REVISTA DA **ABRALIN**

TUTORIAL

# Diferentes análises estatísticas podem levar a conclusões categoricamente distintas

Ronaldo Mangueira LIMA JR
Universidade Federal do Ceará (UFC)

Guilherme Duarte GARCIA
Ball State University (BSU)

OPEN ACCESS

EDITADO POR
- Miguel Oliveira Jr. (UFAL)
- Oliver Niebuhr (SDU)

AVALIADO POR
- Pablo Arantes (UFSCar)
- Livia Oushiro (UNICAMP)
- Ricardo A. de Souza (UFMG)

SOBRE OS AUTORES
- Ronaldo Mangueira Lima Jr. Conceptualização, Curadoria de Dados, Análise Formal, Metodologia, Administração do Projeto, Software, Escrita – rascunho original.
- Guilherme Duarte Garcia Conceptualização, Curadoria de Dados, Análise Formal, Metodologia, Software, Validação, Visualização, Escrita – análise e edição.

DATAS
- Recebido: 28/12/2020
- Aceito: 23/07/2021
- Publicado: 05/08/2021

COMO CITAR
Lima Jr., R. M.; Garcia, G. D. (2021). Diferentes análises estatísticas podem levar a conclusões categoricamente distintas. *Revista da Abralin*, v. 20, n. 1, p. 1-19, 2021.

RESUMO

Neste estudo, demonstramos como significância estatística pode variar a partir da comparação de quatro métodos distintos: teste *t*, ANOVA (seguida de Tukey HSD), modelo linear simples, e modelo linear de efeitos mistos. Em nossa demonstração, modelamos tempos de reação em função de diferentes afixos em dinamarquês, e mostramos como nossas conclusões a respeito do efeito de certos afixos podem mudar categoricamente dependendo de qual dos métodos mencionados acima decidimos utilizar. Por fim, reiteramos o que dizem estudos recentes (e.g., BARR *et al.*, 2013), e sugerimos que modelos de efeitos mistos devam ser a norma sempre que dados agrupados forem analisados. Esperamos, com este estudo, alertar pesquisadores da área para a importância de decisões analíticas bem informadas e éticas em estudos linguísticos.

ABSTRACT

In this study, we illustrate the potential variability of statistical significance by comparing four different methods, namely, *t*-test, ANOVA (followed by Tukey HSD), simple linear regression, and mixed effects linear regression. In our demonstration, we model reaction times as a function of different affixes in Danish, and show how our conclusions regarding the effect of certain affixes can change categorically depending on which of the aforementioned methods we choose to use. Finally, we echo recent

studies (e.g., BARR *et al.*, 2013), and suggest that mixed effects models be the norm whenever grouped data is analyzed. With our comparison, we hope to raise researchers' awareness to the need for well-informed and ethical analytical decisions in linguistic studies.

PALAVRAS-CHAVE
Análise quantitativa de dados. Modelos de regressão. Testes estatísticos. Significância estatística.

KEYWORDS
Quantitative data analysis. Regression models. Statistical tests. Statistical significance.

# Introdução

Neste artigo, demonstramos brevemente como análises estatísticas mais robustas, como modelos de regressão, podem gerar resultados categoricamente distintos daqueles gerados por análises mais básicas, como testes *t*. Para tanto, empregamos os dados de Balling e Baayen (2008), que demonstram efeitos da morfologia no reconhecimento auditivo de palavras do dinamarquês. Os dados estão livremente disponibilizados por meio do pacote *languageR* (BAAYEN, 2007a) do programa R (R CORE TEAM, 2020). Nosso principal objetivo é argumentar a favor de modelos mistos e contra análises mais simples (como testes *t* e ANOVAs), uma vez que esses modelos são mais robustos (e.g., JAEGER, 2008), possuem maior poder explicativo e, ao mesmo tempo, levam em conta a variabilidade de dados linguísticos.

O R é uma plataforma gratuita, livre, e de código aberto, voltada principalmente para a visualização e análise de dados. É possivelmente o programa de análise estatística mais utilizado atualmente nas ciências (CHAMBERS, 2020), por ser rápido e poderoso. Seu código aberto permite aos usuários criarem pacotes com funções customizadas às suas análises (atualmente, há mais de 16.000 pacotes disponíveis). Um exemplo de pacote é justamente o *languageR*, que agrupa bancos de dados e funções que acompanham o livro de introdução a análise de dados linguísticos de Baayen (2007b). Há 51 conjuntos de dados no pacote, entre eles *danish*, que contém 3.326 observações de tempos de reação de decisões lexicais auditivas de palavras complexas do dinamarquês, e que será utilizado para ilustrar as análises deste artigo.

R também se refere à linguagem de programação para computação estatística utilizada no programa homônimo. Indicaremos ao longo do texto as linhas de comando exatas que utilizamos, e o script completo das análises apresentadas está disponível no seguinte repositório OSF (*Open Science Framework*): https://osf.io/awqpu/. Este texto não se propõe, contudo, a ser um tutorial do uso do

*R*, então a explicitação das linhas de comando utilizadas será útil para aqueles minimamente familiarizados com a linguagem. Os leitores ainda não familiarizados com o *R* poderão ignorar esses trechos e, mesmo assim, se beneficiar do conteúdo e das discussões propostas. Para uma introdução ao R, assim como visualização de dados e modelos de regressão, ver Garcia (2021).[1]

Enfatizamos logo no início deste texto que faremos múltiplas análises com os mesmos dados com um propósito exclusivamente didático. Mostraremos que cada análise pode gerar uma interpretação diferente dos efeitos encontrados nos dados. Naturalmente, não defendemos que sejam realizadas diversas análises com o mesmo conjunto de dados até que se alcance o resultado desejado. Pelo contrário, argumentaremos que é necessário que pesquisadores em linguística se capacitem em análises quantitativas a fim de selecionarem o modelo de sua análise na fase de planejamento de seu estudo, de maneira criteriosa e bem informada.

Manipular os dados e/ou as análises está entre as práticas que caracterizam a má conduta científica que ficou conhecida como *p-hacking* (e.g., NUZZO, 2014). Além disso, cada rodada de análise com os mesmos dados aumenta a chance de erro do tipo I, ou seja, a rejeição da hipótese nula quando esta é verdadeira. Em outras palavras, cometemos erro do tipo I quando concluímos que um efeito é significativo embora este tenha se dado ao acaso.

Demonstraremos as diferentes análises seguindo a sequência que acreditamos ser aquela de uma capacitação natural de cientistas em relação a análise quantitativa de dados, iniciando com a estatística descritiva dos dados (seção 1), seguida de um teste de hipótese (seção 2), uma ANOVA com comparações múltiplas *post-hoc* (seção 3), um modelo de regressão linear (seção 4), e um modelo de efeitos mistos (seção 5). Cada uma dessas análises é superior à anterior em termos de poder explicativo. Por exemplo, uma análise descritiva dos dados é melhor do que uma mera descrição qualitativa dos dados quando há dados quantitativos envolvidos. Testes estatísticos de hipótese, por sua vez, adicionam uma camada de informação ao avaliarem a probabilidade de se observar determinada distribuição dos dados coletados em caso de não haver um real efeito de correlação entre variáveis. No entanto, devido às diversas limitações desses testes, que serão expostas abaixo, discutiremos por que análises por meio de modelos de regressão, em especial os de efeitos mistos, são superiores a testes de hipótese.

