



CEFET/RJ

Avaliação Experimental



Eduardo Ogasawara

eduardo.ogasawara@cefet-rj.br

<https://eic.cefet-rj.br/~eogasawara>

Introdução à avaliação experimental

- A experimentação é um dos pilares fundamentais da ciência
- Na computação, os experimentos são frequentemente conduzidos com datasets para testar hipóteses sobre algoritmos e sistemas
- Para que um experimento tenha validade científica, ele deve ser bem concebido e demonstrado
- Exemplo:
 - Para validar um novo classificador de imagens, um pesquisador pode testar sua implementação em datasets padrão, como MNIST ou imagenet, comparando os resultados com modelos já estabelecidos

Tipos de avaliação experimental

- Processos automatizados
 - Algoritmos, workflows, simulações
 - Exemplo: avaliação de desempenho de um modelo de reconhecimento facial em diferentes condições de iluminação
- Processos manuais
 - Questionários, experimentos com humanos, estudos qualitativos
 - Exemplo: estudo de usabilidade de uma nova interface de software

Ética em avaliação experimental

- Quando um experimento envolve humanos, é necessário passar por um comitê de ética
- Exemplos de casos que exigem aprovação ética:
 - Estudos de usabilidade em interfaces de usuário
 - Experimentos envolvendo testes psicológicos ou comportamentais
 - Pesquisas que coletam dados sensíveis (como reconhecimento facial)

Procedimento para avaliações baseadas em formulário

- Passos essenciais para avaliações com humanos:
 1. Elaborar um questionário bem estruturado
 2. Obter aprovação do comitê de ética
 3. Coletar o termo de consentimento livre e esclarecido (TCLE)
 4. Definir um número adequado de participantes
 5. Analisar os dados qualitativa e quantitativamente
- Exemplo:
 - Pesquisa sobre experiência do usuário (UX) em um novo aplicativo

Termo de consentimento livre e esclarecido (TCLE)

- Por que é necessário?
- Assegura que os participantes compreendam os objetivos da pesquisa
- Define riscos, benefícios e confidencialidade dos dados
- Deve ser assinado antes da coleta dos dados
- Referência: diretrizes éticas do conselho nacional de saúde

Definição de baseline

- O baseline deve ser o estado da arte para o problema estudado
- Exemplo:
 - No processamento de linguagem natural (PLN), transformers substituíram word embeddings e redes LSTM como baseline
- Em problemas complexos, um único baseline pode ser insuficiente
- Diferentes métodos podem ser usados como referência
- Exemplo:
 - Em previsões meteorológicas, pode-se usar modelos estatísticos e redes neurais recorrentes como baseline

Table 2

Performance results for temperature forecasting using the previous five observations (grids) to predict the next five observations ($5 \rightarrow 5$), and the next 15 observations ($5 \rightarrow 15$). We highlight the lowest values among the models.

5 \rightarrow 5					
Model	RMSE	MAE	Memory usage (MB)	Mean training time	Training time/epoch
ARIMA	2.1880	1.9005	–	–	–
ConvLSTM [19]	1.8555 ± 0.0033	1.2843 ± 0.0028	922	02:38:27	00:02:21
PredRNN [25]	1.6962 ± 0.0038	1.1885 ± 0.0020	2880	06:59:34	00:05:52
MIM [26]	1.6731 ± 0.0099	1.1790 ± 0.0055	4145	11:05:37	00:10:43
STConvS2S-C (ours)	1.3699 ± 0.0024	0.9434 ± 0.0020	1040	03:34:52	00:02:48
STConvS2S-R (ours)	1.2692 ± 0.0031	0.8552 ± 0.0018	895	03:15:12	00:02:13
5 \rightarrow 15					
Model	RMSE	MAE	Memory usage (MB)	Mean training time	Training time/epoch
ARIMA	2.2481	1.9077	–	–	–
ConvLSTM [19]	2.0728 ± 0.0069	1.4558 ± 0.0076	1810	5:29:30	00:07:32
PredRNN [25]	2.0237 ± 0.0067	1.4311 ± 0.0149	7415	11:45:48	00:17:03
MIM [26]	2.0287 ± 0.0361	1.4330 ± 0.0250	10673	19:19:00	00:31:19
STConvS2S-C (ours)	1.8739 ± 0.0107	1.2946 ± 0.0061	1457	03:12:24	00:05:17
STConvS2S-R (ours)	1.8051 ± 0.0040	1.2404 ± 0.0068	1312	03:15:42	00:05:03

