

# PHCM9795 Foundations of Biostatistics

Pilot Notes for R

30 May, 2022

## Contents

<b>Contents</b>	<b>1</b>
<b>Introduction</b>	<b>3</b>
<b>1 Introduction to R and RStudio</b>	<b>5</b>
Learning outcomes . . . . .	5
Part 1: An introduction to R . . . . .	5
1.1 R vs RStudio . . . . .	5
1.2 Installing R and RStudio . . . . .	6
1.3 Recommended setup . . . . .	7
1.4 A simple R analysis . . . . .	11
1.5 The RStudio environment . . . . .	13
1.6 Some R basics . . . . .	14
1.7 What is this thing called the tidyverse? . . . . .	19
Part 2: Obtaining basic descriptive statistics . . . . .	20
1.8 Set up your data . . . . .	21
1.9 Reading a data file . . . . .	22
1.10 Summarising continuous variables . . . . .	23
1.11 Producing a histogram . . . . .	25
1.12 Producing a boxplot . . . . .	27
1.13 Producing a one-way frequency table . . . . .	28
1.14 Producing a two-way frequency table . . . . .	29
1.15 Saving data in R . . . . .	31
1.16 Copying output from R . . . . .	31
Part 3: Creating other types of graphs . . . . .	31
1.17 Bar graphs . . . . .	32
1.18 Creating line graphs . . . . .	35

<b>2</b>	<b>Probability and probability distributions</b>	<b>39</b>
2.1	Importing data into R . . . . .	39
2.2	Checking your data for errors in R . . . . .	40
2.3	Overlaying a Normal curve on a histogram . . . . .	42
2.4	Descriptive statistics for checking normality . . . . .	44
2.5	Importing Excel data into R . . . . .	45
2.6	Generating new variables . . . . .	45
2.7	Summarising data by another variable . . . . .	47
2.8	Plotting data by another variable . . . . .	48
2.9	Recoding data . . . . .	49
2.10	Computing binomial probabilities using R . . . . .	50
2.11	Computing probabilities from a Normal distribution . . . . .	51
<b>3</b>	<b>Precision: R notes</b>	<b>53</b>
3.1	Calculating a 95% confidence interval of a mean . . . . .	53
<b>4</b>	<b>Hypothesis testing</b>	<b>55</b>
4.1	One sample t-test . . . . .	55
	<b>Bibliography</b>	<b>57</b>

# Introduction

These notes provide an introduction to R and instructions on how to conduct the analyses introduced in Foundations of Biostatistics.

These notes are currently under development, with sections being added and revised as the course progresses.

This is the first year that R has been offered as an option. I am keen to receive feedback about the notes and your experience learning R. Please get in touch if anything is unclear, or you have any questions or suggestions.

## Changelog

**2022-05-30**

[Changed]

- Module 1: Typo in R Preferences (Section 1.3.1)

[Added]

- Section 1.12: Instructions to plot a histogram with relative frequencies (i.e. percents) instead of frequencies

**2022-05-27**

[Changed]

- Module 1: Fixed bar-charts that were not plotted correctly

**2022-05-27**

[Added]

- Section 1.2.1: Added a note about using the “patched” version of R 4.2.0 for Windows
- Section 1.14: Instructions for creating two-way tables using the `contTables()` function in the `jmv` package

**2022-05-23**

[Added]

- Section 1.9: Explicit instructions to install `jmv` and `summarytools` when working in Module 1

[Changed]

- Section 1.9: Changed location of `pbcc.dat` from `examples` to `activities` folder for consistency

**2022-05-19**

Initial release

# Module 1

## Introduction to R and RStudio

### Learning outcomes

By the end of this Module, you will be able to:

- understand the difference between R and RStudio
- navigate the RStudio interface
- input and import data into R
- use R to summarise data
- perform basic data transformations
- understand the difference between saving R data and saving R output
- copy R output to a standard word processing package

### Part 1: An introduction to R

“R is a language and environment for statistical computing and graphics.” [Link](#). It is an open-source programming language, used mainly for statistics (including biostatistics) and data science.

The aim of these notes is to introduce the R language within the RStudio environment, and to introduce the commands and procedures that are directly relevant to this course. There is so much more to R than we can cover in these notes. Relevant information will be provided throughout the course, and we will provide further references that you can explore if you are interested.

#### 1.1 R vs RStudio

At its heart, R is a programming language. When you install R on your computer, you are installing the language and its resources, as well as a very basic interface for using R. You can write and run R code using the basic R app, but it's not recommended.

RStudio is an “Integrated Development Environment” that runs R while also providing useful tools to help you write code and analyse data. You can think of R as an engine which does the work, and RStudio as a car that uses the engine, but also provides useful tools like GPS navigation and reversing cameras that help you drive.

Note: even though we recommend that you use RStudio, you still need install R. **RStudio will not run without R installed.**

---

R: Don't run this	RStudio: Run this instead
-------------------	---------------------------

---

## 1.2 Installing R and RStudio

### 1.2.1 To install R on your computer

1. Download the R installer from:

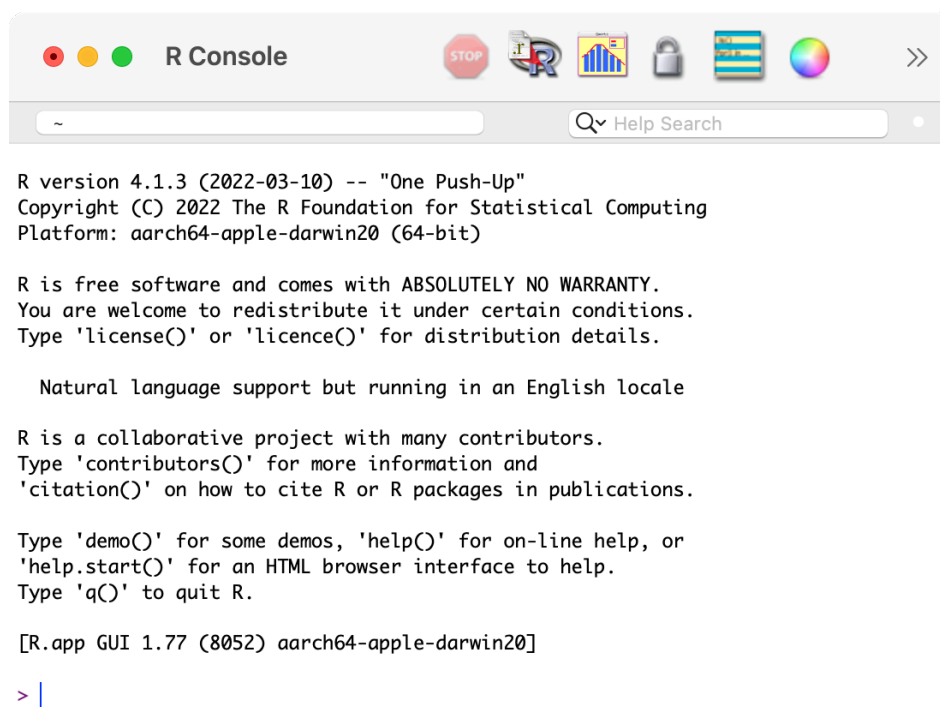
- a. for Windows: <https://cran.r-project.org/bin/windows/base/>
- b. for MacOS: <https://cran.r-project.org/bin/macosx/>

- **Note for Windows users:** as at May 27, 2022, R Version 4.2.0 has compatability issues with RStudio. You should download and install R from <https://cran.r-project.org/bin/windows/base/rpatched.html>

2. Install R by running the installer and following the installation instructions. The default settings are fine.

- **Note for macOS:** if you are running macOS 10.8 or later, you will need to install an additional application called XQuartz, which is available at <https://www.xquartz.org/>. Download the latest installer (XQuartz-2.8.1.dmg as of April 2022), and install it in the usual way.

3. Open the R program. You should see a screen as below:



Near the bottom of the R screen, you will find the “>” symbol which represents the command line. If you type `1 + 2` into the command line and then hit enter you should get:

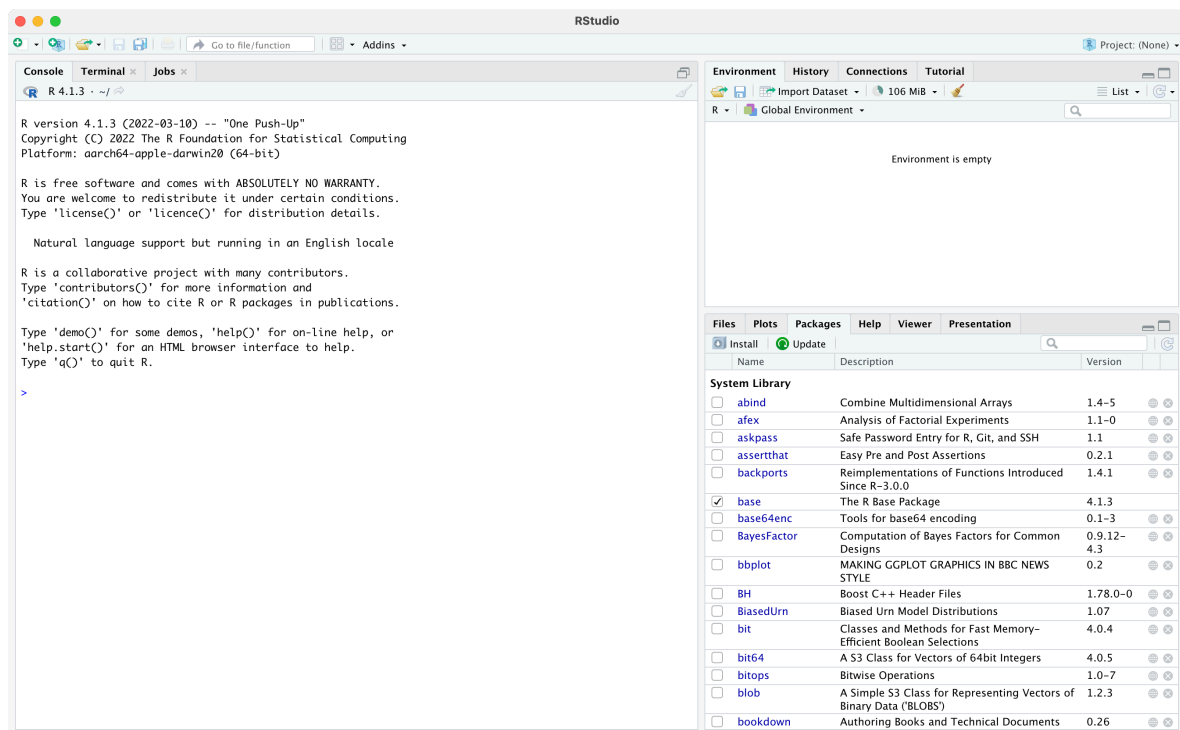
```
[1] 3
```

This is R performing your calculation, with the `[1]` indicating that the solution to  $1 + 2$  is a single number (the number 3).

At this point, close R - we will not interact with R like this in the future. You can close R by typing `quit()` at the command prompt, followed by the return key, or in the usual way of closing an application in your operating system. There is no need to save anything here if prompted.

## 1.2.2 To install RStudio on your computer

1. Make sure you have already installed R, and verified that it is working.
2. Download the RStudio desktop installer at:  
<https://www.rstudio.com/products/rstudio/download>. Ensure that you select the RStudio Desktop (Free) installer in the first column.
3. Install RStudio by running the installer and following the installation instructions. The default settings are fine.
4. Open RStudio, which will appear as below:



Locate the command line symbol `>` at the bottom of the left-hand panel. Type  $1 + 2$  into the command line and hit enter, and you will see:

```
[1] 3
```

This confirms that RStudio is running correctly, and can use the R language to correctly calculate the sum between 1 and 2!

RStudio currently comprises three window panes, and we will discuss these later.

## 1.3 Recommended setup

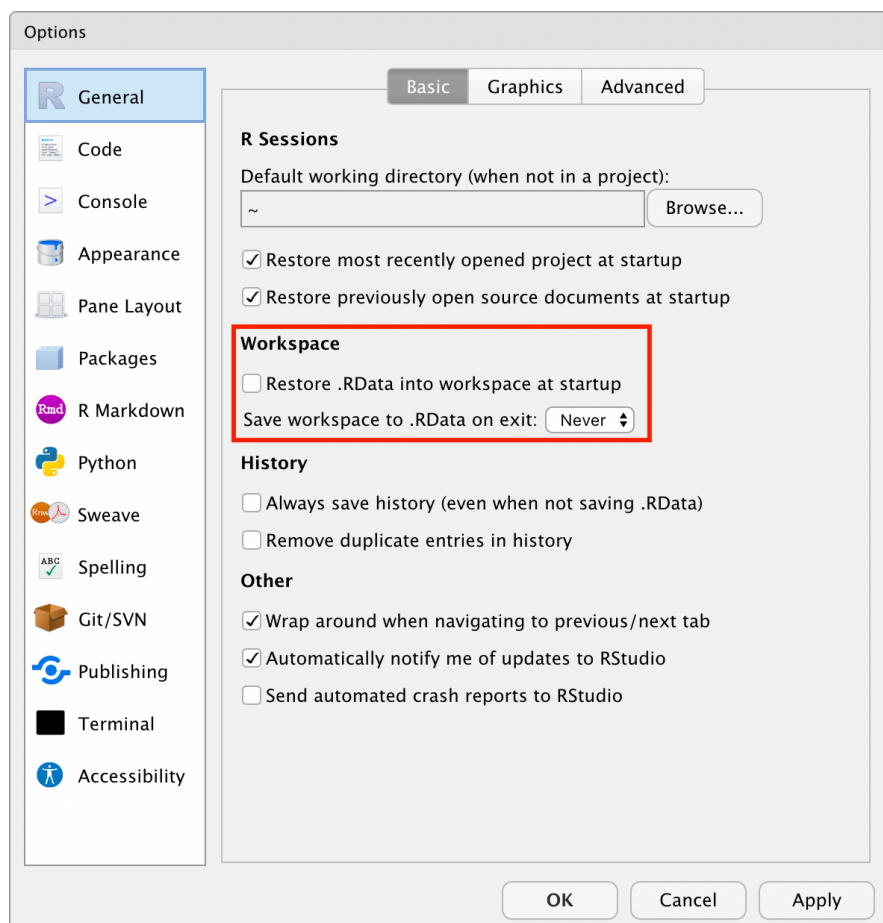
I will provide a recommended setup for R and RStudio in this section. You are free to use alternative workflows and setup, but this setup works well in practice.

### 1.3.1 RStudio preferences

By default, RStudio will retain data, scripts and other objects when you quit your RStudio session. Relying on this can cause headaches, so I recommend that you set up RStudio so that it does not preserve your workspace between sessions. Open the RStudio options:

- Mac: **RStudio > Preferences**
- Windows: **Tools > Options**

and **deselect “Restore .RData into workplace at startup”**, and choose: **“Save workspace to .RData on exit: Never”**.



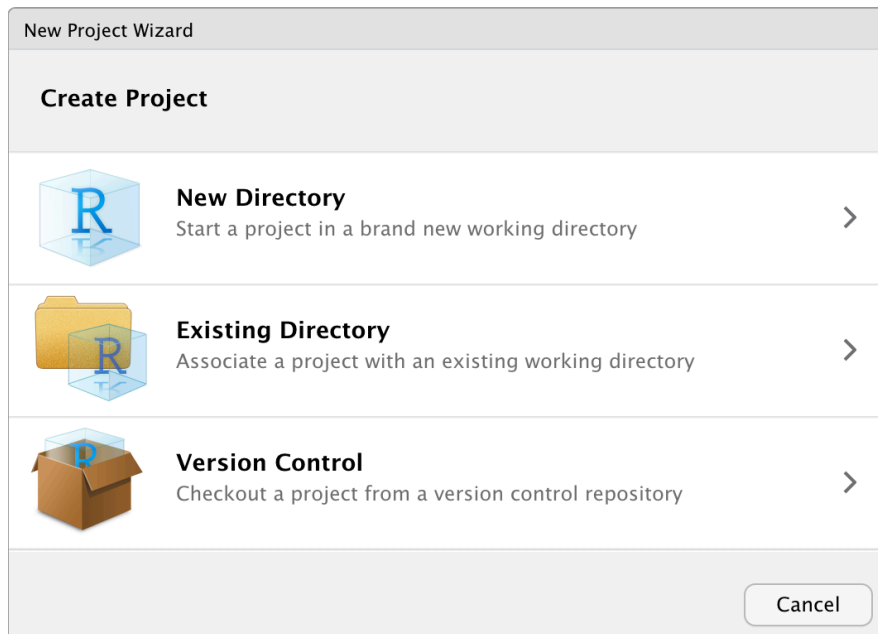
### 1.3.2 Set up a project

A project in RStudio is a folder that RStudio recognises as a place to store R scripts, data files, figures that are common to an analysis project. Setting up a folder allows much more simple navigation and specification of data files and output. More detail can be found in Chapter 8 of the excellent text: *R for Data Science*. Using projects is not necessary, but I recommend working with projects from day one.