# 1. Estatística descritiva

O primeiro passo comumente dado por pesquisadores em direção a uma capacitação em análise quantitativa de dados é a compreensão de elementos de estatística descritiva, como medidas de tendência central (média, mediana e moda) e de dispersão (como desvio padrão e erro padrão). Essa etapa é comumente chamada de Análise Exploratória de Dados (AED, ou EDA - *Exploratory Data Analysis*), principalmente se acompanhada de inspeções visuais dos dados, e deve estar entre os

[1] Informações sobre o livro podem ser consultadas em https://guilhermegarcia.github.io/dvaslr.

primeiros passos de um/a pesquisador/a independentemente da complexidade dos testes ou modelos que irá utilizar em sua análise inferencial. Essa etapa foi fortemente defendida por John Tukey (matemático americano) justamente por permitir que pesquisadores percebam tendências de suas variáveis e identifiquem erros de medição, por exemplo.

O conjunto de dados *danish* do pacote *languageR* contém a variável de resposta LogRT, referente ao tempo de reação (latência) em escala logarítmica[2] de 3.326 observações do tempo de reação de 22 participantes a 156 palavras complexas do dinamarquês, e diversas variáveis preditoras, como sexo do participante, frequência das palavras, erro ou acerto na tentativa anterior, afixos das palavras complexas, frequência dos afixos, entre outras. Para ilustrar uma AED e as análises estatísticas, focaremos na variável de resposta 'tempo de reação' (LogRT) e no preditor 'afixo' (Affix). Há 16 afixos diferentes utilizados na coleta de dados, mas, para simplificar nossa exemplificação, utilizaremos apenas cinco, selecionados na linha 8 do quadro a seguir, a saber, 'bar', 'ende', 'ede', 'ere' e 'lig'. A questão de pesquisa hipotética, então, é se esses afixos têm efeito sobre o tempo de reação na identificação das palavras.

```
1| library(tidyverse)
2| library(languageR)

3| data(danish)
4| dan = danish %>%
5|   select(Subject, Word, Affix, LogRT) %>%
6|   as_tibble()

7| dan = dan %>%
8|   filter(Affix %in% c("bar", "ende", "ede", "ere", "lig")) %>% droplevels()
9| dan

 Subject Word      Affix   LogRT
 <fct>   <fct>     <fct>   <dbl>
 2s14    appetitlig  lig     6.45
 2s17    appetitlig  lig     6.84
 2s15    appetitlig  lig     6.84
 2s04    appetitlig  lig     6.83
 ...

10| ggplot(data = dan, aes(x = fct_reorder(Affix, LogRT, .fun = mean),
    y = LogRT)) +
11|   geom_boxplot() +
12|   stat_summary(shape = 17) +
13|   theme_bw()
```

QUADRO 1 - Linhas de comando para carregar pacotes, carregar e filtrar os dados a serem analisados, e gerar um gráfico de caixas.
Fonte: elaborado pelos autores.

[2] Tempos de reação raramente seguem uma distribuição normal, por nenhum valor ser zero ou inferior a zero. A transformação logarítmica em questão aproxima essa distribuição à normalidade, reduzindo a assimetria das caudas.

As duas primeiras linhas carregam os pacotes necessários para essa etapa, e as linhas de 3 a 8 carregam e filtram os dados para que contenham apenas as variáveis que serão exploradas. Ao rodar a linha 9, nossa variável de dados, visualizamos as 10 primeiras linhas do conjunto de dados para uma breve conferência se o carregamento funcionou como esperado. As linhas 10 a 13 geram a figura 1, que apresenta os gráficos de caixa (*boxplots*) do tempo de reação em escala logarítmica no eixo $y$ em função de cada um dos cinco afixos no eixo $x$. Os triângulos dentro de cada caixa representam a média do tempo de reação para aquele afixo, uma vez que a linha central de cada caixa representa a mediana.

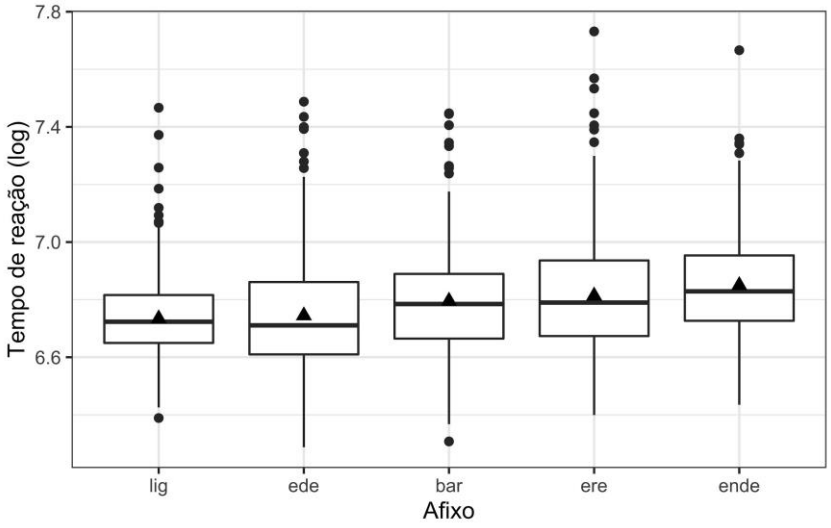


FIGURA 1 - Gráficos de caixa do tempo de reação em função dos cinco afixos selecionados
Fonte: elaborado pelos autores.

Podemos ver que a média no tempo de reação cresce na ordem dos afixos apresentados na figura 1, ou seja: 'lig' < 'ede' < 'bar' < 'ere'< 'ende'. O quadro 2 apresenta as linhas de comando utilizadas para gerar a média, o desvio-padrão e o erro padrão do tempo de reação por afixo, bem como o output desse comando. Vemos, portanto, que a média do tempo de reação de fato aumenta para cada afixo, conforme visualizado no gráfico, e que essa média pode variar de 6,73 em 'lig' a 6,85 em 'ende' (em escala logarítmica).

```
14| dan %>%
15|   group_by(Affix) %>%
16|   summarize(meanRT = mean(LogRT),
17|             sdRT = sd(LogRT),
18|             seRT = sdRT / sqrt(n())) %>%
19|   arrange(meanRT)

Affix   meanRT  sdRT   seRT
<fct>   <dbl>    <dbl>  <dbl>
lig      6.73     0.161  0.0110
ede      6.74     0.217  0.0148
bar      6.80     0.202  0.0140
ere      6.81     0.215  0.0147
ende     6.85     0.186  0.0135
```

QUADRO 2 - Linhas de comando para gerar a média, o desvio padrão e o erro padrão do tempo de reação por afixo, e o output do comando.
Fonte: elaborado pelos autores.

