Lecture 2
Introduction to serological data analyses in R

May 21, 2025

Seroanalytics Training
Blantyre, Malawi

Lecture outline

- What is R and RStudio?
- What is a R script and R markdown?
- What is a ‘wide’ versus ‘long’ data frame?
- What are the different types of data frames used in serology?
- Review the *Intro to R.Rmd*

What is R?

- R is a programming language for statistical computing and graphics.
- Open-source and maintained by the R Core Team.
- Used for data analysis, modeling, visualization, and more.
- Runs in a command-line interface or through scripts.

What is RStudio?

- RStudio is an integrated development environment (IDE) for R.
- Provides a user-friendly interface for writing and running R code.
- Includes script editor, console, environment viewer, and plot pane.
- Helps manage workflows and boosts productivity.

Key Differences Between R and RStudio

- R is the **language**; RStudio is an **interface** to use R.
- R can be **used alone**; RStudio **requires R** to run.
- R handles the **computation**; RStudio provides **tools** to interact with R.
- RStudio **enhances usability** with projects, tabs, and debugging tools.

R

R
Studio®

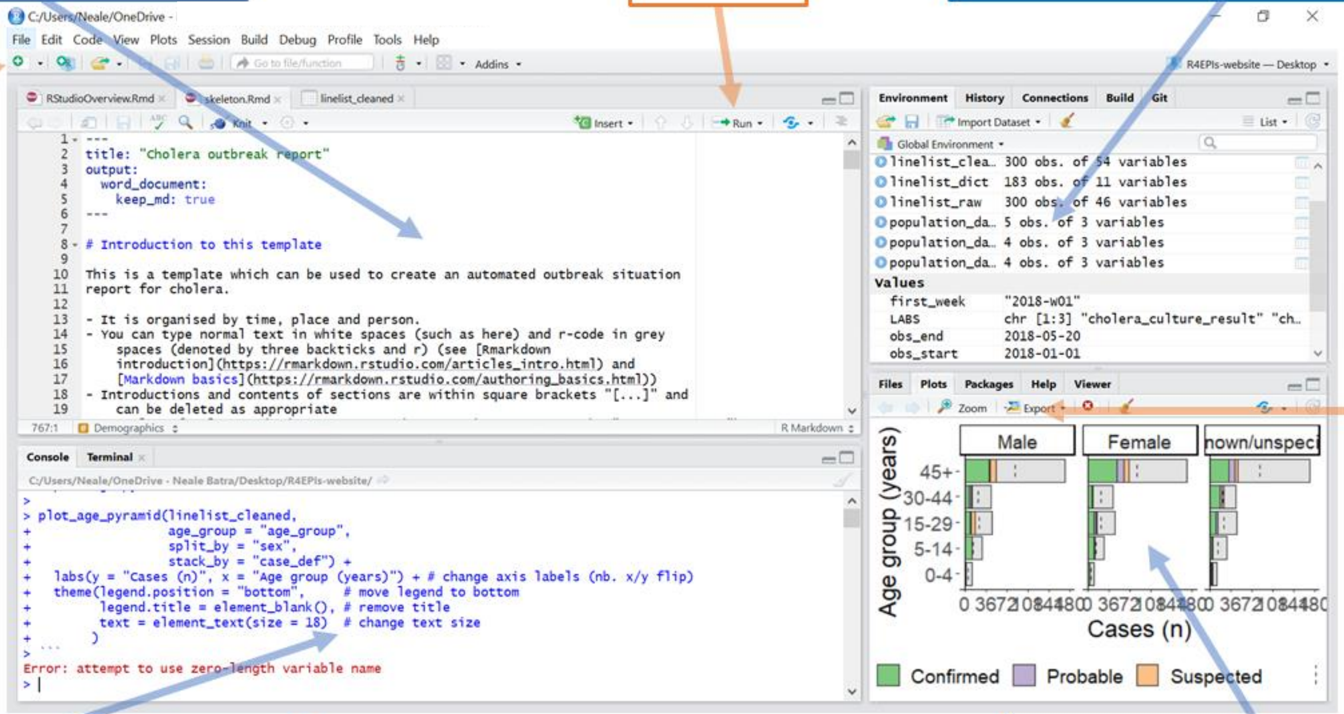
RStudio

Source Pane
Edit and run scripts (e.g. Rmarkdown templates), and view datasets

Tip: Run script

Environment Pane
Overview of objects (datasets, parameters, lists, etc.) you have imported or created.

