ESTATÍSTICA con R

PARA LINGÜISTAS

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

INSTITUTO DA LINGUA GALEGA

R project

ESTATÍSTICA

Módulo I (Introdución a R)

I) Introdución
II) Axuda, paquetes e operacións aritméticas

Módulo II (A información en R)

I) Lectura de bases de datos
II) Obxectos e estrutura da información
III) Escritura de bases de datos

Módulo III (Estatística)

I) Conceptos básicos

Módulo IV (Estatística descritiva)

I) Tipos de variables
II) Variables cuantitativas
    i. Representación gráfica
    ii. Descrición dos datos
III) Variables cualitativas
    i. Representación gráfica
    ii. Descrición dos datos
IV) Descritiva bivariante

Módulo V (Estatística inferencial)

I) Introdución
II) Inferencia
    i. Estimación puntual
    ii. Intervalos de confianza
    iii. Contrastes de hipóteses

Módulo I – Introdución a R

I) Introdución

i) Que é?

ii) Por que utilizar R?

iii) Interface

Módulo I – Introdución a R

I) Introdución

i) Que é?

ii) Por que utilizar R?

iii) Interface

R project

I) Que é?

Software estatístico libre e gratuíto

- **Linguaxe de programación orientado a obxectos:**
  As variables, datos, resultados, funcións,... almacénanse na área de traballo mediante **obxectos** cun nome.

- **Paquete estatístico** que permite:
  - Manexo de bases de datos
  - Análises estatísticas
  - Representacións gráficas

Módulo I – Introdución a R

I) Introdución

i) Que é?

ii) Por que utilizar R?

iii) Interface

R project

II) Por que utilizar R?

- **R** está avalado por unha comunidade académica que proporciona unha gran **variedade de paquetes** que permiten estimar e solucionar unha ampla gama de problemas.
- **R** é **multiplataforma** (funciona en Mac, Windows ou Linux).
- **R** traballa de **maneira integrada** con **outro tipo de linguaxes**.

Vainos permitir…

- **Flexibilidade** para realizar as **análises estatísticas** (ó contrario doutros paquetes que se manipulan con ventás ou pestanas).
- **Representacións gráficas** de calidade e variadas.

R project

II) Por que utilizar R?

Representacións gráficas variadas…

<u>Meteoroloxía</u>

Exemplo: Representar a <u>traxectoria</u> do Furacán "Andrew" (1992)

Datos: "Andrew"

Paquete: googleVis

II) Por que utilizar R?

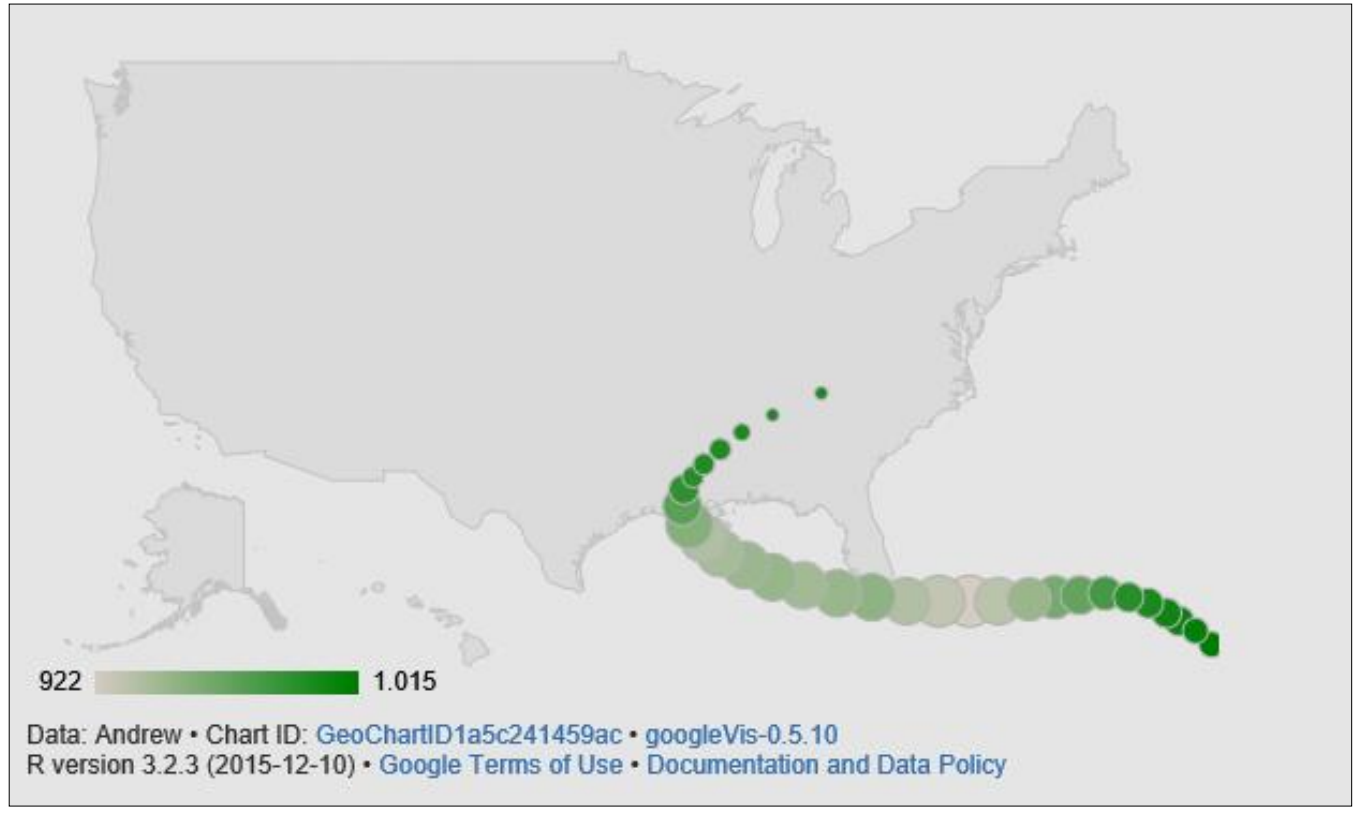
Representacións gráficas variadas…

Meteoroloxía

Exemplo: Representar a traxectoria e a presión do Furacán "Andrew" (1992)

Datos: "Andrew"

Paquete: googleVis



R project

II) Por que utilizar R?

Representacións gráficas variadas…

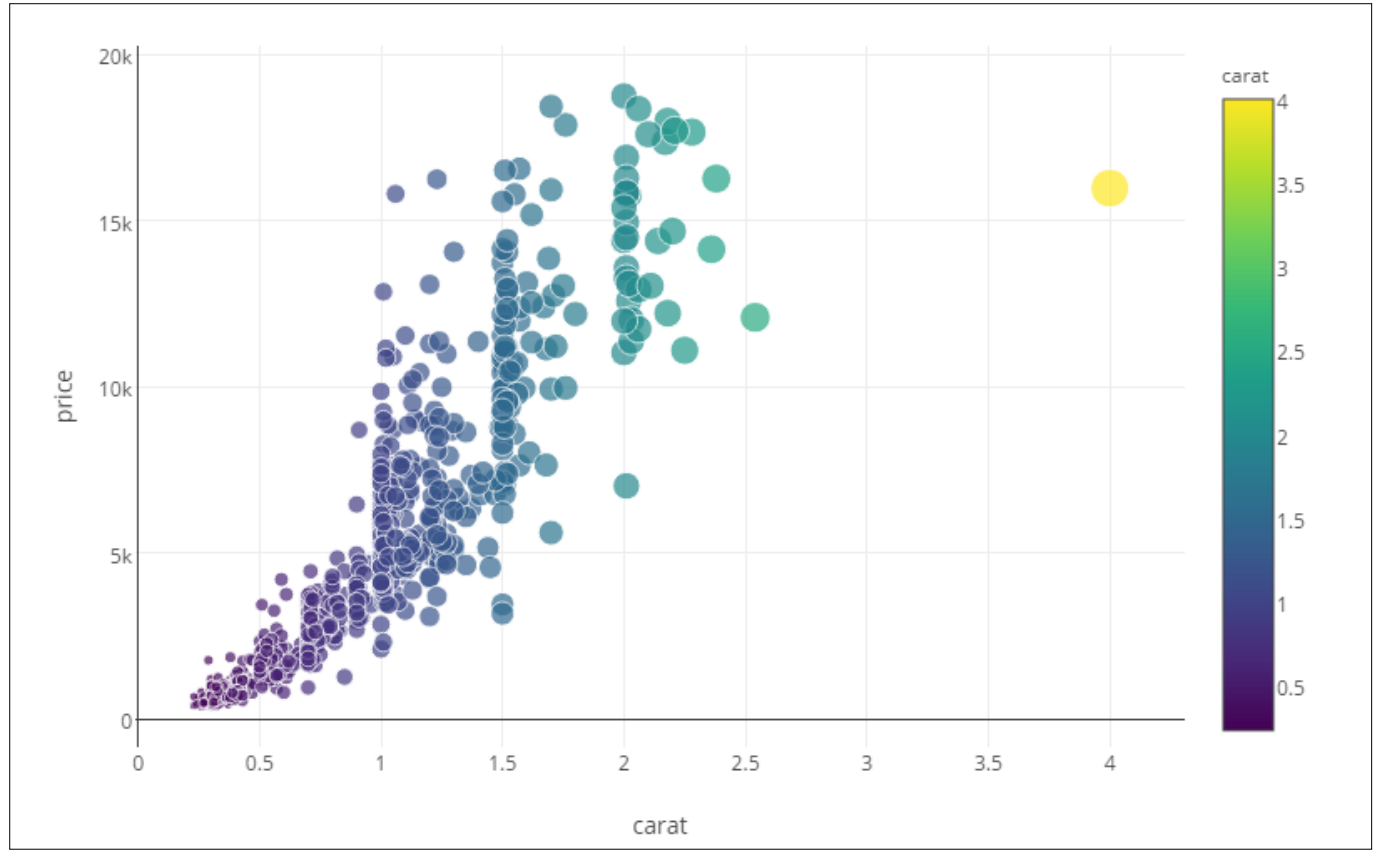
<u>Economía</u>

Exemplo: Representar a relación entre o <u>tamaño</u> e o <u>prezo</u> do diamante

Datos: “diamonds”

- price - Prezo en dólares ($326-$18,823)
- carat - tamaño do diamante (0.2--5.01)

Paquete: plotly



R project

II) Por que utilizar R?

Representacións gráficas variadas…

<u>Criminoloxía</u>

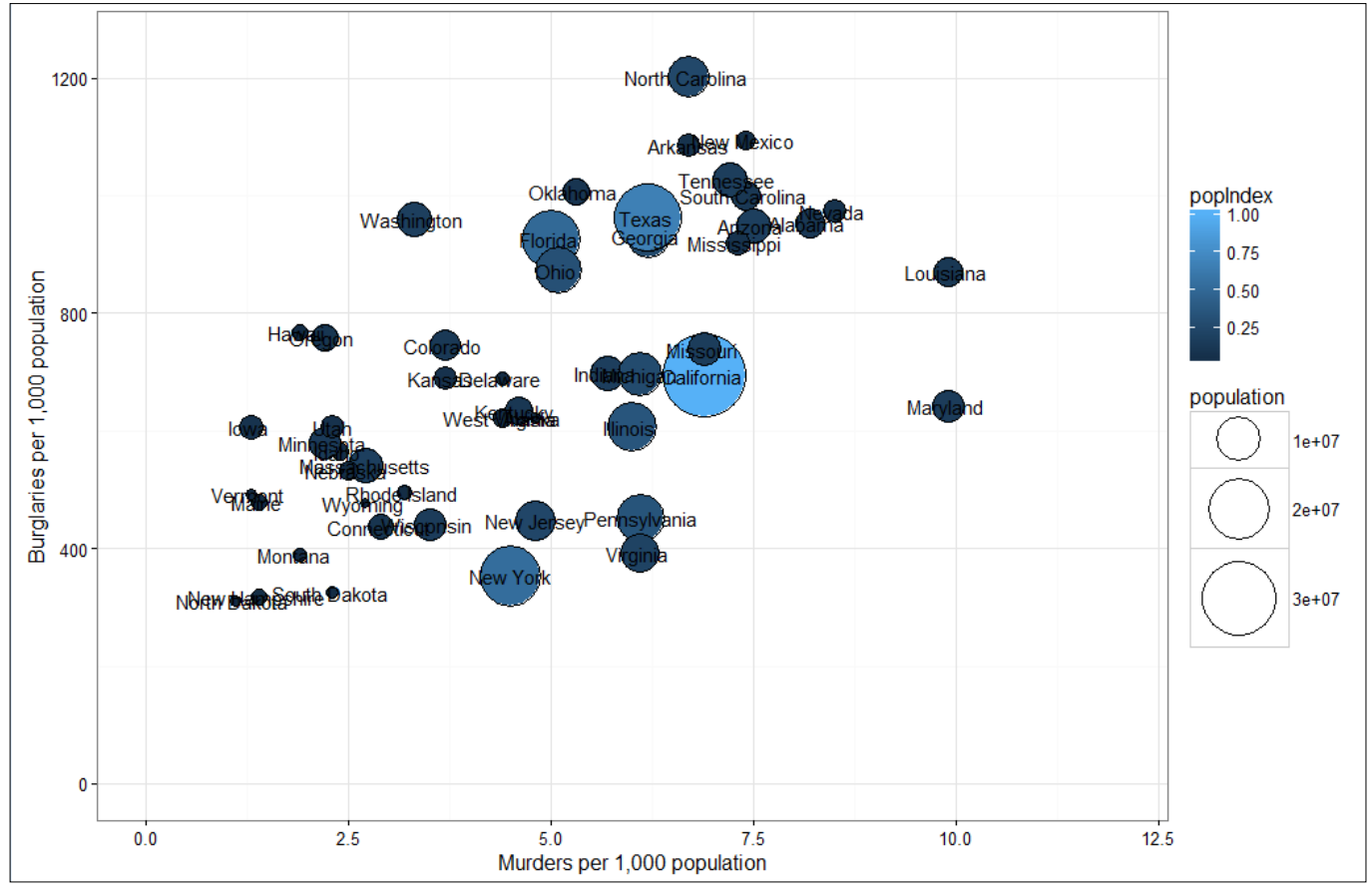
Exemplo: Representar a <u>relación</u> entre os roubos e os asasinatos mentres que se observa o <u>tamaño</u> da poboación (Estados Unidos)

Datos: “crimeRatesByState2005.tsv”

- Roubos
- Asasinatos
- Poboación
- Índice de poboación

  popindex=poboación/máx(pobación)