Para saber os valores em milissegundos, basta exponenciar o valor em logaritmo, o que no *R* pode ser feito com o comando `exp()`. Sendo assim, o 6,73 de 'lig' se refere a 837 milissegundos, e o 6,85 de 'ende' se refere a 944 milissegundos, perfazendo uma diferença de 107 milissegundos – diferença considerável em estudos psicolinguísticos que medem o tempo de reação a ponto de motivar uma investigação sobre o tamanho dessa diferença por meio de estatística inferencial.

Apesar de se tratar de números, uma exposição de dados descritivos como a média, a mediana e o desvio-padrão, por exemplo, ainda pode ser considerada uma análise qualitativa dos dados. Isso se dá porque, apesar de podermos constatar que o tempo de reação na identificação de palavras com o afixo 'ede' é 8 milissegundos maior que aquelas com o afixo que obteve o menor tempo de reação nos dados, 'lig',[3] não conseguimos inferir se esses 8 milissegundos de fato representam um maior tempo de processamento das palavras com 'ede', pois essa diferença pode ter acontecido "ao acaso". Isto é, será que uma pesquisa com outros participantes não geraria um tempo de reação de palavras com 'ede' igual ou até menor do que o aquele para palavras com 'lig'? Apenas com a análise descritiva não é possível responder a essa pergunta.

## 2. Teste de hipótese

O próximo passo natural na busca por conhecimento de análise de dados é em direção a testes estatísticos, que, de fato, acrescentam informações inferenciais à análise, mas que, como será demonstrado, têm sérias limitações. Um dos primeiros testes de hipótese aprendidos é o teste *t*, que compara duas médias (de dois grupos, de um mesmo grupo em momentos diferentes, ou de um grupo em relação a um valor pré-determinado) e informa, por meio do *valor de p*, a probabilidade de

[3] Calculado no *R* com exp(6.74) - exp(6.73).

encontrarmos uma diferença de médias igual ou superior à encontrada caso os dois grupos fossem iguais (i.e., hipótese nula). Caso essa probabilidade seja muito pequena, abaixo de 5%, o/a pesquisador/a infere que as médias são significativamente diferentes. Apesar de categórico e arbitrário, esse limite de 5% é amplamente aceito na linguística, bem como em outras áreas de ciências sociais.

Desta forma, um/a linguista pode olhar para o aumento no tempo de reação entre palavras com os afixos 'ede' e 'bar', que são os que tiveram o maior salto no gráfico de caixas da figura e nas médias apresentadas no quadro 2, e se perguntar se essa diferença de 52 milissegundos[4] é estatisticamente significativa ou não. Para isso, ele/a conduz um teste *t* de duas amostras não pareado com os comandos das linhas 20-22 do quadro 3, e conclui, pelo pequeno valor de p apresentado no output (p-value = 0.01268), que esses dois afixos geram tempos de reação significativamente diferentes. Isto é, a probabilidade de se encontrar uma diferença de 52 milissegundos ou mais caso não haja diferença real de tempos de reação entre os dois afixos é de apenas 1,3%, abaixo do limite de 5% pré-estabelecido e aceito pela comunidade acadêmica.

```
20| dan %>%
21|   filter(Affix %in% c("ede", "bar")) %>%
22|   t.test(LogRT ~ Affix, data = .)

Welch Two Sample t-test[5]

data:  LogRT by Affix
t = 2.5034, df = 421.23, p-value = 0.01268
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.01095439 0.09103241
sample estimates:
mean in group bar mean in group ede
        6.795550         6.744556
```

QUADRO 3 - Linhas de comando e output para um teste t entre as médias de tempo de reação para os afixos 'ede' e 'bar'.
Fonte: elaborado pelos autores.

Note que, se partíssemos do pressuposto de que 'bar' naturalmente sempre leva a tempos de reação mais altos do que 'ede', poderíamos conduzir um teste *t* unicaudal, adicionando o argumento alternative = greater nos parênteses da linha 22. O resultado de um teste unicaudal sempre gera um valor de p que corresponde à metade daquele para o teste bicaudal – neste caso, o teste unicaudal geraria um valor de p de 0,006, metade do apresentado no quadro 3. Isso mostra como o valor de p é sensível às assunções prévias do/a pesquisador/a – no caso dessa decisão por um teste unicaudal, reduzindo o valor de p para a sua metade.

---

[4] Calculada no R com exp(6.8) - exp(6.74).

[5] Em sua configuração *default*, o R conduz um teste *t* de Welch, que não assume variâncias homogêneas nas duas amostras testadas, e é por isso que o grau de liberdade (df) pode ter decimais. Para rodar um teste *t* tradicional (de Student), basta adicionar o argumento var.equal = T dentro dos parênteses da linha 22, cujo resultado dará um grau de liberdade de 422 (424 dados - 2 grupos).

# 3. ANOVA com comparações múltiplas

Uma segunda opção de comparação entre médias, que normalmente corresponde ao próximo passo na aquisição de conhecimentos estatísticos, é a condução de uma Análise de Variância (ANOVA) para se comparar as médias de mais de 2 grupos.[6] No caso dos dados que estamos utilizando, um/a pesquisador/a poderia ter o interesse de saber se há pelo menos uma diferença significativa entre as médias do tempo de reação dos cinco afixos e, caso haja, investigar quais pares de afixos apresentam diferenças estatisticamente significativas.

Para isso, o/a pesquisador/a pode utilizar as linhas 23 e 24 do quadro 4 para rodar uma ANOVA e visualizar o seu resumo. Com o valor de p baixo ($p < 0{,}001$; Pr(>F) = 1.32e-09 no *output*), o/a pesquisador/a rejeita a hipótese nula de que os afixos não geram tempos de reação diferentes, e infere que há diferença em pelo menos um dos pares. Mais especificamente, o valor de p em questão é de 0,00000000132, muito abaixo do limite de 0,05.

```
23| anovaDan = aov(LogRT ~ Affix, data = dan)
24| summary(anovaDan)

            Df  Sum    Sq Mean   Sq F value   Pr(>F)
Affix        4   1.89    0.4717     12.09        1.32e-09 ***
Residuals   1035  40.39  0.0390
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

25| TukeyHSD(anovaDan)

Tukey multiple comparisons of means
  95% family-wise confidence level

Fit: aov(formula = LogRT ~ Affix, data = dan)

$Affix
             diff               lwr               upr               p adj
ede-bar  -0.05099340   -1.034311e-01  0.001444266  0.0612021
ende-bar  0.05413199   -5.440768e-05  0.108318388  0.0503757
ere-bar    0.01781355   -3.468447e-02  0.070311574  0.8863835
lig-bar    -0.06136498   -1.139239e-01 -0.008806100  0.0127046
ende-ede  0.10512539    5.129924e-02  0.158951542  0.0000012
ere-ede    0.06880695    1.668084e-02  0.120933061  0.0029895
lig-ede    -0.01037158   -6.255898e-02  0.041815820  0.9827775
ere-ende  -0.03631844   -9.020339e-02  0.017566517  0.3500510
lig-ende   -0.11549697   -1.694412e-01 -0.061552722  0.0000001
lig-ere    -0.07917853   -1.314266e-01 -0.026930484  0.0003598
```

QUADRO 4 - Análise de Variância (ANOVA) seguida do teste pareado *post hoc Tukey Honest Significant Differences*.
Fonte: elaborado pelos autores.

