

CURSO R - 2023



—
NAIARA ALCANTARA





CURSO R - 2023

7- ANÁLISES INFERENCIAIS

1º Análise bivariada

Testes paramétricos e não paramétricos

NAIARA ALCANTARA

TÓPICOS

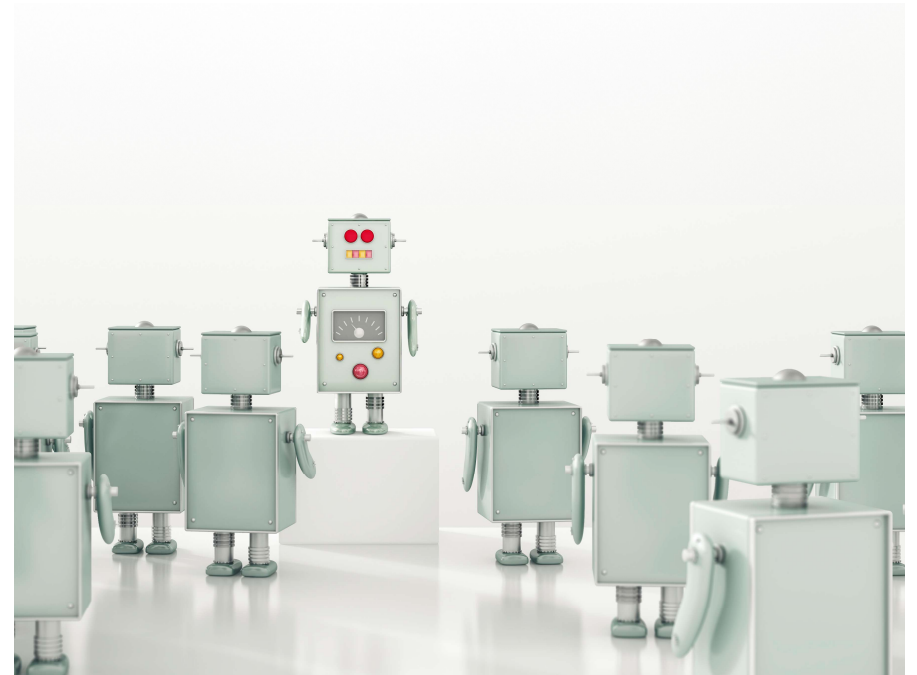
- 1-INTRODUÇÃO AO R
- 2- ANÁLISE EXPLORATÓRIA E MANIPULAÇÃO DOS DADOS
- 3- SALVAMENTO E ABERTURA
- 4- ANÁLISES DESCRITIVAS
- 5- APRESENTAÇÃO GRÁFICA
- 6- PROCESSAMENTO DE DADOS
- 7- ANÁLISES INFERENCIAIS

Análise bivariada

Teste de correlação

Teste de regressão linear
simples e múltiplo

Teste de regressão
logística simples e múltiplo



ANÁLISES INFERENCIAIS: BIVARIADA

- A primeira atividade que faremos, antes mesmo de entender o que é uma análise de dados bivariada, é aprender a usar mais um tipo de base de dados (dados eleitorais).
- Então já aprendemos a baixar dados sobre opinião pública lá do site do LAPOP, no tópico 3, e trabalhamos com esses dados nos tópicos seguintes (descrevendo e analisando graficamente).
- Agora usaremos dados do TSE. Dessa forma, não precisaremos justificar teoricamente nossas escolhas a todo momento.
- Isso é muito importante porque a escolha do material empírico deve ser feita com base na teoria.
- Usaremos a base de dados para senador, cuja eleições ocorrem a cada 4 anos.
- Quem não se lembra quais são os tipos de variáveis é indicado de busque os tópicos iniciais do curso.

