MACQUARIE University
SYDNEY·AUSTRALIA

STAT2170 AND STAT6180
APPLIED STATISTICS

SGTA Week 2 : Introduction to R/RStudio 2022

Part 1

To install R (if it has not been done yet),

- For windows, please go to
  - https://cran.r-project.org/bin/windows/base/
- For Mac, please go to
  - https://cran.r-project.org/bin/macosx/

To install Rstudio (if it has not been done yet),

- For all systems, please go to the "all installer" section of
  - https://www.rstudio.com/products/rstudio/download/#download

Now we review some common commands in `R`.

Question 1: Install packages.

Packages include reusable `R` functions and sample data; for example, `stats` package incoporates many useful statistical functions.

- Please correct all the errors in the codes below, which try to install the `stats` package

```
install.package(stats)
```

- Please try to, with function `library`, add to the library the `stats` package you installed
- Please try to install the `rmarkdown` package, if you have not done it yet

Question 2: Algebraic calculation

At its most basic level, `R` can be viewed as a fancy calculator.

- Calculate 2^{10}.
- Calculate $\log_2 10$ and round it to 2 decimal places.
- Please correct all the errors in the codes below, which try to calculate $\cos(2\pi)$, $\arcsin(1/2)$, and the upper 95% quantile for the t-distribution with 10 degrees of freedom.

```
cos(2pi)
arcsin(1/2)
qt(0.95)
```

Question 3: Naming rules

`R` only allows certain combination of characters to be the names of variables and the rules are

- A variable name must start with a letter and can be a combination of letters, digits, period(.) and underscore(_). If it starts with period(.), it cannot be followed by a digit.
- A variable name cannot start with a number or underscore (_).
- Variable names are case-sensitive (`age`, `Age` and `AGE` are three different variables).
- Reserved words cannot be used as variables (such as `TRUE`, `FALSE`, `NULL`, `if`).

Which of the followings are invalid variable names? Justify your answer. You can then run the code to confirm.

```
1X <- 2
X1 <- 3
X_ <- 4
_X <- 5
.X <- 6
X. <- 7
```

Question 4: Vector operations

`R` can store a series of numbers as a vector with function `c()`

- The daily average temperatures (in Celsius) in Sydney from Jan to Apr:

  $$23.5,\ 23.4,\ 22.1,\ 19.5$$

- Please put these into a vector called `temp_C`.
- Calculate the mean, median, and standard deviation of `temp_C`.
- Convert it into Fahrenheit temperatures and store it in `temp_F`.
- What's your guess on the sample correlation coefficient between `temp_C` and `temp_F`?
- Find the function in `R` that calculates the sample correlation coefficient between two variables. Use it to check your guess :)
- Please correct all the errors in the codes below, which try to calculate the mean of $1, \ldots, 10$

```
x <- 1:10
Mean(x)
```

Question 5: Vector operations

`R` can also generate vectors with `seq()`, `rep()`, and random number generators.

- Please generate 10 independent and identically distributed standard normal random variables and stored it in `x`.
- Calculate the sum of the squares of the entries of `x`.
- Generate the vector of $1, 3, 5, 7, \ldots, 17, 19$ and store it in `y`.
- Calculate the inner product:

$$x \cdot y = \sum_{i=1}^{n} x_i \times y_i,$$

  a.k.a., dot product, of `x` and `y`.

Part 2

Next, this SGTA involves analysing the melting points for 60 samples of a particular hydrogenated fat (taken from Sokal and Rohlf (1995)). The slight differences in impurities in the samples results in the actual melting point varying from sample to sample. The first six observations are shown in the table below

meltpoint	burner
94.10	b
94.17	b
94.67	a
93.83	b
94.82	b
94.16	a

The data are stored in a text file worksheet called `melt.dat` available on iLearn. Download this file and save it to a directory (preferably your working directory) on your computer that you wish to work in.

Create a `*.rproj` file as per the instruction in Week 1 Lecture and that should bring you back to your working directory every single time.

