INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

TWO MAIN USES OF R IN STATISTICS PORTUGAL: SAMPLING AND CONFIDENTIALITY

Pedro Sousa
Inês Rodrigues, Maria Ferreira, Pedro Campos

Department of Methodology and Information System | Statistical Methods Unit

(2018)

Program

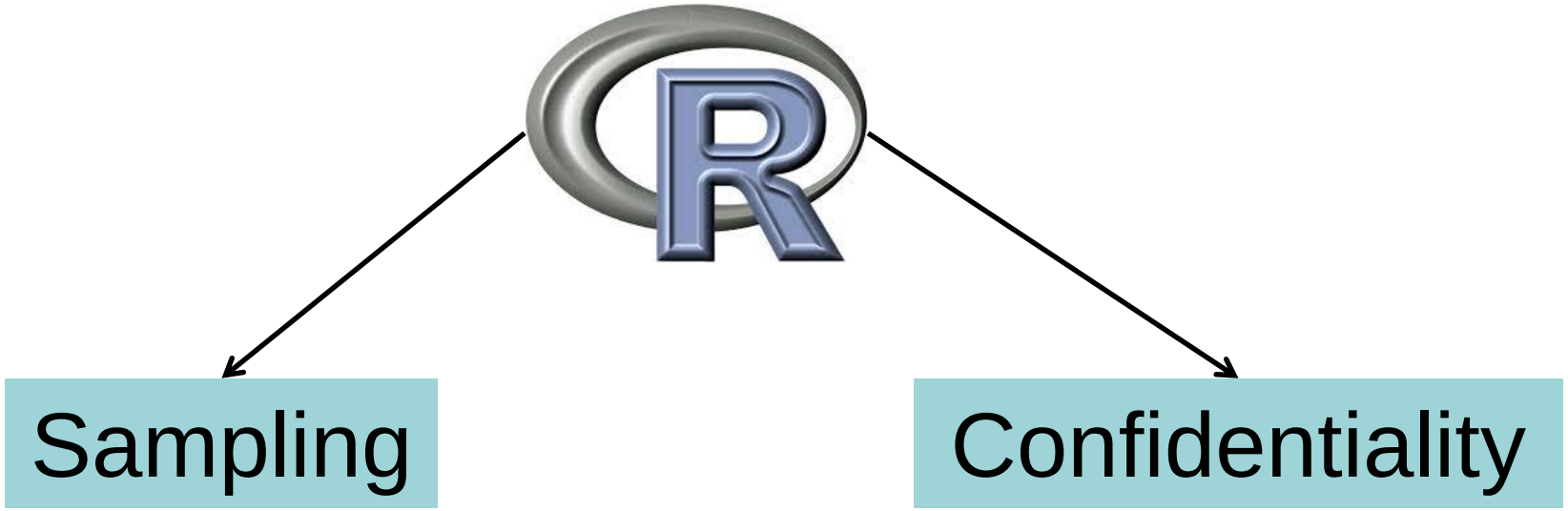
- Introduction
- Sampling with `survey` package
- Statistical disclosure control
- Other use cases of R

Introduction

Software used at Statistics Portugal:

Introduction

R at the Statistical Methods Unit



Sampling

Estimates on sampled designs

Sampling with `survey` package

How official statistics are produced?

- **Census**

Surveys where all individuals of the population must be observed.

- **Sample Surveys**

Surveys where only part (non-probabilistic or **probabilistic sample**) of the individuals are observed.

