NZSSN Courses: Introduction to R

Session 1 – Introduction

Statistical Consulting Centre

consulting@stat.auckland.ac.nz
The Department of Statistics
The University of Auckland

1 March, 2017

SCIENCE
DEPARTMENT OF STATISTICS

Wednesday

Each session comprises two parts: lecture and practice.

Session	Time	Session
1	09:00am - 10:30am	Introduction
	10:30am - 10:50am	Break
2	10:50am - 01:00pm	Subsetting data
	01:00pm - 02:00pm	Lunch break
3	02:00pm - 03:00pm	Data manipulation
	03:00pm - 03:20pm	Break
4	03:20pm - 04:30pm	Data exploration

Thursday

Session	Time	Session
1	09:00am - 10:30am	Graphics
	10:30am - 10:50am	Break
2	10:50am - 12:30pm	Advanced Graphics
	12:30pm - 01:30pm	Lunch break
3	01:30pm - 03:00pm	Simple analysis
	03:00pm - 03:20pm	Break
4	03:20pm - 04:30pm	Advanced analysis

R and UoA's Department of Statistics

- R was initially written by Robert Gentleman and Ross Ihaka – *R & R* – of the **Department of Statistics, University of Auckland**.
- Three members of the *R Development Core Team* are in UoA's Department of Statistics.

SCIENCE
DEPARTMENT OF STATISTICS

R and UoA's Department of Statistics

Ross Ihaka

R and UoA's Department of Statistics

Robert Gentleman (no longer in our department)

R and UoA's Department of Statistics

Paul Murrell

R and UoA's Department of Statistics

Thomas Lumley

R and UoA's Department of Statistics

What does this mean?

If you want to learn R, you are talking to the right people!

Chris Triggs
Director Consulting Services
Phone: +64 9 373 7599 ext 88856
Email: triggs@stat.auckland.ac.nz

For more information, please see Chris's profile.

Yannan Jiang
Senior Research Fellow
Phone: +64 9 373 7599 ext 84725
Email: y.jiang@auckland.ac.nz

For more information, please see Yannan's profile.

Kathy Ruggiero
Senior Lecturer
Phone: +64 9 373 7599 ext 89456
Email: k.ruggiero@auckland.ac.nz

For more information, please see Kathy's profile.

Jessica McLay
Research Fellow
Phone: +64 9 373 7599 ext 73678 or 85313
Email: jessica.mclay@auckland.ac.nz

For more information, please see Jessica's profile.

Rachel Chen
Research Fellow
Phone: +64 9 373 7599 ext 89384
Email: rachel.chen@auckland.ac.nz

For more information, please see Rachel's profile.

Avinesh Pillai
Research Fellow
Phone: +64 9 373 7599 ext 82368 (Mon-Wed) or ext 81169 (Thurs & Fri)
Email: a.pillai@auckland.ac.nz

For more information, please see Avinesh's profile.

