

Estatística Descritiva

(II)

Exemplo 1:

Dados *CEA06P16*, do projeto *Perfil Evolutivo da fluência da fala de falantes do português brasileiro*

- Estudo realizado pela Faculdade de Medicina – USP e Faculdade de Filosofia, Letras e Ciências Humanas – USP
- Ano de realização da análise: 2006
- Finalidade: doutorado
- Análise Estatística: Centro de Estatística Aplicada (CEA), IME-USP

Exemplo 1: *fluência da fala*

- **Objetivo:** avaliar o perfil de fluência da fala de acordo com sexo, idade e grau de escolaridade.
- **Amostra:** 594 indivíduos residentes na Grande São Paulo, com idades entre 2 e 99 anos.
- **Amostras de fala auto-expressiva:** o indivíduo era apresentado a uma figura e orientado a discorrer sobre a mesma durante um tempo mínimo de 3 minutos e máximo de 6 minutos. Para crianças de 2 e 3 anos, as amostras foram obtidas com a colaboração dos pais.

Exemplo 1: *fluência da fala*

Algumas variáveis:

- Sexo (1 - Fem e 2 - Masc);
- Idade (em anos);
- Grau de escolaridade (de pré-escola a superior completo);
- Fluxo de palavras por minuto (FPM);
- Fluxo de sílabas por minuto (FSM);
- Número de interjeições durante o discurso (INTERJ);
- Número de palavras não terminadas durante o discurso (PNT);
- Número de pausas durante o discurso (PAUSA).

APOIO COMPUTACIONAL

Software sugerido: *R*



- Vantagem: *software* livre
- *Download*: <http://www.r-project.org/>
 - Escolher opção *Download R*
 - Seguir os passos de instalação
 - Material de apoio: www.ime.usp.br/~mae116

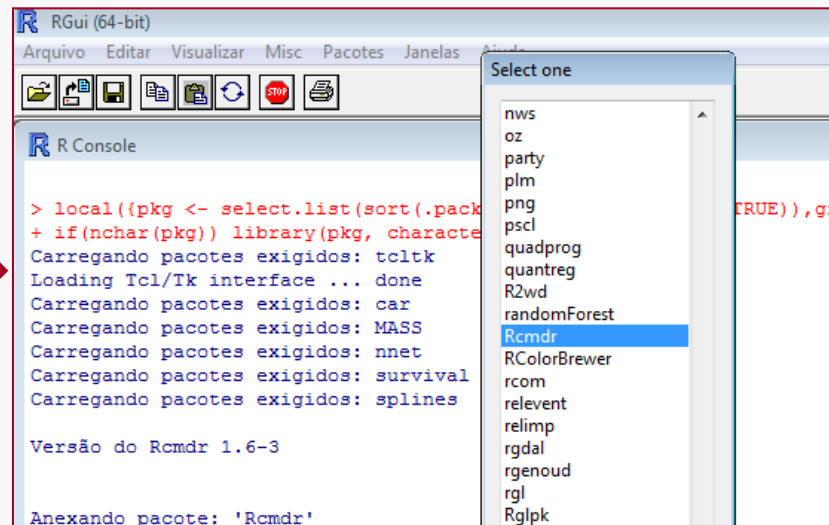
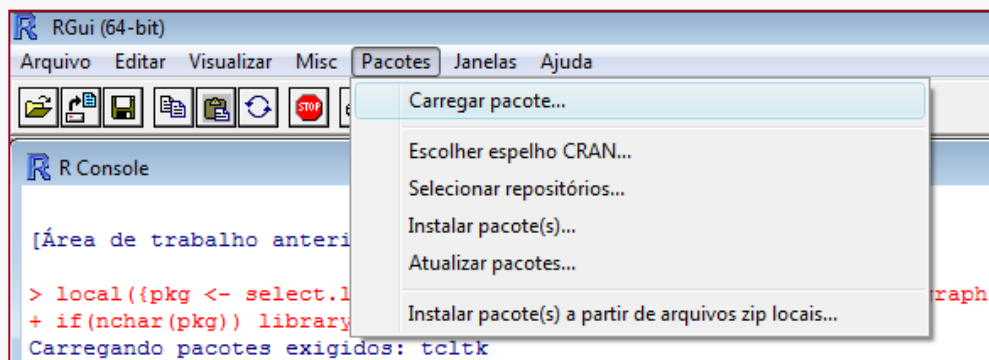
Biblioteca Rcmdr



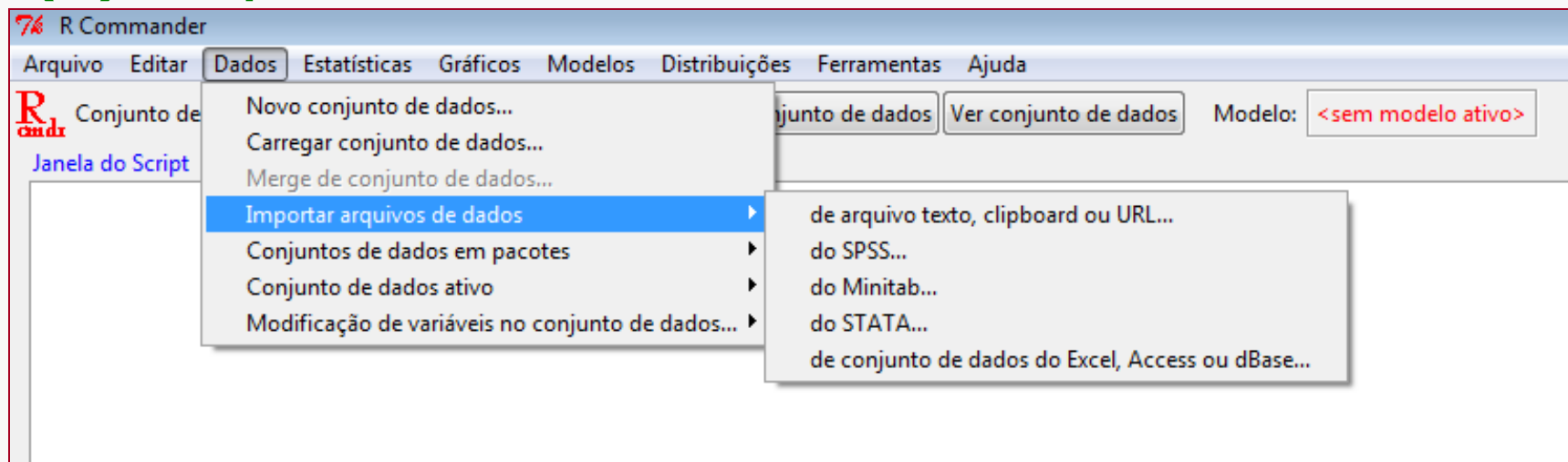
- Ambiente baseado em menus
- Deve ser instalada após instalação do R
- Instruções de instalação no material de apoio

Arquivo *CEA06P16*: Carregando dados no R

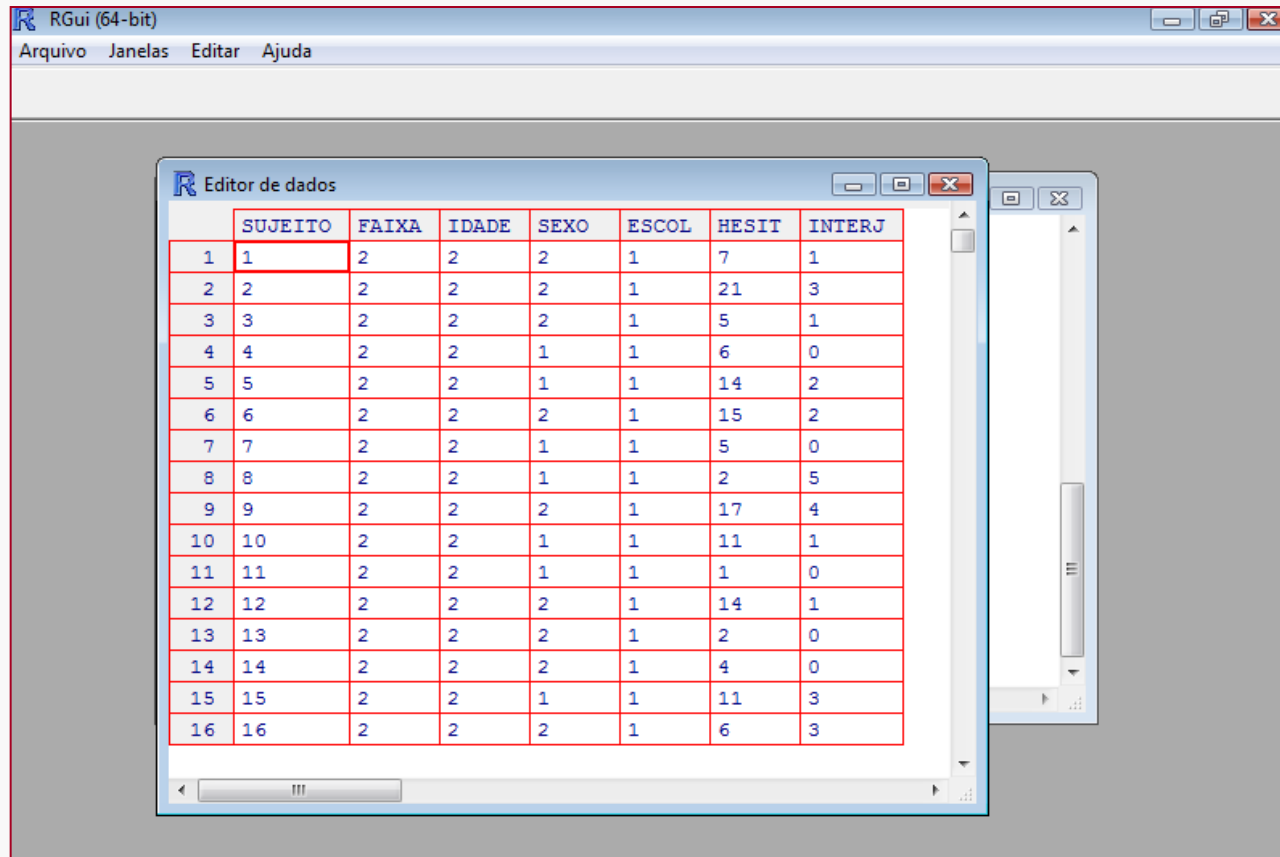
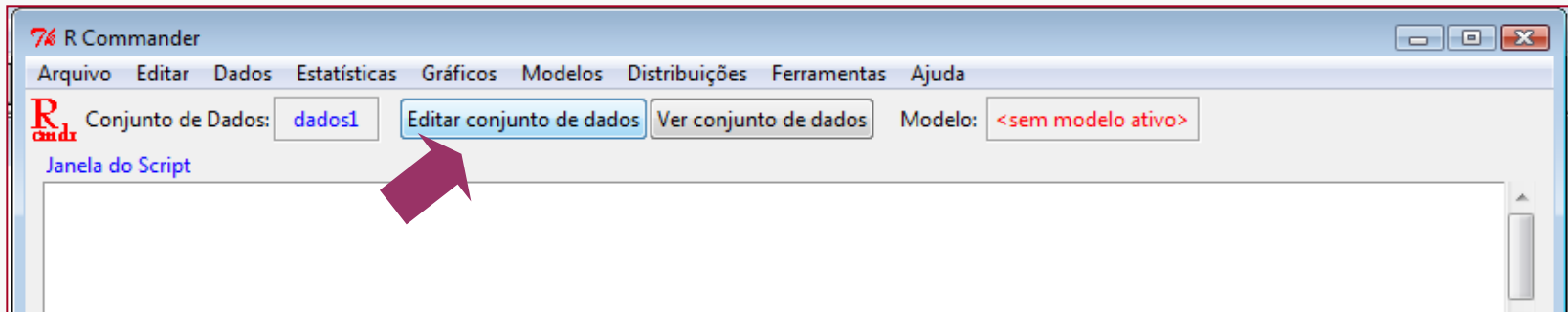
(1) Carregar Rcmdr:



(2) Importar dados:



Arquivo CEA06P16: Visualizar/editar dados



Variáveis qualitativas



Sexo
Grau de
escolaridade

Nominal

Ordinal

Variáveis quantitativas



Num. palavras
não term.
Fluxo sil./min
Fluxo
palavras/min

Discreta

Contínuas

Variáveis Quantitativas

Medidas de posição

Média (\bar{x})

Mediana (md)

Quartis ($Q1, Q3$)

Máximo ($máx$)

Mínimo (min)

Medidas de dispersão

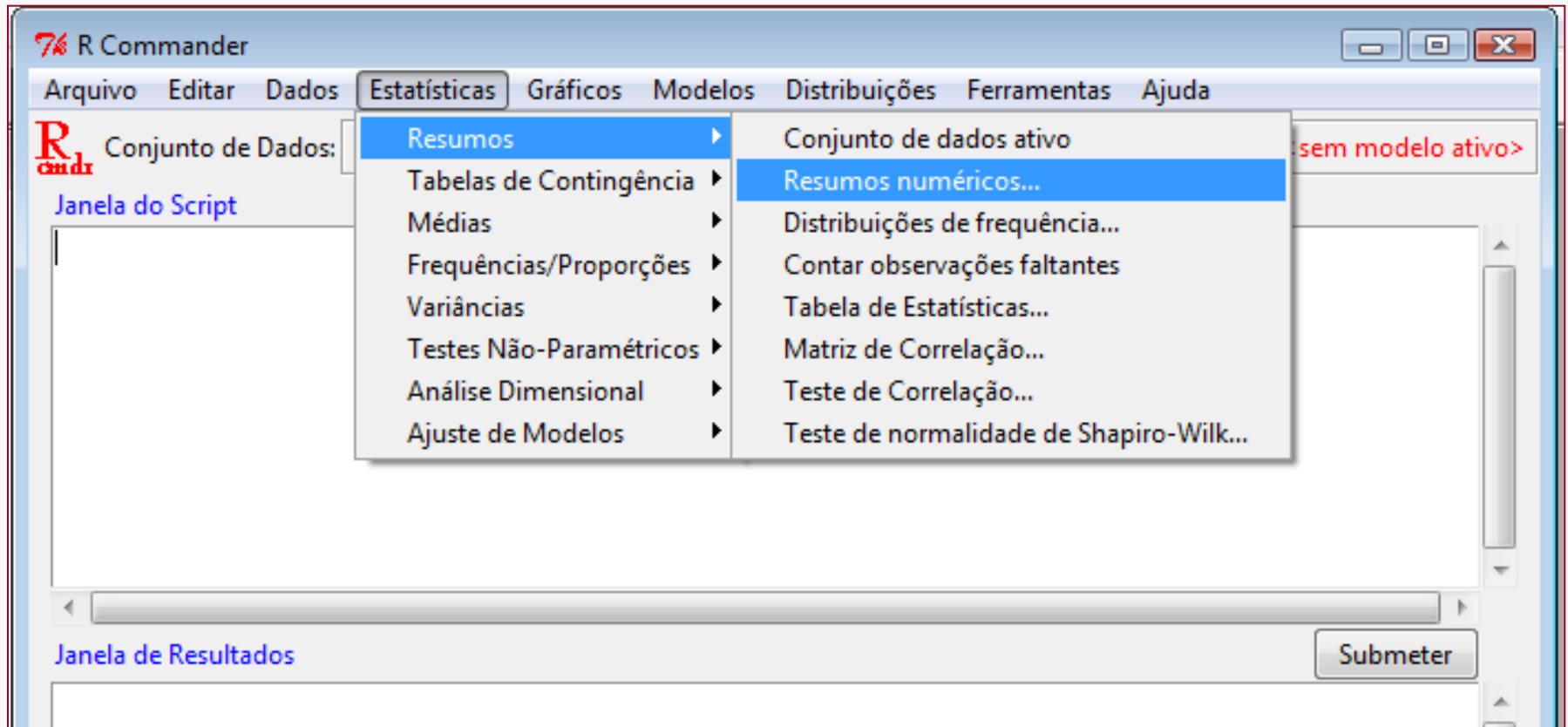
Variância (s^2)

Desvio padrão (s)

Intervalo-interquartil ($Q3 - Q1$)

Coefficiente de variação (CV)

Rcmdr



Estatísticas → Resumos Numéricos

Medidas descriptivas

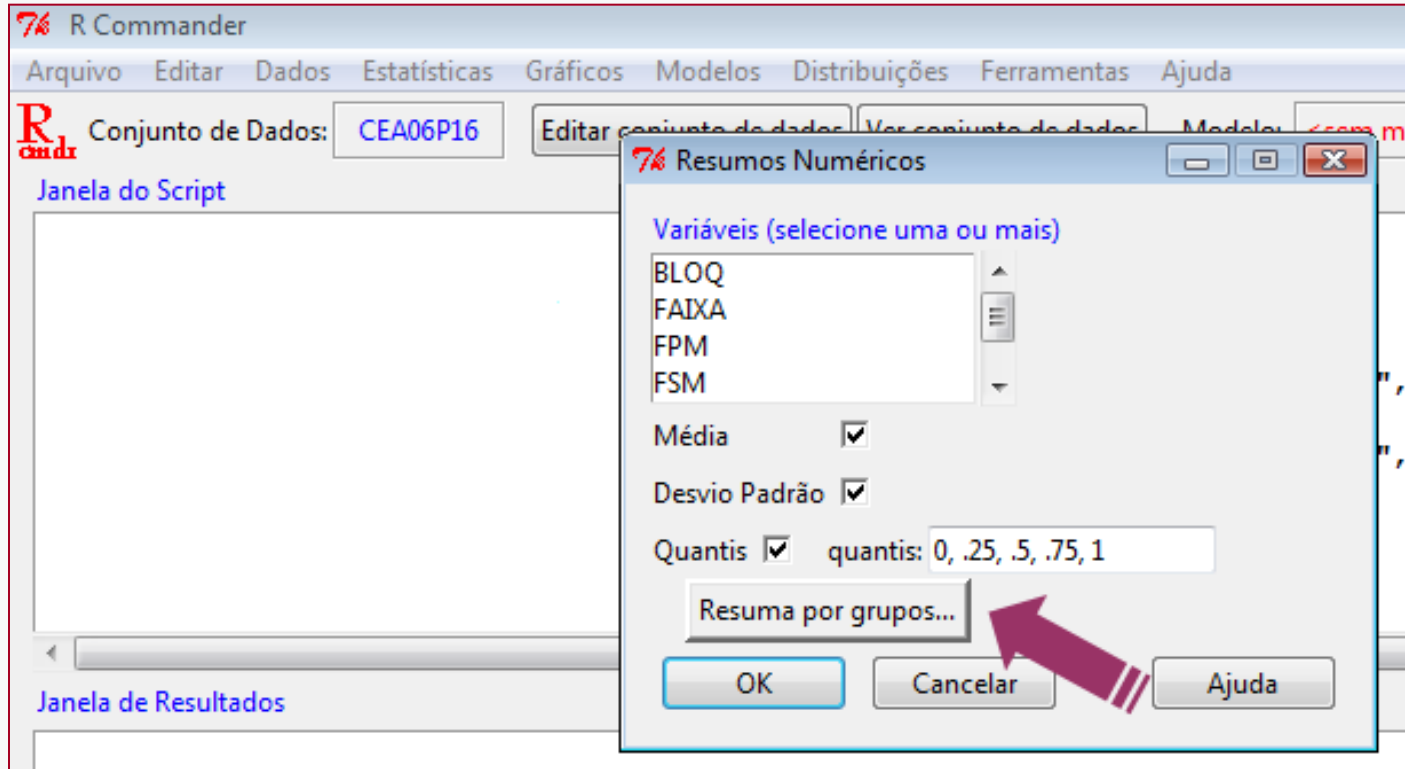
| | mean | sd | 0% | 25% | cv | |
|---------------|-------------|-----------|------------|------------|---------------------|----------|
| FPM | 98.76 | 29.94 | 28.7 | 78.05 | $29.94/98.76=0.30$ | |
| FSM | 179.29 | 54.71 | 53.8 | 142.80 | $54.71/179.29=0.30$ | |
| INTERJ | 4.36 | 4.41 | 0 | 1 | $4.41/4.36=1.01$ | |
| | 50% | | 75% | | 100% | n |
| FPM | 96.945 | | 117.98 | | 209.09 | 594 |
| FSM | 176.470 | | 214.29 | | 364.64 | 594 |
| INTERJ | 3 | | 6 | | 25 | 594 |

Alguns comentários:

- 50% dos indivíduos falaram até 3 *interjeições*;
- 25% dos entrevistados tiveram um *fluxo de palavras* menor ou igual a 78,05 palavras por minuto;
- o *fluxo de sílabas* de 75% dos indivíduos foi igual ou menor a 214,29 sílabas por minuto;
- A variável com maior dispersão em relação à média é *número de interjeições*;
- *Fluxo de sílabas* e *fluxo de palavras* apresentam dispersão em relação à média praticamente iguais.

