Desafio Inteligência - Brasil Cursinhos

Gabriel Ribeiro Freitas

gabriel.ufrgs@ufrgs.br

***Resumo.*** *Neste trabalho foram realizadas duas análises principais utilizando a linguagem de programação R. A primeira análise consiste num estudo geral do simulado em si utilizando modelos de Teoria de Resposta ao Item, de forma a obter curvas características de itens e informação do teste. Já a segunda análise possui aplicação mais voltada ao Cursinho 7, elaborando correlações e comparações entre ele e a população geral.*

**Análise geral do simulado utilizando modelos 2PL e 3PL de TRI.**

**1) Organização e agrupamento dos bancos de dados:**

Antes de realizar qualquer coisa é necessário organizar e agrupar os dados iniciais. Para isso, o arquivo com os gabaritos foi modificado com o auxílio do Excel, transformando-o em dois arquivos .csv distintos, um com o gabarito com espanhol e outro com o gabarito com inglês.

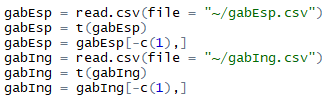
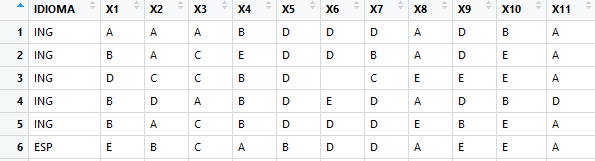


Fig. 1: Leitura dos gabaritos.

Após isso, os arquivos contendo os dados referentes ao dia 1 e dia 2 foram agrupados numa tabela contendo apenas os resultados, figura 2, com finalidade de transformá-la numa tabela binária com os acertos individuais de cada questão, figura 3. Para isso, foi utilizado um algoritmo simples de iteração, mostrado na figura 4, comparando as respostas de cada indivíduo com o gabarito transposto anteriormente.

Fig. 2: Tabela com resultados e idioma.

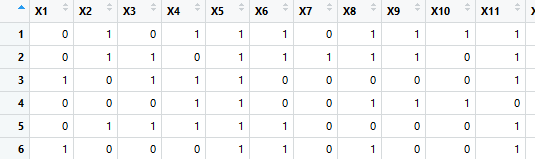


Fig. 3: Matriz binária de acertos.

Os casos em que a resposta não estava presente, ou estava apresentada de forma que não pode ser distinta (*i.e. “D|E”*) foram tratadas simplesmente como sendo erros.

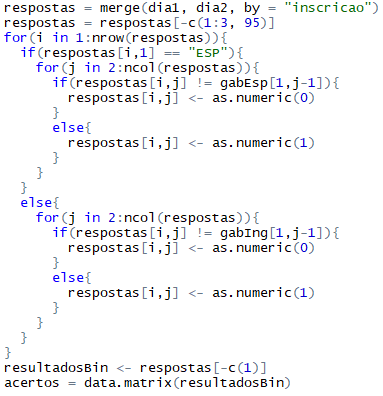


Fig. 4: Algoritmo de criação da matriz *acertos,* contendo as respostas de forma binária.

Após criada a matriz binária de acertos, podemos partir para o modelo de TRI.

**2) Escolha do modelo a ser utilizado:**

Existem três modelos logísticos principais para a Teoria de Resposta ao Item, o modelo de *Rasch, o* modelo com dois parâmetros logísticos e o com três parâmetros logísticos. Para realizar uma análise é necessário escolher o modelo a ser utilizado.

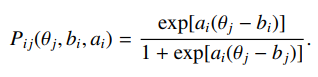
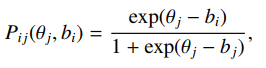
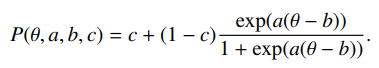
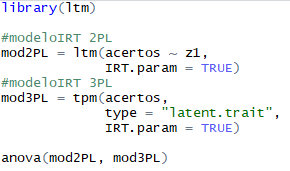
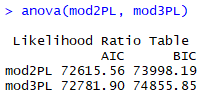
Fig. 5: Modelo de *Rasch*, apresentando um parâmetro logístico (***b - dificuldade do item***), e θ (habilidade de cada indivíduo).

Fig. 6: Modelo 2PL, apresentando dois parâmetros logísticos (***b - dificuldade do item***, e ***a - discriminante***), e θ (habilidade de cada indivíduo).

Fig. 7: Modelo 3PL, apresentando três parâmetros logísticos (***b - dificuldade do item***, ***a - discriminante*** e ***c - parâmetro de chute***), e θ (habilidade de cada indivíduo).