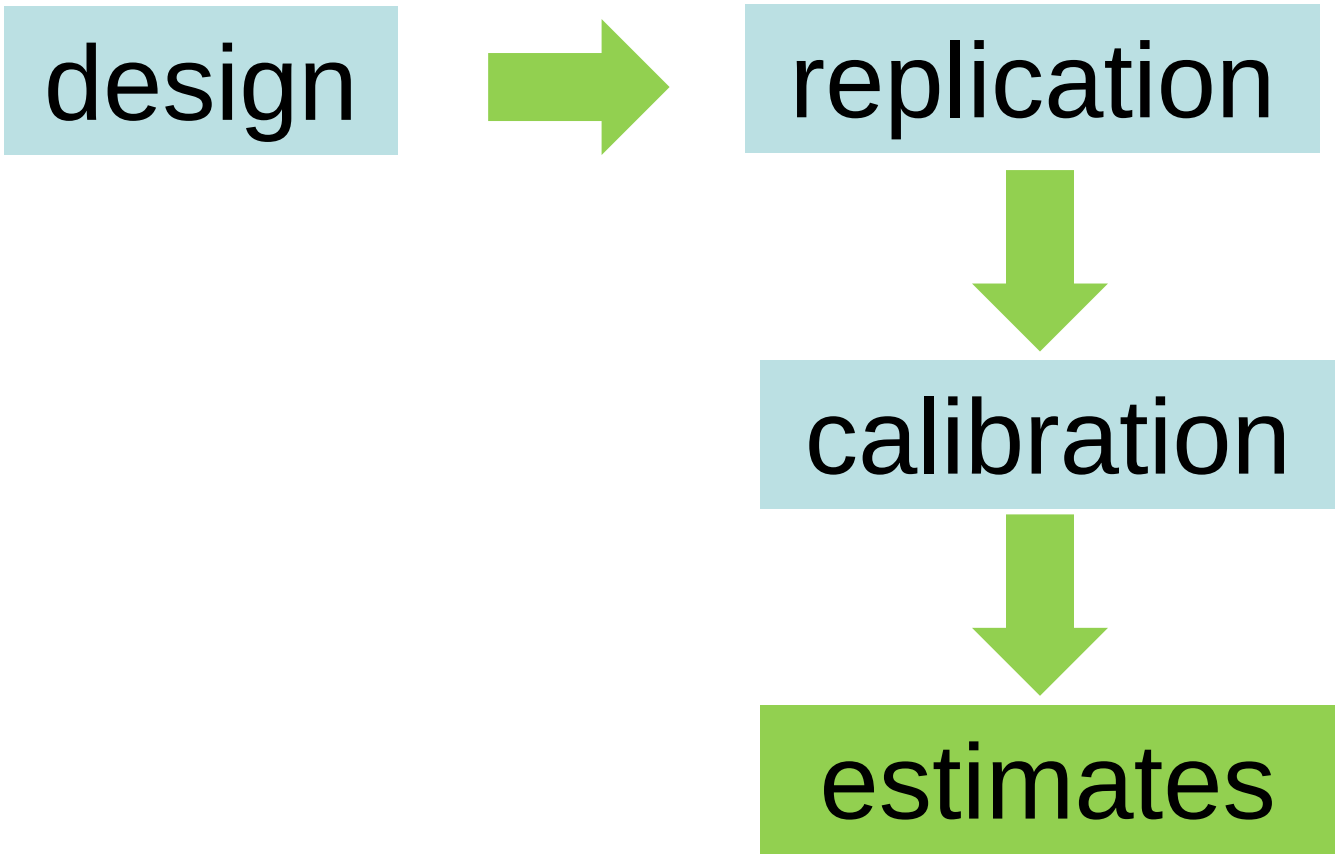
- **Administrative sources**

Data from administrative procedures is used for statistical purposes.

Sampling with `survey` package

LFS – Monthly unemployment rate estimate – 2014-…

Labor Market indicators



Sampling with `survey` package

Survey design: `svydesign()`

```
> desenho <- svydesign(id = ~AREA, weights = ~PESOIN, data = basef)
> desenho
1 - level Cluster Sampling design (with replacement) With (324) clusters.
svydesign(id = ~AREA, weights =~PESOIN, data = basef)
```

- AREA is the PSU (geographic area >300 households)
- PESOIN a variable containing the initial sampling weights
- basef is the dataframe that contains the actual data

Sampling with `survey` package

Replicate weights: `as.svrepdesign()`

```
> desenho_jk <- as.svrepdesign(desenho, type = "JK1")
```

• JK1 creates multiple subsamples using JackKnife method omit one PSU at a time

Sampling with `survey` package

Calibration: `calibrate()`.

```
> calibra_jk<-calibrate(desenho_jk,make.formula(c(E1,E2,E3)), est_pop, aggregate.index=~ALOJ, bounds=c(0.25,4), calfun="logit", epsilon=1e-9)

> calibra_jk

Call: calibrate(desenho_jk, make.formula(c(E1, E2, E3)), est_estr, aggregate.index=~ALOJ, bounds = c(0.25,4), calfun="logit", epsilon=1e-9) Unstratified cluster jacknife (JK1) with 324 replicates.
```

- Adjusting the weights according to the known population total margins for these disaggregation variables: E1 - NUTS2, sex and 5-years age groups; E2 - NUTS3 (or groups of NUTS3) by six age groups; E3 - NUTS3 (or groups of NUTS3) by sex
- **`logit`** – calibration method with range limits on the weights defined by the bounds
- **`est_pop`** – known population estimates
- **`ALOJ -`** households IDs