Medidas descritivas por sexo

Rcmdr



**Estatísticas → Resumos Numéricos →
Resuma por grupos**

Medidas descriptivas por sexo

Variable: FPM

| | mean | sd | 0% | 25% | 50% | 75% | 100% | n |
|---|-------|-------|------|-------|-------|--------|--------|-----|
| F | 99.34 | 29.69 | 28.7 | 79.52 | 98.70 | 118.93 | 209.09 | 349 |
| M | 97.95 | 30.33 | 34.3 | 76.20 | 96.39 | 117.80 | 181.62 | 245 |

Variable: INTERJ

| | mean | sd | 0% | 25% | 50% | 75% | 100% | n |
|---|------|------|----|-----|-----|-----|------|-----|
| F | 4.52 | 4.55 | 0 | 1 | 4 | 6 | 25 | 349 |
| M | 4.13 | 4.19 | 0 | 1 | 3 | 6 | 25 | 245 |

Alguns comentários:

- Medidas de posição: mulheres apresentam medidas um pouco maiores do que homens tanto para o *fluxo de palavras por minuto* quanto *número de interjeições* utilizadas.
- Medidas de dispersão: mulheres apresentam dispersão muito próxima à de homens para o *fluxo de palavras por minuto*. A dispersão relativamente à média também está muito próxima (0,30 e 0,31 para mulheres e homens, respectivamente).

Descrevendo o fluxo de palavras por minuto por escolaridade

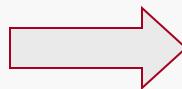
| ESCOLARIDADE | mean | sd | 0% | 25% | 50% | 75% | 100% | n |
|----------------------|---------------|---------------|--------------|--------------|---------------|---------------|---------------|------------|
| Pre-escola | 84.79 | 25.12 | 28.70 | 68.97 | 83.65 | 97.92 | 157.50 | 100 |
| Fund. incomp. | 86.95 | 28.72 | 29.60 | 64.66 | 85.30 | 105.20 | 209.09 | 165 |
| Fundamental | 112.14 | 29.66 | 61.12 | 91.18 | 108.10 | 133.56 | 181.62 | 48 |
| Medio incomp. | 109.33 | 26.19 | 54.24 | 92.96 | 105.94 | 124.62 | 174.00 | 65 |
| Medio | 105.19 | 24.45 | 60.63 | 85.38 | 102.72 | 128.53 | 150.67 | 44 |
| Superior | 105.17 | 25.74 | 67.34 | 85.84 | 102.24 | 117.47 | 172.31 | 44 |
| Sem resposta | SR | 110.12 | 30.19 | 44.68 | 89.58 | 107.80 | 125.66 | 128 |

Os dados também podem ser resumidos construindo-se uma tabela de distribuição de frequências.

Distribuição de frequências de uma variável é uma lista dos valores individuais ou dos intervalos de valores que a variável pode assumir, com as respectivas frequências de ocorrência.

No arquivo *CEA06P16*

Variável
IDADE



Não há perda
de informação



Distribuição de frequências, var. cont.

Idade Freq.Abs. Porcent.

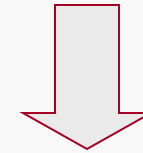
| | | |
|----|---|------|
| . | . | . |
| . | . | . |
| . | . | . |
| 29 | 5 | 0.84 |
| 30 | 5 | 0.84 |
| 31 | 5 | 0.84 |
| 32 | 4 | 0.67 |
| 33 | 4 | 0.67 |
| 34 | 2 | 0.34 |
| 35 | 2 | 0.34 |
| 36 | 1 | 0.17 |
| 37 | 2 | 0.34 |
| 38 | 1 | 0.17 |
| 39 | 4 | 0.67 |
| 40 | 2 | 0.34 |
| 41 | 2 | 0.34 |
| 42 | 4 | 0.67 |
| 43 | 5 | 0.84 |
| 44 | 1 | 0.17 |
| 45 | 6 | 1.01 |
| 46 | 5 | 0.84 |
| 47 | 4 | 0.67 |
| 48 | 4 | 0.67 |
| . | . | . |
| . | . | . |
| . | . | . |

N= 594

Alternativa: construir intervalos de classe

| Classe de Idade | frequência |
|------------------------|-------------------|
| 2,0 - 4,0 | 60 |
| 4,0 - 6,0 | 40 |
| 6,0 - 9,0 | 60 |
| 9,0 - 11,0 | 40 |
| 11,0 - 14,0 | 65 |
| 14,0 - 16,0 | 40 |
| 16,0 - 22,9 | 41 |
| 22,9 - 36,3 | 50 |
| 36,6 - 50,0 | 51 |
| 50,0 - 68,0 | 51 |
| 68,0 - 78,0 | 50 |
| 78,0 - 97,0 | 46 |
| Total | 594 |

Informações mais resumidas



Perda de informação

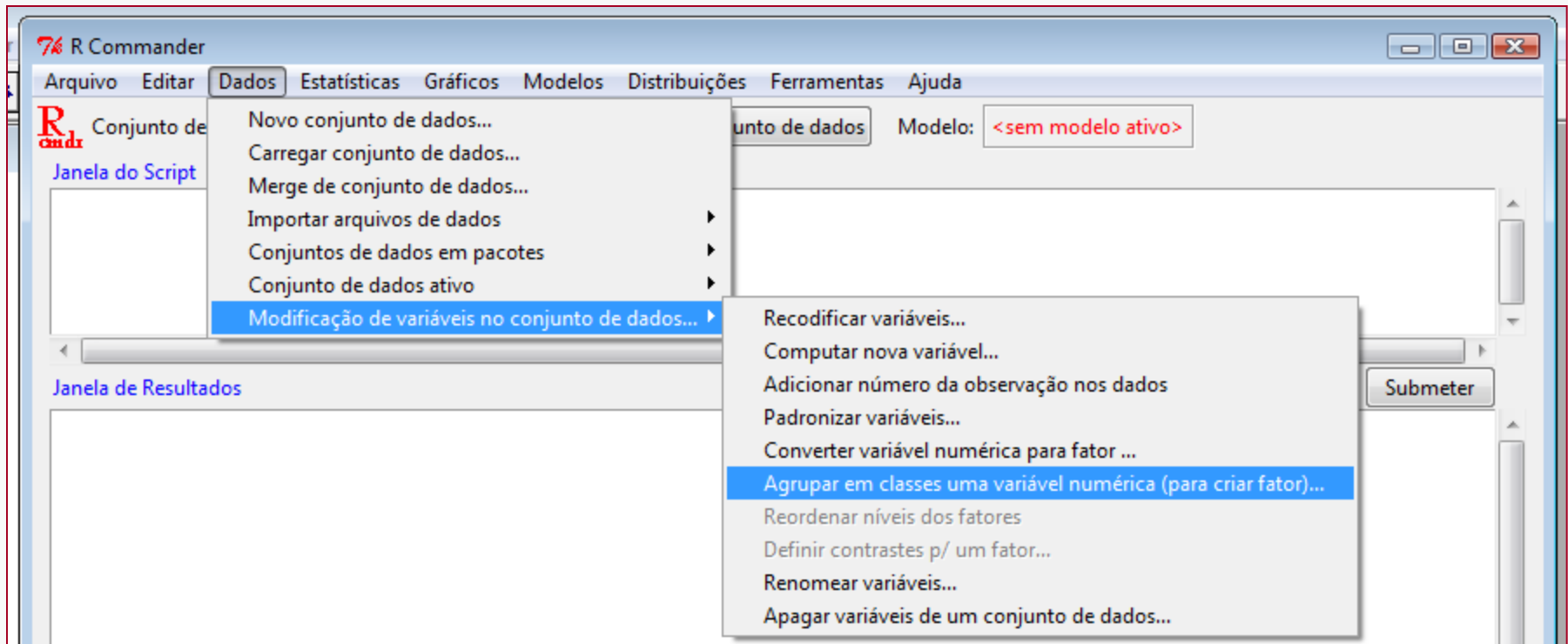
Exemplo 1:

Variável: Número de interjeições (INTERJ)

→ quantitativa →

Construir intervalos de classe

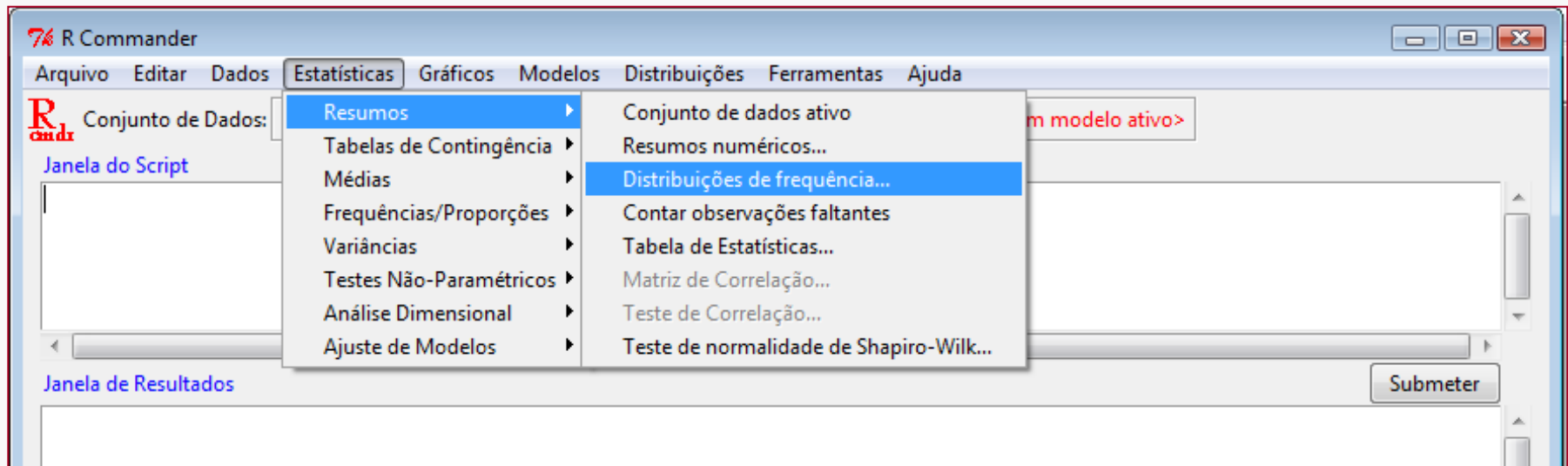
Rcmdr: (1) deve-se criar nova variável



Exemplo 1:

Rcmdr:

(2) deve-se obter a distribuição de frequências da nova variável



Exemplo 1: Variável número de interjeições

Distribuição de frequência para INTERJ

| Classes de INTERJ | f | fr (%) |
|--------------------------|------------|---------------|
| 0 - 5 | 360 | 60,61 |
| 5 - 10 | 165 | 27,78 |
| 10 - 16 | 54 | 9,09 |
| 16 - 21 | 10 | 1,68 |
| 21 - 25 | 5 | 0,84 |
| Total | 594 | 100 |

Variáveis Quantitativas

Gráficos mais comuns

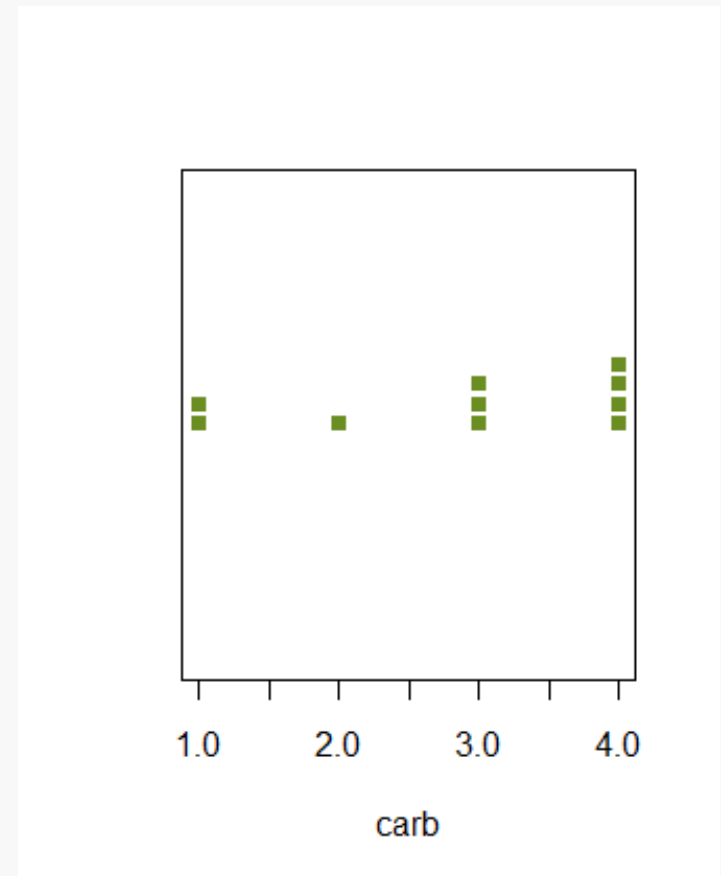
- “Strip Chart” ou “Dotplot”
- “Boxplot”
- Histograma

STRIP CHART ou DOT PLOT

Exemplo: Dados de *performance* e *design* de 10 modelos de carros (1973-74) retirados do arquivo *mtcars* (disponível no R)

Variável: número de carburadores

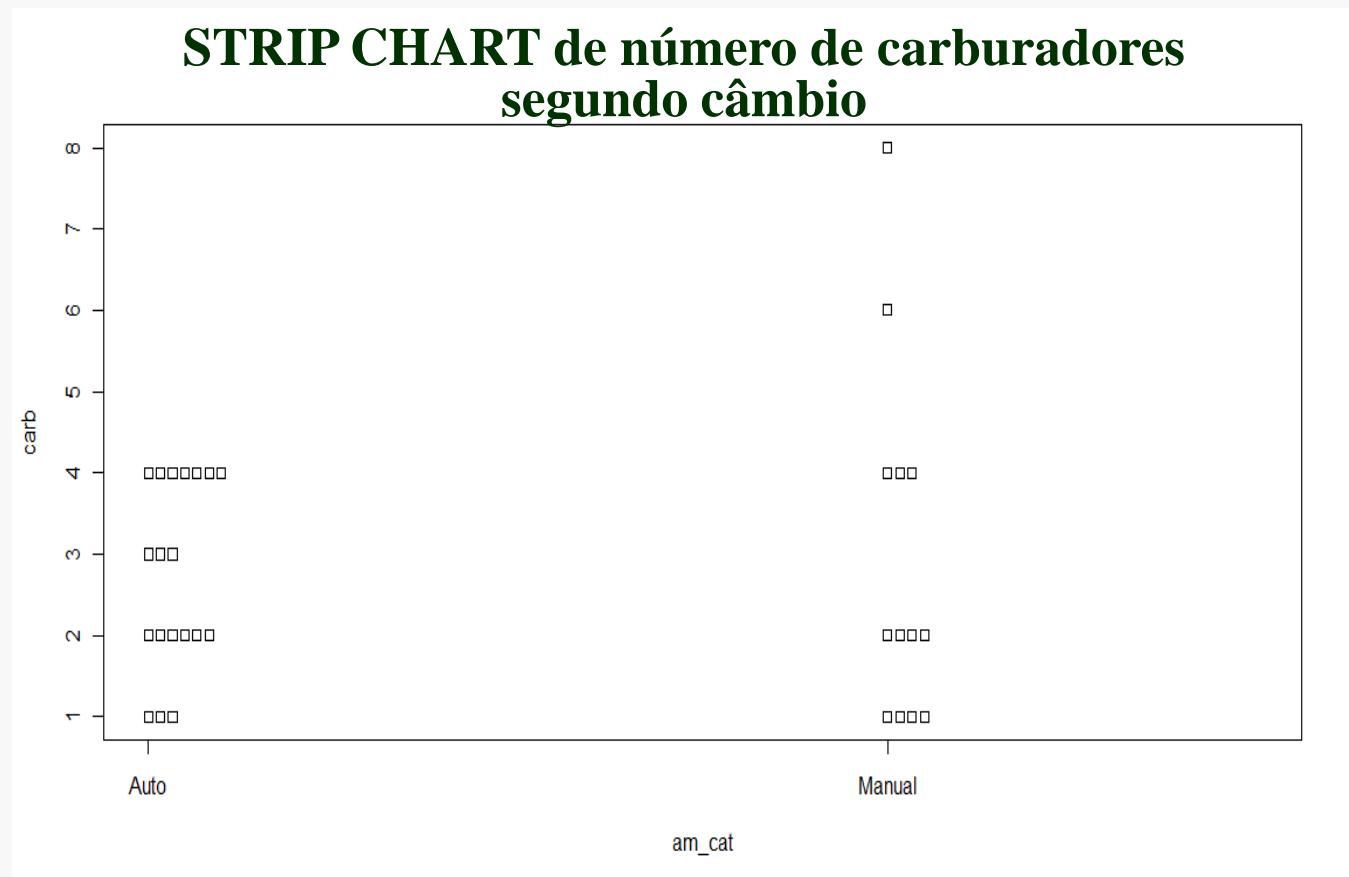
| | row.names | carb |
|----|---------------------|------|
| 11 | Merc 280C | 4 |
| 12 | Merc 450SE | 3 |
| 13 | Merc 450SL | 3 |
| 14 | Merc 450SLC | 3 |
| 15 | Cadillac Fleetwood | 4 |
| 16 | Lincoln Continental | 4 |
| 17 | Chrysler Imperial | 4 |
| 18 | Fiat 128 | 1 |
| 19 | Honda Civic | 2 |
| 20 | Toyota Corolla | 1 |



Exemplo: Dados de *performance e design* de **32** modelos de carros (1973-74) retirados do arquivo *mtcars* (disponível no R)

Variáveis:

- Número de carburadores
- Câmbio: manual ou automático



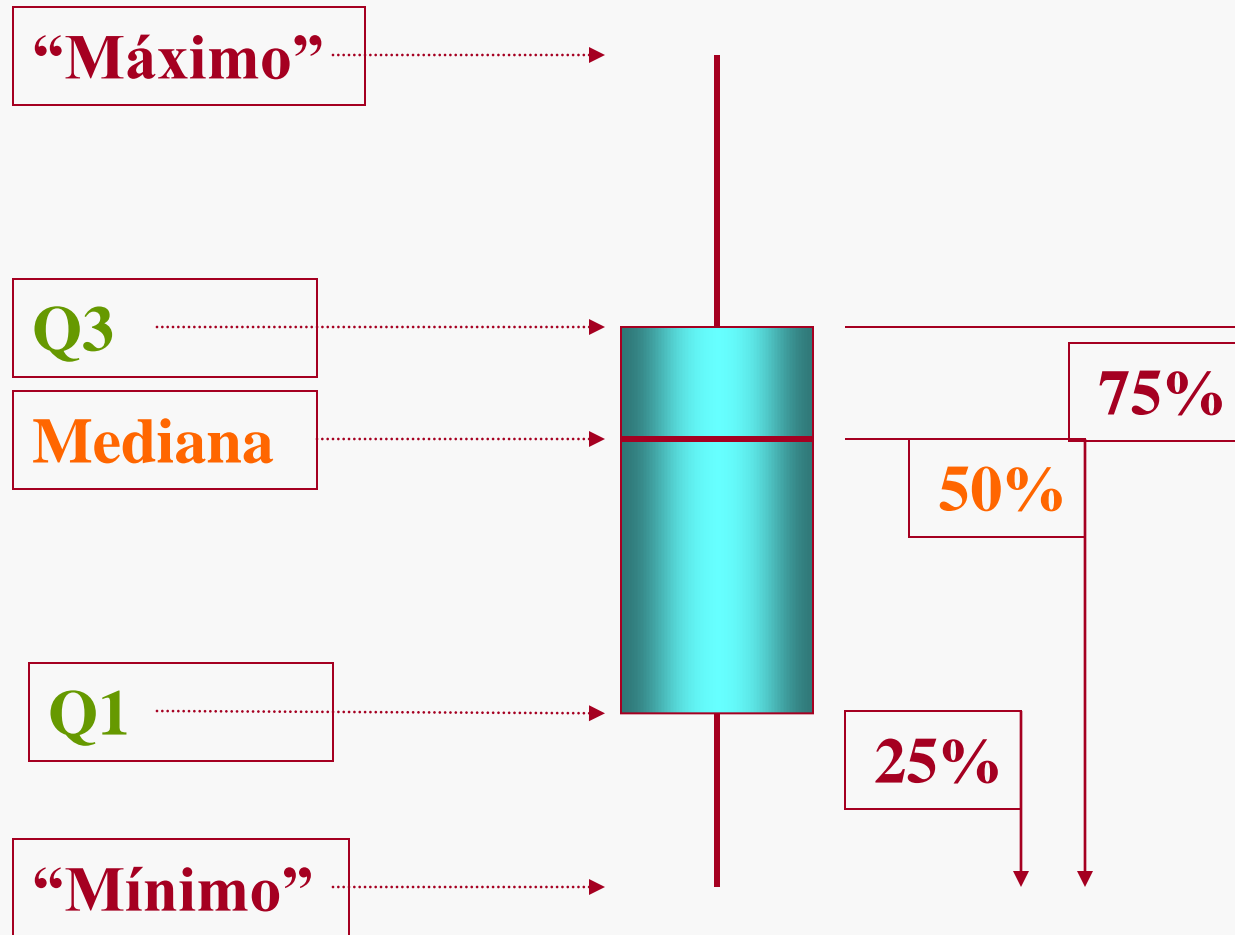
Notar que os *Strip Charts* são construídos na mesma escala.