Paquete: ggplot2



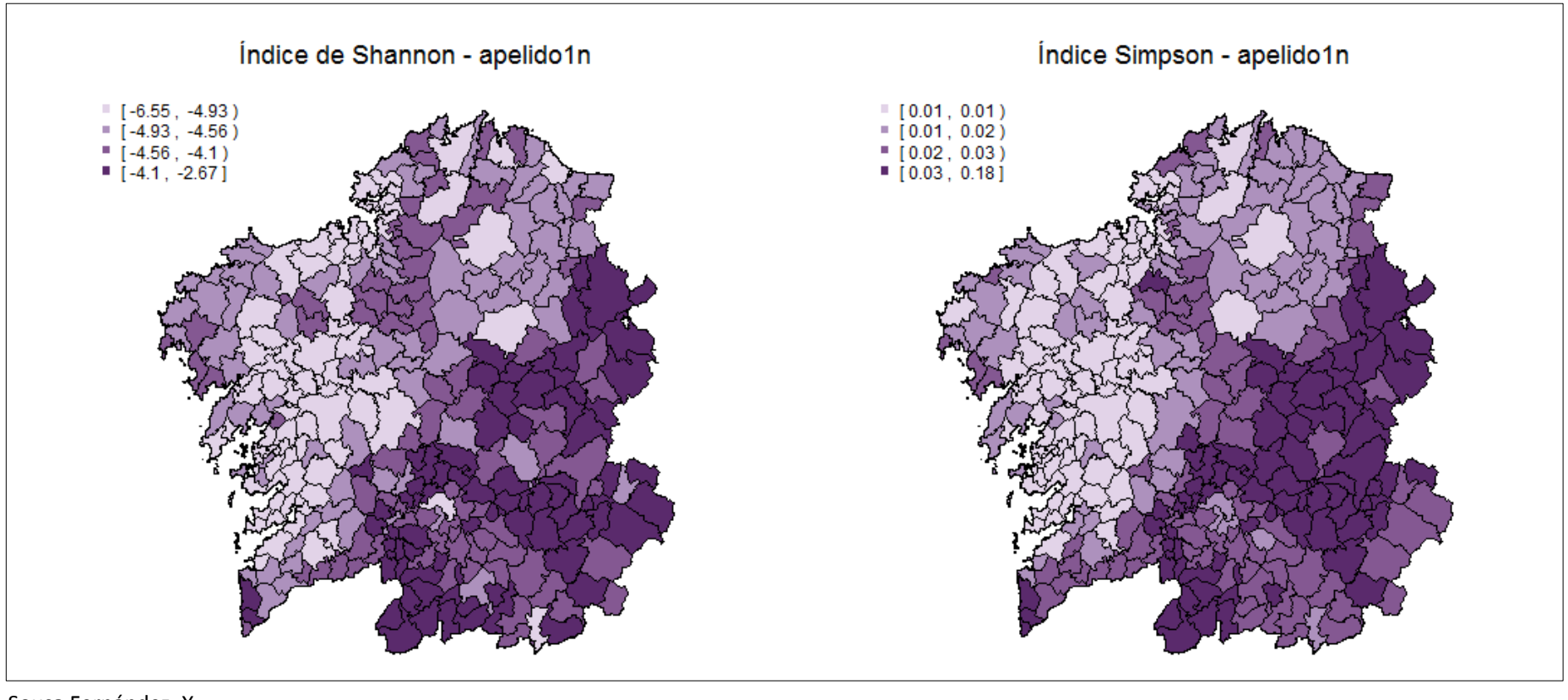
R project

II) Por que utilizar R?

Representacións gráficas variadas…

Tononimia e cartografía

Exemplo: Mostra a diversidade de apelidos nos concellos galegos



Sousa Fernández, X.

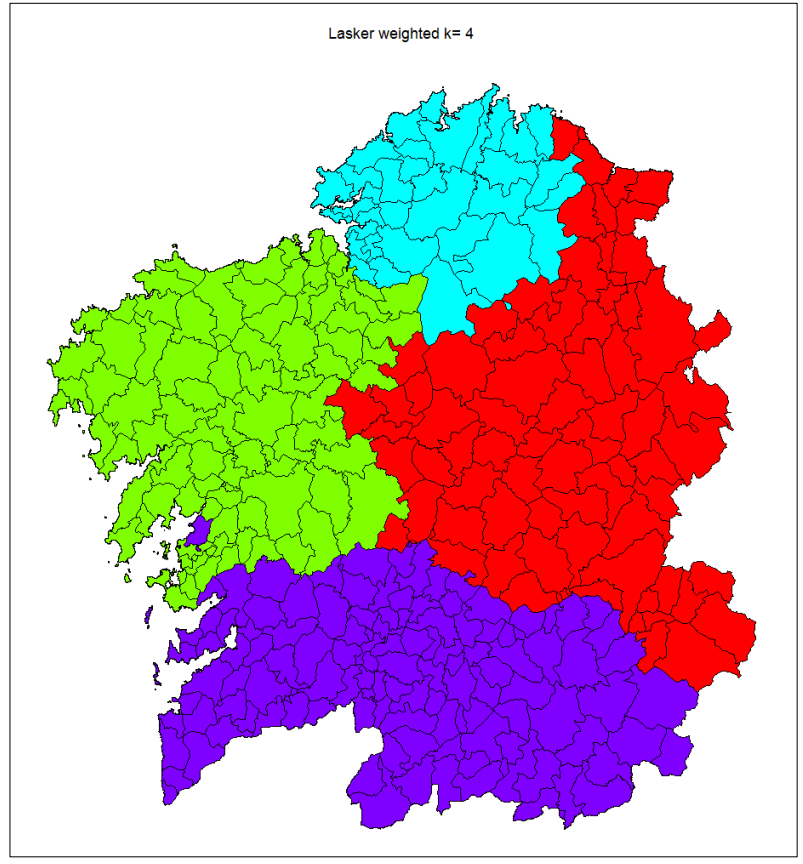
R project

II) Por que utilizar R?

Representacións gráficas variadas…

Toponimia e cartografía (rexións dos apelidos)

Exemplo: Mostra o resultado dunha análise clúster dos apelidos galegos (áreas xeográficas que comparten unha serie de apelidos)



Sousa Fernández, X.

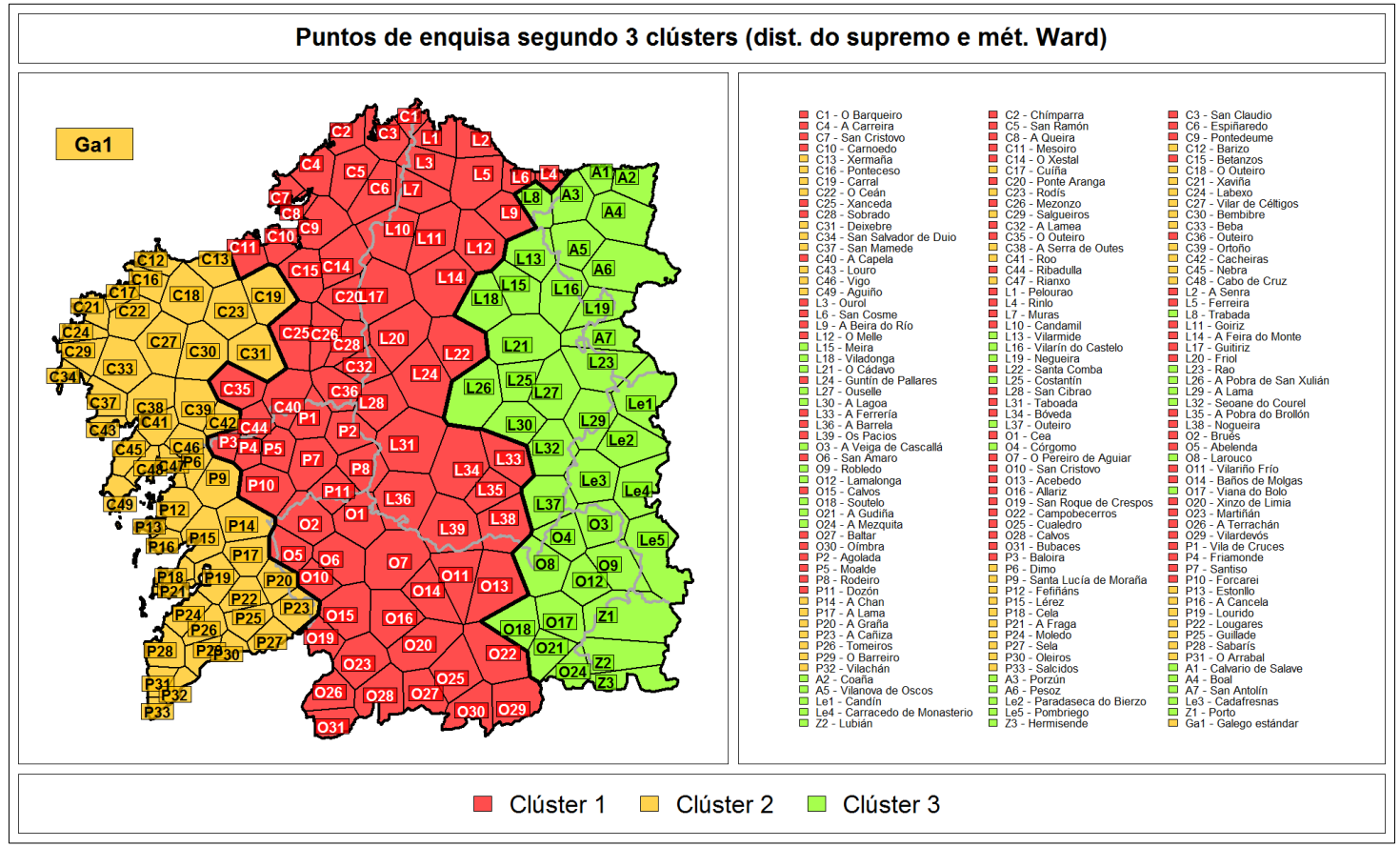
R project

II) Por que utilizar R?

Representacións gráficas variadas…

Xeolingüística

Exemplo: Mostra o resultado dunha análise clúster das variedades dialectais do galego a partir de variables morfosintácticas.



Sousa Fernández, X.

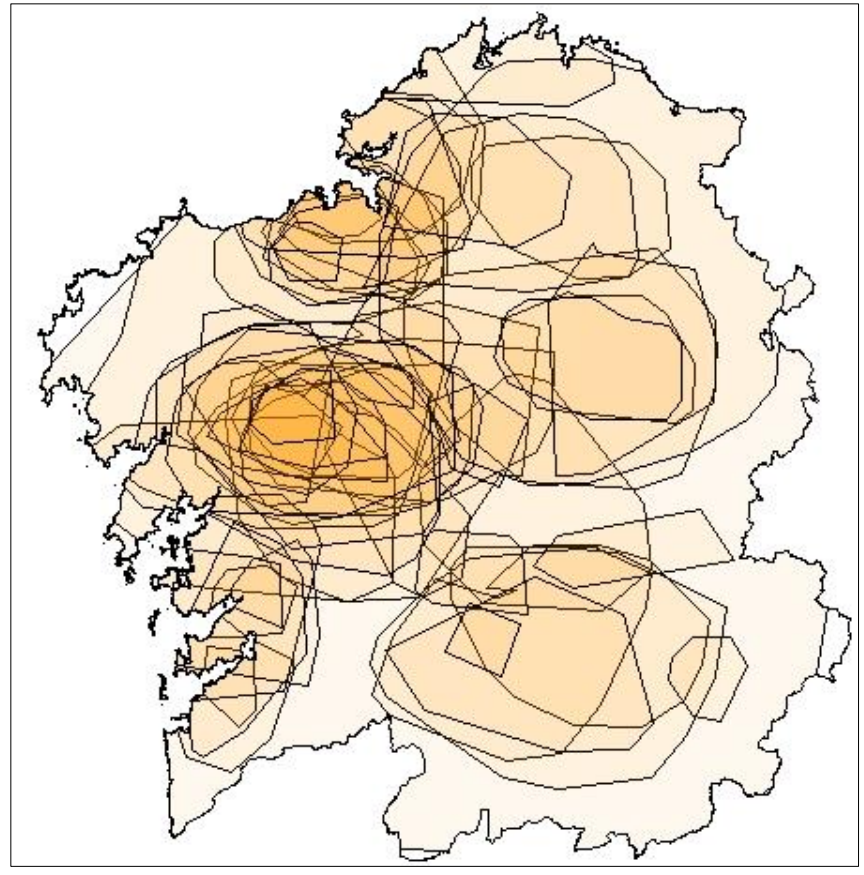
R project

II) Por que utilizar R?

Representacións gráficas variadas…

<u>Xeolingüística</u>

Exemplo: Na área de Dialectoloxía perceptiva, permite mostrar as diferentes percepcións



Datos correspondentes a:
Suárez Quintas, S. (2015-2016)
Tese en curso: A percepción da variación lingüística en galego: os falantes e os dialectos.

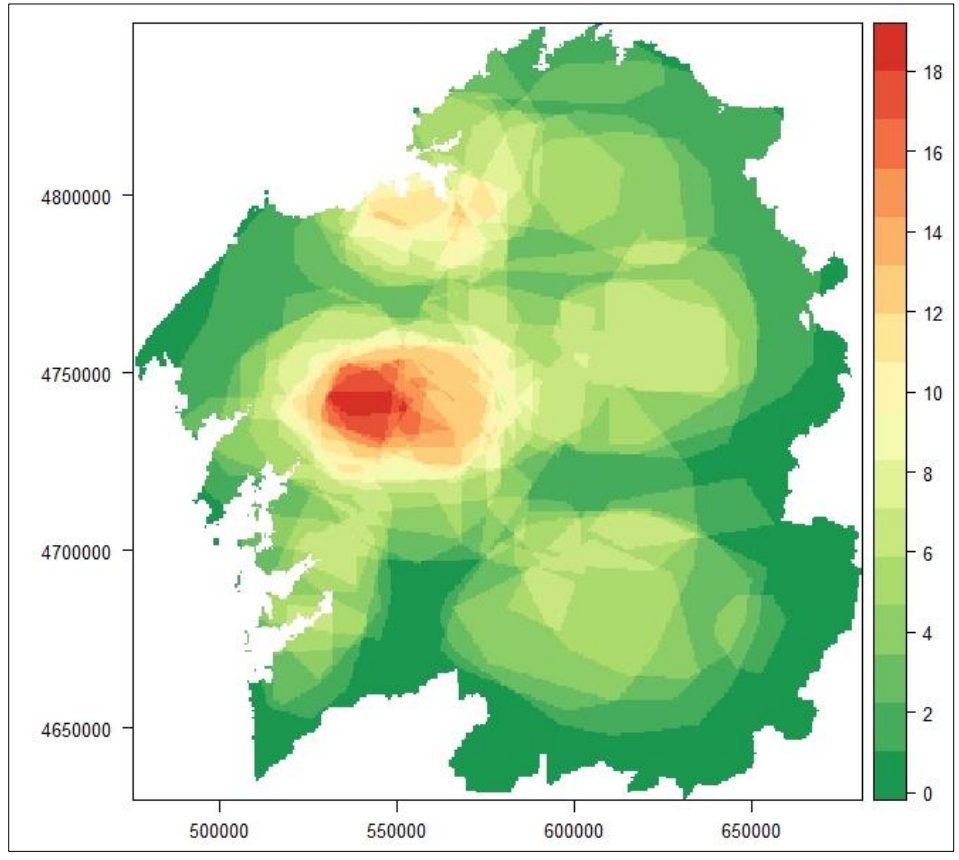
R project

II) Por que utilizar R?

