

# Curso de Pós-Graduação em Ciências Veterinárias - UFRRJ

Métodos Estatísticos

---

Prof: Wagner Tassinari

wagner.tassinari@ini.fiocruz.br

Estatística Descritiva

# **Estatística Descritiva**

---

- Organização e descrição dos dados;
- Identificação de valores que represente o elemento típico;
- Avaliação e quantificação da variabilidade do conjunto de dados;
- Familiarização com os dados; forma da distribuição dos dados;
- Identificar estruturas interessantes, como a de valores atípicos.

## Formas de sumarizar os dados:

- Tabelas
- Gráficos
- Medidas-resumo

# Medidas de Tendência Central (Medidas de Centro)

---

## Medidas de Tendência Central (Medidas de Centro)

- Caracterizam o conjunto de dados por valores que representem todos os outros valores da amostra
- É uma forma de resumir o conjunto de dados em um único valor
- Medidas: **média, mediana e moda.**

- Somam-se todos os  $n$  valores da amostra e divide-se pela quantidade total de valores  $n$  da amostra.
- O valor da média não necessariamente pertence ao conjunto original de valores.
- Não é uma medida robusta  $\rightarrow$  influenciada por valores extremos.
- É expressa por:  $\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$

- **Exemplo:** Pressão sistólica de uma amostra de 5 pacientes

Pacientes	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$
Pressão Sistólica	15	20	14	14	12

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$

$$\bar{X} = \frac{15 + 20 + 14 + 14 + 12}{5} = \frac{75}{5} = 15$$



Exemplo: Influência de valores extremos na média

Pacientes	1	2	3	4	5	6	7
Dados	2	2	5	7	6	4	5

$$\bar{X} = \frac{2 + 2 + 5 + 7 + 6 + 4 + 5}{7} = \frac{32}{7} = 4,4$$

Pacientes	1	2	3	4	5	6	7
Dados	2	2	5	7	6	4	55

$$\bar{X} = \frac{2 + 2 + 5 + 7 + 6 + 4 + 55}{7} = \frac{81}{7} = 11,6$$

- Definição: valor que divide o conjunto de dados em duas partes iguais
- 50% das observações ficam acima da mediana e 50% ficam abaixo
- Medida mais robusta → não sofre influência de valores extremos.

# Mediana

- Colocar os valores em ordem e, em seguida, aplicar um dos dois processos abaixo:
- Se o número de valores é **ímpar**, a posição da mediana é dada pelo elemento de ordem:  $\frac{n+1}{2}$ 
    - $x_1, x_2, x_3 \rightarrow \frac{3+1}{2} = 2 \rightarrow md = x_2$ , ou seja, elemento de ordem 2
  - Se o número de valores é **par**, a mediana é dada pela média dos elementos de ordem  $\frac{n}{2}$  e  $\frac{n+2}{2}$ :
    - $x_1, x_2, x_3, x_4 \rightarrow md = \frac{x_2 + x_3}{2}$

- **Exemplo 1:** (1, 2, 5, 6, 7)
  - Número **ímpar** de elementos  $\rightarrow$  mediana é dada pelo valor que ocupa a terceira posição  $\frac{5+1}{2}$ , que é igual a 5.
- **Exemplo 2:** (1, 2, 5, 6, 7, 7)
  - Número par de elementos  $\rightarrow$  mediana será dada por  $md = \frac{5+6}{2} = 5,5$

- **Exemplo:** Influência de valores extremos na mediana
- (2, 2, 4, 5, 6, 7)
  - Número **ímpar** de elementos → mediana é dada pelo valor que ocupa a quarta posição  $\frac{7+1}{2}$ , que é igual a 5.
- (2, 2, 4, 5, 6, 7, 55)
  - Número **ímpar** de elementos → mediana é dada pelo valor que ocupa a quarta posição  $\frac{7+1}{2}$ , que é igual a 5.