Boxplot

Representa os dados através de um retângulo construído com os **quartis** e fornece várias informações, incluindo a existência de **valores extremos**.

Construção

$$LS=Q3+1,5(Q3-Q1)$$



$$LI=Q1-1,5(Q3-Q1)$$

“Máximo” é o maior valor menor que LS ;

“Mínimo” é o menor valor maior que LI .

Exemplo: Tempo de sobrevivência (dias) após cirurgia

Dados ordenados ($n=36$)

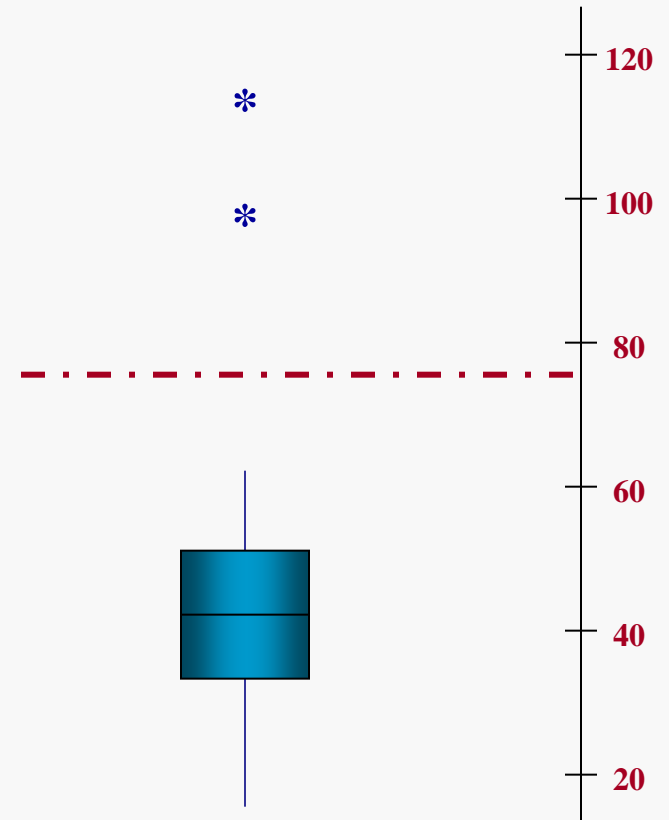
md = 41,5 Q1 = 30,25 Q3 = 49,5

| | | | | | |
|----|----|----|----|----|-----|
| 18 | 21 | 21 | 23 | 23 | 25 |
| 27 | 29 | 30 | 31 | 32 | 32 |
| 32 | 34 | 35 | 36 | 38 | 41 |
| 42 | 42 | 43 | 44 | 45 | 46 |
| 46 | 47 | 48 | 50 | 54 | 56 |
| 57 | 58 | 60 | 61 | 98 | 116 |

Observações discrepantes?

$$LI = Q1 - 1,5(Q3 - Q1) = 1,38$$

$$LS = Q3 + 1,5(Q3 - Q1) = 78,38$$



Exemplo 2:

Dados *CEA06P24* do projeto *Caracterização Postural de Crianças de 7 e 8 anos das Escolas Municipais da Cidade de Amparo/SP*

- Estudo realizado pelo Departamento de Fisioterapia, Fonoaudiologia e Terapia Ocupacional da Faculdade de Medicina da USP
- Ano de realização da análise: 2006
- Finalidade: mestrado
- Análise Estatística: Centro de Estatística Aplicada (CEA), IME-USP



Exemplo 2: *Caracterização Postural*

- Variações de postura da criança, associadas aos estágios de crescimento: resposta aos problemas de equilíbrio devido às mudanças nas proporções do corpo.
- **Objetivo:** caracterizar a postura de crianças da cidade de Amparo/SP, entre sete e oito anos, de ambos os sexos
- **Amostra:** 230 crianças com 7 e 8 anos.
- Medidas de postura das crianças foram obtidas.

Exemplo 2: *Estudo Caracterização Postural*

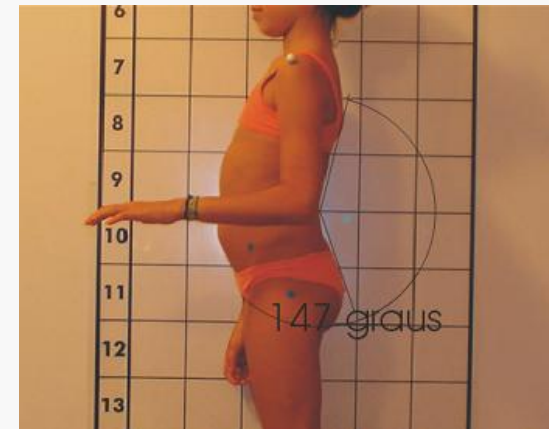
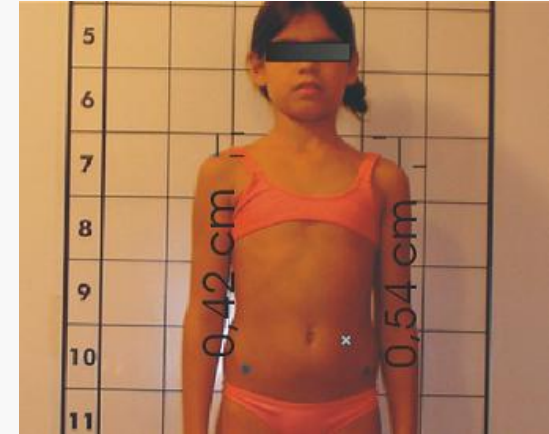
Algumas variáveis:

- Sexo (feminino, masculino);
- Peso (em kg);
- Altura (em metros);
- Índice de Massa Corpórea - IMC(em kg/m^2);
- Atividade Física (em horas/semana);
- Tipo de Mochila Utilizada (com fixação escapular, com fixação lateral, de carrinho, outros);
- Dominância (destro, canhoto);
- Região da escola;

Exemplo 2: *Estudo Caracterização Postural*

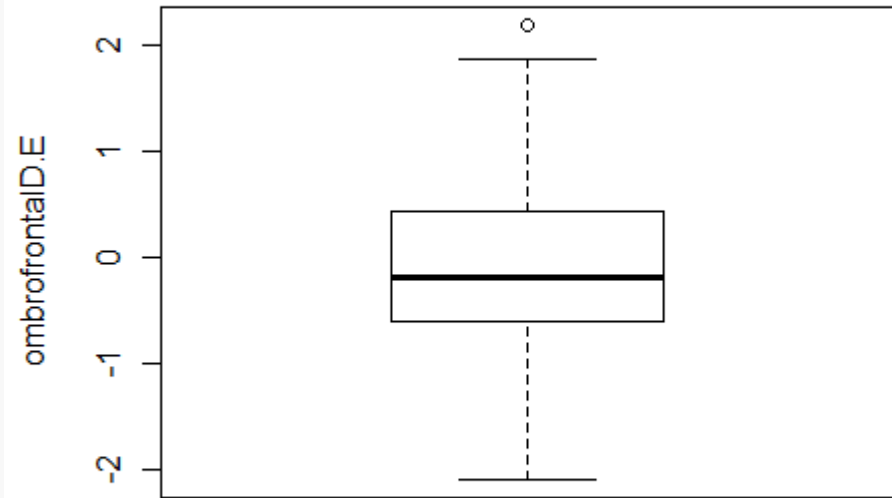
Algumas variáveis relativas a postura:

- Postura do ombro no plano frontal (cm): avaliado pelo desnível entre os ombros, conforme figura; anota-se a diferença Direito-Esquerdo;
- Lordose Lombar (graus): avaliada pelo aumento e diminuição (retificação) da lordose lombar, medindo-se o ângulo formado entre os pontos de maior convexidade da coluna torácica e da região glútea e o ponto de maior concavidade da coluna lombar, em ambos lados (Direito e Esquerdo).