Representacións gráficas variadas…

<u>Xeolingüística</u>

Exemplo: Na área de Dialectoloxía perceptiva, permite mostrar as diferentes percepcións (mapas de calor)



Datos correspondentes a:
Suárez Quintas, S. (2015-2016)
Tese en curso: A percepción da variación lingüística en galego: os falantes e os dialectos.

Módulo I – Introdución a R

I) Introdución

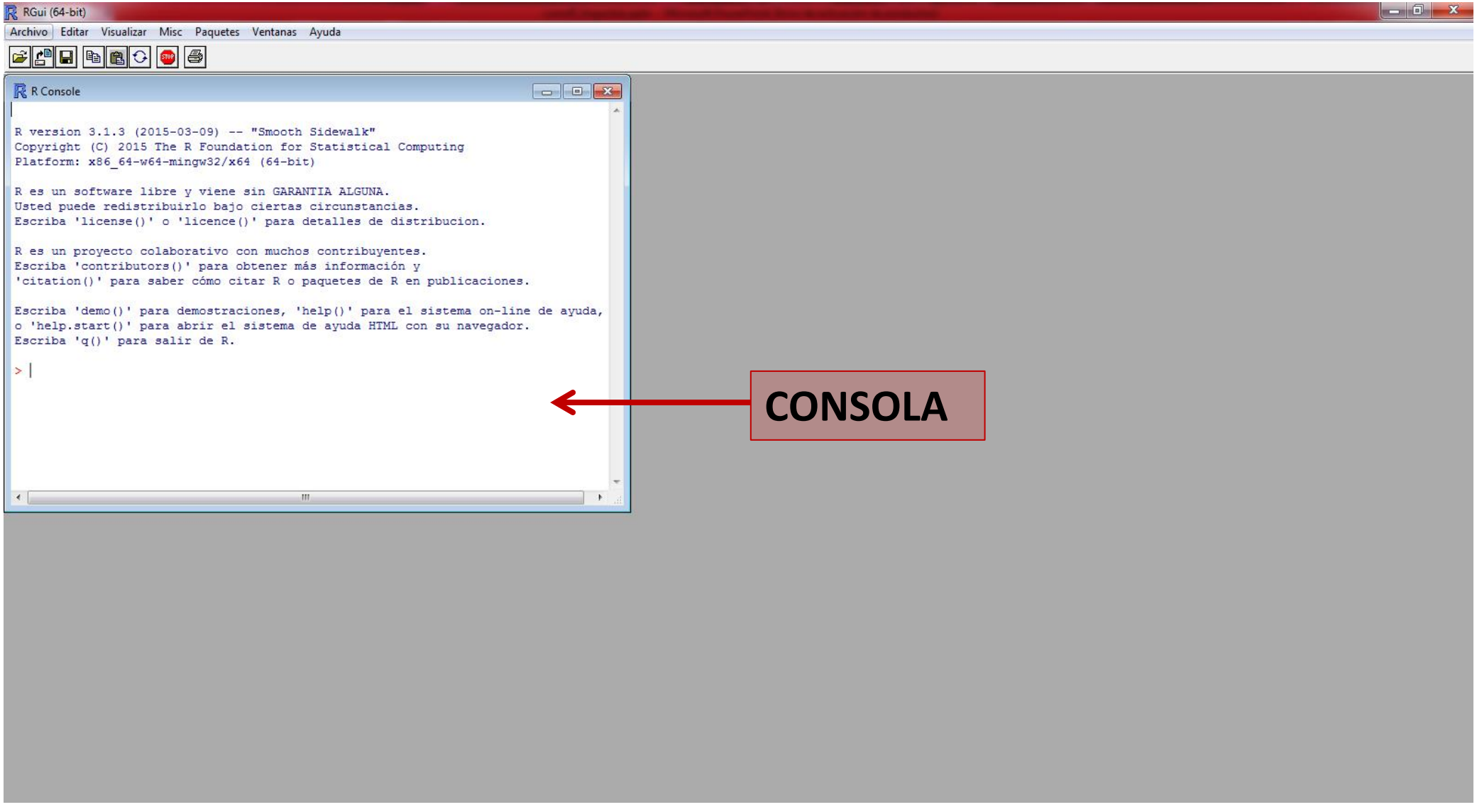
i) Que é?

ii) Por que utilizar R?

iii) Interface

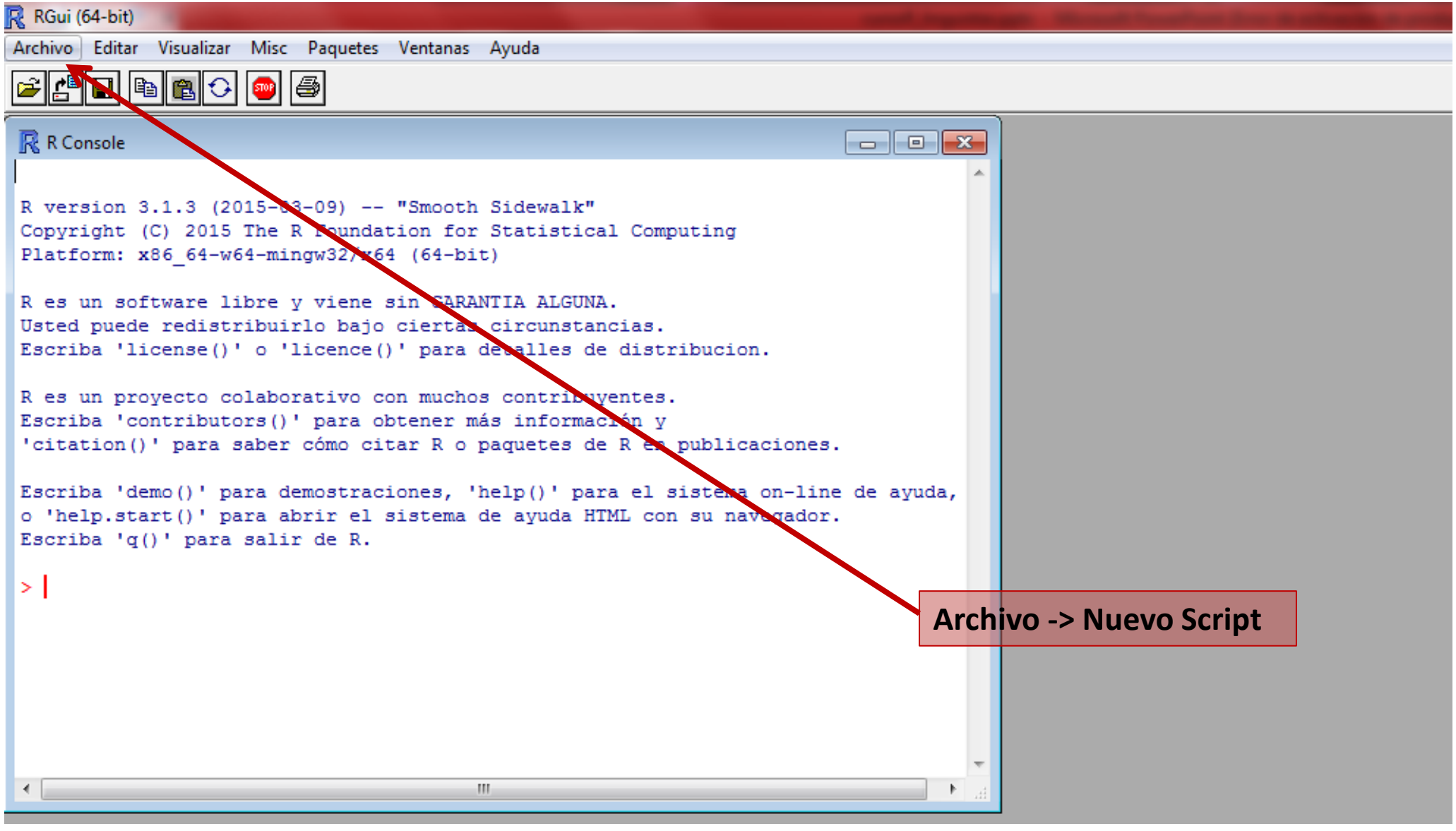
R project

III) Interface



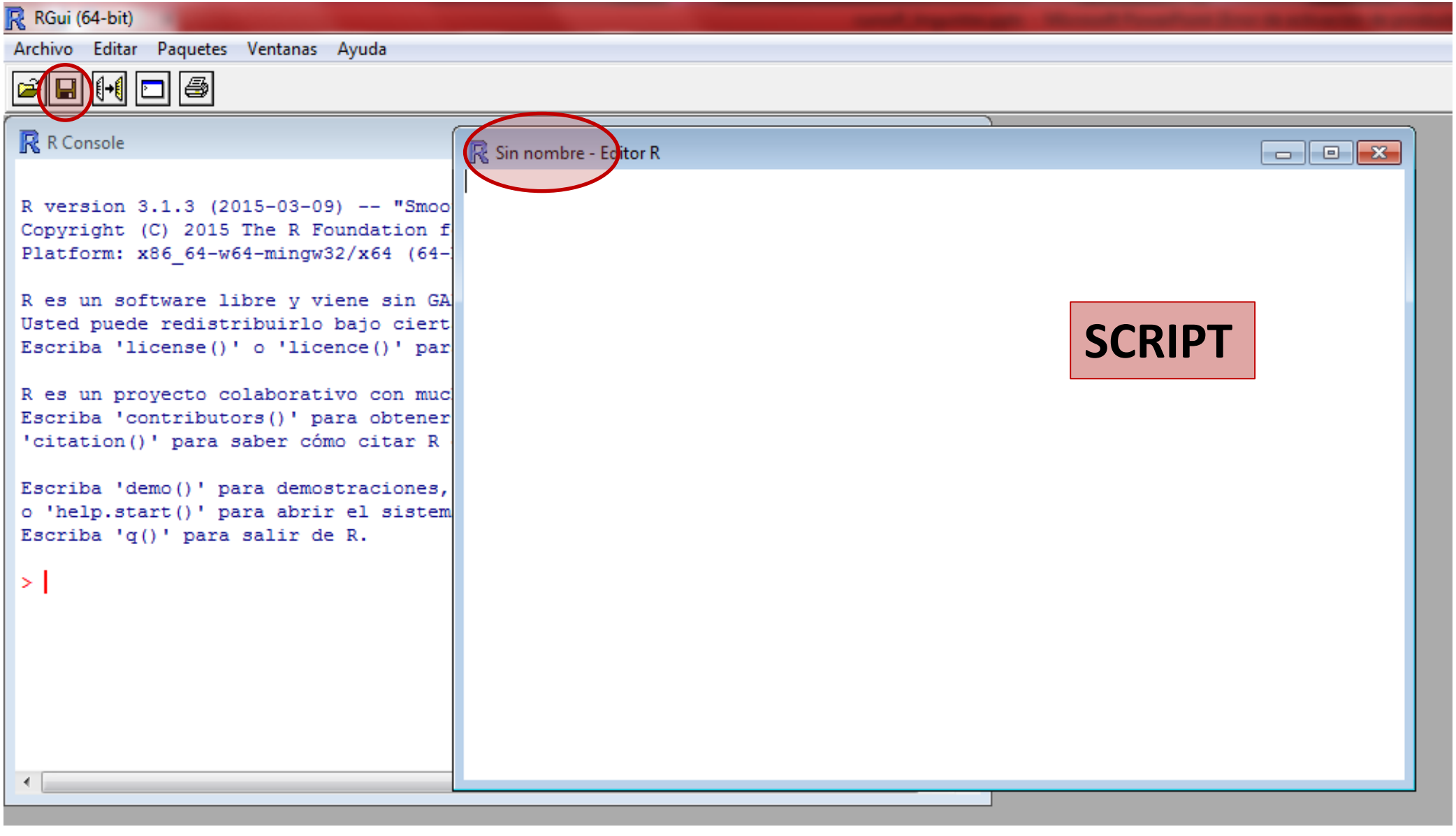
R project

III) Interface



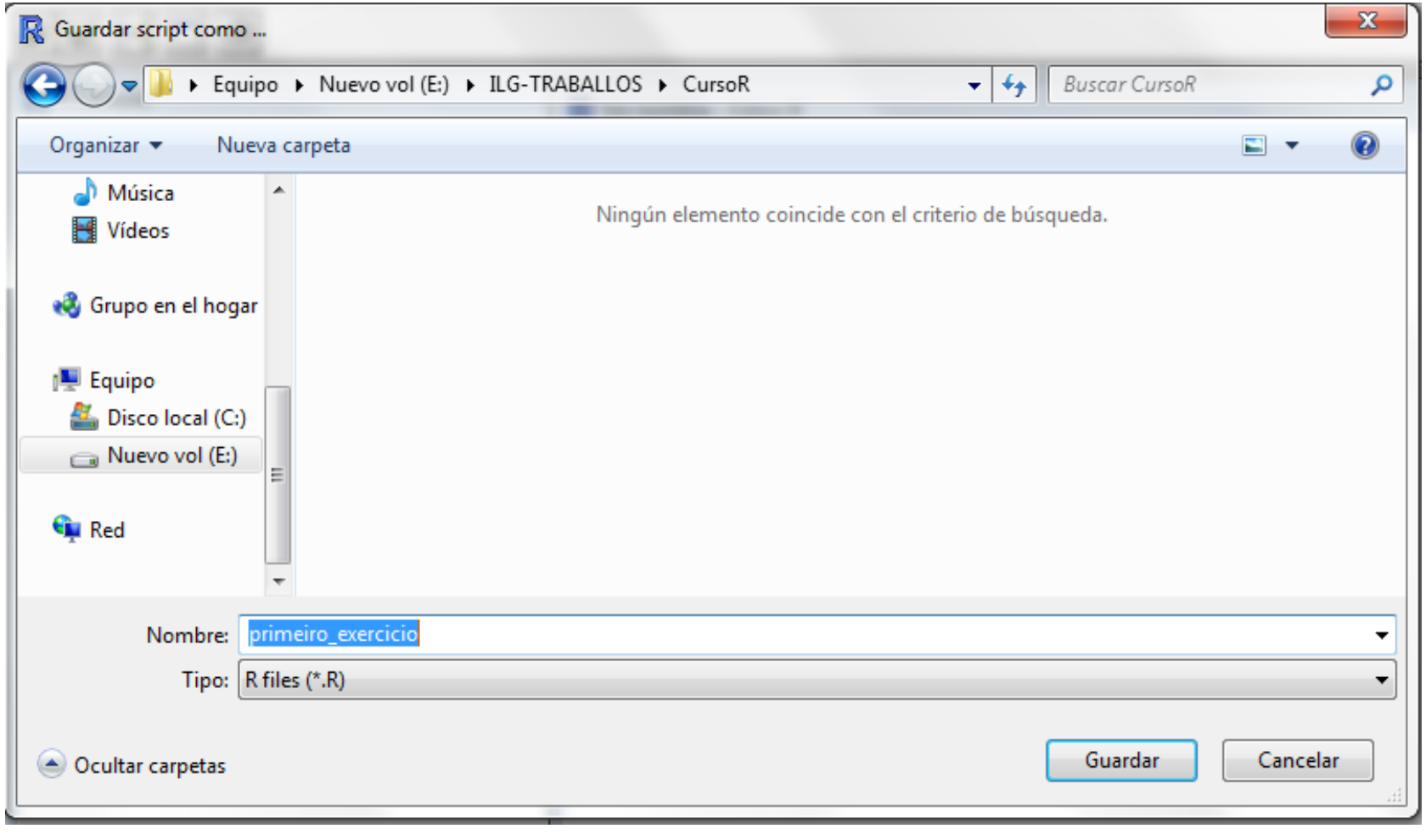
R project

III) Interface



R project

III) Interface



R project

III) Interface

OLLO: É necesario controlar o directorio!