Coleta e análise de dados

- Perguntas essenciais antes da análise:
 - Quais são as fontes dos dados?
 - Foram aplicadas técnicas de pré-processamento?
 - Os dados contêm viés ou falhas? Como mitigá-los?
- Exemplo:
 - Em detecção de fraudes bancárias, um dataset enviesado pode não representar todos os tipos de fraude

Data papers

- A criação de datasets é um esforço significativo em diversas áreas
- Data papers são publicações focadas na documentação de datasets
- Eles possibilitam reprodutibilidade e incentivam novas pesquisas
- Exemplo:
 - O dataset *integrated dataset of brazilian flights* permitiu estudos sobre padrões de atraso e eficiência no tráfego aéreo no Brasil

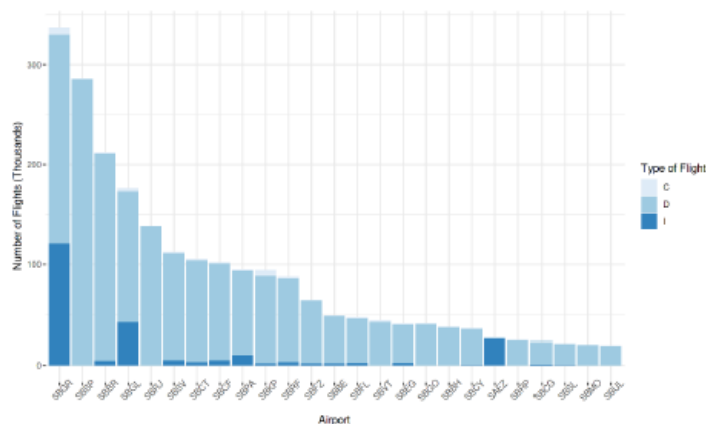


Figure 2. Number of flights per airport, for the top-25 most active airports

Métodos de análise qualitativa e quantitativa

- **Análise quantitativa**
 - Baseada em métricas objetivas e estatísticas
 - Exemplo: comparação de precisão entre dois algoritmos
- **Análise qualitativa**
 - Interpretação subjetiva dos dados (entrevistas, questionários)
 - Exemplo: pesquisa sobre percepção de usuários sobre uma interface

Interpretação de resultados

- A análise dos resultados deve considerar múltiplas interpretações
- Testes adicionais podem ser necessários para eliminar interpretações equivocadas
- Exemplo:
 - Um modelo de aprendizado de máquina que falha ao superar o baseline pode indicar que a abordagem não é adequada ou que a implementação não foi otimizada corretamente

Erro experimental e controle de variáveis

- O que é erro experimental?
 - Diferenças nos resultados devido a flutuações no processo
 - Pode ser reduzido com múltiplas replicações e planejamento estatístico
- Erro sistemático vs. Erro aleatório
 - Erro aleatório: pequenas variações nos resultados entre execuções, devido a ruído ou incertezas naturais
 - Erro sistemático: desvio constante causado por problemas no projeto experimental (exemplo: viés na coleta de dados)
- Como reduzir o erro experimental?
 - Replicação interna: realizar múltiplas execuções sob as mesmas condições
 - Replicação externa: outro pesquisador repetir o experimento para verificar se os achados se mantêm
 - Controle de variáveis: garantir que fatores externos não alterem os resultados
- Exemplo:
 - Em reconhecimento facial, treinar um modelo apenas com imagens de uma etnia pode levar a um erro sistemático, fazendo com que o modelo tenha desempenho ruim em outras populações.

