Análise Quantitativa de Dados em Linguística

Teste t e ANOVA

Ronaldo Lima Jr.

ronaldojr@letras.ufc.br

ronaldolimajr.github.io

Universidade Federal do Ceará

Roteiro

1. Teste t
2. ANOVA
3. Testes não paramétricos
4. Tamanho do efeito

Disclaimer!

→ Ensinarei alguns testes de hipótese (NHST – *Null-Hypothesis Significance Tests*: teste de proporção, teste de qui-quadrado, teste-t e ANOVA) para que compreendam leituras que os envolvam, mas não recomendo utilizá-los nas análises

→ No lugar deles, recomendo a utilização de Modelos de Regressão

- Se quiserem muito utilizar testes de hipótese, lembrem-se de analisar (e reportar!) também:
  - tamanho do efeito
  - intervalo de confiança
  - poder estatístico

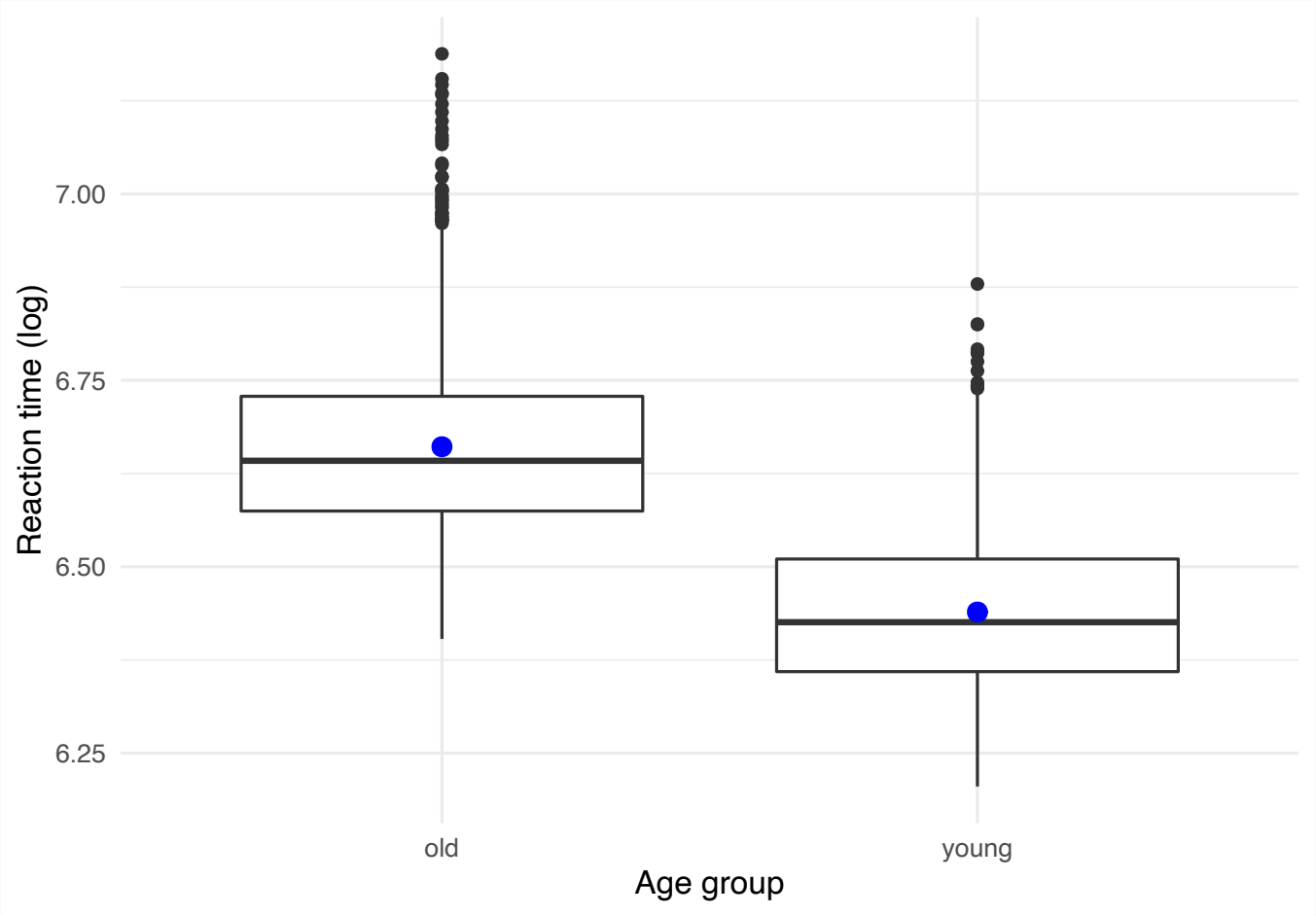
Teste t

Teste t

- Testar a diferença entre duas médias
  - média de uma amostra vs média da população (e.g., média de uma turma no Enem vs média nacional do Enem)
  - médias de duas amostras independentes (e.g., médias de turmas diferentes)
  - médias de uma mesma amostra em dois momentos distintos (e.g., médias de uma mesma turma antes e depois de uma intervenção)
- Variável resposta contínua
- Variável preditora categórica com 2 níveis

Exemplo 1

Dados “english”, tempo de reação (RTlexdec) de jovens e de velhos (young, old)



Exemplo 1

```
1 english %>%
2   group_by(AgeSubject) %>%
3   summarize(MeanRT = mean(RTlexdec))

4   AgeSubject MeanRT
5 1 old          6.66
6 2 young        6.44
```

- H_0: participantes jovens e velhos têm o mesmo tempo de reação
- H_1: participantes jovens e velhos têm tempos de reação diferentes