Tip: Start new script



```
---
title: "Cholera outbreak report"
output:
  word_document:
    keep_md: true
---

# Introduction to this template

This is a template which can be used to create an automated outbreak situation
report for cholera.

- It is organised by time, place and person.
- You can type normal text in white spaces (such as here) and r-code in grey
    spaces (denoted by three backticks and r) (see [Rmarkdown
    introduction](https://rmarkdown.rstudio.com/articles_intro.html) and
    [Markdown basics](https://rmarkdown.rstudio.com/authoring_basics.html))
- Introductions and contents of sections are within square brackets "[...]" and
    can be deleted as appropriate
```

```
linelist_clea… 300 obs. of 54 variables
linelist_dict  183 obs. of 11 variables
linelist_raw   300 obs. of 46 variables
population_da… 5 obs. of 3 variables
population_da… 4 obs. of 3 variables
population_da… 4 obs. of 3 variables
Values
  first_week    "2018-W01"
  LABS          chr [1:3] "cholera_culture_result" "ch…
  obs_end       2018-05-20
  obs_start     2018-01-01
```

```
C:/Users/Neale/OneDrive - Neale Batra/Desktop/R4EPIs-website/
>
> plot_age_pyramid(linelist_cleaned,
+                  age_group = "age_group",
+                  split_by = "sex",
+                  stack_by = "case_def") +
+   labs(y = "Cases (n)", x = "Age group (years)") + # change axis labels (nb. x/y flip)
+   theme(legend.position = "bottom",     # move legend to bottom
+         legend.title = element_blank(), # remove title
+         text = element_text(size = 18)  # change text size
+        )
+ ...
>
Error: attempt to use zero-length variable name
> |
```

Tip: Zoom and export plots

R Console Pane
R commands run are shown here, and non-graphic output and errors are displayed

Plots, Packages, and Help Pane
Commonly used to view graphics, install packages, and view help

What is an R Script?

- A plain text file containing R code (**.R** extension).
- Used for writing, running, and saving R commands.
- Best for quick scripting, data analysis, and model building.
- Cannot produce formatted output documents directly.

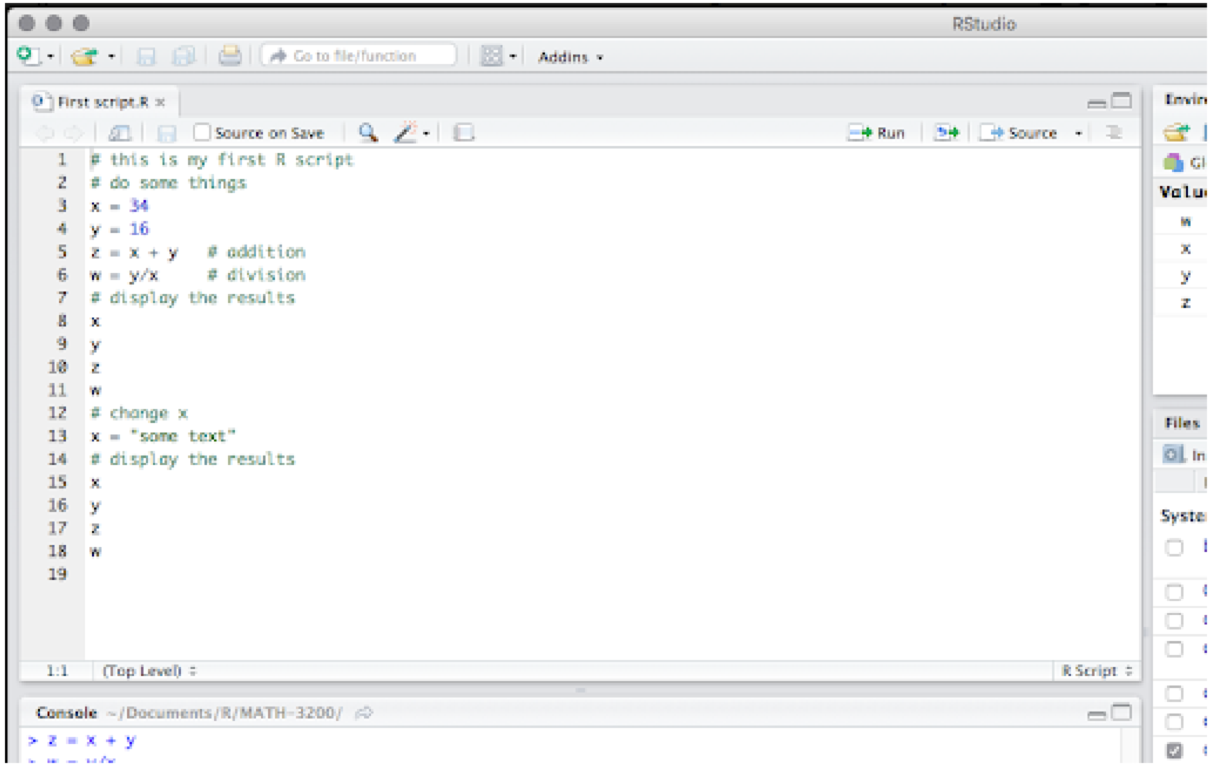
What is an R Markdown File?

