

# Curso de Pós-Graduação em Ciências Veterinárias - UFRRJ

Métodos Estatísticos

---

Prof: Wagner Tassinari

wagner.tassinari@ini.fiocruz.br

Inferência Estatística

Tamanho da Amostra

Introdução à Inferencia Estatística

Testes de Normalidade

# Tamanho da Amostra

---

## Tamanho da Amostra ( $n$ ) - Para Média

- Qual deve ser o tamanho da amostra para que se tenha um dado um erro máximo de  $|\bar{x} - \mu| \leq \varepsilon$  e uma confiança de  $100(1 - \alpha)\%$  ?

$$n = \left[ \frac{z_{(\alpha/2)} \sigma}{\varepsilon} \right]^2$$

- Correção para populações finitas ( $N \leq n$ )

$$n' = \frac{n}{1 + \frac{n}{N}}$$

## Exemplo 1:

Vamos supor que queremos amostrar alguns ratinhos da raça *Wistar* com 30 dias de idade para realizar um determinado experimento. Conforme já estabelecido pela literatura, sabemos que o peso médio desses ratos é de 65g. Determine o  $n$ , ao nível de 5% de significância, para esse experimento, supondo um erro máximo de 0,02 ( $|\bar{x} - \mu| \leq 0,02$ ) e  $\sigma = 1$ . Temos então:

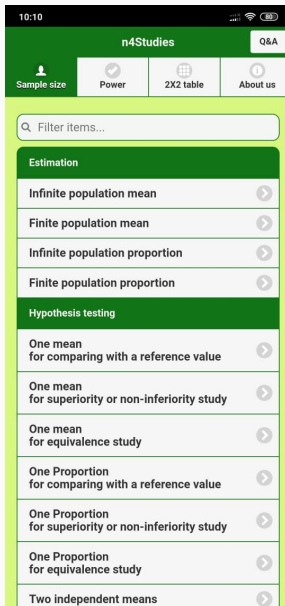
$$n = \left[ \frac{z_{(\alpha/2)}\sigma}{\varepsilon} \right]^2 = \left[ \frac{1,96 \cdot 1}{0,02} \right]^2 \simeq 9.604 \text{ ratinhos}$$

## Exemplo 2:

Mas só dispomos de uma população com 300 ratinhos experimentais.

$$n' = \frac{n}{1 + \frac{n}{N}} = \frac{9604}{1 + \frac{9604}{300}} \simeq 291 \text{ ratinhos}$$

# Solução no app n4Studies



# Solução no app n4Studies

10:57

Back n4Studies Help

Sample size Power 2X2 table About us

Formula<sup>[ref]</sup>:

$$n = \frac{N \sigma^2 z_{1-\frac{\alpha}{2}}^2}{d^2 (N-1) + \sigma^2 z_{1-\frac{\alpha}{2}}^2}$$

Population size (N) =

300

Standard deviation ( $\sigma$ ) =

1

Error (d) =

0.02

Alpha ( $\alpha$ ) =

0.05

Cluster sampling?

No

Calculate Clear

Output:

Sample size = 291



### n4Studies: Sample Size Calculation for an Epidemiological Study on a Smart Device

Chetta Ngamjarus, M.Sc.\*,\*\*, Virasakdi Chongsuvivatwong, Ph.D.\*, Edward McNeil, M.Sc.\*

\* Epidemiology Unit, Faculty of Medicine, Prince of Songkla University, Songkhla 90110, \*\*Department of Biostatistics and Demography, Khon Kaen University, Khon Kaen 40002, Thailand.

#### ABSTRACT

**Objective:** This study was to develop a sample size application (called “n4Studies”) for free use on iPhone and Android devices and to compare sample size functions between n4Studies with other applications and software.

**Methods:** Objective-C programming language was used to create the application for the iPhone OS (operating system) while JavaScript, jquery mobile, PhoneGap and jstat were used to develop it for Android phones. Other sample size applications were searched from the Apple app and Google play stores. The applications’ characteristics and sample size functions were collected. Spearman’s rank correlation was used to investigate the relationship between number of sample size functions and price.

