

Pôster: Geração aleatória e correção automática de questões através do R/`exams`

Julho 2022

Introdução

- Planejar e executar avaliações é um dos grandes desafios no processo de ensino e aprendizagem.
 - Em turmas de ensino à distância (EAD) esse aspecto necessita de maior atenção.
- Importante criar mecanismos avaliativos que evitem a repetição de questões entre os alunos e inibam potenciais cópias/plágios.
- Se destacam a criação de questões e a sua correção, bem como a análise dos resultados e o retorno da avaliação ao estudante.
- No nosso Departamento as turmas EAD de Probabilidade e Estatística atendem entre 300 e 400 alunos nos últimos semestres. ufrgs.br/probabilidade-estatistica

Porque R e o pacote `exams`?

- Utilizar o conhecimento em R para auxiliar na criação de questões aleatorizadas.
- Aproveitar as facilidades do formato de arquivos `Rmarkdown` (integrar códigos em R e texto).
- Possibilidade de gerar provas impressas ou online, com correção automática.

Motivação

- Mecanismos que auxiliem na elaboração e correção automática das avaliações possibilitam que professores tenham mais tempo para planejar o instrumento de avaliação e analisar o desempenho da turma.
- A criação das questões é parte essencial no processo de flexibilização das avaliações e exige conhecimento de programação, no software R, pacotes `Markdown`, linguagem `Latex` e da teoria estatística envolvida.

Contribuições

- Criação do banco de questões no formato `exams`, códigos em R para geração e correção automática de avaliações, no formato impresso ou XML.
- Otimização do processo de avaliação, com o auxílio de recursos que minimizem atividades repetitivas na geração, correção, análise e divulgação dos resultados, presenciais ou à distância.

Desenvolvimento

Criação das Questões

- O desenvolvimento das questões é uma tarefa complexa que exige conhecimentos sobre construção de provas múltipla escolha bem como programação no software R.

- Juntamente com a criação dos enunciados e alternativas de respostas, a construção do gabarito da questão já é feita de maneira integrada com o pacote **exams**.
- Diferentes tipos de formatos para respostas: abertas e única/múltipla escolha
 - formatos do exams: numérica (**numeric**), discursiva (**string**), única escolha (**schoice**), múltiplas escolhas (**mchoice**) e combinações entre os tipos.

Questão 1
Incorreto
Valor: 1,00 (pontuação)

Em certo banco de dados, o tempo para realização das buscas é aproximadamente normal, com média de 59 s e desvio padrão de 11 s. Depois de realizadas algumas modificações no sistema, observou-se que, em 28 consultas, o tempo médio caiu para 54,2 s. Há evidência de melhoria? Admita que as 28 observações possam ser consideradas uma amostra aleatória e que não houve alteração na variância. Use o nível de significância de 0,01.

Escolha uma opção:

☐ a. Não, não há evidência de melhoria; rejeitamos H_0 .

☒ b. O teste não se aplica à situação. ✗ Incorreta. O teste não se aplica à situação.

☐ c. Sim, há evidência de melhoria; rejeitamos H_1 .

☐ d. Sim, há evidência de melhoria; rejeitamos H_0 .

☐ e. Não, não há evidência de melhoria; não se rejeita H_0 .

Questão 1
Incorreto
Valor: 0,00 (pontuação)

Em certo banco de dados, o tempo para realização das buscas é aproximadamente normal, com média de 52 s e desvio padrão de 14 s. Depois de realizadas algumas modificações no sistema, afirmou-se de diminuir o tempo, observou-se que, em 24 consultas, o tempo médio mudou para 53,4 s. Admita que as 24 observações possam ser consideradas uma amostra aleatória e que não houve alteração na variância. Use o nível de significância de 0,1. Responda:

OBS: Os números entre parênteses no início das questões referem-se ao número de casas decimais da resposta e à tolerância considerada.

a. Indique a alternativa abaixo que possui as hipóteses adequadas ✗

• [1] $H_0 : \mu = 52$ contra $H_1 : \mu \neq 52$.

• [2] $H_0 : \mu = 52$ contra $H_1 : \mu > 52$.

• [3] $H_0 : \mu = 52$ contra $H_1 : \mu \leq 52$.

b. Qual a estatística de teste apropriada? ✗ $F = F$ de Snedecor.

c. (2; 0,05) Qual o valor crítico do teste (o valor que define a região crítica)? ✗ 1,96

d. (2; 0,05) O valor da estatística de teste é ✗ 2

e. Existe evidência de melhoria? Decida e conclua sobre o teste.

☐ O teste não se aplica à situação. ✗

Do enunciado temos que $\bar{x} = 54,2$, $\sigma = 11$, $\alpha = 0,01$, $n = 28$ e $\mu_0 = 56$.