Arquivo *CEA06P24* – *Boxplot* do desnível dos ombros

Gráficos → Box Plot

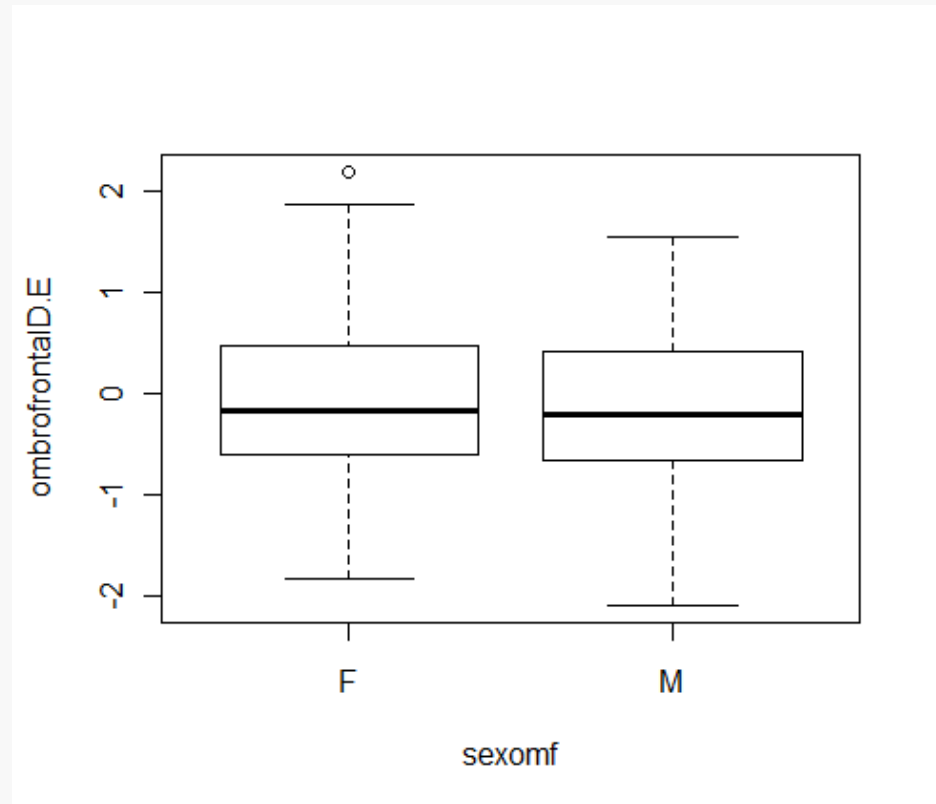


Alguns Comentários:

- há uma observação discrepante;
- a distribuição dos valores parece um pouco assimétrica.

Arquivo *CEA06P24* – *Boxplots* do desnível dos ombros segundo Sexo

Gráficos → Box Plot
→ Gráfico por grupos



Alguns Comentários:

- há uma observação discrepante para meninas;
- não há observações discrepantes para meninos;
- medidas de posição tendem a ser próximas para os dois sexos.

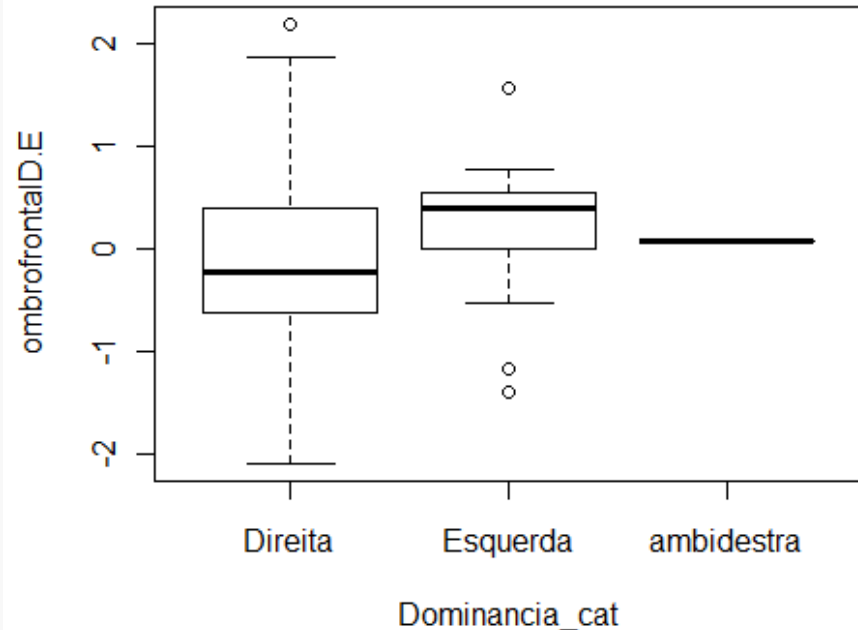
Arquivo *CEA06P24* – *Boxplots* do desnível dos ombros segundo Dominância

Frequências:

Direita 212

Esquerda 17

Ambidestra 1



Alguns Comentários:

- Note que só há uma criança ambidestra;
- Há observações discrepantes para dominância esquerda e direita;
- Distribuição dos valores bem diferente para as duas dominâncias.

Histograma

Agrupar os dados em intervalos de classes
(distribuição de frequências)

Bases iguais

Construir um retângulo para cada classe, com base igual ao tamanho da classe e *altura proporcional à frequência da classe (f)*.

Bases diferentes

Construir um retângulo para cada classe, com base igual ao tamanho da classe e *área do retângulo igual a frequência relativa da classe (fr)*. A altura será dada por $h = fr/\text{base}$ (densidade de frequência).

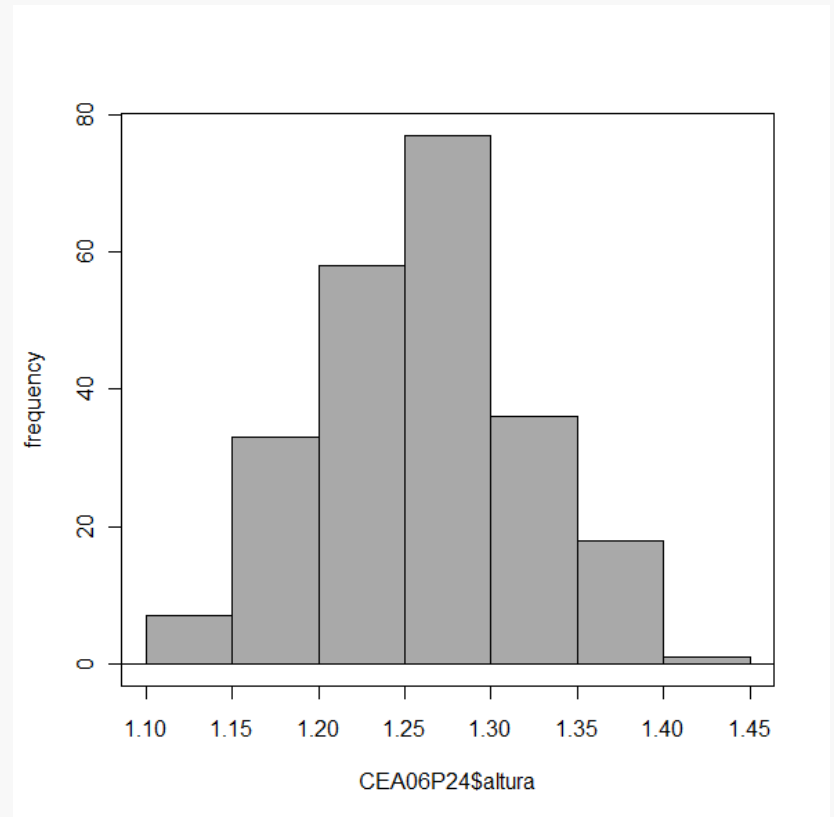
Arquivo *CEA06P24*– Histograma da altura

Distribuição de frequências para altura (arquivo *CEA06P24*)

| Classe de altura | f | fr (%) |
|------------------|---|--------|
|------------------|---|--------|

| | | |
|-------------|----|-------|
| 1,10 – 1,15 | 7 | 3,04 |
| 1,15 – 1,20 | 33 | 14,35 |
| 1,20 – 1,25 | 58 | 25,22 |
| 1,25 – 1,30 | 77 | 33,48 |
| 1,30 – 1,35 | 36 | 15,65 |
| 1,35 – 1,40 | 18 | 7,83 |
| 1,40 – 1,45 | 1 | 0,43 |

| | | |
|--------------|------------|------------|
| Total | 230 | 100 |
|--------------|------------|------------|

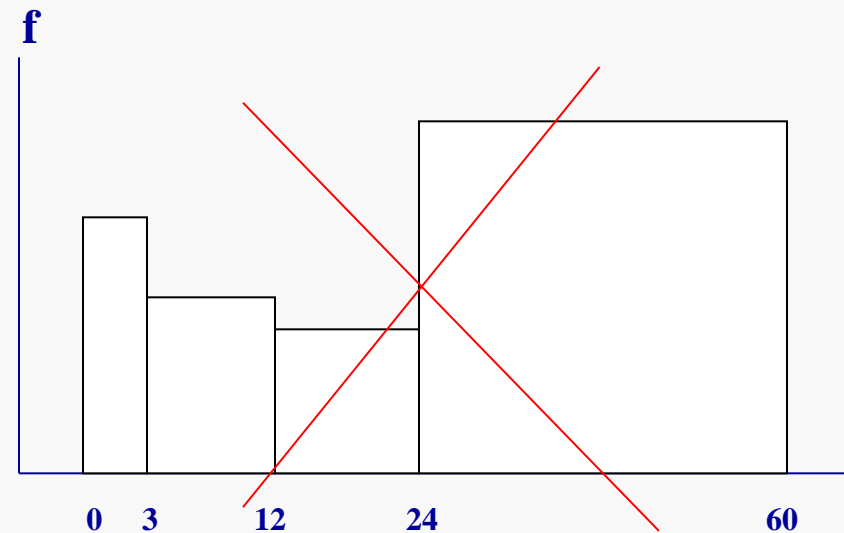
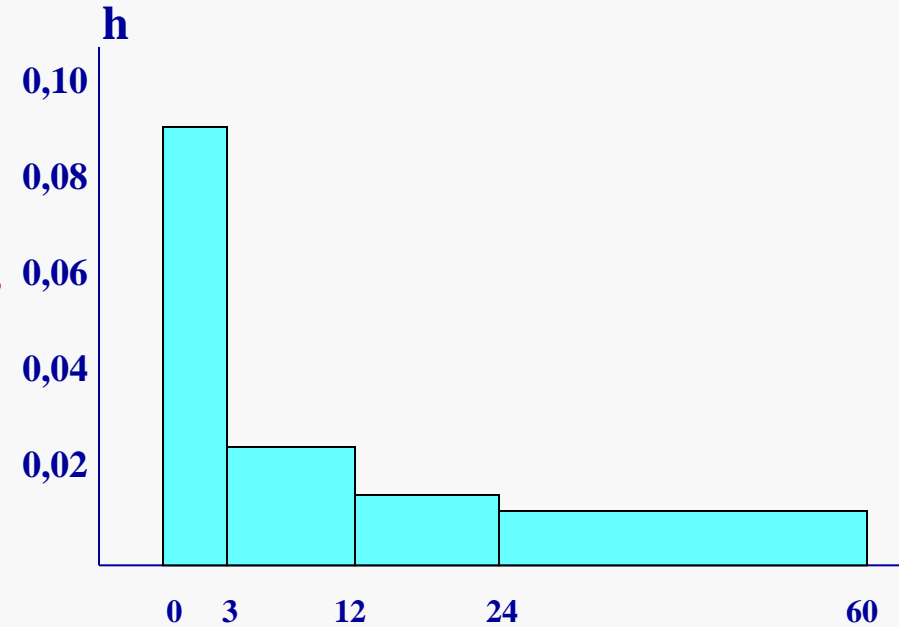


Gráficos → Histograma...

