Passo-a-passo

ETAPA 5. Estatísticas Descritivas

Prof. Pedro Feliú

Instituto de Relações Internacionais
Universidade de São Paulo

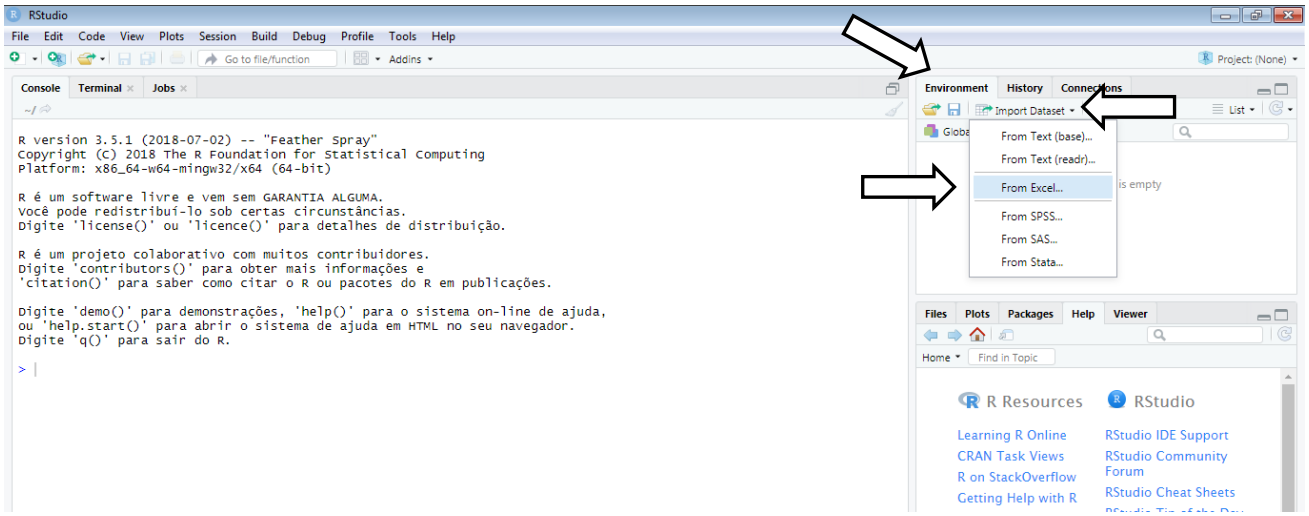
INTRODUÇÃO

Nesta etapa 5, vamos aprender a importar um banco de dados em Excel, alguns comandos para manipulação do banco de dados, a construção de uma tabela de frequências e uma tabela de estatísticas de tendência central e dispersão. As opções do Rstudio são muitas, nesta etapa selecionamos algumas funções introdutórias.