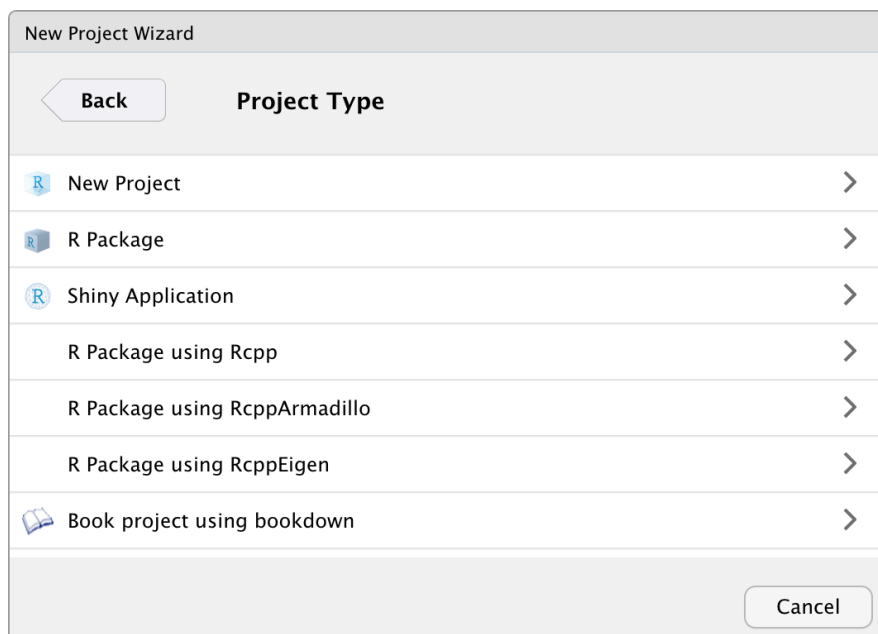
We will create a project called **PHCM9795** to store all the data you will use and scripts that you will write in this course. First, think about where you want to store your project folder: this could be somewhere in your *Documents* folder.

Step 1: Choose **File > New Project...** in RStudio to open the **Create Project** dialog box:

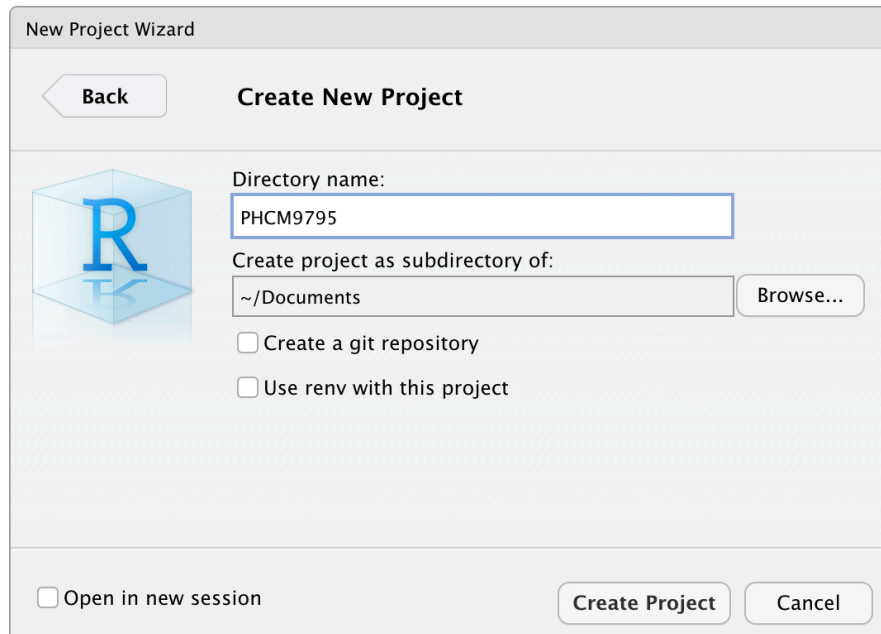




Step 2: Click the first option to create a project in a **New directory**



Step 3: Click the first option: **New Project**. Give the project a name, by typing PHCM9795 in the "Directory name", and choose where you want to store the project by clicking the **Browse** button.



New Project Wizard

Back Create New Project

Directory name:  
PHCM9795

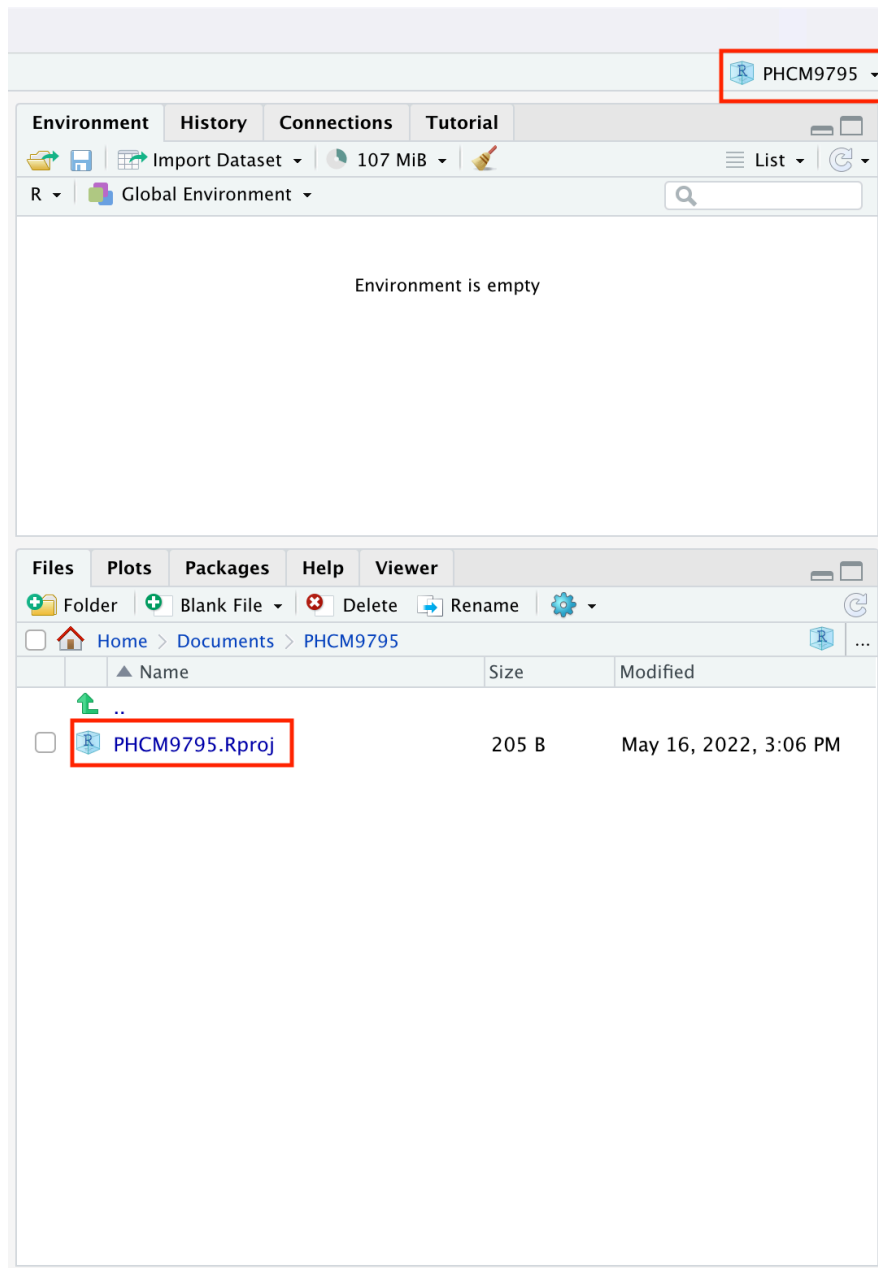
Create project as subdirectory of:  
~/Documents Browse...

☐ Create a git repository  
☐ Use renv with this project

☐ Open in new session Create Project Cancel

Step 4: Click **Create** to create your project.

You will now have a new folder in your directory, which contains only one file: PHCM9795.Rproj, and the two right-hand panes of RStudio will appear as below:



The top-right menu bar is showing that you are working within the PHCM9795 project, and the bottom-right window is showing the contents of that window: the single PHCM9795.Rproj file. We will add some more files to this project later.

## 1.4 A simple R analysis

In this very brief section, we will introduce R by calculating the average of six ages.

To begin, open a new R Script by choosing **File > New file > R Script**. A script (or a program) is a collection of commands that are sequentially processed by R. You can also type Ctrl+Shift+N in Windows, or Command+Shift+N in MacOS to open a new script in RStudio, or click the **New File** button at the top of the RStudio window.

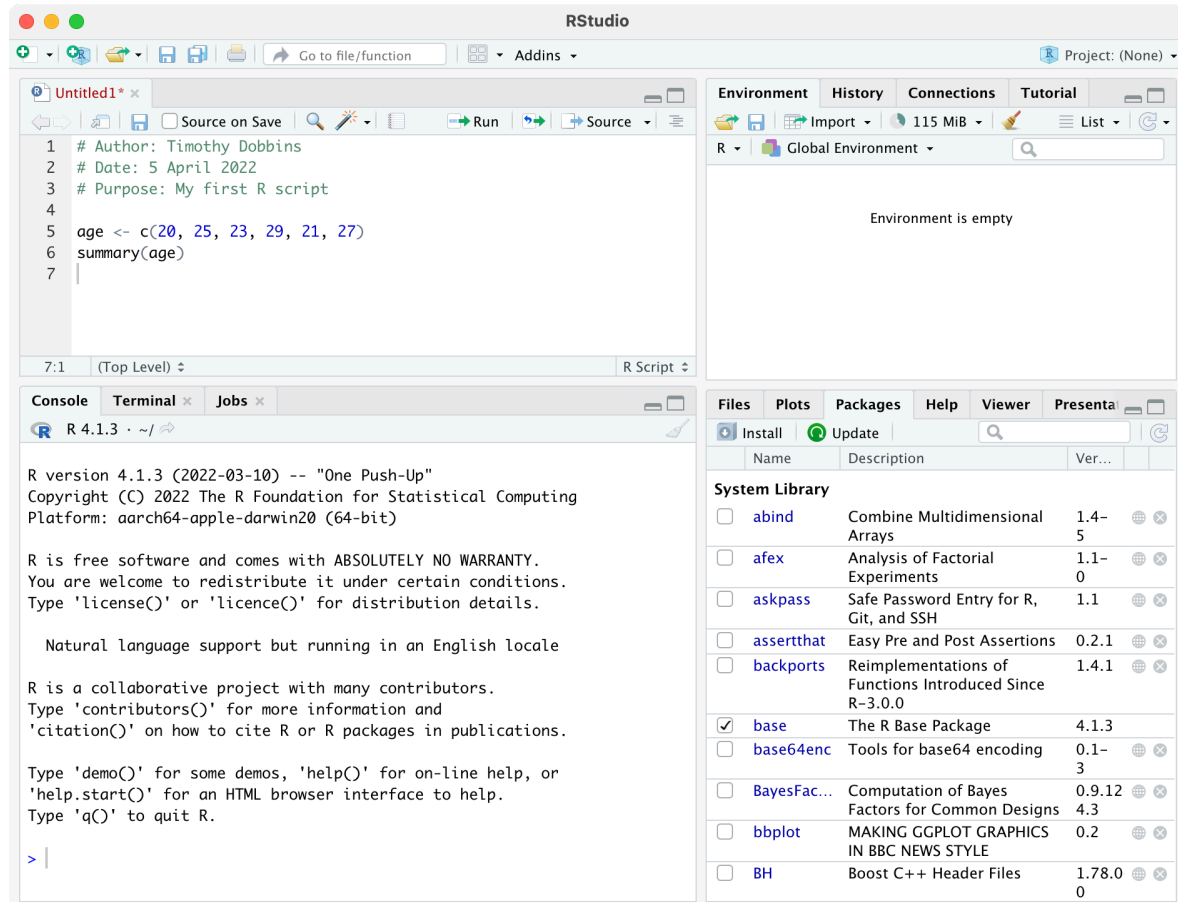
You should now see four window panes, as below. In the top-left window, type the following (replacing my name with yours, and including today's date):

```
# Author: Timothy Dobbins
# Date: 5 April 2022
# Purpose: My first R script

age <- c(20, 25, 23, 29, 21, 27)
summary(age)
```

**Note:** R is case-sensitive, so you should enter the text exactly as written in these notes.

Your screen should look something like:



To run your script, choose **Code > Run Region > Run All**. You will see your code appear in the bottom-left window, with the following output:

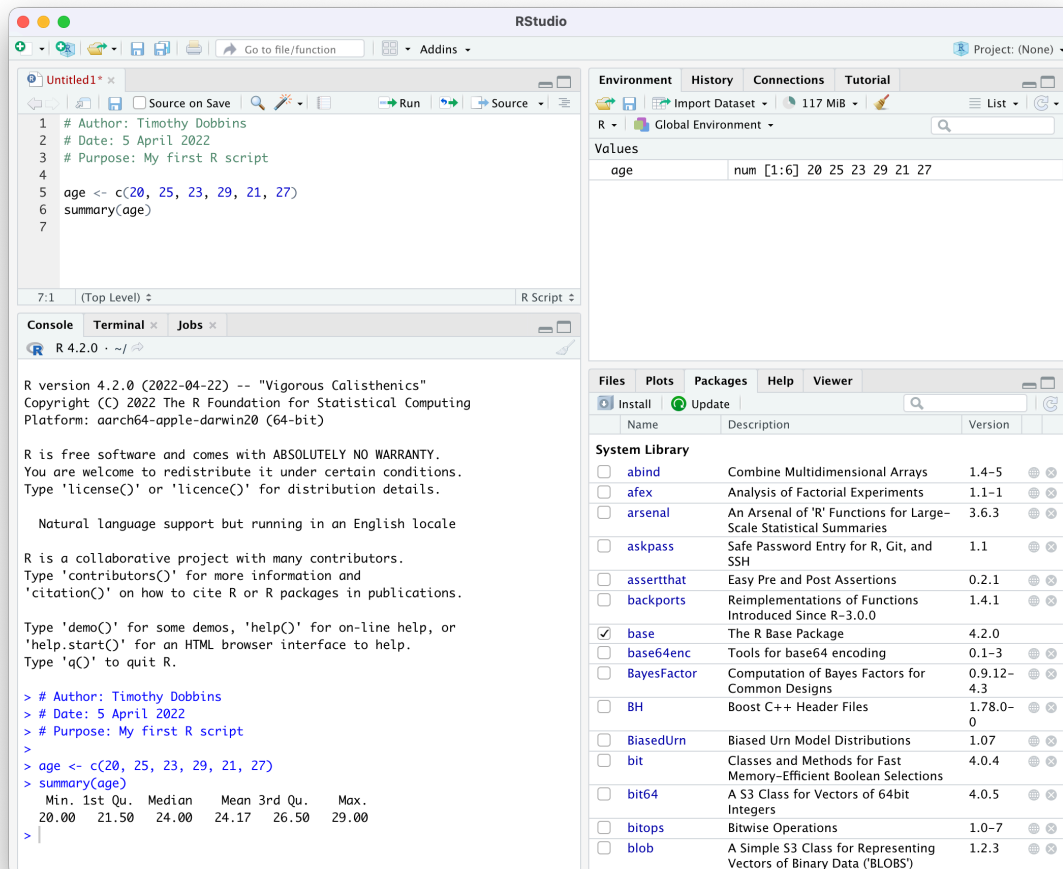
```
> # Author: Timothy Dobbins
> # Date: 5 April 2022
> # Purpose: My first R script
>
> age <- c(20, 25, 23, 29, 21, 27)
>
> summary(age)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 20.00  21.50   24.00   24.17  26.50   29.00
```

We will explain the key parts of this script later, but for now, you have entered six ages and calculated the mean age (along with five other summary statistics).

Save your script within the PHCM9795 project by using **File > Save As**, using the name `my_first_analysis.R`.

## 1.5 The RStudio environment

Now that we have seen a simple example of how to use R within RStudio, let's describe the RStudio environment. Let's assume that you have just run your first R script, and you have four windows as below:



The top-left window is called the **Source** window, and is where you write and edit your R scripts. Scripts can be saved by clicking **File > Save As** or by clicking on the symbol of a floppy disk at the top of the script. The file will have an extension of `.R`, for example `script.R`. Remember to give your script a meaningful title and remember to periodically save as you go.

In RStudio, the name of the script will be black when it has been saved, and will change to red if you have any unsaved changes.

The **Console** window, at the bottom left, contains the command line which is indicated with the symbol `>`. You can type commands here, but anything executed directly from the console is not saved and therefore is lost when the session ends (when you exit RStudio). You should always run your commands from a script file which you can save and use again later. When you run commands from a script, the output and any notes/errors are shown in the console. The Terminal and Jobs tabs will not be used in this course.

The **Environment** window at the top-right shows a list of objects that have been created during your session. When you close your RStudio session these objects will disappear. We will not use the History or Connections tabs in this course.

The bottom right corner contains some useful tabs, in particular the **Help** tab. When you are troubleshooting errors or learning how to use a function, the Help tab should be the first place you visit. Here you can search the help documents for all the packages you have installed. Whenever you create plots in R, these will be shown in the **Plots** tab. The **Packages** tab contains a list of installed packages and indicates which ones are currently in use (we will learn about packages later). Packages which are loaded, i.e. in use, are indicated with a tick. Some packages are in use by default when you begin a new session. You can access information about a package by clicking on its name. The **Files** tab provides a shortcut to access your files. The Viewer tab will not be used in this course.

## 1.6 Some R basics

While we use R as a statistics package, R is a programming language. In order to use R effectively, we need to define some basics.