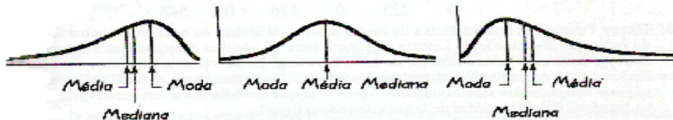
- Definição: valor que ocorre com maior frequência;
- A moda sempre pertence ao conjunto original de valores;
- Uma distribuição pode ser unimodal, bimodal, multimodal ou amodal.
- Exemplos:
  - $(1, 1, 1, 3, 5, 6, 8) \rightarrow \text{Moda} = 1$
  - $(1, 1, 2, 2, 3, 4, 5) \rightarrow \text{Moda} = 1 \text{ e } 2$
  - $(M, F, M, M, M, F) \rightarrow \text{Moda} = M$
  - $(1, 2, 5, 9, 11) \rightarrow \text{Amodal}$

# Mediana versus Média - Qual medida escolher?

- Média
  - Medida mais usada na prática;
  - Facilidade de tratamento estatístico;
  - Muito influenciada por valores extremos.
- Mediana
  - Não é tão influenciada por valores extremos;
  - Utiliza no máximo dois valores da amostra (desvantagem).

# Forma da Distribuição das Medidas de Tendência Central

- Uma distribuição de dados é simétrica se a metade esquerda do seu histograma é praticamente uma imagem espelhada de sua imagem direita.
- A distribuição de dados é assimétrica quando se estende mais para um lado que para o outro.



Assimétrica à esquerda

Simétrica

Assimétrica à direita



# Prática

---

## Vamos praticar ?

- Qual a média de pesos de recém nascidos na Maternidade N. S. da Luz no dia de ontem ?
  - Bebê 1 = 3,2 Kg
  - Bebê 2 = 2,8 Kg
  - Bebê 3 = 2,7 Kg
  - Bebê 4 = 3,4 Kg
  - Bebê 5 = 3,1 Kg

## Vamos praticar ?

- Qual a média de pesos de recém nascidos na Maternidade N. S. da Luz no dia de ontem ?

$$\bar{X} = \frac{3,2 + 2,8 + 2,7 + 3,4 + 3,1}{5} = 3,04Kg$$

## Vamos praticar ?

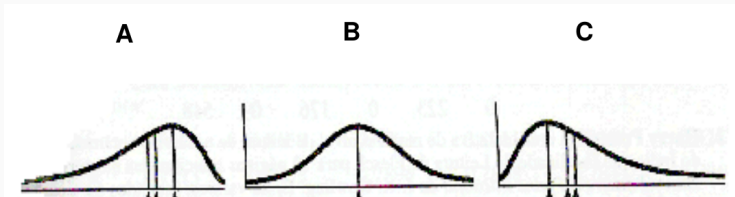
- Qual a mediana dos dados abaixo ?
  - Dados 1: (1, 2, 4, 6, 7)
  - Dados 2: (1, 2, 5, 6, 7, 7)

## Vamos praticar ?

- Colocando os valores em ordem, temos:
1. Dados 1 (**ímpar**), a posição da mediana é dada pelo elemento de ordem:  $\frac{n+1}{2}$ 
    - $(1, 2, 4, 6, 7) \rightarrow \frac{5+1}{2} = 3 \rightarrow md = x_3$ , ou seja, elemento de ordem 3
    - $md = 4$
  2. Se o número de valores é **par**, a mediana é dada pela média dos elementos de ordem  $\frac{n}{2}$  e  $\frac{n+2}{2}$ :
    - $(1, 2, 5, 6, 7, 7) \rightarrow md = \frac{x_2 + x_3}{2} = \frac{5 + 6}{2}$
    - $md = 5,5$

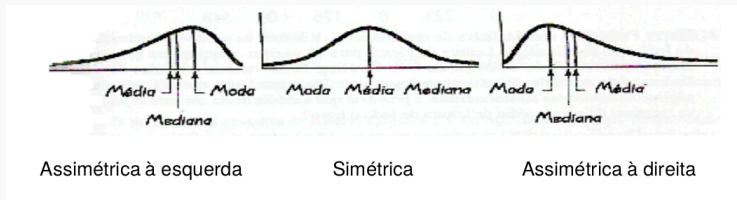
## Vamos praticar ?