Nada máis comezar a traballar o primeiro que faremos é ir a:

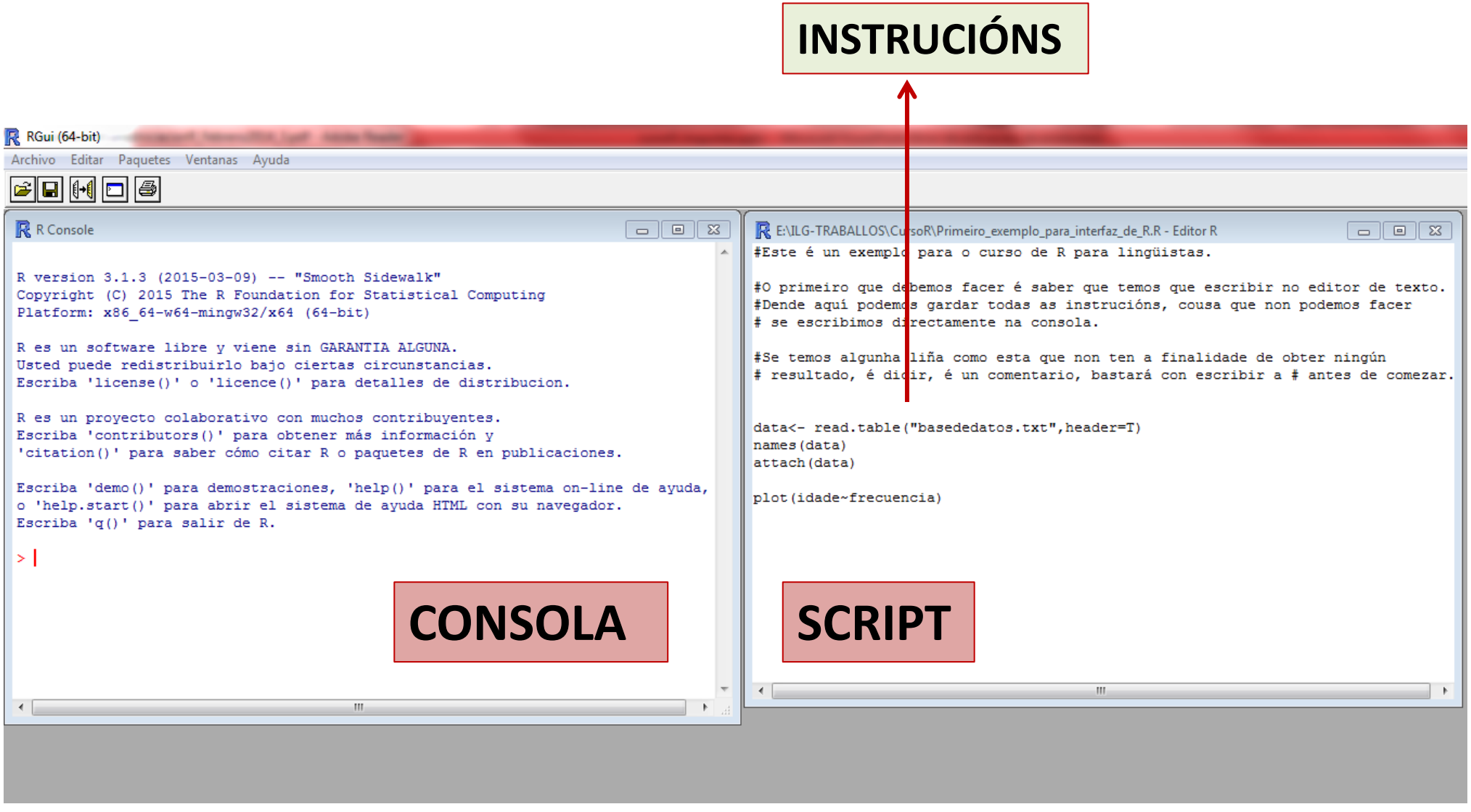
Archivo -> Cambiar dir... -> e coller a ruta onde imos traballar

Debemos ter en conta que:

1. Se traballamos cun script xa elaborado debe estar nesa ruta
2. Se temos unha base de datos debe estar nesa mesma ruta

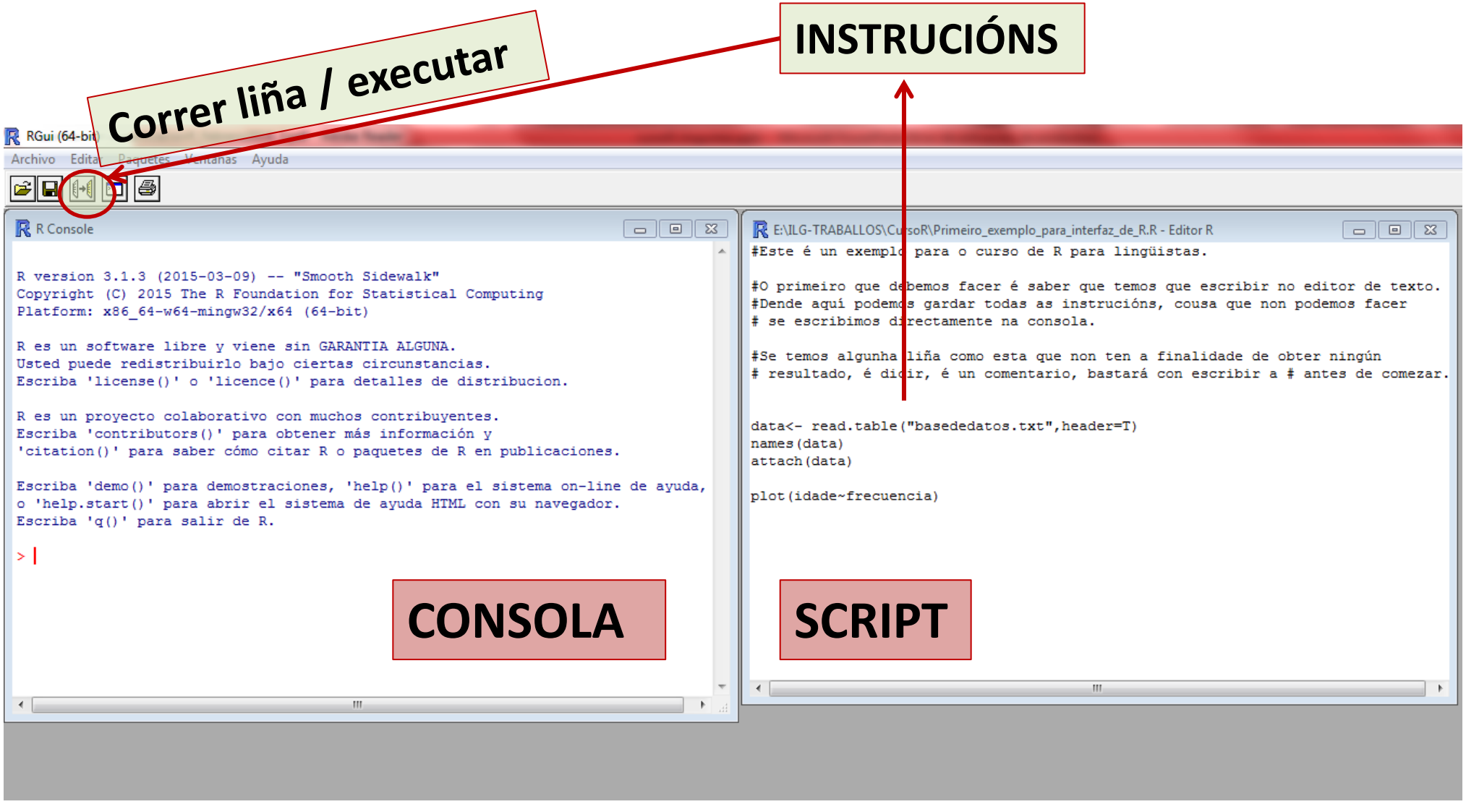
R project

III) Interface



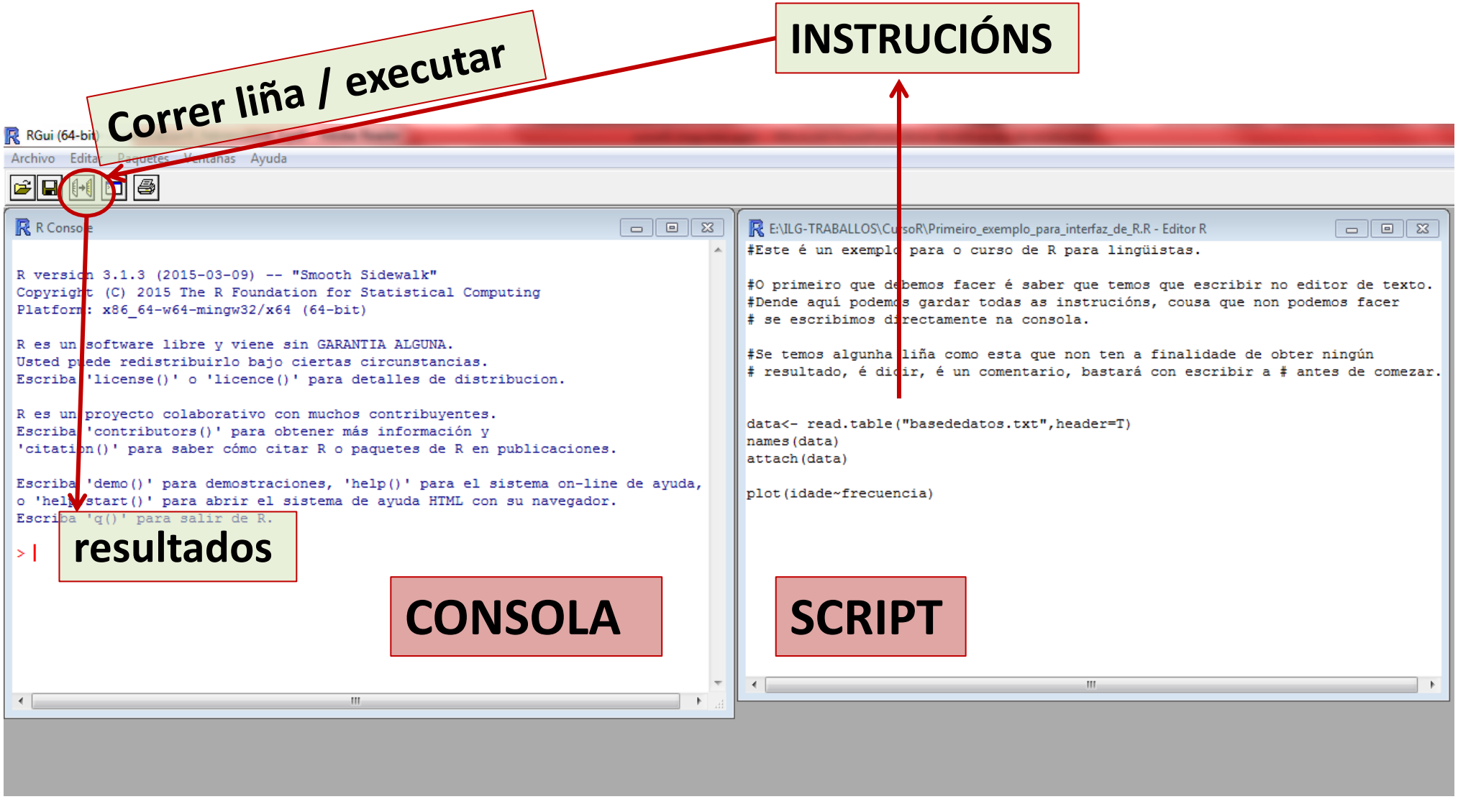
R project

III) Interface



R project

III) Interface



Módulo I – Introdución a R

II) Antes de comezar...

i) Axuda

ii) Paquetes

iii) Operacións aritméticas

Módulo I – Introdución a R

II) Antes de comezar...

i) Axuda

ii) Paquetes

iii) Operacións aritméticas

R project

I) Axuda

- Para obter axuda sobre cada comando
  help(comando) ou ?comando
- Para obter exemplos de uso do comando:
  example(comando)
- Para obter unha lista de comandos relacionados cun tema:
  help.search("tema")
- Para abrir unha ventá de axuda HTML:
  help.start()

Exemplo:
help(plot)
example(plot)
help.search("regression")

Módulo I – Introdución a R

II) Antes de comezar...