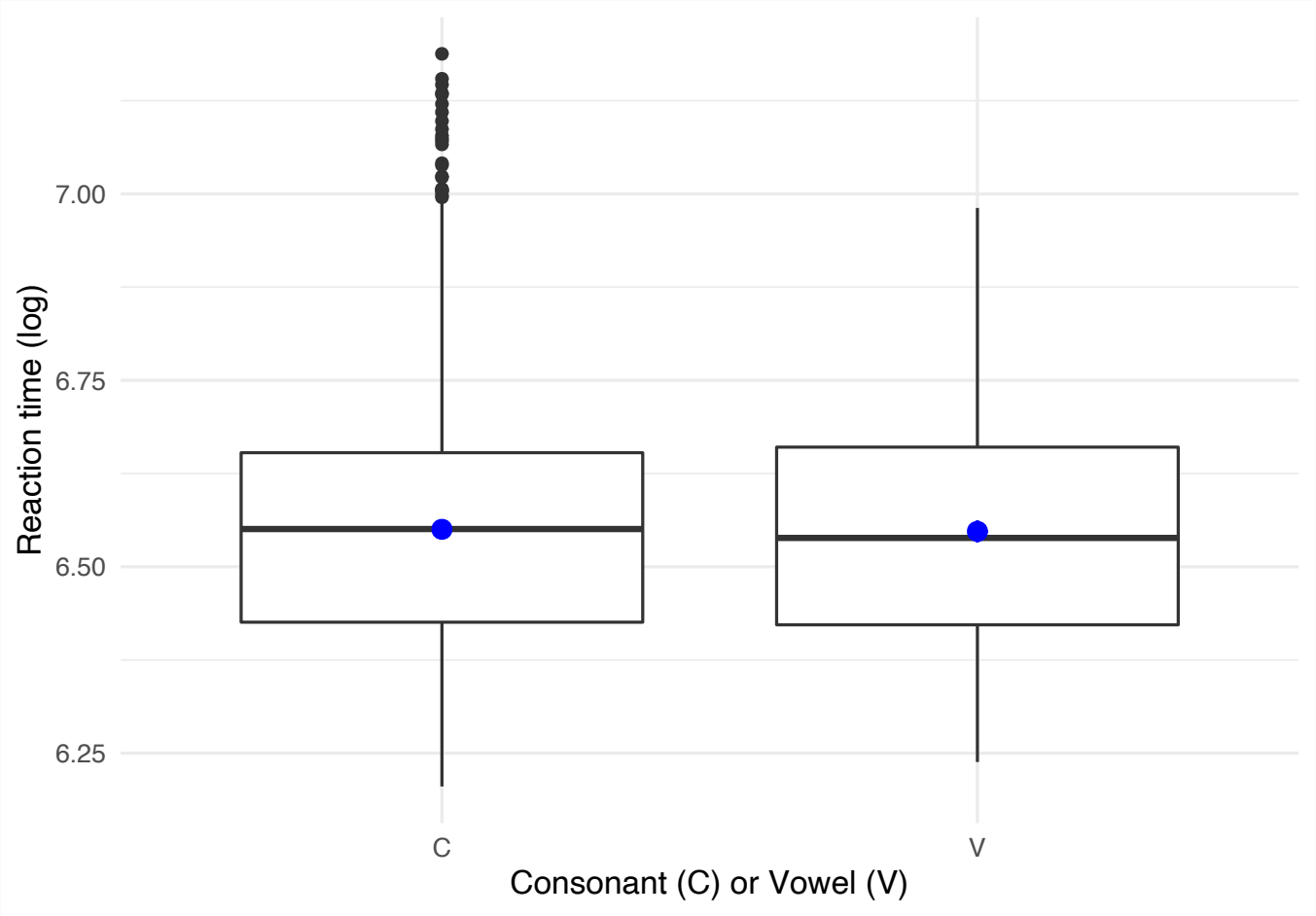
Exemplo 1

```
t.test(RTlexdec ~ AgeSubject, english)

        Welch Two Sample t-test

data:  RTlexdec by AgeSubject
t = 67.468, df = 4534.6, p-value < 2.2e-16
alternative hypothesis: true difference in means between group old and group young is not
    equal to 0
95 percent confidence interval:
 0.2152787 0.2281642
sample estimates:
  mean in group old mean in group young
           6.660958            6.439237
```

- $t(4535) = 68, p < 0.001$
- ~~H_0: participantes jovens e velhos têm o mesmo tempo de reação~~
- H_1: participantes jovens e velhos têm tempos de reação diferentes

Exemplo 2

Dados “english”, tempo de reação (RTlexdec) de palavras começando com consoantes ou vogais (CV)



Exemplo 2

```
  CV    MeanRT
1 C       6.55
2 V       6.55
```

- H_0: palavras começando com C ou V levam ao mesmo tempo de reação
- H_1: palavras começando com C ou V levam a tempos de reação diferentes

Exemplo 2

```
t.test(RTlexdec ~ CV, english)

        Welch Two Sample t-test

data:  RTlexdec by CV
t = 0.18295, df = 127.09, p-value = 0.8551
alternative hypothesis: true difference in means between group C and group V is not equal
   to 0
95 percent confidence interval:
 -0.02708861  0.03260793
sample estimates:
mean in group C mean in group V
       6.550171        6.547411
```

- $t(127) = 0.18, p < 0.855$
- H_0: palavras começando com C ou V mesmo tempo de reação
- H_1: palavras começando com C ou V tempos de reação diferentes

$\rightarrow$ Não podemos rejeitar a hipótese nula

Exemplo 3

Dados “english”, tempo de reação (RTlexdec) substantivos e verbos (WordCategory)