- Em qual gráfico a moda é maior do que a mediana e a média ?



# Vamos praticar ?

- Em qual gráfico a moda é maior do que a mediana e a média ?
  - Gráfico **A**



# Medidas Separatrizes

---



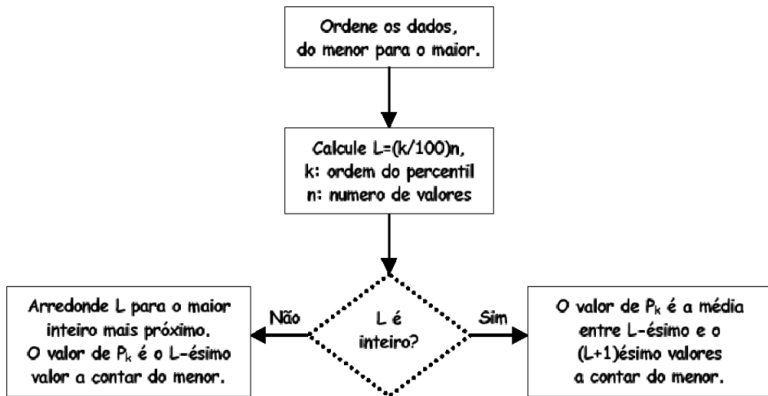
- **Percentil:** O percentil de ordem  $k$  (onde  $k$  é qualquer valor entre 0 e 100), denotado por  $P_k$ , é o valor tal que  $k\%$  dos valores do conjunto de dados são menores ou iguais a ele. Divide a distribuição em 100 partes iguais em um conjunto ordenado de valores.
- **Quartil:** Divide a distribuição em 4 partes iguais em um conjunto ordenado de valores.
- **Decil:** Divide a distribuição em 10 partes iguais em um conjunto ordenado de valores.

# Medidas Separatrizes

- Percentis: 10, 20, 30, ..., 90 → Decis
- Percentil 25 → Primeiro quartil ( $Q_1$ )
- Percentil 50 → Segundo quartil ( $Q_2$ ) → Mediana
- Percentil 75 → Terceiro quartil ( $Q_3$ )

50%				Me	50%			
P <sub>10</sub>	P <sub>25</sub>			P <sub>50</sub>			P <sub>75</sub>	P <sub>90</sub>
	Q <sub>1</sub>			Q <sub>2</sub>			Q <sub>3</sub>	
D <sub>1</sub>	D <sub>2</sub>	D <sub>3</sub>	D <sub>4</sub>	D <sub>5</sub>	D <sub>6</sub>	D <sub>7</sub>	D <sub>8</sub>	D <sub>9</sub>

Figura 3.5 - Determinação do Percentil de ordem K (Triola, 1996).



## Medidas Separatrizes

- **Exemplo:** A tabela abaixo lista 40 níveis ordenados de nicotina para fumantes.

0	1	1	3	17	32	35	44	48	86
87	103	112	121	123	130	131	149	164	167
173	173	198	208	210	222	227	234	245	250
253	265	266	277	284	289	290	313	477	491

- Ache o percentil 30.

$$L_{30} = \frac{30}{100} \cdot 40 = 12$$

- Como o  $L$  é inteiro, tiramos a média entre o elemento  $L = 12$  e  $L + 1 = 13$
- Assim,  $P_{30} = \frac{103 + 112}{2} = 107,5$

# Medidas de Dispersão ou Variabilidade

---

- A dispersão fornece uma medida da proximidade da série de dados em torno de um valor de tendência central, tomado como comparação.
- Medidas para avaliar a dispersão de um conjunto de dados: **Amplitude Total, Variância, Desvio Padrão e Coeficiente de Variação.**

$$AT = x_{mximo} - x_{mnimo}$$

- Maior amplitude total  $\rightarrow$  maior dispersão.
- **Problema:** somente são usados os extremos do conjunto.
- **Elemento auxiliar na análise**  $\rightarrow$  mostra a faixa de variação onde encontramos todos os elementos do conjunto.