**Results:** “n4Studies” provides several functions for sample size and power calculations for various epidemiological study designs. It can be downloaded from the Apple App and Google play store. Comparing n4Studies with other applications, it covers several more types of epidemiological study designs, gives similar results for estimation of infinite/finite population mean and infinite/finite proportion from GRANMO, for comparing two independent means from BioStats, for comparing two independent proportions from EpiCal application. When using the same parameters, n4Studies gives similar results to STATA, epicalc package in R, PS, G\*Power, and OpenEpi.

**Conclusion:** “n4Studies” can be an alternative tool for calculating the sample size. It may be useful to students,

## Tamanho da Amostra ( $n$ ) - Para Proporção

- Qual deve ser o tamanho da amostra para que se tenha um dado um erro máximo de  $|p - \pi| \leq \varepsilon$  e uma confiança de  $100(1 - \alpha)\%$  ?

$$n = \frac{[z_{(\alpha/2)}]^2 \pi(1 - \pi)}{\varepsilon^2}$$

- Correção para populações finitas ( $N \leq n$ )

$$n' = \frac{n}{1 + \frac{n}{N}}$$

## Exemplo:

Um epidemiologista veterinário quer investigar a prevalência de febre *aftosa* em uma fazenda que possui 10 mil cabeças de gado da raça *nelore* na região Centro-Oeste do Brasil. Segundo a literatura, a prevalência de febre *aftosa* nesta região no ano passado foi de 50%, determine o  $n$ , ao nível de 5% de significância para este *survey* supondo um erro máximo de 0,04 ( $|p - \pi| \leq 0,04$ ). Temos então:

$$n = \frac{[z_{(\alpha/2)}]^2 \pi(1 - \pi)}{\varepsilon^2} = \frac{(1,96)^2 0.5(1 - 0.5)}{0.04^2} \simeq 600 \text{ bois}$$

## Solução do exemplo no R:

```
library(epiDisplay)
n.for.survey(p=.5, delta =.04, alpha = 0.05)
```

Sample size for survey.

Assumptions:

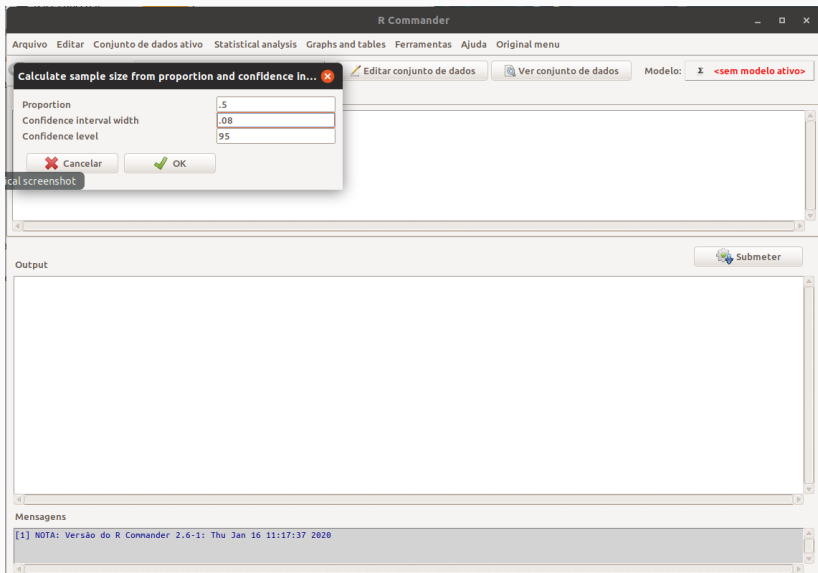
Proportion = 0.5

Confidence limit = 95 %

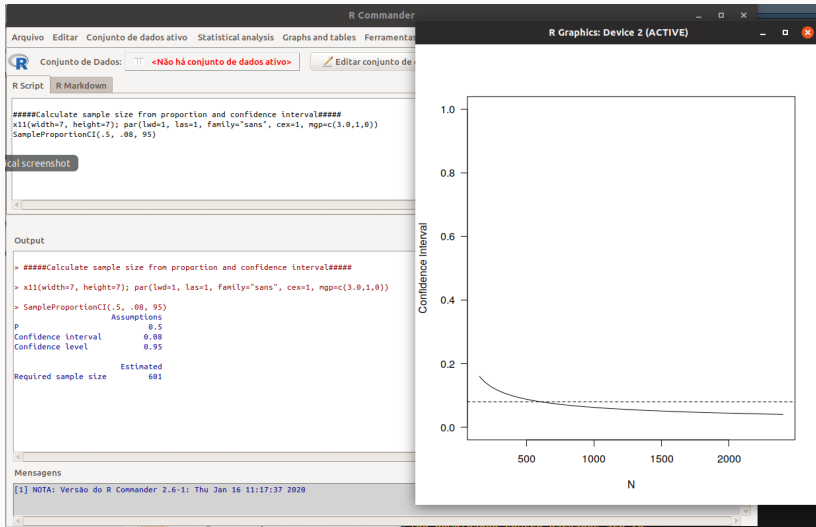
Delta = 0.04 from the estimate.

