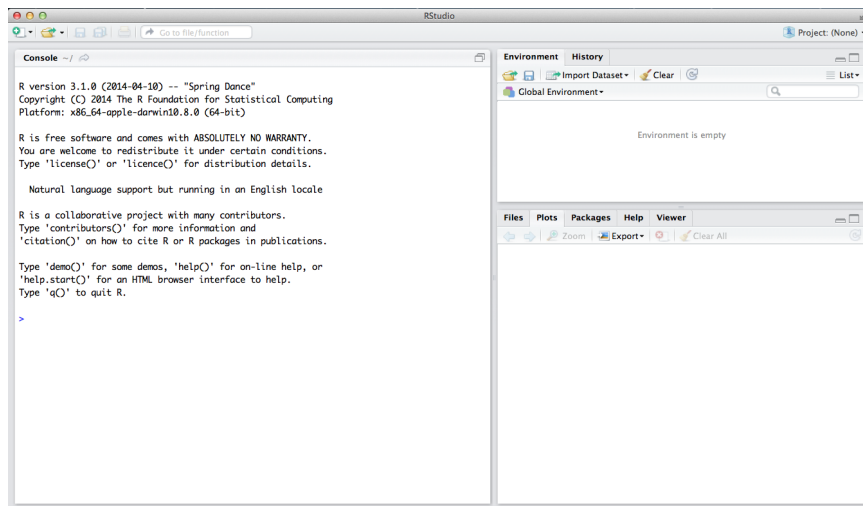


Atividade 1: Primeiros passos no R e RStudio

Introdução

O objetivo deste laboratório é apresentar o R e o RStudio, que são os programas que você usará ao longo do curso para analisar dados reais e chegar a conclusões informadas. Para já distinguir qual é qual: R é o nome da linguagem de programação e RStudio é uma interface gráfica conveniente para utilizar o R.

À medida que as atividades avançarem, você é encorajado a explorar além do que os laboratórios propõem; a vontade de experimentar o fará um programador muito melhor. Antes de chegarmos a este estágio, contudo, você precisa desenvolver alguma fluência básica em R. Hoje nós começaremos com os blocos fundamentais do R e do RStudio: a interface, importação de dados, e comandos básicos.



O painel na parte superior-direita contém seu *espaço de trabalho* e também um histórico dos comandos que você utilizou anteriormente. Quaisquer gráficos que você gerar aparecerão no painel no canto inferior direito.

O painel à esquerda é onde a ação acontece. Ele é chamado de *console*. Toda vez que você iniciar o RStudio, ele terá o mesmo texto no topo do console dizendo qual versão do R você está rodando. Abaixo desta informação está a *linha de comando*. Como o nome sugere, ela interpreta qualquer entrada como um comando a ser executado. Inicialmente, a interação com o R é feita principalmente pela digitação de comandos e a interpretação dos resultados. Esses comandos e sua sintaxe evoluíram ao longo de décadas e agora proporcionam o que muitos usuários acreditam ser um forma bastante natural de acessar dados, organizar, descrever e invocar computações estatísticas.

Para iniciar, inclua o seguinte comando no R (i.e. logo depois do `>` na linha de comando). Você pode digitar o comando manualmente ou copiar deste documento e colar na linha de comando.

```
source("http://www.openintro.org/stat/data/arbuthnot.R")
```

Este comando instrui o R a acessar o website OpenIntro e buscar alguns dados: a contagem de batismos de meninos e meninas coletada por Arbuthnot. Você deve ver que a área do espaço de trabalho no canto superior direito da janela do RStudio agora lista um conjunto de dados chamado `arbuthnot` que tem 82 observações de três variáveis. À medida que você interage com o R, você criará uma série de objetos. Às

Este é um produto da OpenIntro que é distribuído sob uma Licença Creative Commons Atribuição – Compartilhamento pela Mesma Licença 3.0 (<http://creativecommons.org/licenses/by-sa/3.0>). Este laboratório foi adaptado para a OpenIntro por Andrew Bray e Mine Çetinkaya-Rundel de um laboratório escrito por Mark Hansen do departamento de Estatística da UCLA. Tradução para o português por Erikson Kaszubowski. Adaptado para a disciplina ME111 por Samara Kiihl

vezes você os carregará como nós fizemos aqui, e às vezes você os criará por conta própria como o produto de uma computação ou alguma análise que você realizou. Preste atenção que, por você estar acessando os dados a partir da internet, esse comando (e todas as tarefas) funcionará apenas se você estiver conectado à internet.