- **Exemplo:** Pressão sistólica de uma amostra de 5 pacientes

Pacientes	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$
Pressão Sistólica	15	20	14	14	12

$$AT = 20 - 14 = 8$$



- Poderíamos então pensar na soma das diferenças entre cada valor do conjunto de dados e a média, mas:

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \dots + (x_n - \bar{x}) = \sum_{i=1}^n (x_i - \bar{x}) = 0$$

- Então essa medida não serve como medida de dispersão. Segundo ela, todos os conjuntos de dados teriam variabilidade nula.

- Medida direta da dispersão → conjunto com os dados mais dispersos terá maior variância.
- A variância mede a variabilidade ao redor da média, fornecendo o grau de precisão da média.
- Medida em unidade quadrada (exemplo: anos<sup>2</sup>) → o que dificulta a sua interpretação.

## Variância e Desvio padrão

- A **Variância** é dada por:

$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

- **Desvio padrão** → é obtido por meio da extração da raiz quadrada da variância. Representa o desvio médio dos valores em relação a média. Dado por:

$$S = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

- O desvio-padrão possui a mesma unidade de medida que os dados originais.

## Variância e Desvio padrão

- **Exemplo:** média = 15

Pressão sistólica	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
15	$15 - 15 = 0$	$0^2 = 0$
20	$20 - 15 = 5$	$5^2 = 25$
14	$14 - 15 = -1$	$(-1)^2 = 1$
14	$14 - 15 = -1$	$(-1)^2 = 1$
12	$12 - 15 = -3$	$(-3)^2 = 9$

$$S^2 = \frac{36}{4} = 9 \text{ e } S = \sqrt{9} = 3$$

## Desvio padrão - Interpretação

- Uma pergunta que pode surgir é se um desvio padrão é grande ou pequeno  $\rightarrow$  depende da ordem de grandeza da variável.
- Um desvio padrão de 10 unidades é grande ou pequeno ?
- Se a média é 10.000  $\rightarrow$  desvio é pequeno (0,1% da média).
- Se a média é 100  $\rightarrow$  desvio é grande (10% da média).

## Coeficiente de variação

- É uma medida de dispersão relativa (%) que mede a variação do desvio padrão em relação à média aritmética;
- **Vantagem:** permite a comparação entre variáveis ou populações distintas
- Quanto menor é o coeficiente de variação de um conjunto de dados, menor é a sua variabilidade.
- **Medida adimensional**
- O **Coeficiente de Variação** é dado por:

$$CV(\%) = \frac{S}{\bar{X}} \cdot 100$$

## Coeficiente de variação

- Exemplo:

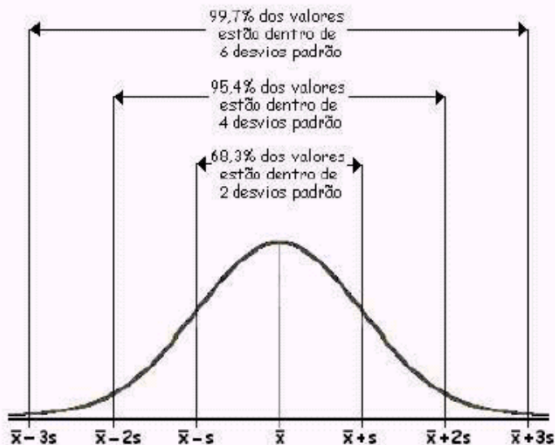
Pressão sistólica	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
15	$15 - 15 = 0$	$0^2 = 0$
20	$20 - 15 = 5$	$5^2 = 25$
14	$14 - 15 = -1$	$(-1)^2 = 1$
14	$14 - 15 = -1$	$(-1)^2 = 1$
12	$12 - 15 = -3$	$(-3)^2 = 9$