Sample size = 600

# Solução do exemplo utilizando o plugin “RcmdrPlugin.EZR”



# Solução do exemplo utilizando o plugin “RcmdrPlugin.EZR”



# Introdução à Inferencia Estatística

---

- O processo de inferência (indução) realizado sobre uma população pode ser feito, a partir de uma amostra de duas formas:
  1. **Estimação:** É o processo que usa os resultados extraídos da amostra para produzir inferências de um certo parâmetro populacional. Tal processo pode ser feito de duas formas: Estimação Pontual e Intervalar;
  2. **Testes de Hipótese:** É o processo que usa os resultados extraídos da amostra para testar valores de certos parâmetros da população (Testes Paramétricos e Não-Paramétricos).



- Testes de hipóteses permitem mensurar a evidência em favor ou contra valores específicos de um parâmetro de interesse;
- Para além de uma estimativa pontual de um determinado parâmetro de interesse, em muitas situações, é importante dispôr de alguma forma de do intervalo que indique a confiança (1) que se pode depositar na estimativa pontual.

# Elementos Básicos de um Teste de Hipótese

1. Hipótese Nula ( $H_0$ )
2. Hipótese Alternativa ( $H_1$ )
3. Nível de significância ( $\alpha$ )
4. Estatística do teste
5. Valor-p ou p-valor
6. Regra de decisão - Rejeita-se ou não  $H_0$

- Valor-p ou p-valor é a probabilidade estimada de se rejeitar  $H_0$ , quando  $H_0$  é verdadeira.

$$P(\text{rejeitar } H_0 | H_0 \text{ verdadeira}) = P(\text{erro do tipo I})$$

# Testes Paramétricos e Não-Paramétricos

- Os testes estatísticos são fundamentalmente utilizados em pesquisas que tem como objetivo comparar condições experimentais;
- Eles fornecem um respaldo científico às pesquisas para que estas tenham validade e tenham aceitabilidade no meio científico.
- Os testes podem ser divididos em paramétricos e não-paramétricos.

- Conforme Callegari-Jacques (2003), nos testes paramétricos os valores da variável estudada devem ter distribuição normal ou aproximação normal.
- Já os testes não-paramétricos, também chamados por testes de distribuição livre, não têm exigências quanto ao conhecimento da distribuição da variável na população.

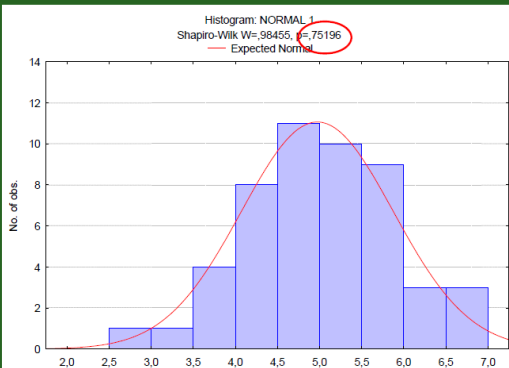
# Testes de Normalidade

---

- O teste de *Shapiro-Wilk* tem sido o mais utilizado para testar a normalidade;
- Se o valor calculado de  $W$  é estatisticamente significativo (para  $pvalor < \alpha$ ) rejeita-se a hipótese que a distribuição estudada é normal, ou seja, para a distribuição ser considerada Normal o valor de  $p$  deve ser maior que  $\alpha$ .