i) Axuda

ii) Paquetes

iii) Operacións aritméticas

R project

II) Paquetes

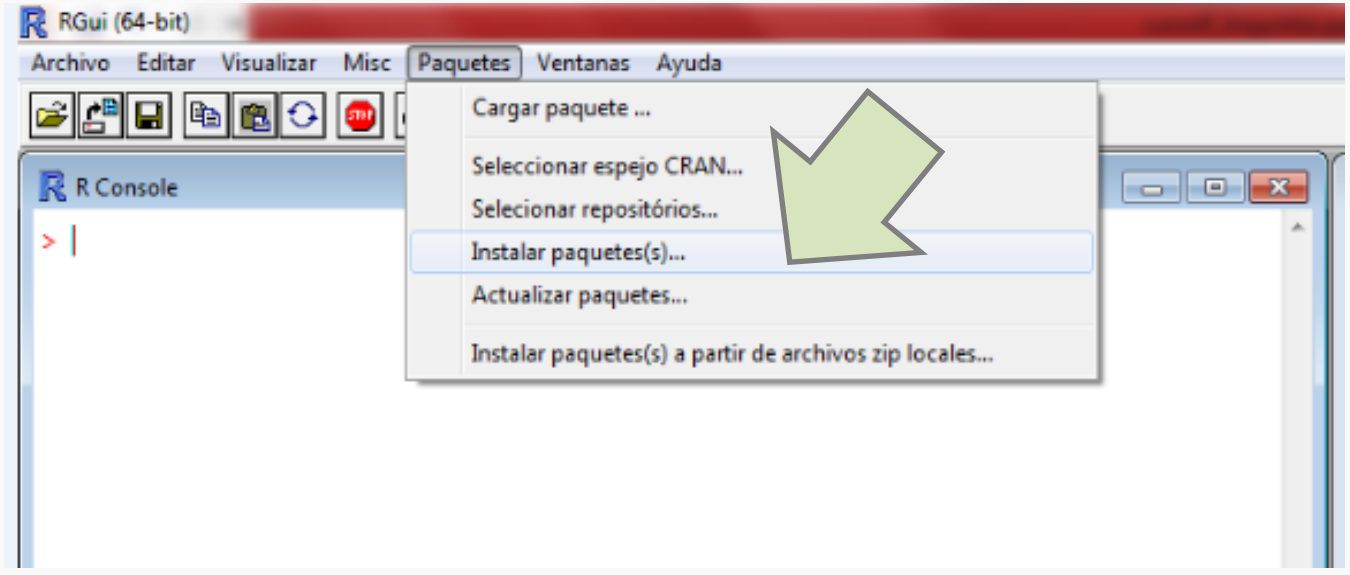
- **A información en R (métodos estatísticos e funcións) está estruturada en paquetes ou librarías**
- **Algunhas funcións xa veñen instaladas por defecto:** min(), max(), log()...

Como podemos ter acceso a estes paquetes?

R project

II) Paquetes

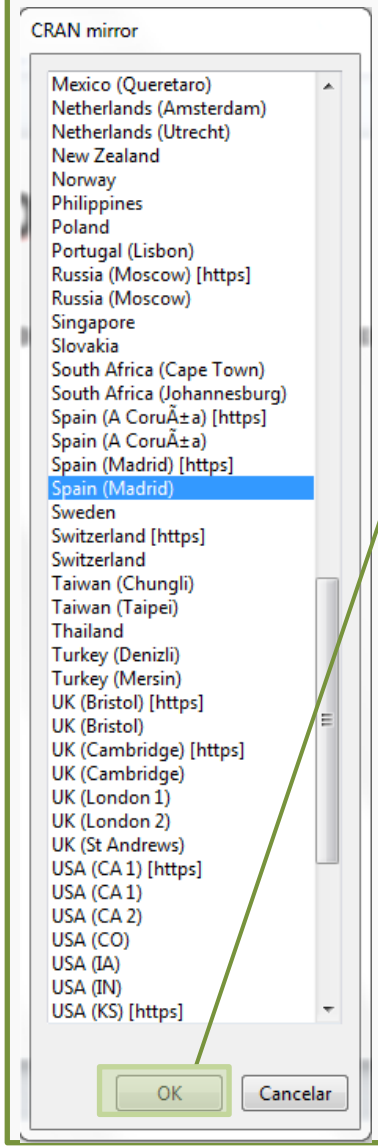
INSTALAR → **CARGAR**

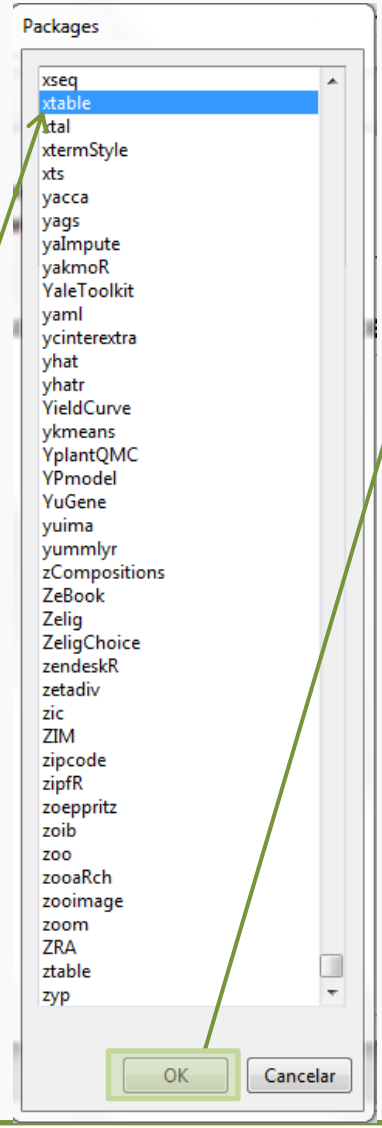


R project

II) Paquetes

INSTALAR





Na consola aparece o proceso de instalación:

```
> utils:::menuInstallPkgs()
--- Please select a CRAN mirror for use in this session ---
probando la URL 'http://cran.rediris.es/bin/windows/contrib/3.1/xtable_1.8-0.zi$
Content type 'application/zip' length 326813 bytes (319 KB)
URL abierta
downloaded 319 KB

package 'xtable' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Laura\AppData\Local\Temp\Rtmp8wuoPX\downloaded packages
```

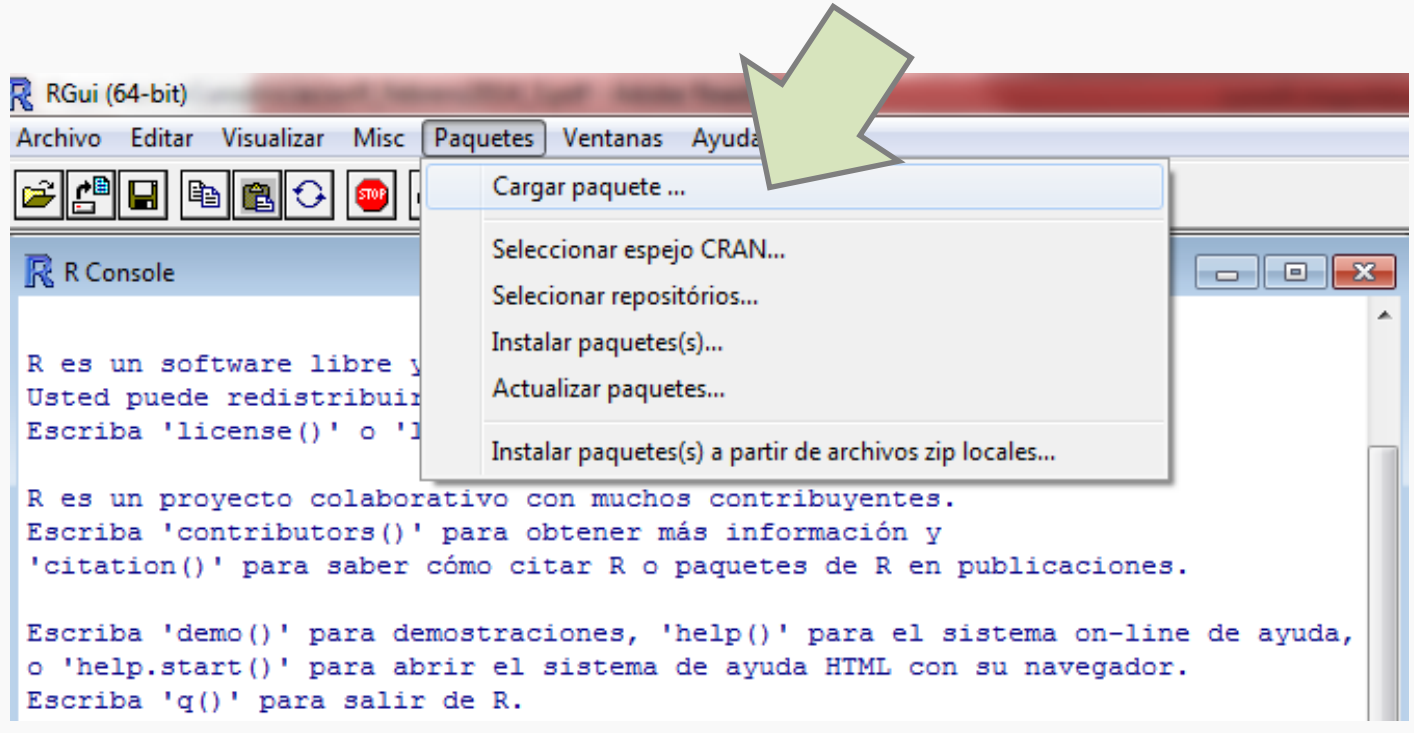
CARGAR

R project

II) Paquetes

INSTALAR →

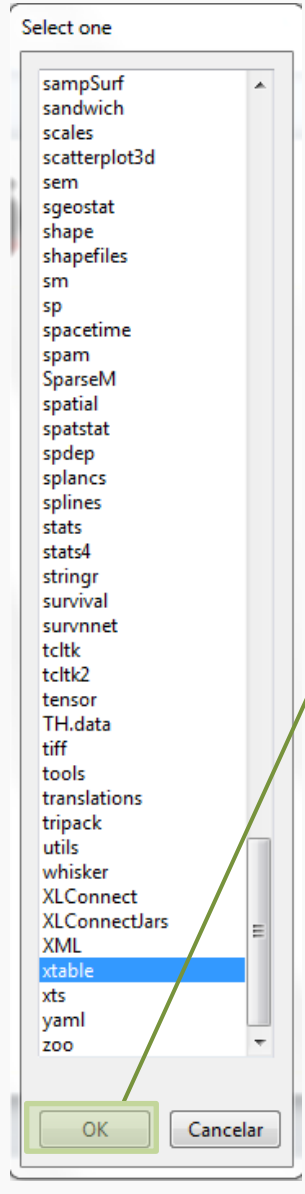
CARGAR



R project

II) Paquetes

INSTALAR



CARGAR

Na consola aparece que o paquete foi cargado:

```
> local({pkg <- select.list(sort(.packages(all.available = TRUE)),graphics=TRUE)
+ if(nchar(pkg)) library(pkg, character.only=TRUE)})
```

R project

II) Paquetes

Este mesmo procedemento de instalación e carga pódese facer en liña de comandos:

INSTALAR

install.packages("Nombre paquete")

CARGAR

library(Nombre paquete)

Lista de todos os paquetes dispoñibles que podemos cargar:

library()

Exemplo:

Instalación do paquete:

> install.packages("languageR")

Cargar o paquete:

> library(languageR)

Módulo I – Introdución a R

II) Antes de comezar...

i) Axuda

ii) Paquetes

iii) Operacións aritméticas

R project

III) Operacións aritméticas

R como unha calculadora

Suma	2+2
Resta	10-5
Multiplicación	2*2
División	10/2
Potencias	3^2
Raíz cadrada	4^(1/2); sqrt(4)
Raíz cúbica ; raíz n-esima	8^(1/3) ; 8^(1/n)
Logaritmo neperiano	log(e)
Logaritmo en base 10	log(1,10)

Conserva a orde das operacións:

2*(3+4) + 1/2 * (3 +5)

Módulo II – A información en R

I) Lectura/ Importación de datos

R project

I) Importación de datos

Nada máis comezar a traballar o primeiro que faremos é ir a:

Archivo -> Cambiar dir... -> e coller a ruta onde imos traballar

Debemos ter en conta que:

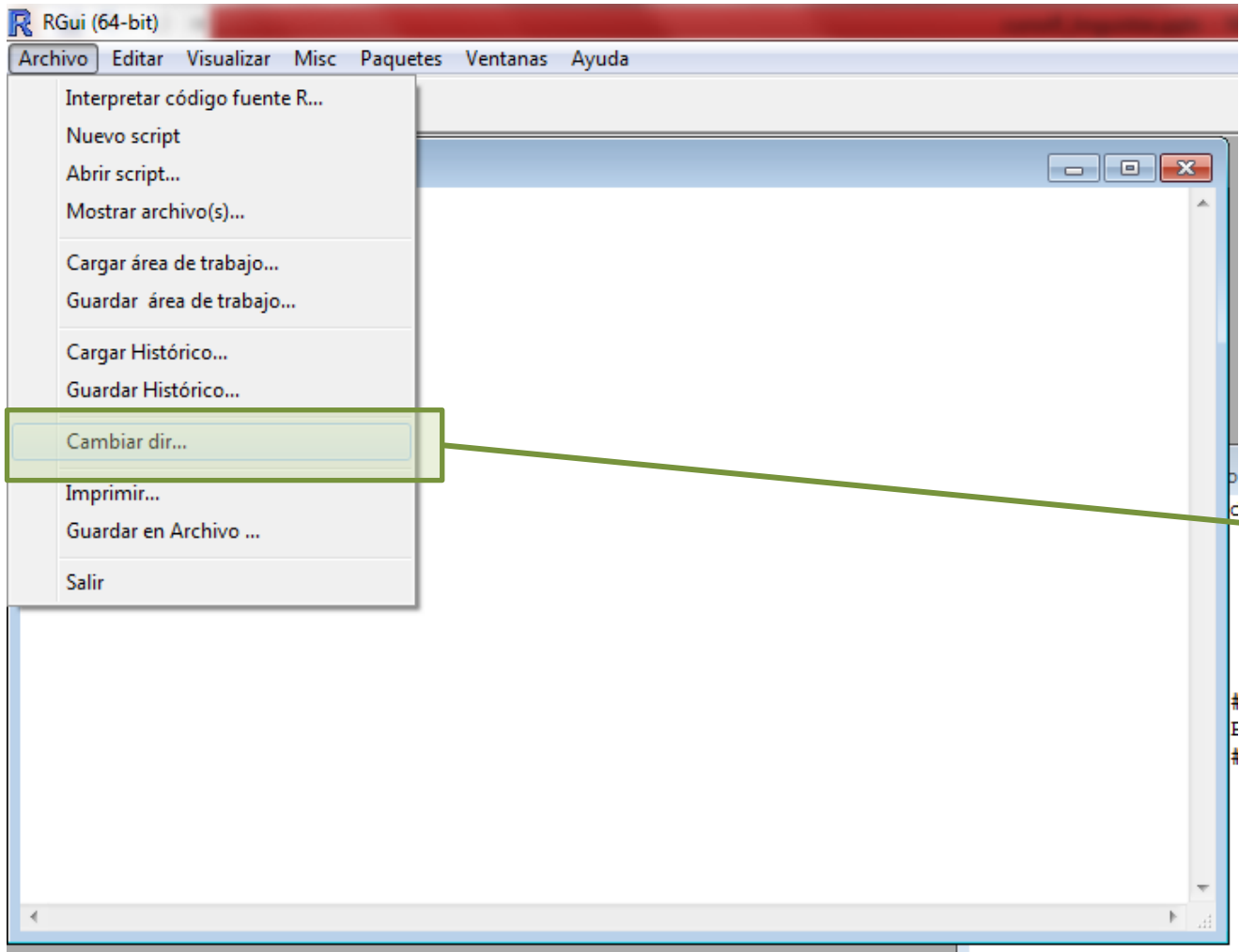
1. Se traballamos cun script xa elaborado debe estar nesa ruta
2. **Se temos unha base de datos coa que queiramos traballar debe estar nesa mesma ruta**