$\rightarrow$ Tarefa de casa

Teste t

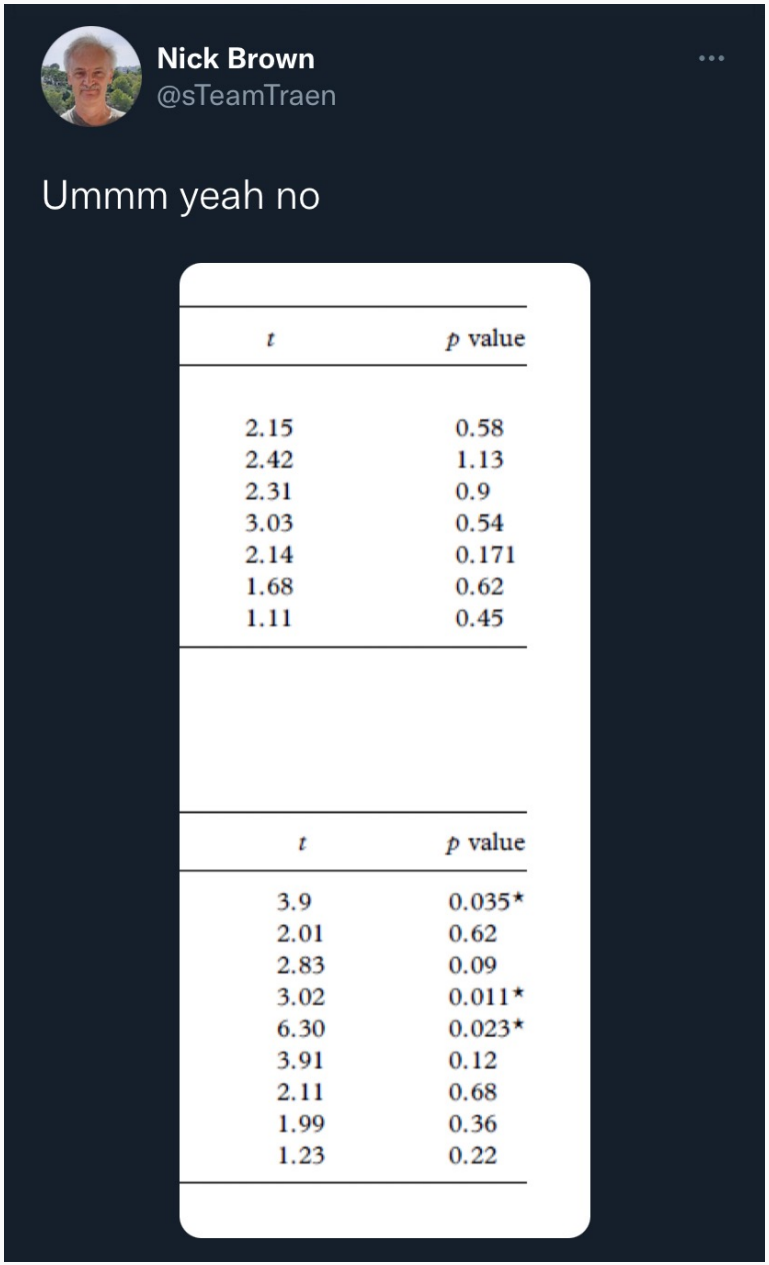
- Para teste unicaudal, acrescentar `altervative = "greater"` (ou "less")
  - Tempo de reação por idade poderia ser unicaudal (mas o valor de p já deu tão baixo que não faria diferença dividir por 2)
  - Tempo de reação por CV não se justifica ser unicaudal (não tem motivação teórica para isso)
- Para teste t pareado (mesmo grupo em 2 momentos diferentes), acrescentar `paired = T`
- Para test t "padrão" (não de Welch), acrescentar `var.equal = T`

valor de p

Vamos simular a probabilidade de cometer um erro de tipo II (ignorar um efeito) em um teste t com base no tamanho (e dispersão) da diferença, e o tamanho da amostra:

