

## Aula Prática de Estatística

Os roteiros das aulas práticas no R são um misto de diversos materiais disponíveis na internet sobre cursos básicos do software. Dentre alguns, citam-se as notas dos professores Paulo Justiniano Ribeiro Jr (UFPR), Trevor Bailey (University of Exeter, UK), Bill Venables (CSIRO Austrália), Antônio Guilherme Fonseca Pacheco e Oswaldo Cruz (FIOCRUZ).

### Instalando o R e pacote Rcmdr:

- 1) Download do programa R para Windows: <http://cran.r-project.org/bin/windows/base/>

O R é uma linguagem e ambiente, de acesso livre e código aberto, que fornece ampla variedade de técnicas estatísticas e de gráficos. Vamos começar tendo uma ideia de seus recursos e da forma de trabalhar com este programa.

Siga os seguintes passos:

1. Inicie o R em seu computador;
2. Você verá uma janela de comandos com o símbolo `>`, que é o prompt do R, indicando que o programa está pronto para receber comandos;
- 2) Instalando o pacote Rcmdr: na linha de comandos, digite: `install.packages("Rcmdr",dep=TRUE)` e escolha o espelho – "*cran mirror*" como Brasil(RJ);
- 3) Chame a biblioteca com o comando: `library(Rcmdr)`.

### Alguns comandos básicos no R:

No *R Commander*, digite os comandos na janela **Janela da Script**. Com o cursor do mouse em cada linha de comando de interesse, clique no botão **Submeter**

- 1) Usando o R como uma calculadora: operações aritméticas: Você pode usar o R para avaliar algumas expressões aritméticas simples.

```
1 + 2 + 3
2 + 3 * 4
3/2 + 1
4 * 3**3
sqrt(2)
```

- 2) O R é uma linguagem orientada a objetos: variáveis, dados, matrizes, funções, etc, que são armazenados na memória ativa do computador na forma de objetos. Por exemplo, se um objeto `x` tem o valor 10, ao digitarmos o seu nome, o programa exibe o valor do objeto:

```
x <- 10
x
y <- sqrt(5)
y + x
```

**IMPORTANTE: O R distingue letras maiúsculas e minúsculas nos nomes dos objetos.**

- 3) Tipos de objetos: Os tipos básicos de objetos do R são: vetores; matrizes e *arrays*; *data-frames*; listas; e funções. Os quatro primeiros se diferem na forma de armazenar e operar com os dados. A função é um tipo de objeto especial, que recebe algum “input” e produz um “output”.

- Vetores: Vetores são o tipo básico e mais simples de objeto para armazenar dados no R. A função `c()` (“c” de concatenar) é usada para criar um vetor. Os colchetes `[ ]` são usados para indicar seleção de elementos. As funções `rep()`, `seq()` e o símbolo `:` são usadas para facilitar a criação de vetores que tenham alguma lei de formação.

```
x2 <- c(1, 3, 6)
x2[1]
x2[2]
length(x2)
x3 <- seq(0, 1, by = 0.1)
x3[x3 > 0.5]
x4 <- rep(1, 5)
```

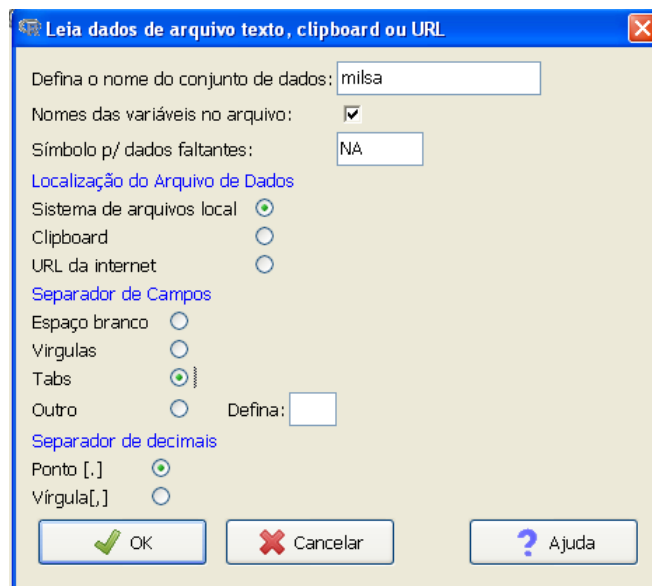
- Data-frames: Vetores, matrizes e *arrays* comportam elementos do mesmo “tipo” (numérico ou caracter). O “*data-frame*” é uma estrutura semelhante a uma matriz, porém com cada coluna sendo tratada separadamente. Podemos ter colunas de valores numéricos e de caracteres no mesmo objeto, mas em uma mesma coluna, os elementos ainda devem ser do mesmo tipo.

```
d1 <- data.frame(X = 1:10, Y = c(51, 54, 61, 67, 68, 75, 77, 75, 80, 82))
names(d1)
d1$X
d1$Y
```