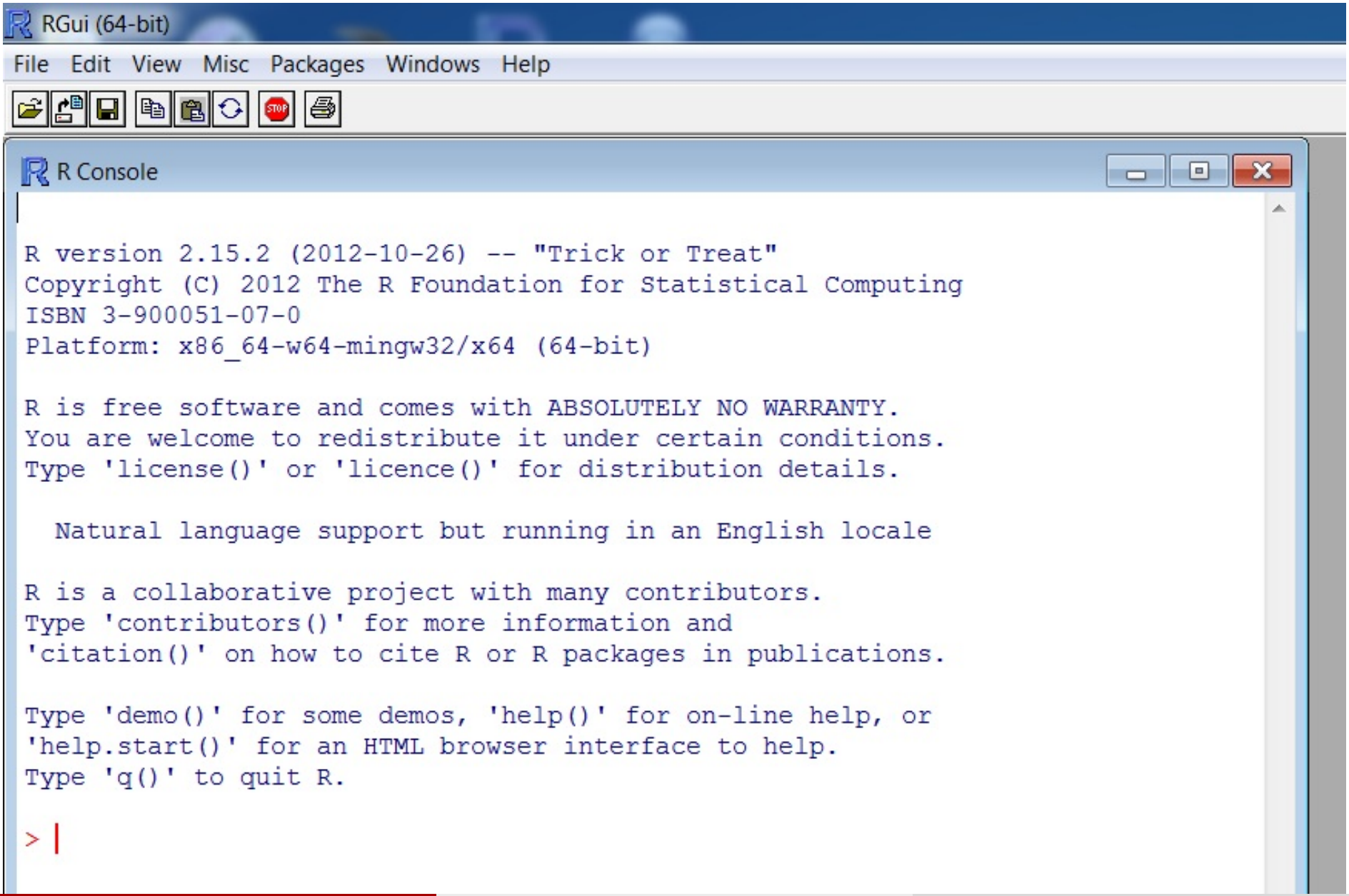
What is R?

"R is a free software environment for statistical computing and graphics"

Key words:

- FREE!!!!!
- Statistical computing
- Graphics (much more flexible than SAS, SPSS, JMP, etc.)
- Support from communities of different fields, i.e. R packages. https://cran.r-project.org/web/views/.
- Even Microsoft is in it: Microsoft R Open. https://mran.microsoft.com/open/.

The R Graphical User Interface (GUI)



How to download and install R

1. Go to the CRAN (Comprehensive R Archive Network) `cran.stat.auckland.ac.nz`.
2. Download the relevant version for Linux/Mac/Windows.
   - We will only look at R in the Windows environment today.
3. Install it on your computer (for Windows only):
   - Choose "Yes (customized startup)" in Startup options.
   - Choose "SDI (separate windows)" in Display mode.
   - Choose "HTML help" in Help .

Using the R editor

- The R GUI is not menu driven.
- Commands can be typed at the console.
  - OK for simple calculations requiring few lines of code
  - Painful for anything more!
- We *strongly* recommend using an R editor
  - Great for reproducible analyses and research!!
  - Best editor for you depends on whether you are a(n)...
    1. Beginner: Built-in R editor,
    2. Advanced user: Rstudio, Tinn-R, Notepad++, and many others.
    3. R geek: Emacs

Using R as a calculator

```
1+2
```

```
[1] 3
```

```
1 + 3^2
```

```
[1] 10
```

```
log(15) - sqrt(3.4)
```

```
[1] 0.8641413
```

```
pnorm(1.96)
```

```
[1] 0.9750021
```

Using R as a calculator

- "<-" is the "assign to" operator, made up of "<" and "-" without a space.
- E.g., `x <- 2` is read as "The value 2 is assigned to the object `x`".

```
x <- 2
y <- 3
x^2 - 3*y + 5
```

```
[1] 0
```

- <- has a direction, from right to left, `x <- 2` means assigning 2 to x,

Using R as a calculator

- -> operates from left to right, assigning x to 2.
  2 is a real value so you can not do that.

```
x -> 2
```

```
Error in 2 <- x:  invalid (do_set) left-hand side to
assignment
```

- = has no direction and can be confusing sometimes.
- It is good programming practice to use <-.

