# Inferência bayesiana de dados fonéticos

Webinário 2022

**Ronaldo Lima Jr.**
ronaldojr@letras.ufc.br
ronaldolimajr.github.io

Universidade Federal do Ceará

# Roteiro

1. Dados Individuais x Agrupados

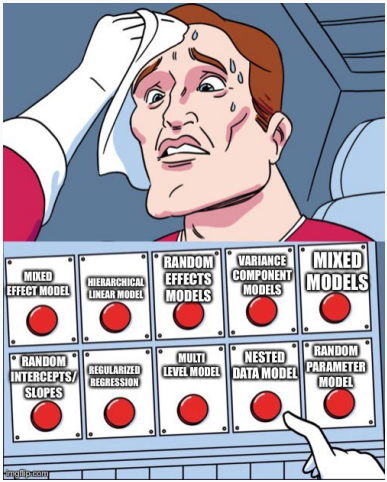
2. Sugestão 1: efeitos mistos/aleatórios

3. Sugestão 2: modelos bayesianos

# Possível caminho de progressão

- Testes de hipótese (valor de $p$)
- Testes de hipótese (com intervalos de confiança, tamanho de efeito, análise de poder)
- Modelos de regressão
- Modelos de regressão com efeitos mistos
- Modelos bayesianos (de efeitos mistos)

# Terminologia

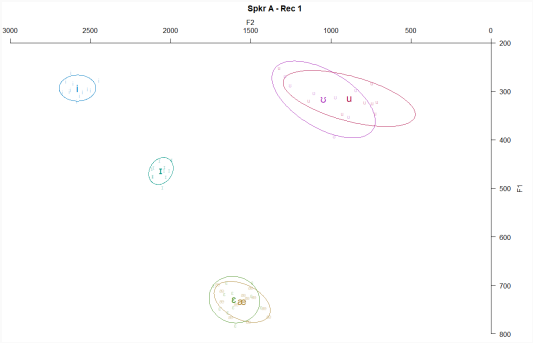


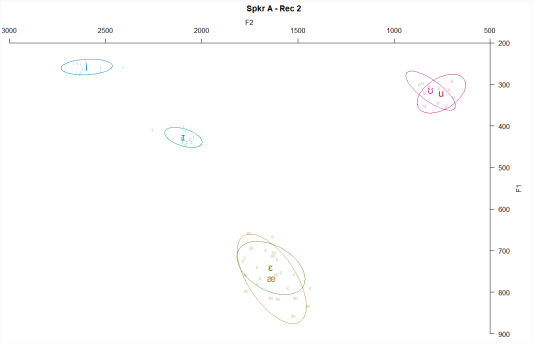

- Modelo de efeitos mistos
- Modelo misto
- Modelo com efeitos aleatórios
- Modelo hierárquico
- Modelo multinível

https://twitter.com/chelseaparlett/status/1458461737431146500?s=21&t=6A3Ftp2BDfdT5U99k5qzBA