O modelo de *Rasch* foi descartado logo de início por ser muito simples para um simulado com um número grande de questões. Para comparar os modelos 2PL e 3PL então, foi realizada uma análise de variância, técnica estatística que visa verificar se existe uma diferença significativa entre as médias de cada modelo e se os fatores adicionais (no caso de 2PL para 3PL a inserção do parâmetro **c**) exercem influência em alguma variável independente.



É possível então, verificar que o modelo 3PL é o modelo mais adequado para a situação, pois apresenta um *p-value* bastante significativo (*p = 0.23*). É possível também ver dessa análise que a o número de chutes em um simulado é extremamente alto, como esperado.

**3) Análise de itens e informação:**

Com o modelo escolhido em mãos, agora podemos analisar detalhadamente as áreas da prova individualmente em seus itens e informações.

Primeiramente pode ser realizada uma análise qualitativa dos itens de cada área, simplesmente analisando o formato do gráfico e curvas foras da média.

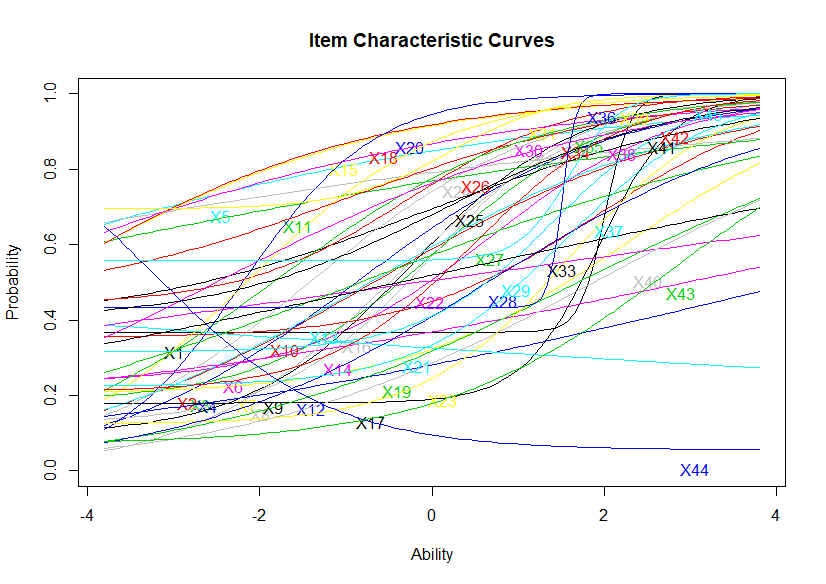


Fig. 8: Curvas características dos itens da área de linguagens, códigos e suas tecnologias.

Na área de linguagens, podemos perceber uma alta variabilidade presente nível de dificuldade, porém com discriminantes não tão bons, com exceção de algumas questões. É possível também analisar que várias questões possuem um parâmetro de chute elevados, com assíntotas à esquerda com valores altos, chegando a quase 70% de chance de acerto no chute em algumas questões. É necessário analisar as questões e reformular as respostas.