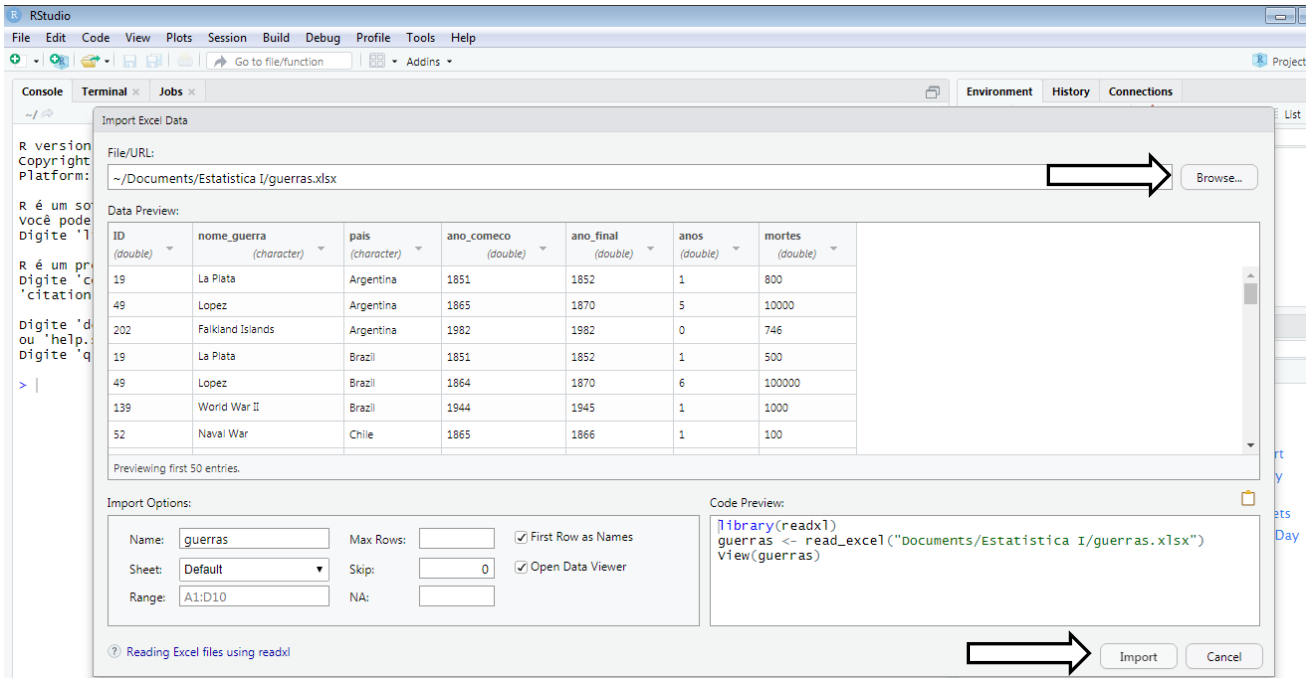
PASSO 1: Importar o banco de dados "guerras"

Começamos importando o banco de dados denominado "guerras", presente na página do moodle da disciplina, logo no começo da página, na pasta denominada Bases de Dados. Uma vez aberta a referida pasta do moodle, clique no arquivo "guerras.xlsx" (Excel) e salve o arquivo no seu computador. Já salvo o arquivo de Excel em seu computador, vamos abrir o Rstudio e começar a importação dos dados do Excel para o Rstudio.

No quadrante superior da direita na tela inicial do Rstudio, clique na aba "Environment" e logo abaixo clique em "Import Dataset", como as setas indicam na figura abaixo. Feito isso, clique na opção "From Excel...", iluminada em azul na figura abaixo.



Cumprida essa etapa, aparecerá uma janela com a opção de abrir o arquivo "guerras" salvo em seu computador. Para tanto clique no botão Browse e selecione o arquivo guerras em seu computador. Após isso, aparecerão: uma pre-visualização do banco de dados, algumas opções abaixo, como ler a primeira linha do banco como nome da variável e não a própria variável. Na visualização dos dados isso fica claro na primeira linha, cujos nomes das variáveis são: ID, nome_guerra, pais, ano_comeco, ano_final, anos, mortes. Todas bastante intuitivas, os nomes já dizem a que se referem. Na caixinha inferior a esquerda dessa janela encontram-se os comandos necessários para essa tarefa (realizada por clique aqui no exemplo, mas poderia ser feito por comando direto). Para finalizar a importação, clique no botão "Import", como indica a figura abaixo.