https://guilherme.shinyapps.io/nhst/

valor de p



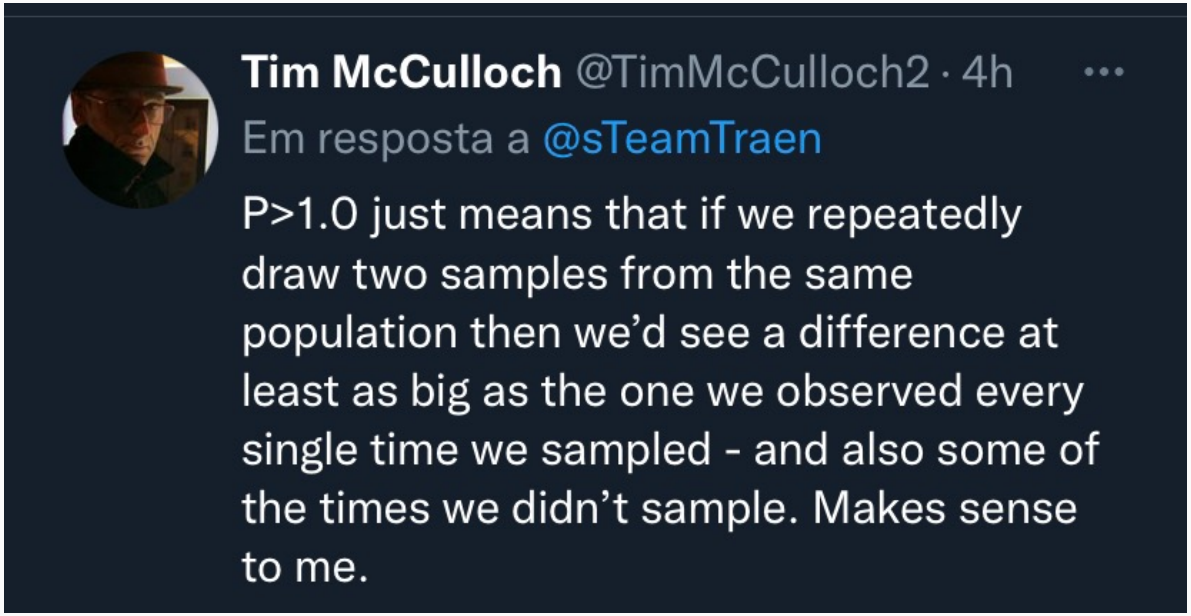
Nick Brown
@sTeamTraen

Ummm yeah no

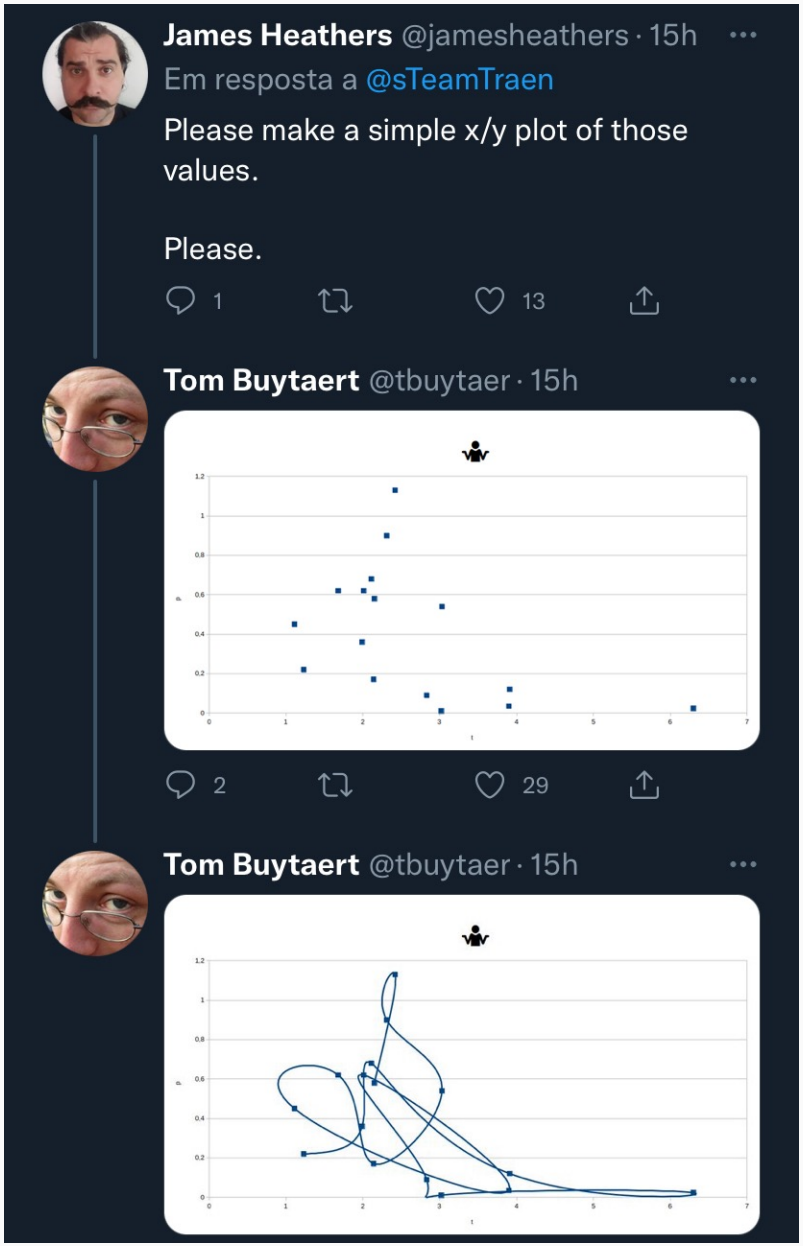
t	p value
2.15	0.58
2.42	1.13
2.31	0.9
3.03	0.54
2.14	0.171
1.68	0.62
1.11	0.45

t	p value
3.9	0.035*
2.01	0.62
2.83	0.09
3.02	0.011*
6.30	0.023*
3.91	0.12
2.11	0.68
1.99	0.36
1.23	0.22

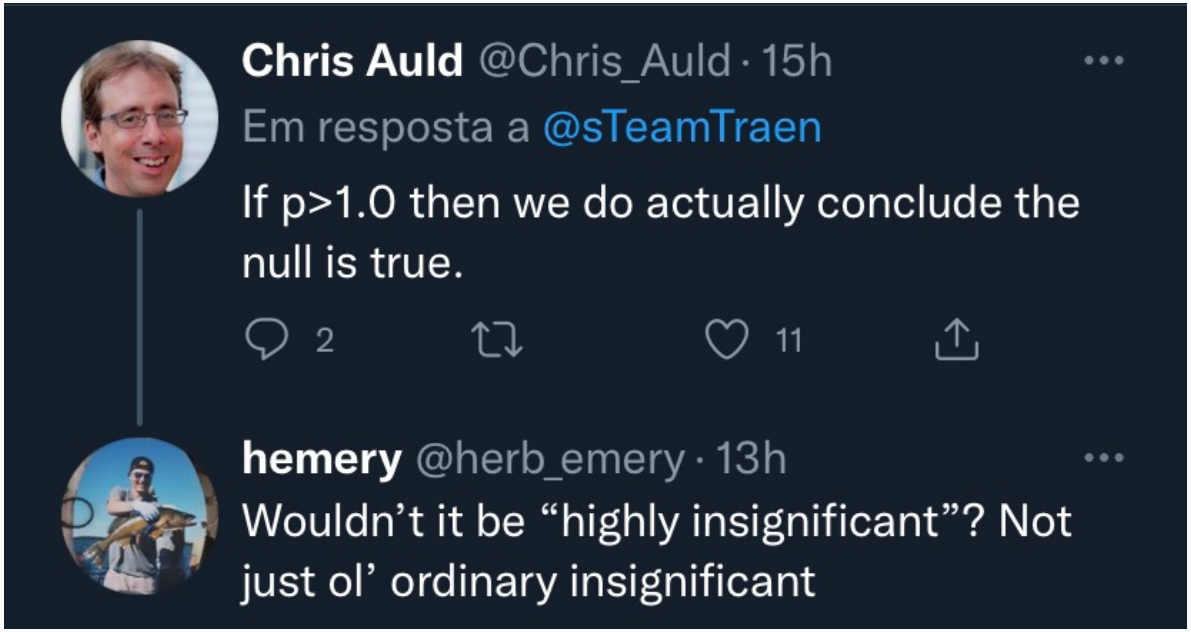
valor de p


Tim McCulloch @TimMcCulloch2 · 4h
Em resposta a @sTeamTraen
P>1.0 just means that if we repeatedly draw two samples from the same population then we'd see a difference at least as big as the one we observed every single time we sampled - and also some of the times we didn't sample. Makes sense to me.

valor de p


James Heathers @jamesheathers · 15h
Em resposta a @sTeamTraen
Please make a simple x/y plot of those values.
Please.
1
13
Tom Buytaert @tbuytaer · 15h
2
29
Tom Buytaert @tbuytaer · 15h

valor de p


Chris Auld @Chris_Auld · 15h
Em resposta a @sTeamTraen
If p>1.0 then we do actually conclude the null is true.
2
11
hemery @herb_emery · 13h
Wouldn't it be "highly insignificant"? Not just ol' ordinary insignificant