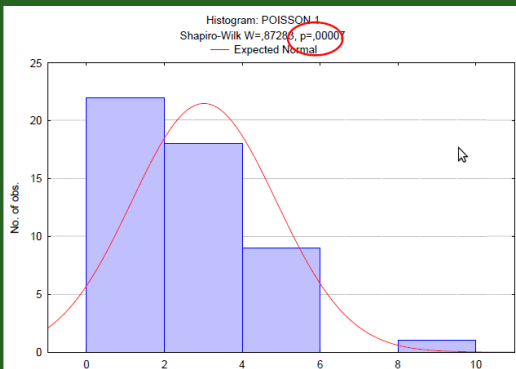
# Teste de Shapiro-Wilk

## DISTRIBUIÇÃO NORMAL





## DISTRIBUIÇÃO NÃO NORMAL



## Teste de Kolmogorov-Smirnov

- O teste de *Kolmogorov-Smirnov* baseia-se na máxima diferença entre a distribuição acumulada da amostra e distribuição acumulada esperada.
- Se o valor calculado de  $D$  é estatisticamente significativo (para  $p < 0,05$ ) rejeita-se a hipótese que a distribuição estudada é normal, ou seja, para a distribuição ser considerada Normal o valor de  $p$  deve ser maior que 0,05.

## Exemplo de aplicação

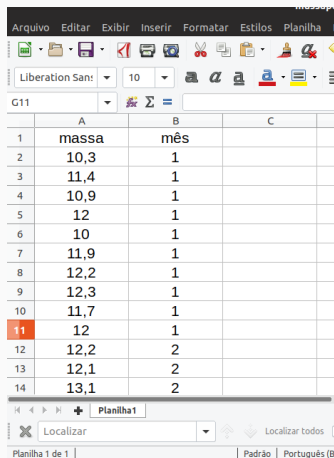
- A massa de 10 pássaros migratórios foi medida em duas ocasiões, primeiro em agosto e os mesmos pássaros (marcados individualmente e recapturados) foram remedidos em setembro.
- Os dados apresentam normalidade ?

## Exemplo de aplicação

Agosto	Setembro
10,3	12,2
11,4	12,1
10,9	13,1
12,0	11,9
10,0	12,0
11,9	12,9
12,2	11,4
12,3	12,1
11,7	13,5
12,0	12,3

# Solução do exemplo utilizando o Rcommander

Importar o banco **massapassaros.xlsx** para o R



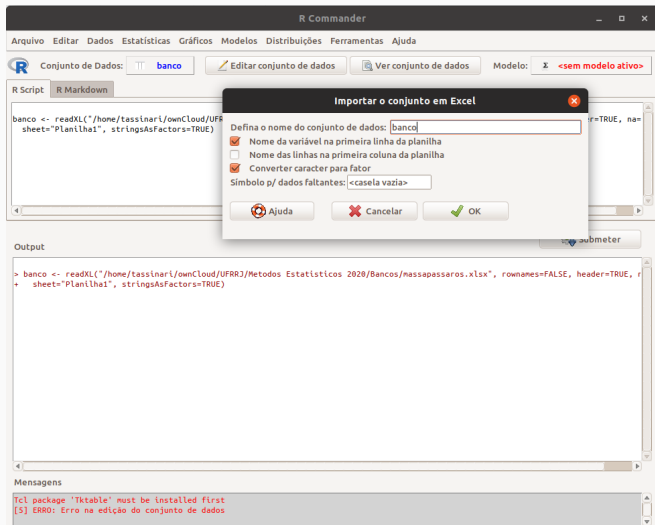
The screenshot shows a spreadsheet application window with a menu bar (Arquivo, Editar, Exibir, Inserir, Formatar, Estilos, Planilha) and a toolbar. The spreadsheet contains a table with two columns: 'massa' (mass) and 'mês' (month). The data is as follows:

	A	B	C
1	massa	mês	
2	10,3	1	
3	11,4	1	
4	10,9	1	
5	12	1	
6	10	1	
7	11,9	1	
8	12,2	1	
9	12,3	1	
10	11,7	1	
11	12	1	
12	12,2	2	
13	12,1	2	
14	13,1	2	