Para saber entre quais pares há uma diferença significativa, o/a pesquisador/a conduz, então, um teste pareado *post-hoc*, que, na verdade, funciona como múltiplos testes *t* que incluem valores de p

[6] Naturalmente, ANOVAs também podem ser utilizadas na comparação de médias de 2 grupos.

ajustados para múltiplas testagens – como já mencionado na introdução, conduzir múltiplos testes com os mesmos dados aumenta as chances de erro do tipo I. Em um cenário em que utilizamos o valor de 5% como limite para se inferir diferença significativa, e conduzimos 10 testes com os mesmos dados, há uma probabilidade de 40% de encontrarmos pelo menos *um* falso positivo ($1 - (1 - 0.05)^{10} = 0.401$). Por isso, nas múltiplas testagens pareadas *post-hoc* comumente conduzidas após uma ANOVA, é feito um ajuste no valor de p, deixando o limite de rejeição da hipótese nula mais rígido.

Uma opção de teste pareado *post-hoc* é o *Tukey Honest Significant Differences*[7], conduzido com a linha 25 do quadro 4. Ao analisar o output do teste de *Tukey*, é possível notar que o teste apresenta diferenças significativas entre os afixos 'lig-bar', 'ende-ede', 'ere-ede', 'lig-ende' e 'lig-ere', já que os valores de p ajustado (*p adj*) foram inferiores a 0,05. Utilizando a ordem dos afixos apresentados na figura 1, infere-se, portanto, que o primeiro afixo, 'lig', gera tempos de reação significativamente mais baixos do que o terceiro, o quarto e o quinto afixos; o segundo afixo, 'ede', gera tempos de reação significativamente mais baixos do que o quarto e quinto afixos; e o terceiro afixo, 'bar', difere significativamente do primeiro afixo.

Perceba, contudo, que o par de afixos comparados no teste *t* do quadro 3, 'bar' e 'ede', não se mostrou significativamente diferente no teste pareado *post-hoc* (p adj = 0,06), resultado que contradiz o que revelou o teste *t* acima (p = 0,01). Naturalmente, o tamanho da diferença em milissegundos ou na escala logarítmica do tempo de reação de palavras com 'bar' e com 'ede' foi o mesmo nas duas análises, uma vez que estão baseadas nos mesmos dados. Porém, o valor de p de cada uma levaria pesquisadores a conclusões completamente diferentes: uma indicando efeito desses afixos, e outra não. Isso acontece por causa das comparações múltiplas efetuadas no teste *post-hoc* em questão. Afinal, no teste *t*, a intenção era unicamente comparar *dois* afixos específicos; na ANOVA, por outro lado, a intenção era comparar *todos* os afixos e, mais tarde, avaliar cada comparação individual, o que nos levou a comparações múltiplas e ao devido ajuste de valor de p. Ou seja, o valor de p da comparação entre 'bar' e 'ede' (e, portanto, a conclusão sobre essa comparação) depende das intenções do/a pesquisador/a.

São diversas e antigas as críticas à ênfase cega ao valor de p (e.g., BERGER; SELLKE, 1987; LOFTUS, 1993; COHEN, 1994; JOHNSON, 1999; WAGENMAKERS, 2007; NUZZO, 2014; HALSEY *et al*, 2015; KRUSCHKE, 2015). Começamos destacando que o valor de p representa apenas a probabilidade de se encontrar dados tão ou mais extremos caso a hipótese nula (até aqui de que não há diferença entre grupos) seja verdadeira. Ele não diz nada sobre a probabilidade de a hipótese alternativa (até aqui de que os grupos são diferentes) ser verdadeira, tampouco nos informa sobre o tamanho do efeito da variável 'afixo' no tempo de reação em nosso exemplo: é possível encontrar valor de p abaixo de 0,05 em estudos com tamanhos de efeitos minúsculos e/ou baixo poder.[8]

---

[7] Outras alternativas comuns incluem Bonferroni, Holm, Gabriel, Scheffé e Duncan.

[8] O poder de um teste estatístico é a probabilidade de que ele rejeite a hipótese nula quando de fato ela é falsa. Ele é calculado utilizando-se o tamanho do efeito, o tamanho da amostra, o desvio-padrão e o nível de significância considerado (que é o limite de 5% comumente aceito).

Além disso, o uso do valor de p impõe uma decisão binária e categórica (rejeitar a H0 se p<0,05 e não rejeitar se p>0,05), embora dados reais normalmente sejam mais complexos e incluam variâncias e gradiências incompatíveis com uma decisão tão categórica – perceba também que o valor de significância comumente adotado, de 0,05, é um valor arbitrário. Por fim, a ênfase no valor de p para tomadas de decisão inferenciais sobre o efeito de variáveis, advindo em grande parte do *publication bias*, que hipervaloriza estudos com resultados "estatisticamente significativos" (com base sobretudo no valor de p), pode levar à má conduta conhecida como *p-hacking*, quando os dados ou as análises são manipuladas a fim de se chegar a um valor de p baixo. Isso pode envolver desde a retirada de dados que estejam mais nas extremidades da distribuição (mesmo não sendo erros de medição) até manipulações mais drásticas dos dados ou dos parâmetros das análises estatísticas.

No nosso exemplo, um/a pesquisador/a que tivesse interesse especial no efeito dos afixos 'bar' e 'ede', e que tivesse conduzido tanto o teste *t* como a ANOVA, poderia intencionalmente escolher não reportar os resultados da ANOVA e reportar apenas o teste *t*, já que nele a diferença entre as médias resultou em p<0,05 – mesmo que uma Análise de Variância com posterior teste pareado seja mais informativa por incorporar mais dados à análise. Isso seria um tipo de *p-hacking*. Por outro lado, se o pesquisador/a quisesse focar sua análise desde o início apenas nos afixos 'bar' e 'ede', o valor de p do teste *t* seria legítimo (uma vez que os resultados da ANOVA e do teste *post-hoc* seriam desconhecidos nesse cenário hipotético).

Por fim, é importante lembrar que ANOVAs podem ser descritas como modelos de regressão em que as variáveis preditoras são categóricas. Ou seja, são um tipo específico de regressão. Apesar disso, o output tradicional de uma ANOVA no *R* difere do *output* de um modelo de regressão típico, como veremos a seguir.

# 4. Modelo de Regressão (linear)

Modelos de regressão são uma opção mais robusta em relação a testes de hipótese por colocarem a ênfase no tamanho do efeito e por permitirem a elaboração de modelos mais complexos, que incorporam diversas variáveis preditoras (contínuas e/ou categóricas), suas possíveis interações, e até mesmo a natureza aleatória de algumas delas, como será demonstrado na próxima seção. Modelos mais complexos estão mais alinhados à natureza dos fenômenos naturais, como os que envolvem linguagem, que costumam abranger diversos fatores concomitantemente. A explicação de um fenômeno complexo exige, portanto, um modelo que dê conta de sua complexidade.