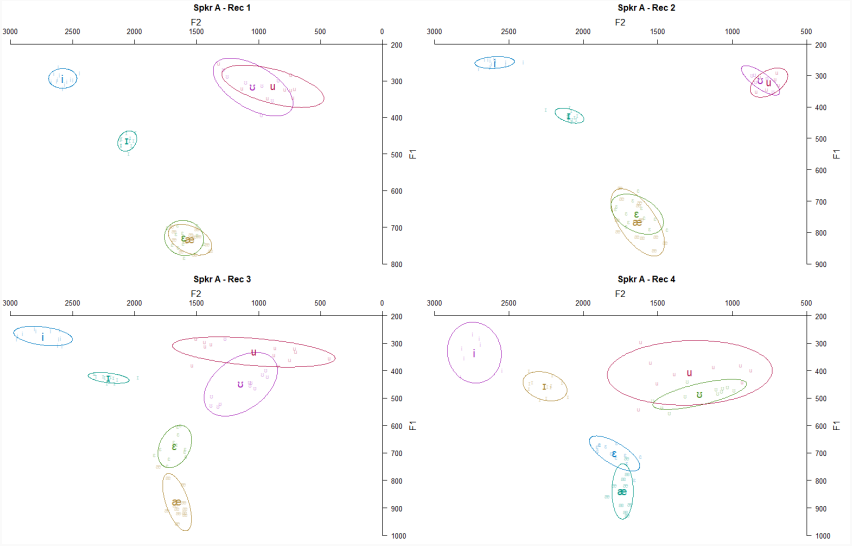
# Dados Individuais x Agrupados

# Dados individuais

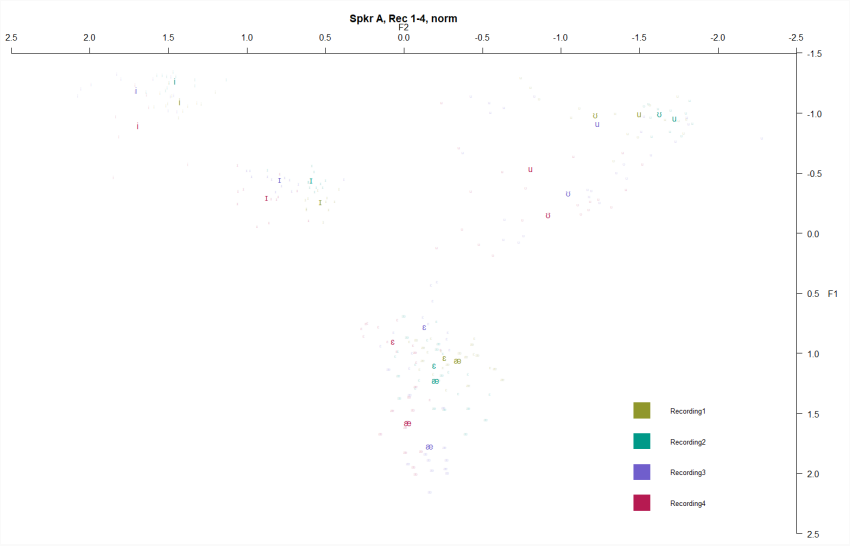


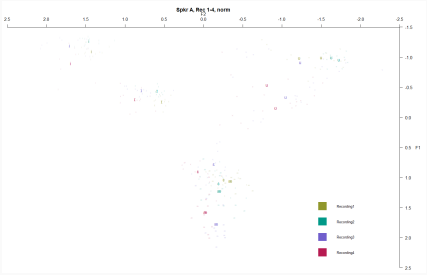



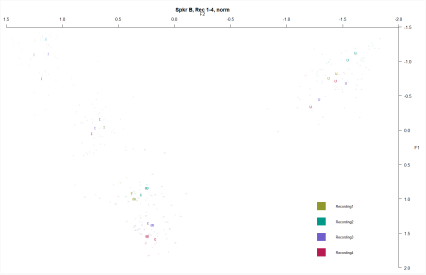
# Dados individuais

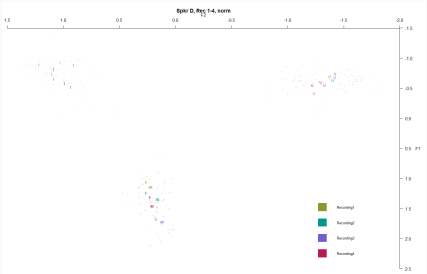


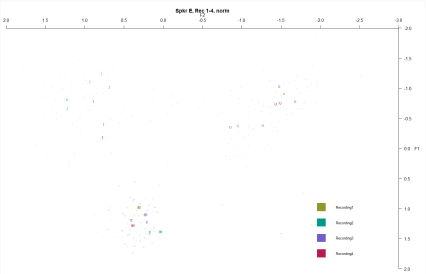
# Dados individuais

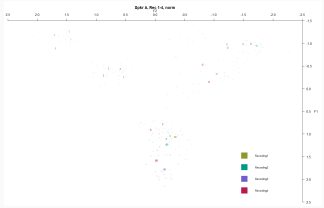


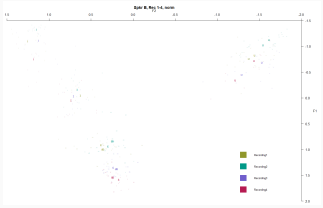

Spkr A, Rec 1-4, norm
F2
F1
Recording1
Recording2
Recording3
Recording4

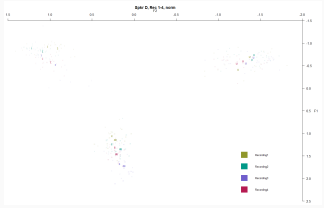

Spkr B, Rec 1-4, norm
F2
F1
Recording1
Recording2
Recording3
Recording4

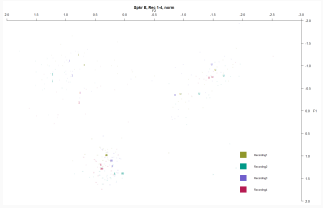

Spkr D, Rec 1-4, norm
F2
F1
Recording1
Recording2
Recording3
Recording4


Spkr E, Rec 1-4, norm
F2
F1
Recording1
Recording2
Recording3
Recording4







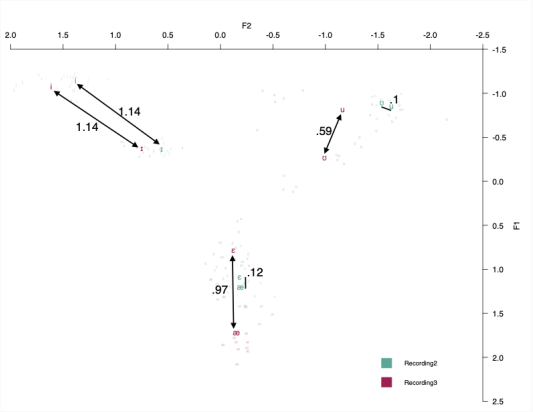


→ 10 falantes!