Getting help

- Google!!!!
  e.g. How to calculate the mean in `R`? The search results tell you that the function `mean()` would be helpful.
- Quick-R: `http://www.statmethods.net/`
- R-bloggers: `https://www.r-bloggers.com/`

Getting help

- `?`
  e.g. `?mean` brings up the help file for this function. It will tell you (almost) everything you need to know to use `mean()`.
- `??`
  e.g. `??mean` searches for everything related to mean in your computer.
- `RSiteSearch(" ")`
  Searches everything on CRAN as well as your computer.

Data, files, statisticians and R

- Statisticians prefer (read: **want**) rectangular data files
  - Each case in its own row
  - Data collected on each variable in its own column
  - Variable names in the first row of each column
  - No blanks, e.g. fill with NA, *, 99999, anything but a blank!
- R likes (read: **needs**) this too!
- R prefers to read data files in Comma Separated Value (CSV) format.
- This does not mean R only reads files stored in csv format.

Getting data into R

Try your best to save your data in a `csv` or `txt` format.

- Most datasets are saved in an Excel spreadsheet.
- Do as much data cleaning as you can in Excel. No comments, no formatting, no colours, no fancy fonts.
- Convert it into `csv` by clicking on `Save As`. Change the `Save as type` from `xlsx` or `xls` into `CSV (Comma Delimited)`.
- `CSV` can have one worksheet only. If you have multiple worksheets, it saves the active worksheet.

issp.df

- International Social Survey Programme (ISSP): 1994 - Family and Changing Gender Roles II (Modified)
- Question 1 to 4, choose from one of the following: Agree strongly, Agree, Neither agree nor disagree, Disagree, Disagree strongly, Can't choose.

1. Both the man and the woman should contribute to the household income.
2. A man's job is to earn money: a woman's job is to look after the home and family.
3. It is not good if the man stays at home and cares for the children and the woman goes out to work.
4. Family life often suffers because men concentrate too much on their work.

issp.df

5. Which of these would you say is more important in preparing children for life?
   to be obedient, to think for themselves, or Can't choose.

- Question 6 to 8, choose one of the following:
  Always wrong, Almost always wrong, Wrong only sometimes, Not wrong at all, Can't choose.

6. Do you think it is wrong or not wrong if a man and a woman have sexual relations before marriage?
7. What if they are in their early teens, say under 16 years old, in that case is it...
8. What about a married person having sexual relations with someone other than his or her husband or wife, it is...

Eight additional variables in `issp.df`

- ID: Identification number.
- Gender.
- Age.
- Marital Status.
- Education: Education level.
- Working hours per week: the average number of hours per week.
- Income: Individual annual income.
- Ethnicity

Read and Check

- Always set a working directory using `setwd()`, this can be a directory where you store the data and/or outputing the results.
- Use `read.csv` to read a CSV file into `R`.
- `dim()`: Returns the number of observations (rows) and variables (columns).
- `head()`/`tail()`: Returns the first/last few rows of a data set.
- `str()`: Returns the structure of the dataset, e.g., dimension, column names, type of data object, first few values of each variable.
- `names()`: Returns the names of the variables contained in a dataset.

Reading data into R

```
setwd("your working directory")
issp.df <- read.csv("issp.csv", stringsAsFactors = FALSE)
head(issp.df)
```

```
      ID                  Q1                       Q2
1 1900073            disagree                    agree
2 1900013 strongly disagree neither agree nor dis
3 1900025            disagree     strongly disagree
4 1900037   cant choose, dk                 disagree
5 1900043            disagree neither agree nor dis
6 1900061            disagree                 disagree
```

`stringsAsFactors` argument is set to `FALSE`, so **character** vectors are not converted to **factor**s. We will cover the factor at Session 3.

dim() and str()

```
dim(issp.df)
str(issp.df)
```

```
[1] 1047   16
'data.frame': 1047 obs. of  10 variables:
 $ ID    : int  1900073 1900013 1900025 1900037 1900043 1900061 190
 $ Q1    : chr  "disagree" "strongly disagree" "disagree" "cant cho
 $ Q2    : chr  "agree" "neither agree nor dis" "strongly disagree"
 $ Q3    : chr  "neither agree nor dis" "disagree" "strongly disagr
 $ Q4    : chr  "agree" "agree" "agree" "agree" ...
 $ Q5    : chr  "think themselves" "think themselves" "think themse
 $ Q6    : chr  "always wrong" "always wrong" "not wrong at all" "n
 $ Q7    : chr  "always wrong" "always wrong" "almost always wrong"
 $ Q8    : chr  "always wrong" "always wrong" "only sometimes wrong
 $ Gender: chr  "Female" "Male" "Female" "Female" ...
```