## Estatística Descritiva

Para ilustrar algumas ferramentas de análise de dados, vamos usar dados hipotéticos de atributos de 36 funcionários da companhia “Milsa” (livro Estatística Básica -W. Bussab e P. Morettin). Estes dados são do tipo *data.frame*, armazenados no arquivo texto “milsa.txt”. Para ler o arquivo no R, siga os passos nas ferramentas do menu:

**Dados -> Importar arquivo de dados -> de arquivo texto, clipboard ou URL...** (arquivo “milsa.txt”).

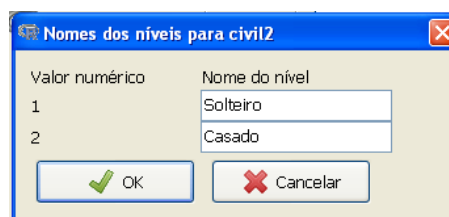
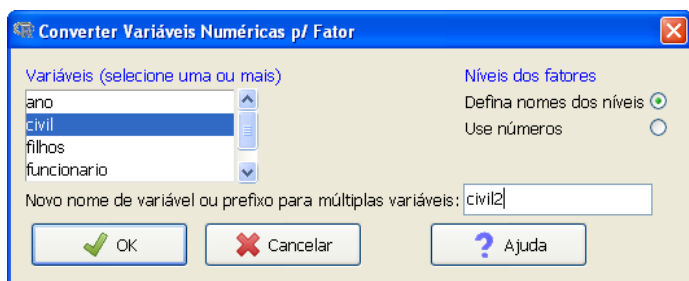


Para ver o banco de dados, clique na aba **Ver conjunto de dados**.

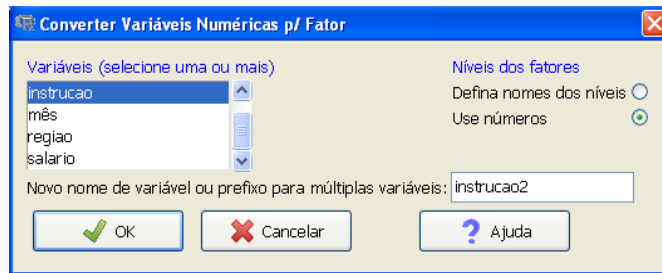
A planilha ainda não está pronta! Precisamos informar para o programa que as variáveis civil, instrução e região NÃO são numéricas e sim categóricas:

**Dados -> Modificação de variáveis no conjunto de dados... -> Converter variável numérica para fator...**

Primeiro, redefinimos a variável civil com os rótulos (*labels*) solteiro e casado associados aos níveis (*levels*) 1 e 2. Na variável região, codificamos assim: 1=interior, 2=capital, 3=outro.

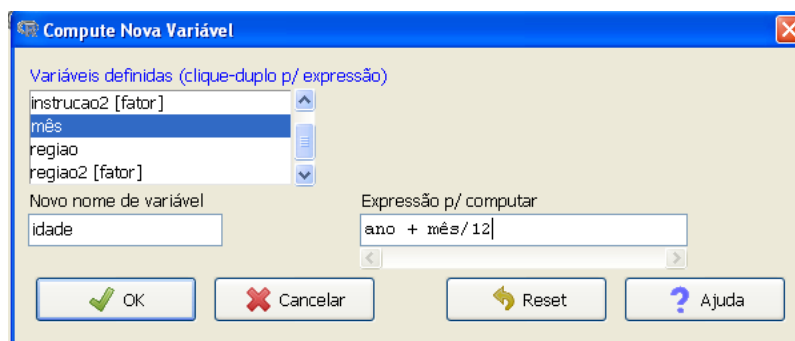


Faça o mesmo para região. Para instrução, apenas defina os níveis como os próprios números:



Vamos agora, definir uma nova variável denominada idade, em anos, a partir das variáveis ano e mês.

**Dados -> Modificação de variáveis no conjunto de dados... -> Computar nova variável**



A seguir, fazemos algumas análises descritivas uni e bi-variadas.

1) Análise Univariada: consiste basicamente em, para cada uma das variáveis individualmente:

- classificar a variável quanto a seu tipo: qualitativa (nominal ou ordinal) ou quantitativa (discreta ou contínua);
- obter tabelas, gráficos e/ou medidas que resumam a distribuição da variável.

Para exemplificar o uso dessas ferramentas, vamos selecionar uma variável de cada tipo.

### Variável Qualitativa Nominal

A variável civil é uma variável qualitativa nominal, assim podemos obter: (i) uma tabela de frequências (absolutas e/ou relativas), (ii) um gráfico de setores.

- (i) **Estatísticas -> Resumos -> Distribuições de frequências** (selecionar variável civil2)
- (ii) **Gráficos -> Gráfico de Pizza** (selecionar variável civil2). **ATENÇÃO:** A janela gráfica será aberta no programa R original.

Como mudar as cores do gráfico?

