STU11002

Statistical Analysis I

Dr. Hannah Comiskey

Learning objectives

- Download R and Rstudio
- Create a project in Rstudio and save it.
- Write and execute a basic R script

Structure

- **Lectures** will take place during weeks 22 to 27 and 29 to 33. Wednesdays' lectures will be held in room CHLLT_0.11 (Chemistry Building), from 10am to 11am, while Thursdays' lectures will take place in room 2039 (Arts Building), from 3pm to 4pm.
- **Labs** will take place every two weeks starting in week 23 (Group A) and in week 24 (Group B):
  - Group A: labs will take place in weeks 23, 25, 27, 30.
  - Group B: labs will take place in weeks 24, 26, 29, 31.

  Check your timetable to see your time slots.

Assessment

- **Continuous assessment** 30% of your final grade for STU11002 will depend on continuous assessment. This will consist of two MCQ tests, which will take place in week 27 and in week 33.
- **Final exam** 70% of your final grade for STU11002 will depend on on a written final exam.

CONTACT comiskeh@tcd.ie

What is statistics?

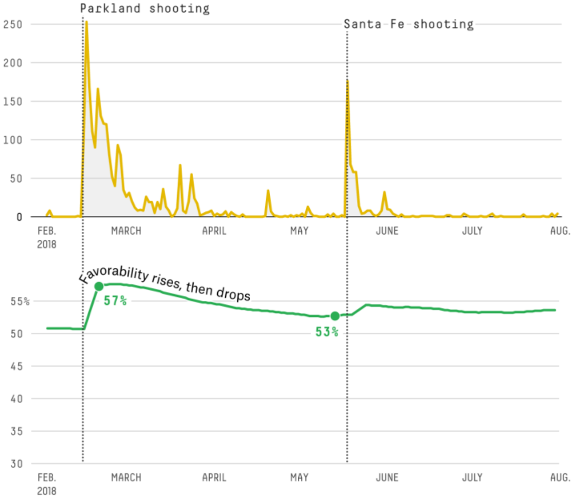
Broadly speaking, Statistics is a discipline that concerns the study of data. Its main areas revolve around:

- Data collection (how do we properly draw a sample from a population of interest?)
- Summarizing and representing the information in data (what types of measures and/or visualization methods suit best the task at hand?)
- Analyse and model the data (how can we postulate and answer research questions using data?)
- Interpretation (of any of the above)

Exploring possible relationships

After school shootings, support for gun laws rises but then drops

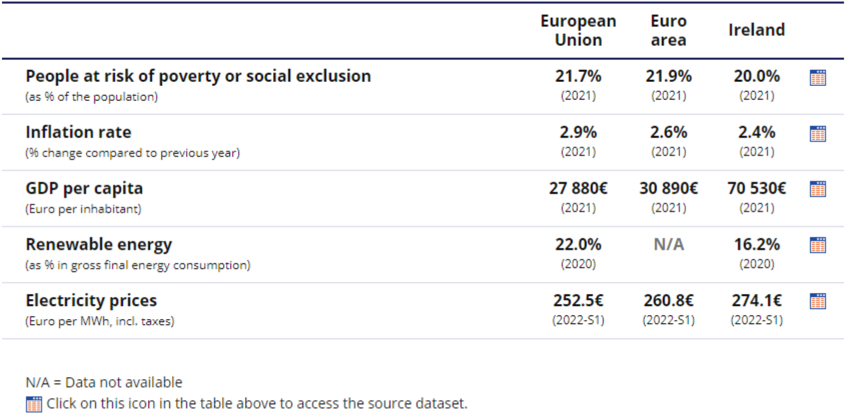
Number of 15-second cable news clips (CNN, Fox News, MSNBC) in which “school shooting” was mentioned, and overall favorability of stricter gun control laws



FiveThirtyEight

SOURCES: TV NEWS ARCHIVES, CIVIQS

Summarizing complex phenomena with indexes



	European Union	Euro area	Ireland	
People at risk of poverty or social exclusion (as % of the population)	**21.7%** (2021)	**21.9%** (2021)	**20.0%** (2021)	▦
Inflation rate (% change compared to previous year)	**2.9%** (2021)	**2.6%** (2021)	**2.4%** (2021)	▦
GDP per capita (Euro per inhabitant)	**27 880€** (2021)	**30 890€** (2021)	**70 530€** (2021)	▦
Renewable energy (as % in gross final energy consumption)	**22.0%** (2020)	N/A	**16.2%** (2020)	▦
Electricity prices (Euro per MWh, incl. taxes)	**252.5€** (2022-S1)	**260.8€** (2022-S1)	**274.1€** (2022-S1)	▦

N/A = Data not available

▦ Click on this icon in the table above to access the source dataset.