- A document combining R code with text and formatting (**.Rmd** extension).
- Supports code chunks, inline results, and narrative text.
- Can generate reports in PDF, Word, or HTML using knitr and pandoc.
- Great for reproducible research and reporting.

Key Differences: R Script vs R Markdown

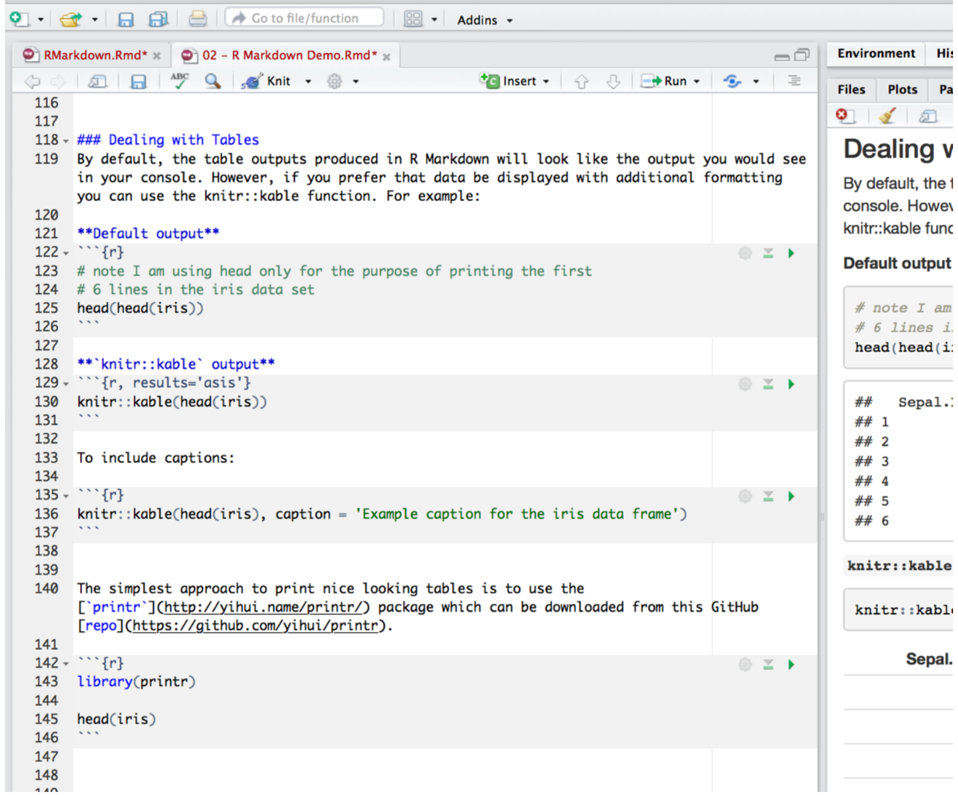
- R Script runs code; **R Markdown runs and documents code**.
- R Script code only; **R Markdown code + formatted text**.
- **R Markdown supports knitting into readable documents** with plots and tables.
- R Scripts are best for development; **R Markdown is best for communication**.

Visual Comparison: R Script vs R Markdown



```r
# this is my first R script
# do some things
x = 34
y = 16
z = x + y   # addition
w = y/x     # division
# display the results
x
y
z
w
# change x
x = "some text"
# display the results
x
y
z
w
```

```
> z = x + y
```



~~~
### Dealing with Tables
By default, the table outputs produced in R Markdown will look like the output you would see
in your console. However, if you prefer that data be displayed with additional formatting
you can use the knitr::kable function. For example:

**Default output**
```{r}
# note I am using head only for the purpose of printing the first
# 6 lines in the iris data set
head(head(iris))
```

**`knitr::kable` output**
```{r, results='asis'}
knitr::kable(head(iris))
```

To include captions:

```{r}
knitr::kable(head(iris), caption = 'Example caption for the iris data frame')
```


The simplest approach to print nice looking tables is to use the
[`printr`](http://yihui.name/printr/) package which can be downloaded from this GitHub
[repo](https://github.com/yihui/printr).