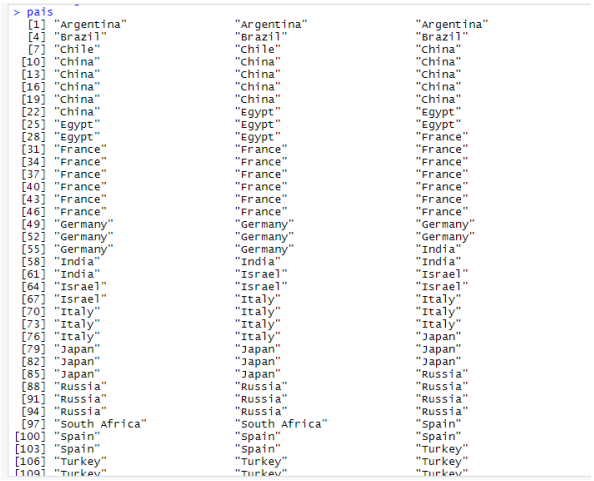
PASSO 2: Utilizar a função **attach()**

Apenas para teste, digite uma variável deste banco de dados, por exemplo, "pais" no console do R.

```
> library(readxl)
Warning message:
package 'readxl' was built under R version 3.5.3
> guerras <- read_excel("Documents/Estatistica I/guerras.xlsx")
> View(guerras)
> pais
Erro: objeto 'pais' não encontrado
>
```

Aparecerá um erro, escrito em vermelho: "objeto 'pais' não encontrado", como na figura acima. Agora digite os seguintes comandos:

attach(guerras)
pais



```
> pais
  [1] "Argentina"           "Argentina"           "Argentina"
  [4] "Brazil"              "Brazil"              "Brazil"
  [7] "Chile"               "Chile"               "China"
 [10] "China"               "China"               "China"
 [13] "China"               "China"               "China"
 [16] "China"               "China"               "China"
 [19] "China"               "China"               "China"
 [22] "China"               "Egypt"               "Egypt"
 [25] "Egypt"               "Egypt"               "Egypt"
 [28] "Egypt"               "Egypt"               "France"
 [31] "France"              "France"              "France"
 [34] "France"              "France"              "France"
 [37] "France"              "France"              "France"
 [40] "France"              "France"              "France"
 [43] "France"              "France"              "France"
 [46] "France"              "France"              "France"
 [49] "Germany"             "Germany"             "Germany"
 [52] "Germany"             "Germany"             "Germany"
 [55] "Germany"             "Germany"             "India"
 [58] "India"               "India"               "India"
 [61] "India"               "Israel"              "Israel"
 [64] "Israel"              "Israel"              "Israel"
 [67] "Israel"              "Italy"               "Italy"
 [70] "Italy"               "Italy"               "Italy"
 [73] "Italy"               "Italy"               "Italy"
 [76] "Italy"               "Italy"               "Japan"
 [79] "Japan"               "Japan"               "Japan"
 [82] "Japan"               "Japan"               "Japan"
 [85] "Japan"               "Japan"               "Russia"
 [88] "Russia"              "Russia"              "Russia"
 [91] "Russia"              "Russia"              "Russia"
 [94] "Russia"              "Russia"              "Russia"
 [97] "South Africa"        "South Africa"        "Spain"
[100] "Spain"               "Spain"               "Spain"
[103] "Spain"               "Spain"               "Turkey"
[106] "Turkey"              "Turkey"              "Turkey"
[109] "Turkey"              "Turkey"              "Turkey"
```

Agora, como podemos observar na figura acima, o Rstudio lê as variáveis do banco de dados enquanto objetos, assim, quando você digita o nome da variável no console do Rstudio, ele retorna os valores desta variável. A função **attach()**, portanto, torna as variáveis acessíveis apenas digitando os nomes delas. Nesse caso, são os nomes dos países do banco de dados.

Há outro caminho sem a utilização da função **attach()** para ler as variáveis. Vejamos:

detach(guerras)
guerras$pais

O primeiro comando, **detach()**, desfaz o **attach()** realizado anteriormente. O segundo comando, **guerras$pais**, indica que queremos ver a variável "pais" de nosso objeto chamado "guerras".