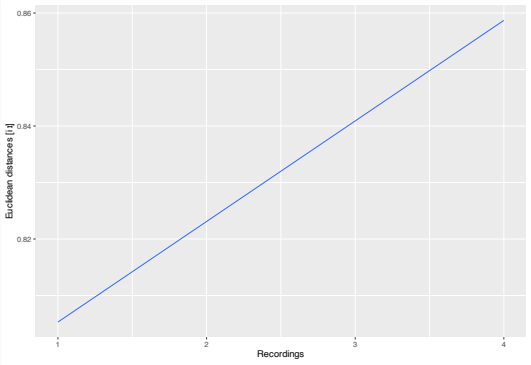
→ Muito difícil de chegar a generalizações (análise qualitativa de dados numéricos)

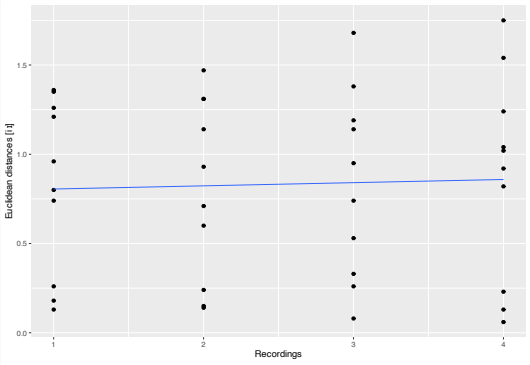
# Dados agrupados

→ Distância (euclidiana) entre as (médias das) vogais de cada par, por falante, em cada gravação



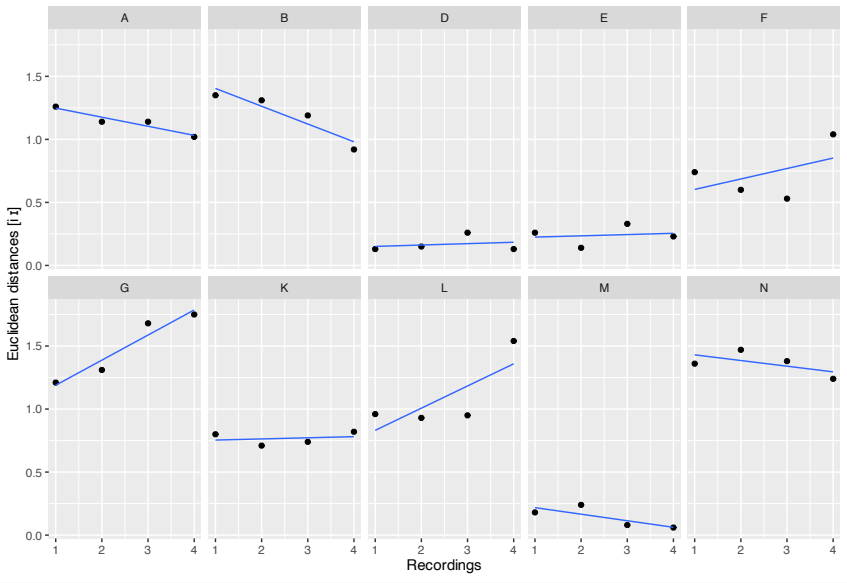
# Dados agrupados de [i ɪ]