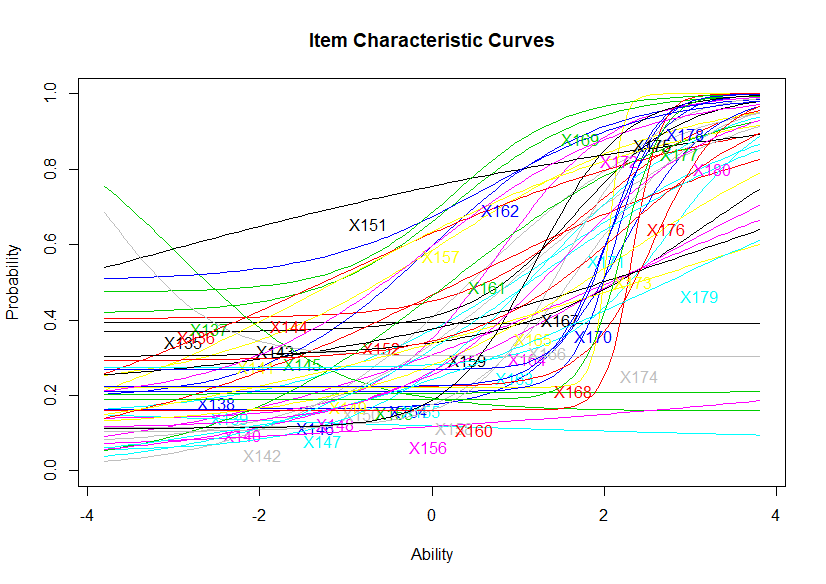
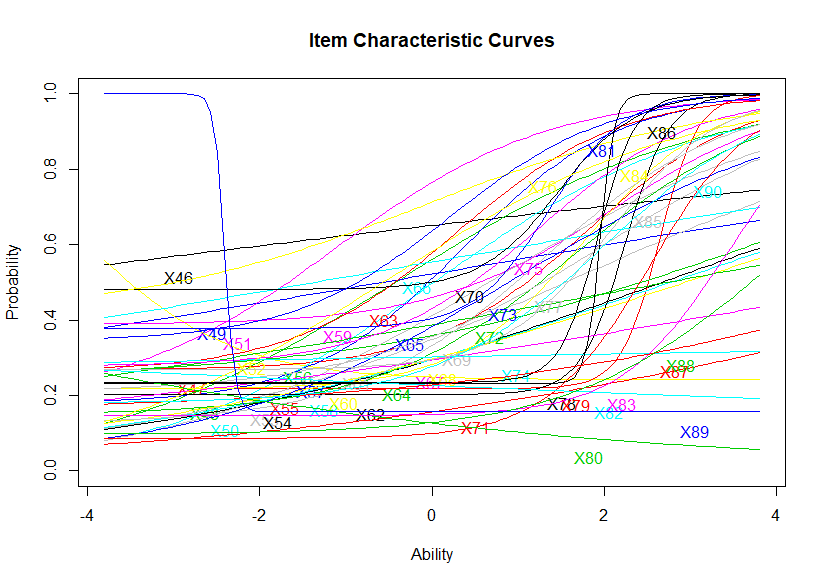
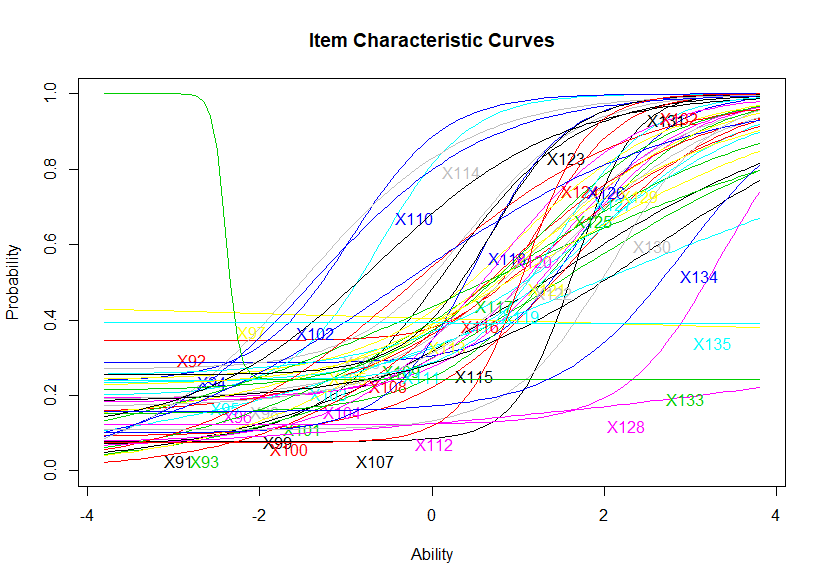


Fig. 9: Curvas características dos itens da área de ciências humanas e suas tecnologias.

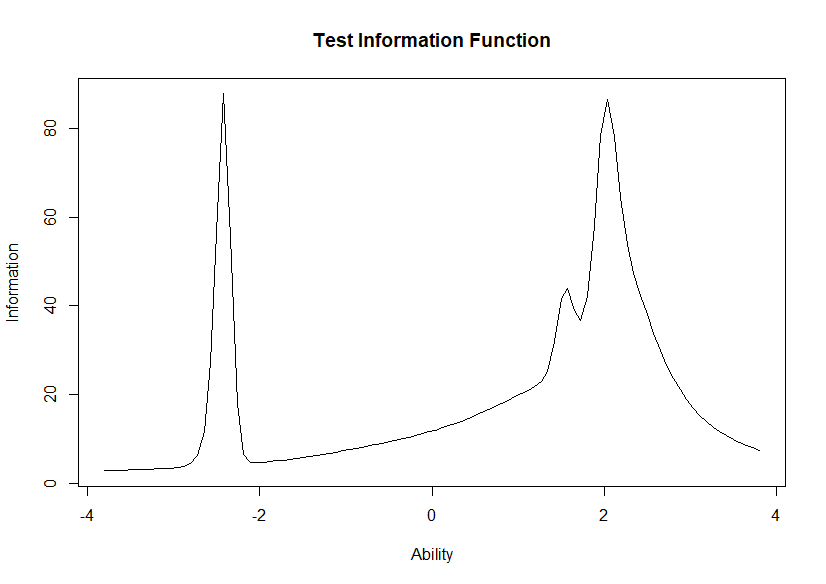
Já na área de ciências humanas a variabilidade no discriminante também é alta, porém contendo bons valores. A prova apresenta dificuldades de habilidade majoritariamente acima de uma dificuldade média. Os parâmetros ***c***  de cada item também não são muito elevador, com maioria chegando, no máximo, a 40%. Poderia se dizer que foi, no geral, uma prova bem construída.