### 1.6.1 Scripts

While R can be run completely from the command line, issuing commands one-by-one, it is most commonly run using **scripts**. A script is simply a list of commands that are processed in order. The simple analysis we conducted earlier is a very simple script. Some things to know about R scripts:

- anything appearing after a # is a comment, and is ignored by R. The first three lines of our script are there for ourselves (either as writers of code, or readers of code). I include comments at the beginning of each of my scripts to describe:
  - who wrote the script (useful if someone else uses your script and wants to ask questions about it);
  - when the script was written;
  - what the script does. This last point may seem odd, but it's useful to describe what this script does, and why it might differ to other scripts being used in the analysis. This is particularly useful if your scripts become long and complex.
- **R is case-sensitive**. So age, AGE and Age could refer to three separate variables (please don't do this!)
- use blank lines and comments to separate sections of your script

### 1.6.2 Objects

If you do some reading about R, you may learn that R is an “object-oriented programming language”. When we enter or import data into R, we are asking R to create **objects** from our data. These objects can be manipulated and transformed by **functions**, to obtain useful insights from our data.

Objects in R are created using the **assignment operator**. The most common form of the assignment operator looks like an arrow: `<-` and is typed as the `<` and `-` symbols. The simplest way of reading `<-` is as the words “is defined as”. Note that it is possible to use `->` and even `=` as assignment operators, but their use is less frequent.

Let's see an example:

```
x <- 42
```

This command creates a new object called `x`, which is defined as the number 42 (or in words, “`x` is defined as 42”). Running this command gives no output in the console, but the new object appears in the top-right **Environment** panel. We can view the object in the console by typing its name:

```
# Print the object x
x
```

```
## [1] 42
```

Now we see the contents of `x` in the console.

This example is rather trivial, and we rarely assign objects of just one value. In fact, we created an object earlier, called `age`, which comprised six values.

### 1.6.3 Data structures

There are two main structures we will use to work with data in this course: **vectors** and **data frames**. A **vector** is a combination of data values, all of the same type. For example, our six ages that we entered earlier is a vector. You could think of a vector as a column of data (even though R prints vectors as rows!) And technically, even an object with only one value is a vector, a vector of size 1.

The easiest way of creating a vector in R is by using the `c()` function, where `c` stands for ‘combine’. In our previous Simple Analysis in R (Section 1.4), we wrote the command:

```
age <- c(20, 25, 23, 29, 21, 27)
```

This command created a new object called `age`, and *combined* the six values of `age` into one vector.

Just as having a vector of size 1 is unusual, having just one column of data to analyse is also pretty unusual. The other structure we will describe here is a **data frame** which is essentially a collection of vectors, each of the same size. You could think of a data frame as being like a spreadsheet, with columns representing variables, and rows representing observations.

There are other structures in R, such as matrices and lists, which we won’t discuss in this course. And you may come across the term **tibble**, which is a type of data frame.

### 1.6.4 Functions

If objects are the nouns of R, functions are the verbs. Essentially, functions transform objects. Functions can transform your data into summary statistics, graphical summaries or analysis results. For example, we used the `summary()` function to display summary statistics for our six ages.

R functions are specified by their arguments (or inputs). The arguments that can be supplied for each function can be inspected by examining the help notes for that function. To obtain help for a function, we can submit `help(summary)` (or equivalently `?summary`) in the console, or we can use the **Help** tab in the bottom-right window of RStudio. For example, the first part of the help notes for `summary` appear as:

summary {base}

R Documentation

## Object Summaries

### Description

summary is a generic function used to produce result summaries of the results of various model fitting functions. The function invokes particular [methods](#) which depend on the [class](#) of the first argument.

### Usage

```
summary(object, ...)
```

```
## Default S3 method:
summary(object, ..., digits, quantile.type = 7)
## S3 method for class 'data.frame'
summary(object, maxsum = 7,
        digits = max(3, getOption("digits")-3), ...)
```

```
## S3 method for class 'factor'
summary(object, maxsum = 100, ...)
```

The help notes in R can be quite cryptic, but the **Usage** section details what inputs should be specified for the function to run. Here, summary requires an object to be specified. In our case, we specified age, which is our object defined as the vector of six ages.

Most help pages also include some examples of how you might use the function. These can be found at the very bottom of the help page.

### Examples

#### [Run examples](#)

```
summary(attenu, digits = 4) #-> summary.data.frame(...), default precision
summary(attenu $ station, maxsum = 20) #-> summary.factor(...)
```

```
lst <- unclass(attenu$station) > 20 # logical with NAs
## summary.default() for logicals -- different from *.factor:
summary(lst)
summary(as.factor(lst))
```

The summary() function is quite simple, in that it only requires one input, the object to be summarised. More complex functions might require a number of inputs. For example, the help notes for the descriptives() function in the jmv package show a large number of inputs can be specified:



descriptives {jmv}

R Documentation

## Descriptives

### Description

Descriptives are an assortment of summarising statistics, and visualizations which allow exploring the shape and distribution of data. It is good practice to explore your data with descriptives before proceeding to more formal tests.

### Usage

```
descriptives(data, vars, splitBy = NULL, freq = FALSE,
  desc = "columns", hist = FALSE, dens = FALSE, bar = FALSE,
  barCounts = FALSE, box = FALSE, violin = FALSE, dot = FALSE,
  dotType = "jitter", boxMean = FALSE, boxLabelOutliers = TRUE,
  qq = FALSE, n = TRUE, missing = TRUE, mean = TRUE,
  median = TRUE, mode = FALSE, sum = FALSE, sd = TRUE,
  variance = FALSE, range = FALSE, min = TRUE, max = TRUE,
  se = FALSE, ci = FALSE, ciWidth = 95, iqr = FALSE,
  skew = FALSE, kurt = FALSE, sw = FALSE, pcEqGr = FALSE,
  pcNEqGr = 4, pc = FALSE, pcValues = "25,50,75", formula)
```

There are two things to note here. First, notice that the first two inputs are listed with no = symbol, but all other inputs are listed with = symbols (with values provided after the = symbol). This means that everything apart from data and vars have **default** values. We are free to not specify values for these inputs if we are happy with the defaults provided. For example, by default the variance is not calculated (as variance = FALSE). To obtain the variance as well as the standard deviation, we can change this default to variance = TRUE:

```
# Only the standard deviation is provided as the measure of variability
descriptives(data=pbcc, vars=age)

# Additionally request the variance to be calculated
descriptives(data=pbcc, vars=age, variance=TRUE)
```

Second, for functions with multiple inputs, we can specify the input name and its value, or we can ignore the input name and specify just the input values **in the order listed in the Usage section**. So the following are equivalent:

```
# We can specify that the dataset to be summarised is pbcc,
# and the variable to summarise is age:
descriptives(data=pbcc, vars=age)

# We can omit the input name, as long as we keep the inputs in the correct order -
# that is, dataset first, variable second:
descriptives(pbcc, age)

# We can change the order of the inputs, as long as we specify the input name:
descriptives(vars=age, data=pbcc)
```

In this course, we will usually provide all the input names, even when they are not required. As you become more familiar with R, you will start to use the shortcut method.

#### 1.6.4.1 The curse of inconsistency

As R is an open-source project, many people have contributed to its development. This has led to a frustrating part of R: some functions require a single object to be specified, but some require you to

specify a data frame and select variables for analysis. Let's see an example.

The help for `summary()` specifies the usage as: `summary(object, ...)`. This means we need to specify a single object to be summarised. An object could be a single column of data (i.e. a vector), or it could be a data frame. If we have a data frame called `pbic` which contains many variables, the command `summary(pbic)` would summarise every variable in the data frame.

What if we only wanted to summarise the age of the participants in the data frame? To select a single variable from a data frame, we can use the following syntax: `dataframe$variable`. So to summarise just age from this data frame, we would use: `summary(pbic$age)`.

Compare this with the `descriptives()` function in the `jmv` package. We saw earlier that the two required inputs for `descriptives()` are `data` (the data frame to be analysed) and `vars` (the variables to be analysed). So to summarise age from the `pbic` data frame, we would specify `descriptives(data=pbic, vars=age)`.

This inconsistency will seem maddening at first, and will continue to be maddening! Reading the **usage** section of the help pages is a useful way to determine whether you should specify an object (like `pbic$age`) or a data frame and a list of variables.

### 1.6.5 Packages

A **package** is a collection of functions, documentation (and sometimes datasets) that extend the capabilities of R. Packages have been written by R users to be freely distributed and used by others. R packages can be obtained from many sources, but the most common source is CRAN: the Comprehensive R Archive Network.

A useful way of thinking about R is that R is like a smartphone, with packages being like apps which are downloaded from CRAN (similar to an app-store). When you first install R, it comes with a basic set of packages (apps) installed. You can do a lot of things with these basic packages, but sometimes you might want to do things differently, or you may want to perform some analyses that can't be done using the default packages. In these cases, you can install a package.

Like installing an app on a smartphone, you only need to *install* a package once. But each time you want to use the package, you need to *load* the package into R.

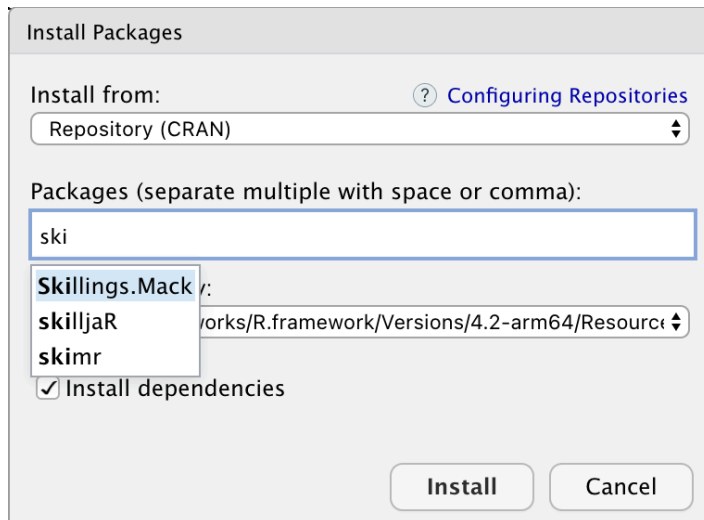
### 1.6.6 How to install a package

There are a couple of ways to install a package. You can use the `install.packages()` function if you know the exact name of the package. Let's use an example of installing the `skimr` package, which gives a very nice, high-level overview of any data frame. We can install `skimr` by typing the following into the console:

```
install.packages("skimr")
```

Note the use of the quotation marks.

Alternatively, RStudio offers a graphical way of installing packages that can be accessed via **Tools > Install Packages**, or via the **Install** button at the top of the **Packages** tab in the bottom-right window. You can begin typing the name of the package in the dialog box that appears, and RStudio will use predictive text to offer possible packages:



While writing code is usually the recommended way to use R, installing packages is an exception. Using **Tools > Install Packages** is perfectly fine, because you only need to install a package once.

### 1.6.7 How to load a package

When you begin a new session in RStudio, i.e. when you open RStudio, only certain core packages are automatically loaded. You can use the `library()` function to load a package that you have previously been installed. For example, now that we have installed `skimr`, we need to load it before we can use it:

```
library(skimr)
```

Note that quotation marks are not required for the `library()` function (although they can be included if you really like quotation marks!).

### Installing vs loading packages

Package installation:

- use the `install.packages()` function (note the 's') or **Tools > Install packages**
- the package name must be surrounded by quotation marks
- only needs to be done once

Package loading

- use the `library()` function
- the package name does not need to be surrounded by quotation marks
- must be done for each R session

## 1.7 What is this thing called the tidyverse?

If you have done much reading about R, you may have come across the tidyverse:

“The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.” <https://www.tidyverse.org/>

Packages in the tidyverse have been designed with a goal to make using R more consistent by defining a “grammar” to manipulate data, examine data and draw conclusions from data. While the tidyverse is a common and powerful set of packages, we will not be teaching the tidyverse in this course for two main reasons:

1. The data we provide have been saved in a relatively tidy way, and do not need much manipulation for analyses to be conducted. The cognitive load in learning the tidyverse in this course is greater than the benefit that could be gained.
2. There are many resources (online, in print etc) that are based on base R, and do not use the tidyverse. It would be difficult to understand these resources if we taught only tidyverse techniques. In particular, the `dataframe$variable` syntax is an important concept that should be understood before moving into the tidyverse.

In saying all of this, I think the tidyverse is an excellent set of packages, which I frequently use. At the completion of this course, you will be well equipped to teach yourself tidyverse using many excellent resources such as: Tidyverse Skills for Data Science and R for Data Science.

## Part 2: Obtaining basic descriptive statistics

In this exercise, we will analyse data to complete a descriptive table from a research study. The data come from a study in primary biliary cirrhosis, a condition of the liver, from Modeling Survival Data: Extending the Cox Model Therneau and Grambsch [2010]. By the end of this exercise, we will have completed the following table.

Table 1.2: Summary of 418 participants from the PBC study (Therneau and Grambsch, 2000)

Characteristic		Summary
Age (years)		Mean (SD) or Median [IQR]
Sex	Male	n (%)
	Female	n (%)
AST* (U/ml)		Mean (SD) or Median [IQR]
Serum bilirubin		Mean (SD) or Median [IQR]
Stage	I	n (%)
	II	n (%)
	III	n (%)
	IIIV	n (%)
Vital status at study end	Alive: no transplant	n (%)
	Alive: transplant	n (%)
	Deceased	n (%)

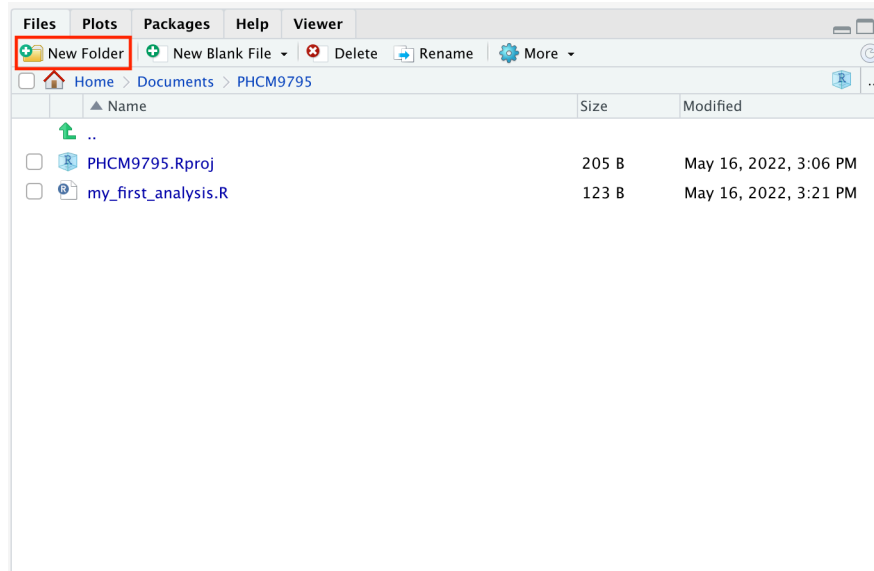
\* asparate aminotransferase

This table is available in Table1.docx, saved on Moodle.

## 1.8 Set up your data

We created a project in the previous step. We will now create a folder to store all the data for this course. Storing the data within the project makes life much easier!

Create a new folder by clicking the **New Folder** icon in the **Files** tab at the bottom-right:

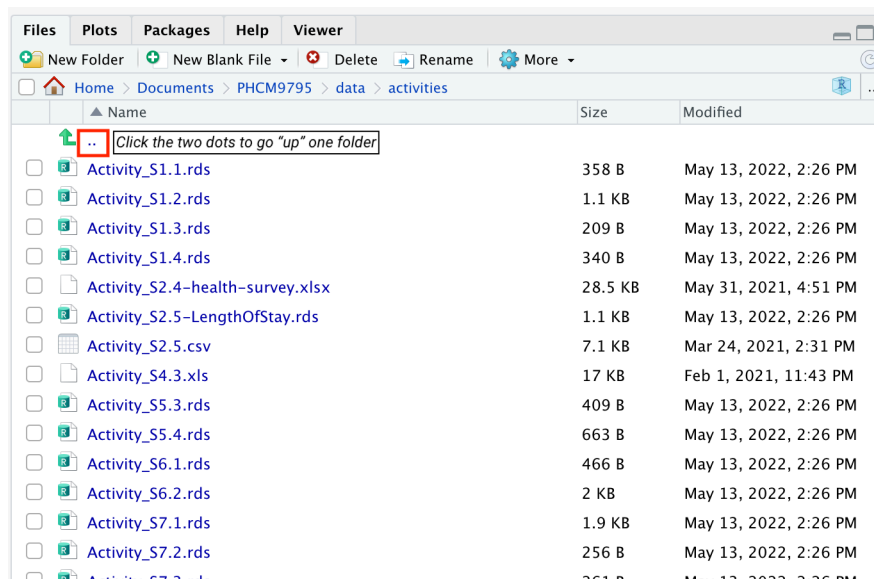


Call the new folder **data**.

Click on this folder to open it, and then create two new folders: **activities** and **examples**.

Download the “Data sets: for learning activities” from Moodle, and use Windows Explorer or MacOS Finder to save these data sets in **activities**. Save the “Data sets: example data from course notes” into the **examples** folder.

Your **activities** folder should look like:



Click the two dots next to the up-arrow at the top of the folder contents to move back up the folder structure. Note that you need to click the dots, and not the up-facing green arrow!

## 1.9 Reading a data file

Typing data directly into R is not common; we usually read data that have been previously saved. In this example, we will read an `.rds` file using the `readRDS()` function, which has only one input: the location of the file.

1 - Confirm that the `pbcrds` file is in the `activities` sub-folder within the `data` folder (as per the previous steps).

