Visualizando e modelando dados em aquisição da linguagem

Guilherme D. Garcia

Université Laval | gdgarcia.ca

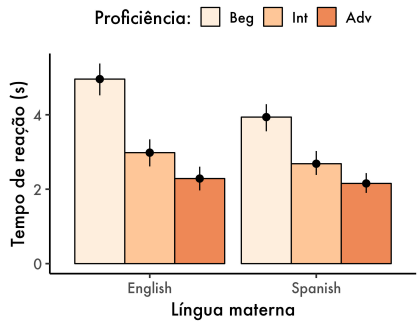
II Seminário Internacional de Pesquisas em Ensino e Aprendizagem de Línguas: abordagens quantitativa e mista

Outubro, 2023

Por que visualizar dados?

Maximizar a compreensão de padrões empíricos

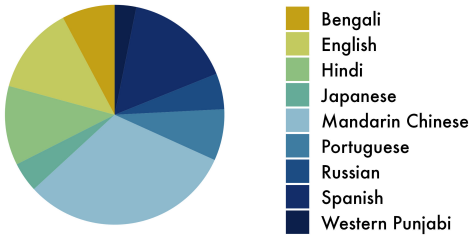
	L1	Proficiency	$\bar{x}$	95% CI
1	English	Beg	4.96	[5.38, 4.53]
2	English	Int	2.98	[3.36, 2.60]
3	English	Adv	2.29	[2.60, 1.98]
4	Spanish	Beg	3.94	[4.30, 3.57]
5	Spanish	Int	2.68	[3.02, 2.35]
6	Spanish	Adv	2.16	[2.42, 1.89]

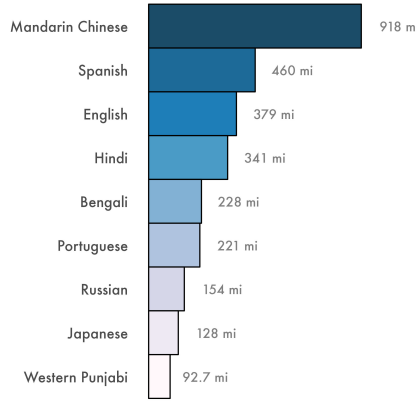


- Dados de Garcia (no prelo): doi.org/10.31219/osf.io/8r4ec

Por que visualizar dados?

Maximizar a compreensão de padrões empíricos





Por que visualizar dados?

Pré-submissão

- Exploração de padrões
- Ajuste de método

Publicação

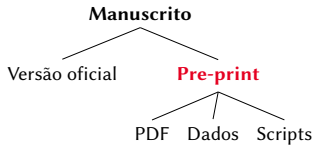
- Clareza metodológica
- Eficiência comunicativa

☞ Estudos de aquisição: **muitos dados + variação** → uma razão específica importante

Antes de começarmos

Uma estrutura recomendada

- A noção de Open science: acesso livre a estudos e publicações



- Repositórios como OSF são altamente recomendados
- Acesso livre, SEO, controle sobre formatação e atualizações, etc.

☞ Algumas revistas (formato físico) tendem a **limitar** tipos de visualização (e.g., cores)

O problema

- Estudos de aquisição subutilizam métodos quantitativos (Plonsky 2011)
- O mesmo pode ser dito sobre **exploração visual** de dados

☞ Dados não explorados costumam ser dados pouco compreendidos
Análises inapropriadas, resultados pouco confiáveis

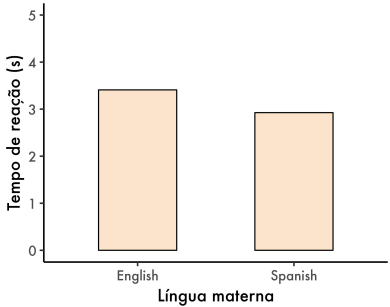
Agravante: modelos complexos são facilmente executados hoje em dia

DADOS CONTÍNUOS

Exemplo

Padrões gerais

- Dois grupos de aprendizes de francês: línguas maternas inglês e espanhol
- Dados: tarefa de escolha forçada → acurácia (0/1), tempo de reação (s) e certeza (1–4)



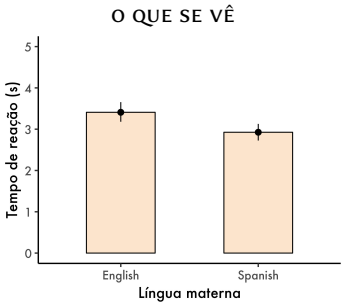
- Pouca informação ($\bar{x}$): duas variáveis
- Ausência de barras de erro