O primeiro modelo que apresentamos costuma ser o primeiro aprendido por um/a pesquisador/a em busca de mais conhecimentos estatísticos. Trata-se de um modelo de regressão linear simples, com apenas uma variável preditora buscando explicar e prever uma variável de resposta contínua – ou seja, um modelo fundamentalmente equivalente à ANOVA discutida acima. Um modelo de regressão cumpre o duplo papel de verificar uma relação entre variáveis preditoras e de resposta, e de, ao mesmo tempo, gerar previsões de dados futuros. No nosso caso, a variável preditora é o

fator 'Affix', que possui cinco níveis, e a variável de resposta o tempo de reação em escala logarítmica. Um modelo de regressão, portanto, buscará identificar uma possível relação entre afixo e tempo de reação (que pode ou não ser causal), e, ao mesmo tempo, estimar o tempo de reação de palavras não necessariamente observadas nos dados. É possível rodar esse modelo e visualizar seus resultados executando as linhas 26 e 27 do quadro a seguir.[9]

```
26| fit1 = lm(LogRT ~ Affix, data = dan)
27| summary(fit1)

Residuals:
    Min      1Q          Median      3Q         Max
-0.48793  -0.12759  -0.02011    0.10379  0.91810

Coefficients:
              Estimate  Std. Error   t value    Pr(>|t|)
(Intercept) 6.79555    0.01366    497.301  < 2e-16 ***
Affixede    -0.05099   0.01919    -2.657     0.00800 **
Affixende    0.05413   0.01983     2.730     0.00644 **
Affixere     0.01781   0.01921     0.927     0.35403
Affixlig     -0.06136  0.01923    -3.190     0.00146 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1976 on 1035 degrees of freedom
Multiple R-squared:  0.04463,     Adjusted R-squared:  0.04094
F-statistic: 12.09 on 4 and 1035 DF,  p-value: 1.321e-09
```

QUADRO 5 - Modelo de regressão linear simples para tempo de reação em função dos afixos.
Fonte: elaborado pelos autores.

A parte principal do *output* são os coeficientes, encontrados na coluna *Estimate*. O primeiro, *Intercept*, informa o valor estimado pelo modelo para o tempo de reação quando o afixo utilizado é o 'bar' (o primeiro nível do fator em questão em ordem alfabética, e o único a não ser explicitamente listado no *output* do modelo), que é um valor de 6,8 na escala logarítmica (894 milissegundos).[10] O baixíssimo valor de p, Pr(>|t|), apresentado ao final da linha que começa com (*Intercept*), <2e-16, indica que o valor dessa estimativa (também chamada de coeficiente linear) difere significativamente de zero, o que era esperado, já que um tempo de reação de zero é impossível (independentemente do afixo usado como referência). As demais estimativas (coeficientes angulares, ou *slopes*) indicam quanto o tempo de reação aumenta ou diminui para cada afixo – relativamente ao afixo 'bar'. Isto é, ao se trocar o afixo de referência 'bar' pelo segundo, 'ede', o tempo de reação diminui 0,051 na escala logarítmica, o que corresponde a um tempo de reação de 849 milissegundos para 'ede'.[11] Além disso,

[9] O leitor pode confirmar que o modelo em questão produz os mesmos resultados produzidos pela ANOVA discutida acima rodando lm(anovaDan).

[10] Calculado no R com exp(6.79555).

[11] Calculado no R com exp(6.79555 - 0.05099).

o baixo valor de p ao final da linha que começa com 'Affixede' (0,008) indica que essa diferença entre 'bar' e 'ede' difere significativamente de zero. O intervalo de 95% de confiança para cada coeficiente, ($CI_{0.95}^{\hat{\beta}_i} = [\hat{\beta}_i \pm 1.96 \times SE(\hat{\beta}_i)]$), pode ser obtido com o comando confint(fit1). No caso da diminuição do tempo de reação de 0,051 de 'bar' para 'ede', o intervalo é de [-0.089 -0.013], e o fato de o intervalo não cruzar zero indica que o efeito é realmente de diminuir o tempo de reação.

Com esse mesmo raciocínio, é possível inferir que o afixo 'ende' tem um efeito significativo de aumentar o tempo de reação na identificação das palavras em relação ao afixo de referência, 'bar', e que o afixo 'lig' tem um efeito de diminuir o tempo de reação em relação a 'bar'. O afixo 'ere' também recebeu um coeficiente linear (*slope*) positivo, indicando aumento no tempo de reação em relação a 'bar'; contudo, com um efeito que não difere estatisticamente de zero, o que levaria o/a pesquisador/a a inferir que esse afixo não apresenta efeito significativo em relação a 'bar' – isto é, não é possível rejeitar a hipótese nula. Note também que o contraste hipoteticamente de interesse no teste *t*, 'bar-ede', que se mostrou significativamente diferente no teste *t* mas não no teste *post hoc* de Tukey, se mostrou novamente significativo nesse primeiro modelo de regressão. Isso ocorre porque no modelo em questão não estamos elaborando comparações múltiplas, e, portanto, nosso valor de p não precisa ser ajustado. Se rodássemos um teste Tukey em nosso modelo (TukeyHSD(aov(fit1))), também chegaríamos ao mesmo resultado não significativo de 'bar-ede' que encontramos em nosso teste *post-hoc*.

É importante reiterar a sutil distinção entre uma ANOVA seguida de Tukey, e de uma regressão linear tradicional rodada com lm() no R. No primeiro caso, a tendência é desejar visualizar comparações múltiplas, possivelmente por haver um interesse em comparar diferenças entre grupos. No segundo caso, a tendência é escolhermos um nível específico, que servirá de *intercept*, e compararmos todos os demais níveis ao *intercept*. Perceba que, apesar de fundamentalmente idênticos, ambos os métodos podem levar a resultados categoricamente distintos: como uma regressão linear simples tipicamente não efetua comparações múltiplas, não será necessário ajustar valores de p. No exemplo em questão, essa distinção nos levaria a concluir um efeito significativo do sufixo 'ede' relativo a 'bar', algo que não podíamos concluir em nossas comparações *post-hoc* acima.

Esse modelo, no entanto, ainda não é suficiente para explicar de maneira muito robusta a variação no tempo de reação. Essa informação está na penúltima linha do output, com um valor de $R^2 = 0,04$ indicando que a apenas 4% da variação em tempo de reação é explicado pela alternância entre afixos. Em uma análise real, o/a pesquisador/a deveria investigar se a inclusão de outras variáveis preditoras no modelo, bem como possíveis interações entre elas, aumentaria o poder explicativo do modelo, buscando chegar à melhor explicação e previsão do tempo de reação de identificação de palavras complexas a partir dos dados coletadas e das variáveis identificadas e registradas. Nesse caso, o modelo seria um de regressão linear múltipla, já que haveria diversas variáveis preditoras no modelo.