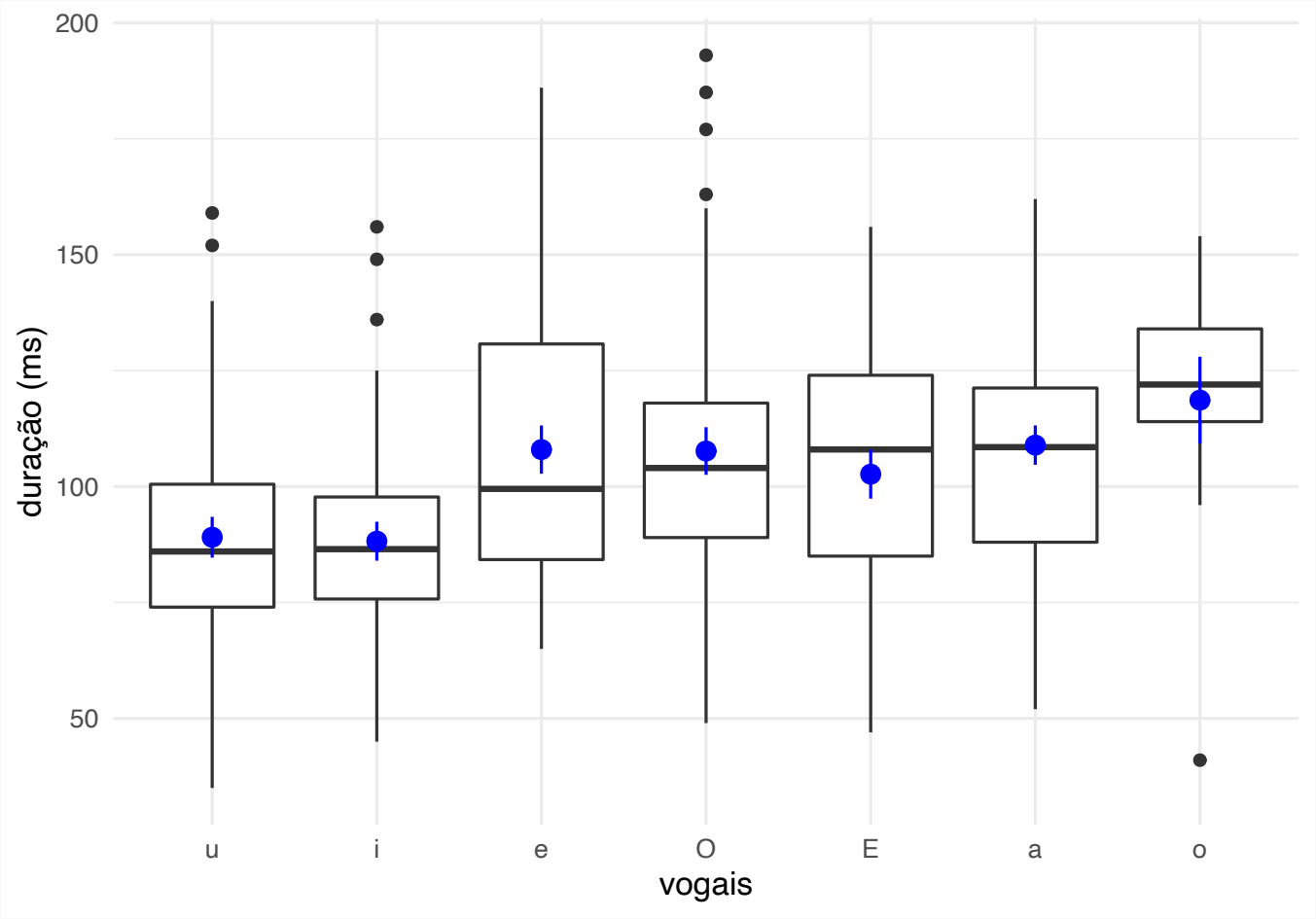
ANOVA

ANOVA - Análise de Variância

- Testar a diferença entre mais de duas médias
  - Médias de mais de dois grupos diferentes
  - Médias do mesmo grupo em mais de 2 momentos distintos (`paired = T`)
- Variável resposta contínua
- Variável preditora categórica com mais de 2 níveis
- Trata-se de um tipo específico de modelo linear
- Serve para 2 níveis também (teste depois rodar os testes t dos exemplos anteriores com uma ANOVA)

Exemplo 1

Dados “vogaisPB”, duração (dur) de cada uma das 7 vogais orais tônicas do português do Brasil



Exemplo 1

```
vogais %>%
  group_by(vogal) %>%
  summarize(meanDur = mean(dur)) %>%
  arrange(meanDur)

  vogal meanDur
  <chr>   <dbl>
1 i        88.2
2 u        89.1
3 E       103.
4 O       108.
5 e       108.
6 a       109.
7 o       119.
```

- H_0: As vogais têm a mesma duração
- H_1: As vogais têm durações diferentes

Exemplo 1

```
anovaVogais = aov(dur ~ vogal, data = vogais)
summary(anovaVogais)

             Df Sum Sq Mean Sq F value  Pr(>F)
vogal         6  21510    3585   4.249 0.00045 ***
Residuals   223 188154     844
```

- $f(6) = 4.3, p < 0.001$
- ~~H_0: As vogais têm a mesma duração~~
- H_1: As vogais têm durações diferentes
- Mas entre quais vogais estão as diferenças?

→ Precisamos de um teste pareado que ajuste o nível de significância (alfa) para diminuir a probabilidade de erro de Tipo I por repetição múltipla de testes