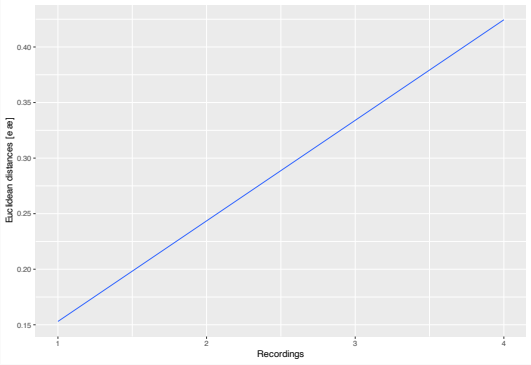



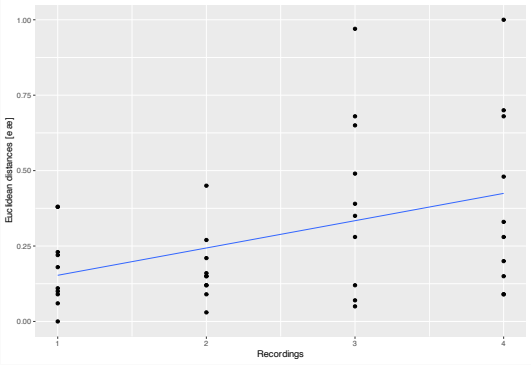

$\rightarrow$ $f(3) = 0.035$; $p = 0.991$

# Porém...


A
B
D
E
F
G
K
L
M
N
Euclidean distances [i,i]
0.0
0.5
1.0
1.5
1
2
3
4
Recordings

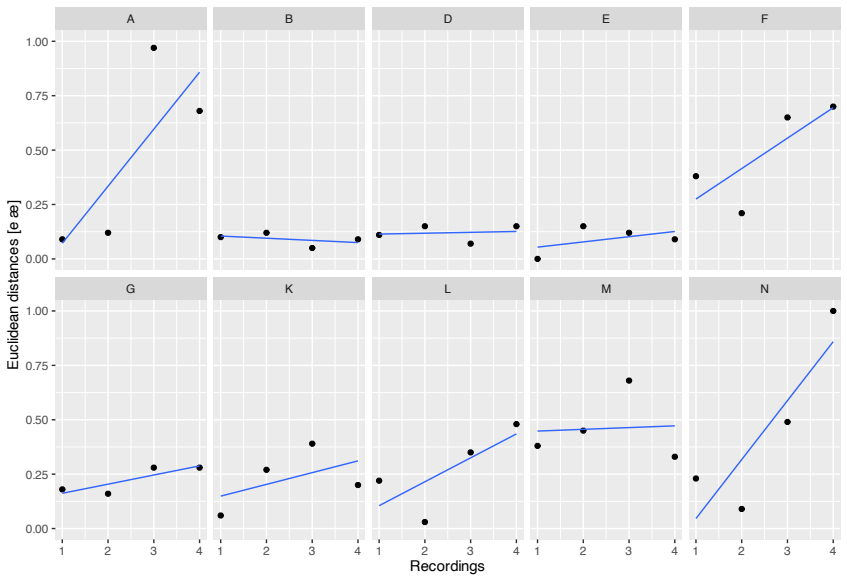
# Dados agrupados de [ɛ ae]



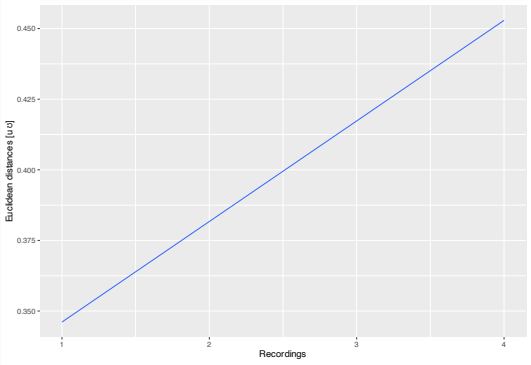



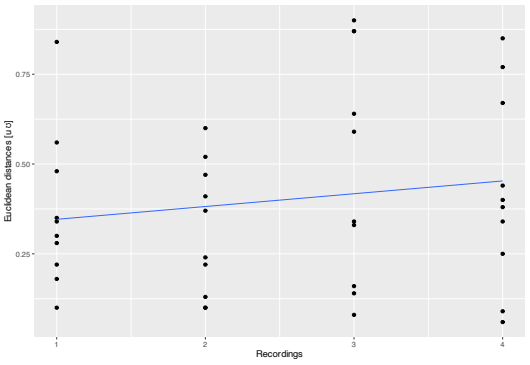

→ $f(3) = 3.234$; $p = 0.0335$, mas sem valores de $p$ significativos em teste pareado post-hoc Tukey HSD

# Porém...


A
B
D
E
F
G
K
L
M
N
Euclidean distances [e æ]
1.00
0.75
0.50
0.25
0.00
1
2
3
4
Recordings

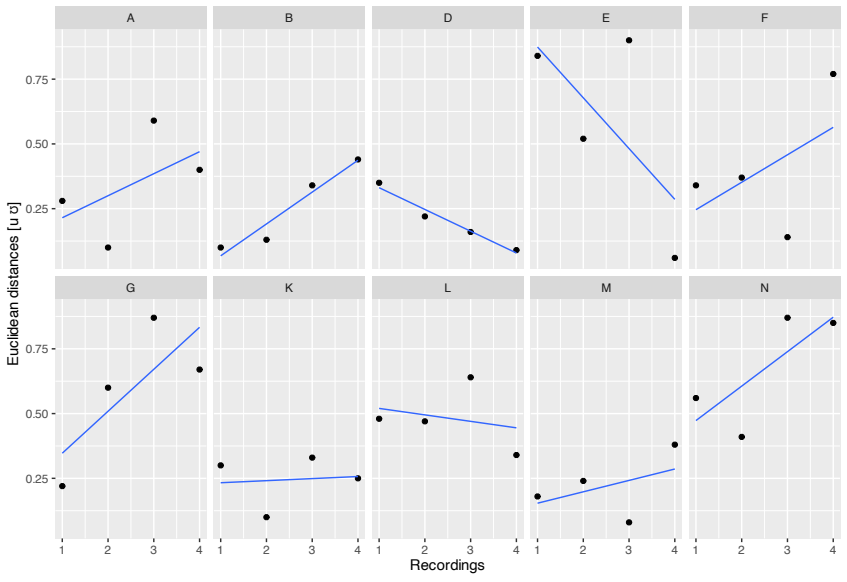
# Dados agrupados de [u ʊ]