2 - Install the packages: `jmv`, `skmr` and `summarytools` using **Tools > Install packages**, or by typing into the console:

```
install.packages("jmv")
install.packages("skmr")
install.packages("summarytools")
```

3 - Load the `skmr` package, and use the `readRDS()` function to read the file into R, assigning it to a data frame called `pbcrds`. Because we set up our project, we can locate our data easily by telling R to use the file: `"data/activities/pbcrds"`, which translates as: the file `pbcrds` which is located in the `activities` sub-folder within the `data` folder.

```
library(skmr)

pbcrds <- readRDS("data/activities/pbcrds")
```

4 - We can now use the `summary()` function to examine the `pbcrds` dataset:

```
summary(pbcrds)
```

##	id	time	status	trt
##	Min. : 1.0	Min. : 41	Min. : 0.0000	Min. : 1.000
##	1st Qu.: 105.2	1st Qu.: 1093	1st Qu.: 0.0000	1st Qu.: 1.000
##	Median : 209.5	Median : 1730	Median : 0.0000	Median : 1.000
##	Mean : 209.5	Mean : 1918	Mean : 0.8301	Mean : 1.494
##	3rd Qu.: 313.8	3rd Qu.: 2614	3rd Qu.: 2.0000	3rd Qu.: 2.000
##	Max. : 418.0	Max. : 4795	Max. : 2.0000	Max. : 2.000
##				NA's : 106
##	age	sex	ascites	hepato
##	Min. : 26.28	Min. : 1.000	Min. : 0.00000	Min. : 0.0000
##	1st Qu.: 42.83	1st Qu.: 2.000	1st Qu.: 0.00000	1st Qu.: 0.0000
##	Median : 51.00	Median : 2.000	Median : 0.00000	Median : 1.0000
##	Mean : 50.74	Mean : 1.895	Mean : 0.07692	Mean : 0.5128
##	3rd Qu.: 58.24	3rd Qu.: 2.000	3rd Qu.: 0.00000	3rd Qu.: 1.0000
##	Max. : 78.44	Max. : 2.000	Max. : 1.00000	Max. : 1.0000
##			NA's : 106	NA's : 106
##	spiders	edema	bili	chol
##	Min. : 0.0000	Min. : 0.0000	Min. : 0.300	Min. : 120.0
##	1st Qu.: 0.0000	1st Qu.: 0.0000	1st Qu.: 0.800	1st Qu.: 249.5
##	Median : 0.0000	Median : 0.0000	Median : 1.400	Median : 309.5
##	Mean : 0.2885	Mean : 0.1005	Mean : 3.221	Mean : 369.5
##	3rd Qu.: 1.0000	3rd Qu.: 0.0000	3rd Qu.: 3.400	3rd Qu.: 400.0
##	Max. : 1.0000	Max. : 1.0000	Max. : 28.000	Max. : 1775.0
##	NA's : 106			NA's : 134
##	albumin	copper	alkphos	ast
##	Min. : 1.960	Min. : 4.00	Min. : 289.0	Min. : 26.35

```
## 1st Qu.:3.243 1st Qu.: 41.25 1st Qu.: 871.5 1st Qu.: 80.60
## Median :3.530 Median : 73.00 Median : 1259.0 Median :114.70
## Mean :3.497 Mean : 97.65 Mean : 1982.7 Mean :122.56
## 3rd Qu.:3.770 3rd Qu.:123.00 3rd Qu.: 1980.0 3rd Qu.:151.90
## Max. :4.640 Max. :588.00 Max. :13862.4 Max. :457.25
## NA's :108 NA's :106 NA's :106
## trig platelet protime stage
## Min. : 33.00 Min. : 62.0 Min. : 9.00 Min. :1.000
## 1st Qu.: 84.25 1st Qu.:188.5 1st Qu.:10.00 1st Qu.:2.000
## Median :108.00 Median :251.0 Median :10.60 Median :3.000
## Mean :124.70 Mean :257.0 Mean :10.73 Mean :3.024
## 3rd Qu.:151.00 3rd Qu.:318.0 3rd Qu.:11.10 3rd Qu.:4.000
## Max. :598.00 Max. :721.0 Max. :18.00 Max. :4.000
## NA's :136 NA's :11 NA's :2 NA's :6
```

An alternative to the `summary()` function is the `skim()` function in the `skimr` package, which produces summary statistics as well as rudimentary histograms:

```
skim(pbc)
```

```
— Data Summary —————
Name                pbc
Number of rows      418
Number of columns    20

Column type frequency:
numeric             20

Group variables      None

— Variable type: numeric —————
skim_variable n_missing complete_rate mean sd p0 p25 p50 p75 p100 hist
1 id           0           1      210. 121. 1 105. 210. 314. 418
2 time         0           1    1918. 1105. 41 1093. 1730 2614. 4795
3 status       0           1      0.830 0.956 0 0 0 2 2
4 trt         106         0.746 1.49 0.501 1 1 1 2 2
5 age          0           1      50.7 10.4 26.3 42.8 51.0 58.2 78.4
6 sex          0           1      1.89 0.307 1 2 2 2 2
7 ascites     106         0.746 0.0769 0.267 0 0 0 0 1
8 hepato      106         0.746 0.513 0.501 0 0 1 1 1
9 spiders     106         0.746 0.288 0.454 0 0 0 1 1
10 edema       0           1      0.100 0.253 0 0 0 0 1
11 bili        0           1      3.22 4.41 0.3 0.8 1.4 3.4 28
12 chol       134         0.679 370. 232. 120 250. 310. 400 1775
13 albumin     0           1      3.50 0.425 1.96 3.24 3.53 3.77 4.64
14 copper      108         0.742 97.6 85.6 4 41.2 73 123 588
15 alkphos    106         0.746 1983. 2140. 289 872. 1259 1980 13862.
16 ast        106         0.746 123. 56.7 26.4 80.6 115. 152. 457.
17 trig       136         0.675 125. 65.1 33 84.2 108 151 598
18 platelet    11         0.974 257. 98.3 62 188. 251 318 721
19 protime     2         0.995 10.7 1.02 9 10 10.6 11.1 18
20 stage       6         0.986 3.02 0.882 1 2 3 4 4
```

The `summary()` and `skim()` functions are useful to give a quick overview of a dataset: how many variables are included, how variables are coded, which variables contain missing data and a crude histogram showing the distribution of numeric variables.

## 1.10 Summarising continuous variables

One of the most flexible functions for summarising continuous variables is the `descriptives()` function from the `jmv` package. The function is specified as `descriptives(data=, vars=)` where:

- data specifies the dataframe to be analysed
- vars specifies the variable(s) of interest, with multiple variables combined using the c() function

We can summarise the three continuous variables in the pbc data: age, AST and serum bilirubin, as shown below.

```
library(jmv)

descriptives(data=pbcc, vars=c(age, ast, bili))
```

```
##
##  DESCRIPTIVES
##
##  Descriptives
##
##           age           ast           bili
##
##  N              418             312             418
##  Missing           0             106             0
##  Mean          50.74155        122.5563         3.220813
##  Median         51.00068        114.7000         1.400000
##  Standard deviation 10.44721        56.69952         4.407506
##  Minimum        26.27789        26.35000         0.3000000
##  Maximum        78.43943        457.2500         28.00000
##
```

By default, the descriptives function presents the mean, median, standard deviation, minimum and maximum. We can request additional statistics, such as the quartiles (which are called the percentiles, or pc, in the descriptives function):

```
descriptives(data=pbcc, vars=c(age, ast, bili), pc=TRUE)
```

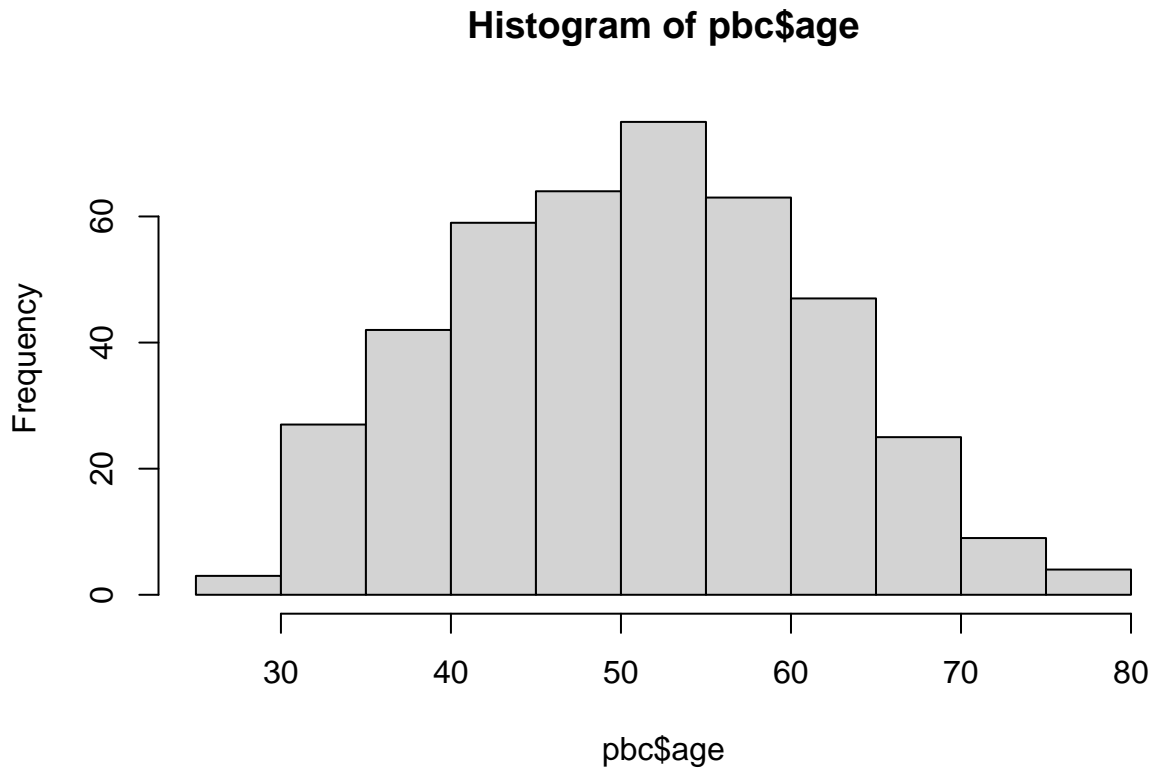
```
##
##  DESCRIPTIVES
##
##  Descriptives
##
##           age           ast           bili
##
##  N              418             312             418
##  Missing           0             106             0
##  Mean          50.74155        122.5563         3.220813
##  Median         51.00068        114.7000         1.400000
##  Standard deviation 10.44721        56.69952         4.407506
##  Minimum        26.27789        26.35000         0.3000000
##  Maximum        78.43943        457.2500         28.00000
##  25th percentile  42.83231         80.60000         0.8000000
##  50th percentile  51.00068        114.7000         1.400000
##  75th percentile  58.24093        151.9000         3.400000
##
```



### 1.11 Producing a histogram

We can use the `hist()` function to produce a histogram, specifying the dataframe to use and the variable to be plotted as `dataframe$variable`:

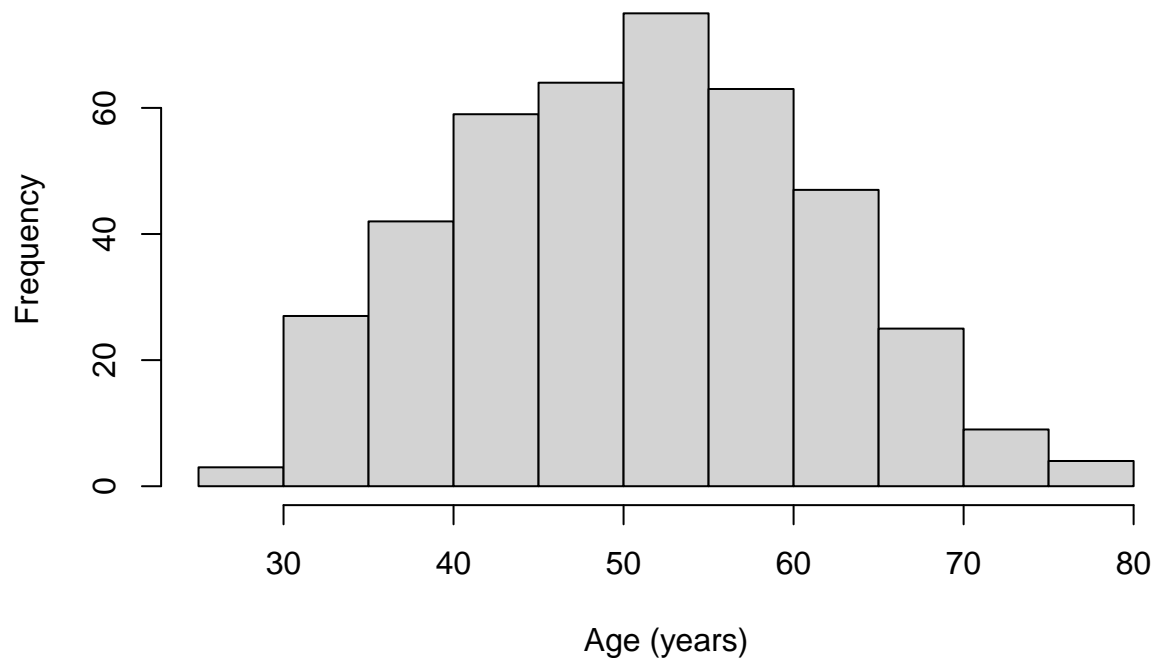
```
hist(pbc$age)
```



The histogram function does a remarkably good job of choosing cutpoints and binwidths, and these rarely need to be changed. However, the labelling of the histogram should be improved by using `xlab=" "` and `main=" "` to assign labels for the x-axis and overall title respectively:

```
hist(pbc$age, xlab="Age (years)",  
     main="Histogram of participant age from pbc study data")
```

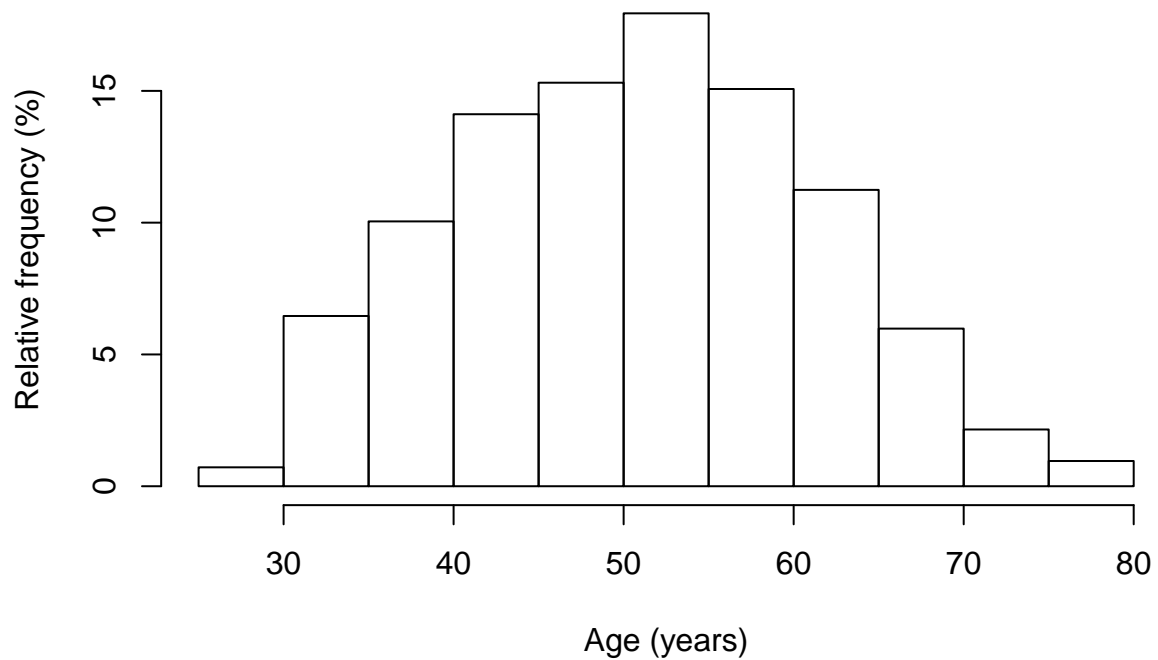
## Histogram of participant age from pbc study data



By default, the `hist()` function plots a **frequency histogram**, with counts on the y-axis. We can tweak the histogram using the following code to plot a histogram of the **relative frequencies**:

```
h <- hist(pbc$age, plot=FALSE)
h$density <- h$counts/sum(h$counts)*100
plot(h, freq=FALSE,
      xlab="Age (years)",
      ylab="Relative frequency (%)",
      main="Histogram of participant age from pbc study data")
```