Fig. 10: Curvas características dos itens da área de matemática e suas tecnologias.

Podemos ver um comportamento da prova de matemática muito similar a de ciências humanas, com bons parâmetros de dificuldade e discriminação. Além, conta com índices de chute ainda menores que a área anterior, as respostas foram melhor construídas - existe o viés de ser matemática, porém.

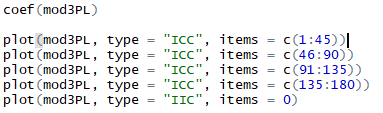
Fig. 11: Curvas características dos itens da área de ciências da natureza e suas tecnologias.

A prova de ciências da natureza possui alto grau de chute, como pode ser verificado nas assíntotas à esquerda. A prova possui ótimos discriminantes, contando com uma dificuldade alta em várias questões. Assim como todas as provas, é possível encontrar itens que precisam ser reformulados tanto na dificuldade, quanto nas múltiplas escolhas.

Fig. 11: Gráfico da função informação do teste.

O gráfico da função informação do teste mostra quando informação podemos tirar do simulado em função da habilidade individual. Podemos perceber dois picos claros, o primeiro pode ser interpretado numa primeira hipótese como sendo o altos valores do parâmetro ***c***, a prova possui diversas questões com respostas fáceis de serem chutadas. Já o segundo pico mostra que o teste consegue discriminar bem indivíduos com habilidade acima de 0, mostrando que a prova estava difícil, apesar dos altos valores do parâmetro ***c.***

Todos os gráficos e análises de coeficientes podem ser obtidas utilizando os seguintes comandos.

Fig. 12: Chamadas de funções para coeficientes e plotagem de gráficos.

Mais análises podem ser feitas, como o cálculo de escore de fatores, ajustamento de indivíduo e de itens, contudo, para uma análise breve e geral, a análise qualitativa dos gráficos é suficiente.

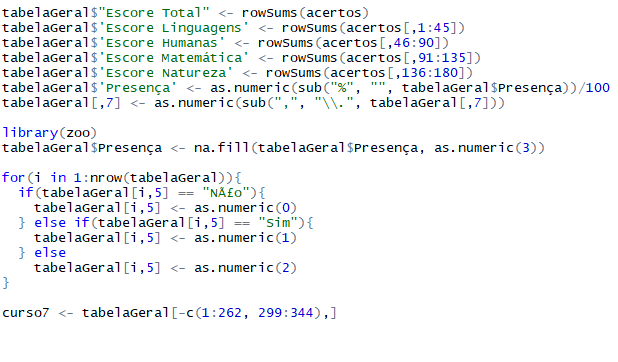


Fig. 13: Formatação da tabela geral e específica.

**Análise geral do simulado utilizando modelos 2PL e 3PL de TRI.**

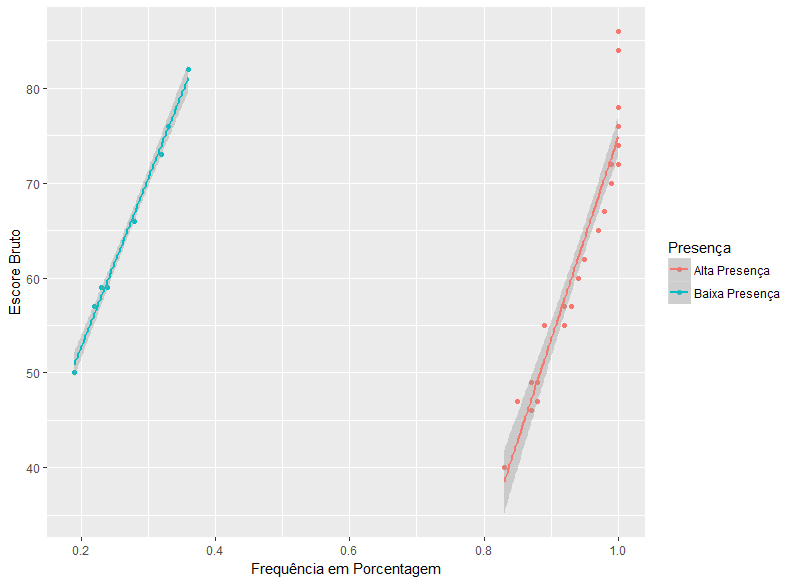
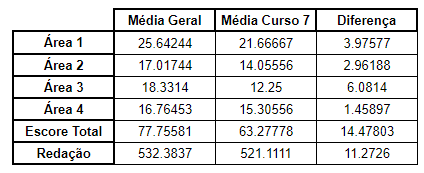
**1) Organização e agrupamento dos bancos de dados:**