ANÁLISES INFERENCIAIS: BIVARIADA

- O que é uma amostra ?
É uma pequena fração do universo

Amostra NÃO é:
Mínimo de 10% da população

Amostragem probabilística

- Amostra Aleatória simples
- Amostra Sistemática
- Amostra Estratificada
- Amostra por Conglomerado



Amostragem não-probabilísticas

- Amostra por julgamento
- Amostra por cotas
- Amostra bola de neve
- Amostra desproporcional




Para entender mais sobre amostra, ler: Métodos quantitativos para iniciantes

<https://cpop.ufpr.br/publicacoes-cpop/>

ANÁLISES INFERENCIAIS: BIVARIADA

- Para realizar uma análise inferencial nós temos que utilizar amostras que sejam estatisticamente representativas da população amostrada/Universo.
- Não é possível realizar análises e fazer inferências sobre amostras que não podem ser extrapoladas para o Universo.
- Em uma pesquisa sempre devemos partir de uma hipótese de pesquisa (H_0)-hipótese nula-de que não existe uma relação estatística entre as questões que estamos estudando. Se for possível rejeitar essa hipótese nula, poderemos confirmar sua hipótese contrária (H_1).

Exemplo:

- Minha hipótese de pesquisa é de anos de escolaridade influenciam no aumento médio da renda da população.
 - Se eu confirmar essa hipótese de pesquisa (H_1), estarei rejeitando a hipótese de que anos de escolaridade não aumentam a média de renda da população (H_0).
- 

ANÁLISES INFERENCIAIS: BIVARIADA

- Para saber se a hipótese nula deve ser confirmada ou rejeitada, devemos realizar testes inferenciais.
- Esses testes irão fornecer um valor de probabilidade que irá indicar se a H_0 deve ou não ser rejeitada.
- Em geral devemos rejeitar a H_0 , quando p for < 0.05

Probabilidade de se obter um resultado igual (ou mais extremo) que o obtido, dado que a hipótese nula é verdadeira

$P < 0.001$ Altamente significativa
 $p < 0.01$ Razoavelmente significativa
 $p \leq 0.05$ Pouco significativa
 $p > 0.05$ pouca evidência de existência de significância.



ANÁLISES INFERENCIAIS: BIVARIADA

Teste qui- quadrado

Apropriado para variáveis qualitativas/categóricas não ordenadas, que também podem ser dicotômicas, isto é, com apenas 2 respostas válidas (exemplo de variáveis que podem ser utilizadas nesse teste: sexo, gosto musical, cor dos olhos, preferencias,...)

Assim como para qualquer teste estatístico inferencial é interessante que as amostras sejam relativamente grandes.


Esse teste somente indica se existe um relacionamento estatisticamente significativo entre as variáveis, portanto não é possível saber a direção da associação.

Teste não paramétrico: não depende de parâmetros populacionais (média e variância)



ANÁLISES INFERENCIAIS: BIVARIADA

```
##Teste de qui-quadrado####  
#Diferença de média entre eleitos e não eleitos por cor e dps sexo|  
Teste1 ← table(BASE_SEN_2022$DS_SIT_TOT_TURN0,  
                BASE_SEN_2022$DS_COR_RACA)  
  
chisq.test(Teste1)  
  
#  
# Pearson's Chi-squared test  
#  
# data:  Teste1  
# X-squared = 15.701, df = 5, p-value = 0.00775
```



Apesar das variáveis escolhidas cumprirem o requisito para o teste, essa base de dados não é a mais adequada para o qui-quadrado.

Por que?

Porque fere o pressuposto de terem amostras grandes e semelhantes.

Explicar melhor.

```
Teste2 ← table(BASE_SEN_2022$DS_SIT_TOT_TURN0,  
               BASE_SEN_2022$DS_GENERO)  
  
chisq.test(Teste2)  
  
# Pearson's Chi-squared test with Yates' continuity correction  
#  
# data:  Teste2  
# X-squared = 1.0059, df = 1, p-value = 0.3159
```

- o título do teste,
- quais variáveis foram usadas dentro do objeto criado
- os graus de liberdade e o p valor do teste.

O teste qui-quadrado permite apenas verificar se há relação estatística entre as variáveis, através da rejeição ou confirmação da H_0

ANÁLISES INFERENCIAIS: BIVARIADA

Teste lambda (λ)

A estatística de Goodman-Kruskal Lambda é uma medida de associação entre variáveis categóricas, especialmente quando uma das variáveis é ordinal (ou seja, tem uma ordem específica). Mas também pode ser utilizado para variáveis sem ordenação.

Essa estatística varia entre 0 e 1, onde 0 indica que não há associação, e 1 indica uma associação perfeita.

Apresenta um pouco mais de informação que o qui-quadrado, porque permite analisar também a direção da relação entre as variáveis testadas.