Sampling with `survey` package

Weights after calibration: `weights()`

```
> wk <- weights(calibra_jk, type = "sampling")
```

NUTS2	AREA	ALOJ	SEXO	Q_ETARIOS	G_ETARIOS	AGREG_N3	IDADEE	dk	wk	wk/dk
1	2018	2018_0147	1	1	1	1	2	298,953	538,055	1,800
1	2041	2041_0144	2	12	5	1	2	353,493	380,617	1,077
1	2318	2318_0492	1	3	1	5	2	279,081	236,067	0,846
2	2641	2641_0206	1	14	5	6	2	368,747	324,778	0,881
2	2473	2473_0538	1	8	3	7	2	354,955	720,957	2,031
2	2537	2537_0425	2	11	5	9	2	317,138	312,688	0,986
3	2745	2745_0591	1	16	6	13	2	392,416	302,046	0,770
3	2894	2894_1433	2	2	1	13	2	442,975	426,179	0,962
3	2894	2894_1433	2	10	4	13	2	442,975	426,179	0,962
3	2894	2894_1597	2	9	4	13	2	442,975	492,916	1,113
4	2918	2918_0039	1	1	1	14	2	141,125	223,426	1,583
4	2930	2930_0213	1	3	1	14	2	141,125	124,039	0,879
4	2919	2919_0516	2	14	5	14	2	111,226	61,093	0,549
5	2325	2325_0521	2	4	2	19	2	109,424	133,470	1,220
5	2325	2325_0141	2	7	3	19	2	109,424	198,410	1,813
6	3095	3095_0309	2	6	2	20	1	72,923	53,174	0,729
6	3109	3109_0017	1	1	1	20	2	80,346	75,876	0,944

Example of the Calibration process

Sampling with `survey` package

Analysis of the variables estimates : `svytotal()`

```
> svyby(POP_ACT, NUTS2 + SEXO, calibra_jk, svytotal, vartype = c("var", "cvpct"))
```

NUTS2	SEXO	POP_ACT	VAR	CV(%)
1	1	940006,27	64224866,07	0,85
1	2	887776,39	91145381,88	1,08
2	1	600436,45	51376995,54	1,19
2	2	548571,36	59188803,81	1,40
3	1	692965,68	44375526,61	0,96
3	2	723397,40	54725716,17	1,02
4	1	184209,38	4663305,22	1,17
4	2	163538,13	7626837,24	1,69
5	1	109588,99	1705448,76	1,19
5	2	110648,40	2211930,20	1,34
6	1	66573,32	1299742,85	1,71
6	2	55847,93	2377161,95	2,76
7	1	66909,97	1787817,69	2,00
7	2	66356,72	2409876,77	2,34

Variance and coefficient of variation of the total active population by NUTS2 and sex

Confidentiality

Statistical disclosure control (SDC)

Statistical disclosure control (SDC)

Access to confidential data for scientific purposes

- Research entity 
- Research proposal 

Prepare microdata file

Analyse / measure (re-)identification risk of statistical units:

- disclosure scenarios (cross-tabulations of key variables)

Apply SDC methods

SDC of microdata (R package `sdcMicro`)

Estimation of (re-)identification risk:

- `freqCalc()` - compute/estimate sample and population frequency counts
- `indivRisk()` - estimate the risk for each observation

```
> fre <- freqCalc(bd, keyVars = subset(ind[1:4], ind[1:4]!=0),w = ind[5])
> ind <- indivRisk(fre)
> max <- max(ind$rk)                  # maximum individual risk
> count <- sum(ind$rk > 0.04)         # n of records whose individual risk is
                                       above a given threshold
```

Statistical disclosure control (SDC)

SDC of microdata (R package `sdcMicro`)

Implementation of SDC methods(suppression, global or top/down recoding)