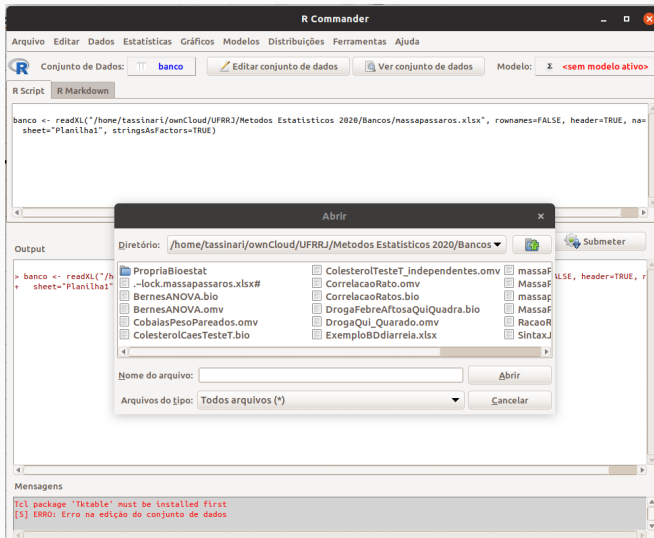
The spreadsheet interface includes a status bar at the bottom showing 'Planilha 1 de 1' and 'Padrão Português (B)'. The cell '11' is highlighted in orange.

# Solução do exemplo utilizando o Rcommander

Rcommander → Dados → Importar arquivo de dados → do arquivo Excel → massapassaros.xlsx



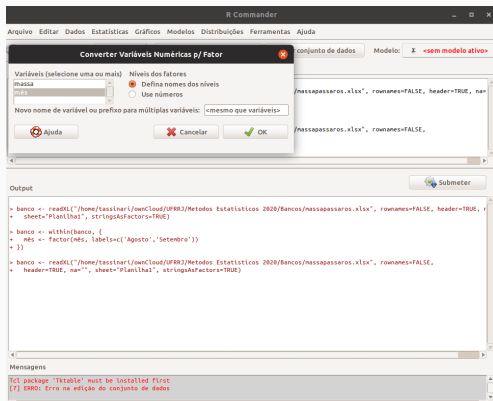
# Solução do exemplo utilizando o Rcommander



# Solução do exemplo utilizando o Rcommander

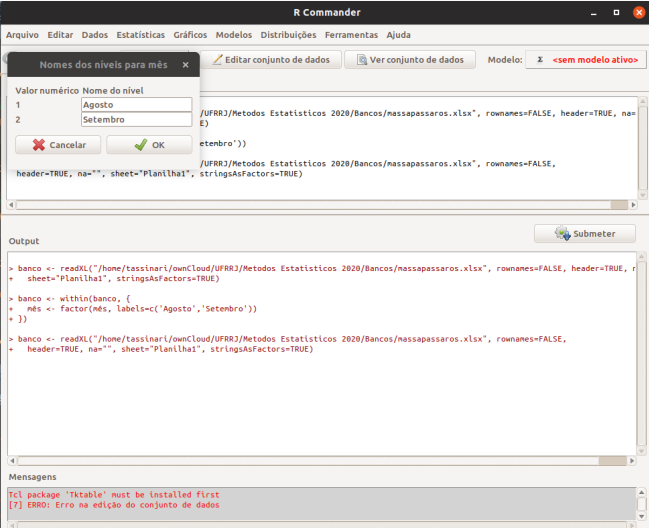
Colocar os *labels* na variável categórica (qualitativa)

Rcommander → Dados → Modificação de variáveis no conjunto de dados... → Converter variável numérica para fator...





# Solução do exemplo utilizando o Rcommander



The screenshot shows the R Commander application window. A dialog box titled "Nomes dos níveis para mês" is open, allowing the user to assign names to factor levels. The dialog has two rows: the first row has "1" in the "Valor numérico" column and "Agosto" in the "Nome do nível" column; the second row has "2" in the "Valor numérico" column and "Setembro" in the "Nome do nível" column. There are "Cancel" and "OK" buttons at the bottom of the dialog.