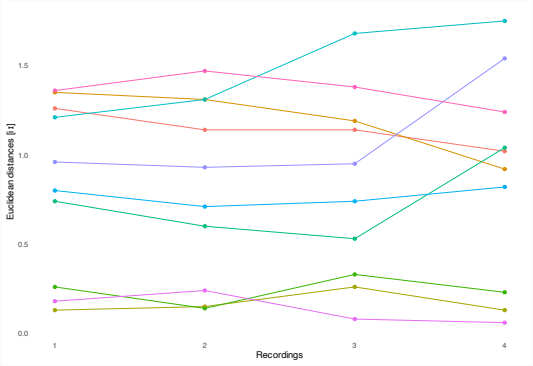

→ $f(3) = 0.907$; $p = 0.447$

# Porém...

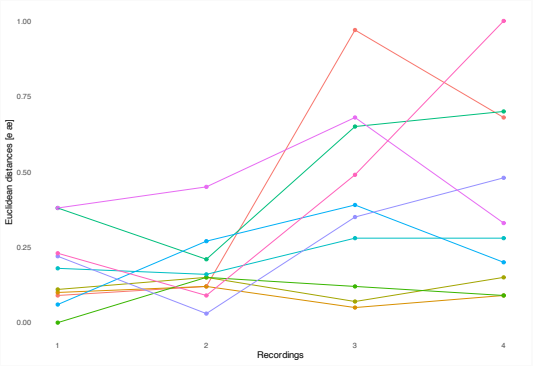

A
B
D
E
F
G
K
L
M
N
Euclidean distances [u.u]
0.75
0.50
0.25
1
2
3
4
Recordings

# Linha reta?

[i ɪ]

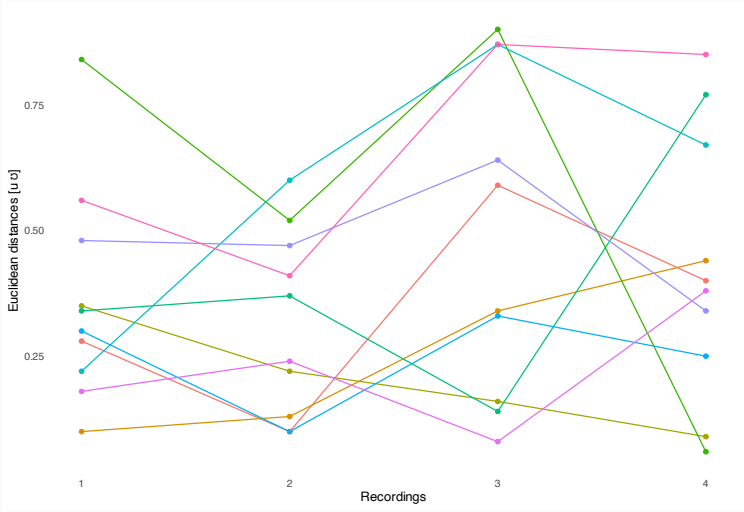



[ɛ ae]



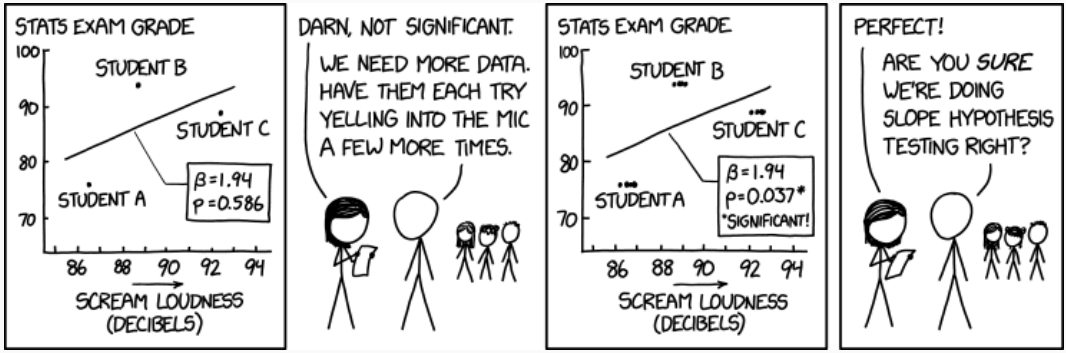
# Linha reta?

[u ʊ]



# Sugestão 1: efeitos mistos/aleatórios

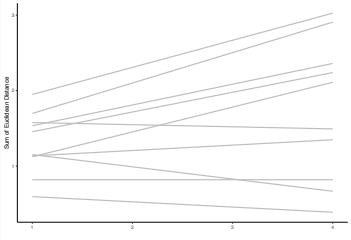
# Efeitos mistos/aleatórios



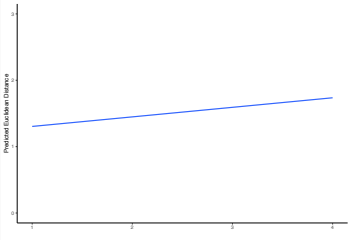

https://xkcd.com/2533/

# Soma das distância euclidianas

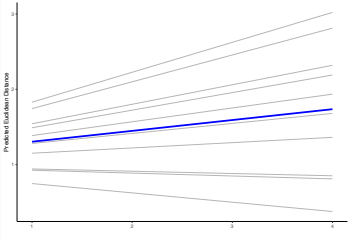
Linhas de tendência individuais (lm)




Modelo linear (sem efeitos mistos)




Modelo com *intercepts* e *slopes* variáveis




→ Adicionar os efeitos variáveis alterou a linha?

Não, mas mudou a confiança do modelo sobre a linha:

# Soma das distância euclidianas

→ Adicionar os efeitos variáveis alterou a linha?