Example: **Microaggregation:** `microaggregation()`

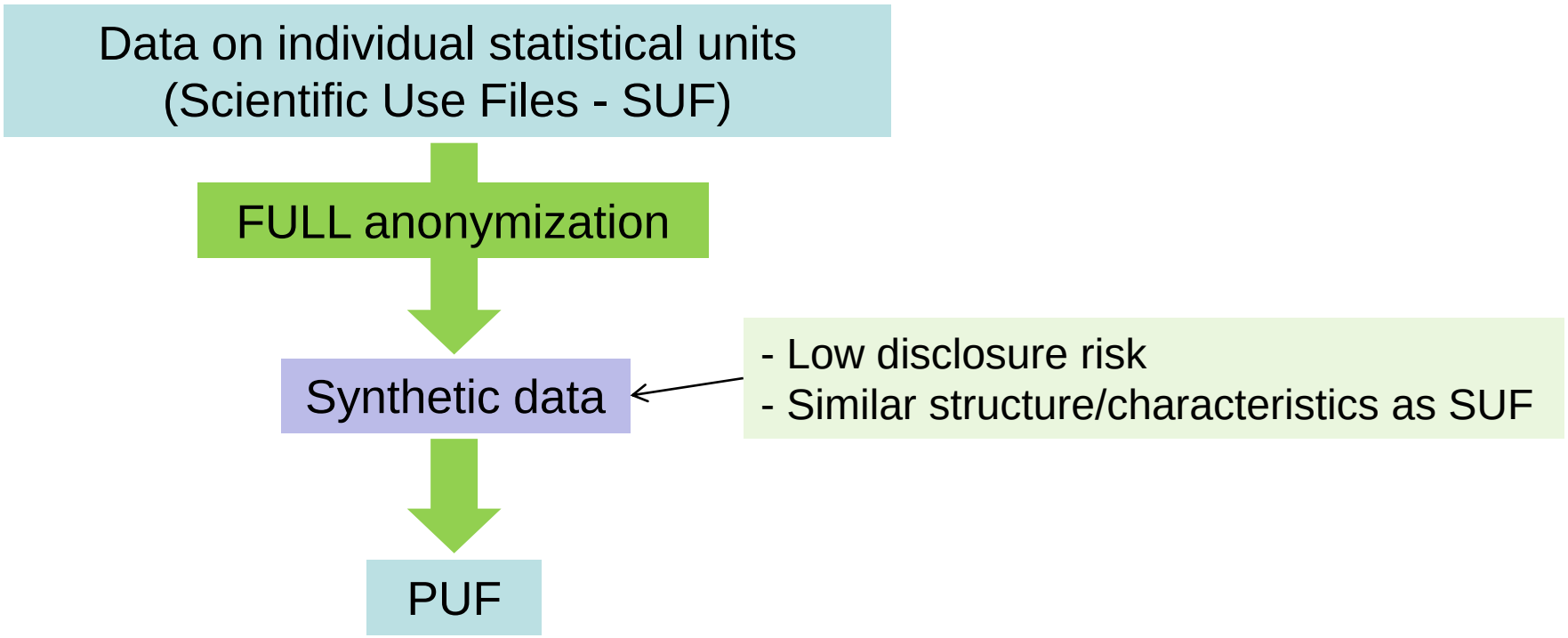
```
> bd <- data.frame(id = 1:8, v = round(rnorm(mean = 1000, sd =
150, 8), 0))

> ma_v <- microaggregation(bd[!is.na(bd$v),], variables =
c("v"), aggr = 3, method = "onedims", measure = "mean")
```

```
> ma_v$x   # Original values
id v
1 813
2 999
3 959
4 766
5 1112
6 914
7 1129
8 967
```


```
> ma_v$mx  # Microaggregated values
id v
1 831.0
2 1033.2
3 1033.2
4 831.0
5 1033.2
6 831.0
7 1033.2
8 1033.2
```

Statistical disclosure control (SDC)

Public Use Files (PUF) for the Household Budget Survey (HBS)



Statistical disclosure control (SDC)

Public Use Files (PUF) for the Household Budget Survey (HBS)

Synthetic data file as a PUF:

- generate the synthetic data based on its sample distribution

Draw from univariate or conditional (conditioned on a factor variable) sample distribution

For a set of main variables: keep main multivariate relationships

Parametric models

multinomial logistic and log-linear regressions

`nnet::multinom`, `MASS::polr`, `Hmisc::rMultinom`

Classification and Regression Trees -CART

`Rpart::`, `partykit`

- re-calibrate sample weights: `simPop::calibSample()`

Statistical disclosure control (SDC)

Public Use Files (PUF) for the Household Budget Survey (HBS)