The main script editor contains the following R code:

```
/UFRRJ/Metodos Estatísticos 2020/Bancos/massapassaros.xlsx", rownames=FALSE, header=TRUE, na=""  
E)  
setenbro'))  
/UFRRJ/Metodos Estatísticos 2020/Bancos/massapassaros.xlsx", rownames=FALSE,  
header=TRUE, na="", sheet="Planilha1", stringsAsFactors=TRUE)
```

The Output window shows the execution of the code:

```
> banco <- readXL("/home/tassinari/ownCloud/UFRRJ/Metodos Estatísticos 2020/Bancos/massapassaros.xlsx", rownames=FALSE, header=TRUE, r  
+ sheet="Planilha1", stringsAsFactors=TRUE)  
  
> banco <- within(banco, {  
+   mês <- factor(mês, labels=c('Agosto','Setenbro'))  
+ })  
  
> banco <- readXL("/home/tassinari/ownCloud/UFRRJ/Metodos Estatísticos 2020/Bancos/massapassaros.xlsx", rownames=FALSE,  
+ header=TRUE, na="", sheet="Planilha1", stringsAsFactors=TRUE)
```

The Mensagens window shows an error message:

```
Tcl package 'tktable' must be installed first  
[7] ERRO: Erro na edição do conjunto de dados
```

# Solução do exemplo utilizando o Rcommander

The screenshot displays the R Commander application window. The main menu bar includes Arquivo, Editar, Dados, Estatísticas, Gráficos, Modelos, Distribuições, Ferramentas, and Ajuda. Below the menu, there are buttons for 'Conjunto de Dados: banco', 'Editar conjunto de dados', 'Ver conjunto de dados', and 'Modelo: <sem modelo ativo>'. The 'R Script' tab is active, showing the following R code:

```
sheet="Planilha1", stringsAsFactors=TRUE)
banco <- within(banco, {
  mês <- factor(mês, labels=c('Agosto','Setembro'))
})
banco <- readXL("/home/tassinari/ownCloud/UFRRJ/2020/Bancos/nassapassaros.xlsx", rownames=FALSE,
header=TRUE, na="", sheet="Planilha1", stringsAsFactors=TRUE)
banco <- within(banco, {
  mês <- factor(mês, labels=c('Agosto','Setembro'))
})
```

A small data preview window is open, showing a table with columns 'nassa' and 'mês'.

	nassa	mês
1	10.3	Agosto
2	11.4	Agosto
3	10.9	Agosto
4	12.0	Agosto
5	18.0	Agosto
6	11.9	Agosto
7	12.2	Agosto
8	12.3	Agosto
9	11.7	Agosto
10	12.0	Agosto
11	12.2	Setembro
12	12.1	Setembro
13	13.1	Setembro
14	11.9	Setembro
15	12.0	Setembro
16	12.9	Setembro
17	11.4	Setembro
18	12.1	Setembro
19	13.5	Setembro
20	12.3	Setembro

The 'Output' window shows the execution of the R code, resulting in the following output:

```
> banco <- readXL("/home/tassinari/ownCloud/UFRRJ/2020/Bancos/nassapassaros.xlsx", rownames=FALSE,
+ sheet="Planilha1", stringsAsFactors=TRUE)
> banco <- within(banco, {
+   mês <- factor(mês, labels=c('Agosto','Setembro'))
+ })
> banco <- readXL("/home/tassinari/ownCloud/UFRRJ/2020/Bancos/nassapassaros.xlsx", rownames=FALSE,
+ header=TRUE, na="", sheet="Planilha1", stringsAsFactors=TRUE)
> banco <- within(banco, {
+   mês <- factor(mês, labels=c('Agosto','Setembro'))
+ })
```

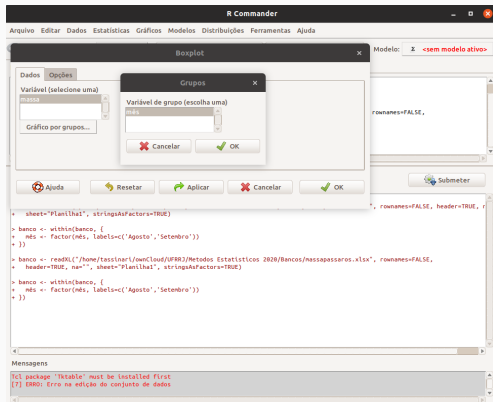
The 'Mensagens' window shows the following error message:

```
Tcl package 'tktable' must be installed first
[7] ERRO: Erro na edição do conjunto de dados
```

