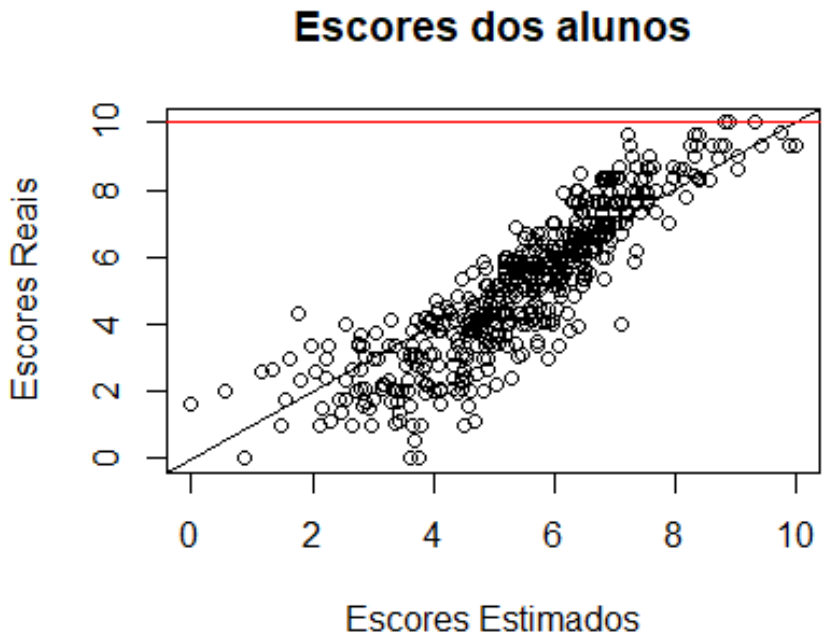


Figura 18 – *Gráfico de dispersão entre os escores reais dos alunos e os escores estimados*

Quando separamos os alunos aprovados dos alunos reprovados, tanto para o método de avaliação tradicional quanto para o método de avaliação via TRI, é observado uma quantidade muito maior de alunos que seriam aprovados pelo novo método. O gráfico 19 mostra que 62 dos alunos reprovados teriam sidos aprovados caso a avaliação fosse realizada seguindo o modelo de TRI (pontos de amarelo), enquanto apenas 7 alunos

aprovados teriam sido reprovados por esse modelo (pontos de laranja). Também vemos que 250 alunos seriam aprovados e 168 seriam reprovados independente do método utilizado para avaliação.

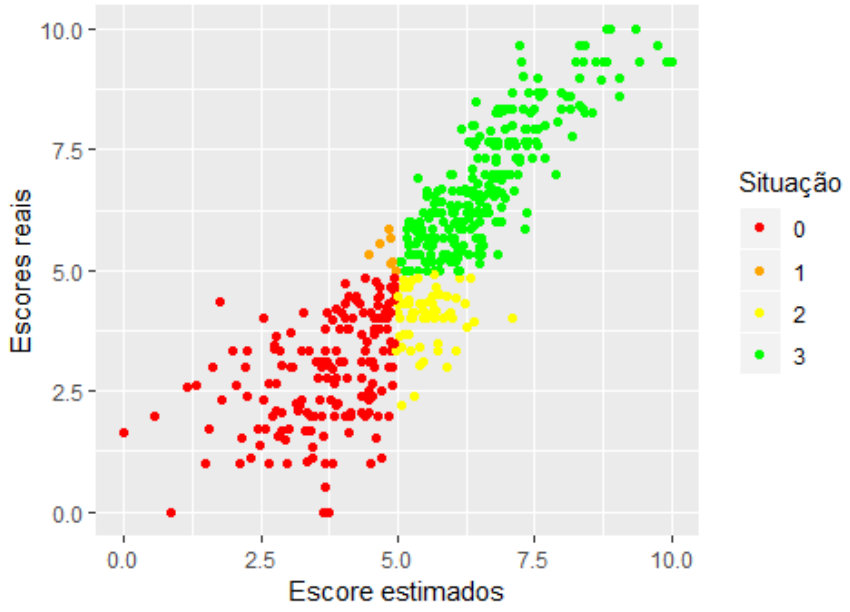
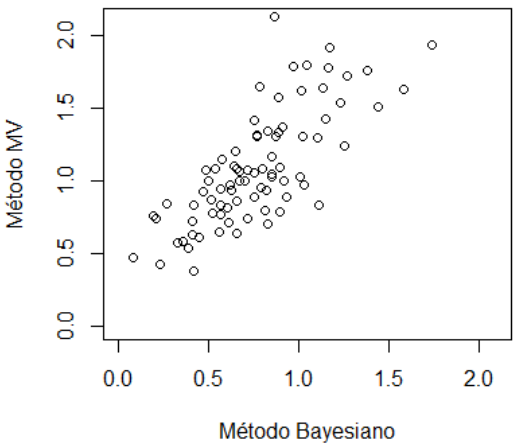


Figura 19 – *Escores reais e estimados, separado por aprovação e reprovação dos alunos.*

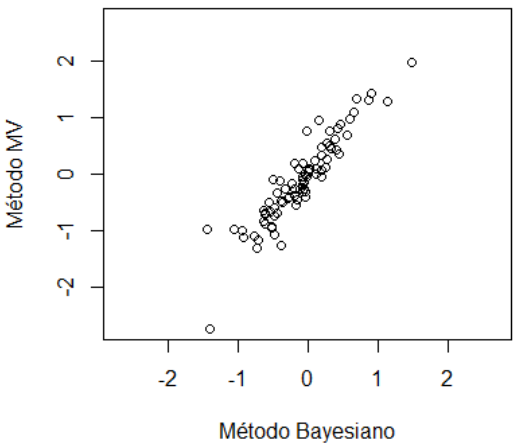
3.3 Comparação entre os métodos Bayesiano e de Máxima Verossimilhança

Para fins de comparação, foi realizada a estimação dos parâmetros pelo método de máxima verossimilhança. No método de MV primeiro estimamos os parâmetros dos itens, através da função *tpm* e depois estimamos as habilidades, pela função *eap*. Mesmo que o foco desse estudo seja para o método Bayesiano, é interessante vermos as estimativas do método MV para podermos comparar ambos métodos e verificar se apresentam discrepâncias e a qualidade deles.