Os Dados: Registro de Batismos do Dr. Arbuthnot

O conjunto de dados Arbuthnot se refere ao Dr. John Arbuthnot, um médico, escritor e matemático do século 18. Para investigar a razão de meninos e meninas recém-nascidos ele coletou os registros de batismo de crianças nascidas em Londres entre os anos de 1629 e 1710. Nós podemos dar uma olhada nos dados digitando seu nome no comando de linha.

```
arbuthnot
```

Você deve ver quatro colunas de números, com cada linha representando um ano diferente: a primeira entrada em cada linha é simplesmente o número da linha (um índice que podemos usar para acessar os dados de anos individuais, se quisermos), a segunda é o ano, e a terceira e a quarta são os números de meninos e meninas batizados naquele ano, respectivamente. Use a barra de rolagem à direita da janela do console para examinar o conjunto de dados completo.

Preste atenção que os números das linhas na primeira coluna não fazem parte dos dados de Arbuthnot. O R os adiciona como parte das impressões em tela para ajudá-lo a fazer comparações visuais. Pense neles como um índice que costuma ficar ao lado esquerdo de uma planilha. A comparação com uma planilha geralmente será útil, de fato. O R armazenou os dados de Arbuthnot em um tipo de planilha ou tabela chamada de *data frame* ou banco de dados.

Você pode ver as dimensões deste banco de dados digitando:

```
dim(arbuthnot)
```

Este comando deve dar como resposta `[1] 82 3`, indicando que há 82 linhas e 3 colunas (nós já voltaremos ao que o `[1]` quer dizer), da mesma forma como está especificado ao lado do objeto em seu espaço de trabalho. Você pode ver os nomes das colunas (ou variáveis) digitando:

```
names(arbuthnot)
```

Você deve ver que o banco de dados contém as colunas `year` (ano), `boys` (meninos), e `girls` (meninas). A essa altura, você deve ter notado que muitos dos comandos no R se parecem muito com funções matemáticas; ou seja, invocar comandos do R significa passar a uma função um certo número de argumentos. Por exemplo, os comandos `dim` e `names` precisaram de um único argumento cada um: o nome do banco de dados.

Uma vantagem do RStudio é que ele vem com um visualizador de dados embutido. Clique no nome `arbuthnot` no canto superior direito da janela que lista os objetos em seu espaço de trabalho. Isso fará com que uma visualização alternativa das contagens de Arbuthnot apareça na janela superior esquerda. Você pode fechar o visualizador de dados clicando no “x” no canto superior esquerdo.

Explorando os dados

Vamos começar a examinar os dados um pouco mais de perto. Nós podemos acessar separadamente os dados de uma única coluna da base de dados usando um comando como

```
arbuthnot$boys
```

Este comando mostrará somente o número de meninos batizados em cada ano.

Exercício 1 Qual comando você utilizaria para extrair somente a contagem de meninas batizadas? Experimente!

Preste atenção que a maneira como o R imprimiu esses dados é diferente. Quando nós visualizamos o banco de dados completo, vimos 82 linhas, uma em cada linha do console. Esses dados não estão mais estruturados em uma tabela com outras variáveis, então eles são dispostos um ao lado do outro. Objetos que são impressos na tela desta maneira são chamados de *vetores*; eles representam um conjunto de números. O R adicionou números em [colchetes] no lado esquerdo dos resultados para indicar localizações dentro do vetor. Por exemplo, 5218 segue [1], indicando que 5218 é a primeira entrada no vetor. E se [43] inicia uma linha, então isso significa que o primeiro número naquela linha representa a 43ª entrada no vetor.

O R tem algumas funções poderosas para criar gráficos. Podemos criar um gráfico simples do número de meninas batizadas por ano com o comando

```
plot(x = arbuthnot$year, y = arbuthnot$girls)
```

Por padrão, o R cria um gráfico de dispersão com cada par x,y indicado por um círculo aberto. O gráfico deve aparecer sob a aba "Plots" no canto inferior direito do RStudio. Repare que o comando acima também se parece com uma função, desta vez com dois argumentos separados por vírgula. O primeiro argumento na função de gráfico especifica a variável para o eixo x e o segundo para o eixo y. Se nós quiséssemos conectar os pontos dos dados com linhas, nós poderíamos adicionar um terceiro argumento, a letra "l" de linha.

```
plot(x = arbuthnot$year, y = arbuthnot$girls, type = "l")
```

Você pode se perguntar como você poderia saber que era possível adicionar aquele terceiro argumento. Felizmente, o R tem documentações extensivas de todas as suas funções. Para ler o que a função faz e aprender os argumentos disponíveis, basta digitar um ponto de interrogação seguido pelo nome da função na qual vocês está interessado. Tente o seguinte.

```
?plot
```

Veja que o arquivo de ajuda substitui o gráfico no painel no canto inferior direito. Você pode alternar entre gráficos e arquivos de ajuda usando as abas no topo daquele painel.