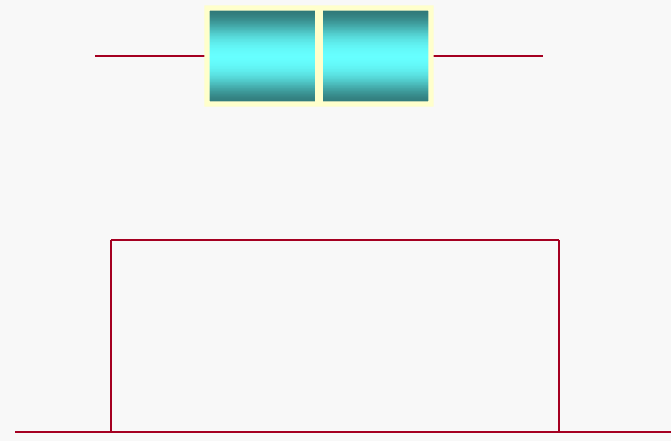
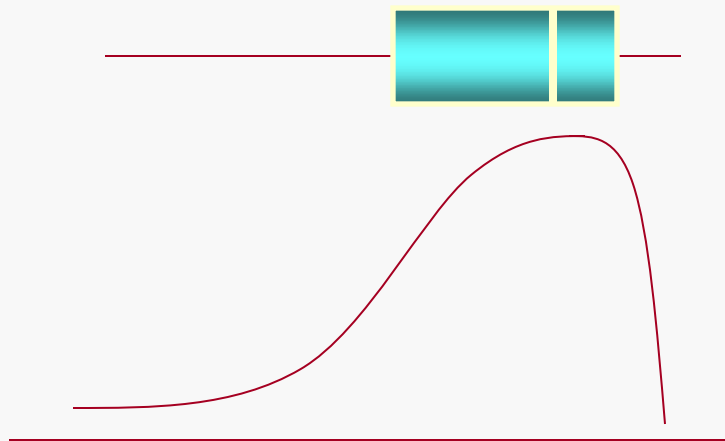
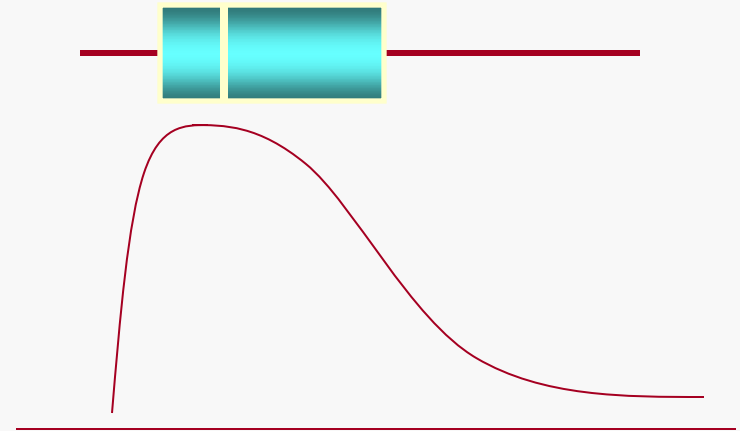
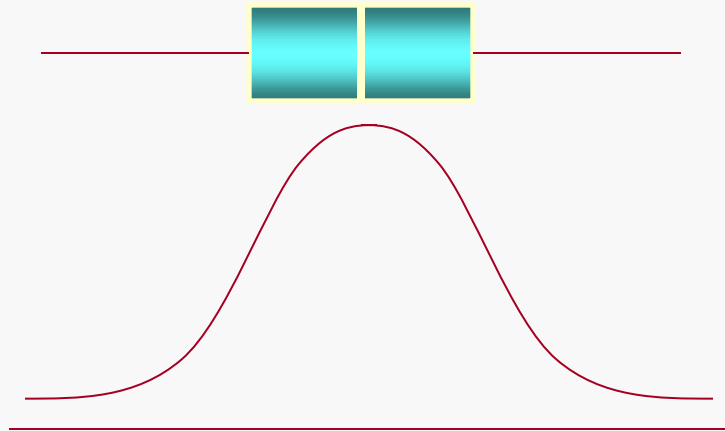
Exemplo: Classes desiguais

Distribuição das idades (em meses) de uma amostra de 500 crianças vacinadas

| Classes (meses) | f | fr | h |
|-----------------|------------|-------------|-------|
| 0 - 3 | 140 | 0,28 | 0,093 |
| 3 - 12 | 100 | 0,20 | 0,022 |
| 12 -24 | 80 | 0,16 | 0,013 |
| 24 -60 | 180 | 0,36 | 0,010 |
| Total | 500 | 1,00 | |



Forma da Distribuição



Variáveis Qualitativas

Os dados podem ser resumidos construindo-se uma tabela de distribuição de frequências, que quantifica a frequência das distintas categorias.

Variáveis qualitativas no arquivo *CEA06P24*
(postura)

Dominância

Sexo

Tipo de mochila

Transformando uma variável quantitativa numa variável qualitativa

The image shows the R Commander interface with the 'Dados' menu open. The 'Converter variável numérica para fator ...' option is selected. Below, two dialog boxes are shown: 'Converter Variáveis Numéricas p/ Fator' and 'Nomes dos níveis para tipom...'. The first dialog shows 'sujeito' and 'tipomochila' selected as variables, with 'Defina nomes dos níveis' chosen. The second dialog shows a mapping of numerical values to categorical labels.

| Valor numérico | Nome do nível |
|----------------|---------------|
| 1 | Escapular |
| 2 | Lateral |
| 3 | Carinho |
| 4 | Outro |

The image shows two windows from the R Commander interface. The left window is the R Console, and the right window is the R Commander main interface.

R Console:

```

Digite 'demo()' para demonstrações, 'help()' para o sistema on-line de ajuda,
ou 'help.start()' para abrir o sistema de ajuda em HTML no seu navegador.
Digite 'q()' para sair do R.

> local(pkg <- select.list(sort(.packages(all.available = TRUE)),graphics=TRUE)
+ if(nchar(pkg)) library(pkg, character.only=TRUE))
Carregando pacotes exigidos: tcltk
Loading Tcl/Tk interface ... done
Carregando pacotes exigidos: car
Carregando pacotes exigidos: MASS
Carregando pacotes exigidos: nnet
Carregando pacotes exigidos: survival
Carregando pacotes exigidos: splines

Versão do Rcmdr 1.6-3

Anexando pacote: 'Rcmdr'

The following object(s) are masked from 'package:tcltk':

  tclvalue

Carregando pacotes exigidos: RODBC
>

```

R Commander:

The R Commander window shows a menu for 'Estadísticas' (Statistics) with 'Distribuições de frequência...' (Frequency Distributions) selected. The console shows the following code and results:

```

CEAO6P24 <- sql
names(CEAO6P24)
library(nbind,
numSummary(CEAO
  statistics=c(
.Table <- table
.Table # counts for sexomf
round(100*.Table/sum(.Table), 2) # percentajes for sexomf
remove(.Table)

Janela de Resultados
Submeter

> numSummary(CEAO6P24[, "dominancia"], groups=CEAO6P24$sexomf,
+ statistics=c("mean", "sd", "quantiles"), quantiles=c(0,.25,.5,.75,1))
  mean      sd 0% 25% 50% 75% 100%  n
F 1.10 0.3258058  1  1  1  1  3  130
M 1.06 0.2386033  1  1  1  1  2  100

> .Table <- table(CEAO6P24$sexomf)

> .Table # counts for sexomf

  F  M
130 100

> round(100*.Table/sum(.Table), 2) # percentajes for sexomf

  F  M
56.52 43.48

> remove(.Table)

```

Cálculo de frequências e porcentagens

Variáveis qualitativas no arquivo *CEA06P24*

Medidas descritivas para variáveis qualitativas

| Sexo | Freq. | Porc. | Dominância | Freq. | Porc. |
|-------------|--------------|--------------|-------------------|--------------|--------------|
| M | 130 | 56,52 | Direita | 212 | 92,17 |
| F | 100 | 43,48 | Esquerda | 17 | 7,39 |
| N= | 230 | | Ambidestra | 1 | 0,43 |
| | | | N= | 230 | |

| Tipo Mochila | Freq. | Porc. |
|---------------------|--------------|--------------|
| Escapular | 123 | 53,48 |
| Lateral | 23 | 10,00 |
| Carrinho | 80 | 34,78 |
| Outros | 4 | 1,74 |
| N= | 230 | |

R Commander

Arquivo Editar Dados Estatísticas Gráficos Modelos Distribuições Ferramentas Ajuda

Conjunto de Dados: [dados] [Novo conjunto de dados] [Modelo: <sem modelo ativo>]

Janela do Script

```

.Table
rowPercents (.Table)
remove (.Table)
CEAO6P24$tipomochila
'Lateral', 'Ca
.Table <- table(
.Table # counts for tipomochila
round(100*.Table/sum(.Table), 2) # perc
remove (.Table)

```

Resumos

Tabelas de Contingência

Médias

Frequências/Proporções

Variâncias

Testes Não-Paramétricos

Análise Dimensional

Ajuste de Modelos

Tabela de dupla entrada...

Tabela multientrada...

Digite e analise tabela de dupla entrada...