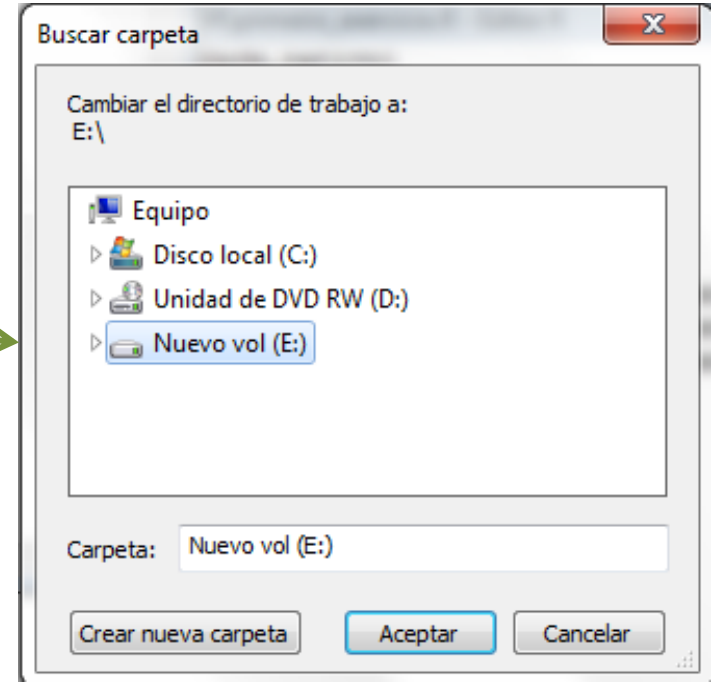
R project

I) Importación de datos

Recordatorio importante!

Archivo -> Cambiar dir... -> e coller a ruta onde imos traballar





R project

I) Importación de datos

Como ler ficheiros de datos en R?

Podemos **ler ficheiros de datos** en formato:

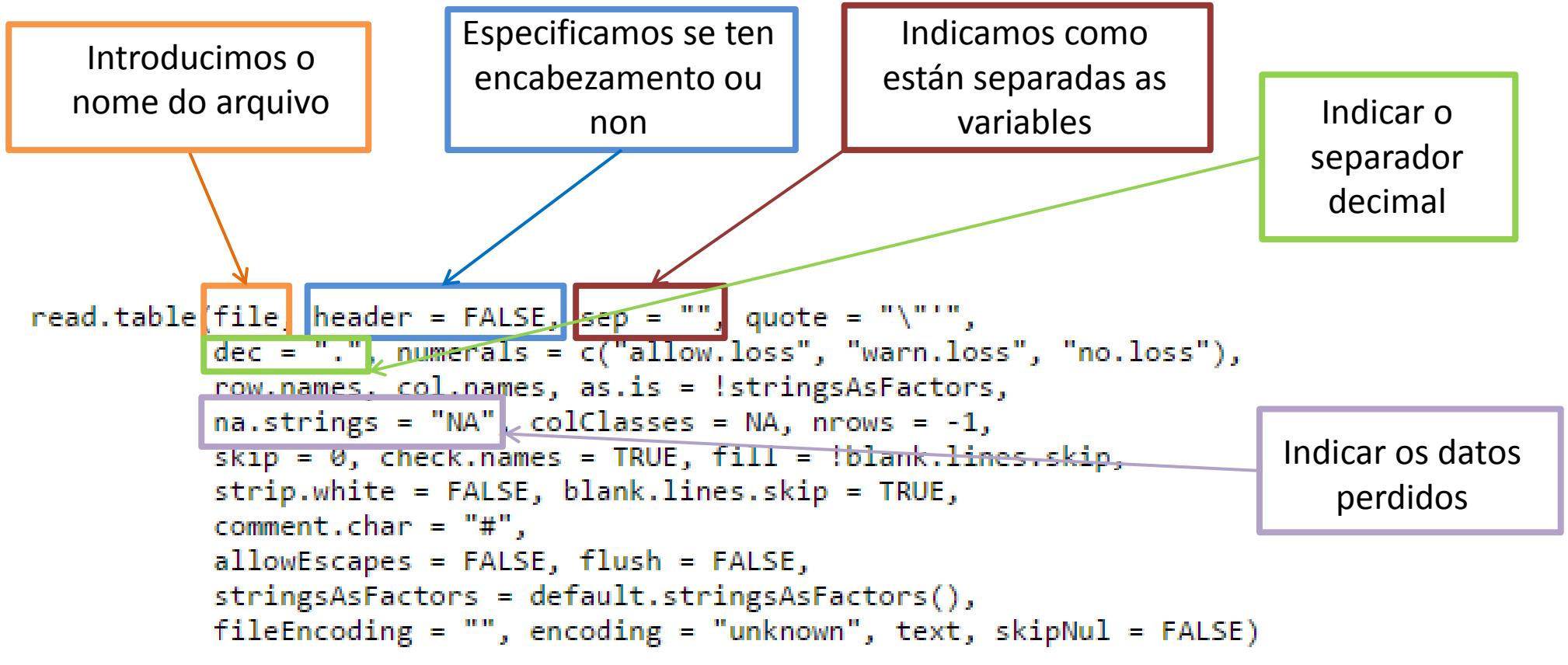
1. **.txt** **(en bloc de notas)**
2. **.xls** **(en excel)**
3. **.csv** **(en excel)**
4. **.sav** **(en spss)**

R project

I) Importación de datos

Como ler ficheiros de datos en R?

1. **Ficheiros de datos** en formato **.txt:** ***read.table()***



R project

I) Importación de datos

Como ler ficheiros de datos en R?

1. **Ficheiros de datos** en formato **.txt:** ***read.table()***

Exemplo 1:

```
xuices<-read.table("xuices.txt",header=TRUE)

View(xuices)
```

	id_xuiz	Idade	Sexo	Concello_da_vivenda_familiar	NATIVO
14	id_14	19	Muller	SANTIAGO_DE_COMPOSTELA	galego
15	id_15	28	Muller	VILAGARCIA_DE_AROUSA	bilingue
16	id_16	31	Home	CORUnA_A	castelan
17	id_17	24	Muller	CERCEDA	galego
18	id_18	28	Home	LUGO	castelan
19	id_19	25	Home	TRAZO	bilingue
20	id_20	20	Muller	VILARMAIOR	bilingue
21	id_21	20	Home	LEIRO	bilingue
22	id_22	19	Home	CANGAS	bilingue

R project

I) Importación de datos

Como ler ficheiros de datos en R?

1. **Ficheiros de datos** en formato **.txt:** ***read.table()***

Exemplo 2:

```
xuices_con_perdidos<-read.table("xuices_con_perdidos.txt",header=TRUE)
Error en scan(file, what, nmax, sep, dec, quote, skip, nlines, na.strings,  :
  la linea 2 no tiene 16 elementos
```

Temos que indicar os argumentos na.strings=“ “ e sep=“\t”:

```
xuices_con_perdidos<-read.table("xuices_con_perdidos.txt",header=TRUE,na.strings=" ",
sep="\t")
```

```
View(xuices_con_perdidos)
```

	id_xuiz	Idade	Sexo	Concello_da_viven
1	id_1	25	Muller	ORDES
2	id_2	31	Home	VILABOA
3	id_3	27	Home	PONTEVEDRA
4	id_4	NA	Muller	GUARDA_A
5	id_5	20	Home	VIGO
6	id_6	23	Muller	ORDES
7	id_7	23	Home	LEIRO

R project

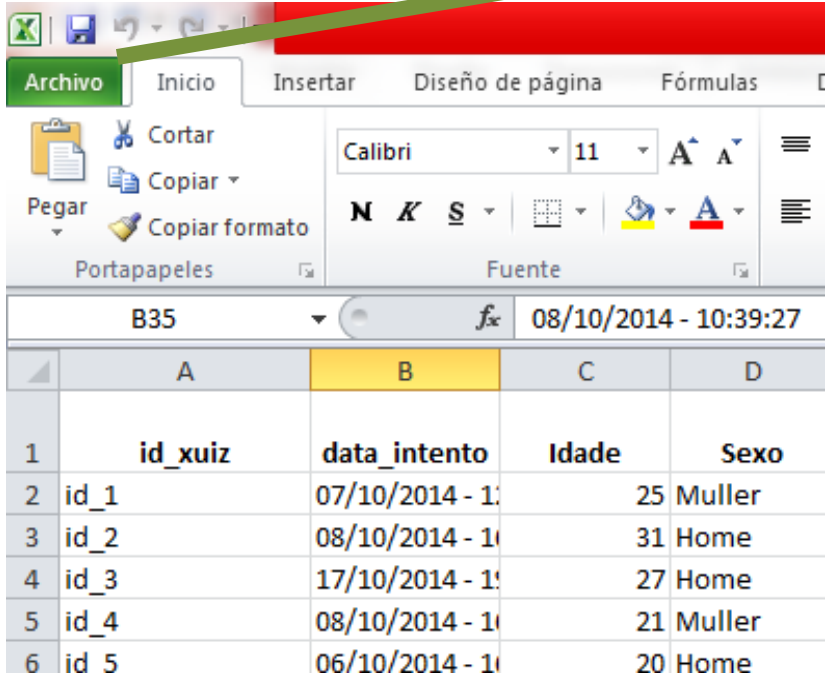
I) Importación de datos

Como ler ficheiros de datos en R?

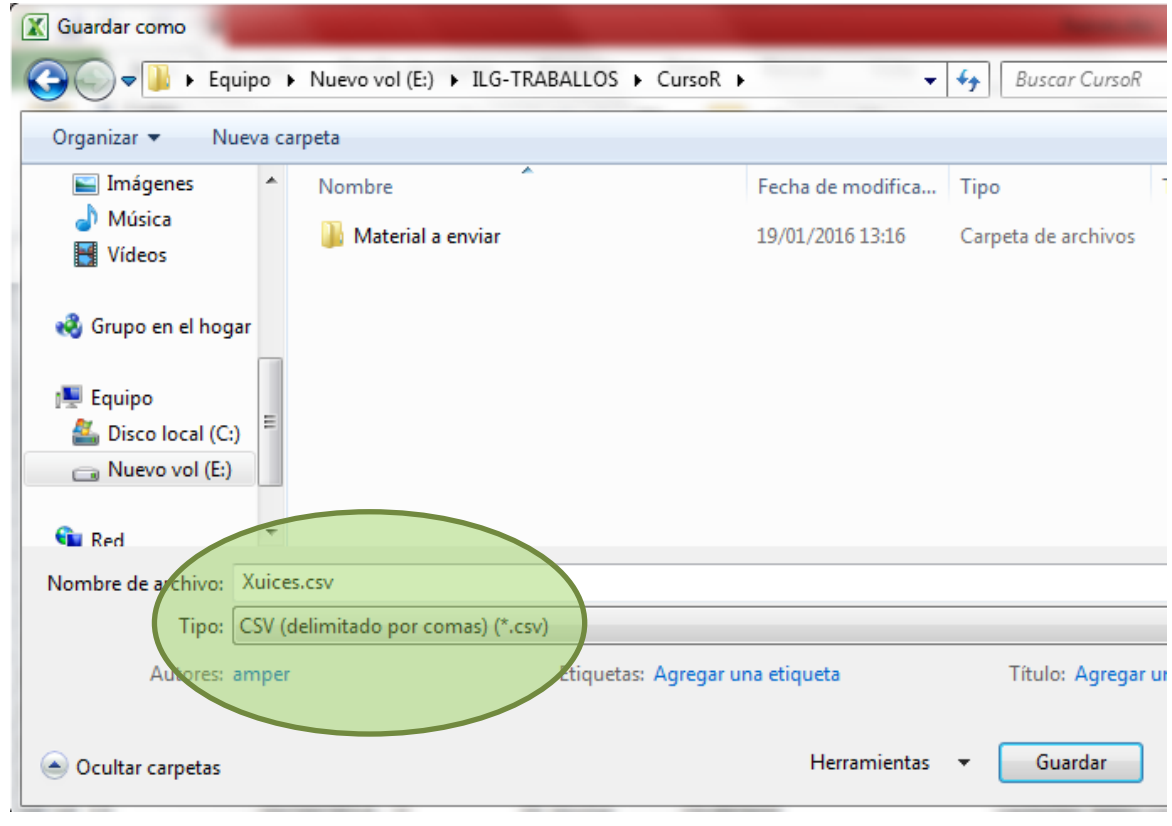
2. **Ficheiros de datos** en formato **.xls : *read.csv()***

En Excel

Archivo -> Guardar Como-> csv (delimitado por comas)



	A	B	C	D
1	id_xuiz	data_intento	Idade	Sexo
2	id_1	07/10/2014 - 1	25	Muller
3	id_2	08/10/2014 - 1	31	Home
4	id_3	17/10/2014 - 1	27	Home
5	id_4	08/10/2014 - 1	21	Muller
6	id_5	06/10/2014 - 1	20	Home



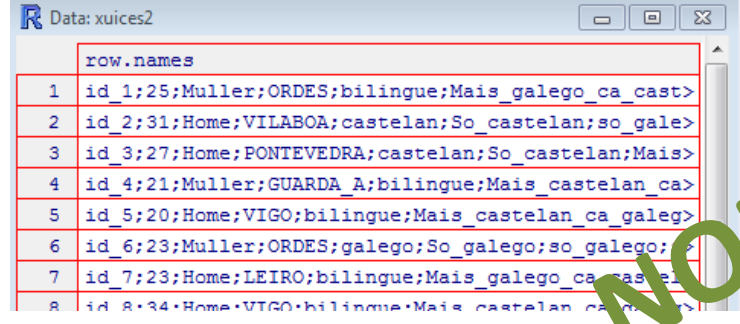
R project

I) Importación de datos

Como ler ficheiros de datos en R?