$$\bar{x} = 15, S^2 = \frac{36}{4} = 9 \text{ e } S = \sqrt{9} = 3$$

$$CV(\%) = \frac{3}{15} \cdot 100 = 20\%$$

# Regra do Desvio-padrão (Distribuições Simétricas)

Figura 3.4 – Ilustração da regra do desvio padrão para dados com distribuição simétrica.

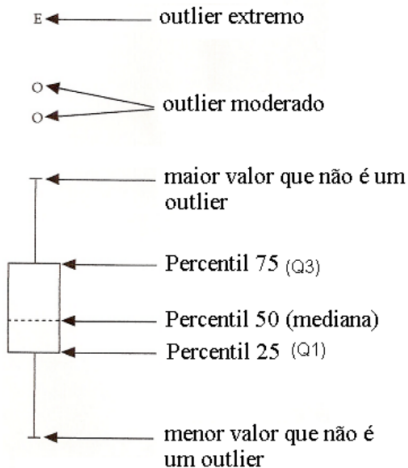




## Gráfico – Boxplot

---

## Gráfico – Boxplot



São úteis na comparação da dispersão de dois ou mais grupos (tamanho da caixa ou distância entre os extremos);

Adicionalmente utilizado para identificar a amplitude dos dados, a presença de pontos discrepantes (*outliers*).

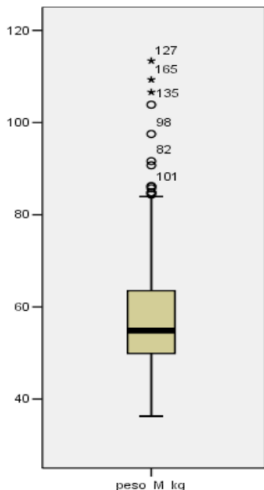
Se não existirem observações discrepantes a distância entre as extremidades do gráfico correspondem a amplitude total.

*Outliers* são marcados com "o" ou "\*" por quase todos os pacotes estatísticos.



- Boxplot e assimetria
  - Simétrica: mediana fica no centro;
  - Assimetria à direita: a mediana fica mais próxima do valor mínimo;
  - Assimetria à esquerda: a mediana fica mais próxima do valor máximo.

## Gráfico – Boxplot



- Banco de dados: *low birth weight*
- Variável: peso materno no último período menstrual, em kg.
- Observa-se a grande variabilidade da variável e a presença de valores extremos (os número identificam os pacientes).
- Obs: *Outliers* moderados - entre 1.5 e 3 x IQR. *Outliers* extremos: acima de 3 x IQR.

IQR = Inter Quartile Range ( $Q3 - Q1$ )

# Variância vs Coeficiente de Variação

- **Variância**

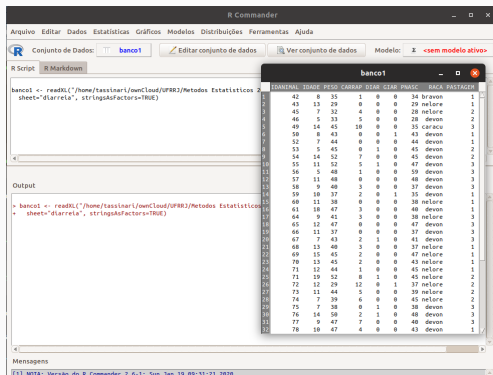
- A variância mede a variabilidade ao redor da média, fornecendo o grau de precisão da média.
- Medida em unidade quadrada.

- **Coeficiente de Variação**

- É uma medida de dispersão relativa (%) que mede a variação do desvio padrão em relação à média aritmética.
- Permite a comparação entre variáveis ou populações distintas.

