STAT1201
Analysis of Scientific Data
Summer Semester 2022

- Introduction to R and RStudio
- Introduction to The Islands
- Different types of variables
- Observational studies and experimental studies
- What is hypothesis testing and how to make decisions from hypothesis testing – important for your paper review (statistical component)

Introduction to R and RStudio

R is a free software programming language for statistical analysis. The software is distributed through CARN (Comprehensive R Achieve Network)

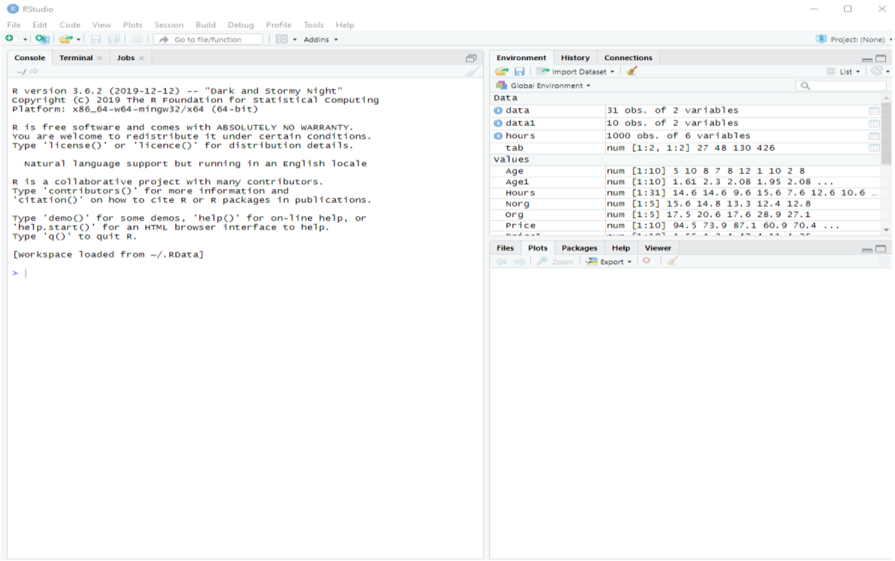
RStudio is an integrated development environment for (IDE) for R. We will be using RStudio in this course.

You need R to use RStudio. R is the base pack that is essential to run RStudio.

Thus, you first need to download R and then RStudio.

Links to download R and RStudio and some other relevant material are available in the “RStudio” folder under Learning Resources on BB.

RStudio



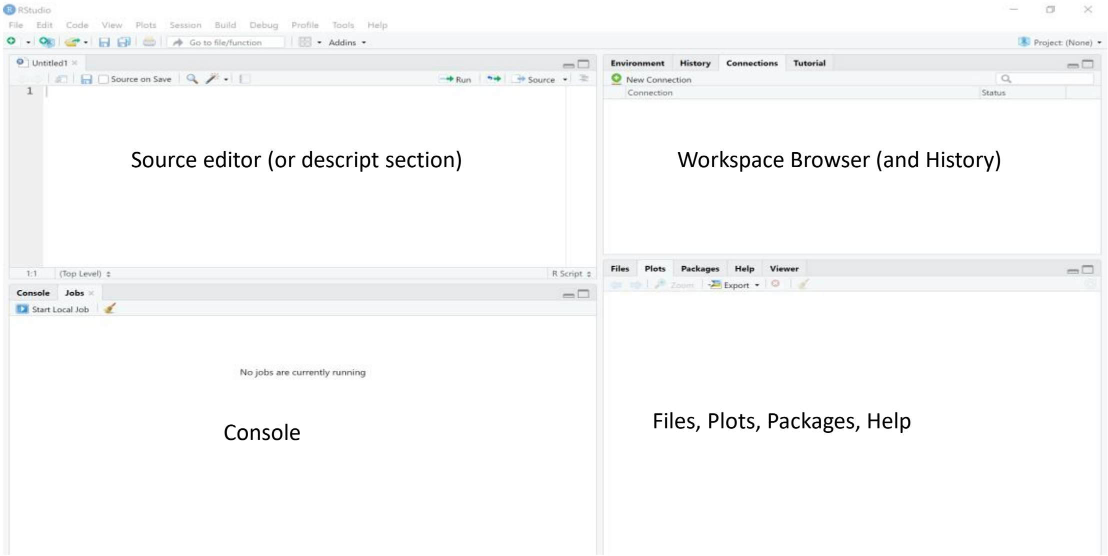
```
R version 3.6.2 (2019-12-12) -- "Dark and Stormy Night"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

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.

  Natural language support but running in an English locale

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]

> |
```

```
Data
  data     31 obs. of 2 variables
  data1    10 obs. of 2 variables
  hours    1000 obs. of 6 variables
  tab      num [1:2, 1:2] 27 48 130 426
Values
  Age      num [1:10] 5 10 8 7 8 12 1 10 2 8
  Age1     num [1:10] 1.61 2.3 2.08 1.95 2.08 ...
  Hours    num [1:31] 14.6 14.6 9.6 15.6 7.6 12.6 10.6 ...
  Norg     num [1:5] 15.6 14.8 13.3 12.4 12.8
  org      num [1:5] 17.5 20.6 17.6 28.9 27.1
  Price    num [1:10] 94.5 73.9 87.1 60.9 70.4 ...
```