Começaremos a análise da mesma forma que a anterior, organizando e agrupando dados considerados relevantes para a análise, no caso, as informações de cada realizante do simulado e seus escores.. Primeiramente é realizado uma limpeza e reorganização dos nomes das colunas, para fins de maior entendimento. Após isso, as colunas contendo números são ajustadas e transformadas em valores numéricos para podermos realizar operações matemáticas e plotar. As operações realizadas estão mostradas na figura abaixo.

A partir da tabela geral, é criada uma sub-tabela contendo apenas os dados dos indivíduos participantes do curso 7.

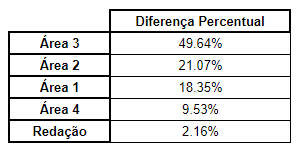
**2) Análise de valores estatísticos:**

Podemos começar a análise calculando a média geral da população e comparando com a amostra do curso 7. Para isso, utilizaremos a função ***mean(x)*** para cada coluna dos escores***.***

Tabela 1: Tabela de médias de acertos, comparando a geral com a do curso 7.

* **Área 1** **-** Linguagens, códigos e suas tecnologias;
* **Área 2** **-** Ciências humanas e suas tecnologias;
* **Área 3** **-** Matemática e suas tecnologias;
* **Área 4** **-** Ciências da natureza e suas tecnologias;

Apesar dos alunos do cursinho terem ficado com média de acertos menor que a população geral em todas as áreas, pode-se verificar quais as áreas que necessitam de mais enfoque. Foi calculado então a diferença percentual entre as áreas e posto em ordem.

Tabela 2: Diferença percentual em ordem das áreas.

As áreas que precisam de maior número de aulas, monitoria ou revisão são então, em ordem, matemática, ciências humanas, linguagens, ciências da natureza e redação. Para uma análise geral e básica, a divisão entre as cinco competências são suficientes, porém, para maior detalhamento, pode ser realizada também uma análise em cada disciplina separadamente, separando, por exemplo, ciências humanas em geografia, história, filosofia, etc.

**2) Análise de correlações no cursinho 7:**

Após a análise comparativa direta, é possível levantar hipóteses correlativas entre variáveis.

**2.1) Escore Bruto x Frequência nas Aulas**

Gráfico 1: Gráfico de dispersão com regressão linear, correlacionando o Escore Bruto x Frequência

Percebe-se inicialmente a existência de dois grupos, um que possui baixa frequência em aula, e outro que possui alta frequência em aula. Em ambos os grupos podemos ver que existe uma correlação entre presença em aula e escore bruto no simulado.

Algo que o cursinho pode fazer é estimular mais os alunos de ambos os grupos frequentarem mais às aulas, mas principalmente o primeiro grupo.

Foram hipotetizadas outras correlações, porém nenhuma outra se mostrou relevante para uma análise básica.

**Conclusão**

Foram feitas análises tanto do simulado quanto do cursinho 7, com o objetivo de comparar a amostra com a população geral. Muitas outras análises específicas e detalhadas podem ser realizadas utilizando o *dataset* inicial, mas, por fins de recursos (tempo incluso), serão realizadas em outra oportunidade. Em relação ao cursinho 7: aulas de matemática deverão ser reforçadas, proporcional a diferença percentual entre as médias apresentadas na tabela 2. Ao mesmo tempo, presença deve ser cobrada mais rigorosamente.

**Notas Adicionais**

* Infelizmente acabei me inscrevendo e descobrindo sobre o desafio dois dias antes do *deadline* do desafio, mais trabalhos específicos poderiam ter sido realizados, mas por questão de tempo apenas foi possível realizar uma análise básica.
* Códigos em R e tabelas utilizadas podem ser encontradas no seguinte link:
  + <https://goo.gl/AVrtTM>
* Qualquer dúvida sobre o trabalho realizado é só entrar em contato! Espero poder trabalhar com vocês :D