Lembramos que a escolha por um modelo de regressão linear, neste caso, se deu porque a variável de resposta é contínua. No caso de uma variável de resposta binária, utilizaríamos um modelo de regressão logística; no caso de uma variável de resposta categórica com mais de dois níveis (não ordenados), a escolha seria por um modelo de regressão multinomial; para variável de resposta

escalar, a escolha seria por um modelo de regressão ordinal; e para variável de resposta de contagem/proporção, o modelo utilizado deveria ser um de regressão de *Poisson*.

# 5. Modelo de Efeitos Mistos

Os modelos mencionados até o momento trabalham com o pressuposto de independência dos dados coletados, o que raramente é o caso em coletas de dados linguísticos, uma vez que geralmente contamos com o mesmo participante para coletar várias observações e repetimos os itens utilizados para a coleta. Esse é o caso dos dados *danish*, que utilizamos para nossas ilustrações. Há 3.326 observações de tempo de reação coletadas de apenas 22 participantes que foram expostos a 156 palavras complexas. O pressuposto de independência só não teria sido violado se as 3.326 observações tivessem sido coletadas de 3.326 pessoas diferentes reagindo a 3.326 palavras diferentes, cada pessoa contribuindo com apenas uma observação de tempo de reação de uma palavra diferente. Para a maioria dos estudos linguísticos, um desenho experimental como esse é impossível (com raras exceções, como Hassemer e Winter (2016)), pois é muito custoso conseguir participantes, e muitas das vezes a própria natureza do estudo impede a elaboração de tantos itens (BAAYEN, DAVIDSON; BATES, 2008).

Sendo assim, cada participante que forneceu várias observações traz aos dados uma variação intrínseca e individual. É possível que algum participante seja mais rápido de maneira geral no teste de tempo de reação por estar mais alerta naquele momento, por ser naturalmente uma pessoa ágil, por ter facilidade ou familiaridade com a tarefa, por estar motivado a "ajudar" o pesquisador, etc. Outro participante pode ser mais lento que os demais de maneira geral também por motivos idiossincráticos. O mesmo ocorre para cada item que foi utilizado diversas vezes na coleta. Uma palavra em especial pode provocar um tempo de reação mais longo (ou mais curto) por características inerentes a ela (frequência, ortografia, familiaridade, etc.). Essas variações são chamadas de efeitos aleatórios, pois alterando os participantes e as palavras, por exemplo, poderíamos encontrar resultados diferentes. Veja na figura 2, gerada com os comandos do quadro 6, o tempo de reação por indivíduo (linhas) sobreposto aos padrões de tempo de reação do grupo como um todo (caixas), e observe como há variação individual.

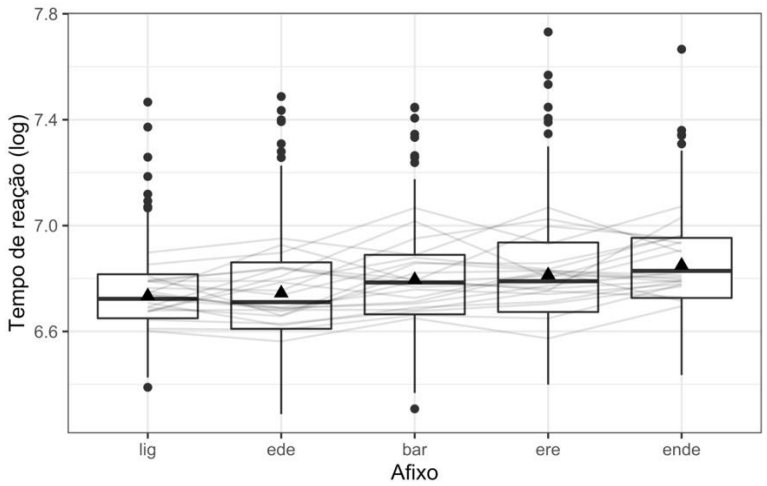


FIGURA 2 - Tempos de reação médios para cada falante (linhas cinzas) sobrepostos a gráfico de caixas.
Fonte: elaborado pelos autores.

```
28| ggplot(data = dan, aes(x = fct_reorder(Affix,LogRT,.fun = mean), y = LogRT)) +
   geom_boxplot() +
   stat_summary(shape = 17) +
   stat_summary(fun = mean, aes(group = Subject), alpha = 0.15, geom = "line") +
   labs(x = "Afixo", y = "Tempo de reação (log)") +
   theme_bw()
```

QUADRO 6 - Linhas de comando para gerar a figura 2.
Fonte: elaborado pelos autores.