O comando que gerou o gráfico foi:

```
with(milsa, pie(table(civil2), labels=levels(civil2), xlab="", ylab="", main="civil2",
col=rainbow_hcl(length(levels(civil2)))))
```

Para mudar as cores precisamos modificar o argumento col (marcado em amarelo).

Digite na janela R script:

```
with(milsa, pie(table(civil2), labels=levels(civil2), xlab="", ylab="", main="civil2", col=c(4,7)))
```

Cores: 1 = preto, 2=vermelho, 3=verde, 4=azul escuro, 5=azul claro, 6=rosa, 7=amarelo, 8=cinza.

Como digitamos 4 e 7 no argumento col o gráfico será azul e amarelo.

Uma outra possibilidade seria escrever o nome das cores entre aspas.

```
with(milsa, pie(table(civil2), labels=levels(civil2), xlab="", ylab="", main="civil2", col=c("purple","pink")))
```

### Variável Qualitativa Ordinal

Para exemplificar como obter análises para uma variável qualitativa ordinal, vamos selecionar a variável instrucao2. As tabelas de frequências são obtidas de forma semelhante à mostrada anteriormente. O gráfico de setores não é adequado para este tipo de variável por não expressar a ordem dos possíveis valores. Usamos então, um gráfico de barras:

#### **Gráficos -> Gráfico de Barras** (selecionar variável instrucao2)

Como modificar os limites do eixo Y? Digite:

```
with(milsa, barplot(table(instrucao2), xlab="instrucao2", ylab="Frequency", ylim=c(0,20)))
```

### Variável quantitativa discreta

Vamos agora usar a variável filhos (número de filhos) para ilustrar algumas análises que podem ser feitas com uma variável quantitativa discreta. Esta deve ser uma variável numérica, e não um fator.

Além das tabelas de frequências para as variáveis numéricas são obtidas usando as funções table(milsa\$filhos) e prop.table(table(milsa\$filhos)), podemos fazer um gráfico de linhas verticais, um gráfico de pontos, além de calcular medidas sumárias:

#### **Estatísticas -> Resumos -> Resumos numéricos** (selecione filhos e as medidas de interesse)

**Gráficos -> Plot discrete numeric variable...**

**Gráficos -> Gráfico de pontos...**

Variável quantitativa Contínua

Para concluir os exemplos para a análise univariada, vamos considerar a variável quantitativa contínua salario. Podemos obter as estatísticas-resumo, além do histograma:

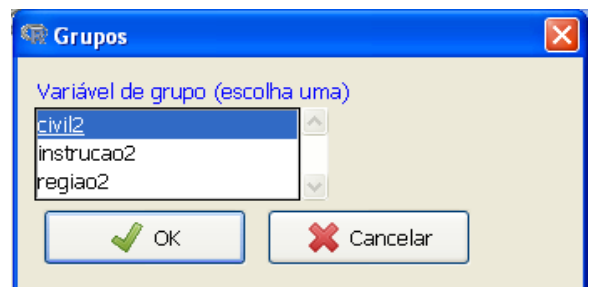
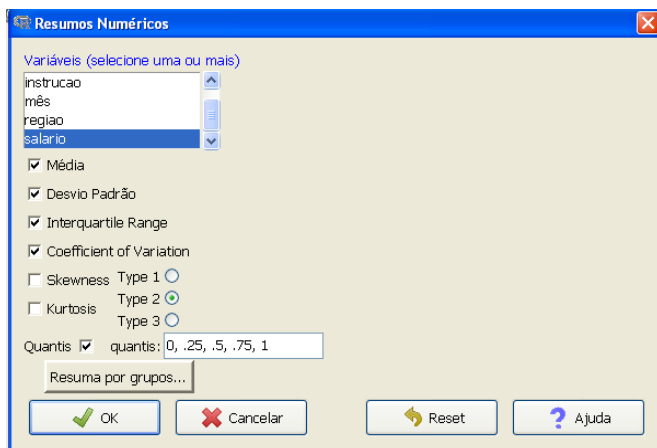
**Estatísticas -> Resumos -> Resumos numéricos** (selecione salario e as medidas de interesse)

**Gráficos -> Histograma...**

2) Análise Bivariada: procuramos identificar relações entre duas variáveis. O tipo de resumo estatístico informativo vai depender dos tipos das variáveis envolvidas.

Para exemplificar este caso vamos considerar as variáveis civil2 e salario. Podemos verificar as estatísticas resumo de salário para solteiros e casados, assim como fazer o boxplot. O teste de hipóteses será visto mais adiante.

**Estatísticas -> Resumos -> Resumos numéricos**



**Gráficos -> Boxplot...** (selecione a variável salário e peça **Gráfico por civil2**)

Para salvar o arquivo com as novas variáveis criadas

**Dados -> Conjunto de dados ativo -> Exportar conjunto de dados ativo...** (colocar como separador de campos o ponto e vírgula (;) pressione "OK", coloque como nome do arquivo "milsa2.txt")