### Histogram of participant age from pbc study data

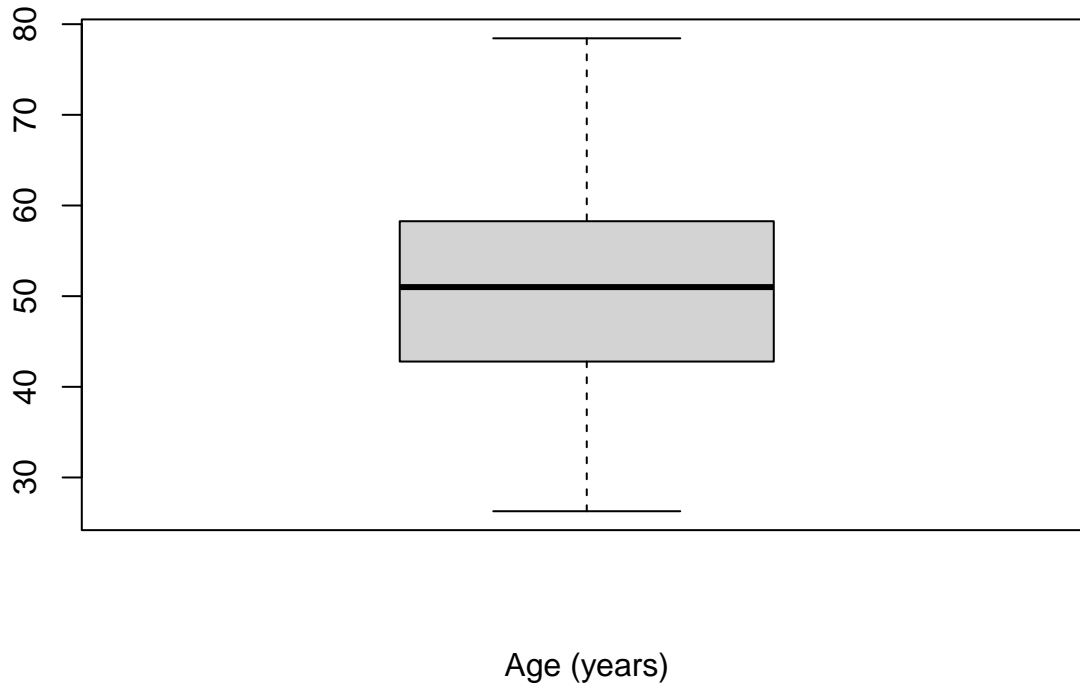


#### 1.12 Producing a boxplot

The `boxplot` function is used to produce boxplots, again specifying the dataframe to use and the variable to be plotted as `dataframe$variable`. Labels can be applied in the same way as the histogram:

```
boxplot(pbc$age, xlab="Age (years)", main="Boxplot of participant age from pbc study data")
```

## Boxplot of participant age from pbc study data



### 1.13 Producing a one-way frequency table

We have three categorical variables to summarise in Table 1: sex, stage and vital status. These variables are best summarised using one-way frequency tables.

```
library(summarytools)

freq(pbc$sex)
```

```
## Frequencies
## pbc$sex
## Type: Numeric
##
##      Freq  % Valid % Valid Cum.  % Total % Total Cum.
## -----
##      1    44   10.53     10.53    10.53    10.53
##      2   374   89.47    100.00    89.47   100.00
##    <NA>     0     0.00     100.00     0.00   100.00
##    Total   418  100.00    100.00   100.00   100.00
```

#### 1.13.1 Defining categorical variables as factors

You will notice that the table above, in its current form, is uninterpretable as the 1 and 2 categories are not labelled. In this course, all variables including categorical variables tend to be numerically coded. To define a categorical variable as such in R, we define it as a **factor** using the factor function:

```
factor(variable=, levels=, labels=)
```

We specify:

- levels: the values the categorical variable can take
- labels: the labels corresponding to each of the levels (entered in the same order as the levels)

To define our variable sex as a factor, we use:

```
pbcs$sex <- factor(pbc$sex, levels=c(1, 2), labels=c("Male", "Female"))
```

We can confirm the coding by re-running a frequency table:

```
freq(pbc$sex)
```

```
## Frequencies
## pbc$sex
## Type: Factor
##
##           Freq  % Valid  % Valid Cum.  % Total  % Total Cum.
## -----
##      Male     44    10.53      10.53    10.53    10.53
##     Female    374    89.47     100.00    89.47    100.00
##      <NA>       0     0.00      0.00     0.00    100.00
##      Total    418   100.00     100.00   100.00    100.00
```

Task: define Stage and Vital Status as factors, and produce one-way frequency tables.  
For example, for Stage:

```
pbcs$stage <- factor(pbc$stage, levels=c(1, 2, 3, 4), labels=c("Stage 1", "Stage 2", "Stage 3", "Stage 4"))
freq(pbc$stage)
```

```
## Frequencies
## pbc$stage
## Type: Factor
##
##           Freq  % Valid  % Valid Cum.  % Total  % Total Cum.
## -----
##    Stage 1     21     5.10      5.10     5.02     5.02
##    Stage 2     92    22.33     27.43    22.01    27.03
##    Stage 3    155    37.62     65.05    37.08    64.11
##    Stage 4    144    34.95    100.00    34.45    98.56
##      <NA>       6     1.44     100.00     1.44   100.00
##      Total    418   100.00     100.00   100.00   100.00
```

## 1.14 Producing a two-way frequency table

To produce tables summarising two categorical variables, we can use the `contTables()` function within the `jmv` package. The minimal inputs to include are `data`: the name of the data frame to be analysed, `rows`: the variable representing the rows of the table, and `cols`: the name of the columns of the table.

For example, to produce a two-way table showing stage of disease by sex using the `pbc` data frame, we use:

```
contTables(data=pbcc, rows=sex, cols=stage)
```

```
##
## CONTINGENCY TABLES
##
## Contingency Tables
##
##      sex      Stage 1      Stage 2      Stage 3      Stage 4      Total
##
##      Male         3         8         16         17         44
##      Female       18        84        139        127        368
##      Total       21        92        155        144        412
##
##
##
## x2 Tests
##
##      Value      df      p
##
##      x2    0.8779873      3    0.8307365
##      N         412
##
```

[The bottom part of the output,  $\chi^2$  Tests, can be ignored for now]

You may notice in the above that the number of observations is now 412. This is because there are missing observations for either sex or stage: which is it, and how would you determine this?

From the cross-tabulation, you can see the individual frequencies of participants in each of the categories in each cell. For example, there are 3 male participants who have Stage 1 disease. You can also read the totals for each row and column. For example, there are 44 males, and 144 participants have Stage 4 disease.

You can also add percentages into your table using `pcCol=TRUE` to include column percents, and `pcRow=TRUE` for row percents. For example, to calculate the relative frequencies (i.e. percentages) of sex within each stage, we would request **column percents** with the option: `pcCol=TRUE`.

```
contTables(data=pbcc, rows=sex, cols=stage, pcCol=TRUE)
```

```
##
## CONTINGENCY TABLES
##
## Contingency Tables
##
##      sex      Stage 1      Stage 2      Stage 3      Stage 4      Total
##
##      Male      Observed         3         8         16         17         44
##      % within column    14.28571    8.69565    10.32258    11.80556    10.67961
##
##      Female      Observed        18        84        139        127        368
##      % within column    85.71429    91.30435    89.67742    88.19444    89.32039
##
##      Total      Observed        21        92        155        144        412
##      % within column   100.00000   100.00000   100.00000   100.00000   100.00000
```

```
##
##
##
##  x2 Tests
##
##           Value      df      p
##
##  x2    0.8779873      3    0.8307365
##  N           412
##
```

We can see that the 3 male participants with Stage 1 disease represent 14% of those with Stage 1 disease.

### 1.15 Saving data in R

There are many ways to save data from R, depending on the type of file you want to save. The recommendation for this course is to save your data using the `.rds` format, using the `saveRDS()` function, which takes two inputs: `saveRDS(object, file)`. Here, `object` is the R object to be saved (usually a data frame), and `file` is the location for the file to be saved (file name and path, including the `.rds` suffix).

It is not necessary to save our PBC data, as we have made only minor changes to the data that can be replicated by rerunning our script. If you had made major changes and wanted to save your data, you could use:

```
saveRDS(pbc, file="pbc_revised.rds")
```

### 1.16 Copying output from R

It is important to note that saving your data or your script in R will not save your output. The easiest way to retain the output of your analyses is to copy the output from the Console into a word processor package (e.g. Microsoft Word) before closing R.

Unfortunately, by default, R is not ideal for creating publication quality tables. There are many packages that will help in this process, such as R Markdown, `bookdown`<sup>1</sup>, `huxtable`, `gt` and `gtsummary`, but their use is beyond the scope of this course. R Markdown for Scientists provides an excellent introduction to R Markdown.

Task: Complete Table 1 using the output generated in this exercise. You should decide on whether to present continuous variables by their means or medians, and present the most appropriate measure of spread. Include footnotes to indicate if any variables contain missing observations.

## Part 3: Creating other types of graphs

The `plot()` function, also known as *base graphics*, is the default method of plotting data in R that can produce publication-quality graphics with minimal coding. There are alternative packages for plotting, with `ggplot2` being one of the most well known. We will present instructions for base graphics in this course, but excellent documentation for `ggplot2` can be found at the `ggplot2`: Elegant Graphics for Data Analysis website, written by the package authors.

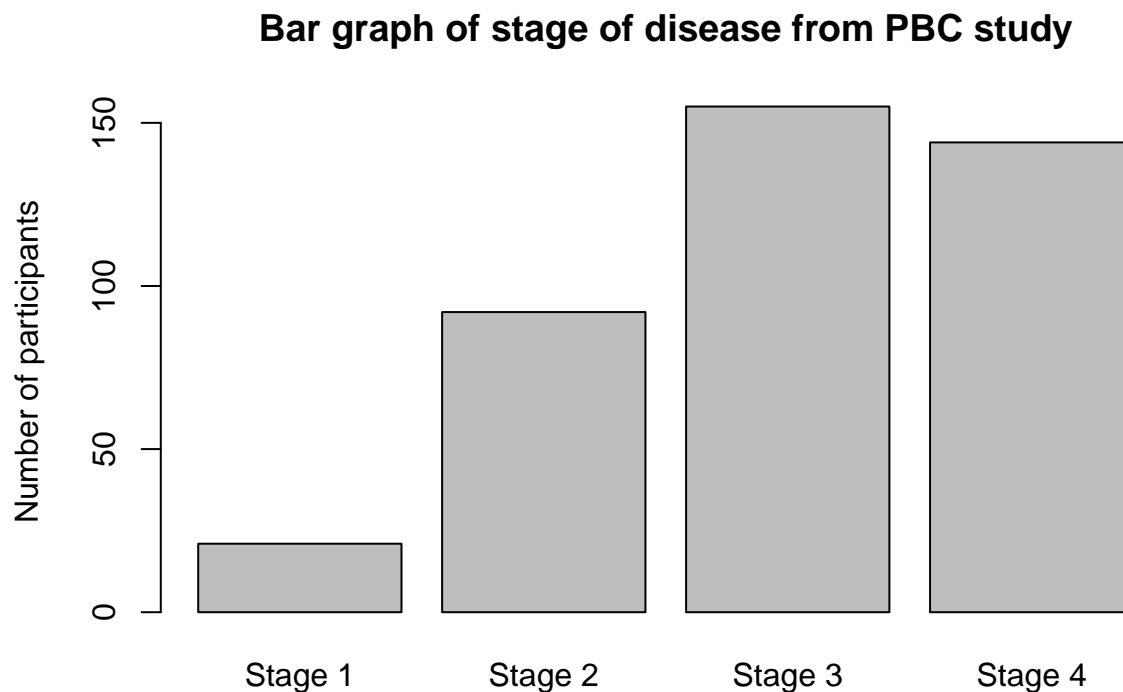
<sup>1</sup>these R notes and the PHCM9795 course notes have been written using `bookdown`

### 1.17 Bar graphs

The simplest way to use the `plot()` function is by specifying an object to be plotted. As with the `hist()` function, to plot a single variable from a data frame, we must define it using: `dataframe$variable`.

Here we will create the bar chart shown in Figure 1.3 of the statistics notes using the `pbcrds` dataset. The x-axis of this graph will be the stage of disease, and the y-axis will show the number of participants in each category.

```
plot(pbc$stage,
     main="Bar graph of stage of disease from PBC study",
     ylab="Number of participants")
```



Note that stage is a categorical variable, that has been defined as a factor (in Section 1.13.1). You **must define categorical data as factors** to plot them in a bar graph.

#### 1.17.1 Clustered bar graph

To create a clustered bar chart as shown in Figure 1.4 of the statistics notes, we need to do a bit of manipulation. We first need to tabulate the data using the `table()` function. We want to plot stage of disease broken down by sex, so we specify sex as the first variable, and stage as the second variable for the `table()` command.

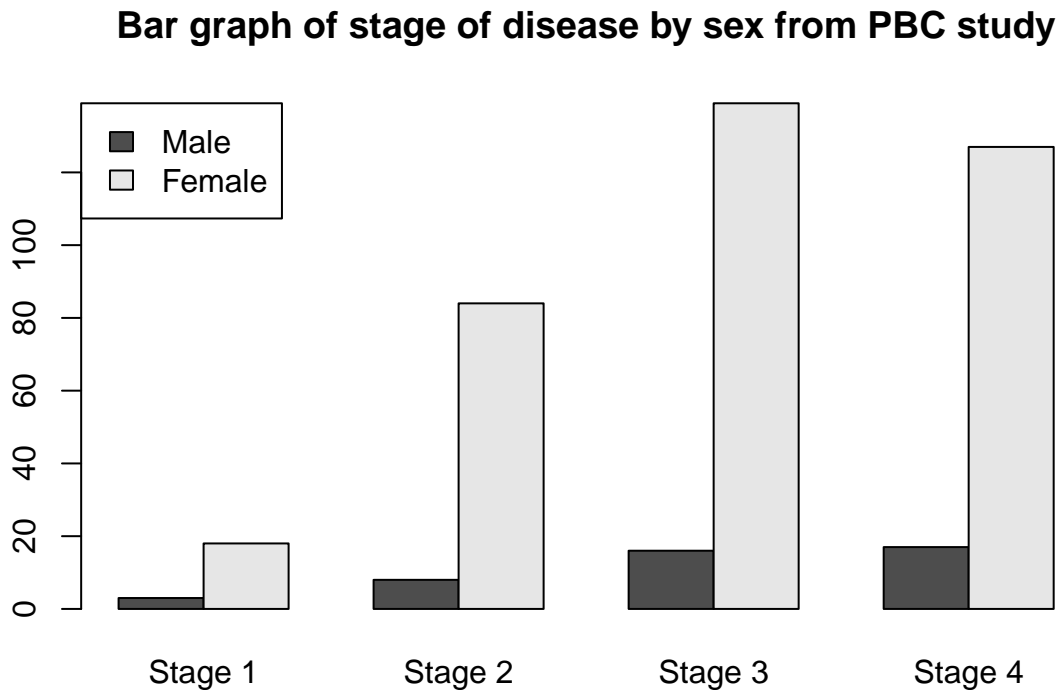
```
counts <- table(pbc$sex, pbc$stage)
counts
```

```
##
##      Stage 1 Stage 2 Stage 3 Stage 4
## Male      3      8      16      17
## Female    18     84     139     127
```



After tabulating the data, we use the `barplot()` function to plot the summarised data. We specify the main title using `main=""`, specify that the stages be plotted separately by sex (`beside=TRUE`), specify the legend be defined by sex, and position the legend in the top-left of the graph:

```
barplot(counts, main="Bar graph of stage of disease by sex from PBC study",  
        beside=TRUE, legend = rownames(counts), args.legend = list(x = "topleft"))
```

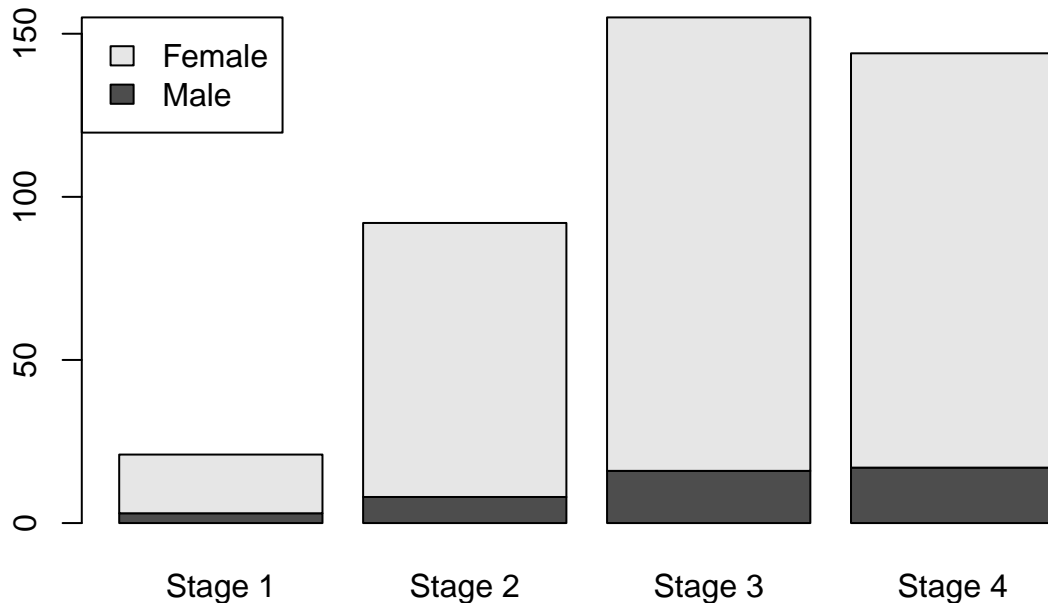


### 1.17.2 Stacked bar graph

A stacked bar graph can be constructed as for the clustered bar graph, but we specify `beside=FALSE`:

```
barplot(counts, main="Bar graph of stage of disease by sex from PBC study",  
        beside=FALSE, legend = rownames(counts), args.legend = list(x = "topleft"))
```

### Bar graph of stage of disease by sex from PBC study



#### 1.17.3 Stacked bar graph of relative frequencies

To plot relative frequencies, we need to transform our table of frequencies (counts) into proportions, by using the `prop.table()` function. The `prop.table()` function takes two arguments: a table of counts, and `margin`, which defines whether we want proportions calculated by row (`margin=1`) or column (`margin=2`).