```{r}
library(printr)

head(iris)
```
~~~

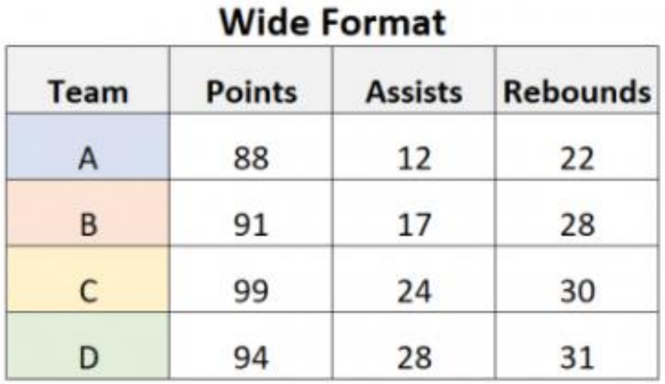
R Script View
(code-only interface)

R Markdown View
(code + formatted output)

Wide vs. Long Data Frames in R

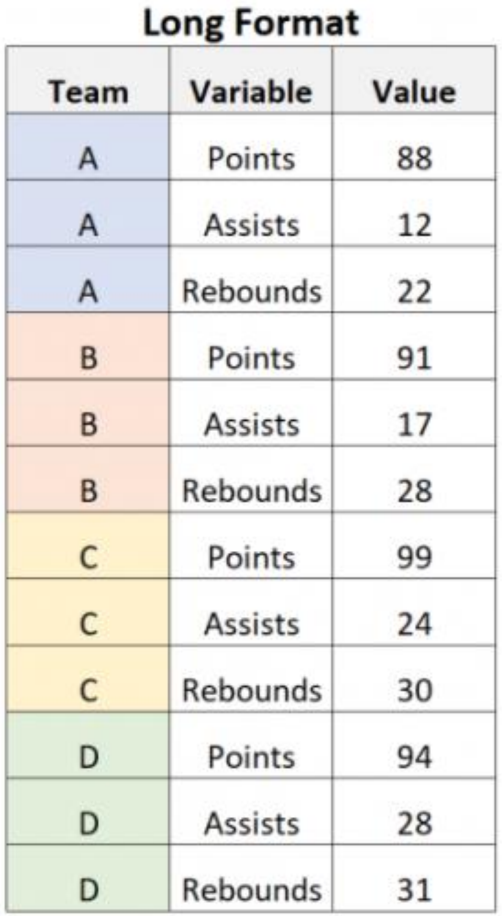
- **Wide** Format
  - Each variable has its own column (e.g., one column per time point)
  - Common in spreadsheet-style data or for visualization (e.g., plots)
  - Easier to read but harder to manipulate in some R functions
- **Long** Format
  - Each observation is a row; variables like 'time' and 'value' are separate columns
  - Preferred for some functions (e.g., ggplot2, dplyr, tidyr)
  - Easier for grouped operations and reshaping

Wide vs. Long Format Data



Wide Format

Team	Points	Assists	Rebounds
A	88	12	22
B	91	17	28
C	99	24	30
D	94	28	31



Long Format

Team	Variable	Value
A	Points	88
A	Assists	12
A	Rebounds	22
B	Points	91
B	Assists	17
B	Rebounds	28
C	Points	99
C	Assists	24
C	Rebounds	30
D	Points	94
D	Assists	28
D	Rebounds	31