Sempre uma das variáveis será considerada dependente, ainda que ainda não tenha sido feito essa distinção.



ANÁLISES INFERENCIAIS: BIVARIADA

```
#Antes de rodar o teste, vamos atribuir os levels
BASE_SEN_2022$DS_COR_RACA <-
  factor(BASE_SEN_2022$DS_COR_RACA,
    levels = c("PRETA", "PARDA", "INDÍGENA",
      "NÃO INFORMADO", "AMARELA", "BRANCA"))

levels(BASE_SEN_2022$DS_COR_RACA)

BASE_SEN_2022$DS_GENERO <-
  factor(BASE_SEN_2022$DS_GENERO,
    levels = c("FEMININO", "MASCULINO"))

levels(BASE_SEN_2022$DS_GENERO)
```

- Não apresenta o valor de p, por isso deve ser feito após o qui-quadrado
- Resultado positivo na coluna, isto é, existe uma associação entre a cor e o sucesso eleitoral que é positivo.

```
lambda.test(Teste1)
```

```
# $row
# [1] 0
#
# $col
# [1] 0.07407407
```

```
lambda.test(Teste2)
```

```
# $row
# [1] 0
#
# $col
# [1] 0
```

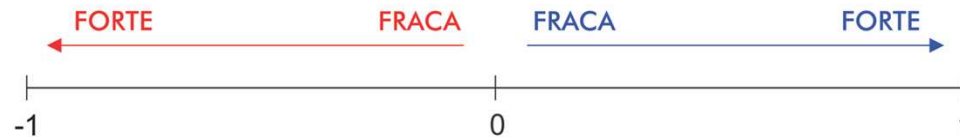
ANÁLISES INFERENCIAIS: BIVARIADA

Gamma (γ)

Apropriado para variáveis qualitativas e ordenadas

Também considera que uma das variáveis é dependente e a outra independente

Apresenta o valor de ρ , e a direção da associação que varia entre:



ANÁLISES INFERENCIAIS: BIVARIADA

```
#Teste de Gamma (Y)
install.packages("vcdExtra")
library(vcdExtra)

#Vamos criar uma variável chamada
#Satisfação com a vida, somente para rodar o teste
#Vamos criá-la, a partir do estdo civil

# table(BASE_SEN_2022$DS_ESTADO_CIVIL)
# CASADO(A)          DIVORCIADO(A)  SEPARADO(A)  JUDICIALMENTE
# 133                36              4
# SOLTEIRO(A)        VIÚVO(A)
# 27                  5
```

```
#1= será o totalmente insatisfeito e 5=totalmente satisfeito
library(memisc)
BASE_SEN_2022$SatVida ← recode(BASE_SEN_2022$DS_ESTADO_CIVIL,
                                "Totalmente insatisfeito" ← "VIÚVO(A)",
                                "Insatisfeito" ← "SEPARADO(A) JUDICIALMENTE",
                                "Meio termo" ← "DIVORCIADO(A)",
                                "Satisfeito" ← "SOLTEIRO(A)",
                                "Totalmente satisfeito" ← "CASADO(A)")

table(BASE_SEN_2022$SatVida)
BASE_SEN_2022$SatVida ← as.character(BASE_SEN_2022$SatVida)
DS_GRAU_INSTRUCAO ← as.character(DS_GRAU_INSTRUCAO)

tab2 ← table(BASE_SEN_2022$SatVida,
              BASE_SEN_2022$DS_GRAU_INSTRUCAO)

GKgamma(tab2)
# gamma      : -0.371
# std. error  : 0.041
# CI          : -0.451 -0.291
```

Interpretação:

1º Verificar se o p é $< 0,05$

Como não é, a hipótese nula não pode ser rejeitada.

2º Com a não rejeição da hipótese nula, não iremos interpretar o resultado do teste.