Source : `https://ec.europa.eu/eurostat`

Getting the probabilities right

Independence matters

- Sally Clark's two infant sons both died of SIDS (cot death), one in 1996, the second in 1998.
- It was estimated that the probability of SIDS in an affluent family, with non-smoking parents, and a mother with over 26 years of age is approximately 0.000117
- SIDS deaths were considered independent, so that the probability of two of them happening in the same family, with the aforementioned conditions, is $(0.000117)^2 \approx 0.00000001$. Too unlikely!
- In November 1999 Sally Clark was convicted and sentenced to life in prison.
- In January 2003 the conviction was overturned on appeal.

Sally Clark's story - details

Evidence for first conviction:

- Sir Roy Meadow, a pediatrician, stated that:
  - the chance of two children from an affluent family suffering SIDS was 1 in 73 million.
  - "one sudden infant death in a family is a tragedy, two is suspicious and three is murder unless proven otherwise"
- Data used:
  - The Clarks were an affluent, non-smoking family
  - The probability of a single cot death was 1 in 8543. Two SIDS were assumed to be independent so that the probability of them occurring in the same family is around 1 in 73 million (8543×8543).
  - Every year in Britain there were approximately 700,000 live births
  - Therefore, a double cot death was expected to occur once every hundred years or so

Sally Clark's story - details

Evidence for appeal:

- Professor of Mathematics Ray Hill stated that
  - "There may well be unknown genetic or environmental factors that predispose families to SIDS"
  - The probability of a child dying from SIDS is 1 in 1300
  - The 1 in 8500 figure takes into account three additional characteristics
  - "conveniently ignored factors such as both the Clark babies being boys – which make cot death more likely"
  - "if the parents are affluent, in a stable relationship and non-smoking, the prosecution will claim that the chances of the death being natural are greatly reduced ... the very same factors which make a family low risk for cot death also make it low risk for murder"

Cherry-picking and spurious correlations

- Does pork give you cancer? `https://fivethirtyeight.com/features/you-cant-trust-what-you-read-about-nutrition/`

- Deaths by Swimming Pool Drowning vs. Nicholas Cage Films `https://www.wnycstudios.org/podcasts/otm/articles/spurious-correlations`

Selection bias

Survival bias

- During World War II, researchers from the Center for Naval Analyses conducted a study of the damage done to aircraft that had returned from missions.
- The researchers recommended that armor be added to the areas that showed the most damage.
- Statistician Abraham Wald: the study only considered the aircraft that had survived their missions
- The holes in the returning aircraft, then, represented areas where a bomber could take damage and still return home safely.

Machine bias

- In the USA, 'COMPAS' is a computer program that predicts the score/likelihood of arrested individuals committing a future crime
- Scores derived from 137 questions (race not included)
- Falsely flags black defendants as future criminals at almost twice the rate as white defendants. White defendants mislabeled as low risk more often than black defendants
- Difficult to construct a score that doesn't include items that can be correlated with race (poverty, social marginalization)
- No transparency (code is not public)

`https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`

Available information and how to use it

Butcher shop example

- Installed sensors outside shop and determined footfall
- Count: How many people stopped to look at window and at sandwich board. Determined that lots of passers by after pubs closed
- Decided to open at that time

Transport for London (TFL)

- Early 2000's, London had a population of around 7 million, which was expected to grow to 10 million
- TFL had two priorities: Planning services and providing information to customers
- In 2003 the Oyster card was introduced (with around 19 million taps a day). It provides TFL with info on when and where people are travelling