# Exemplo utilizando o Rcommander

- Importar o arquivo “ExemploBDdiarreia.xlsx”
  - Rcommander → Dados → Importar arquivos de dados → do arquivo Excel



# Exemplo utilizando o Rcommander

- Sumário (resumo) estatístico de todo o banco
  - Rcommander → Resumos → Conjunto de dados ativo

Output



```
> remove(.Table)
```

```
> summary(banco1)
```

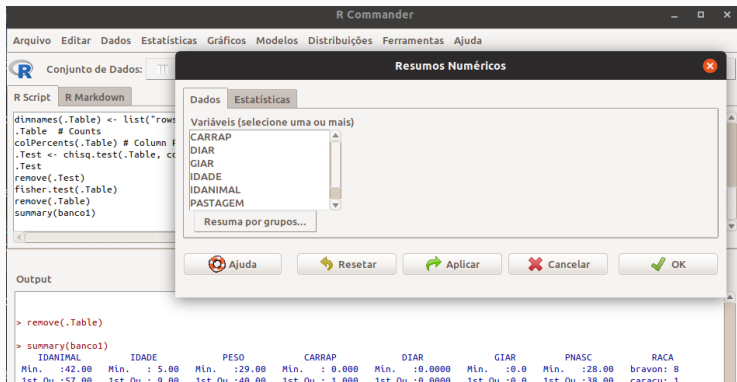
IDANIMAL	IDADE	PESO	CARRAP	DIAR	GIAR	PNASC	RACA
Min. :42.00	Min. : 5.00	Min. :29.00	Min. : 0.000	Min. :0.0000	Min. :0.0	Min. :28.00	bravon: 8
1st Qu.:57.00	1st Qu.: 9.00	1st Qu.:40.00	1st Qu.: 1.000	1st Qu.:0.0000	1st Qu.:0.0	1st Qu.:38.00	caracu: 1
Median :70.00	Median :12.00	Median :45.00	Median : 3.000	Median :0.0000	Median :0.0	Median :40.00	devon:22
Mean :68.47	Mean :11.64	Mean :43.64	Mean : 3.689	Mean :0.1778	Mean :0.2	Mean :41.04	nelore:14
3rd Qu.:80.00	3rd Qu.:14.00	3rd Qu.:47.00	3rd Qu.: 5.000	3rd Qu.:0.0000	3rd Qu.:0.0	3rd Qu.:45.00	
Max. :91.00	Max. :20.00	Max. :54.00	Max. :14.000	Max. :1.0000	Max. :2.0	Max. :59.00	

PASTAGEM
Min. :1.000
1st Qu.:1.000
Median :2.000
Mean :2.022
3rd Qu.:3.000
Max. :3.000

# Exemplo utilizando o Rcommander

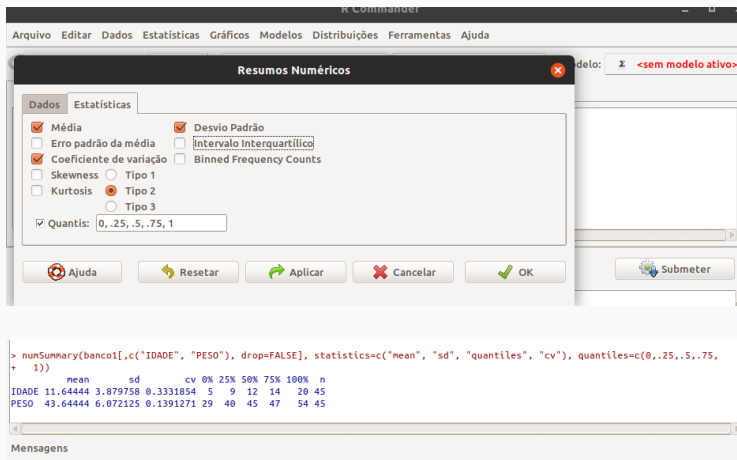
- Sumário (resumo) estatístico de todo o banco
  - Rcommander → Resumos → Resumos numéricos ...