Não, mas mudou a confiança do modelo sobre a linha:

```
> fit 1 = lm(sum ~ recording)
> summary(fit1)

Coefficients:
             Estimate Std. Error t value
(Intercept)    1.1605     0.2767   4.195
recording      0.1439     0.1010   1.424
```

```
> fit2 = lmer(sum ~ recording + (recording|speaker))
> summary(fit2)

Fixed effects:
             Estimate Std. Error       df t value
(Intercept)  1.16050    0.12993 15.49896   8.932
recording    0.14390    0.07027  9.81783   2.048
```

→ Consequentemente:

| Predictors | Estimates | CI | p | Estimates | CI | p |
|---|---|---|---|---|---|---|
| Intercept | 1.16 | 0.60 – 1.72 | <0.001 | 1.16 | 0.90 – 1.42 | <0.001 |
| recording | 0.14 | -0.06 – 0.35 | 0.162 | 0.14 | 0.00 – 0.29 | **0.048** |

# Observações sobre valores de p

- Definição do valor de $p$ (probabilidade dos dados vs probabilidade das hipóteses)
- Problemas/limitações
  - O valor de $p$ não diz nada sobre a hipótese de trabalho (alternativa)
  - O valor de $p$ não diz nada sobre o tamanho do efeito
  - Limiar arbitrário
  - Decisão categórica sobre os dados (porém, comumente utilizada com gradiência – de acordo com a intenção do pesquisador?)
  - É possível obter um valor de $p$ baixo com baixo poder estatístico e/ou com pequeno pequeno tamanho de efeito
  - *p-hacking*
- Modelos com efeitos mistos não geram valores de $p$ (Bates), e diferentes aproximações geram valores diferentes (e.g. `lmerTest` vs `sjPlot`)

# Observações sobre valores de p

```
lmertTest::fit2 = lmer(sum ~ recording + (recording|speaker))
summary(fit2)
```

| Predictors | Estimates | p |
|---|---|---|
| Intercept | 1.16 | <0.001 |
| recording | 0.14 | 0.068 |

1. Satterthwaite's method
2. t-statistics and the normal distribution function
3. conditional F-test with Kenward-Roger approximation

```
sjPlot::tab_model(fit2)
```

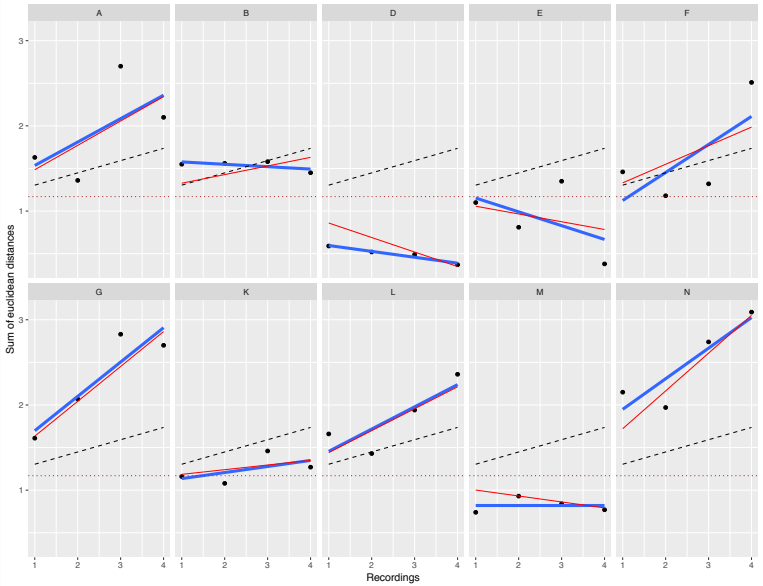
| Predictors | Estimates | p |
|---|---|---|
| Intercept | 1.16 | <0.001 |
| recording | 0.14 | 0.048 |

```
sjPlot::tab_model(fit2, p.val = "kr")
```

| Predictors | Estimates | p |
|---|---|---|
| Intercept | 1.16 | <0.001 |
| recording | 0.14 | 0.071 |

# Intercepts e slopes variáveis

→Modelos de efeitos mistos podem prever valores para participantes individuais



# Sugestão 2: modelos bayesianos

# Por que um modelo bayesiano?

→ Probabilidade dos parâmetros (hipóteses) diante dos dados
(em vez de probabilidade dos dados diante da $H_0$)

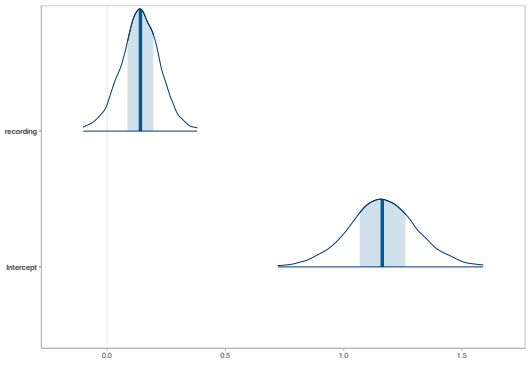
→ Distribuições de probabilidades dos coeficientes
(em vez de *point estimates*)

→ Intervalos de credibilidade
(em vez de intervalos de confiança)