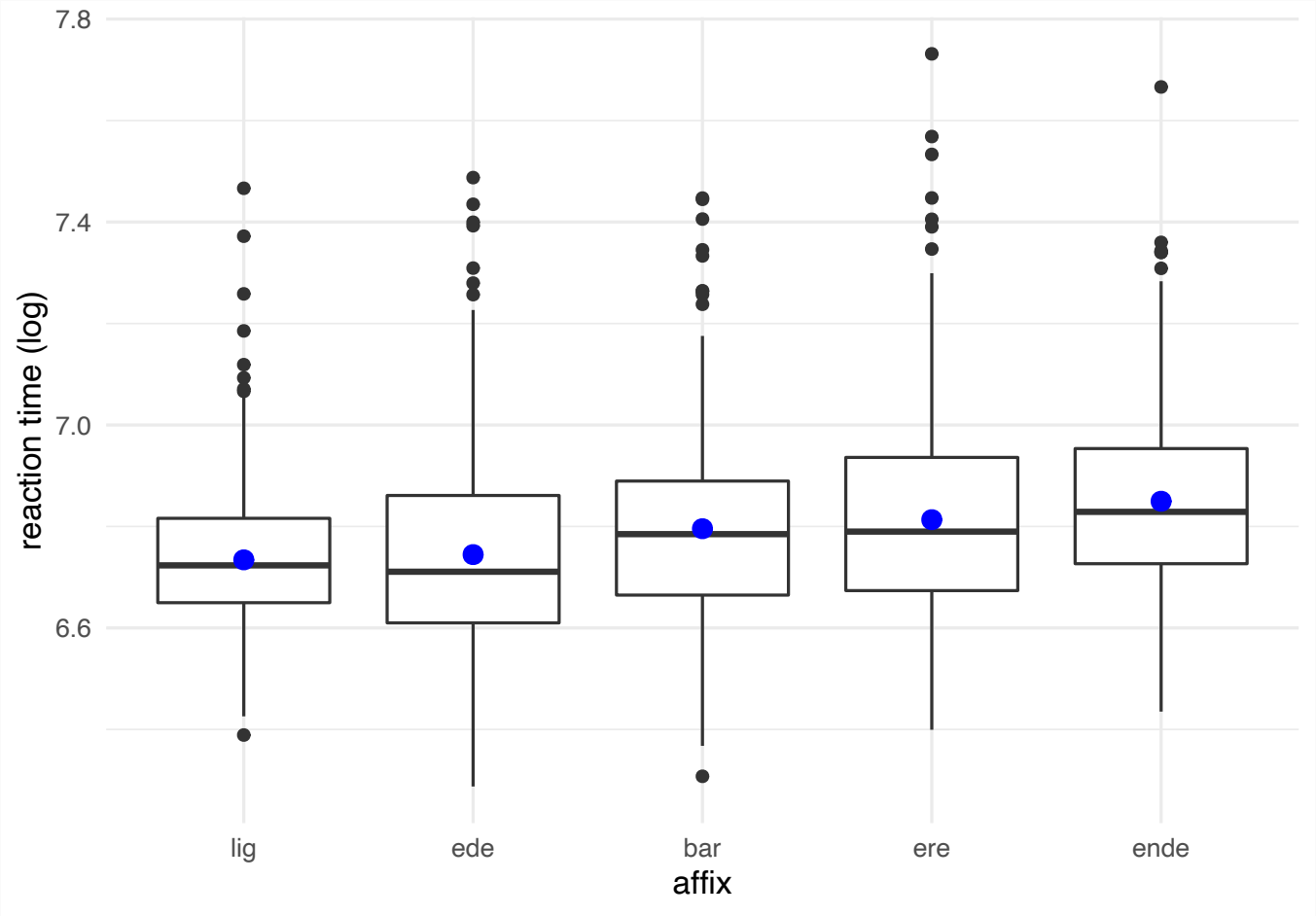
Exemplo 1

```
TukeyHSD(anovaVogais)
```

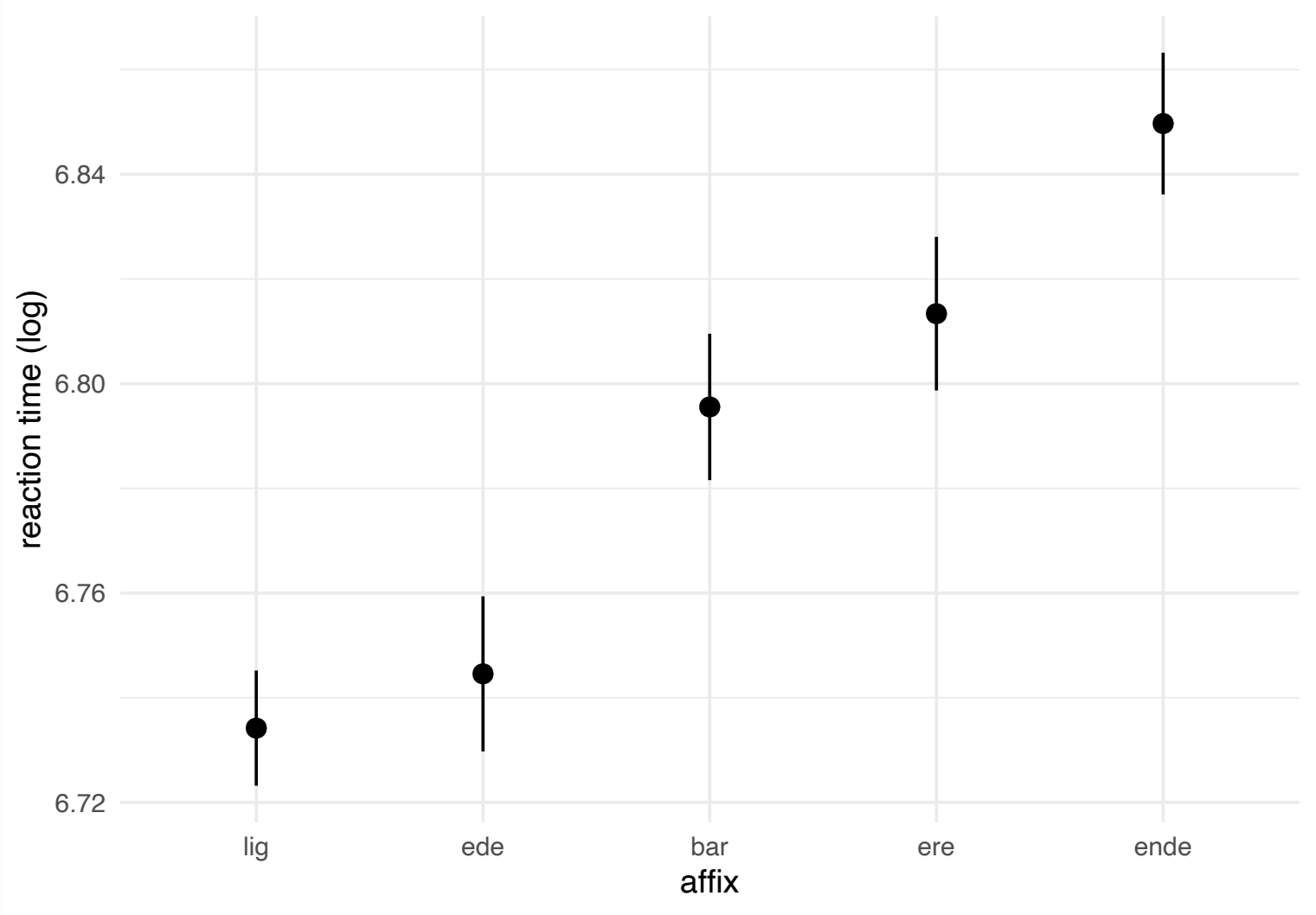
	diff	lwr	upr	p adj
e-a	-0.9638158	-21.70054090	19.772909322	0.9999994
E-a	-6.2708333	-28.85619621	16.314529542	0.9820394
i-a	-20.7269737	-41.46369880	0.009751428	0.0501953
o-a	9.6988636	-20.50905718	39.906784450	0.9627739
O-a	-1.2789634	-21.66592247	19.107995639	0.9999964
u-a	-19.8677326	-40.04578995	0.310324830	0.0567340
E-e	-5.3070175	-27.06110830	16.447073213	0.9908842
i-e	-19.7631579	-39.59125959	0.064943799	0.0513742
o-e	10.6626794	-18.92887897	40.254237822	0.9356272
O-e	-0.3151476	-19.77716057	19.146865320	1.0000000
u-e	-18.9039168	-38.14698951	0.339155969	0.0577787
i-E	-14.4561404	-36.21023111	7.297950406	0.4320733
o-E	15.9696970	-14.94546397	46.884857913	0.7221627
O-E	4.9918699	-16.42907301	26.412812843	0.9928681
u-E	-13.5968992	-34.81912098	7.625322531	0.4782805
o-i	30.4258373	0.83427892	60.017395717	0.0394154
O-i	19.4480103	-0.01400268	38.910023215	0.0502991
u-i	0.8592411	-18.38383161	20.102313863	0.9999995
O-o	-10.9778271	-40.32534244	18.369688336	0.9236470
u-o	-29.5665962	-58.76937962	-0.363812767	0.0450558
u-O	-18.5887691	-37.45440402	0.276865738	0.0564100