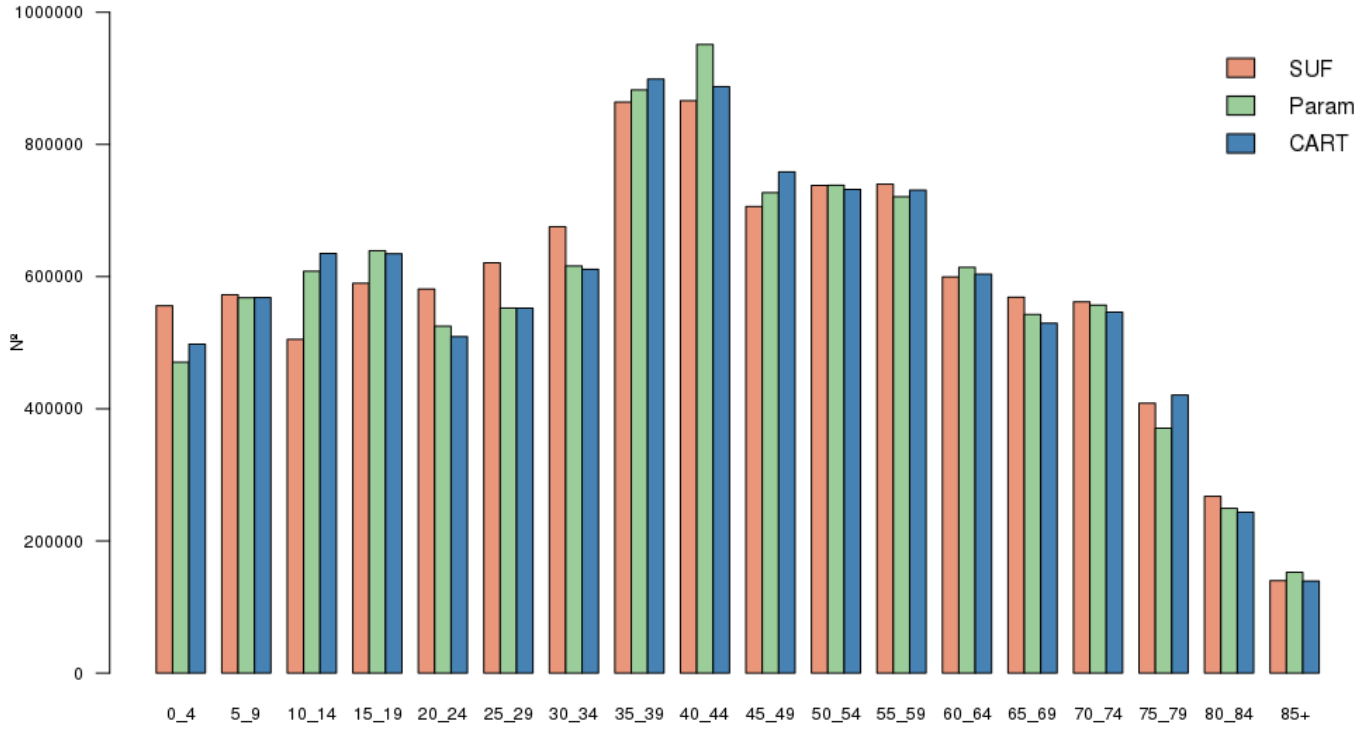
Main indicators – real and synthetic data (HBS 2010/2011)

	Median equivalised disposable income (€)	Mean equivalised disposable income (€)	At-risk-of-poverty threshold (€)	At-risk-of-poverty rate after social transfers (%)	Gini coefficient for equivalised disposable income (%)	Income quintile share ratio (S80/S20 (N.º)	Mean annual household total expense (€)
SUF (real)	11 000	13 750	6 600	14.8	33.2	5.2	20 391
Parametric	11 140	13 100	6 684	19.2	31.7	5.1	19 942
CART	10 800	13 279	6 480	15.5	32.6	5.1	19 661

Statistical disclosure control (SDC)

Public Use Files (PUF) for the Household Budget Survey (HBS)