PASSO 3: Utilizar a função tapply(), ordenar e excluir variáveis

A função **tapply()** é muito útil quando temos várias categorias no banco de dados. A função possui o seguinte formato: **tapply(dados, grupos, função)**. Como exemplo, vamos extrair a soma das mortes contabilizadas nas guerras do banco de dados, que abrangem o século XIX e XX, por país:

attach(guerras)
tapply(mortes,pais,sum)

```
> attach(guerras)
> tapply(mortes,pais,sum)
             Argentina                 Brazil                  Chile                  China
                 11546                 101500                   3376                2617973
                 Egypt                 France                Germany                  India
                 37714                1870348                5330729                  10829
                Israel                  Italy                  Japan                 Russia
                  7850                 896500                2106826                2117997
          South Africa                  Spain                 Turkey         United Kingdom
                  8800                   7797                 657117                1350793
United States of America                   USSR
                651790                7634775
> |
```

Como podemos observar na figura acima, a URSS possui uma soma de 7634775 mortes nas guerras em que se envolveu, os EUA 651790, a China 2617973 e o Brasil 101500, por exemplo.

PASSO 4: Tipo de banco de dados e tipos de variáveis.

Para saber o tipo de base de dados que temos, assim como o tipo de variável utilizamos o comando **str()** abaixo. Aqui vamos ver a classificação, feita pelo R, das variáveis (quantitativa contínua e discreta; qualitativa nominal e ordinal).

str(guerras)

```
> str(guerras)
Classes 'tbl_df', 'tbl' and 'data.frame':       148 obs. of  7 variables:
 $ ID         : num  19 49 202 19 49 139 52 64 67 73 ...
 $ nome_guerra: chr  "La Plata" "Lopez" "Falkland Islands" "La Plata" ...
 $ pais       : chr  "Argentina" "Argentina" "Argentina" "Brazil" ...
 $ ano_comeco : num  1851 1865 1982 1851 1864 ...
 $ ano_final  : num  1852 1870 1982 1852 1870 ...
 $ anos       : num  1 5 0 1 6 1 1 4 1 1 ...
 $ mortes     : num  800 10000 746 500 100000 ...
> |
```

A função **str(guerras)**, como pode ser observado na figura acima, retorna quantas observações temos no banco de dados (148 guerras) e quantas variáveis (7). Ela também classifica os tipos de variáveis de nosso banco: “chr” (caractere) e “num” (numérica). chr (caractere) equivale a variável qualitativa nominal. No caso dessa base de dados são os nomes dos países que entraram em guerra. “num” significa uma variável quantitativa contínua ou discreta. Notem que o Rstudio não difere as quantitativas, entre discretas e contínuas, tratando ambas como a mesma coisa já que para efeitos estatísticos ambos são números. Não há nesse banco de dados um exemplo de variável qualitativa ordinal.

PASSO 5: Tabela de Frequência

Vamos agora fazer uma tabela de frequência utilizada no vídeo da aula 1. Para tanto precisamos instalar o pacote "summarytools" com os comandos já conhecidos nas etapas anteriores

install.packages("summarytools")
library(summarytools)

Agora vamos fazer uma tabela de frequência da quantidade de guerras travadas por cada país da base de dados. Vamos utilizar o seguinte comando abaixo, onde freq indica a função do pacote carregado, pais a variável qualitativa de interesse, cumul indica que queremos a frequência acumulada, totals indica que não queremos os totais e order =”freq” indica que queremos a tabela ordenada do maior (mais guerras) para o menor (menos guerras).

freq(pais, cumul = TRUE, totals = FALSE, order = "freq")

```
> freq(pais, cumul = TRUE, totals = FALSE, order = "freq")
Frequencies
pais
Type: Character

                              Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
---------------------------- ------ --------- -------------- --------- --------------
                          France     19     12.84          12.84     12.84          12.84
                           China     14      9.46          22.30      9.46          22.30
                  United Kingdom     13      8.78          31.08      8.78          31.08
      United States of America     13      8.78          39.86      8.78          39.86
                          Turkey     12      8.11          47.97      8.11          47.97
                           Italy     10      6.76          54.73      6.76          54.73
                          Russia     10      6.76          61.49      6.76          61.49
                           Japan      9      6.08          67.57      6.08          67.57
                         Germany      8      5.41          72.97      5.41          72.97
                           Egypt      7      4.73          77.70      4.73          77.70
                          Israel      6      4.05          81.76      4.05          81.76
                           Spain      6      4.05          85.81      4.05          85.81
                            USSR      6      4.05          89.86      4.05          89.86
                           India      5      3.38          93.24      3.38          93.24
                       Argentina      3      2.03          95.27      2.03          95.27
                          Brazil      3      2.03          97.30      2.03          97.30
                           Chile      2      1.35          98.65      1.35          98.65
                    South Africa      2      1.35         100.00      1.35         100.00
                            <NA>      0                                0.00         100.00
>
```