- Alternatively, change the R/RStudio working directory to where you downloaded the file. To do this, navigate to the following, *Session* → *Set Working Directory* → *Choose Directory* and select the directory you wish to work in.

If you are using `RStudio` via https://rstudio.cloud. You will need to upload `melt.dat` to your account first. You can find the `Upload` button within the `Files` pane in `RStudio`.

To load the data into R/Rstudio try **EITHER** of the following:

a. Read in the data by typing the following

```
melt = read.table("melt.dat", header = TRUE)
```

   You can use `?read.table` to see what other functions are available to read in other types of text data.

b. You can use an importer to read in the data. Select the *Import Dataset* dropdown from the *Environment* tab and select *From Text (base)...*

- Select the `melt.dat` file from the directory you saved the file.
- Click *Import*.

Please refer to this article on how to use the importer to read in other file types.

To verify you have imported the data successfully, you should be able to view the data. You can do this in certain ways.

- Double click the name of the dataset in the *Environment tab* in `RStudio`.
- Type the name of the dataset in the `R` prompt
- **NOTE**: The variable `meltpoint` lives inside the dataset
- Reference the data

```
# Note the difference between melt and melt$meltpoint
head(melt) # Shows the dataset with meltpoint and burner variables inside it
```

```
##   meltpoint burner
## 1     94.10      b
## 2     94.17      b
## 3     94.67      a
## 4     93.83      b
## 5     94.82      b
## 6     94.16      a
```

```
head(melt$meltpoint) # Shows only the values of the meltpoint variable
```

```
## [1] 94.10 94.17 94.67 93.83 94.82 94.16
```

You can avoid having to reference `meltpoint` using the `$` operator with `attach(melt)` but this is not recommended in practice (if you have the same variable name inside two datasets that are both attached, it will cause problems since the variable name can only be used once)

Question 1

Fully examine the data set and summarize it both numerically and graphically. Use the output to decide what distribution best describes the data and how valid you think it is to assume that distribution is Normal. Is there anything unusual about the data (outliers for example) that you should consider?

a. Obtain a Histogram, a Boxplot and a Stem and Leaf plot of the data.

Hint: The commands `hist(x)`, `boxplot(x)` and `stem(x)` will create the appropriate graphs of a variable called `x`. Note our variable is called `meltpoint` inside the `melt` dataset.

b. Obtain numerical summaries (Quartiles, the standard deviation, the mean of the data and the number of observations)

Hint: For a variable, `x`; `summary(x)`, `mean(x)`, `sd(x)` and `length(x)` will return a summary, the mean, the standard deviation, and the number of observations in the variable x in R/RStudio. **Note**: our variable is called `meltpoint` inside the `melt` dataset.

Question 2

For health related reasons, the legal requirement is that the melting point should be 94.4 degrees Celsius. Carry out an appropriate statistical test to investigate whether there is substantial evidence that the average melting point for this source of the hydrogenated fat is not 94.4°C. Use a significance level of 5%. Would changing the level of significance to 1% have any effect on your conclusions?

Remember that in writing up any test of significance you need to state clearly the hypotheses, the test statistic, the p-value (probability) and your conclusions. Make sure that you understand how each of the values in the `R` output was obtained.

Hint: Use the `t.test` command. E.g. For a variable `x` and $H_0 : \mu = 94.4$ with two sided alternative, one would type

```
t.test(x, mu = 94.4)
```

Check the help page to see how to change the level of α (or $1 - \alpha$ confidence level in the confidence interval). You can find the help by typing

```
? t.test
```

at the `R` prompt.

Question 3

Obtain the 95% and 99% confidence intervals for the melting point for this particular fat. Comment on the difference between the two confidence intervals.

Confidence intervals are available as part of the hypothesis test output of the `t.test` command.

Question 4

If we had carried out the test at the 5% level of significance for whether the melting point is significantly **lower** than 94.4°, would this have affected our conclusions? Why?

References

Sokal, RR, and FJ Rohlf. 1995. *Biometry: The Principles of Statistics in Biological Research.* New York, NY: WH Freeman; Co.