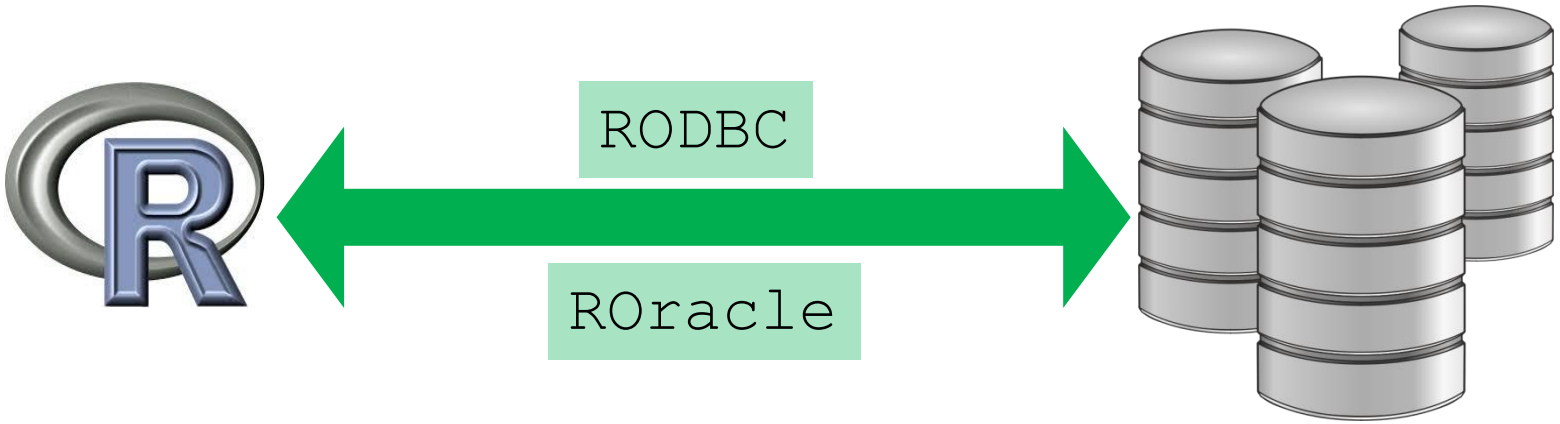
Age group distribution from real and synthetic data (HBS 2010/2011)



Other use cases of R

Other use cases of R

Data handling



Other use cases of R

Data handling

An RODBC example:

```
> library(RODBC)
> con<-odbcConnect("ORACLE-PROD",uid="user.id",pwd="passwd")
> result<- sqlQuery(con,"select * from UNIV_ACT Where YEAR=2018”)
> odbcClose(con)
```

An ROracle example:

```
> drv <- dbDriver("Oracle")

> connect.string<-"(DESCRIPTION=(ADDRESS=(PROTOCOL=tcp)(HOST= “ORACLE-PROD”)
(PORT=1521))(CONNECT_DATA=(SID=dw)))"

> con <- dbConnect(drv, username = "user.id", password = "passwd",
dbname=connect.string)

> result <- dbSendQuery(con, "select * from UNIV_ACT Where YEAR = 2018“)
```

Other use cases of R

Data handling

Handling with bigger datasets (`data.table` and `dplyr`)

- better performance
- more readable code
- better memory usage

```
> mtr <- tbl(con, "METERS")
> mtr_result <- mtr
%>% group_by(ID,DATA)
%>% summarise(CONST = sum(CONS,na.rm = T))
```

```
> mtr_result
# source: lazy query [?? x 3]
# Database: OraConnection
# Groups: ID
ID DATA CONST
<chr> <chr> <dbl>
1 1 2017-03-01 1428
2 1 2017-03-02 1476
3 1 2017-03-03 1428
4 1 2017-03-04 1060
5 1 2017-03-05 1068
6 1 2017-03-06 1728
7 1 2017-03-07 1744
8 1 2017-03-08 1664
9 1 2017-03-09 1476
10 1 2017-03-10 1668
# ... with more rows
```

Other use cases of R

Data handling

Handling with files from other software packages (`rio`)

```
> library(rio)
> data<-import(file="file.sas7bdat")
> export(data, file="output.sav")
> export(data, file="output.tmp", format="SPSS")
```

Other use cases of R

“r” R courses

Four days courses to Statistics Portugal’s collaborators.

First Level course:

- R essentials (R interface (the adopted IDE is RStudio))
- Some basic commands and functionalities
- Syntax rules and principal operators
- Working with arrays (sequence, index and order)
- Import and export CSV files into R
- Data Analysis (some basic descriptive statistics, graphics and statistical inference)

Other use cases of R

“r” R courses

On the **second level** of the R courses, some of the previous points are debated on a more advanced manner with additional focus on these three points:

• Database Access (connecting to databases using packages such a RODBC or ROracle)

• Data Visualization (Working with plot, ggplot2 package (Wickham, 2016) and some samples with Shiny Dashboard)

• Advanced data analysis (Statistical Inference, variance analysis, linear regression, decision trees and multivariate analysis)

» THANK YOU !

pedro.sousa
ines.rodrigues
maria.ferreira
pedro.campos
@ine.pt