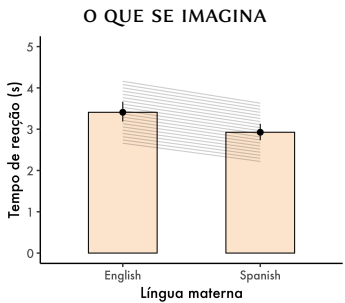
☞ Omissão de **variação**

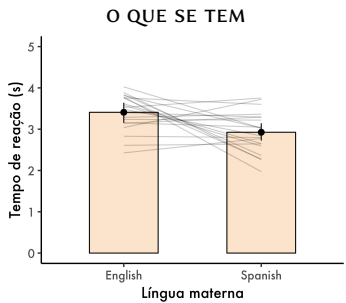
Exploração visual

Visualizando variação: itens

- Comportamento de aprendizes: **raramente** constante para todos os **itens** de um estudo





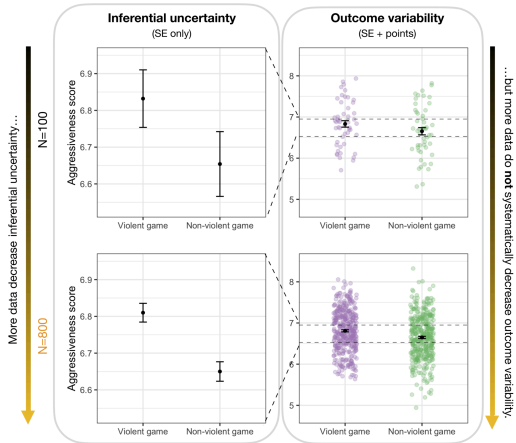


Exploração visual

A ilusão da magnitude de efeitos

- Incerteza inferencial vs variabilidade
- Percepções **bastante** distintas

☞ *n* pode não resolver o problema

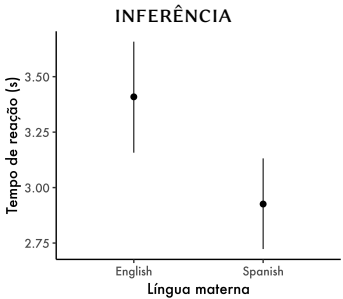


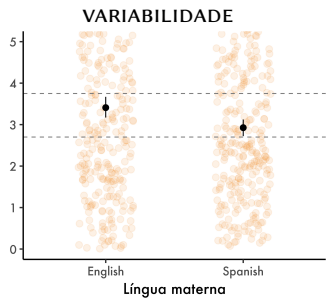
(Zhang et al. 2023, p. 2)

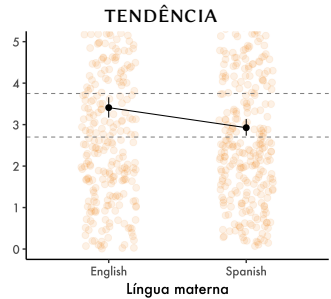
Do geral ao específico

A ilusão da magnitude de efeitos

- Gráficos afetam nossa conclusão sobre o tamanho do efeito de L1

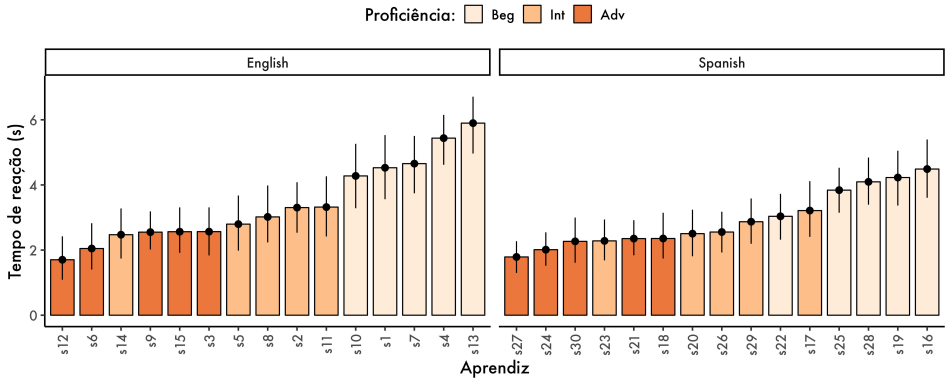






Exploração visual

Visualizando variação: aprendizes

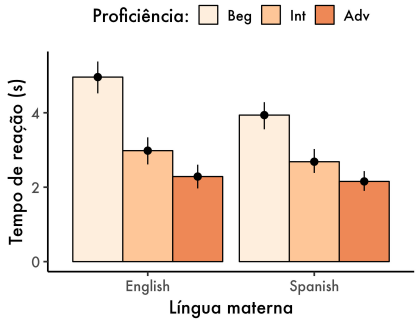


- Variação em tempo de reação: aprendizes, língua materna, e níveis de proficiência