# Exemplo utilizando o Rcommander

- Sumário (resumo) estatístico personalizado por variável

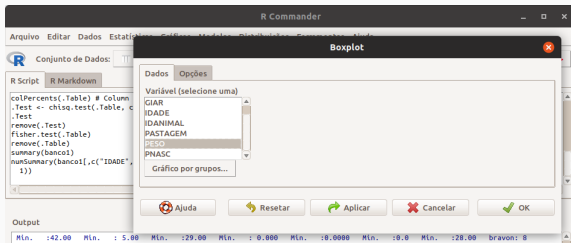


The screenshot shows the R Commander application window. The 'Resumos Numéricos' (Numerical Summaries) dialog box is open, displaying the 'Estatísticas' (Statistics) tab. The following options are checked: Média (Mean), Desvio Padrão (Standard Deviation), Coeficiente de variação (Coefficient of Variation), and Quantis (Quantiles). The 'Intervalo Interquartilico' (Interquartile Range) option is also visible but not checked. The 'Quantis' field is set to 0, .25, .5, .75, 1. The 'Aplicar' (Apply) button is highlighted. The console window at the bottom shows the execution of the `nunSummary` function on the 'IDADE' variable of the 'banco1' dataset, with the following output:

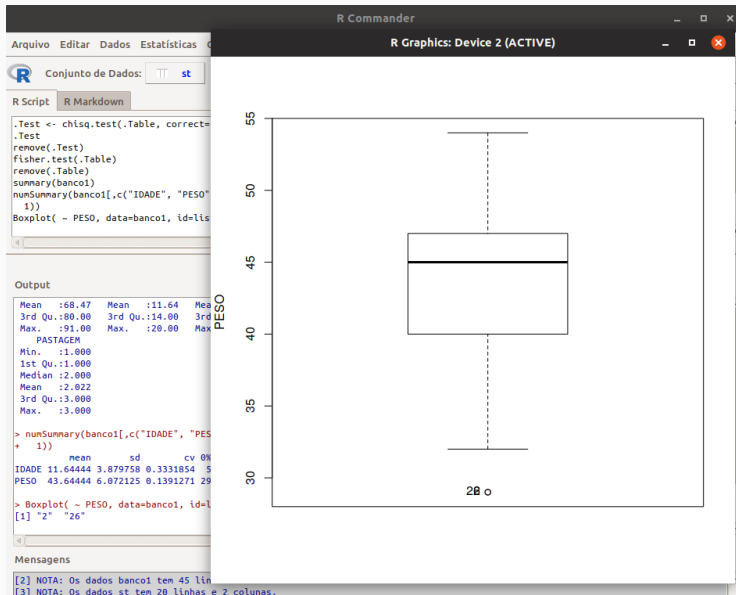
```
> nunSummary(banco1[,c("IDADE", "PESO"), drop=FALSE], statistics=c("mean", "sd", "quantiles", "cv"), quantiles=c(0,.25,.5,.75,
+ 1))
      mean      sd      cv 0% 25% 50% 75% 100% n
IDADE 11.64444 3.879758 0.3331854  5  9 12 14 20 45
PESO  43.64444 6.072125 0.1391271 29 40 45 47 54 45
```

# Exemplo utilizando o Rcommander

- Plotar Boxplot
  - Rcommander → Gráficos → Boxplot

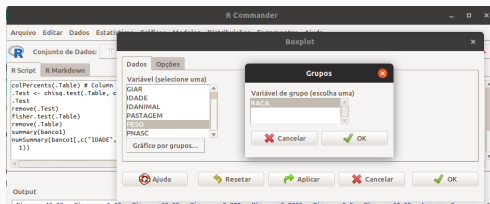
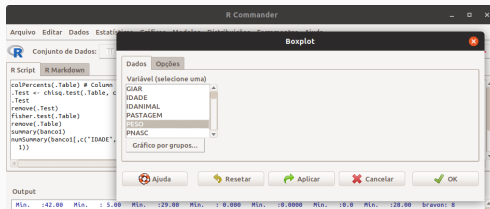


