

# STAT115: Introduction to Biostatistics

University of Otago  
Ōtākou Whakaihu Waka

## Lecture 3: Statistical Software

- How do we interact with data?
- In the past: pen and paper
- Today we use computers
- This lecture will introduce software for data and statistics

# Statistical software

- There are many statistical software packages
  - ▶ R
  - ▶ SAS
  - ▶ Stata
  - ▶ SPSS
  - ▶ ...
- Other software packages are also used
  - ▶ Excel
  - ▶ Python
  - ▶ Julia
  - ▶ ...

# R (and Excel)

- We are going to focus on one of these: R
  - ▶ R has a learning curve
    - Provide support in lectures, tutorials and assignments
- We will also see Excel
  - ▶ Excel is used by many researchers to record data
  - ▶ It is also used by many researchers to analyze data
  - ▶ Excel has many weaknesses for data handling and statistics
    - Data handling: easy to (unintentionally) change/corrupt data
    - Statistical modelling: has basic functionality
  - ▶ Learn how to import data into R

## R: NZ on the world stage

- R was developed at the University of Auckland in the early 90s
  - ▶ Ross Ihaka (Ngati Kahungunu, Rangitane)
  - ▶ Robert Gentleman
- It is used around the world
- Advantages:
  - ▶ Freely available
  - ▶ External packages that extend base functionality <sup>a</sup>
    - Contributed by researchers around the world
    - New methodology often readily implemented in R



---

<sup>a</sup>We may see how to install and use packages later

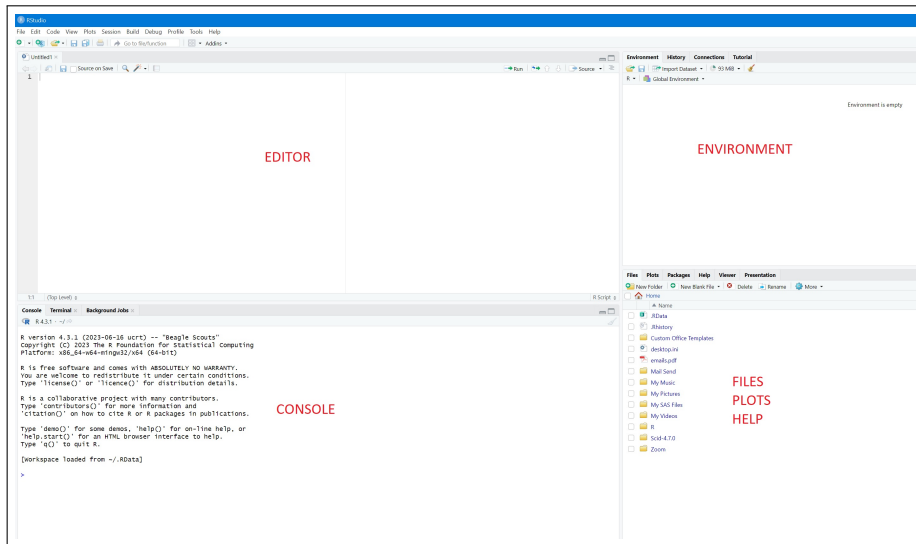
# Getting R and RStudio Up and Running

- We will be using RStudio
  - ▶ R is the language (command line)
  - ▶ Rstudio is an IDE (integrated development environment) for R
    - Provides a more user-friendly experience
- *Option 1 (recommended):* Download and install R and RStudio on own device
  - ▶ See video on Blackboard information for installation instructions
  - ▶ Difficult/impossible on Chromebook or tablet
- *Option 2: Run RStudio using Apps at Otago*
  - ▶ See <https://ask.otago.ac.nz/knowledgebase/article/KA-10005663>
- **Important: get R and RStudio working for you as soon as possible**

# A First Session with R

- Move into Rstudio
- Look at some data
- We will mostly see data in csv (*comma separated values*) files
  - ▶ Comma separated file
  - ▶ Tabular (or rectangular) data
  - ▶ Opened by spreadsheet (like Excel), but is plain text
  - ▶ See video on blackboard for how to obtain a csv from Excel
  - ▶ It is possible to import data directly from Excel
    - It requires installing and loading an additional package
    - Not considered further in this course
  - ▶ Some csv datasets can be imported directly from the URL

# RStudio: Getting Started





# RStudio: Getting Started

## Commentary

- Four panes
  1. LL: Console pane (where R code is run)
    - Start with this today: get things working initially
  2. UL: Editor pane (where we work)
    - Circle back around to how to use editor
    - This is our primary 'work environment'
  3. UR: Environment (etc.) pane (what have we done)
  4. LR: Files (etc.) pane (help, plots, packages)

# RStudio: Importing Data

## General Process

- Download data
  - ▶ By default, this will likely be in the computer's Downloads folder
- Convert to CSV format (if required)
  - ▶ Maybe use Excel to do this
- Import into R studio
  - ▶ Can be done in various ways
  - ▶ `File > Import Dataset > From text (base)`
- View data
  - ▶ Can look in 'Environment' tab (reopen if necessary)