R

Reading material

There is a lot of excellent reading material free online for R. But it is easy to get overwhelmed! I've uploaded two books to blackboard that are freely available through CRAN:

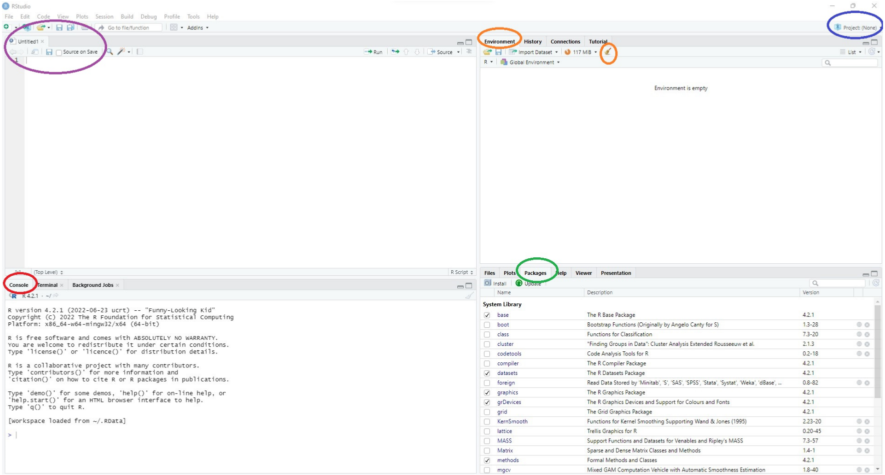
- R For Beginners by Emmanuel Paradis
- An Introduction to R

Getting started - Download R

- Have you downloaded R? No?
- R is available from `https://cran.r-project.org/`
- Most of you will want to click either
  - Download R for Windows
  - Download R for (Mac) OS X
- **Windows :** click on base, then at top of page ‘Download R-4.2.2 for Windows’
- **Mac :** click on the appropriate ‘.pkg’ file for your version of OS X

Getting started - Download R Studio

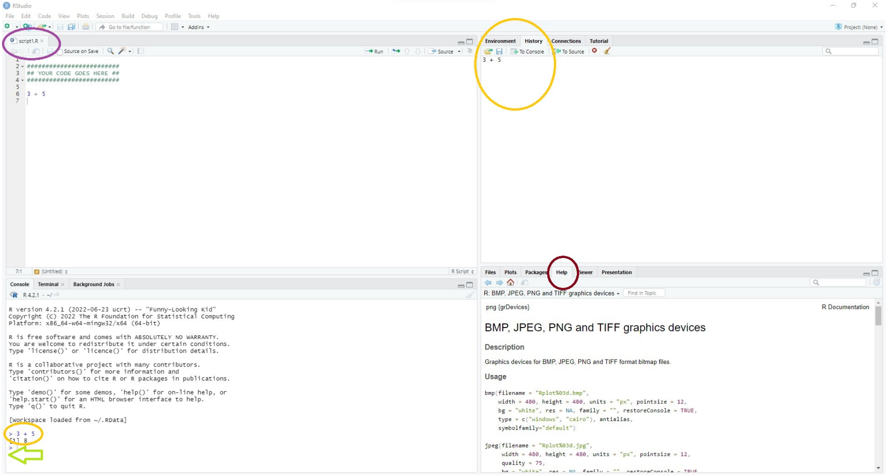
- Have you downloaded R Studio?
- Note that you must first install R before trying to install R Studio... it won't work otherwise
- You can get it from `https://www.rstudio.com/`, following the links to download (top-right of the page).
- Download 'RStudio Desktop - Open Source License - Free'
- When you have these installed, try to start up R Studio



```
R version 4.2.1 (2022-06-23 ucrt) -- "Funny-Looking Kid"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

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

[Workspace loaded from ~/.RData]

> |
```