# Exemplo utilizando o Rcommander

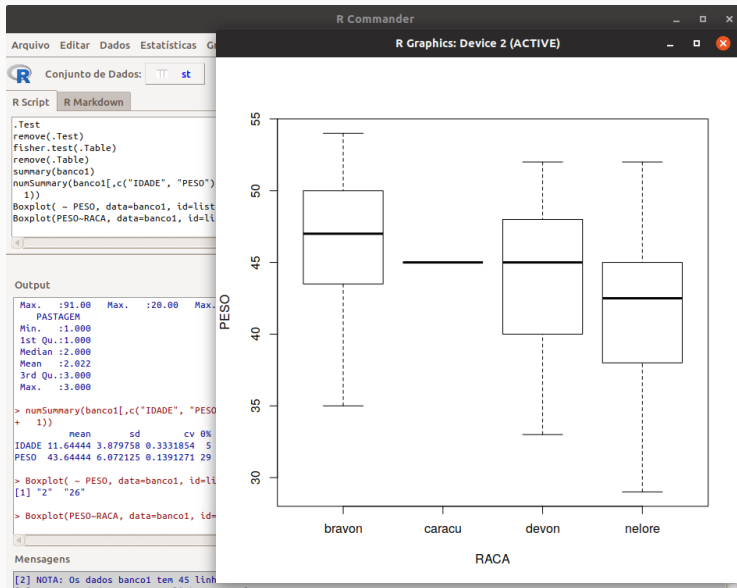


# Exemplo utilizando o Rcommander

- Plotar Boxplot por grupos de variáveis
  - Rcommander → Gráficos → Boxplot

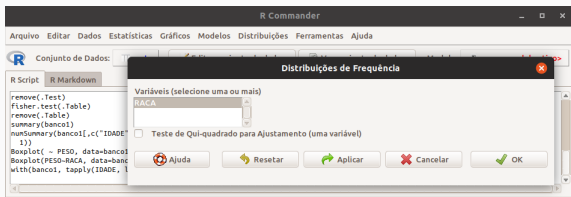


# Exemplo utilizando o Rcommander



# Exemplo utilizando o Rcommander

- Distribuição de frequência da variável RACA
  - Rcommander → Estatísticas → Resumo → Distribuições de frequência ...



Output

```
13.87500 14.00000 10.09091 12.64286

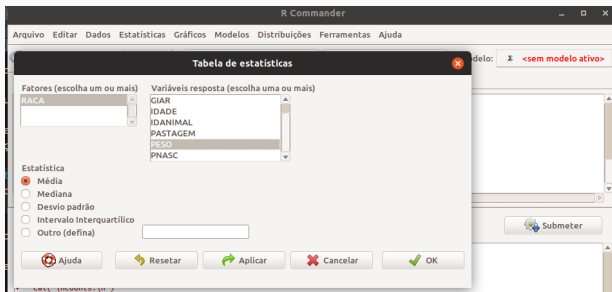
> local({
+   .Table <- with(banco1, table(RACA))
+   cat("\ncounts:\n")
+   print(.Table)
+   cat("\npercentages:\n")
+   print(round(100*.Table/sum(.Table), 2))
+ })

counts:
RACA
bravon caracu devon nelore
8 1 22 14

percentages:
RACA
bravon caracu devon nelore
17.78 2.22 48.89 31.11
```

# Exemplo utilizando o Rcommander

- Média da variável PESO por RACA
  - Rcommander → Estatísticas → Resumo → Tabela de Estatísticas ...



```
> with(banco1, tapply(PESO, list(RACA), mean, na.rm=TRUE))
bravon caracu devon nelore
46.25000 45.00000 44.36364 40.92857
```