Projetando um experimento científico

- Passos essenciais:
 1. Definir as variáveis de resposta
 2. Escolher os fatores e parâmetros
 3. Selecionar o tipo de experimento
 4. Determinar o número de replicações
 5. Selecionar amostras representativas
- Requisitos para um bom experimento:
 - Previsibilidade: os experimentos devem demonstrar tendências gerais, não apenas funcionar em datasets específicos
 - Robustez: os resultados devem ser consistentes e não ambíguos

[1] c. Alexopoulos, 2007, statistical analysis of simulation output: state of the art, in: *proceedings - winter simulation conference*, p. 150–161

[2] a.M. Law, 2007, statistical analysis of simulation output data: the practical state of the art, in: *proceedings - winter simulation conference*, p. 77–83

Variável de resposta, parâmetros e fatores

- Variável de resposta
 - Saída do experimento (variável dependente)
 - Exemplo: a precisão de um classificador de imagens
- Parâmetros
 - Características fixas durante o experimento
 - Exemplo: número de camadas de uma rede neural
- Fatores
 - Variáveis independentes modificáveis
 - Exemplo: diferentes técnicas de pré-processamento de dados

Análise de experimentos automatizados

- Perguntas essenciais:
 - O experimento foi projetado para que os dados gerados possam ser analisados corretamente?
 - Quantas execuções são necessárias para uma inferência estatística confiável?
 - Devo descartar outliers na saída do experimento?

Período de aquecimento ("warm-up period")

- O que é o período de aquecimento?
 - Em simulações, os primeiros instantes podem gerar dados instáveis
 - Esses dados devem ser descartados antes da análise se houver evidências de que não representam o estado estável do sistema
- Como determinar se há um período de aquecimento?
 - Visualização dos dados: plotar as métricas de interesse ao longo do tempo e verificar se há estabilização
 - Teste de convergência estatística: aplicar testes como chow test ou cusum test para detectar mudanças na distribuição dos dados
 - Comparação de múltiplos experimentos: rodar múltiplas execuções e verificar se os primeiros dados são significativamente diferentes dos posteriores
- Exemplo:
 - Em uma simulação de filas, os primeiros minutos podem ter tempos de espera reduzidos porque poucas pessoas chegaram ao sistema. Esse efeito desaparece com o tempo e os tempos médios se estabilizam

Validade interna e externa dos experimentos

- O que é validade interna?
 - Garante que a relação entre causa e efeito dentro do experimento é real
 - Elimina fatores externos que possam influenciar os resultados
- O que é validade externa?
 - Indica se os resultados do experimento podem ser generalizados para outros contextos
 - Depende da representatividade dos dados e das condições experimentais

[1] c. Stoddart, 2016, is there a reproducibility crisis in science?, Nature (may.)

[2] m. Baker, 2016, 1,500 scientists lift the lid on reproducibility, nature, v. 533, n. 7604 (may.), P. 452–454.








Tamanho da amostra e poder estatístico

- Por que o tamanho da amostra é importante?
 - Uma amostra pequena pode gerar resultados enganosos
 - Um número inadequado de testes reduz a confiança nos resultados
- Como determinar o tamanho adequado da amostra?
 - Usar cálculo de poder estatístico
 - Verificar nível de confiança e margem de erro
 - Aplicar validação cruzada para garantir robustez
- Exemplo:
 - Comparar dois algoritmos treinados com apenas 100 exemplos pode levar a conclusões erradas. Com 10.000 exemplos, a análise estatística será muito mais confiável

[1] c. Stoddart, 2016, is there a reproducibility crisis in science?, Nature (may.)

[2] m. Baker, 2016, 1,500 scientists lift the lid on reproducibility, nature, v. 533, n. 7604 (may.), P. 452–454.