System Library

	Name	Description	Version
☑	base	The R Base Package	4.2.1
□	boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-28
□	class	Functions for Classification	7.3-20
□	cluster	"Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al.	2.1.3
□	codetools	Code Analysis Tools for R	0.2-18
□	compiler	The R Compiler Package	4.2.1
☑	datasets	The R Datasets Package	4.2.1
□	foreign	Read Data Stored by 'Minitab', 'S', 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', 'dBase', ...	0.8-82
☑	graphics	The R Graphics Package	4.2.1
☑	grDevices	The R Graphics Devices and Support for Colours and Fonts	4.2.1
□	grid	The Grid Graphics Package	4.2.1
□	KernSmooth	Functions for Kernel Smoothing Supporting Wand & Jones (1995)	2.23-20
□	lattice	Trellis Graphics for R	0.20-45
□	MASS	Support Functions and Datasets for Venables and Ripley's MASS	7.3-57
□	Matrix	Sparse and Dense Matrix Classes and Methods	1.4-1
☑	methods	Formal Methods and Classes	4.2.1
□	mgcv	Mixed GAM Computation Vehicle with Automatic Smoothness Estimation	1.8-40



```
##########################
## YOUR CODE GOES HERE ##
##########################

3 + 5
```

```
3 + 5
```

```
R version 4.2.1 (2022-06-23 ucrt) -- "Funny-Looking Kid"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

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

[workspace loaded from ~/.RData]

> 3 + 5
[1] 8
>
```

png {grDevices}

R Documentation

BMP, JPEG, PNG and TIFF graphics devices

Description

Graphics devices for BMP, JPEG, PNG and TIFF format bitmap files.

Usage

```
bmp(filename = "Rplot%03d.bmp",
    width = 480, height = 480, units = "px", pointsize = 12,
    bg = "white", res = NA, family = "", restoreConsole = TRUE,
    type = c("windows", "cairo"), antialias,
    symbolfamily="default")

jpeg(filename = "Rplot%03d.jpg",
     width = 480, height = 480, units = "px", pointsize = 12,
     quality = 75,
     bg = "white", res = NA, family = "", restoreConsole = TRUE
```

Using R Studio

The top left panel is the **Editor**:

- This is where we write and edit code before running it
- This allows us to save it easily
- If you open a dataset, it will appear in a tab in this panel

The bottom left panel is the **Console**:

- This is where we run our code
- We'll also be able to access the results of our code in the console

Using R Studio

The top right panel has two tabs of note:

- **Environment**: lists all of the objects (datasets, vectors. . .) that you are working with in the console in this R session (since R Studio was started up)
- **History**: lists the things that were last sent to the console and run

Bottom right panel has six tabs:

- **Files** the file system on your computer
- **Plots** where plots will appear when created
- **Packages** The packages you have installed with ticks for inclusion in the session
- **Help**
- **Viewer** and **Presentation**

Creating a Project

- Projects are a “‘“neat way” to work in RStudio
- All files needed for a specific project/ analysis can be stored in the corresponding project folder, allowing to bypass the directory setting step
- Clicking on the icon of the ‘R Project’ file we can quickly access the workspace and the files (in RStudio) for a specific project/ analysis.

Create a project

On the top right panel, click on “Project: (None)”:

1. Click on ‘New Project...’
2. Now click on ‘New Directory’ and then on ‘New Project’
3. Specify a name (for example ‘project1’) under ‘Directory name’ and select where to store your project using “Browse...’

Using the editor

- Let's try to write a script for R in R studio and run it
- Go to the editor and type the following three lines of code (personalized to you)

```
# Hello
name <- "Hannah Comiskey"
cat("\n Hi",name,"welcome to R and R Studio! \n")
```

- `name` is a variable that is set equal to the value `Hannah Comiskey` .
- The name given to the variable is not important- we could call it anything...

Using the editor

... so for example, writing the code like this will do exactly the same thing

```
# Hello
x <- "Hannah Comiskey"
cat("\n Hi",x,"welcome to R and R Studio! \n")
```

- We can write a comment by using # which means the line will be ignored when running
- Congratulations! You've just written an R script!

Running a script

- Organising your files and folders and knowing where your data is/your scripts are is important when working with R
- Create a project named `Rcourse`
- Save the script we just created as `script1.R`. This script will be automatically stored in the folder of the `Rcourse` project.
- When you are working with datasets and scripts together, it is important to know what is where– it can save a lot of time. You should save all files related to the same project within the same R project folder.