Caso a H_0 fosse $< 0,05$, então interpretaríamos o valor do teste para saber se a relação é positiva ou negativa

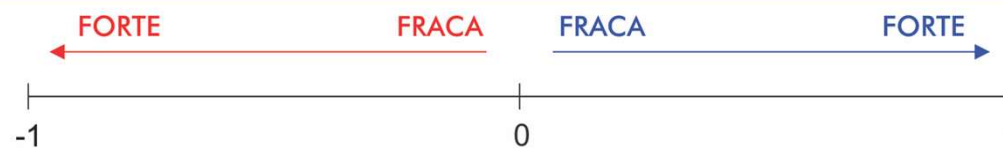
```
chisq.test(tab2, simulate.p.value = T) #Apenas para simular o valor de p
# Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)
#
# data:  tab2
# X-squared = 26.001, df = NA, p-value = 0.3208
```

ANÁLISES INFERENCIAIS: BIVARIADA

Kendall Tau

O teste de Kendall Tau, também conhecido como coeficiente de concordância de Kendall (ou apenas Kendall's tau), é uma medida estatística utilizada para avaliar o grau de concordância ou associação entre classificações ou rankings de duas variáveis. Especificamente, ele mede a correlação entre as ordens dos pares de observações em duas variáveis.

Bastante parecido com o teste de Gamma, isto é, apropriado para variáveis qualitativas ordenadas
Porém apresenta o valor do qui-quadrado junto com teste
Sua estimativa também varia entre:



ANÁLISES INFERENCIAIS: BIVARIADA

```
#Teste de Kendal####  
install.packages("Kendall")  
library(Kendall)  
  
#Utilização da variável criada para realização do teste de gamma  
  
library(memisc)  
BASE_SEN_2022$SatVida <- recode(BASE_SEN_2022$DS_ESTADO_CIVIL,  
                                "Totalmente insatisfeito" <- "VIÚVO(A)",  
                                "Insatisfeito" <- "SEPARADO(A) JUDICIALMENTE",  
                                "Meio termo" <- "DIVORCIADO(A)",  
                                "Satisfeito" <- "SOLTEIRO(A)",  
                                "Totalmente satisfeito" <- "CASADO(A)")  
  
table(BASE_SEN_2022$SatVida)|  
BASE_SEN_2022$SatVida <- as.factor(BASE_SEN_2022$SatVida)  
BASE_SEN_2022$DS_GRAU_INSTRUCAO <- as.factor(BASE_SEN_2022$DS_GRAU_INSTRUCAO)  
  
Kendall(BASE_SEN_2022$SatVida,  
        BASE_SEN_2022$DS_GRAU_INSTRUCAO)  
# tau = 0.0245, 2-sided pvalue =0.7069
```

Interpretação:

1º Tau = 0.0245 - sugere uma correlação muito fraca entre as duas variáveis testadas.

2º p-value = 0.7069 é bastante alto, o que indica que não há evidência estatisticamente significativa para rejeitar a hipótese nula de que não há correlação entre as variáveis.

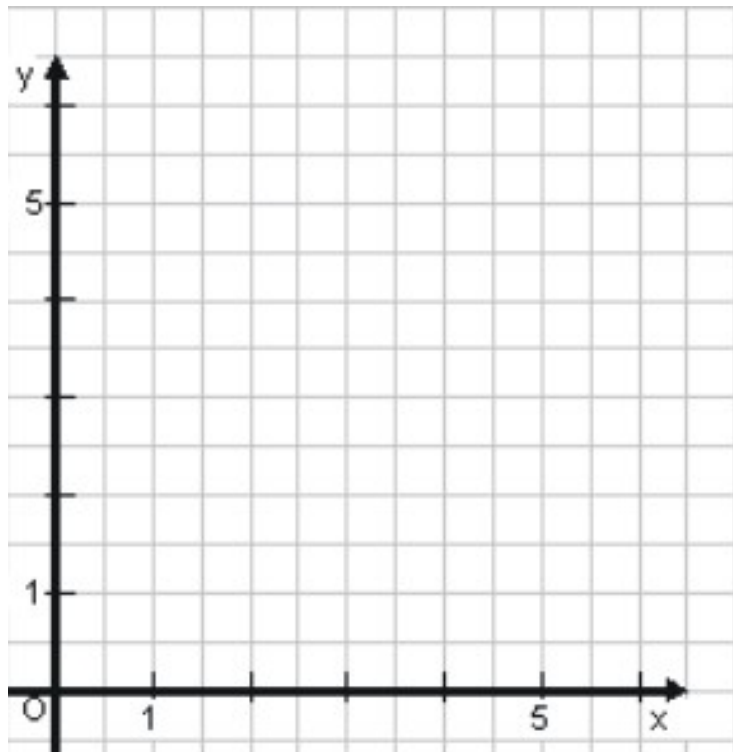


ANÁLISES INFERENCIAIS: BIVARIADA

! NORMALIDADE DOS DADOS

Todos os testes de aprendemos até agora não dependiam de uma distribuição normal dos dados.