Para testar as hipóteses $H_0 : \mu = 56$ contra $H_1 : \mu < 56$, então a estatística de teste calculada é dada por

$$z_{\text{calc}} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{54,2 - 56}{\frac{11}{\sqrt{28}}} = \frac{-1,8}{2,08} = -0,87.$$

Assim, utilizando a tabela da distribuição normal padrão, obtemos $z_{0,01} = z_{0,01} = 2,33$. Então não rejeitamos H_0 , pois no teste unilateral não se rejeita H_0 caso $z_{\text{calc}} \geq -z_{\text{tab}}$. Ou seja, não há evidência de melhoria.

a. Incorreta. Não, não há evidência de melhoria; rejeitamos H_0 .

b. Incorreta. O teste não se aplica à situação.

c. Incorreta. Sim, há evidência de melhoria; rejeitamos H_1 .

d. Incorreta. Sim, há evidência de melhoria; rejeitamos H_0 .

e. Correta. Não, não há evidência de melhoria; não se rejeita H_0 .

A resposta correta é: Não, não há evidência de melhoria; não se rejeita H_0 .

Do enunciado temos que $\bar{x} = 53,4$, $\sigma = 14$, $\alpha = 0,1$, $n = 24$ e $\mu_0 = 52$.

Para testar as hipóteses $H_0 : \mu \geq 52$ contra $H_1 : \mu < 52$, então a estatística de teste calculada é dada por

$$z_{\text{calc}} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{53,4 - 52}{\frac{14}{\sqrt{24}}} = \frac{1,4}{2,86} = 0,49.$$

Assim, utilizando a tabela da distribuição normal padrão, obtemos $z_{0,1} = -1,28$. Então não rejeitamos H_0 , pois no teste unilateral não se rejeita H_0 caso $z_{\text{calc}} \geq -z_{\text{tab}}$. Ou seja, não há evidência de melhoria.

Figure 1: Enunciado e o gabarito de uma questão respondida no Moodle no formato ‘schoice’ (esquerda) e no formato ‘cloze’ (direita).

- Por fim, estressar/testar as questões, para avaliar possíveis erros e reformulações, pode ser facilitada com a função **stresstest_exercise()**. Apenas as questões que já foram completamente testadas devem compor o banco de questões.

Banco de Questões

- A organização das questões é fundamental para a eficiência na criação da avaliações. Um protocolo para criação de questões e inclusão no banco pode auxiliar nessa tarefa.
- Para pré visualizar um questionário/questões:
 - a função **exams2html()** gera um arquivo HTML com a(s) questão(ões);
 - ou a função **exams2pdf()** gera um arquivo pdf, mas precisa ter alguma distribuição LaTeX instalada.

Questionários

- O **exams** utiliza um objeto **list** com nomes das questões/caminho;
 - cada elemento da lista pode ser um único nome ou vetor de nomes de questões;
 - se for um vetor, a aleatorização também ocorrerá entre diferentes questões.
- Para gerar avaliações e provas impressas a função **exams2nops(...)** auxiliam no processo.
- para gerar XML e aplicar provas remotas, no Moodle por exemplo, usar a função **exams2moodle(...)**.
 - Poderíamos criar um arquivo XML com códigos puramente em R sem usar o exams, porém a estrutura e organização criado pelo exams.

Correção da avaliação

- Com a utilização do **exams** a correção das avaliações se torna um trabalho puramente mecânico, também no formato impresso.
 - A correção é feita com a leitura das provas feita em qualquer scanner e o reconhecimento das respostas é realizada pelo próprio pacote **exams** através de um software de reconhecimento óptico de caracteres. A Figura apresenta o resultado da correção no formato impresso.
- Para questões geradas nos formatos digitais, a correção é feita automaticamente após a finalização da atividade, conforme ilustrado na Figura .

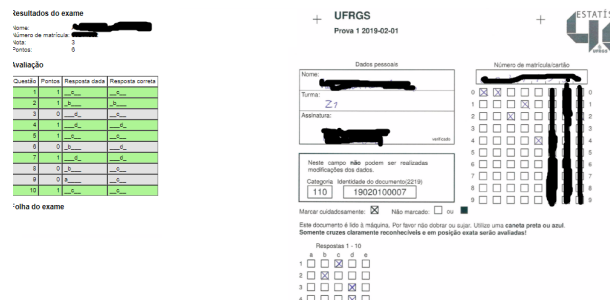
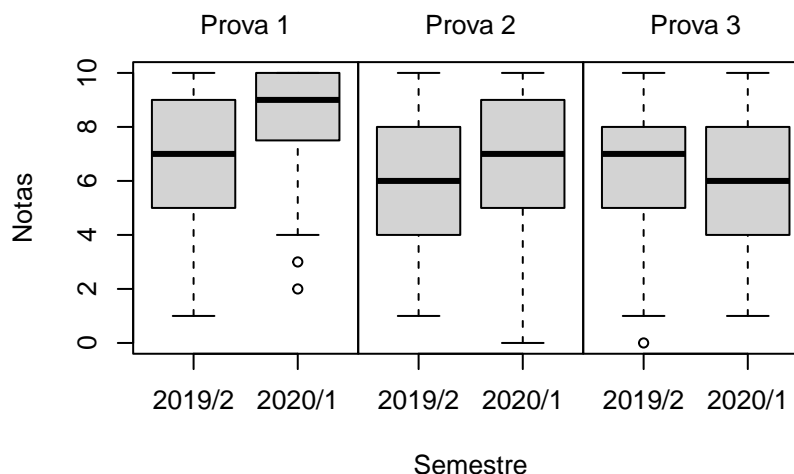