## Example: Pertussis Vaccine Response Data

- Medical researchers constantly searching for better vaccines.
- This example is concerned with vaccine for pertussis (whooping cough)
- Compares antibody response for three different vaccines:
  - ▶ WCV: whole cell pertussis vaccine
  - ▶ APV: pertussis acellular vaccine
  - ▶ DAPV: double dose APV
- $n = 91$  infants age 17–19 months randomly assigned one of the vaccines
- Response, in  $\log(\text{IU/ml})$ , measured one month after immunization

# Example: Pertussis Vaccine Response Data

## Data download

The screenshot shows a Blackboard course page for 'STAT115, S2DN1, 2025 Introduction to Biostatistics'. The 'Datasets' section contains a table with the following data:

Dataset (CSV format)	Description	Reference
<a href="#">10-3-4-shipman-baker-data-x.csv</a>	Shipman deaths	Lecture 1
<a href="#">all_variant_counts.csv</a>	New Zealand COVID genotype data	Lecture 1
<a href="#">pertussis.csv</a>	Pertussis (whooping cough) vaccine data	Lecture 3

A 'Recent download history' pop-up window is open, showing a list of downloaded files:

- [pertussis \(1\).csv](#) (1,135 B - Done)
- [pertussis.csv](#) (1,135 B - Done)
- [10-3-4-shipman-baker-data-x.csv](#) (1 B - 4 minutes ago)
- [all\\_variant\\_counts.csv](#) (1 B - 5 minutes ago)
- [Bordetella\\_pertussis.jpg](#) (858 KB - 2 hours ago)

The pop-up also includes a 'Full download history' link.

# Example: Pertussis Vaccine Response Data

## Data import

The screenshot displays the RStudio interface. On the left, the 'Select File to Export' dialog is open, showing the 'Downloads' folder. Two files, 'pertussis (1).csv' and 'pertussis.csv', are selected. The 'File name' field is set to 'pertussis.csv'. Below the dialog, the R console shows the following text:

```
RStudio 0.3.1 (2023-06-16 ucrt) == "beagle scouts"
right (C) 2023 The R Foundation for Statistical Computing
form: x86_64-v64-rtngs02/x64 (64-bit)

Free software and comes with ABSOLUTELY NO WARRANTY.
are welcome to redistribute it under certain conditions.
"license()" or "licence()" for distribution details.

a collaborative project with many contributors.
"contributors()" for more information and
action() on how to cite R or R packages in publications.

"demo()" for some demos, "help()" for on-line help, or
p.start() for an HTML browser interface to help.
"q()" to quit R.

kspace loaded from ~/Rdata]
```

On the right, the 'Environment' panel shows 'Global Environment' with 'Environment is empty'. The 'Files' panel shows a list of files in the 'Home' directory, including 'RData', 'Rhistory', 'Custom Office Templates', 'desktop.ini', 'emails.pdf', 'Mail Send', 'My Music', 'My Pictures', 'My Recent Places', 'My Videos', 'R', 'Scd-4.7.0', and 'Zoom'.

# Example: Pertussis Vaccine Response Data

## Data view

The screenshot displays the RStudio environment with the 'pertussis' dataset loaded. The console shows the R version (4.3.1) and the 'beagle.scouts' package. The Data viewer shows the 'pertussis' dataset with 91 observations and 2 variables.

**Environment:** pertussis (91 obs. of 2 variables)

**Data:**

antibody	vaccine
3.93	APV
4.06	APV
3.98	APV
5.11	APV
3.27	APV
2.26	APV
3.01	APV
3.78	APV
3.96	APV
4.08	APV
4.74	APV
5.21	APV
5.16	APV
5.17	APV
4.62	APV
4.62	APV

**Console:**

```
R 4.3.1 - ~/ -  
R version 4.3.1 (2023-06-16 ucrt) -- "Beagle Scouts"  
Copyright (C) 2023 The R Foundation for Statistical Computing  
Platform: x86_64-w64-mingw32/x64 (64-bit)  
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.  
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.  
[workspace loaded from ~/RData]  
> pertussis <- read.csv("C:/Users/whazelton/downloads/pertussis.csv")  
> View(pertussis)  
> |
```

## RStudio: Next Steps

- So far we can look at data in a 'spreadsheet'
- To do anything more we have to engage with editor
  - Command line
  - Typing commands to R