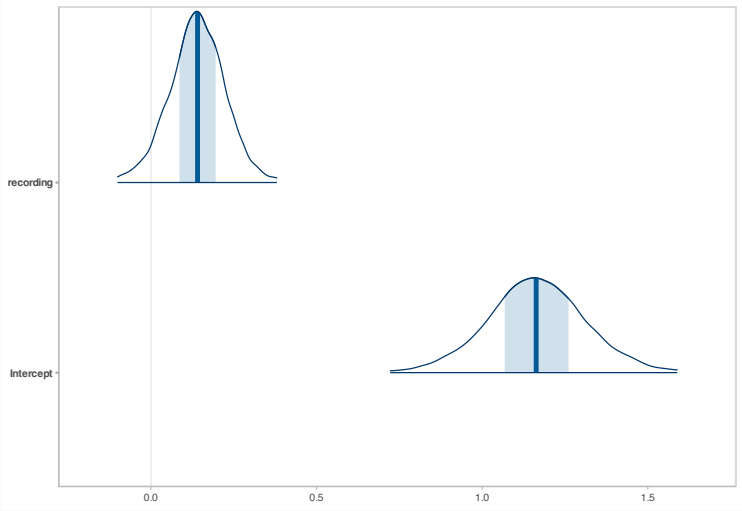
→ Informação/conhecimento prévios no modelo
(em vez de todos os possíveis valores de coeficientes terem a mesma probabilidade *a priori*)

# Modelo bayesiano de efeitos mistos

| Predictors | Estimates | 50% CI | 95% CI |
|---|---|---|---|
| Intercept | 1.16 | 1.07 – 1.26 | 0.86 – 1.47 |
| Recording | 0.14 | 0.09 – 0.20 | -0.04 – 0.31 |



# Modelo bayesiano de efeitos mistos

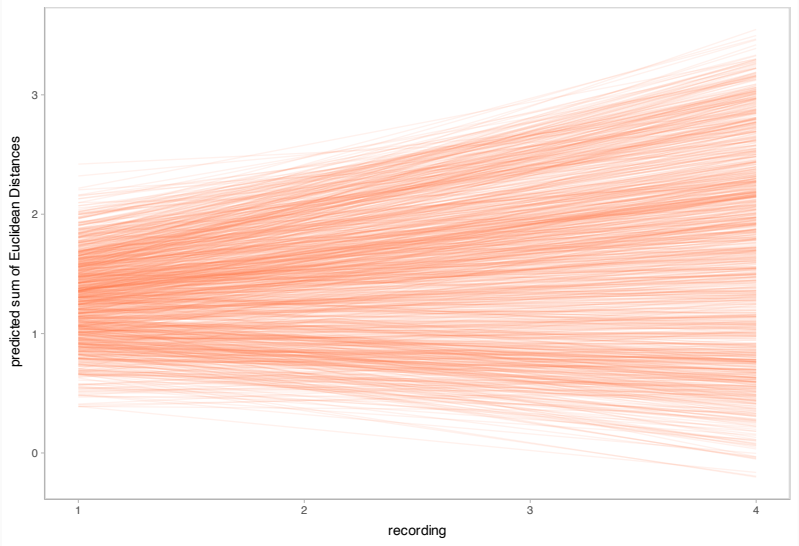



- → 6% da área sob a curva (AUC – area under the curve) abaixo de 0
- → Este tipo de análise adiciona incerteza bem-vinda ao inferir valores (parâmetros) da população com base em uma amostra limitada

# Mais exemplos

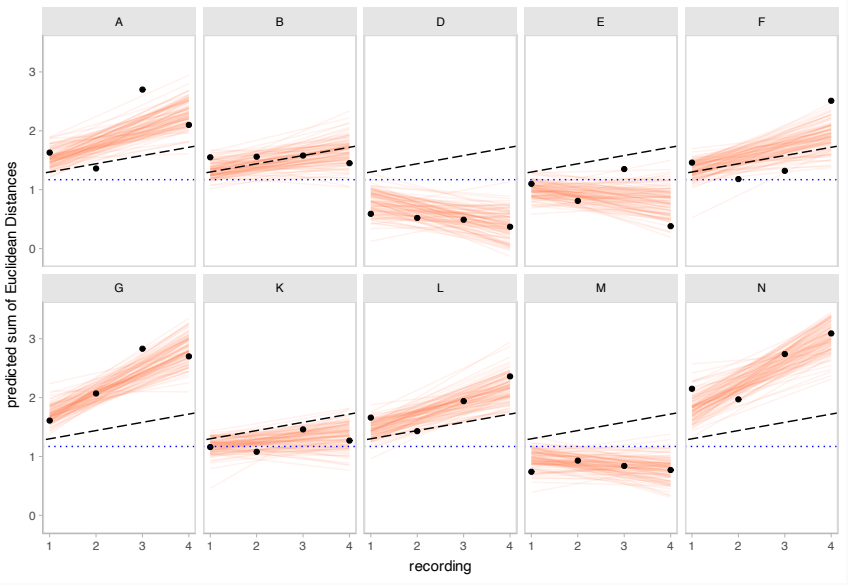
- Alerrandro - Felipe - Pablo - Many Speech Analysis