We want to calculate the relative frequency of sex within each stage category. From our counts table above, this equates to calculating *column* proportions, so we specify `margin=2`. We also multiply the resulting table by 100 to obtain percentages (rather than proportions):

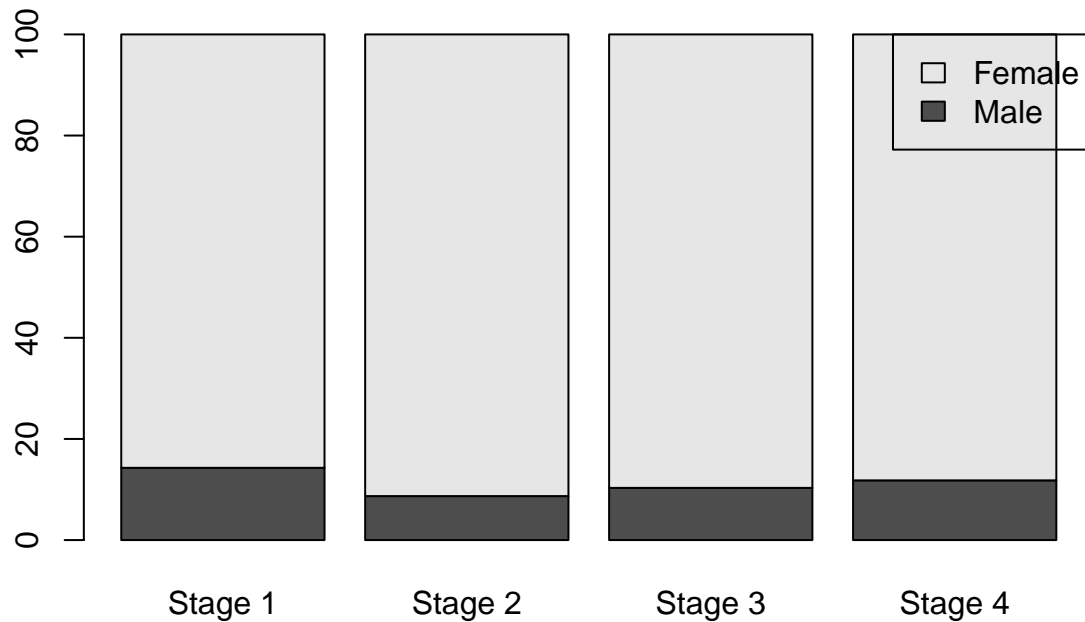
```
percent <- prop.table(counts, margin=2)*100
percent
```

```
##
##           Stage 1  Stage 2  Stage 3  Stage 4
##   Male   14.285714  8.695652 10.322581 11.805556
##   Female 85.714286 91.304348 89.677419 88.194444
```

After calculating the percentages, we use `barplot()` again, similar to the stacked bar graph:

```
barplot(percent,
  main="Relative frequency of sex within stage of disease from PBC study",
  legend = rownames(counts), beside=FALSE, args.legend = list(x = "topright"))
```

## Relative frequency of sex within stage of disease from PBC study



### 1.18 Creating line graphs

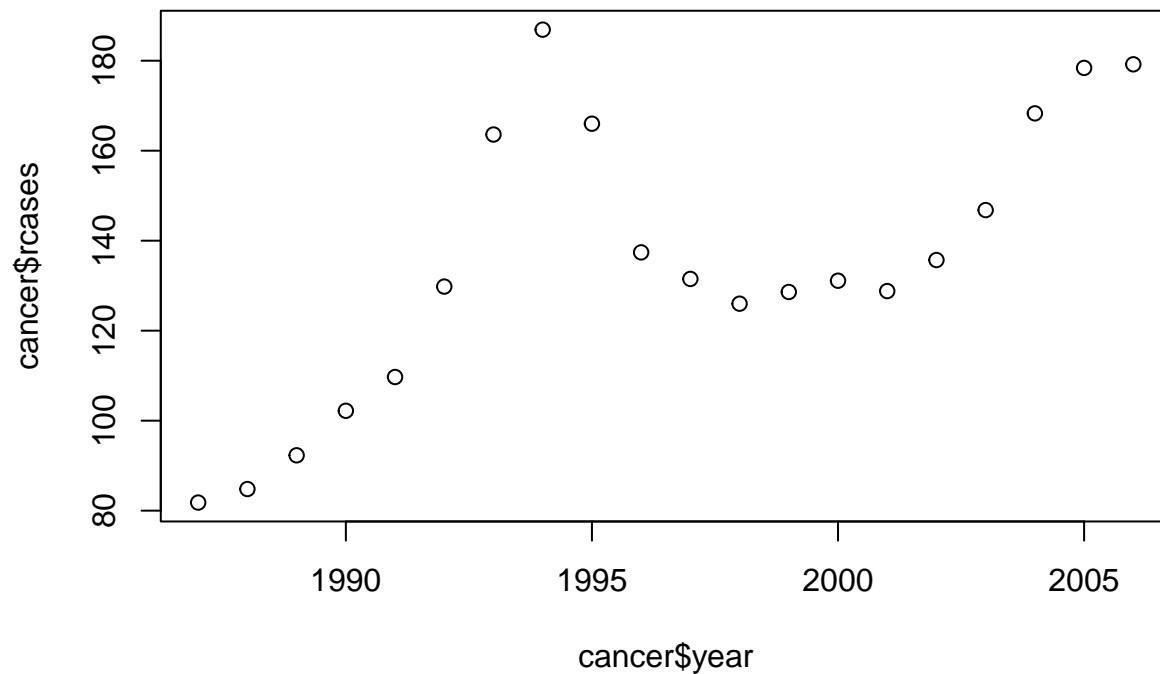
To demonstrate the graphing of aggregate data, we use the data on new cases and deaths from prostate cancer in males in NSW. This data has been entered as `Example_1.2.rds`.

```
cancer <- readRDS("data/examples/Example_1.2.rds")
summary(cancer)
```

```
##      year      ncases      ndeaths      rcases      rdeaths
## Min.   :1987   Min.   :1567   Min.   : 645.0   Min.   : 81.8   Min.   :31.10
## 1st Qu.:1992   1st Qu.:2804   1st Qu.: 788.2   1st Qu.:121.9   1st Qu.:34.67
## Median :1996   Median :3790   Median : 868.0   Median :131.3   Median :36.55
## Mean   :1996   Mean   :3719   Mean   : 855.0   Mean   :135.4   Mean   :37.09
## 3rd Qu.:2001   3rd Qu.:4403   3rd Qu.: 921.0   3rd Qu.:164.2   3rd Qu.:40.38
## Max.   :2006   Max.   :6158   Max.   :1044.0   Max.   :186.9   Max.   :43.80
```

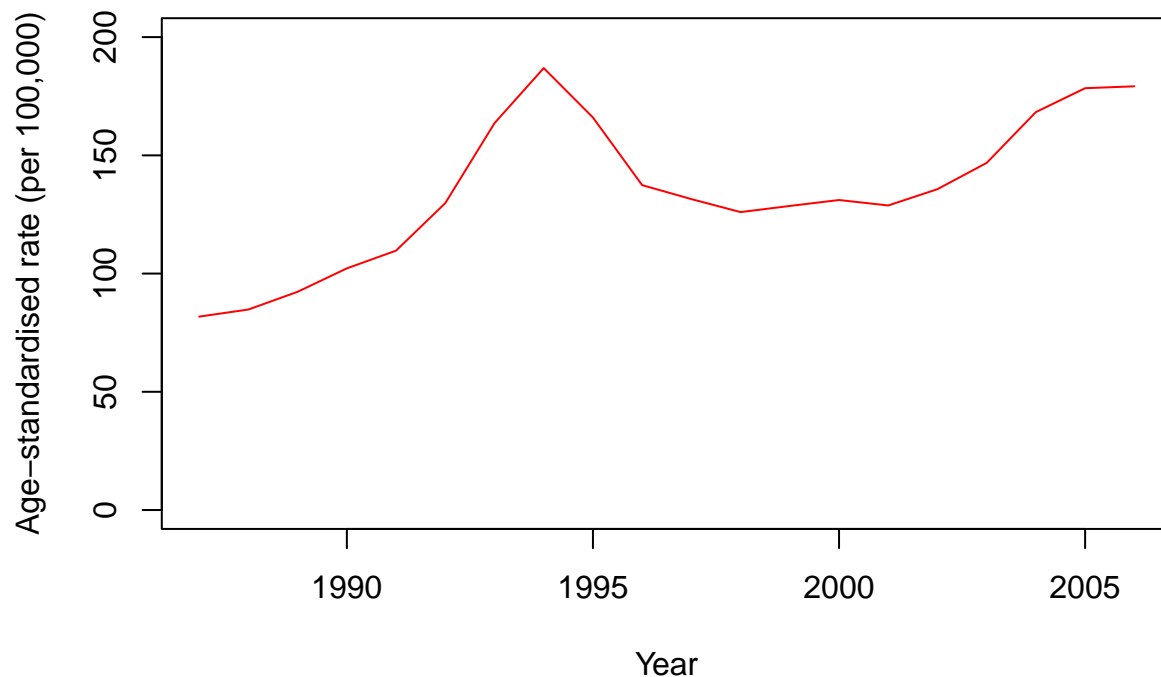
We begin by plotting cancer cases (as the y variable) against year (the x variable).

```
plot(x=cancer$year, y=cancer$rcases)
```



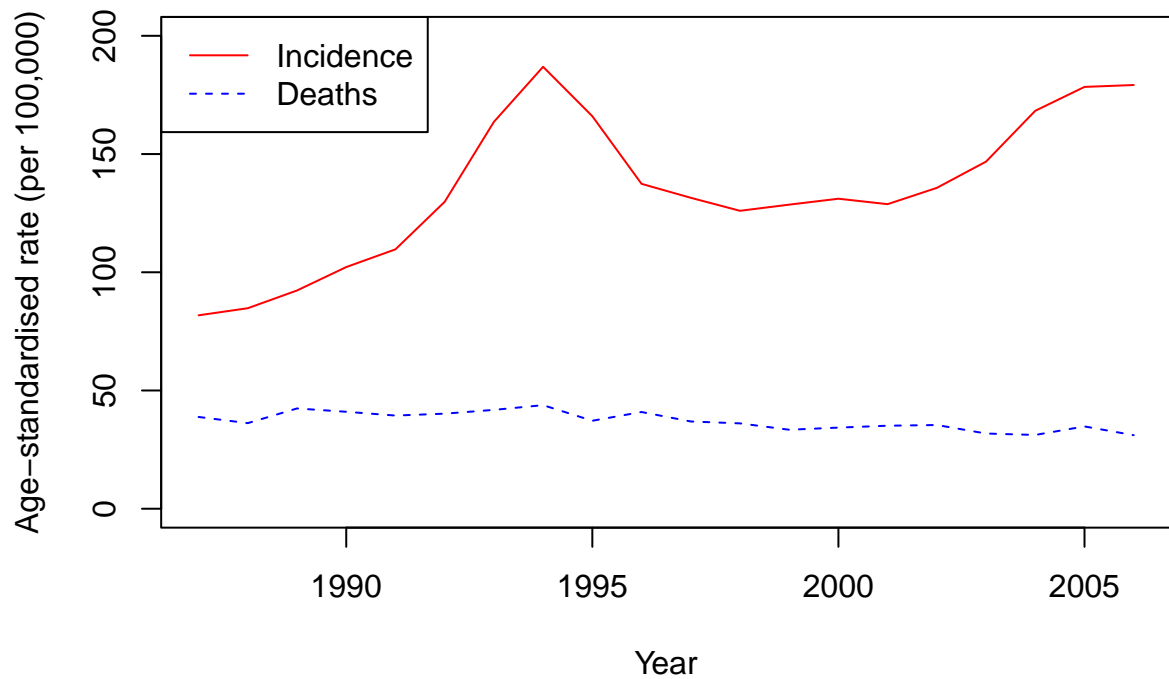
Let's define the plot to be joined by lines (`type="l"`), in the colour red (`col="red"`), providing meaningful labels for the x-axis and y-axis, and changing the scale of the y-axis to be between 0 and 200 (`ylim=c(0,200)`):

```
plot(x=cancer$year, y=cancer$rcases,
     type="l", col = "red",
     xlab = "Year",
     ylab = "Age-standardised rate (per 100,000)", ylim=c(0,200))
```



We can now add a second line to the plot using the `lines()` function, specifying a dashed line (`lty=2`), and add a legend to the plot:

```
plot(x=cancer$year, y=cancer$rcases, type="l", col = "red",  
     xlab = "Year", ylab = "Age-standardised rate (per 100,000)",  
     ylim=c(0,200))  
  
lines(cancer$year, cancer$rdeaths, col = "blue", type = "l", lty = 2)  
  
legend("topleft", legend=c("Incidence", "Deaths"),  
      col=c("red", "blue"), lty = 1:2)
```



Note: coding for graphs is not always straightforward. Two excellent resources for creating graphs in R are: [R Graphics Cookbook](#) and [The R Graph Gallery](#).



# Module 2

## Probability and probability distributions

### 2.1 Importing data into R

We have described previously how to import data that have been saved as R .rds files. It is quite common to have data saved in other file types, such as Microsoft Excel, or plain text files. In this section, we will demonstrate how to import data from other packages into R.

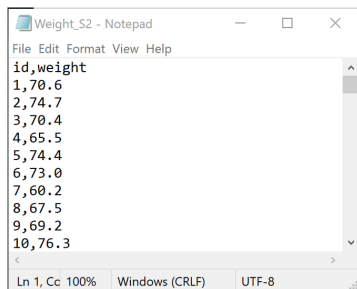
There are two useful packages for importing data into R: haven (for data that have been saved by Stata, SAS or SPSS) and readxl (for data saved by Microsoft Excel). Additionally, the labelled package is useful in working with data that have been labelled in Stata.

#### 2.1.1 Importing plain text data into R

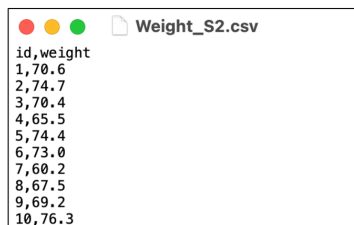
A csv file, or a “comma separated variables” file is commonly used to store data. These files have a very simple structure: they are plain text files, where data are separated by commas. csv files have the advantage that, as they are plain text files, they can be opened by a large number of programs (such as Notepad in Windows, TextEdit in MacOS, Microsoft Excel - even Microsoft Word). While they can be opened by Microsoft Excel, they can be opened by many other programs: the csv file can be thought of as the lingua-franca of data.

In this demonstration, we will use data on the weight of 1000 people entered in a csv file called weight\_s2.csv available on Moodle.

To confirm that the file is readable by any text editor, here are the first ten lines of the file, opened in Notepad on Microsoft Windows, and TextEdit on MacOS.



```
Weight_S2 - Notepad
File Edit Format View Help
id,weight
1,70.6
2,74.7
3,70.4
4,65.5
5,74.4
6,73.0
7,60.2
8,67.5
9,69.2
10,76.3
Ln 1, Cc 100% Windows (CRLF) UTF-8
```



```
Weight_S2.csv
id,weight
1,70.6
2,74.7
3,70.4
4,65.5
5,74.4
6,73.0
7,60.2
8,67.5
9,69.2
10,76.3
```

We can use the read.csv function:

```
sample <- read.csv("data/examples/Weight_s2.csv")
```

Here, the `read.csv` function has the default that the first row of the dataset contains the variable names. If your data do not have column names, you can use `header=FALSE` in the function.

Note: there is an alternative function `read_csv` which is part of the `readr` package (a component of the `tidyverse`). Some would argue that the `read_csv` function is more appropriate to use because of an issue known as `strings.as.factors`. The `strings.as.factors` default was removed in R Version 4.0.0, so it is less important which of the two functions you use to import a `.csv` file. More information about this issue can be found [here](#) and [here](#).

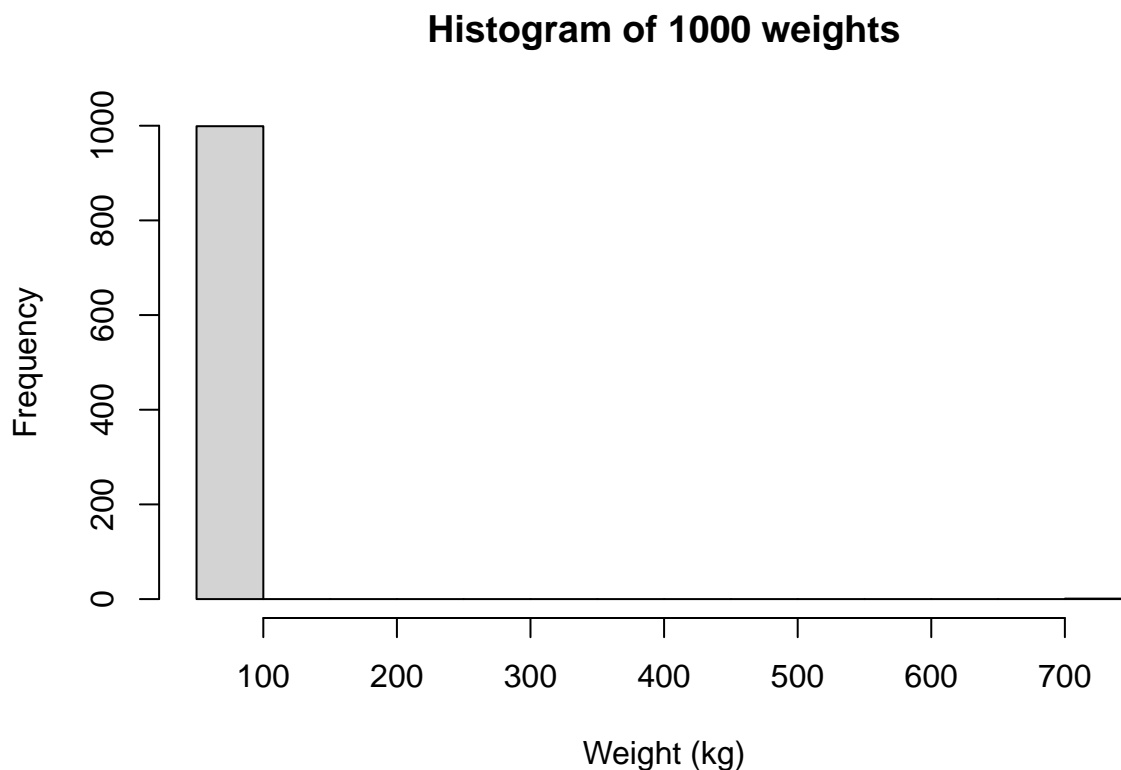
## 2.2 Checking your data for errors in R

Before you start describing and analysing your data, it is important to make sure that no errors have been made during the data entry process. Basically, you are looking for values that are outside the range of possible or plausible values for that variable.