Exemplo 2

Dados "danish" (do pacote `languageR`), tempo de reação (logRT) de palavras com 5 afixos diferentes (`bar, ende, ede, ere, lig`)



Exemplo 2



Exemplo 2

```
danAnova = aov(LogRT ~ Affix, data = danish)
summary(danAnova)
```

```
              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
```

```
TukeyHSD(danAnova)
```

```
                 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
```

→ Note o resultado de `ede-bar`: $p = 0.06$

valor de p

- E se, de início, tivéssemos interessados apenas no contraste ede-bar, por alguma motivação teórica?
- Poderíamos rodar um teste t:

```
edebar = danish %>%
  filter(Affix %in% c("ede", "bar"))

t.test(LogRT ~ Affix, data = edebar)

        Welch Two Sample t-test

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

$\rightarrow p = 0.01$

valor de p

- Fragilidade, simplismo, reducionismo do valor de p como fator único de decisão
- Dicotomia “significativo” vs “não significativo” não é realista
- Margem para erros de análise
- Margem para *p-hacking*

Testes não paramétricos

Assumptions

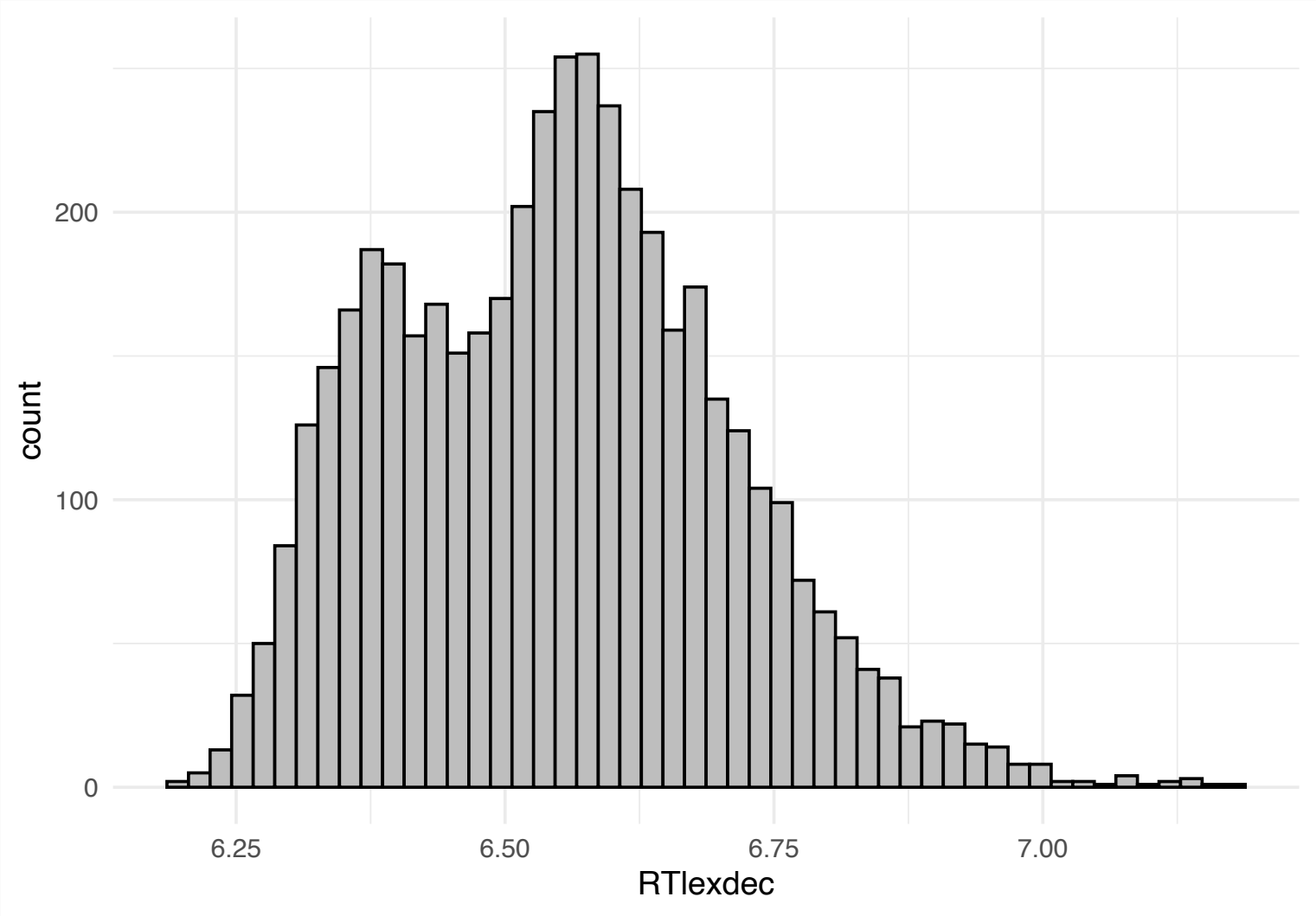
- Pressupostos/Requisitos (*assumptions*) de testes paramétricos
  - Distribuição normal dos dados (variável resposta)
  - Homocedasticidade (homogeneidade de variância)

- Para verificar a normalidade da distribuição dos dados:
  - Histograma
  - Teste de Shapiro

- Para verificar a homocedasticidade:
  - Teste de Levene

Distribuição normal?

Tempo de reação dos dados “english”



Distribuição normal?

- Teste de Shapiro

→ H_0: dados têm de uma distribuição normal

  - $p > 0.05$ = teste paramétrico
  - $p < 0.05$ = teste não paramétrico