Examples of data frames in serology

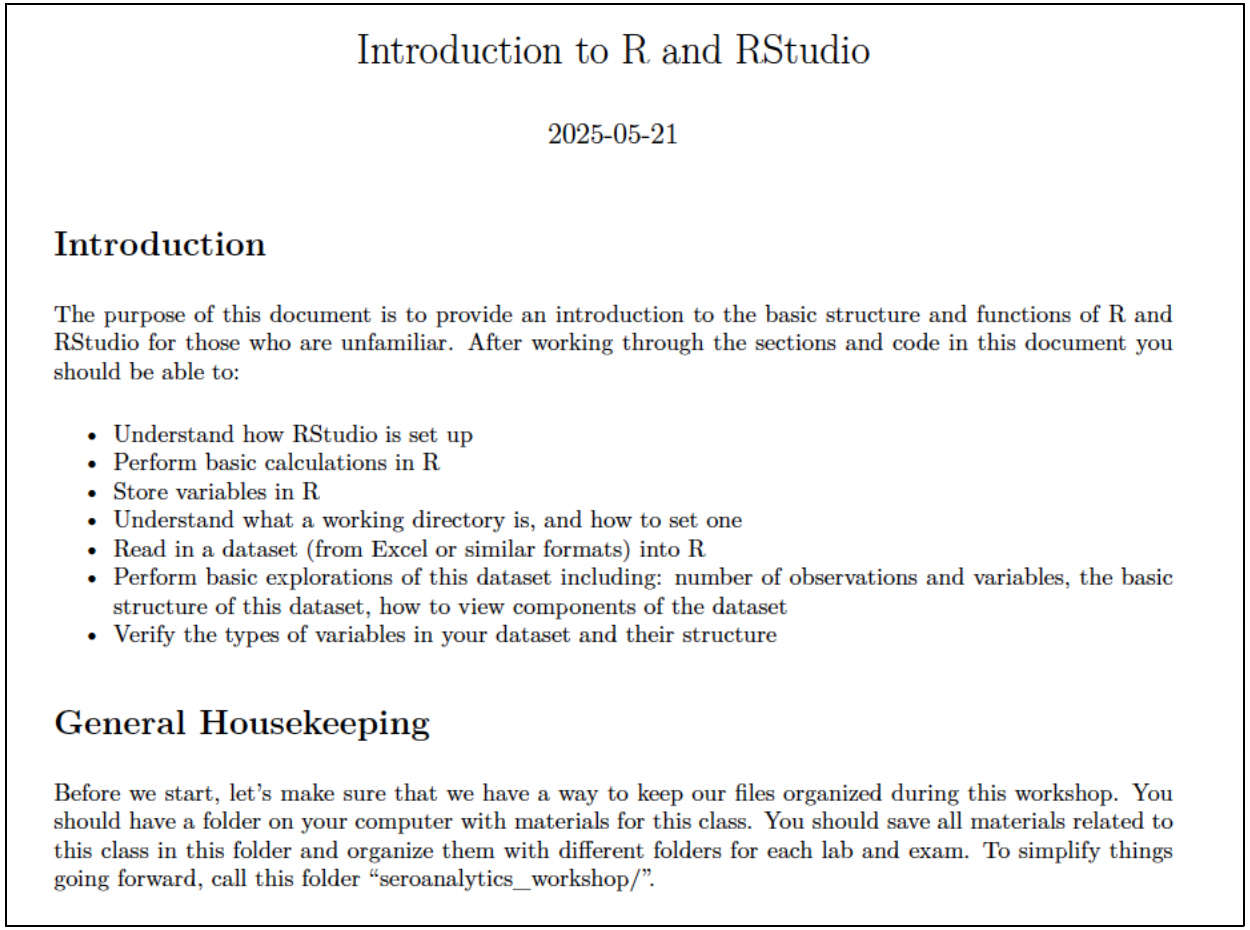
Sample meta-data
Unique sample ID
Sample date*
Specimen information (DBS, plasma)
Manual flag for sample quality
Demographics*: age, sex, geo location, vax status
Survey indicators*: sample weight
(*may store separately)
Manual flag for whether the sample is included in the data analysis

Control meta-data
Unique control ID
Whether it’s a positive vs. negative control, and for which antigen(s) (e.g., “US non-traveler who’s never had Pf malaria”)
Dilution, if part of a standard curve (e.g., “Positive Pool 1:50”)
Short description of what the control is (e.g., “NIBSC XX international standard”)

Experiment meta-data
Plate ID
Date run
Unique sample ID
Well ID
MFI (by antigen)
Net MFI (by antigen)
Bead count (by antigen)

Typically, a wide dataset with each antigen-specific MFI as a column

Follow along in the *Intro to R.Rmd* and *Intro to R.pdf files*



Introduction to R and RStudio

2025-05-21

Introduction

The purpose of this document is to provide an introduction to the basic structure and functions of R and RStudio for those who are unfamiliar. After working through the sections and code in this document you should be able to:

- Understand how RStudio is set up
- Perform basic calculations in R
- Store variables in R
- Understand what a working directory is, and how to set one
- Read in a dataset (from Excel or similar formats) into R
- Perform basic explorations of this dataset including: number of observations and variables, the basic structure of this dataset, how to view components of the dataset
- Verify the types of variables in your dataset and their structure

General Housekeeping

Before we start, let's make sure that we have a way to keep our files organized during this workshop. You should have a folder on your computer with materials for this class. You should save all materials related to this class in this folder and organize them with different folders for each lab and exam. To simplify things going forward, call this folder "seroanalytics_workshop/".