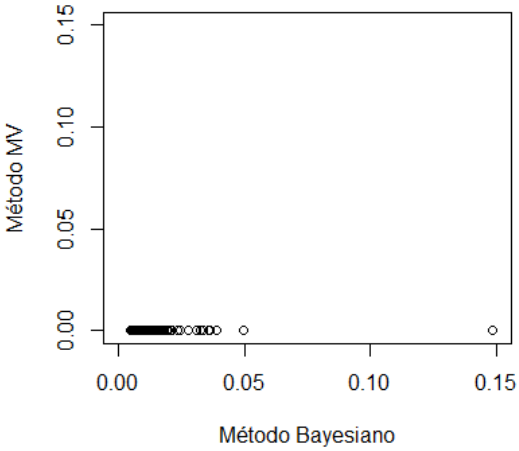
Abaixo temos gráficos de dispersão para cada um dos parâmetros estimados, comparando os métodos Bayesiano e de Máxima Verossimilhança. A correlação é alta e positiva em todos os casos, o que corresponde a uma similariade grande entre os metódos. No caso do parâmetro de "adivinhação", o método Bayesiano estimou apenas 1 item com valor de 0,15 e os demais itens com valores menores que 0,05. É mais provável que os alunos que chutam, ao invés de marcar uma opção aleatoriamente, eles se guiem por um conhecimento prévio ou acabam sendo induzidos a marcarem uma opção errada. O método MV estimou probabilidades de acerto ao acaso iguais a zero para todos os casos.



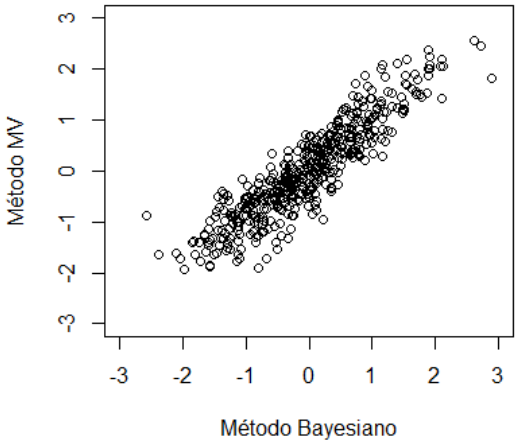
(a) Discriminante



(b) Dificuldade



(c) Probabilidade de acerto ao acaso



(d) Habilidade

Figura 20 – *Comparação entre os métodos Bayesiano e de Máxima Verossimilhança para todos os parâmetros estimados*

4 Conclusão

A realização desse trabalho permitiu encontrar um novo método para avaliação da disciplina de Probabilidade e Estatística da UnB, utilizando um modelo de Teoria de Resposta ao Item. Foi possível observar o nível de dificuldade de cada item que compõe as provas dos alunos bem como seu poder de discriminação. Também identificamos que a Questão 82, apesar de não haver nenhum erro aparente no seu enunciado nem no resultado, se mostrou um item problemático para ser utilizado nas avaliações dos alunos.

Foi visto que, mesmo os alunos que tiraram nota 10 pelo método de avaliação tradicional, não obteriam escore máximo pela avaliação via TRI, uma vez que responderam alguns itens mais fáceis que alunos de outras turmas. O fato dos itens serem sorteados aleatoriamente para compor os testes em cada turma acaba tornando um teste mais fácil que outro, sendo, de certa forma, um pouco injusto. Uma possível minimização para esse problema seria a divisão dos itens do banco de questões em "fáceis","médios"e "difíceis"e então serem sorteados a mesma quantidade de itens com a mesma classificação para cada uma das turmas.

O método de avaliação via modelo de TRI também aprovaria mais alunos do que o método tradicional. É possível para o professor, se for de seu interesse, se basear no método estudado para que alunos muito próximos da aprovação sejam, de fato, aprovados.

O modelo estimado por máxima verossimilhança (MV), se mostrou tão bom quanto o modelo estimado pelo método Bayesiano. De toda forma foi preferível utilizar o método Bayesiano devido ao seu nível de complexidade, precisão na estimação dos parâmetros e técnicas computacionais.

5 Referências

- Andrade, F. D., Tavares, H.R. e Valle, R. C - Teoria de Resposta ao Item: Conceitos e Aplicações, ABE, 2000, Caxambu.

- Brewer, B. J. – *STATS 331 Introduction to Bayesian Statistics*, 2017.

- Baker, F. - *Item Response Theory*, Marcel Dekker, 1992, New York.

- Pasquali, Luiz - *Psicometria: teoria dos testes na psicologia e na educação*, Editora Vozes Limitada, 2017.

- Fox, Jean-Paul - *Bayesian Item Response Modeling*, Springer, 2010.

- Baker, F., Kim, Seock-Ho - *The Basics of Item Response Theory Using R*, Springer, 2017.