Exploração visual

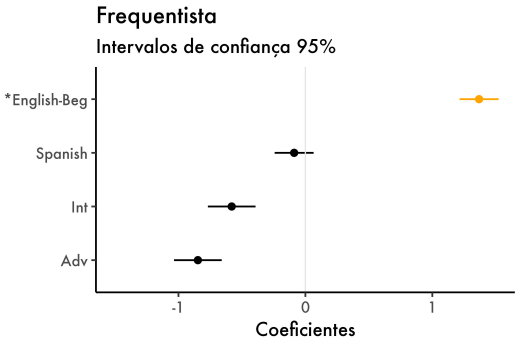
- Tanto língua materna quanto proficiência são relevantes em nossa exploração

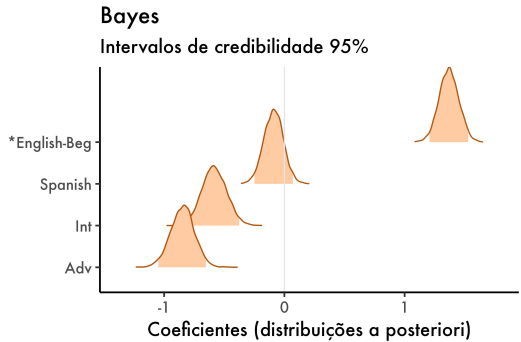
☞ Variável principal: **proficiência** (L1 *provavelmente* não terá um efeito substancial)



Análise

Visualização de modelos estatísticos





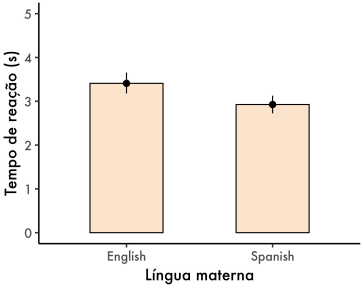
Análise

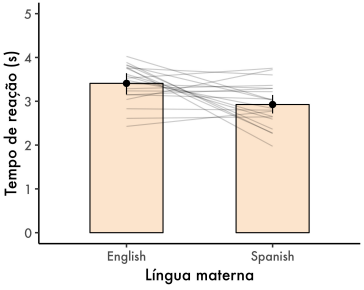
Visualização de modelos estatísticos

- Nem toda variação precisa ser incluída em um modelo: aqui, **fit singular**
- Complexidade excessiva quando consideramos variáveis principais + efeitos aleatórios

☞ Aqui, **L1** + **Proficiência** ≻ **variabilidade de item e aprendiz**

Dúvida: qual das duas figuras será mais apropriada?





DADOS ORDINAIS

Exploração visual

Escalas

- Escalas são frequentemente usadas em estudos de aquisição

Qual o seu grau de certeza sobre sua resposta?

1	2	3	4

☞ Como variáveis binárias, variáveis ordinais muitas vezes exigem **transformações**

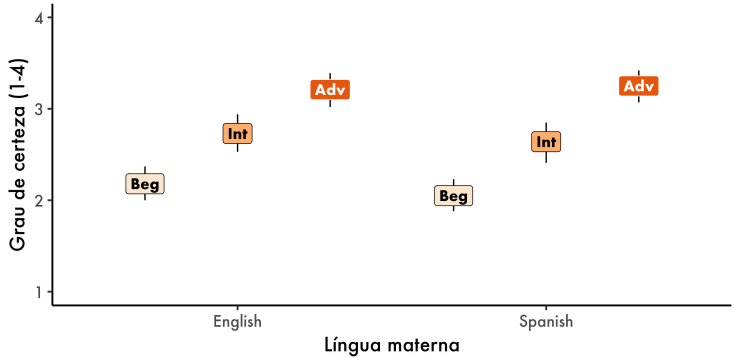
Como visualizar escalas...?