Porque os testes de normalidade se aplicam somente a variáveis quantitativas



- Quando analisamos variáveis qualitativas ou quantitativas que não possuem distribuição normal, utilizamos os testes não paramétricos, como os que eu ensinei ou então: teste de Mann-Whitney, Kruskal-Wallis, Teste de Friedman.
- Quando analisamos variáveis quantitativas com distribuição normal podemos utilizar o teste t independente, teste t pareado, teste de anova

ANÁLISES INFERENCIAIS: BIVARIADA

Teste T

O teste t de Student (ou simplesmente teste t) compara duas médias e mostra se as diferenças entre essas médias são significativas

Permitindo que você avalie se essas diferenças ocorreram por um mero por acaso ou não.

Exemplos de quando é interessante comparar médias:

Para amostras diferentes (independentes)

- Quero saber se existe diferença de médias de votos em homens e mulheres
- Quero saber se existe diferença de média entre as notas de crianças de escolas públicas e privadas
- Quero saber se existe diferença de média entre desenvolvimento de doenças entre fumantes e não fumantes

Para mesma amostra (pareadas)

- Quero saber se existe diferença de média entre as notas dos alunos no 1º semestre e 3º semestre
- Quero saber se existe diferença de média para o exame de uma doença x, antes e depois de tomar uma determinada substância
- Quero saber se existe diferença de média entre um grupo antes e depois de assistirem uma palestra

Muitas vezes a gente acha que existe uma diferença óbvia, mas como saber se existe mesmo? Se não é somente uma percepção construída por influência social?

Através de um teste estatístico



ANÁLISES INFERENCIAIS: BIVARIADA

Teste T não pareado –APROPRIADO PARA AMOSTRAS INDEPENDENTES

```
1257 #Teste T####
1258
1259 options(scipen = 999, digits = 1)
1260
1261 HOMEM ← subset(BASE_SEN_2022, CD_GENERO == 2)
1262 summary(HOMEM$TOTAL_VOTOS)
1263
1264 MULHER ← subset(BASE_SEN_2022, CD_GENERO == 4)
1265 summary(MULHER$TOTAL_VOTOS)
1266
1267
1268 t.test(MULHER$TOTAL_VOTOS, HOMEM$TOTAL_VOTOS)
1269 #
1270 # Welch Two Sample t-test
1271 #
1272 # data: MULHER$TOTAL_VOTOS and HOMEM$TOTAL_VOTOS
1273 # t = -2, df = 203, p-value = 0.02
1274 # alternative hypothesis: true difference in means is not equal to 0
1275 # 95 percent confidence interval:
1276 # -506729 -35829
1277 # sample estimates:
1278 # mean of x mean of y
1279 # 280697 551975
```

Primeira linha: Nome do teste que foi feito

O teste-t de Welch é uma adaptação do teste t de Student, que é mais confiável quando as duas amostras têm variâncias desiguais e tamanhos de amostra desiguais.

A segunda linha informa de onde foram extraídos os dados para o teste:

Total de votos de homens e mulheres.

A terceira apresenta: o valor da estatística t: $t = -2$ /os graus de liberdade da curva de distribuição t: $df = 203$ / o valor de p: $p\text{-value} = 0.02$

A quarta linha informa qual a hipótese alternativa do teste: true difference in means is not equal to 0

A hipótese alternativa é que as médias das amostras são diferentes. Podemos então inferir que a hipótese nula do teste é que as médias das amostras são iguais. Essa hipótese foi rejeitada.

A quinta e a sexta linha informam o intervalo de confiança do teste.

A sétima, oitava e nona linha informam os dados da amostra: a média de cada amostra 280.697 e 551.975

ANÁLISES INFERENCIAIS: BIVARIADA

Teste T pareado –APROPRIADO PARA AMOSTRAS DEPENDENTES

Como nossa base de dados do senado não tem nenhuma variável que seja comparada entre o tempo em relação ao mesmo grupo, iremos criar uma pequena base de dados com dados inventados sobre a média da quantidade de mulheres candidatas ao senado.