```
shapiro.test(english$RTlexdec)

        Shapiro-Wilk normality test

data:  english$RTlexdec
W = 0.98595, p-value < 2.2e-16
```

Homogeneidade de variância?

- Teste de Levene

→ H_0: Todas as variâncias são iguais

  - $p > 0.05$ = teste paramétrico
  - $p < 0.05$ = teste não paramétrico

```
1 library(car)

2 leveneTest(RTlexdec ~ AgeSubject, english)

3 Levene's Test for Homogeneity of Variance (center = median)
4         Df F value  Pr(>F)
5 group    1  5.1615 0.02314 *
```

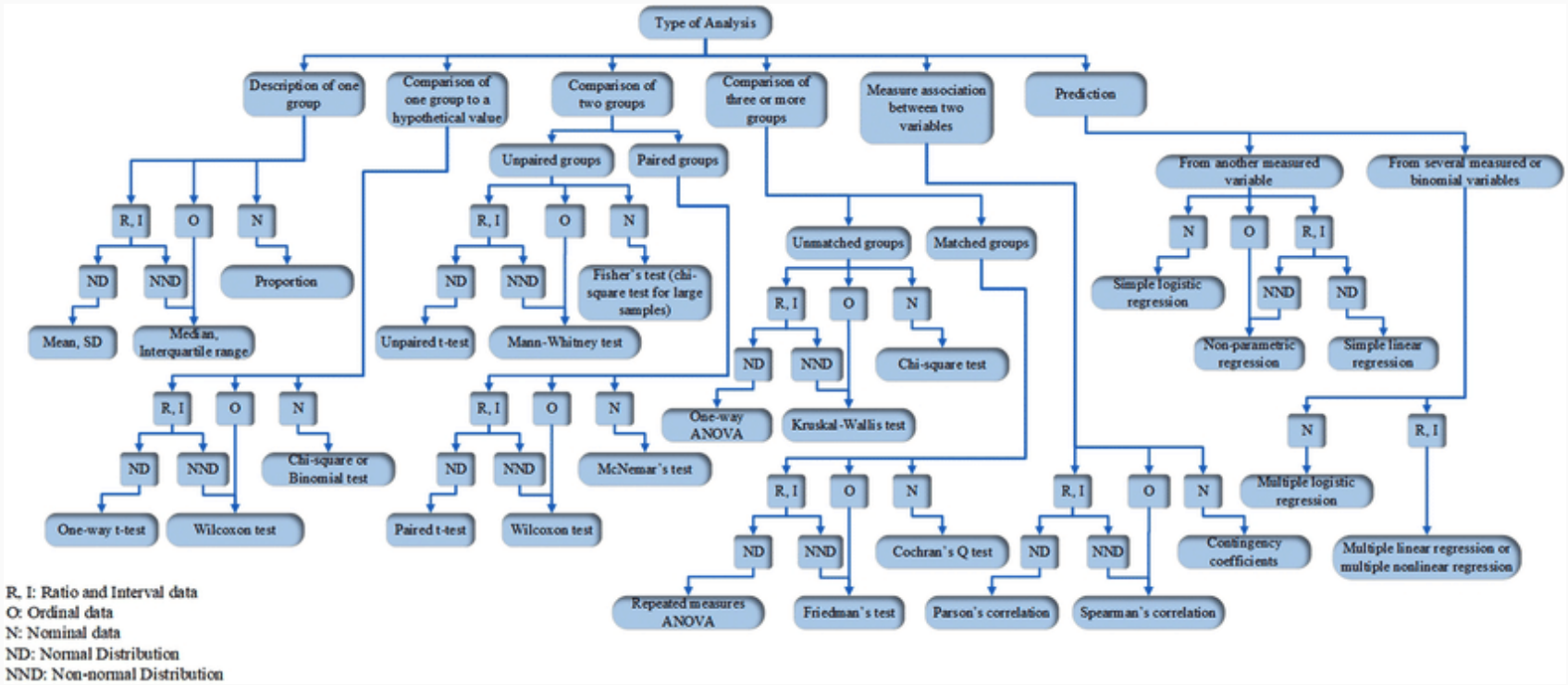
Versões não paramétricas do teste t

- 2 grupos independentes: Wilcoxon test
  → `wilcox.test()`
- 2 grupos pareados: Mann-Whitney (Wilcoxon rank-sum test)
  → `wilcox.test( , paired = T)`

- Mais de 2 grupos independentes: Kruskal-Wallis
  → `kruskal.test()`
- Mais de 2 grupos pareados: Friedman test
  → `friedman.test()`
- `pairwise.wilcox.test()` para verificar as comparações pareadas

- Obs! O teste-t padrão do R (Welch) não assume variância homogênea
  - O padrão é `var.equal = F`
  - Para rodar um teste t "clássico", acrescentar `var.equal = T`

Statistical Test Selection (Crazy!)



Tamanho do efeito

Tamanho do efeito

- Cohen's d para teste t
- Eta-squared para ANOVA

```
library(effectsize)

cohens_d(RTlexdec ~ AgeSubject, data = english)

Cohen's d |       95% CI
------------------------
2.00      | [1.76, 2.31]
```

- $d < 0.2$ "negligible"
- $d < 0.5$ "small"
- $d < 0.8$ "medium"
- $d > 0.8$ "large"

Tamanho do efeito

```
eta_squared(anovaVogais)

Parameter | Eta2 |       95% CI
------------------------------
vogal     | 0.10 | [0.03, 1.00]
```

- $\eta^2 = 0.01$ “small”
- $\eta^2 = 0.06$ “medium”
- $\eta^2 = 0.14$ “large”

Perguntas?