☞ Discussão baseada em Garcia (2021, caps. 5 e 8) e Garcia (no prelo)

Exploração visual

Escalas

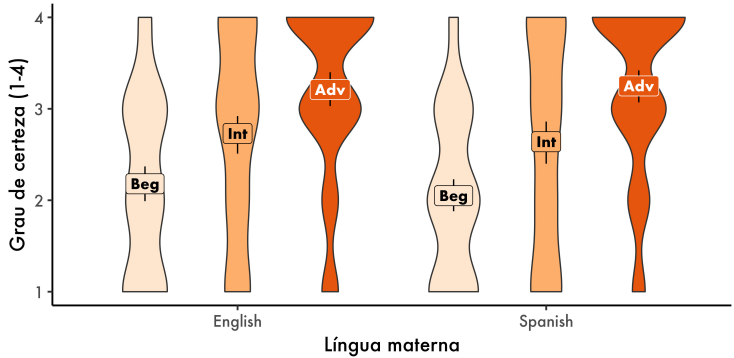
- Dados ordinais não são contínuos (`ordered factor`)
- Distribuição raramente normal → médias são pouco representativas



Exploração visual

Escalas

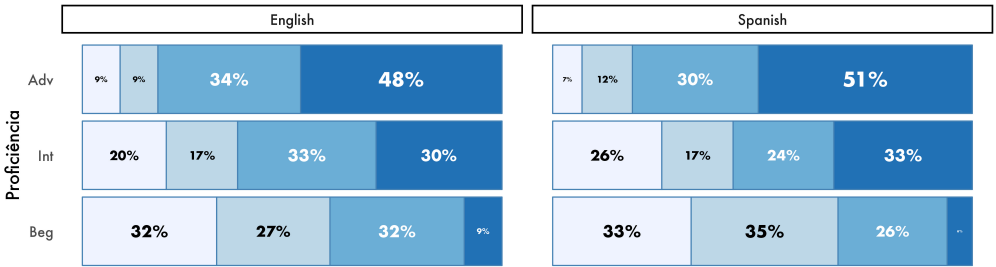
- Dados ordinais não são contínuos (`ordered factor`)
- Distribuição **raramente normal** → médias são pouco representativas



Exploração visual

Escalas

- Visualização **sem** transformação → melhor alinhamento com modelos ordinais
- Barras e cores que espelham escala original (adaptado de Garcia 2021, p. 100)



Escala de certeza: 1–4

Exploração visual

Escalas

- Escala de cinza para publicações físicas (note que, mais uma vez, proficiência ≻ L1)
- Fácil adaptação com diferentes paletas no `ggplot2`

Escala de certeza: 1–4

Exploração visual

Escalas

Preparação de dados

1. Agrupar variáveis relevantes
2. Contar *n* para cada ponto da escala
3. Calcular porcentagens

code

```
1 prop = viz ▷
2   summarize(n = n(),
3                 .by = c(L1, Proficiency, Certainty)) ▷
4   mutate(Prop = n / sum(n),
5                 .by = c(L1, Proficiency),
6          Dark = if_else(Certainty %in% c("3", "4"),
7          "yes", "no"))
```

☞ Utilização de *dummy variables* para maior personalização em gráficos (linhas 6–7)

PREVISIBILIDADE DE DADOS

Qual o propósito de uma análise?

1. Examinar o papel de variáveis dentro de um estudo com bases teóricas específicas
2. Gerar o melhor modelo possível para prever novos dados (e.g., *machine learning*)
3. ...

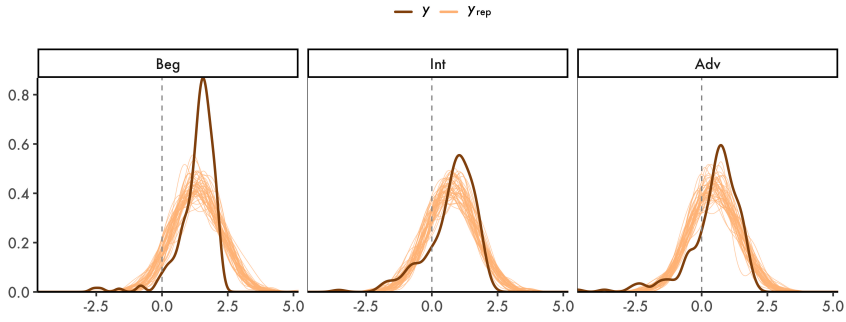
☞ Normalmente, o desejo de previsibilidade está relacionado ao objetivo 2

- Mas visualizar previsões de nossos modelos pode ser bastante informativo

Qual o propósito de uma análise?