3

RStudio

Four (4) main windows (or fields).



No jobs are currently running

Source editor (or descript section)

Workspace Browser (and History)

Console

Files, Plots, Packages, Help

4

RStudio

Source editor (or descript section) – Use this to write R script files and save them. Create a new script file by going to **File > New File > R script**

Use "Run" button to run the script codes. This will send command down into the console and execute the command.

Workspace browser has 4 tabs.

The environment tab will show you what objects are currently loaded into your working space.

The history tab will keep a record of all your commands that have been issued into the console.

Console – You will type commands here. For example, try a simple equation like 2*3 or 8/2 and see the return

Files, Plots, Packages, Help

Please refer next slide

5

RStudio

Files, Plots, Packages, Help

Files – Your Home folder contents are listed here. By default, this will be your documents folder. This is also called the "Working Directory". R will look for files in this Home folder when requested to import into the environment.

If you want to change your working directory (i.e. document folder by default), click on the three dots (…) on the right-hand side of the panel, browse the directory for you to work and save your script files and data files.

You can set this new folder as your working directory by clicking on the drop-down menu of "More" and then click on "Set As Working Directory".

Plots – Shows the latest plot that was generated through the console. Use arrow keys if you have generated multiple plots. Use Export to download the plot.

Packages – This is an overview of all the packages that are installed.

Help – Can be used to get help using R.

6

Read a “.csv” data file in RStudio and Basic functions in R

The data file should be saved in the working directory.

Suppose that your data file is saved in documents folder or in your working directory with a file name “SurveyData.csv”. Read it into RStudio and rename as “survey”.

survey = read.csv(“SurveyData.csv”)

Refer “R-Summary.pdf” file located in the RStudio folder under Learning Resources on BB for a handy summary of the functions in R that you will meet in the course this semester.

Examples

log() – for natural logs

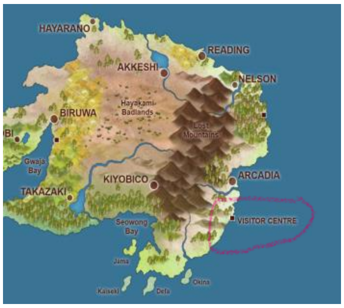
mean() – calculates the simple average

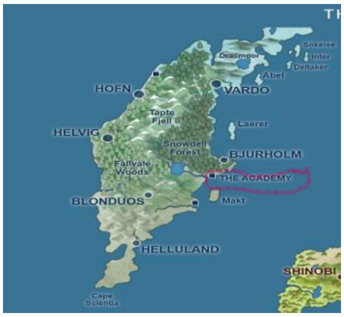
sd() – calculates the standard deviation

lm(Dependent-variable ~ Independent_variable) – Simple linear regression

Introduction to The Islands

- “The Islands” link is located under Learning Resources” on BB.
- The Islands is a population of virtual human subjects.
- You will use this population for quizzes and the research project.
- VISITOR CENTRE (south of Arcadia) contains useful information and guidelines how to access this population.
- THE ACADEMY (south of Bjurholm) provides studies based on the population and their inhabitants. This will be useful for your research projects.





Data Analysis

Important : Read the LearnX and complete questions.

Why do we need data analysis?

Data is a universal language.

We need to extract value from data.

We need to sort out what is important and what is not.

We need to analyse scientific research data to support our conclusions.

Data analysis is a systematic process of utilising data to address research questions. It involves Data collection, Data processing and Data modelling.