If an error is found, the best method for correcting the error is to go back to the original data e.g. the hard copy questionnaire, to obtain the original value, entering the correct value into R. If the original data is not available or the original data is also incorrect, the erroneous value is often excluded from the dataset.

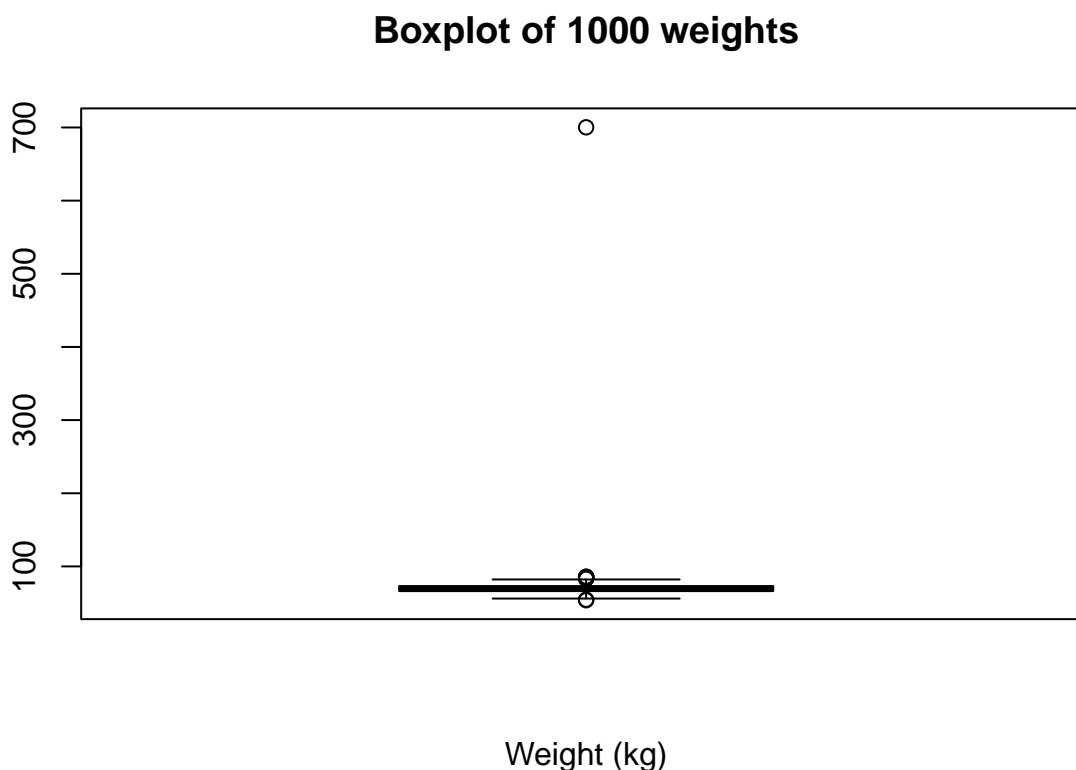
For continuous variables, the easiest methods are to examine a boxplot and histogram. For example, a boxplot and histogram for the weight variable we just imported appear as:

```
hist(sample$weight, xlab="Weight (kg)", main="Histogram of 1000 weights")
```



```
boxplot(sample$weight, xlab="Weight (kg)", main="Boxplot of 1000 weights")
```





There is a clear outlying point shown in the boxplot. Although not obvious, the same point is shown in the histogram as a bar around 700 with a very short height.

We can identify any outlying observations in the dataset using the `subset` function. You will need to decide if these values are a data entry error or are biologically plausible. If an extreme value or “outlier”, is biologically plausible, it should be included in all analyses.

For example, to list any observations from the `sample` dataset with a weight larger than 200:

```
subset(sample, weight>200)
```

Table 2.1

id	weight
58	700

We see that there is a very high value of 700.2kg. A value as high as 700kg is likely to be a data entry error (e.g. error in entering an extra zero) and is not a plausible weight value. Here, **you should check your original data**.

You might find that the original weight was recorded in medical records as 70.2kg. You can change this in R by writing code.

**Note:** many statistical packages will allow you to view a spreadsheet version of your data and edit values in that spreadsheet. This is not best practice, as corrected observations may revert to their original values depending on whether the edited data have been saved or not. By using code-based recoding, the changes will be reproduced the next time the code is run.

We will use an `ifelse` statement to recode the incorrect weight of 700.2kg into 70.2kg. The form of the `ifelse` statement is as follows:

```
ifelse(test, value_if_true, value_if_false)
```

Our code will create a new column (called `weight_clean`) in the `sample` dataframe. We will test whether `weight` is equal to 700.2; if this is true, we will assign `weight_clean` to be 70.2, otherwise `weight_clean` will equal the value of `weight`.

Putting it all together:

```
sample$weight_clean = ifelse(sample$weight==700.2, 70.2, sample$weight)
```

**Note:** if an extreme value lies within the range of biological plausibility it should not be removed from analysis.

Once you have checked your data for errors, you are ready to start analysing your data.

### 2.2.1 What on earth: == ?

In R, the test of equality is denoted by two equal signs: `==`. So we would use `==` to test whether an observation is equal to a certain value. Let's see an example:

```
# Test whether 6 is equal to 6
6 == 6
```

```
## [1] TRUE
```

```
# Test whether 6 is equal to 42
6 == 42
```

```
## [1] FALSE
```

You can read the `==` as "is equal to". So the code `sample$weight == 700.2` is read as: "is the value of `weight` from the data frame `sample` equal to 700.2?". In our `ifelse` statement above, if this condition is true, we replace `weight` by 70.2; if it is false, we leave `weight` as is.

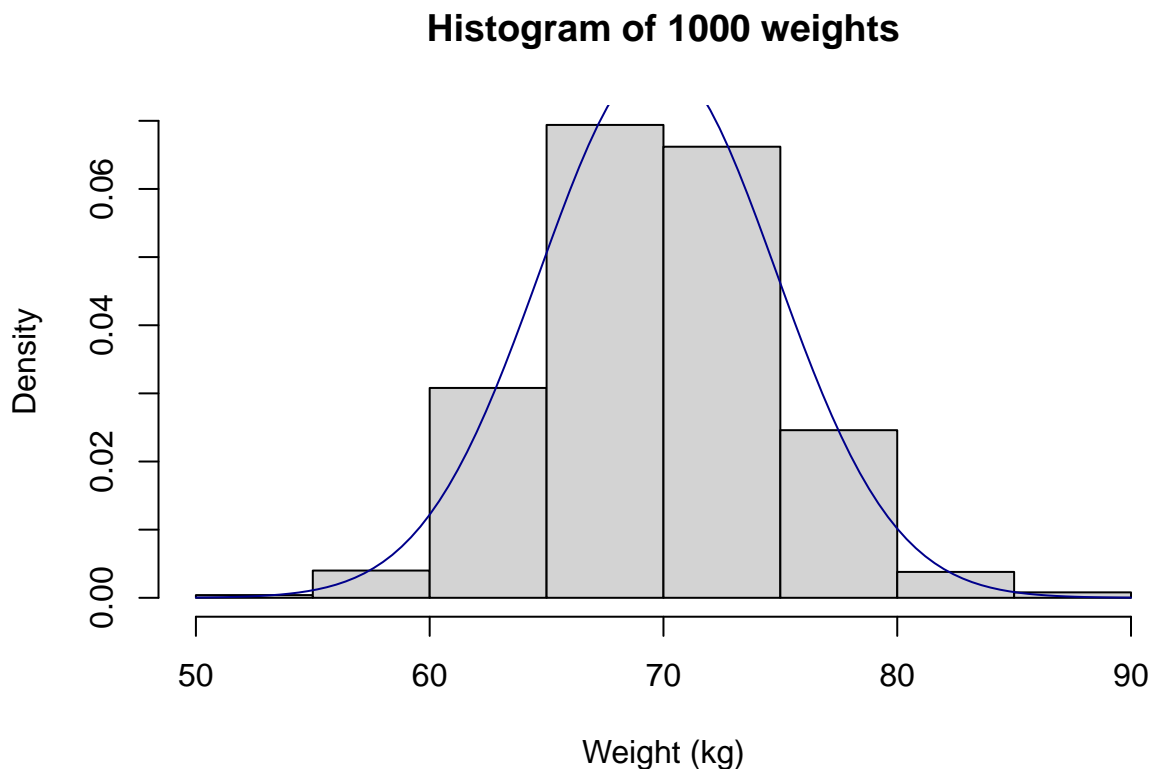
## 2.3 Overlaying a Normal curve on a histogram

It can be useful to produce a histogram with an overlayed Normal curve to assess whether our sample appears approximately Normally distributed. We can do this by plotting a histogram using the `hist()` function. As we're overlaying a probability distribution, we request the histogram be plotted on a probability scale, rather than a frequency scale, using `probability=TRUE`.

We then request a curve be overlayed using the `curve()` function:

- the curve should be based on the Normal distribution (`dnorm`);
  - with a mean equal to the mean of the cleaned weight: `mean(sample$weight_clean)`;
  - and a standard deviation equal to the standard deviation of the cleaned weight: `sd(sample$weight_clean)`
- using a dark-blue colour;
- and added to the previous histogram (rather than plotting the curve by itself): `add=TRUE`

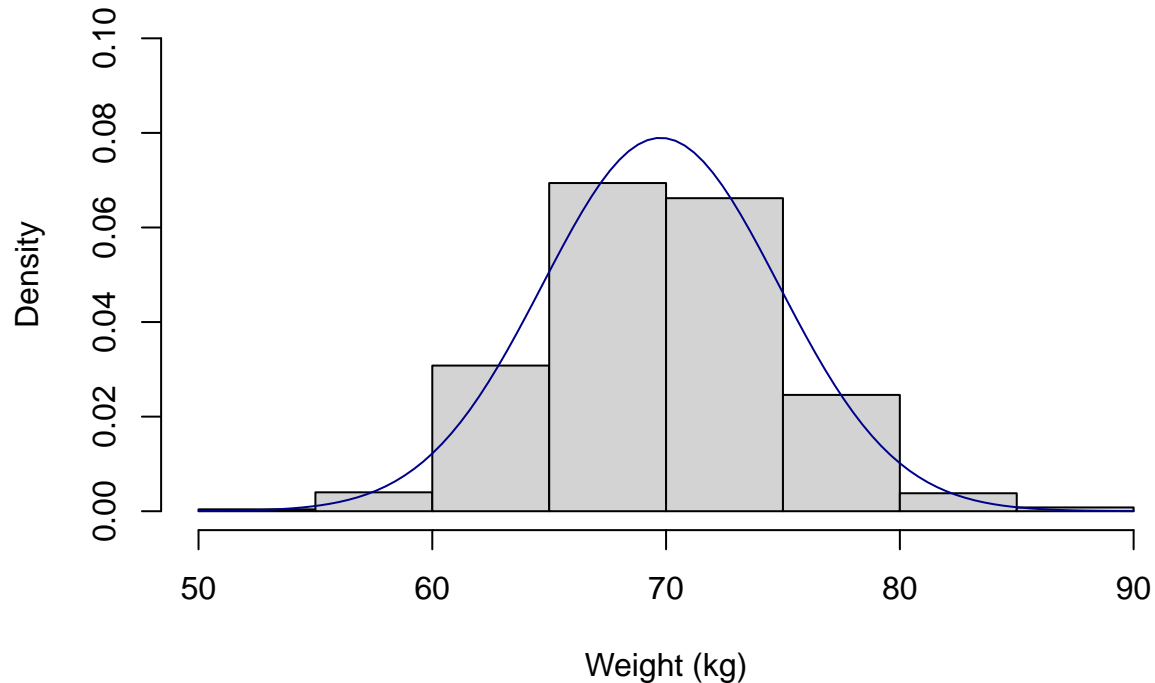
```
hist(sample$weight_clean,  
      xlab="Weight (kg)",  
      main="Histogram of 1000 weights",  
      probability = TRUE)  
  
curve(dnorm(x,  
            mean=mean(sample$weight_clean),  
            sd=sd(sample$weight_clean)),  
      col="darkblue",  
      add=TRUE)
```



Notice that the top of the curve is chopped off. We can plot the whole curve by extending the y-axis of the histogram to 0.1:

```
hist(sample$weight_clean,  
      xlab="Weight (kg)",  
      main="Histogram of 1000 weights",  
      probability = TRUE,  
      ylim=c(0,0.1))  
  
curve(dnorm(x,  
            mean=mean(sample$weight_clean),  
            sd=sd(sample$weight_clean)),  
      col="darkblue",  
      add=TRUE)
```

## Histogram of 1000 weights



### 2.4 Descriptive statistics for checking normality

All the descriptive statistics including *skewness* and *kurtosis* discussed in this module can be obtained using the `descriptives` function from the `jmv` package. In particular, skewness and kurtosis can be requested in addition to the default statistics by including: `skew=TRUE`, `kurt=TRUE`:

```
library(jmv)

descriptives(data=sample, vars=weight_clean, skew=TRUE, kurt=TRUE)
```

```
##
##  DESCRIPTIVES
##
##  Descriptives
##
##              weight_clean
##
##      N                1000
##      Missing            0
##      Mean             69.76450
##      Median           69.80000
##      Standard deviation  5.052676
##      Minimum          53.80000
##      Maximum          85.80000
##      Skewness          0.07360659
##      Std. error skewness 0.07734382
##      Kurtosis          0.05418774
##      Std. error kurtosis 0.1545343
##
```

## 2.5 Importing Excel data into R

Another common type of file that data are stored in is a Microsoft Excel file (.xls or .xlsx). In this demonstration, we will import a selection of records from a large health survey, stored in the file `health-survey.xlsx`.

The health survey data contains 1140 records, comprising:

- sex: 1 = respondent identifies as male; 2 = respondent identifies as female
- height: height in meters
- weight: weight in kilograms

To import data from Microsoft Excel, we can use the `read_excel()` function in the `readxl` package.

```
library(readxl)

survey <- read_excel("data/examples/health-survey.xlsx")
summary(survey)
```

```
##      sex      height      weight
## Min.   :1.00   Min.   :1.220   Min.   : 22.70
## 1st Qu.:1.00   1st Qu.:1.630   1st Qu.: 68.00
## Median :2.00   Median :1.700   Median : 79.40
## Mean   :1.55   Mean   :1.698   Mean   : 81.19
## 3rd Qu.:2.00   3rd Qu.:1.780   3rd Qu.: 90.70
## Max.   :2.00   Max.   :2.010   Max.   :213.20
```

We can see that sex has been entered as a numeric variable. We should transform it into a factor so that we can assign labels to each category:

```
survey$sex <- factor(survey$sex, level=c(1,2), labels=c("Male", "Female"))
summary(survey$sex)
```

```
##      Male Female
##      513     627
```

We also note that height looks like it has been entered as meters, and weight as kilograms.