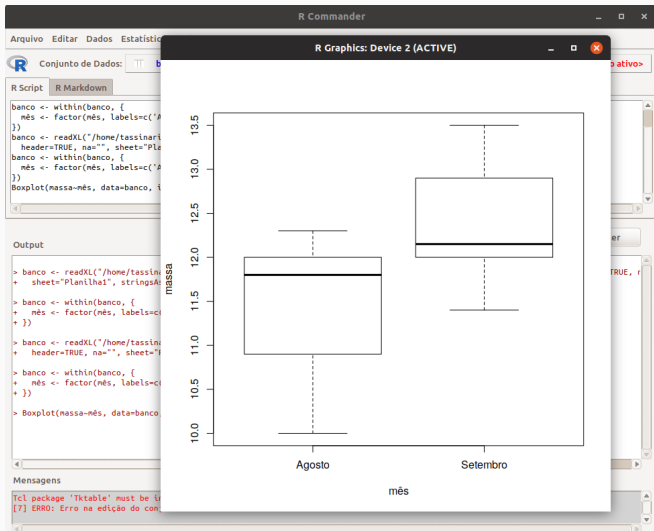
# Solução do exemplo utilizando o Rcommander

Plotando os boxplots

Rcommander → Gráficos → Boxplot



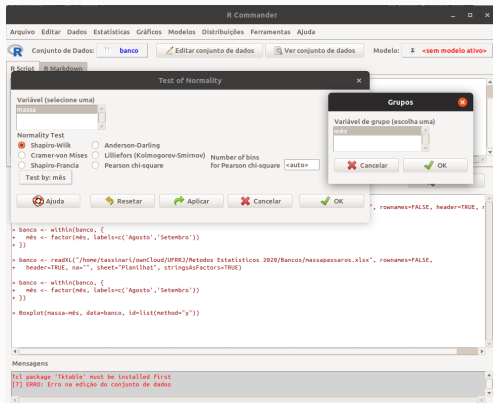
# Solução do exemplo utilizando o Rcommander



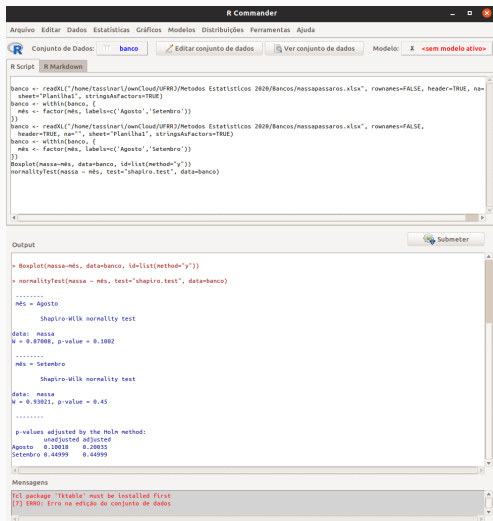
# Solução do exemplo utilizando o Rcommander

Testando a normalidade

Rcommander → Estatísticas → Resumos → Test of normality...



# Solução do exemplo utilizando o Rcommander



The screenshot shows the R Commander application window. The menu bar includes Arquivo, Editar, Dados, Estatísticas, Gráficos, Modelos, Distribuições, Ferramentas, and Ajuda. The toolbar has buttons for Conjunção de Dados, Editar conjunto de dados, Ver conjunto de dados, and Modelo. The main window is divided into three panes: R Script, R Markdown, and Output. The R Script pane contains the following code:

```
banco <- readXL("/home/tassinari/ownCloud/UFRRJ/Metodos Estatisticos 2020/Bancos/massapassaros.xlsx", rownames=FALSE, header=TRUE, na=
sheet="Planilha1", stringsAsFactors=TRUE)
banco <- within(banco, {
  mês <- factor(mês, labels=c('Agosto', 'Setembro'))
})
banco <- readXL("/home/tassinari/ownCloud/UFRRJ/Metodos Estatisticos 2020/Bancos/massapassaros.xlsx", rownames=FALSE,
header=TRUE, na="", sheet="Planilha1", stringsAsFactors=TRUE)
banco <- within(banco, {
  mês <- factor(mês, labels=c('Agosto', 'Setembro'))
})
boxplot(massa~mês, data=banco, id=list(method="y"))
normalityTest(massa ~ mês, test="shapiro.test", data=banco)
```

The Output pane shows the results of the boxplot and normality tests:

```
> boxplot(massa~mês, data=banco, id=list(method="y"))
> normalityTest(massa ~ mês, test="shapiro.test", data=banco)

-----
mês = Agosto
      Shapiro-Wilk normality test

data: massa
W = 0.87008, p-value = 0.1002

-----
mês = Setembro
      Shapiro-Wilk normality test

data: massa
W = 0.93021, p-value = 0.45

-----
p-values adjusted by the Holm method:
unadjusted adjusted
Agosto  0.10018  0.20035
Setembro 0.44999  0.44999
```

The Mensagens pane shows a warning message:

```
lcl package 'lcltable' must be installed first
[7] ERRO: Erro na edição do conjunto de dados
```

# Trasformação de Variáveis

A transformação dos dados é uma possível forma de contornar os problemas da alta dispersão e o pressuposto da não normalidade dos dados.