names(issp.df)

```
#Names of the variables
names(issp.df)

 [1] "ID"                          "Q1"
 [3] "Q2"                          "Q3"
 [5] "Q4"                          "Q5"
 [7] "Q6"                          "Q7"
 [9] "Q8"                          "Gender"
[11] "Age"                         "Marital.Status"
[13] "Education"                   "Working.hours.per.week"
[15] "Income"                      "Ethnicity"
```

- Anything following the # symbol is treated as a comment and ignored by R.
- Writing comments is a very good habit to develop!

Descriptive statistics

Calculate the mean of `Age`:

```
mean(Age)
```

```
Error in mean(Age):  object 'Age' not found
```

You must tell `R` that `Age` is a variable (column) *within* `issp.df`, i.e.

```
mean(issp.df$Age)
```

```
[1] NA
```

You must also tell `R` how to deal with missing values: remove them before calculating the mean, i.e.

```
mean(issp.df$Age, na.rm = TRUE)
```

```
[1] 45.77179
```

table of counts

```
# One-way table of counts

table(issp.df$Gender)
```

```
     Female        Male NA, refused
        607         418           22
```

table of proportions

```
# Total count
total <- sum(table(issp.df$Gender))
total

[1] 1047

# Proportions of total
table(issp.df$Gender)/total


     Female        Male NA, refused
 0.57975167  0.39923591  0.02101242
```

One-way tables with less typing

Tired of typing `issp.df$` over and over again? Use the `with` function.

```
gender.table <- with(issp.df, table(Gender))
gender.table

Gender
     Female        Male NA, refused
        607         418           22

total <- sum(gender.table)
gender.table/total

Gender
     Female        Male NA, refused
 0.57975167  0.39923591  0.02101242
```

One-way tables with less typing

```
#Convert to percentages
gender.pct <- 100*gender.table/total
gender.pct

Gender
     Female        Male NA, refused
  57.975167   39.923591    2.101242

# Round to 1 decimal place
round(gender.pct, 1)

Gender
     Female        Male NA, refused
       58.0        39.9          2.1
```

Two-way frequency tables

```
income.gender.tab <- with(issp.df, table(Income, Gender))
income.gender.tab
```

	Gender		
Income	Female	Male	NA, refused
$10000 or less	177	57	4
$10001-$15000	115	35	2
$15001-$20000	49	29	2
$20001-$25000	65	50	0
$25001-$30000	71	48	2
$30001-$40000	59	70	4
$40001-$50000	27	47	2
$50001-$70000	7	27	1
$70001-$100000	4	35	2
NAV; NAP No own income	33	20	3

Two-way frequency tables

```
# Calculate proportion with respect to 'margin' total
# margin = 1 (row total) or 2 (column total)
perc.income.gender <- prop.table(income.gender.tab, margin=2)
perc.income.gender
```

```
                          Gender
Income                          Female        Male
  $10000 or less          0.291598023 0.136363636
  $10001-$15000           0.189456343 0.083732057
  $15001-$20000           0.080724876 0.069377990
  $20001-$25000           0.107084020 0.119617225
  $25001-$30000           0.116968699 0.114832536
  $30001-$40000           0.097199341 0.167464115
  $40001-$50000           0.044481054 0.112440191
  $50001-$70000           0.011532125 0.064593301
  $70001-$100000          0.006589786 0.083732057
  NAV; NAP No own income  0.054365733 0.047846890
                          Gender
Income                     NA, refused
```

Two-way frequency tables

```
# Tabulate as percentages
round(100*perc.income.gender, 1)
```

```
                                Gender
Income                           Female Male NA, refused
  $10000 or less                   29.2 13.6         18.2
  $10001-$15000                    18.9  8.4          9.1
  $15001-$20000                     8.1  6.9          9.1
  $20001-$25000                    10.7 12.0          0.0
  $25001-$30000                    11.7 11.5          9.1
  $30001-$40000                     9.7 16.7         18.2
  $40001-$50000                     4.4 11.2          9.1
  $50001-$70000                     1.2  6.5          4.5
  $70001-$100000                    0.7  8.4          9.1
  NAV; NAP No own income            5.4  4.8         13.6
```

Summary

- Quick introduction to R
- Getting data into R
- Frequency tables