Comparação correta de métodos e erros comuns em avaliação experimental

-  Erros comuns ao avaliar modelos:
 -  comparar métodos sem usar o mesmo conjunto de dados
 -  Não considerar variabilidade dos resultados
 -  Ignorar métricas estatísticas relevantes (média, desvio padrão, IC95%)
 -  Testar em amostras muito pequenas que não representam o problema
 -  Fairness: não avaliar se um modelo apresenta viés para diferentes grupos
-  Boas práticas:
 - Usar mesmo conjunto de dados e métricas para todos os métodos
 - Testar sob as mesmas condições de hardware/software
 - Aplicar validação cruzada para evitar overfitting
 - Reportar intervalos de confiança para indicar incerteza nos resultados
- Exemplo:
 - Comparar um modelo treinado com 1 milhão de amostras contra outro treinado com apenas 10.000 amostras não é uma comparação justa

[1] c. Stoddart, 2016, is there a reproducibility crisis in science?, Nature (may.)

[2] m. Baker, 2016, 1,500 scientists lift the lid on reproducibility, nature, v. 533, n. 7604 (may.), P. 452–454.

Testes estatísticos e interpretação de resultados

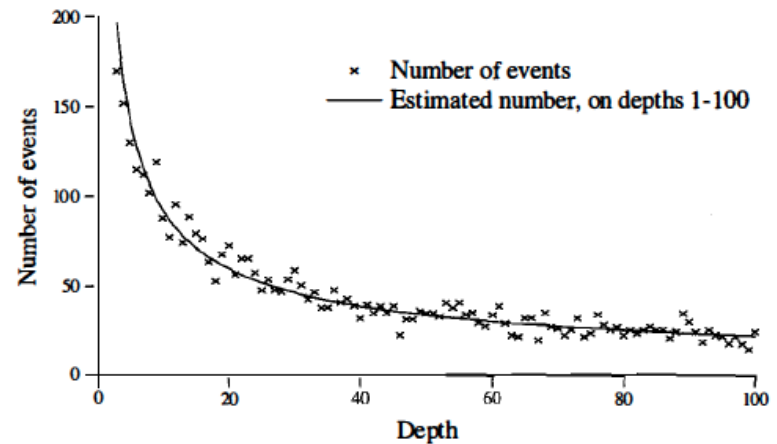
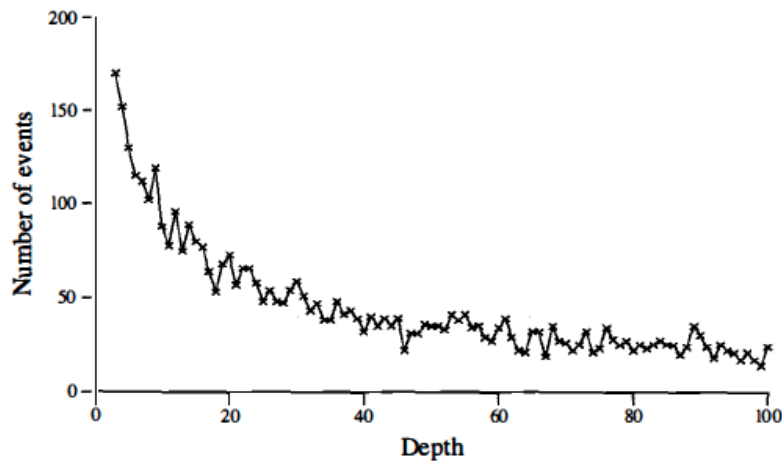
- Quando usar cada teste?
 - Teste t pareado: comparação entre dois métodos no mesmo conjunto de dados
 - Teste de wilcoxon: para amostras pequenas e não paramétricas
 - Anova: para comparar múltiplos métodos simultaneamente
- Como interpretar os testes estatísticos corretamente?
 - Intervalo de confiança (ic95%) → mostra a faixa onde os valores reais provavelmente se encontram
 - P-value < 0.05 → indica evidência estatística forte de que um método é superior
 - Variabilidade dos resultados → usar desvio padrão, boxplot e distribuição dos dados
- Exemplo:
 - Dois algoritmos de IA têm precisão de 92% e 94%. Será que a diferença é estatisticamente significativa? Um teste t pareado e um IC95% podem responder

[1] c. Stoddart, 2016, is there a reproducibility crisis in science?, Nature (may.)