Teremos uma média para o ano de 2020 e uma para o ano 2050, são dados inventados, por isso as datas também são inventadas.

```
#Criação da base de dados para a realização do teste t
BaseHipSen ← data.frame(
  "Regiões" = c("Norte", "Nordeste", "Centro-Oeste", "Sudeste",
               "Sul", "Argentina"),
  "Can_2020" = c(5, 10, 7, 6, 11, 8),
  "Can_2050" = c(25, 15, 4, 12, 30, 17)
)
View(BaseHipSen)
```

ANÁLISES INFERENCIAIS: BIVARIADA

Teste T pareado –APROPRIADO PARA AMOSTRAS DEPENDENTES

```
#Realização do teste
t.test(BaseHipSen$Can_2020, BaseHipSen$Can_2050, paired = T)

#saída do teste

# Paired t-test
#
# data:  BaseHipSen$Can_2020 and BaseHipSen$Can_2050
# t = -2.5908, df = 5, p-value = 0.04879
# alternative hypothesis: true mean
# difference is not equal to 0
# 95 percent confidence interval:
#   -18.59377288  -0.07289378
# sample estimates:
#   mean difference
#   -9.333333
```

A interpretação é muito semelhante ao do teste não pareado
Primeira linha: Nome do teste que foi feito, indicando que é pareado

A segunda linha informa de onde foram extraídos os dados para o teste

A terceira apresenta: o valor da estatística t: $t = -2$ /os graus de liberdade da curva de distribuição t: $df = 5$ / o valor de p: $p\text{-value} = 0.04$

A quarta linha informa qual a hipótese alternativa do teste: true difference in means is not equal to 0

A hipótese alternativa é que as médias das amostras são diferentes. Podemos então inferir que a hipótese nula do teste é que as médias das amostras são iguais. Essa hipótese foi rejeitada.

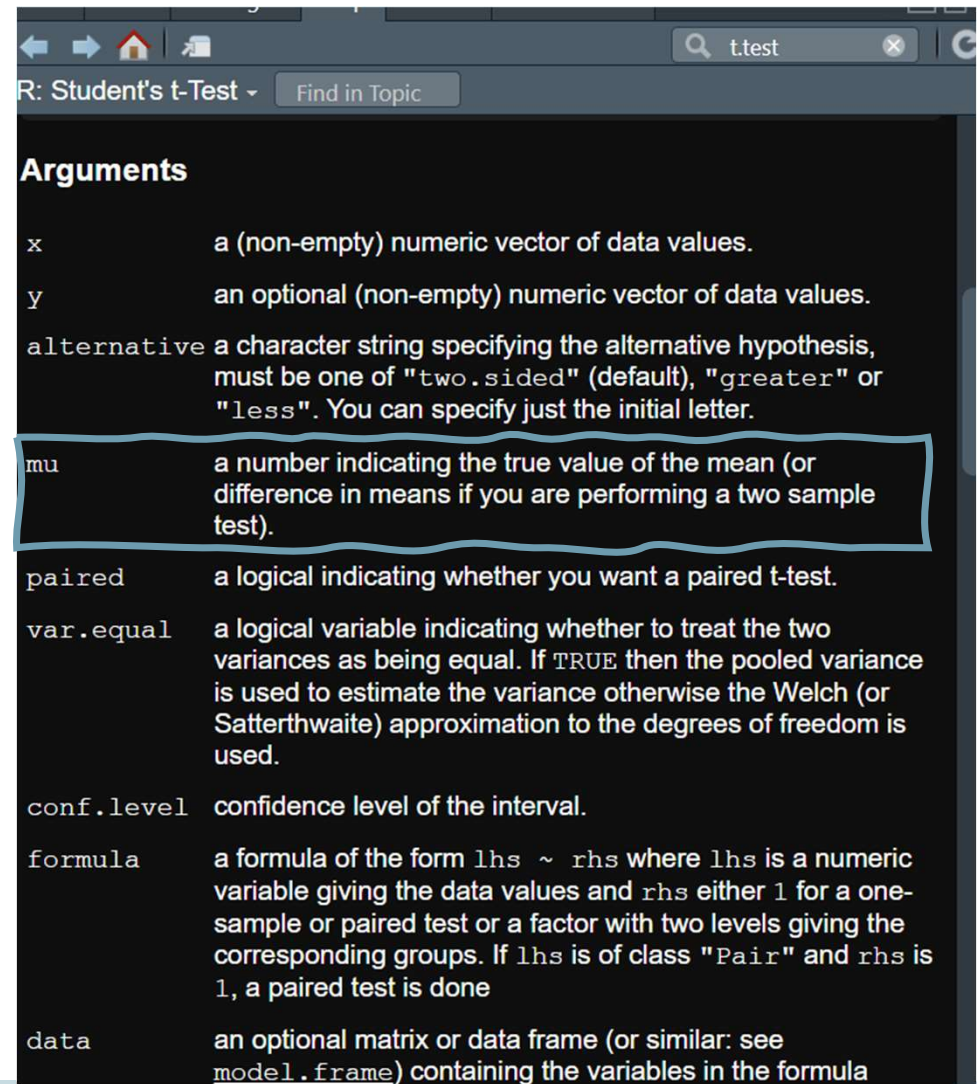
A quinta e a sexta linha informam o intervalo de confiança do teste.

