Lab 0: Introduction to R Markdown

Kevin Chen

September 2024

Orientation

This is an R Markdown (Rmd) document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. It was designed to simplify creating HTML documents, but the labs for this text are written to generate PDF documents. Before continuing, be sure to install a TeX distribution; one easy option is the TinyTex distribution that can be installed via R (https://yihui.org/tinytex/).[1]

To install the TinyTex distribution, run the following commands. The commands appear in a grey box; click the green play button in the upper right hand corner to run both lines:

```r
install.packages('tinytex')
tinytex::install_tinytex()
```

To install the *oibiostat* data package, run the code in the following chunk. The package only needs to be installed once.

```r
install.packages("devtools")
devtools::install_github("OI-Biostat/oi_biostat_data")
```

The content of an R Markdown document is created in the R Studio script editor. Formatting commands in the text are converted to a PDF document when you click on the *Knit* button, located on the toolbar at the top of the script editor.

In R Studio, the menu item *File > New File > R Markdown…* choice produces a dialog box for output type. To create a new document, select *Document* from the left side of the dialog box, enter a name and title, choose PDF, and select OK. A template will open in the script editor, with file extension .Rmd. Templates will be provided for all the lab exercises, and you will start by simply opening the template files, so there will almost never be a need to start a new file 'from scratch' while working on these labs.

Getting Started

The first six lines of this file are referred to as the **header**. R Markdown is very fussy about the form of the header—the three dashed lines above and below lines 2-5 must appear exactly as in this document, and there must be a blank space between the colon and the descriptive text. Also, the title, author, and date fields must be kept within double quotes. The output line specifies the output format as PDF.

1. First, rename this file to include your first initial and last name – e.g., 01_intro_to_Rmd_j_vu.Rmd. (Use *File > Save As…*)

[1] Alternatively, a full TeX distribution can be downloaded from https://www.latex-project.org/get/.

2. In this document, edit the header to include your name and today's date. Click *Knit PDF*. This should produce a PDF file located in the same folder as the Rmd file, with a name like 01_intro_to_Rmd_j_vu.pdf. Note that the file name for a PDF created from an Rmd document will be the same, except with a different file extension. The file name and title of the document, however, can be different.

Plain text is prepared in paragraphs, as in the first part of this document. *Text* enclosed in asterisks is *italicized* in the PDF output. **Text** enclosed in double asterisks appears in **bold font**. There must be no space between the asterisks and the enclosed text.

3. Write a brief paragraph describing previous coursework in statistics (if any) and share your motivation for learning statistics with R. *Knit* the document. Note that each time you *knit* the document, the output overwrites the previous version.

During my academic journey, I have encountered various courses that required the application of statistical concepts, but none provided a comprehensive understanding of statistics tailored for health informatics. This gap in my knowledge has motivated me to pursue applied statistics, as I recognize its significance in analyzing and interpreting data accurately within the healthcare sector. Learning statistics with R excites me because it offers a powerful, open-source tool capable of managing complex datasets, which is essential for making data-driven decisions in health policy and management. My goal is to strengthen my analytical skills to contribute meaningfully to research and policy formulation.

Bulleted lists are produced using the formatting syntax:

- Item 1
- Item 2
    - Item 2a
    - Item 2b

The list must be preceded by a blank line, and 4 spaces should be used before sub-items.

4. Write a bulleted list giving your year of graduation, your field of study, and the country you are from. Under the entry for your country, prepare subitems with the name of your state and city. *Knit* the document and inspect it to make sure the PDF is correctly produced.

- 2022
- Bachleor of dental surgery
- India
    - State: Uttar Pradesh
    - City: Lucknow

Additional formatting commands will be introduced gradually throughout the rest of the labs.

5. To start a new page in the PDF document, enter the text 'newpage' preceded by a backslash, as in… (new page coming in the PDF!)

Using R with R Markdown

The real power of R Markdown is that it allows for R programs to be included in the Rmd file, with both the program and its output automatically being produced in the PDF document.

R programs in an Rmd file are located in **code chunks**, which appear as grey-shaded blocks as shown below starting on line 60. You can embed an R code chunk by either typing the three backticks (‘) followed by an “r” enclosed in braces, then the additional three backticks to close the chunk, or simply press the *Insert Chunk* button from the *Chunks* menu on the far right of the toolbar on top of the script editor.