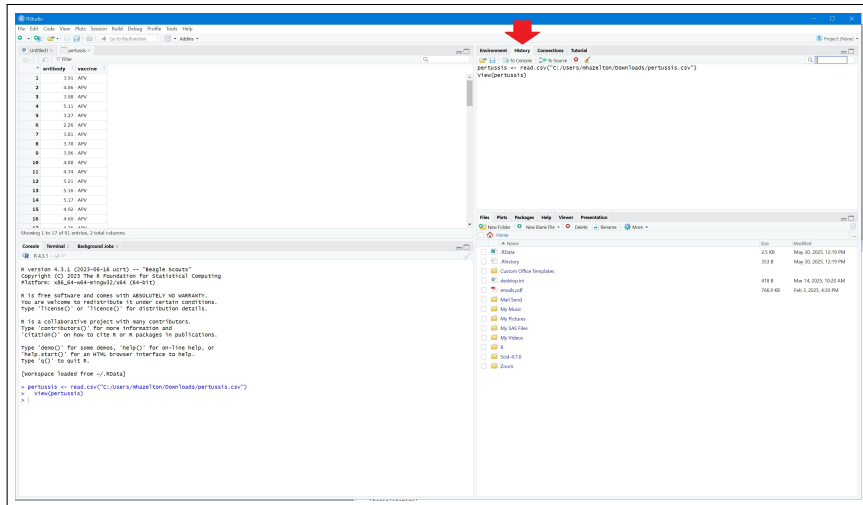
# RStudio: Workflow

- If we exit out of Rstudio
  - ▶ Lose most of what we have done
  - ▶ Start again
  - ▶ Frustrating: assignments and bigger projects
- Solution is to work in the editor
  - ▶ It can be intimidating at first
  - ▶ Rstudio itself helps out
    - 'History'



# RStudio: Workflow

## History



The screenshot displays the RStudio environment. The top-left pane shows a data frame with columns 'antibody' and 'baseline'. The bottom-left pane shows the R console with the following text:

```
R 4.3.1 -- ~~~~~  
R version 4.3.1 (2023-06-16 ucrt) -- "Beeble Scouts"  
Copyright (C) 2023 The R Foundation for Statistical Computing  
Platform: x86_64-w64-mingw32/x64 (64-bit)  
  
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.  
  
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.  
  
[workspace loaded from ~/.RData]  
  
> pertussis <- read.csv("C:/Users/mhazeltan/Downloads/pertussis.csv")  
> view(pertussis)  
>
```

The top-right pane shows the History tab, which contains the command `view(pertussis)`. A red arrow points to this command. The bottom-right pane shows the Files tab, displaying the file explorer.

# Rstudio: Getting Started with Editor

- Instructions for importing data onto editor (alternative method)
  - ▶ 'History' tab shows the R commands for what we have done
    - Put this in the editor window (for when we come back next time)
  - ▶ Care is needed with file structures
    - I suggest creating a STAT115 folder
    - Use this as a 'working directory'

## Rstudio: Getting Started with Editor

- The working directory is the folder (on your computer) that R uses
- Change the working directory:
  - ▶ `Session > Set Working Directory > Choose Directory`
  - ▶ Equivalent command line expression
- Many of the mistakes we see with 100-level students
  - ▶ Asking R to find a file, but you're in the wrong folder
- First ensure in the correct folder
  - ▶ Then import the data

# Example: Pertussis Vaccine Response Data

## Initial exploration

- The data has information about two variables
  - ▶ Antibody response
  - ▶ Vaccine type
- What if we want to select one of these?
  - ▶ Use `$` : allows us to access specific variables by name
  - ▶ Use `[,1]` : allows us to access columns of the data frame by number

```
pertussis$antibody
```

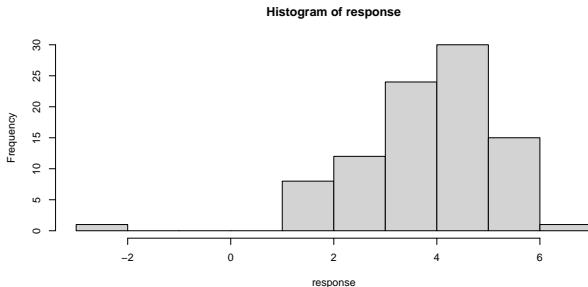
- Assign the new variable to response
  - ▶ Use `=` or `<=`
  - ▶ Use these values later (next slide!)

```
response = pertussis$antibody
```

## Example: Some Summaries of Pertussis Vaccine Data

- We can now look at numeric summaries of antibody response, e.g.
  - ▶ mean: `mean(response)`
  - ▶ median: `median(response)`
  - ▶ standard deviation: `sd(response)`
- We can also look at graphical summaries of antibody response, e.g. histogram

```
hist(response)
```



# RStudio: Help!

- How would we know that in R?
  - ▶ `mean`: calculate the mean
  - ▶ `hist`: plot a histogram
- There is internal help: probably not the first place to look
- For you in STAT115:
  - ▶ Lecture slides
  - ▶ Assignments
  - ▶ Tutorials
  - ▶ Google: e.g. 'Finding an average in R'
  - ▶ AI (e.g. chatgpt)<sup>1</sup>

---

<sup>1</sup>A word of caution: AI tools are excellent for helping you get started with R. AI tools are not a replacement for thinking, but can be helpful tools for learning.

# R Code

- R code will be displayed on lecture slides as follows:

```
mean(response)  
## [1] 3.818
```

- These commands can be copied and pasted
  - ▶ Focus on understanding what the R code is doing
  - ▶ Support for RStudio in tutorials

# Summary

- We will be using R/Rstudio in STAT115
- Free, powerful, and widely used
- We saw how:
  - ▶ Change our working directory
  - ▶ Import data
  - ▶ Subset one variable (antibody response)
  - ▶ Summarize that variable
    - Numerically
    - Graphically