O comando retorna a tabela de frequência exibida na figura acima. A primeira coluna retorna os países da base de dados. A segunda coluna retorna a frequência absoluta. Podemos ver que França (19 guerras), China (14 guerras), Reino Unido (13 guerras) e EUA (13 guerras) são os países que mais entraram em guerras. Se somarmos URSS e Rússia (16 guerras), temos todos os membros permanentes do conselho de segurança da ONU. Nada é à toa nessa vida! Vamos explorar o tema na aula 2. Do outro lado, África do Sul (2), Chile (2), Brasil (3) e Argentina (3) são os países que menos travaram guerras. A terceira coluna indica a frequência relativa (% valid). Só a França, é responsável por 12, 84% das guerras dos dois últimos séculos, enquanto os EUA 8,78%. Brasil e Argentina ficam com 2,03% das guerras do período. Por fim, a quarta coluna (% valid cum.) retorna a frequência acumulada. Essa frequência acumulada vem em termos de porcentagem, mas segue exatamente a mesma lógica já discutida no vídeo sobre política internacional. França, China, Reino Unido, EUA, Turquia e Itália, têm mais da metade (54,73%) das guerras travadas no último século. Com Rússia, Japão e Alemanha, temos 2/3 das guerras do mundo no período (72,97%).

PASSO 6: Medidas de Dispersão e Tendência Central

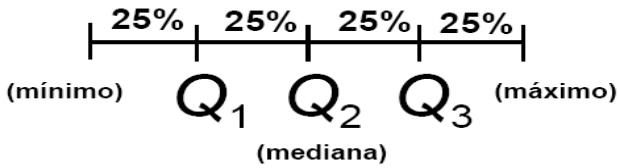
Vamos iniciar esse passo utilizando o pacote "summarytools" recém utilizado. Os comandos que conhecemos de instalação e convocação do pacote (install.packages e library) são descritos novamente. Como utilizamos já esse pacote no passo anterior, vocês podem apenas utilizar o comando library() ou se já tiverem executado esses comandos, ir direto aos novos comandos. Esse pacote possui comandos simples para já retornar todas as principais medidas descritivas em formato de tabela pronto. A função **descr** faz esse trabalho. O comando, na terceira linha abaixo, dentro dos parênteses inclui: o nosso objeto (que é um dataframe) **guerras**, com os dados já explorados anteriormente. Dentro dos colchetes [ ], indicamos ao R que queremos as variáveis da coluna 6 e 7, respectivamente anos e mortes. O comando **style** define o estilo da tabela gerada com as estatísticas descritivas. Escolhemos o "rmarkdown" como estilo. Vocês podem experimentar o estilo grid, por exemplo, substituindo no comando.

```
install.packages("summarytools")
library(summarytools)
descr(guerras[ ,6:7], style='rmarkdown')
```

```
> descr(guerras[ ,6:7], style='rmarkdown')
### Descriptive Statistics
#### guerras
**N:** 148

|            |   anos |     mortes |
|----------------:|-------:|-----------:|
|         **Mean** |   1.25 |  171785.54 |
|      **Std.Dev** |   1.68 |  736191.65 |
|          **Min** |   0.00 |       0.00 |
|           **Q1** |   0.00 |     375.50 |
|       **Median** |   1.00 |    2874.00 |
|           **Q3** |   2.00 |   26000.00 |
|          **Max** |   8.00 | 7500000.00 |
|          **MAD** |   1.48 |    4226.89 |
|          **IQR** |   2.00 |   23624.25 |
|           **CV** |   1.35 |       4.29 |
|     **Skewness** |   1.52 |       7.55 |
|  **SE.Skewness** |   0.20 |       0.20 |
|     **Kurtosis** |   1.78 |      66.68 |
|      **N.Valid** | 148.00 |     148.00 |
|    **Pct.Valid** | 100.00 |     100.00 |
>
```