## 2.6 Generating new variables

Our health survey data contains information on height and weight. We often summarise body size using BMI: body mass index which is calculated as:  $\frac{\text{weight (kg)}}{(\text{height (m)})^2}$

We can create a new column in our data frame in many ways, such as using the following approach:

```
dataframe$new_column = <formula>
```

For example:

```
survey$bmi = survey$weight / (survey$height^2)
```

We should check the construction of the new variable by examining some records. The `head()` and `tail()` functions list the first and last 6 records in any dataset. We can also examine a histogram and boxplot:

```
head(survey)
```

Table 2.2

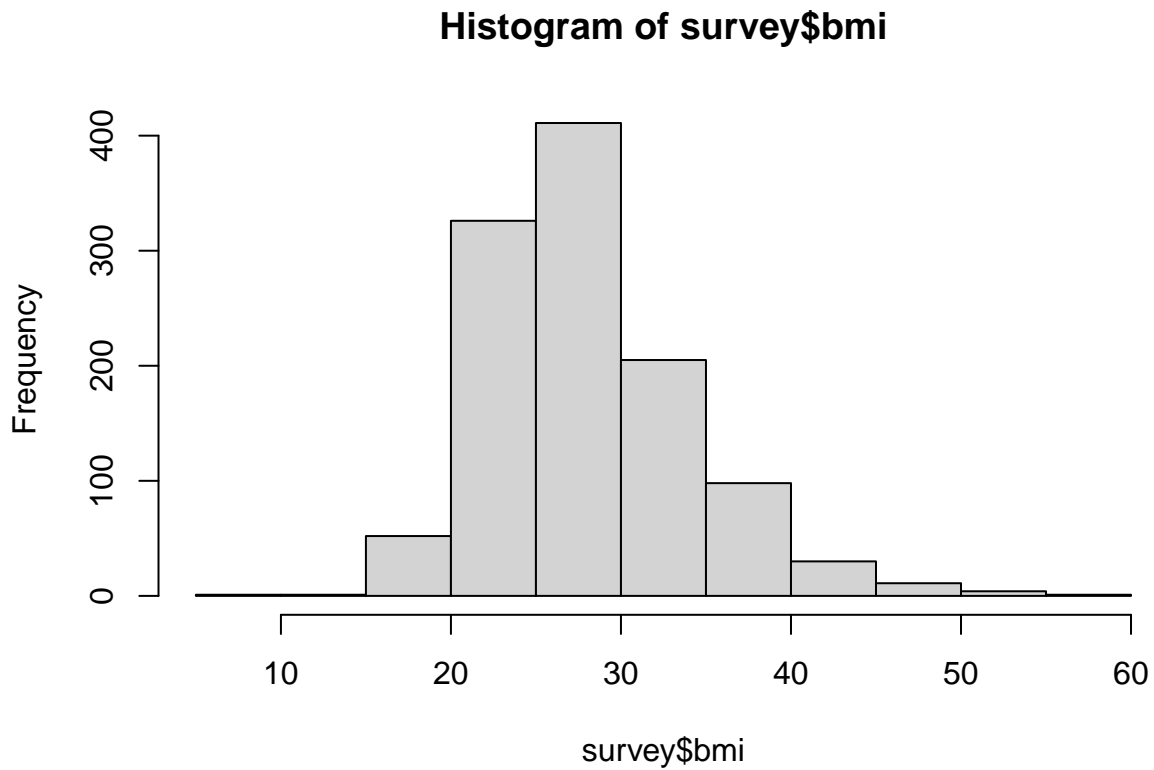
sex	height	weight	bmi
Male	1.63	81.7	30.8
Male	1.63	68	25.6
Male	1.85	97.1	28.4
Male	1.78	89.8	28.3
Male	1.73	70.3	23.5
Female	1.57	85.7	34.8

```
tail(survey)
```

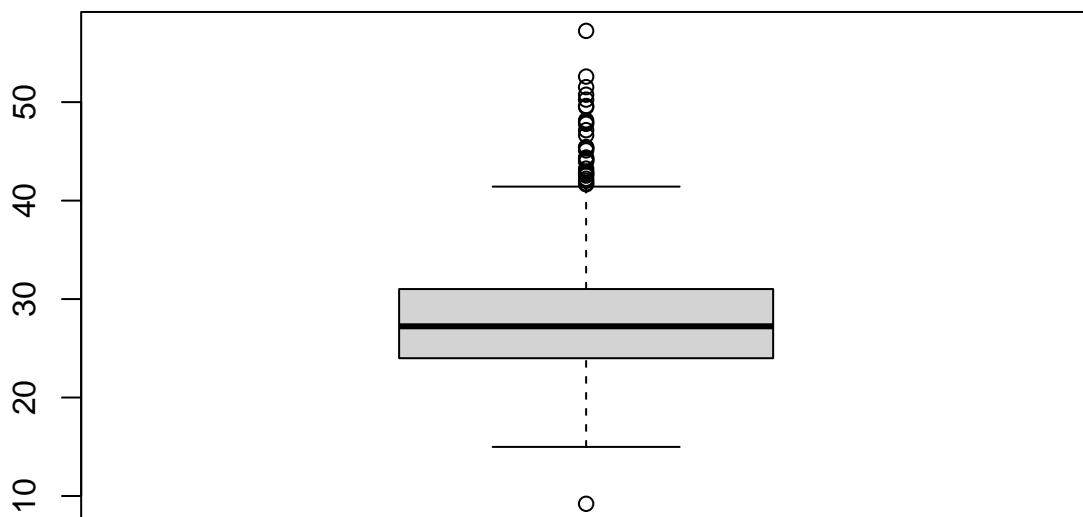
Table 2.3

sex	height	weight	bmi
Female	1.65	95.7	35.2
Male	1.8	79.4	24.5
Female	1.73	83	27.7
Female	1.57	61.2	24.8
Male	1.7	73	25.3
Female	1.55	91.2	38

```
hist(survey$bmi)
```



```
boxplot(survey$bmi)
```



In the general population, BMI ranges between about 15 to 30. It appears that BMI has been correctly generated in this example. We should investigate the very low and some of the very high values of BMI, but this will be left for another time.

## 2.7 Summarising data by another variable

We will often want to calculate the same summary statistics by another variable. For example, we might want to calculate summary statistics for BMI for males and females separately. We can do this in the `descriptives` function by defining `sex` as a `splitBy` variable:

```
library(jmv)
descriptives(data=survey, vars=bmi, splitBy = sex)
```

```
##
##  DESCRIPTIVES
##
##  Descriptives
##
##              sex      bmi
##
##  N              Male      513
##              Female      627
##  Missing        Male        0
##              Female        0
##  Mean           Male    28.29561
##              Female    27.81434
##  Median         Male    27.39592
##              Female    26.66667
##  Standard deviation Male    5.204975
##              Female    6.380523
##  Minimum        Male    16.47519
##              Female    9.209299
##  Maximum        Male    57.23644
##              Female    52.59516
##
```

## 2.8 Plotting data by another variable

Unfortunately, it is not straight-forward to create separate plots for every level of another variable. We will demonstrate by plotting BMI by sex using our health survey data.

The following steps are not the most efficient way of doing this, but are easy to follow and understand. We first begin by creating two new data frames, for males and females separately, using the `subset()` function:

```
survey_males <- subset(survey, sex=="Male")
survey_females <- subset(survey, sex=="Female")
```

Note that we use the **label** for sex, not the underlying numeric value, as sex is a **factor**.

We can now create histograms and boxplots of BMI for males and females separately. To place the graphs next to each other in a single figure, we can use the `par` function, which sets the *graphics parameters*. Essentially, we want to tell R to split a plot window into a matrix with *nr* rows and *nc* columns, and we fill the cells by rows (`mfrow`) or columns (`mfcols`).

For example, to plot four figures in a single plot, filled by rows, we use `par(mfrow=c(2,2))`.

When we are done plotting multiple graphs, we can reset the graphics parameters by submitting `par(mfrow=c(1,1))`.

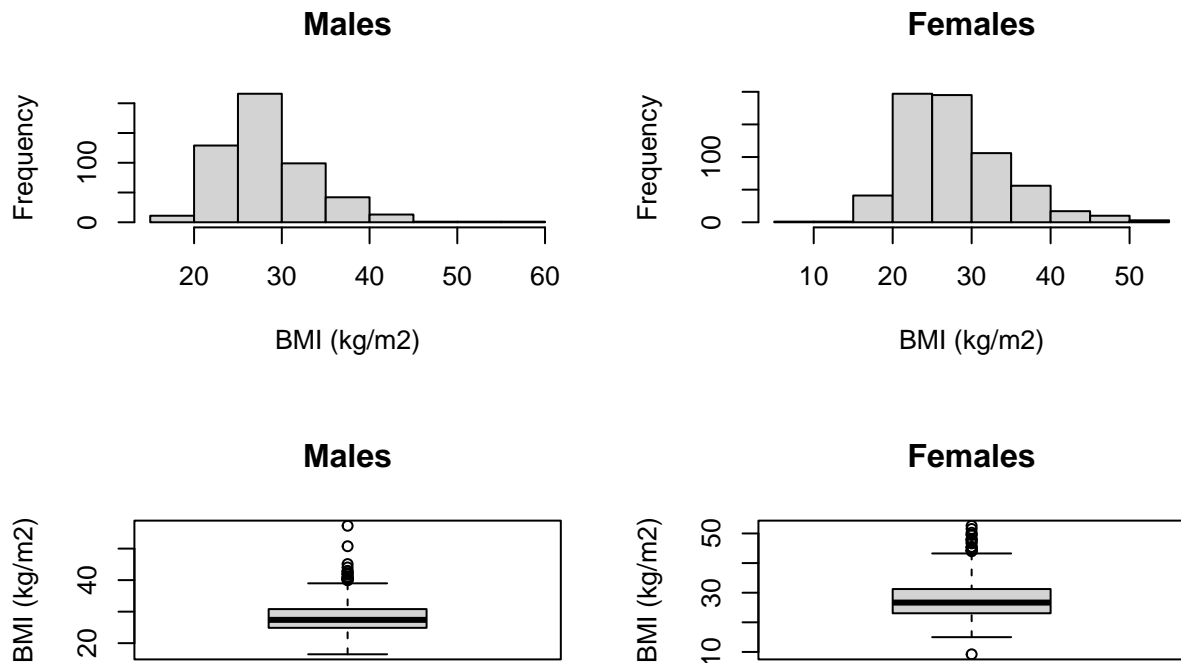
```
# Set the graphics parameters to plot 2 rows and 2 columns:
par(mfrow=c(2,2))
```

```
# Specify each plot separately
hist(survey_males$bmi, xlab="BMI (kg/m2)", main="Males")
```



```
hist(survey_females$bmi, xlab="BMI (kg/m2)", main="Females")

boxplot(survey_males$bmi, ylab="BMI (kg/m2)", main="Males")
boxplot(survey_females$bmi, ylab="BMI (kg/m2)", main="Females")
```



```
# Reset graphics parameters
par(mfrow=c(1,1))
```

## 2.9 Recoding data

One task that is common in statistical computing is to recode variables. For example, we might want to group some categories of a categorical variable, or to present a continuous variable in a categorical way.

In this example, we can recode BMI into the following categories as suggested by the World Health Organisation [footnote]:

- Underweight:  $\text{BMI} < 18.5$
- Normal weight:  $18.5 \leq \text{BMI} < 25$
- Pre-obesity:  $25 \leq \text{BMI} < 30$
- Obesity Class I:  $30 \leq \text{BMI} < 35$
- Obesity Class II:  $35 \leq \text{BMI} < 40$
- Obesity Class III:  $\text{BMI} \geq 40$

The quickest way to recode a continuous variable into categories is to use the `cut` command which takes a continuous variable, and “cuts” it into groups based on the specified “cutpoints”:

```
survey$bmi_cat <- cut(survey$bmi, c(0, 18.5, 25, 30, 35, 40, 100))
```

Notice that lower (BMI=0) and upper (BMI=100) bounds have been specified, as both a lower and upper limit must be defined for each group.

If we examine the new `bmi_cat` variable:

```
summary(survey$bmi_cat)
```

```
## (0,18.5] (18.5,25] (25,30] (30,35] (35,40] (40,100]
##      18      362      411      205      97      47
```

we see that each group has been labelled (a, b]. This notation is equivalent to: greater than a, and less than or equal to b. The `cut` function excludes the lower limit, but includes the upper limit. Our BMI ranges have been defined to include the lower limit, and exclude the upper limit (for example, greater than or equal to 30 and less than 35).

We can specify this recoding using the `right=FALSE` option:

```
survey$bmi_cat <- cut(survey$bmi, c(0, 18.5, 25, 30, 35, 40, 100), right=FALSE)
summary(survey$bmi_cat)
```

```
## [0,18.5) [18.5,25) [25,30) [30,35) [35,40) [40,100)
##      18      362      411      201      101      47
```

## 2.10 Computing binomial probabilities using R

There are two R functions that we can use to calculate probabilities based on the binomial distribution: `dbinom` and `pbinom`:

- `dbinom(x, size, prob)` gives the probability of obtaining  $x$  successes from  $size$  trials when the probability of a success on one trial is `prob`;
- `pbinom(q, size, prob)` gives the probability of obtaining  $q$  **or fewer** successes from  $size$  trials when the probability of a success on one trial is `prob`;
- `pbinom(q, size, prob, lower.tail=FALSE)` gives the probability of obtaining **more than**  $q$  successes from  $size$  trials when the probability of a success on one trial is `prob`.

To do the computation for part (a) in Worked Example 2.1, we will use the `dbinom` function with:

- $x$  is the number of successes, here, the number of smokers (i.e.  $k=3$ );
- $size$  is the number of trials (i.e.  $n=6$ );
- and  $prob$  is probability of drawing a smoker from the population, which is 19.8% (i.e.  $p=0.198$ ).

Replace each of these with the appropriate number into the formula:

```
dbinom(x=3, size=6, prob=0.198)
```

```
## [1] 0.08008454
```

To calculate the upper tail of probability in part (b), we use the `pbinom(lower.tail=FALSE)` function. Note that the `pbinom(lower.tail=FALSE)` function **does not include**  $q$ , so to obtain 4 or more successes, we need to enter  $q=3$ :

```
pbinom(q=3, size=6, prob=0.198, lower.tail=FALSE)
```

```
## [1] 0.01635325
```

For the lower tail for part (c), we use the pbinom function:

```
pbinom(q=2, size=6, prob=0.198)
```

```
## [1] 0.9035622
```

## 2.11 Computing probabilities from a Normal distribution

We can use the pnorm function to calculate probabilities from a Normal distribution:

- pnorm(q, mean, sd) calculates the probability of observing a value of q or less, from a Normal distribution with a mean of mean and a standard deviation of sd. Note that if mean and sd are not entered, they are assumed to be 0 and 1 respectively (i.e. a standard normal distribution.)
- pnorm(q, mean, sd, lower.tail=FALSE) calculates the probability of observing a value of q or more, from a Normal distribution with a mean of mean and a standard deviation of sd.

To obtain the probability of obtaining 0.5 or greater from a standard normal distribution:

```
pnorm(0.5, lower.tail=FALSE)
```

```
## [1] 0.3085375
```

To calculate the worked example: Assume that the mean diastolic blood pressure for men is 77.9 mmHg, with a standard deviation of 11. What is the probability that a man selected at random will have high blood pressure (i.e. diastolic blood pressure greater than or equal to 90)?

```
pnorm(90, mean=77.9, sd=11, lower.tail=FALSE)
```

```
## [1] 0.1356661
```



# Module 3

## Precision: R notes

### 3.1 Calculating a 95% confidence interval of a mean

#### 3.1.1 Individual data

To demonstrate the computation of the 95% confidence interval of a mean we have used data from Example\_1.3.rds which contains the weights of 30 students:

```
library(jmv)

students <- readRDS("data/examples/Example_1.3.rds")

summary(students)
```

```
##      weight      gender
## Min.   :60.00   Male   :16
## 1st Qu.:67.50   Female:14
## Median :70.00
## Mean   :70.00
## 3rd Qu.:74.38
## Max.   :80.00
```

The mean and its 95% confidence interval can be obtained many ways in R. One way is to use the `descriptives` function in the `jmv` package. By default, `descriptives` does not provide a confidence interval, but we can request it by specifying `ci=TRUE`:

```
descriptives(data=students, vars=weight, ci=TRUE)
```

```
##
## DESCRIPTIVES
##
## Descriptives
##
##              weight
##
## N              30
## Missing         0
## Mean           70.00000
## 95% CI mean lower bound 68.19545
```

```
##      95% CI mean upper bound    71.80455
##      Median                    70.00000
##      Standard deviation         5.042919
##      Minimum                    60.00000
##      Maximum                    80.00000
##
```

### 3.1.2 Summarised data

For Worked Example 3.2 where we are given the sample mean, sample standard deviation and sample size. R does not have a built-in function to calculate a confidence interval from summarised data, but we can write our own.

**Note: writing your own functions is beyond the scope of this course. You should copy and paste the code provided to do this.**

```
### Copy this section
ci_mean <- function(n, mean, sd, width=0.95, digits=3){
  lcl <- mean - qt(p=(1 - (1-width)/2), df=n-1) * sd/sqrt(n)
  ucl <- mean + qt(p=(1 - (1-width)/2), df=n-1) * sd/sqrt(n)

  print(paste0(width*100, "%", " CI: ", format(round(lcl, digits=digits), nsmall = digits),
    " to ", format(round(ucl, digits=digits), nsmall = digits) ))
}
### End of copy

ci_mean(n=30, mean=70, sd=6, width=0.95)
```

```
## [1] "95% CI: 67.760 to 72.240"
```

```
ci_mean(n=30, mean=70, sd=6, width=0.99)
```

```
## [1] "99% CI: 66.981 to 73.019"
```

# Module 4

## Hypothesis testing

### 4.1 One sample t-test

We will use data from `Example_4.1.rds` to demonstrate how a one-sample t-test is conducted in R.

```
bloodpressure <- readRDS("data/examples/Example_4.1.rds")  
  
summary(bloodpressure)
```

```
##          dbp  
## Min.      : 24.00  
## 1st Qu.: 64.00  
## Median : 72.00  
## Mean   : 72.41  
## 3rd Qu.: 80.00  
## Max.    :122.00  
## NA's    :35
```

To test whether the mean diastolic blood pressure of the population from which the sample was drawn is equal to 71, we can use the `t.test` command:

```
t.test(bloodpressure$dbp, mu=71)
```

```
##  
## One Sample t-test  
##  
## data: bloodpressure$dbp  
## t = 3.0725, df = 732, p-value = 0.002202  
## alternative hypothesis: true mean is not equal to 71  
## 95 percent confidence interval:  
## 71.50732 73.30305  
## sample estimates:  
## mean of x  
## 72.40518
```

The output gives a test statistic, degrees of freedom and a P values from the two-sided test. The mean of the sample is provided, as well as the 95% confidence interval.





# Bibliography

Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, New York Berlin Heidelberg, December 2010. ISBN 978-1-4419-3161-0.