Example – Suppose we measure the heights of STAT1201 students in Summer Semester 2022. What can we do with those height measures?

9

Sources of variability

Would you find all the heights you measured are the same?

If no, why?

Sources of variability take two forms.

Natural variability – students heights are different to each other

Measurement variability – due to measurement errors

10

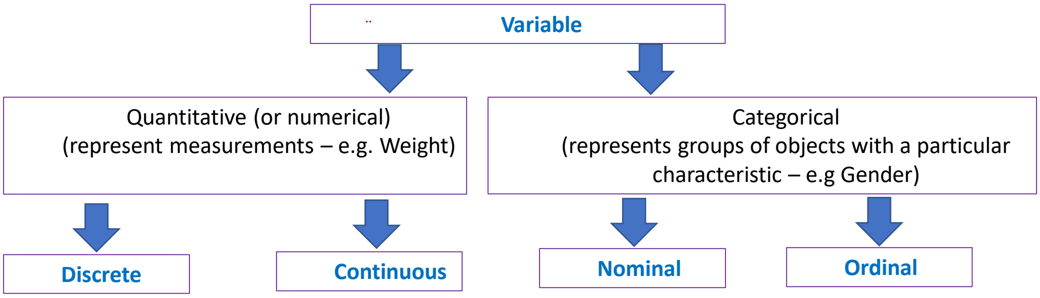
Different types of variables

Data – observations and measurements which have been collected in some way (often through research). Data are usually organised by variables and observational units.

Variable – a characteristic or an attribute that you are observing, measuring and recording data.

Examples – Height, Weight, Eye colour, Blood pressure, Age, Sex

Different types of variables


Variable
Quantitative (or numerical)
(represent measurements – e.g. Weight)
Categorical
(represents groups of objects with a particular characteristic – e.g Gender)
Discrete
Continuous
Nominal
Ordinal

Continuous variable - variable that is measurable, can have any value over some range, includes numerical values with decimal places and can be counting numbers

e.g: Height – 5.2”, 5.55”, 5”, 6”, 5.11”, 5.9” …

Discrete variable – variable that can have only whole counting numbers such as 0, 1 2, 45, 907 and so on

e.g: Number of phone calls – 1, 5, 8, 2, 15 …

Nominal variable – the groups or categories do not have an order.

e.g. marital status – Never married, married, divorced

Ordinal variable – categories or groups have an order.

e.g. Grades for STAT1201 – HD, D, C, P, F; Satisfaction level

Different types of variables

Poll Question 1

The variables country of birth and body mass are examples of

a) Categorical (ordinal) and categorical (nominal) variables respectively
b) Categorical (ordinal) and quantitative (discrete) variables respectively
c) Categorical (nominal) and quantitative (continuous) variables respectively
d) Categorical (ordinal) and quantitative (continuous) variables respectively

Poll Question 2

The variables Age (in whole years) is an example of a

a) Categorical (ordinal) variable
b) Quantitative (discrete) variable
c) Quantitative (continuous) variable
d) Categorical (ordinal) variable

13

Observational and Experimental studies *(Module 10 will provide more information)*

Observational Study	Experimental Study
The researcher observes part of population and measures the characteristics of interest. Make conclusions based on the observations but does not influence to change the existing conditions or does not try to affect them.	The researcher assigns subjects to groups and apply some treatment(s) to group(s) and the other group does not receive the treatment. Can be designed as a blind or a double-blind study. When an experiment involves both comparison and randomization then we call it as a randomized comparative experiment.
Example: Examine the effect of smoking on lung cancer	Example: Examine the effect of caffeinated drinks on blood pressure

14

Observational and Experimental studies

Poll Question 3

A researcher took a random sample of STAT1201 Summer semester 2022 students and examined their social media usage to determine their happiness. Each student was classified as either heavy, moderate or light social media user. Which type of study method is this?

a) Observational study

b) Experimental study

c) Randomised comparative blind study

d) Randomised comparative double-blind study

Observational and Experimental studies

Discussion Question

Use the Ed Discussion on BB to exchange your views to the following question.

Suppose that you design an experiment to examine the effect of caffeine on blood pressure among middle-aged adults (36 – 55 years). You randomly chose 20 adults in this age group (10 males and 10 females). The caffeinated drink is given to males and decaffeinated drink is given to females. Blood pressure is measured before and after the drink.