# Modelo bayesiano de efeitos mistos: valores previstos



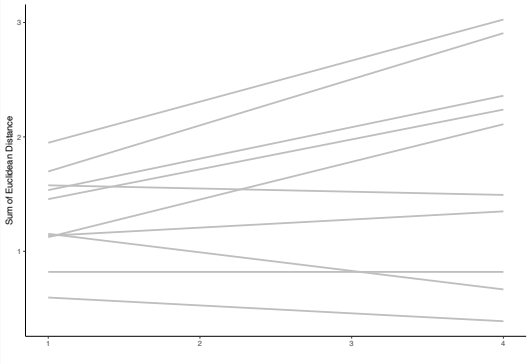
# Modelo bayesiano de efeitos mistos

→Várias linhas prováveis (neste caso, 100) previstas pelo modelo amostradas da distribuição a posteriori (em vez de uma única linha)

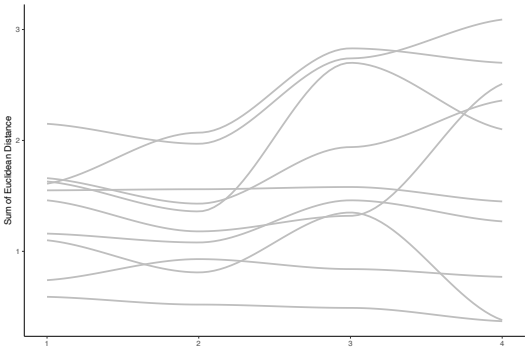


# Não precisam ser linhas retas

```
1 | geom_smooth(method = lm)
```




```
1 | geom_smooth(method = loess)
```



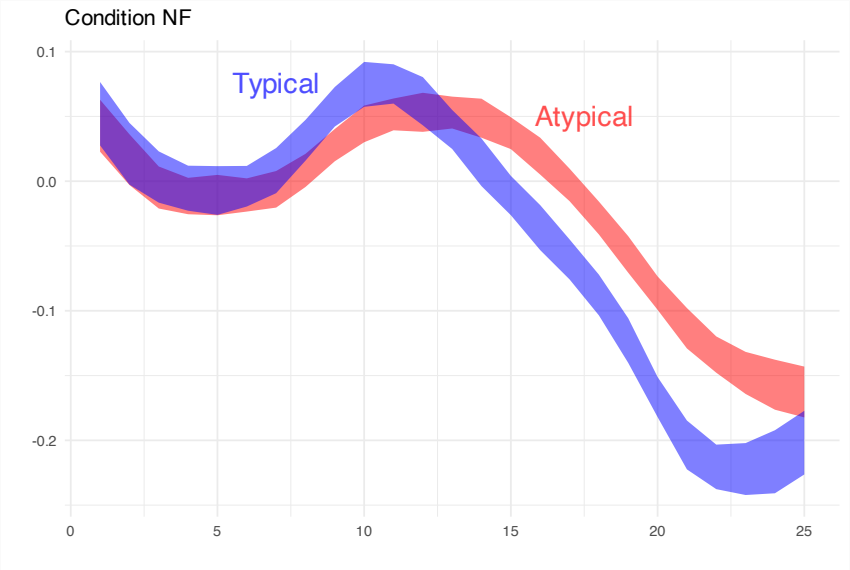
# Não precisam ser linhas retas

- "linear" em matemática não significa ter uma relação 1:1, nem ser uma linha reta

→ Significa adição de termos

- Há modelos de regressão (linear) que preveem curvas ao somar termos específicos à fórmula de regressão. E.g.:
  - Regressões polinomiais (quadrática, cúbica, etc.)
  - *Splines*
  - Modelos aditivos generalizados (GAMs – Generalized Additive Models)

# Não precisam ser linhas retas



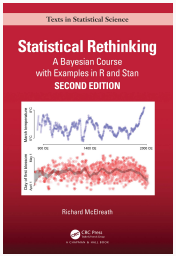
# Finalizando

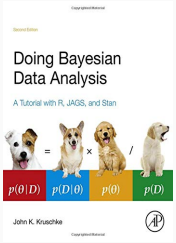
- Desvantagens de modelos bayesianos:
  - Curva de aprendizagem
  - Demanda computacional

- Onde aprender?
  - McElreath, R. (2020). Statistical rethinking: A Bayesian course with examples in R and Stan. Chapman and Hall/CRC.

  → `https://youtube.com/playlist?list=PLDcUM9US4XdMROZ57-OIRtIK0aOynbgZN`

  - Kruschke, J. (2014). Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.



- Lima Jr, R. M., & Garcia, G. D. (2021). Diferentes análises estatísticas podem levar a conclusões categoricamente distintas. Revista Da ABRALIN, 20(1), 1-19. https://doi.org/10.25189/rabralin.v20i1.1790

- Garcia, G. D., & Lima Jr, R. M. (2021). Introdução à estatística bayesiana aplicada à linguística. Revista Da ABRALIN, 20(2), 1-24. https://doi.org/10.25189/rabralin.v20i2.1914

# Perguntas?