6. Datasets for the labs are contained within the *oibiostat* package. To view which datasets are included in the package, go to the “Packages” tab and scroll down to *oibiostat*; clicking the blue hyperlink opens the documentation page that lists the datasets included in the package.

7. The following code chunks use data included in the package. To load the package, use the `library` command. This command only needs to be used once in a document, in the first R chunk that requires data from the package. The `data()` command can be used to load a particular dataset from the package; once the command is run, the name of the dataset will appear in the Environment tab.

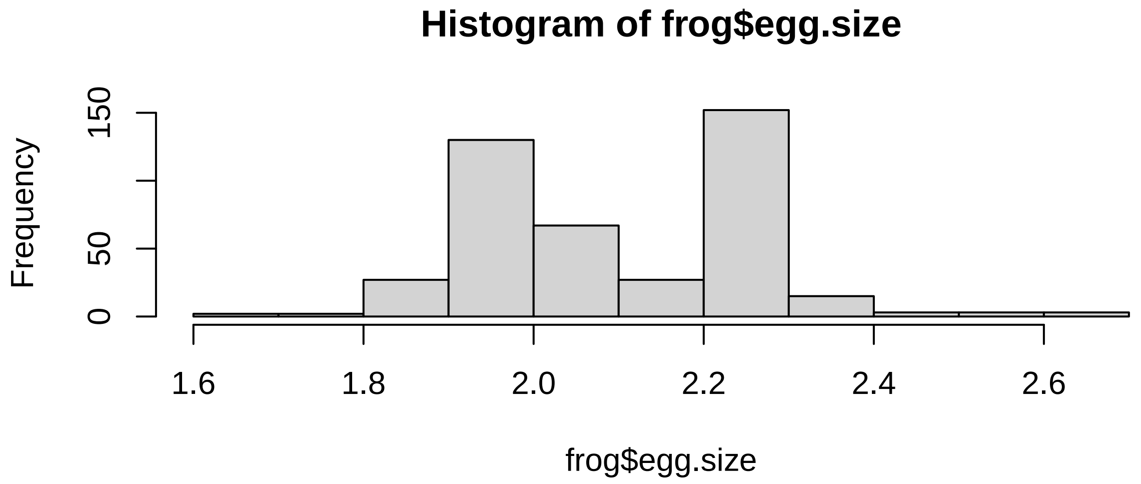
```
library(oibiostat)    #loads the package
data(frog)            #loads the frog dataset
mean(frog$egg.size)
```

```
## [1] 2.114216
```

8. It is also possible to view the output of a code chunk without having to *Knit* the entire document.

   - To run a single line, place your cursor on the line you want to run and press Ctrl/Cmd + Enter. Try this out with the lines in the following chunk. The output appears directly below the code. To clear the output, press the X in the upper right hand corner of the preview.
   - To run an entire chunk, place your cursor within the chunk and press Ctrl/Cmd + Shift + Enter. Try this out in the following chunk. The output is now accessible between the two panes.
   - The Run drop-down menu in the upper right hand corner of the script editor provides other options for running chunks, such as running all chunks above or below a certain point. The gear drop-down menu, next to the *Knit* button, provides options for expanding or collapsing all output in a document.

```r
hist(frog$egg.size)
```



```r
median(frog$egg.size)
```

```
## [1] 2.089296
```

Loading Datasets

9. Datasets are not always in R packages; in most cases, datasets are downloaded to the local computer from alternate sources or generated from R. The following code creates a small dataset called *sample.data* that consists of the numbers 1:9, arranged in a matrix with three rows and three columns and saves it as a file called *sample_data.Rdata*. Run the following code chunk and confirm that the dataset appears in the same folder as where this Rmd document is saved.