[2] m. Baker, 2016, 1,500 scientists lift the lid on reproducibility, nature, v. 533, n. 7604 (may.), P. 452–454.

Como apresentar resultados?

- Inclua sempre:
 - Objetivo do experimento
 - Dados utilizados
 - Resultados principais
- Erro comum
 - Mostrar uma linha interpolando pontos sem significado estatístico



[1] c. Stoddart, 2016, is there a reproducibility crisis in science?, *Nature* (may.)

[2] m. Baker, 2016, 1,500 scientists lift the lid on reproducibility, *nature*, v. 533, n. 7604 (may.), P. 452–454.

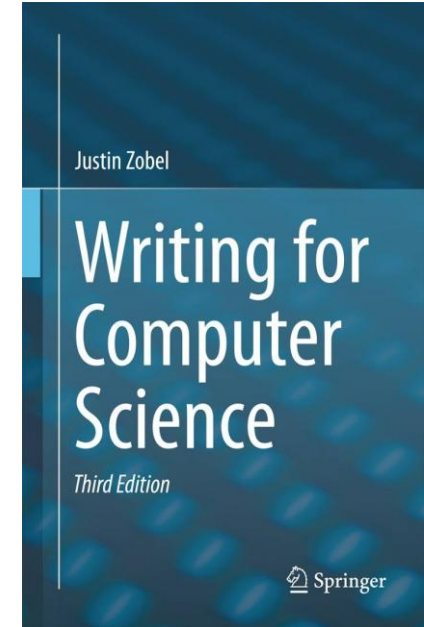
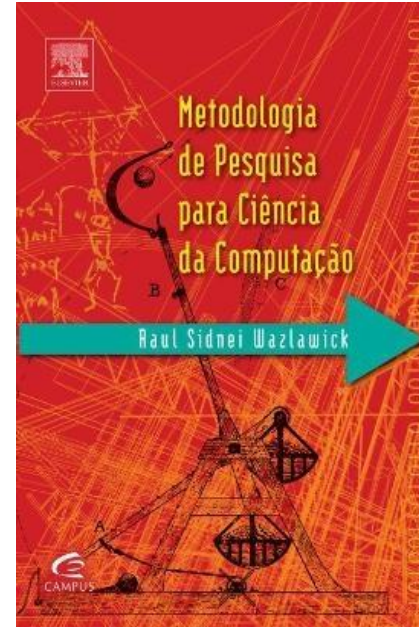
Reprodutibilidade e melhores práticas

- O que melhora a reprodutibilidade?
 - Compartilhamento de código e datasets para facilitar replicação
 - Uso de notebooks interativos (jupyter, rmarkdown) para documentar experimentos
 - Ambientes reprodutíveis (docker, conda, mlflow)
 - Controle de versão (github, gitlab) para rastrear mudanças
 - Automatização de experimentos para minimizar variação manual
- Erros que comprometem a reprodutibilidade:
 - Depender de configurações específicas de hardware que não são replicáveis
 - Não documentar versões de bibliotecas usadas
 - Modificar dados e experimentos sem controle de versão
- Exemplo:
 - Um estudo feito com python 3.8 e tensorflow 2.4 pode gerar resultados diferentes quando testado com python 3.11 e tensorflow 2.10, se não houver controle de versão

[1] c. Stoddart, 2016, is there a reproducibility crisis in science?, *Nature* (may.)

[2] m. Baker, 2016, 1,500 scientists lift the lid on reproducibility, *nature*, v. 533, n. 7604 (may.), P. 452–454.

Referências



- [1] D. G. Perovano, Manual de metodologia da pesquisa científica. Editora Intersaberes, 2016.
- [2] A. L. Cervo, P. A. Bervian, e R. da Silva, Metodologia Científica. Pearson Universidades, 2006.
- [3] R. Wazlawick, 2017, Metodologia de Pesquisa para Ciência da Computação. Elsevier Brasil.
- [4] J. Zobel, 2015, Writing for Computer Science. Springer.