Figure 2: Arquivo HTML gerado na correção automática da avaliação impressa (acima) e a folha de respostas preenchida (abaixo).

- O retorno da atividade/correção aos estudantes é realizado por envio automático de email com o resultado da avaliação para cada aluno, no formato impresso.

Análise: comparação da distribuição das notas anteriormente e no período do Ensino Remoto Emergencial (ERE).

- Podemos considerar cada ano/semestre uma coorte.
 - Em 2019 as avaliações semanais foram online e valiam presença. As provas presenciais, com questões no formato de múltiplas alternativas e única escolha.
 - Em 2020 e 2021 tivemos o regime de Ensino Remoto Emergencial (ERE) e as avaliações semanais valiam nota. As provas foram remotas, sem supervisão, com questões em outros formatos também, como numéricas, ou de associação, sendo incorporadas aos questionários.
- Ilustração:* O que a distribuição das notas de 2019/2 e 2020/1 nos sugerem?
 - Uma análise cuidadosa é necessária para concluirmos sobre diferenças nos instrumentos/formatos de avaliações. Estudantes e colaboradora(o)s são bem vinda(o)s.
 - O gráfico abaixo é apenas uma ilustração com dados limitados. Na prova 1 ainda estávamos impementando diferentes formatos do que os de única escolha utilizados nas avaliações presenciais. Ao longo do semestre 2020/1 fomos adotando diferentes formatos de questões, na prova 3 estávamos com mais questões abertas. Como sera a comparacao com semestres anteriores e posteriores ao período ilustrado?



Bilgin e Lin (2022) Usam o pacote `exams` para criar questões e comparam o desempenho dos estudantes antes e durante a pandemia causada pela Covid-19. Concluem não haver diferença expressiva nos resultados das avaliações, que os instrumentos foram muito semelhantes na mensuração do conhecimento, sugerindo manter a integridade acadêmica embora no segundo período os testes tenham sido online e sem a supervisão do professor.

Conclusões

- Facilitar o processo de geração de questões e correção em disciplinas de massa é necessário.
- Colaboração e organização na criação de avaliações e banco de questões também são fundamentais.
- A avaliação
 - **online** tem maior flexibilidade de formatos, porém menor controle. A aleatorização como forma de evitar cópias/plágios
 - **presencial** se torna escalável com baixo custo, a aplicação e geração de avaliações com a leitura óptica economiza muito tempo na correção das avaliações.
- Nas provas impressas, o retorno da prova corrigida para o aluno, via email, faz com que o estudante se sinta mais integrado à disciplina.
- O processo torna a criação, correção e análise dos resultados de avaliações mais eficiente, dessa forma o docente terá uma ferramenta para facilitar na condução da disciplina.
- É possível criar um processo operacional, para que as ferramentas desenvolvidas sejam estendidas para diferentes formatos de cursos/disciplinas, criar parcerias com outras Universidades, tudo via R, software livre e gratuito.

Trabalhos futuros

- Juntamente à análise do desempenho dos alunos podemos aplicar a metodologia de teoria de resposta ao item (TRI), para analisar o grau de dificuldade das questões auxiliando na criação de questões e composição dos questionários.

Agradecimentos

Ao auxílio da Secretaria de Educação à Distância (SEAD/UFRGS). A(o)s monitora(e)s envolvida(o)s no projeto.

- **Dúvidas, colaborações e acesso às questões enviar email para... QR code...**
- ufrgs.br/probabilidade-estatistica

Referências

Achim Zeileis, Nikolaus Umlauf, Friedrich Leisch (2014). Flexible Generation of E-Learning Exams in R: Moodle Quizzes, OLAT Assessments, and Beyond. *Journal of Statistical Software* 58(1), 1-36.

Ayse Aysin Bilgin, Huan Lin (2022). Designing assessment tasks to prevent cheating in a large first-year statistics unit (2022) Conference paper DOI: 10.52041/iase.errob

Bettina Gruen, Achim Zeileis (2009). Automatic Generation of Exams in R. *Journal of Statistical Software* 29(10), 1-14

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.