Posterior predictive check

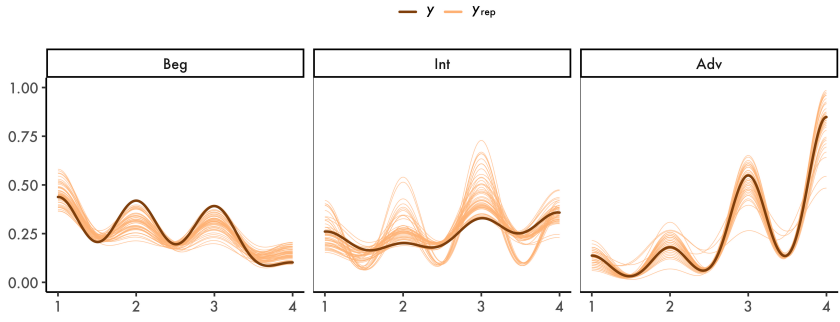
- Comparação entre dados reais e dados simulados a partir do modelo
- Modelo que examina **tempos de reação** (log) em função de L1 e proficiência:



Qual o propósito de uma análise?

Posterior predictive check

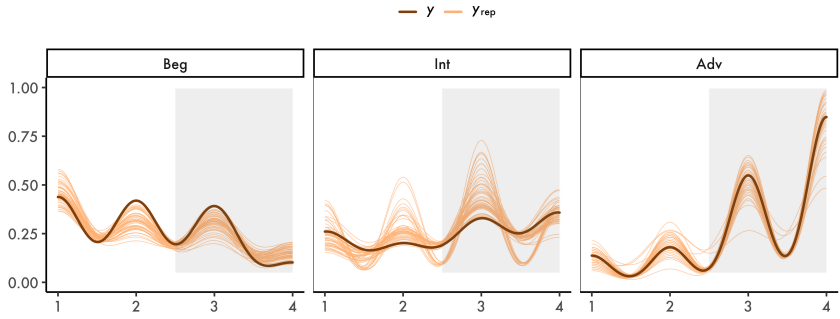
- Comparação entre dados reais e dados simulados a partir do modelo
- Modelo que examina **certeza** em função de L1 e proficiência:



Qual o propósito de uma análise?

Posterior predictive check

- Comparação entre dados reais e dados simulados a partir do modelo
- Modelo que examina **certeza** em função de L1 e proficiência:



Considerações e sugestões finais

- Visualizar dados antes, durante, e depois da análise
- Considerar aspectos estéticos com cuidado: cores, tamanhos, quantidade de informação
- Alinhar visualização com análise e objetivos: maximizar eficiência
- Ter presença online para divulgação de materiais (pre-prints + extras)

Recursos visuais

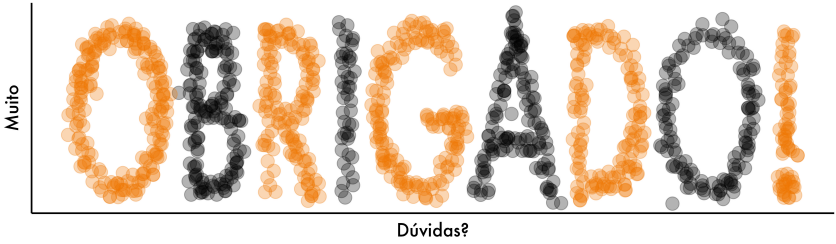
Pacotes em R

- `tidyverse` (`ggplot2`)
- `plotly` (gráficos interativos)
- `MoMAColors`
- `RColorBrewer`

Links e livros

- r4ds.hadley.nz
- gdgarcia.ca + blog
- Garcia (2021, 2023)
- Winter (2019)

☞ Estes slides já estão disponíveis em gdgarcia.ca


OBRIGADO!
Muito
Dúvidas?

References I

Garcia, G. D. (2021). *Data visualization and analysis in second language research.* New York, NY: Routledge.

Garcia, G. D. (2023). Quantitative data visualization. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics.* John Wiley & Sons, Ltd. To appear. Pre-print, data and code available at `https://doi.org/10.31219/osf.io/8r4ec`.

Plonsky, L. (2011). *Study quality in SLA: A cumulative and developmental assessment of designs, analyses, reporting practices, and outcomes in quantitative L2 research.* Ph. D. thesis, Michigan State University.

Winter, B. (2019). *Statistics for linguists: an introduction using R.* New York: Routledge.

Zhang, S., P. R. Heck, M. N. Meyer, C. F. Chabris, D. G. Goldstein, and J. M. Hofman (2023). An illusion of predictability in scientific results: Even experts confuse inferential uncertainty and outcome variability. *Proceedings of the National Academy of Sciences 120*(33), e2302491120.