Criando tabelas de contingência

Tabelas de dupla entrada

Variável linha (escolha uma)

regiao
sexomf
tipodedistjoelho
tipomochila

Variável coluna (escolha uma)

regiao
sexomf
tipodedistjoelho
tipomochila

Computar Percentagens

Percentual nas linhas

Percentual nas colunas

Percentagens do total

Sem percentual

Testes de Hipótese

Teste de independência de Qui-Quadrado

Componentes da estatística do Qui-quadrado

Apresente frequências esperadas

Teste exato de Fisher

Expressão (subset expression)

<todos casos válidos>

OK Cancelar Ajuda

Tabelas de dupla entrada

Variável linha (escolha uma)

regiao
sexomf
tipodedistjoelho
tipomochila

Variável coluna (escolha uma)

regiao
sexomf
tipodedistjoelho
tipomochila

Computar Percentagens

Percentual nas linhas

Percentual nas colunas

Percentagens do total

Sem percentual

Testes de Hipótese

Teste de independência de Qui-Quadrado

Componentes da estatística do Qui-quadrado

Apresente frequências esperadas

Teste exato de Fisher

Expressão (subset expression)

<todos casos válidos>

OK Cancelar Ajuda

Percentuais linha e Percentuais coluna

```
> rowPercents(.Table) # Row Percentages
```

```
tipomochila
```

| sexomf | Escapular | Lateral | Carrinho | Outro | Total | Count |
|--------|-----------|---------|----------|-------|-------|-------|
| F | 45.4 | 12.3 | 40.8 | 1.5 | 100 | 130 |
| M | 64.0 | 7.0 | 27.0 | 2.0 | 100 | 100 |

```
> colPercents(.Table) # Column Percentages
```

```
tipomochila
```

| sexomf | Escapular | Lateral | Carinho | Outro |
|--------|-----------|---------|---------|-------|
| F | 48 | 69.6 | 66.2 | 50 |
| M | 52 | 30.4 | 33.8 | 50 |
| Total | 100 | 100.0 | 100.0 | 100 |
| Count | 123 | 23.0 | 80.0 | 4 |

Variáveis Qualitativas

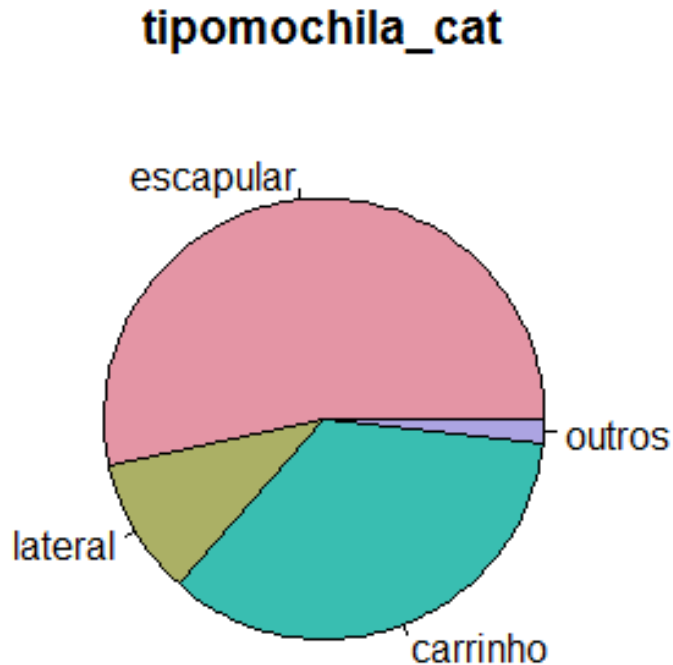
Gráficos

- Gráfico de setores
- Gráfico de barras

Gráfico de setores

Um círculo é dividido em tantos setores quantas forem as categorias da variável. A área de cada setor é proporcional à frequência da categoria.

Arquivo *CEA06P24* — Gráfico de setores para a variável Tipo de Mochila



Gráficos →
Gráfico de Pizza

Arquivo *CEA06P24* — Gráfico de setores para a variável Região da Escola

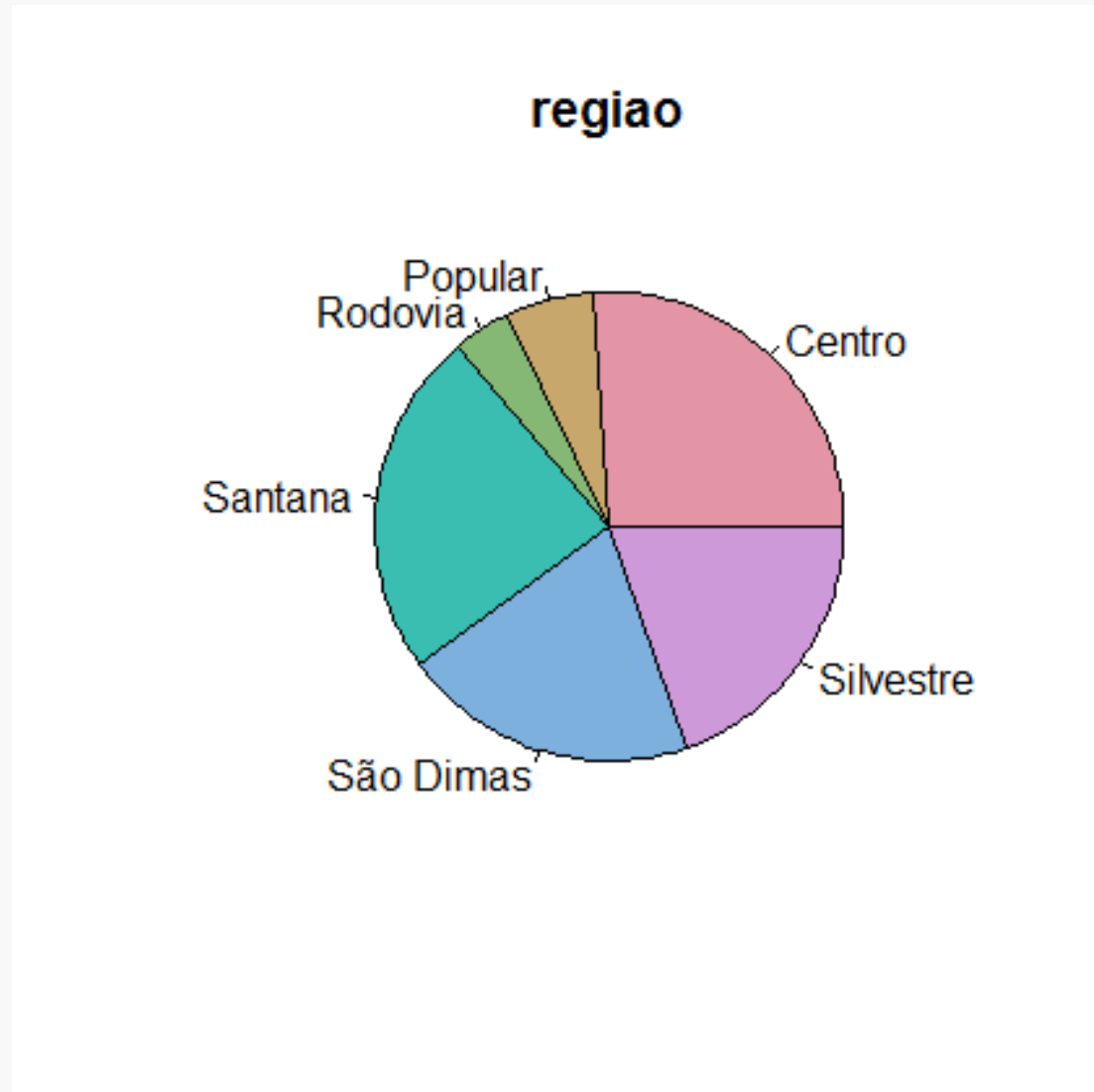
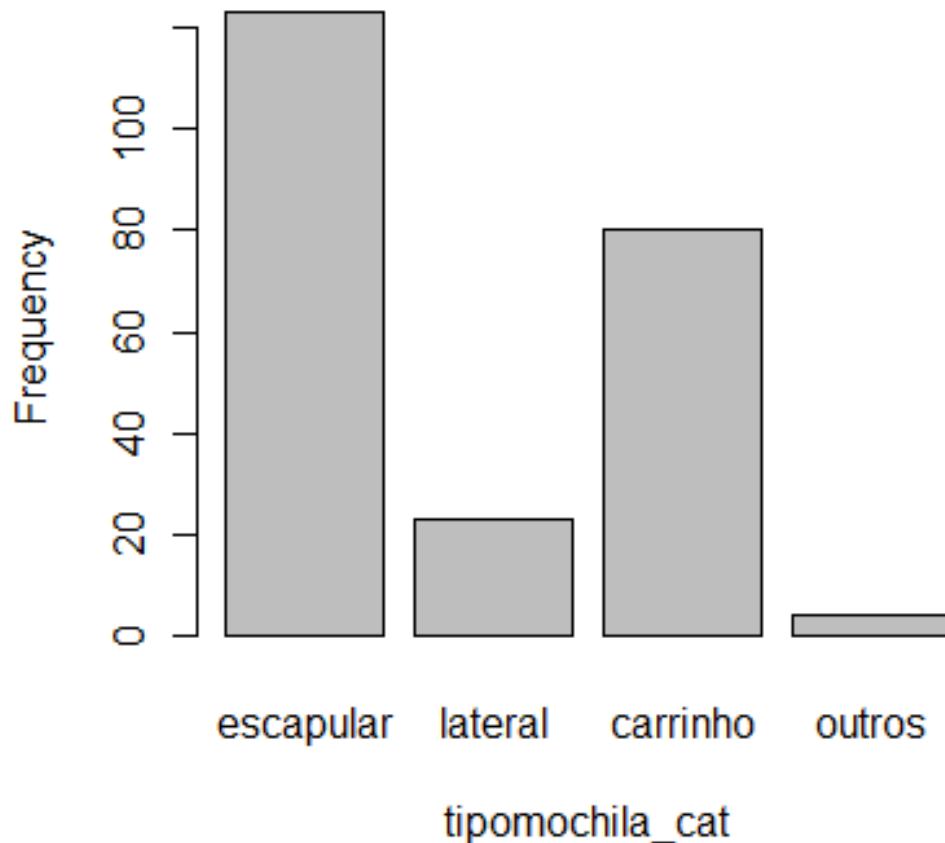


Gráfico de barras

Sobre um eixo, são representados retângulos, um para cada categoria da variável. A altura do retângulo é proporcional à frequência da categoria

Arquivo *CEA06P24* — Gráfico de barras para a variável *Tipo de mochila*



Gráficos → Gráfico de Barras

Arquivo *CEA06P24* - Gráfico de barras para a variável *Região da Escola*