3. **Ficheiros de datos** en formato **.csv:** ***read.csv()***

```
xuices2<-read.csv("xuices.csv",header=T)
View(xuices2)
```



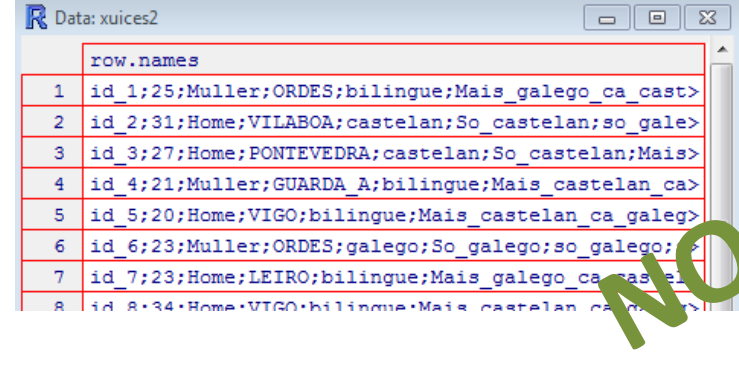
R project

I) Importación de datos

Como ler ficheiros de datos en R?

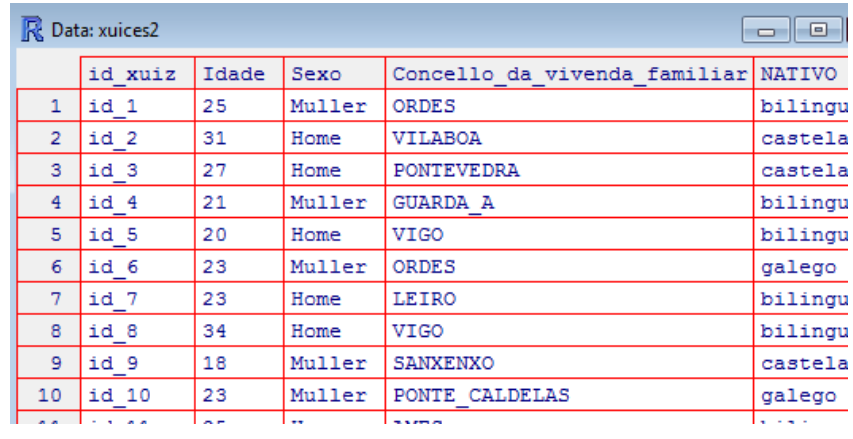
3. **Ficheiros de datos** en formato **.csv:** ***read.csv()***

```
xuices2<-read.csv("xuices.csv",header=T)
View(xuices2)
```



Temos que indicar o argumento sep=";":

```
xuices2<-read.csv("xuices.csv",header=T,sep=";")
View(xuices2)
```



	id_xuiz	Idade	Sexo	Concello_da_vivenda_familiar	NATIVO
1	id_1	25	Muller	ORDES	bilingu
2	id_2	31	Home	VILABOA	castela
3	id_3	27	Home	PONTEVEDRA	castela
4	id_4	21	Muller	GUARDA_A	bilingu
5	id_5	20	Home	VIGO	bilingu
6	id_6	23	Muller	ORDES	galego
7	id_7	23	Home	LEIRO	bilingu
8	id_8	34	Home	VIGO	bilingu
9	id_9	18	Muller	SANXENXO	castela
10	id_10	23	Muller	PONTE_CALDELAS	galego

R project

I) Importación de datos

Como ler ficheiros de datos en R?

4. **Ficheiros de datos** en SPSS, formato **.sav : *read.spss()***

OLLO: É necesario instalar e cargar o paquete «foreign»:

```
install.packages("foreign")
library(foreign)
?read.spss
Xuizo3<-read.spss("pretonica_e.sav",to.data.frame=TRUE)
View(xuizo3)
```

Data: xuizo3

	falante	sexo	palabra	cod_palabra	stress	posicion	d_frase	duracion
1	1	M	emitir	Emitir	1	2	200	96
2	1	M	emitir	Emitir1	1	1	188	85
3	1	M	emitir	Emitir2	1	1	171	78
4	2	M	emitir	Emitir	1	2	193	54
5	2	M	emitir	Emitir1	1	1	180	43
6	2	M	emitir	Emitir2	1	1	207	48
7	3	M	emitir	Emitir	1	2	202	64
8	3	M	emitir	Emitir1	1	1	196	76
9	3	M	emitir	Emitir2	1	1	202	60

R project

I) Importación de datos

Exercicio 1

Como podemos ler esta información?

Temos as seguintes bases coas que queremos traballar que están no material enviado:

tempos_compostos_galego_medieval.csv

1NT004916.txt

Como podemos ler esta información desde R project?

R project

I) Importación de datos

Exercicio 1

Solución

tempos_compostos_galego_medieval.csv

```
tempos_compostos<-read.csv("tempos_compostos_galego_medieval.csv",header=T,sep=";")
View(tempos_compostos)
```

Data: tempos_compostos

	tipo_de_verbo	verbo	auxiliar	num_aparicion
1	paso_de_tempo	durar	aver	3
2	paso_de_tempo	passar	ser	13
3	procesos_fisicos	(de)mudar	ser	5
4	procesos_fisicos	desecar	ser	1
5	procesos_fisicos	dormir	aver	3
6	procesos_fisicos	enloqueçer	ser	1
7	procesos_fisicos	escalentar	ser	1
8	procesos_fisicos	finar	ser	9
9	procesos_fisicos	guarir,guaresçer	ser	5
10	procesos_fisicos	morir	ser	82
11	procesos_fisicos	nasçer	ser	7
12	procesos_fisicos	parir	ser	1
13	suceso	acaeçer	ser	3
14	suceso	acaeçer	aver	3
15	suceso	conteçer	aver	4
16	suceso	passar	aver	3
17	permanencia	albergar	aver	1
18	permanencia	fincar	aver	9
19	permanencia	folgar	aver	1
20	permanencia	morar	aver	7
21	permanencia	posar	ser	1

Exercicio 1

Solución

1NT004916.txt

```
obra_demos<-read.table("1NT004916.txt",header=T)
View(obra_demos)
```

Data: obra_demos

	demostrativo	contaxe
1	este	6
2	estas	3
3	Esta	3
4	aquel	3
5	esa	2
6	Este	2
7	esta	2
8	eses	2
9	aquila	2
10	isto	1
11	ista	1
12	aquil	1
13	Estas	1
14	iso	1
15	aquela	1
16	aquelas	1
17	esto	1
18	esas	1
19	Ese	1

Módulo II – A información en R

II) Obxectos e estrutura da información

- i) Obxectos
- ii) A información estruturada en:
  - Vector
  - Matriz
  - Lista
  - Conxunto de datos

Módulo II – A información en R

II) Obxectos e estrutura da información

i) Obxectos

ii) A información estruturada en:

- Vector
- Matriz
- Lista
- Conxunto de datos

R project

I) Obxectos

R - Linguaxe de programación orientado a obxectos:

As variables, datos, resultados, funcións,... almacénanse na área de traballo mediante **obxectos** cun nome.

Asignación/creación de obxectos

- O **operador asignación** de valores é «=» ou «<-»

  Exemplos: a=2; a<-2
- O **nome dos obxectos** comezan por unha letra e poden conter números e símbolos (agás operadores aritméticos ou lóxicos)

  Exemplos: a_4=2; aBB<-3
- Sobre os obxectos poden actuar **funcións**

  Exemplos: a<-2+4

Recoñece maiúsculas e minúsculas

R project

I) Obxectos

- Para ver o **listado de obxectos** que temos creado:

  ls() (ou **objects()**)

- Para **borrar un obxecto**:

  rm(nombre obxecto)

  No exemplo anterior: **rm(a)**

- Para **borrar todos os obxectos** (da área de traballo):

  rm(list=ls())

- **Acceso ó contido dun obxecto:**

  Escribir o nome do obxecto:

  a<- 2+4

  a

R project

I) Obxectos

- **integer** : números enteiros (..,-2,-1,0,1,2,…) **a1<-4**
- **numeric** : números reais (1.2; 1.4; 2; ….) **a2<-1.2**
- **logical** : TRUE, FALSE **a3<-FALSE**
- **character** : Cadena de caracteres **a4<-"oso"**

Cada obxecto ten uns **atributos** que determinan as súas propiedades:

- Para ver o **tipo** de elementos dun obxecto:

  mode(obxecto) ou *class(obxecto)*

  Exemplo: mode(a2) ou class(a2)

  [1] "numeric"

- Para ver o **número** de elementos dun obxecto:

  length(obxecto)

  Exemplo: length(a2)

  [1] 1

Módulo II – A información en R

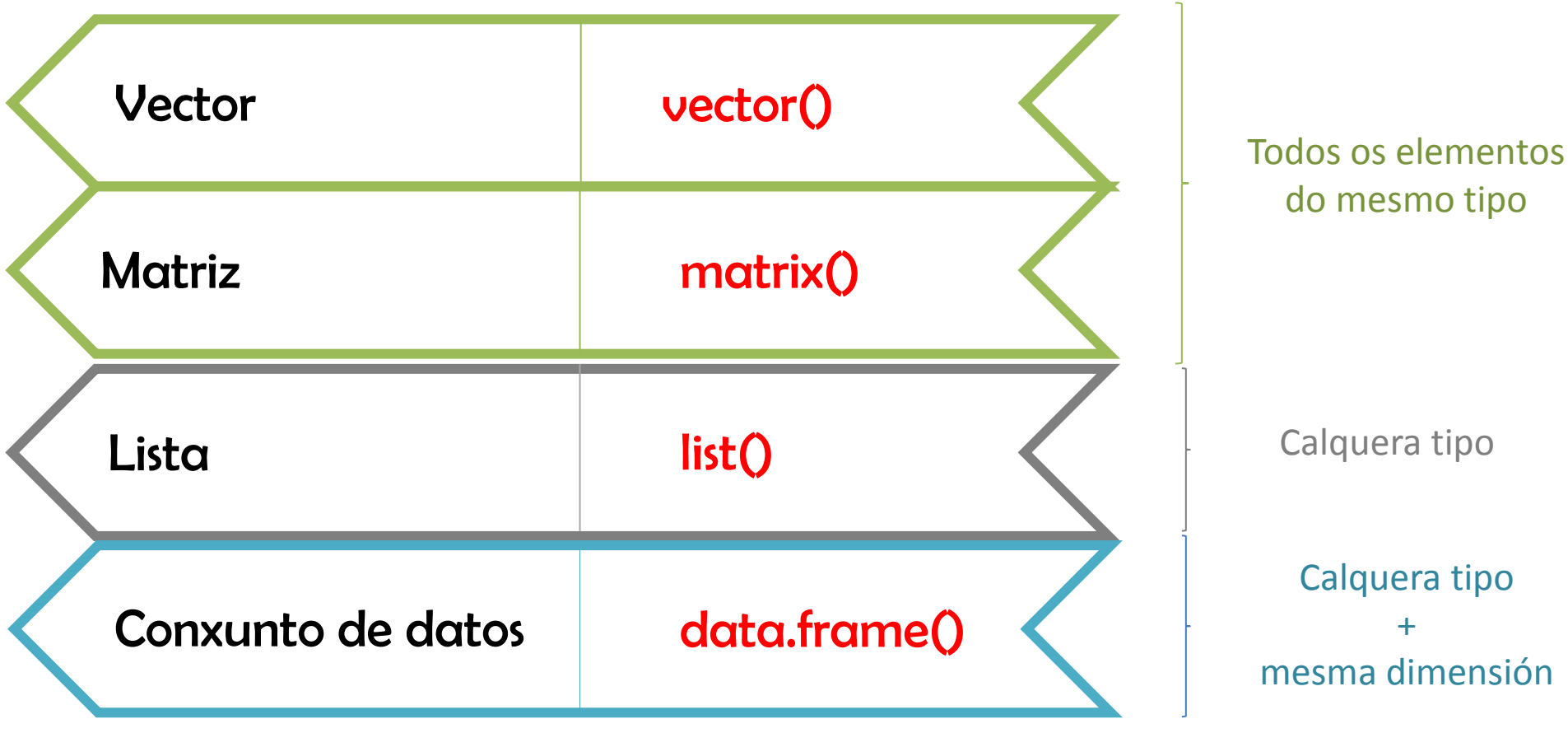
II) Obxectos e estrutura da información

- i) Obxectos
- ii) A información estruturada en:
  - Vector
  - Matriz
  - Lista
  - Conxunto de datos

R project

II) Estrutura da información

ALMACENAR MÁIS DUN VALOR



R project

II) Estrutura da información

Vector - Creación

Conxunto de elementos do mesmo tipo e dunha lonxitude determinada

1) Inicializando un vector: *vector()*

Axuda: ?vector

- **mode,class** : tipo de obxectos
- **length**: lonxitude do vector

Exemplos:

```
vector(mode="logical",length=2)
[1] FALSE FALSE
vector(mode="numeric",length=2)
[1] 0 0
vector(mode="character",length=2)
[1] "" ""
vector(mode="integer",length=2)
[1] 0 0
```

2) Concatenando elementos: *c()*

(o que se precisa normalmente)

Axuda: ?c

Exemplos:

```
a5<-c(TRUE,FALSE); a5; class(a5);length(a5)
[1]  TRUE FALSE
[1] "logical"
[1] 2
a6<-c(1.2,1.3); a6; mode(a6)
[1] 1.2 1.3
a7<-c("home","muller"); a7
[1] "home"   "muller"
a8<-c(3,4); a8
[1] 3 4
```

R project

II) Estrutura da información

Vector - Creación

Conxunto de elementos do mesmo tipo e dunha lonxitude determinada

3) Repetindo elementos: *rep()*

Axuda: ?rep

rep(x, nº de veces)

- x= un ou varios elementos

- nº de veces= un número ou un vector onde se defina o nº de veces que se repite cada número

Exemplos:

```
a9<- rep(2,4); a9
[1] 2 2 2 2
a10<- rep(2:5,3) ; a10
 [1] 2 3 4 5 2 3 4 5 2 3 4 5
a11<- rep(2:5,each=3) ; a11
 [1] 2 2 2 3 3 3 4 4 4 5 5 5
a12<- rep(2:5,c(2,1,4,1)) ; a12
[1] 2 2 3 4 4 4 4 5
```

4) Secuencia de elementos: *seq()*

Axuda: ?seq

seq(from, to, by)

-from= desde

-to = ata

-by = lonxitude do paso

Exemplos:

```
a13<- seq(0,1,by=0.1); a13
 [1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
a14<- 1:10; a14
 [1]  1  2  3  4  5  6  7  8  9 10
a15<- seq(10); a15
[1]  1  2  3  4  5  6  7  8  9 10
```

Exercicio 2

Como podemos gardar esta información?

a) **Realizouse un estudo no que se tivo que entrevistar a persoas de:**

carballo, sarria, arteixo, arteixo, carballo, ponteareas, arteixo, carballo, carballo, boiro, sarria, noia, noia, cangas, noia, sarria, boiro, boiro, sarria

Crea un obxecto (neste caso, estamos definindo unha variable) que se chame «lugar» no que se almacene esta información.

b) **A cada unha delas preguntóuselles cantas linguas falaba, e obtivemos as seguintes respostas:**

1, 2, 2, 1, 3, 2, 4, 3, 2, 3, 2, 2, 2, 2, 5, 2, 3, 2, 4

Crea un obxecto (neste caso, estamos definindo unha variable) que se chame «linguasfaladas» no que se almacene esta información.