O intuito de se ajustar um modelo de regressão não é alcançar resultados válidos apenas para aquela amostra estudada, mas de generalizar seus resultados para a população. Sendo assim, é importante informar ao modelo a existência de efeitos aleatórios para que ele calibre/ajuste os resultados dos efeitos fixos, que são os de interesse. Para se fazer isso, é preciso utilizar um modelo de efeitos mistos, que consideramos ser um próximo tópico a ser buscado por pesquisadores que já dominam modelos de regressão (ver, por exemplo, GODOY; NUNES, 2020). O quadro 7 apresenta as linhas de comando e parte do *output* (os efeitos aleatórios e os efeitos fixos) de um modelo similar ao da seção anterior, ajustado com os comandos do quadro 5, mas, desta vez, prevendo coeficientes lineares aleatórios (*random intercepts*) para sujeitos e palavras.

```
29| library(lmerTest)
30| fit2 = lmer(LogRT ~ Affix + (1 | Subject) + (1 | Word), data = dan)
31| summary(fit2)

Random effects:
 Groups   Name        Variance   Std.Dev.
 Word     (Intercept) 0.004886   0.06990
 Subject  (Intercept) 0.007325   0.08559
 Residual             0.027676 0.16636
Number of obs: 1040, groups:  Word, 49; Subject, 22

Fixed effects:
             Estimate  Std. Error      df          t value    Pr(>|t|)
(Intercept)  6.80096   0.03091       61.07064   220.051   <2e-16 ***
Affixede    -0.05612   0.03521       43.93137  -1.594       0.1182
Affixende    0.05002   0.03622       44.15087   1.381       0.1743
Affixere     0.01091   0.03522       43.96980   0.310       0.7583
Affixlig    -0.06635   0.03523        44.00575  -1.884       0.0662 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

QUADRO 7 - Linhas de comando e parte do output (efeitos aleatórios e efeitos fixos) de um modelo de efeitos mistos para tempo de reação em função dos cinco afixos selecionados, prevendo *random intercepts* para sujeitos e palavras.
Fonte: elaborado pelos autores.

Em poucas palavras, esse modelo considera a variabilidade no valor do coeficiente linear (*intercept*) para cada sujeito e para cada palavra. Trata-se, portanto, de um modelo mais robusto, pois leva em consideração a não independência da coleta de dados. Note que o contraste entre 'bar' e 'ede', que foi significativo no teste *t*, não significativo no teste de Tukey, e novamente significativo no modelo de regressão linear simples, volta a ser não significativo no modelo de efeitos mistos.

O leitor deve ter percebido a semelhança entre os coeficientes de fit1 e de fit2. De fato, em boa parte dos casos em que os dados analisados sejam relativamente balanceados, um modelo de efeitos mistos não gerará grandes alterações nos valores dos coeficientes se comparado a um modelo sem efeitos aleatórios equivalente. Ou seja, ao comparar os coeficientes de fit1 e fit2, um/a pesquisador/a poderia subestimar a importância de efeitos mistos. É apenas quando observamos os erros padrões dos coeficientes que percebemos a grande diferença entre fit1 e fit2: nosso modelo com efeitos aleatórios quase dobrou o erro padrão de cada coeficiente. Esse aumento diminui o valor de *t* de cada coeficiente (uma vez que *t* = Estimate / Std Error), e, consequentemente, afeta o valor de p dos efeitos em questão. Portanto, modelos de efeitos mistos afetarão principalmente o erro padrão dos coeficientes estimados. Geralmente, o erro padrão aumentará (trata-se, assim, de um modelo mais conservador, que reduz a chance de erro do tipo I). Às vezes, contudo, é possível que o erro padrão diminua, o que reduz a chance de erro do tipo II – um exemplo é discutido no capítulo 9 de Garcia (2021).

Reforçamos que estamos realizando múltiplas análises com esses dados por motivo exclusivamente didático, e que não recomendamos que um/a pesquisador/a o faça pelos motivos já elencados na introdução, e, principalmente, se o motivo for escolher a análise que gere o resultado esperado (*p-hacking*), já que a cada nova análise estamos mostrando como a significância do contraste entre 'bar' e 'ede' muda. Até este ponto, o modelo com efeitos mistos é o modelo mais robusto, e que deveria ser o escolhido logo na fase de planejamento do estudo. Essa escolha, sem levar em

consideração as demais variáveis disponíveis para análise, levaria à inferência de que a troca entre esses dois sufixos não leva a mudanças no tempo de reação na identificação de palavras.

Em uma análise real com modelos de efeitos mistos, é importante buscar incorporar ao modelo não apenas coeficientes lineares aleatórios (*random intercepts*), mas também coeficientes angulares aleatórios (*random slopes*). *Random slopes* são importantes porque raramente encontramos o mesmo efeito de uma dada variável para diferentes participantes em um experimento, por exemplo. Para entendermos a diferença entre esses dois tipos de efeitos mistos, observe a comparação entre 'bar' e 'ede' na figura 2 (lembre-se de que nosso afixo de referência é 'bar'). As linhas que representam os 22 falantes em nossos dados partem de pontos diferentes no eixo *y* (tanto para 'bar' quanto para 'ede'): para 'bar', alguns falantes apresentam um tempo de reação acima de 6.8 log(ms), outros abaixo de 6.8 log(ms). Essa é a motivação para adicionarmos *random intercepts* para nossos participantes, ou seja, `(1 | Subject)`, uma vez que diferentes falantes partem de valores diferentes, o que nos mostra que o *intercept* do nosso modelo não é o mesmo para todos os falantes. Em segundo lugar, observe como as linhas não são paralelas entre 'bar' e 'ede'. Isso nos mostra que o efeito de 'ede' (relativo a 'bar') não é constante para os 22 participantes: para alguns falantes, 'ede' gera um tempo de reação menor do que 'bar'; para outros, gera um tempo de reação maior. Essa é a motivação para adicionarmos *random slopes* para Affix para nossos participantes, ou seja, `(1 + Affix | Subject)`. Para maiores informações sobre modelos com efeitos mistos, sugerimos os capítulos 11 a 17 de Gelman e Hill (2006), assim como o capítulo 9 de Garcia (2021).

Com nossos dados, um modelo com *random slopes* para afixo (por falante) gera um aviso de *singularidade*[12] por ser complexo demais para a quantidade de dados.[13] Além disso, o modelo não proporciona uma melhor *fit* relativo a `fit2` acima ($p = 0.26$).

# 6. Conclusão

O objetivo deste artigo foi demonstrar que diferentes análises, neste caso seguindo a sequência que acreditamos ser aquela de uma capacitação natural de linguistas em relação a análise quantitativa de dados, podem levar a conclusões categoricamente distintas, e que modelos mais robustos, como os de efeitos mistos, são mais realistas e informativos. Com uma análise descritiva dos dados (seção 1), focamos nossa atenção nos afixos 'ede' e 'bar'. Apenas olhando para a diferença entre as médias de tempo de reação de ambos os afixos, não é possível inferir se essa diferença se deu ao acaso ou se há de fato um efeito desses afixos sobre o tempo de reação. Um teste *t* (seção 2) nos levou à conclusão de que a diferença entre esses dois afixos é significativa, havendo, portanto, um efeito de 'bar' *vs.* 'ede' sobre o tempo de reação. Já uma ANOVA seguida de testes *post-hoc* pareados (seção 3),

[12] Consulte Winter (2019, capítulo 15) para uma explicação detalhada sobre esse aviso e instruções de como lidar com ele.

[13] É possível resolver o problema rodando o mesmo modelo em Bayes.

por causa do ajuste nos valores de p devido às múltiplas comparações, não revelou uma diferença significativa entre 'ede' e 'bar'. No modelo de regressão linear (seção 4), a diferença entre esses dois afixos volta a ser significativa; e, ao adicionarmos sujeito e palavra como efeitos aleatórios no modelo de efeitos mistos (seção 5), a diferença volta a não ser significativa.