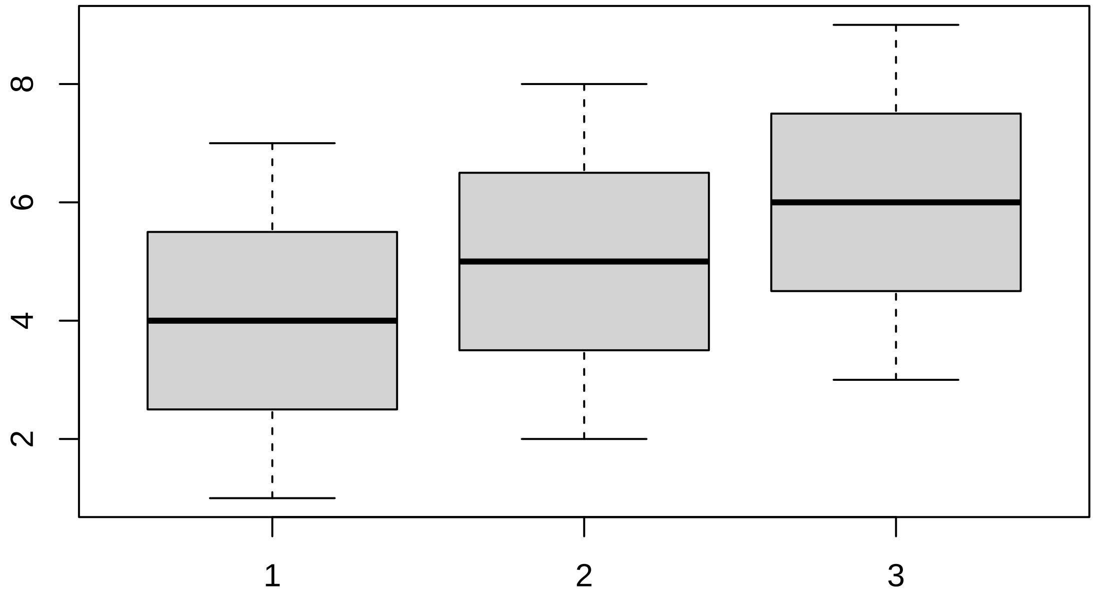
```r
sample.data = matrix(1:9, nrow = 3, byrow = T)  #create sample dataset
save(sample.data, file = "sample_data.Rdata")   #save the file
rm(list = ls()) #clears the environment, equivalent to clicking the broom icon
```

10. To download the dataset, use the `load()` command. Note how the command uses the file name, with the .Rdata extension. In the Environment pane, the dataset appears as *sample.data*, the name that the matrix was given in the previous code chunk.

```r
load("sample_data.Rdata")
```

11. The following code will produce a side-by-side boxplot (one for each column of `sample.data`). To run the code, change eval=FALSE to eval=TRUE. *Knit* the document. The plot should now be visible in the PDF output. Any new R chunks you create default to the eval = TRUE option.

```r
#produce a side-by-side boxplot
boxplot(sample.data)
```



12. When you run a code chunk in an R Markdown file, the commands get sent to the console to be executed. For example, running the command to load `sample_data.Rdata` results in `sample.data` appearing in the Environment pane.

    There is another way to load datasets: clicking the name of an `.Rdata` file in the Files pane also loads data into the global environment.[2] However, the important thing to remember is that while this will allow you to work with the data by running code chunks within an R Markdown file, it does *not* load the data during the knitting process.

13. Let's test that out. On the Environment pane, clikc the broom icon next to the Import Dataset button, then choose Yes from the confirmation pop-up; this clears objects such as variables and datasets from the workspace. Place a # symbol in front of the line in the Step 10 code chunk to turn it into a comment that R will not evaluate.

    Load `sample.data` by clicking the name of the file in the Files pane; confirm that the dataset name appears in the Environment pane and that running the chunk from Step 11 still produces a plot.

    Try to *Knit* the document. This should return an error message reporting that the object 'sample.data' is not found.

14. Remove the # symbol from the Step 10 code chunk. *Knit* the document. During the knitting process, the document is being compiled line-by-line. If there is no command to load a dataset within the R Markdown file, any commands that require the dataset will not be able to run.

    Additionally, the document is compiled in order; i.e., from top to bottom. If, for example, the command to load `sample.data` appeared *after* the histogram command, the same error message would appear. Once the error is detected, the compiler does not "read further" into the document.

[2] A dialog box will appear asking for confirmation; choose Yes to load the dataset into *RStudio*.

> Thus, when working with data files, be sure to check that they are loaded within the R Markdown file itself. There are different functions for loading data depending on whether the data are in `.Rdata` versus `.csv` format, and whether the data are stored in a package like `oibiostat`.

R Script versus R Markdown

An R Script file can be thought of as one large code chunk; R script files are only meant to contain R commands with short plain-text comments. R Markdown files consist of both code, output, and plain text.

R script files are useful when the primary goal is to explore a dataset or test short R programs. R Markdown files are ideal for preparing documents that need to contain both explanatory text and statistical output, such as a lab report.

Congratulations! You have reached the end of Lab 0.