Temos acima que a média (**Mean**) de duração das guerras é 1,25 anos, assim como temos 171785 mortos em média. As medidas de dispersão descrevem o grau de heterogeneidade dos dados. O primeiro mais evidente são os valores mínimos (**Min**) e máximos (**Min**). A guerra com menos mortos teve 0 mortes assim como a mais rápida menos de um ano (zero). A guerra com mais soldados mortos teve 75000000 baixas. Temos 2 anos de diferença (**IQR**) entre o primeiro quartil (**Q1**) e o terceiro quartil (**Q3**) da distribuição, retornando o intervalo interquartil. A distribuição por quartis acrescenta informação aos valores extremos ao "cortar" a distribuição em 4 partes iguais. A figura abaixo representa graficamente a distribuição em quartis:



No nosso exemplo, o primeiro quartil (**Q1**) é igual a 375 mortes; o segundo quartil ou mediana (**Median**) é 2874 mortos; o terceiro quartil (**Q3**) é 26000.

O desvio padrão (**Std.Dev**) é basicamente a raiz quadrada da variância, medindo o grau de heterogeneidade dos valores analisados, indicando quantas unidades as observações desviam da média:

$$Dp = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n}}$$

No exemplo acima, observamos que o desvio padrão de mortes nas guerras é 736191,65 soldados. O desvio padrão nos retorna o valor na mesma unidade da variável, em mortos. Deste modo, um desvio padrão de 736191,65 significa o quão afastado da média é a distribuição das mortes em guerras. Isso demonstra elevada heterogeneidade no nível de intensidade das guerras medidas por mortos. Algumas guerras concentram muitos mortos, tendo uma escala de conflito bem maior que a média.

PASSO 7: Selecionar apenas o Brasil

Vamos utilizar uma função do R par criar um pedaço (**subset**) da nossa base de dados guerras. Vamos separar apenas as guerras em que o Brasil participou no período dos dados com a função subset:

brazil<- subset(guerras, pais == "Brazil")

O comando acima cria um novo objeto, que chamei de **brazil**. Esse objeto seleciona do banco de dados guerras apenas um país, o Brazil. Assim, temos nesse novo objeto, que também é um dataframe, as três guerras que o Brasil participou no período: a guerra contra a Argentina em 1851 (La Plata no banco de dados), a guerra da tríplice aliança ou guerra do Paraguai em 1864 (chamada Lopez na base) e a Segunda Guerra Mundial (II world War na base).

```
> brazil<- subset(guerras, pais == "Brazil")
> descr(brazil[ ,6:7], style='rmarkdown')
### Descriptive Statistics
#### brazil
**N:** 3

|             |    anos |     mortes |
|-----------------:|--------:|-----------:|
|         **Mean** |    2.67 |   33833.33 |
|      **Std.Dev** |    2.89 |   57302.56 |
|          **Min** |    1.00 |     500.00 |
|           **Q1** |    1.00 |     500.00 |
|       **Median** |    1.00 |    1000.00 |
|           **Q3** |    6.00 |  100000.00 |
|          **Max** |    6.00 |  100000.00 |
|          **MAD** |    0.00 |     741.30 |
|          **IQR** |    2.50 |   49750.00 |
|           **CV** |    1.08 |       1.69 |
|     **Skewness** |    0.38 |       0.38 |
|  **SE.Skewness** |    1.22 |       1.22 |
|     **Kurtosis** |   -2.33 |      -2.33 |
|      **N.Valid** |    3.00 |       3.00 |
|    **Pct.Valid** |  100.00 |     100.00 |
>
```

Temos na figura acima as mesmas estatísticas descritivas, mas agora apenas sobre as guerras do Brasil.