Essa alternância nos valores de p quanto à diferença entre 'ede' e 'bar' foi o foco deste artigo. Foram quatro análises inferenciais realizadas (teste *t*, ANOVA + testes pareados, modelo de regressão linear, modelo de efeitos mistos), e as conclusões sobre um possível efeito de 'ede' *vs.* 'bar' sobre os tempos de reação foram, portanto, sim > não > sim > não. O problema da divulgação de uma conclusão equivocada dos dados pode vir de duas fontes: da má conduta de pesquisadores que, conhecendo a possibilidade de diferentes análises, conduzem várias e escolhem as que geram resultados mais convenientes; ou da falta de conhecimento em análise quantitativa dos dados por parte de pesquisadores, que se limitam a análises menos completas. Acreditamos que a capacitação de pesquisadores pode auxiliar na solução de ambas as fontes. Pesquisadores mais conhecedores de análise quantitativa de dados são capazes de conduzir análises mais robustas e de tecer críticas a trabalhos mal intencionados.

Sendo assim, deixamos registrados alguns livros e materiais on-line para estudantes ou pesquisadores da área de linguística que desejarem se capacitar em análise quantitativa de dados. Entre os manuais mais básicos de estatística para linguistas, há os livros de Larson-Hall (2015) e de Loerts, Lowie e Seton (2020). Para uma leitura intermediária, um pouco mais voltada para os modelos de regressão utilizados neste artigo, há os manuais de Gries (2013), Levshina (2015) e Winter (2019). Como opção para quem queira focar exclusivamente em modelos de regressão, com ênfase na visualização de dados, e com uma breve introdução a análises bayesianas, recomendamos o livro de Garcia (2021). Entre os materiais on-line, gratuitos, destacamos o curso de Lima Jr, Garcia e Angele (2020),[14] bem como os materiais de Oushiro (2021);[15] Sonderegger, Wagner e Torreira (2018);[16] Garcia (2019);[17] e Godoy (2019).[18]

Por fim, acreditamos que o próximo passo analítico na busca por capacitação em análise quantitativa de dados seja uma análise bayesiana, em que o/a pesquisador/a possa customizar seus modelos a partir de distribuições *a priori* que melhor caracterizem o conhecimento já estabelecido na área. Embora estejam fora do escopo deste artigo, modelos bayesianos oferecem inúmeras vantagens: não dependem de valores de p; oferecem um *output* mais completo, em que uma distribuição de valores (*a posteriori*) é fornecida (e não apenas uma *point estimate*, como vimos nos modelos acima); apresentam uma interpretação mais intuitiva; geram a probabilidade de um valor de

---

[14] https://ead.abralin.org/course/view.php?id=10 e https://guilhermegarcia.github.io/rling.html

[15] https://rpubs.com/oushiro/iel e 10.5281/zenodo.822069

[16] http://people.linguistics.mcgill.ca/~morgan/book/

[17] https://guilhermegarcia.github.io/rWorkshop/garcia_rWorkshop_complete.html

[18] https://github.com/mahayanag/intro_estatistica_linguistica#readme e https://mahayana.me/mlm/

parâmetro considerando os dados, e não a probabilidade dos dados considerando um valor de parâmetro (como nos modelos frequentistas vistos neste artigo).

Esperamos que os exemplos acima tenham demonstrado como as conclusões de uma análise estatística podem ser sensíveis ao tipo de análise que realizamos. Naturalmente, métodos quantitativos mais avançados não promovem um estudo fraco à excelência, mas certamente consolidam o potencial de um estudo promissor.

## REFERÊNCIAS


BAAYEN, Rolf Harald. *languageR*: v 1.0, 2007a.

BAAYEN, Rolf Harald. *Analyzing Linguistic Data*: A practical introduction to statistics using R, Cambridge: Cambridge University Press, 2007b.

BAAYEN, Rolf Harald; DAVIDSON, Doug; BATES, Douglas. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, v. 59, n. 4, p. 390–412, 2008. https://doi.org/10.1016/j.jml.2007.12.005

BALLING, Laura Winther; BAAYEN, Rolf Harald. Morphological effects in auditory word recognition: Evidence from Danish. *Language and Cognitive Processes*, v. 23, n. 7–8, p. 1159–1190, 2008. https://doi.org/10.1080/01690960802201010

BARR, D.; LEVY, R.; SCHEEPERS, C.; TILY, H. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278. 2013.

BERGER, James O.; SELLKE, Thomas. Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, v. 82, n. 397, p. 112–122, 1987. https://doi.org/10.2307/2289138

CHAMBERS, John, M. S, R, and Data Science. *The R Journal*, v. 12, n. 1, p. 462-476, 2020. DOI 10.32614/RJ-2020-028 Acesso em 24 novembro 2020.

COHEN, Jacob. The earth is round (p < .05). *American Psychologist*, v. 9, n.12, p. 997–1003, 1994. https://doi.org/10.1037/0003-066X.49.12.997

GARCIA, Guilherme Duarte. *Introduction to data analysis using* R, 2019. Disponível em https://guilhermegarcia.github.io/rWorkshop/garcia_rWorkshop_complete.html.

GARCIA, Guilherme D. *Data visualization and analysis in second language research.* NY: Routledge, 2021.

GELMAN, A.; HILL, J. *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press. 2006.

GODOY, Mahayana Cristina. *Introdução aos modelos lineares mistos para os estudos da linguagem*. PsyArXiv, 2019. https://doi.org/10.17605/OSF.IO/9T8UR

GODOY, Mahayana C.; NUNES, Marcus A. Uma comparação entre ANOVA e modelos lineares mistos para análise de dados de tempo de resposta. *Revista da ABRALIN*, v. 19, n. 1, pp. 1–23, 17 jul. 2020. https://doi.org/10.25189/rabralin.v19i1.1388

GRIES, Stefan Th. *Statistics for linguistics with R*: A practical introduction. Berlin: Walter de Gruyter, 2013.

HALSEY, Lewis G.; CURRAN-EVERETT, Douglas; VOWLER, Sarah L.; DRUMMOND, Gordon B. The fickle P value generates irreproducible results. *Nature methods*, v. 12, n. 3, p. 179–185, 2015. https://doi.org/10.1038/nmeth.3288

HASSEMER, Julius; WINTER, Bodo. Producing and perceiving gestures conveying height or shape. *Gesture*, v. 15, n. 3, pp. 404–424, 2016. https://doi.org/10.1075/gest.15.3.07has

JAEGER, T. Florian. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of memory and language*, 59(4), 434-446. 2008.

JOHNSON, Douglas H. The insignificance of statistical significance testing. *The Journal of Wildlife Management*, p. 763–772, 1999. https://doi.org/10.2307/3802789

KRUSCHKE, John K. *Doing Bayesian data analysis*: a tutorial with R, JAGS, and Stan, 2a edição. Elsevier, 2015.

LARSON-HALL, Jenifer. *A guide to doing statistics in second language research using SPSS and R*. Routledge, 2015.

LEVSHINA, Natalia. *How to do linguistics with R*: Data exploration and statistical analysis. Amsterdam: John Benjamins Publishing Company, 2015.

LIMA JR, Ronaldo Mangueira; GARCIA, Guilherme Duarte; ANGELE, Bernhard. *Introdução a modelos de regressão para linguistas no R*, 2020. Disponível em https://guilhermegarcia.github.io/rling.html

LOERTS, Hanneke; LOWIE, Wander; SETON, Bregtje. *Essential Statistics for Applied Linguistics*: Using R Or JASP. Amsterdam: Macmillan International, Red Globe Press, 2020.

LOFTUS, Geoffrey R. A picture is worth a thousand p values: On the irrelevance of hypothesis testing in the microcomputer age. *Behavior Research Methods, Instruments, & Computers*, v. 25, n. 2, p. 250–256, 1993. https://doi.org/10.3758/BF03204506

NUZZO, Regina. Scientific method: statistical errors. *Nature News*, v. 506, n. 7487, p. 150, 2014.

OUSHIRO, Livia. Introdução à Estatística para Linguistas (Version 2.0.3). Zenodo. http://doi.org/10.5281/zenodo.4755739. 2021.

R CORE TEAM. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Acesso http://www.R-project.org/ Acesso em: 15 jun. 2020.

SONDEREGGER, Morgan; WAGNER, Michael; TORREIRA, Francisco. *Quantitative Methods for Linguistic Data*. v. 1.0 (out/2018), 2018. Disponível em http://people.linguistics.mcgill.ca/~morgan/book/

WAGENMAKERS, Eric-Jan. A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review*, v. 14, n. 5, p. 779-804, 2007. https://doi.org/10.3758/BF03194105

WINTER, Bodo. *Statistics for linguists*: An introduction using R. Routledge, 2019.