A sétima, oitava e nona linha informam a diferença de média entre uma amostra e outra que é de -9

ANÁLISES INFERENCIAIS: BIVARIADA

Teste T – Argumentos do teste

Para verificar todos os argumentos que podemos inserir em um teste t basta escrever no help “t.teste” que ele irá mostrar um modelo do teste e uma lista com todos os argumentos que podemos inserir.



R: Student's t-Test - Find in Topic

Arguments

<code>x</code>	a (non-empty) numeric vector of data values.
<code>y</code>	an optional (non-empty) numeric vector of data values.
<code>alternative</code>	a character string specifying the alternative hypothesis, must be one of "two.sided" (default), "greater" or "less". You can specify just the initial letter.
<code>mu</code>	a number indicating the true value of the mean (or difference in means if you are performing a two sample test).
<code>paired</code>	a logical indicating whether you want a paired t-test.
<code>var.equal</code>	a logical variable indicating whether to treat the two variances as being equal. If TRUE then the pooled variance is used to estimate the variance otherwise the Welch (or Satterthwaite) approximation to the degrees of freedom is used.
<code>conf.level</code>	confidence level of the interval.
<code>formula</code>	a formula of the form <code>lhs ~ rhs</code> where <code>lhs</code> is a numeric variable giving the data values and <code>rhs</code> either 1 for a one-sample or paired test or a factor with two levels giving the corresponding groups. If <code>lhs</code> is of class "Pair" and <code>rhs</code> is 1, a paired test is done
<code>data</code>	an optional matrix or data frame (or similar: see model.frame) containing the variables in the formula

ANÁLISES INFERENCIAIS: BIVARIADA

Teste T para uma amostra –compara a medida média de um grupo com a média da população.

- Vamos supor que queremos entender se a média de mulheres candidatas ao senado em 2050 é superior ou inferior que a média na Argentina.
- Nessa suposição nós sabemos qual é a média de candidatas na Argentina em 2050.

```
#Rodamos o teste:  
t.test(BaseMenor$Can_2050, mu=17)
```

```
One Sample t-test  
  
data: BaseMenor$Can_2050  
t = 0.043093, df = 4, p-value = 0.9677  
alternative hypothesis: true mean is not equal to 17  
95 percent confidence interval:  
 4.314184 30.085816  
sample estimates:  
mean of x  
 17.2
```

Primeira linha: Nome do teste que foi feito

A segunda linha informa de onde foram extraídos os dados para o teste

A terceira apresenta: o valor da estatística t: $t = 0.04$ /os graus de liberdade da curva de distribuição t: $df = 4$ / o valor de p: $p\text{-value} = 0.9$

A quarta linha informa qual a hipótese alternativa do teste: true difference in means is not equal to 0

A hipótese alternativa é que as médias das amostras são diferentes. Podemos então inferir que a hipótese nula do teste é que as médias das amostras são iguais. Essa hipótese não pode ser rejeitada

A quinta e a sexta linha informam o intervalo de confiança do teste.

A sétima, oitava e nona linha informam os dados da amostra: a média estimada de x

! Nesse caso não teria como dar significativo, por que ?