What are the problems with this comparative experiment design? How can you overcome the identified problems?

The Language of Hypothesis Testing

Hypothesis test is used to make conclusions about the population using sample data.

Suppose that you design an experimental study to test whether caffeinated drinks increase pulse rate among young adults. You randomly use 30 students aged between 20 and 25 and divide them into 2 groups. One group is receiving caffeinated drink (250ml) and the other group is receiving decaffeinated drink (250ml). The participants don't know which drink they are receiving. Pulse rate is measured and recorded before and after the drink and calculate the change. [This is an example of a bling randomised comparative experiment]. The summary results are as follows.

Group	***Mean increase in pulse rate***
Caffeinated drinks	14.8 beats per minute
Decaffeinated drinks	4.1 beats per minute

Thus, the difference = 10.7 beats per minute

Question – **Can you conclude that caffeinated drinks increase pulse rates for all young adults based on this experiment data?**

To answer this question, we need to do a hypothesis test.

17

The Language of Hypothesis Testing

Two types of Hypotheses.

Null hypothesis (H_0)

Usually a statement of "no effect". Also, refer to the **status quo** (no change from the past, the old standard still correct).

Either reject or do not reject H_0

In our caffeinated drink example, the null hypothesis is as follows.

H_0: the population mean increase in pulse rate is the same for caffeinated and decaffeinated drinkers among young adults (or caffeinated drinks has **no effect** on pulse rate among young adults)

Alternative hypothesis (H_1)

Usually a statement of "an effect". Also refers challenges to the status quo (something new is now occurring compared to the past).

If we reject H_0 we conclude there is sufficient evidence to accept the alternative hypothesis.

In our caffeinated drink example, the alternative hypothesis is as follows.

H_1: the population mean increase in pulse rate is higher for caffeinated drinkers among young adults (or caffeinated drinks **increase** the pulse rate among young adults)

18

The Language of Hypothesis Testing

Back to our caffeinated drinks example – The observed mean increase of 10.7 beats per minute can be explained in two ways.

1) The increase could be due to sampling errors or anxiety of participants or natural increase in pulse rate from drinking or combination of these.
2) The observed increase is because caffeine does increase pulse rate.

Question - which explanation is correct?

We use the concept of *p*-value to answer this question. If we reject the null hypothesis explanation 2) is correct. If we do not reject the null hypothesis, we do not have evidence to accept explanation 2) and explanation 1) could be correct.

(Note: How to calculate the p-value will be discussed in later modules)

19

The Language of Hypothesis Testing

The concept of *p*-value

- We use the concept of *p*-value to reject or do not reject the null hypothesis.
- This *p*-value is always reported in scientific papers that use hypothesis testing.
- *p*-value is mostly denoted by *p*.

- If *p*-value is small, we reject the null hypothesis and conclude that we have evidence to accept the alternative hypothesis.
- If *p*-value is large, we do not reject the null hypothesis and conclude that we do not have evidence to accept the alternative hypothesis.

20

The Language of Hypothesis Testing

The concept of *p*-value

The strength of evidence against the null hypothesis is determined by the magnitude of the *p*-value.

$p < 0.01$	–	strong evidence against H_0
$0.01 \le p < 0.05$	–	moderate evidence against H_0
$0.05 \le p < 0.1$	–	weak evidence against H_0
$p \ge 0.1$	–	no evidence against H_0

The commonly used threshold is 0.05. If we find $p < 0.05$, then we say that the results are significant at 5% level of significance. You will see in scientific journal articles “the results were found to be significant ($p < 0.05$)”.

21

The Language of Hypothesis Testing

Poll Question 4

In our caffeinated drink experiment the *p* value was 0.04. You can conclude that there is

a) strong evidence to suggest that caffeinated drinks increase the mean pulse rate of young adults.

b) moderate evidence to suggest that caffeinated drinks increase the mean pulse rate of young adults.

c) weak evidence to suggest that caffeinated drinks increase the mean pulse rate of young adults.

d) no evidence to suggest that caffeinated drinks increase the mean pulse rate of young adults.

22

Reminders

Quiz 1

- Open now.
- Close – Monday, 05 December 2022 at 3:00 pm

Lecture 2 – Module 2: Exploratory Data Analysis
Thursday, 01 December 2022 at 12:00 via Zoom (818 1453 7986)

23

Thank you!

24