Exercício 2 Há alguma tendência aparente no número de meninas batizadas ao longo dos anos? Como você a descreveria?

Agora, vamos supôr que queiramos fazer um gráfico com o número total de batismos. Para calcular isso, nós podemos nos aproveitar do fato de que o R é, na verdade, apenas uma grande calculadora. Nós podemos digitar expressões matemáticas como

```
5218 + 4683
```

para ver o número total de batismos em 1629. Nós podemos repetir isso para cada ano, mas há um modo mais rápido. Se adicionarmos o vetor de batismo para meninos e meninas, o R irá computar todas as somas simultaneamente.

```
arbuthnot$boys + arbuthnot$girls
```

O que você verá são 82 números (naquela exibição compacta, porque não estamos analisando um banco de dados), cada um representando a soma que nós queremos. Dê uma olhada em alguns deles e verifique se eles estão corretos. Portanto, nós podemos criar um gráfico com o total de batismos por ano com o comando

```
plot(arbuthnot$year, arbuthnot$boys + arbuthnot$girls, type = "l")
```

Desta vez, veja que nós deixamos de fora os nomes dos dois primeiros argumentos. Nós podemos fazer isso porque o arquivo de ajuda mostra que o padrão para o comando `plot` é ter a variável `x` como primeiro argumento e a variável `y` como segundo argumento. Você pode inverter a ordem e obter o mesmo resultado, desde que defina quem é `x` e quem é `y`:

```
plot(y = arbuthnot$boys + arbuthnot$girls, x = arbuthnot$year, type = "l")
```

De maneira similar como calculamos a proporção de meninos, podemos computar a razão entre o número de meninos e o número de meninas batizadas em 1629 com

```
5218 / 4683
```

ou podemos utilizar os vetores completos com a expressão

```
arbuthnot$boys / arbuthnot$girls
```

A proporção de recém-nascidos que são meninos

```
5218 / (5218 + 4683)
```

ou também pode ser calculado para todos os anos simultaneamente:

```
arbuthnot$boys / (arbuthnot$boys + arbuthnot$girls)
```

Preste atenção que usando o R como sua calculadora, você precisa prestar atenção da ordem das operações. Aqui, nós queremos dividir o número de meninos pelo total de recém-nascidos, portanto precisamos usar parênteses. Sem eles, o R efetuará primeiro a divisão, depois a adição, dando como resultado algo que não é uma proporção.

Exercício 3 Agora, crie um gráfico das proporções dos meninos com relação ao tempo. O que você percebe? Dica: se você usar as teclas de flecha para cima e para baixo, você pode retomar os comandos prévios, chamado de histórico de comandos. Você também pode acessá-lo clicando na aba "history" no painel no canto superior direito. Isto irá lhe economizar várias

digitações no futuro!

Por fim, além de operadores matemáticos simples como subtração e divisão, você pode pedir para o R fazer comparações como maior que, $>$, menor que, $<$, e igualdade, $==$. Por exemplo, podemos perguntar se o número de meninos é maior que de meninas em cada ano com a expressão

```
arbuthnot$boys > arbuthnot$girls
```

Este comando retorna 82 valores ou do tipo **TRUE** (verdadeiro) se aquele ano teve mais meninos batizados do que meninas, ou **FALSE** (falso) se naquele ano foi o contrário (a resposta pode surpreendê-lo). Esse resultado mostra um tipo diferente de variável daquelas que vimos até agora. No banco de dados **arbuthnot** nossos dados são numéricos (o ano, o número de meninos e meninas). Aqui, nós pedimos para o R criar dados *lógicos*, dados cujos valores são **TRUE** (verdadeiro) ou **FALSE** (falso). De modo geral, a análise de dados envolverá vários tipos diferentes de dados, e uma razão para usar o R é que ele consegue representar e realizar computações com vários tipos de dados.

Já é o bastante para seu primeiro laboratório, então vamos parar por aqui. Para sair do RStudio você pode clicar no “x” no canto superior direito da janela do aplicativo. Você será questionado se quer salvar seu espaço de trabalho. Se você clicar em “save” (salvar), o RStudio salvará seu histórico e todos os objetos de seu espaço de trabalho para que na próxima vez que você inicializar o RStudio, você verá o objeto **arbuthnot** e você terá acesso aos comando que você digitou nas suas sessões prévias. Por enquanto, clique em “Don’t save”.

Esta foi uma curta introdução ao R e ao RStudio, mas nós forneceremos mais funções ao longo do curso. Sinta-se livre para procurar na internet pelo R <http://www.r-project.org> e o RStudio <http://rstudio.org>, ou encontre mais laboratórios para praticar em <http://openintro.org>.