Alguns exemplos de transformações na variável desfecho:

- $\sqrt{y_i}$
- $\log(y_i)$
- $\exp(y_i)$
- $1/y_i$
- $y_i - \bar{y}$
- $\frac{y_i - \bar{y}}{\sigma_y}$

## Exemplo de aplicação:

- Supondo o peso dos ratinhos abaixo, verifique se existe normalidade nos dados e caso essa condição não seja satisfeita faça algumas transformações na variável peso.
- Os dados são: 79; 10; 11; 86; 12; 81; 90; 99; 29; 19; 11 e 98.



## Solução do exemplo utilizando o Rcommander

Para entrar com um conjunto de dados, é possível ser feito da seguinte forma:

Rcommander → Dados → Novo conjunto de dados...

Ou importar o arquivo *transforme\_normal.xlsx*

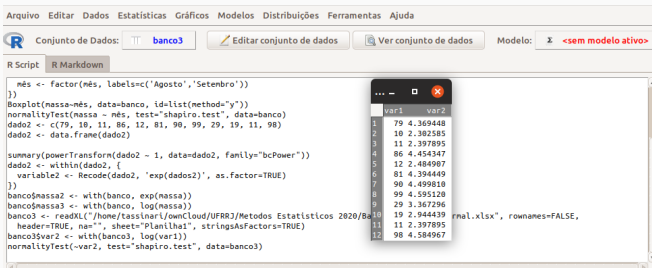
# Solução do exemplo utilizando o Rcommander

Recalcular a variável quantitativa com a função desejada desejada (ex:  $\exp()$ ,  $\log()$ , etc)

Rcommander → Dados → Modificação de variáveis no conjunto de dados... → Computar nova variável...



# Solução do exemplo utilizando o Rcommander



The screenshot shows the R Commander application window. The menu bar includes Arquivo, Editar, Dados, Estatísticas, Gráficos, Modelos, Distribuições, Ferramentas, and Ajuda. The toolbar shows buttons for 'Conjunto de Dados: banco3', 'Editar conjunto de dados', 'Ver conjunto de dados', and 'Modelo: <sem modelo ativo>'. The 'R Script' tab is active, displaying the following R code:

```
mês <- factor(mês, labels=c('Agosto','Setembro'))
})
Boxplot(massa~mês, data=banco, id=list(method="y"))
normalityTest(massa ~ mês, test="shapiro.test", data=banco)
dado2 <- c(79, 10, 11, 86, 12, 81, 90, 99, 29, 19, 11, 98)
dado2 <- data.frame(dado2)

summary(powerTransform(dado2 ~ 1, data=dado2, family="bcPower"))
dado2 <- within(dado2, {
  variable2 <- Recode(dado2, 'exp(dados2)', as.factor=TRUE)
})
banco$massa2 <- with(banco, exp(massa))
banco$massa3 <- with(banco, log(massa))
banco3 <- readXL("/home/tassinari/ownCloud/UFRJ/Metodos Estatisticos 2020/8a...
  header=TRUE, na="", sheet="Planilha1", stringsAsFactors=TRUE)
banco3$var2 <- with(banco3, log(var1))
normalityTest(~var2, test="shapiro.test", data=banco3)
```

Overlaid on the code window is a small data viewer window showing a table with two columns, 'var1' and 'var2', and 12 rows of data:

	var1	var2
1	79	4.369448
2	10	2.302585
3	11	2.397895
4	86	4.454347
5	12	2.484967
6	81	4.394449
7	90	4.499810
8	99	4.595120
9	29	3.367296
10	19	2.944439
11	11	2.397895
12	98	4.584967

A cada variável transformada, basta ir testando a normalidade.