Exercicio 2
Solución

a) Crea un obxecto (neste caso, estamos definindo unha variable) que se chame «lugar» no que se almacene esta información:

```
lugar<-c("carballo", "sarria", "arteixo", "arteixo", "carballo", "ponteareas",
                    "arteixo", "carballo", "carballo", "boiro", "sarria", "noia",
                   "noia", "cangas", "noia", "sarria", "boiro", "boiro", "sarria")
lugar
[1] "carballo"   "sarria"    "arteixo"   "arteixo"   "carballo"
 [6] "ponteareas" "arteixo"   "carballo"  "carballo"  "boiro"
[11] "sarria"    "noia"      "noia"      "cangas"    "noia"
[16] "sarria"    "boiro"     "boiro"     "sarria"
class(lugar)
[1] "character"
length(lugar)
[1] 19
```

Exercicio 2
Solución

b) Crea un obxecto (neste caso, estamos definindo unha variable) que se chame «linguasfaladas» no que se almacene esta información:

```
linguasfaladas=c(1, 2, 2, 1, 3, 2, 4, 3, 2, 3, 2, 2, 2, 2, 5, 2, 3, 2, 4)
linguasfaladas
 [1] 1 2 2 1 3 2 4 3 2 3 2 2 2 2 5 2 3 2 4
 class(linguasfaladas)
[1] "numeric"
length(linguasfaladas)
[1] 19
```

R project

II) Estrutura da información

Exercicio 3

Como podemos gardar esta información?

Realizouse un estudo no que se quixo observar o número de persoas que saben falar 6 linguas estranxeiras segundo o país de procedencia:

País de procedencia	Número de persoas
Finlandia	103
Francia	35
España	23
Portugal	24
Italia	20

Crea un obxecto que se chame «pais» onde se garden estes datos.

Exercicio 3

Solución

Crea un obxecto que se chame «pais» onde se garden estes datos

```
pais=c(rep("Finlandia",103),rep("Francia",35),rep("Espana",23),rep("Portugal",25))
pais
```

```
 [1] "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia"
 [16] "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia"
 [31] "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia"
 [46] "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia"
 [61] "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia"
 [76] "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia"
 [91] "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Finlandia" "Francia"  "Francia"
[106] "Francia"  "Francia"  "Francia"  "Francia"  "Francia"  "Francia"  "Francia"  "Francia"  "Francia"  "Francia"  "Francia"  "Francia"  "Francia"  "Francia"  "Francia"
[121] "Francia"  "Francia"  "Francia"  "Francia"  "Francia"  "Francia"  "Francia"  "Francia"  "Francia"  "Francia"  "Francia"  "Francia"  "Francia"  "Francia"  "Francia"
[136] "Francia"  "Francia"  "Francia"  "Espana"  "Espana"  "Espana"  "Espana"  "Espana"  "Espana"  "Espana"  "Espana"  "Espana"  "Espana"  "Espana"  "Espana"
[151] "Espana"  "Espana"  "Espana"  "Espana"  "Espana"  "Espana"  "Espana"  "Espana"  "Espana"  "Espana"  "Espana"  "Portugal" "Portugal" "Portugal" "Portugal"
[166] "Portugal" "Portugal" "Portugal" "Portugal" "Portugal" "Portugal" "Portugal" "Portugal" "Portugal" "Portugal" "Portugal" "Portugal" "Portugal" "Portugal" "Portugal"
[181] "Portugal" "Portugal" "Portugal" "Portugal" "Portugal" "Italia"  "Italia"  "Italia"  "Italia"  "Italia"  "Italia"  "Italia"  "Italia"  "Italia"  "Italia"
[196] "Italia"  "Italia"  "Italia"  "Italia"  "Italia"  "Italia"  "Italia"  "Italia"  "Italia"  "Italia"
```

```
class(pais)
```

```
[1] "character"
```

```
length(pais)
```

```
[1] 205
```

Vector - Tipo de elementos

is. vector() :para comprobar se é vector

Exemplo

```
is.vector(lugar)
```

OLLO! Todos os elementos dun vector teñen que ser do mesmo tipo

Aínda que R nos permite escribir diferentes tipos de elementos nun vector, este ó final almacénaos do mesmo tipo

Exemplo

```
a16 <- c(2,TRUE, "sandra"); a16
[1] "2"     "TRUE"   "sandra"
class(a16)
[1] "character"
```

Vector - Acceso

Como acceder a unha ou varias compoñentes do vector

Tiñamos definido un obxecto «**lugar**» :

```
lugar<-c("carballo", "sarria", "arteixo", "arteixo", "carballo", "ponteareas", "arteixo", "carballo", "carballo",
"boiro", "sarria", "noia", "noia", "cangas", "noia", "sarria", "boiro", "boiro", "sarria")
```

```
lugar[2]                # consultar unha das compoñentes utilizando a súa posición
[1] "sarria''
```

```
lugar[-2]               # se queremos sacar unha das compoñentes (un dato) utilizando a súa posición
[1] "carballo"  "arteixo"  "arteixo"  "carballo"  "ponteareas" "arteixo"  "carballo"  "carballo"  "boiro"
[10] "sarria"   "noia"     "noia"     "cangas"    "noia"     "sarria"    "boiro"     "boiro"     "sarria"
```

```
lugar[2:5]              # consultar varias compoñentes consecutivas
[1] "sarria"   "arteixo"  "arteixo"  "carballo“
```

```
lugar[c(1,3,7)]         # consultar compoñentes alternadas
[1] "carballo" "arteixo"  "arteixo"
```

R project

II) Estrutura da información

Vector - Acceso

Como acceder ou coñecer as compoñentes que cumpren unha condición lóxica

***which()* : permite coñecer as posicións nun obxecto**

Tiñamos definido un obxecto «**lugar**» :

```
lugar<-c("carballo", "sarria", "arteixo", "arteixo", "carballo", "ponteareas", "arteixo", "carballo", "carballo",
"boiro", "sarria", "noia", "noia", "cangas", "noia", "sarria", "boiro", "boiro", "sarria")
```

Quérese ver cal é a posición na que cadra «sarria» no noso obxecto «lugar»:

```
lugar=="sarria"
 [1] FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[11] TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE
which(lugar=="sarria")
[1]  2 11 16 19
lugar[which(lugar=="sarria")]
[1] "sarria" "sarria" "sarria" "sarria"
which(lugar=="sarria"|lugar=="carballo")
[1]  1  2  5  8  9 11 16 19
```

Condicións lóxicas:

«igual a» : ==
«distinto de» : !=
«menor ou igual que, ≤» : <=
«menor que, <» : <
«maior ou igual que, ≥» : >=
«maior que, >» : >
«e» : &
«ou» : |

R project

II) Estrutura da información

Vector - Acceso

Como acceder ou coñecer as compoñentes que cumpren unha condición lóxica

***which()* : permite coñecer as posicións nun obxecto**

Quérese ver cal é a posición na que cadra «sarria» no noso obxecto «lugar»:

```
lugar!="sarria"
[1]  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
[13]  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE
which(lugar!="sarria")
 [1]  1  3  4  5  6  7  8  9 10 12 13 14 15 17 18
lugar[which(lugar!="sarria")]
 [1] "carballo"   "arteixo"    "arteixo"    "carballo"   "pontеareas"
 [6] "arteixo"    "carballo"   "carballo"   "boiro"      "noia"
[11] "noia"       "cangas"     "noia"       "boiro"      "boiro"
```

Condicións lóxicas:

«igual a» : ==
«distinto de» : !=
«menor ou igual que, ≤» : <=
«menor que, <» : <
«maior ou igual que, ≥» : >=
«maior que, >» : >
«e» : &
«ou» : |

R project

II) Estrutura da información

Vector - Acceso

Como acceder ou coñecer as compoñentes que cumpren unha condición lóxica

***which()* : permite coñecer as posicións nun obxecto**

Definamos un novo obxecto numérico:

```
a17=c(1:10, rep(8,3),rep(9,5), 6:14); a17
[1]  1  2  3  4  5  6  7  8  9 10  8  8  8  9  9  9  9  9  6  7  8  9 10 11 12 13 14
```

```
a17<=8
[1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE
[11]  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE
[21]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
which(a17<=8)
[1]  1  2  3  4  5  6  7  8 11 12 13 19 20 21
a17[which(a17<=8)]
[1] 1 2 3 4 5 6 7 8 8 8 8 6 7 8
```

Condicións lóxicas:

«igual a» : ==
«distinto de» : !=
«menor ou igual que, ≤» : <=
«menor que, <» : <
«maior ou igual que, ≥» : >=
«maior que, >» : >
«e» : &
«ou» : |

Exercicio 4

Como podemos seleccionar información?

Creamos un vector:

exercicio3=(1,2,3,4,3,2,1,5,1,2,3,4,5,6,7,8,9,10)

a) De que tipo é o dito obxecto?
b) Que lonxitude ten?
c) Ver en que posición toma o valor 1
d) Ver en que posicións toma un valor maior que 5
e) Ver en que posicións toma un valor menor que 4
f) Ver en que posicións toma valores distintos de 3 e de 4

R project

II) Estrutura da información

Exercicio 4

Solución

Creamos un vector:

```
exercicio3=c(1,2,3,4,3,2,1,5,1:10)
```

a) **De que tipo é o dito obxecto?**

```
class(exercicio3)
[1] "numeric"
```

b) **Que lonxitude ten?**

```
length(exercicio3)
[1] 18
```

c) **Ver que posicións toma o valor 1**

```
which(exercicio3==1)
[1] 1 7 9
```

d) **Ver en que posicións toma un valor maior que 5**

```
which(exercicio3>5)
[1] 14 15 16 17 18
```

e) **Ver en que posicións toma un valor menor que 4**

```
which(exercicio3<4)
[1]  1  2  3  5  6  7  9 10 11
```

f) **Ver en que posicións toma valores distintos de 3 e de 4**

```
which(exercicio3!=3&exercicio3!=4)
[1]  1  2  6  7  8  9 10 13 14 15 16 17 18
```

R project

II) Estrutura da información

Matriz

Conxunto de elementos do mesmo tipo estruturado en filas e columnas

Exemplo de matriz:

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{pmatrix} = (a_{ij})_{3x4}$$

Como definila en R?

? matrix

matrix(data = NA, nrow = 1, ncol = 1, byrow = FALSE, dimnames = NULL)

- Os datos como vector
- Número de filas
- Número de columnas
- Como entran os datos? Opcións:
  - Por filas
  - Por columnas
- Vector que dá nome ás filas e ás columnas

Matriz - Creación

Conxunto de elementos do mesmo tipo estruturado en filas e columnas

```
m1=matrix(1:8, nrow=2,ncol=4,byrow=F); m1
     [,1] [,2] [,3] [,4]
[1,]    1    3    5    7
[2,]    2    4    6    8
dim(m1)
[1] 2 4
```

Coñecemos as dimensións da matriz co comando dim(), isto é, as filas e as columnas

```
 m2=matrix(1:8, nrow=2,ncol=4,byrow=T); m2
      [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]    5    6    7    8
```

```
m3=matrix(1:8, nrow=2,ncol=4,dimnames=list(c("fila1","fila2"),c("col1","col2","col3","col4"))); m3
      col1 col2 col3 col4
fila1    1    3    5    7
fila2    2    4    6    8
```

R project

II) Estrutura da información

Exercicio 4

Como podemos gardar esta información?

	PUNTO	MODO
p	BILABIAL	OCLUSIVA
t	DENTAL	OCLUSIVA
m	BILABIAL	NASAL
n	ALVEOLAR	NASAL

a) **Garda esta información nun obxecto chamado «fonemas». Cal é a dimensión da nosa matriz?**

Exercicio 4
Solución

a) Garda esta información nun obxecto chamado «fonemas». Cal é a dimensión da nosa matriz?

```
fonemas<-matrix(c("bilabial","dental","bilabial","alveolar","oclusiva","oclusiva","nasal","nasal"),
nrow=4,ncol=2,byrow=F,dimnames=list(c("p","t","m","n"),c("PUNTO","MODO"))); fonemas
     PUNTO      MODO
p  "bilabial"  "oclusiva"
t   "dental"    "oclusiva"
m "bilabial"  "nasal"
n  "alveolar" "nasal"

dim(fonemas)
[1] 4 2
```

Matriz - Creación

Conxunto de elementos do mesmo tipo estruturado en filas e columnas

Outra forma, por **concatenación**:

```
x<-1:4 ; x
[1] 1 2 3 4
y<-5:8 ; y
[1] 5 6 7 8
```

← Definimos os vectores que queremos unir

```
m4=rbind(x,y);m4
  [,1] [,2] [,3] [,4]
x    1    2    3    4
y    5    6    7    8
```

← Con *rbind()* xuntamos os elementos por filas

```
 m5=cbind(x,y);m5
      x y
[1,]  1 5
[2,]  2 6
[3,]  3 7
[4,]  4 8
```

← Con *cbind()* xuntamos os elementos por columnas

R project

II) Estrutura da información

Exercicio 5

Como podemos gardar esta información?

	PUNTO	MODO
p	BILABIAL	OCLUSIVA
t	DENTAL	OCLUSIVA
m	BILABIAL	NASAL
n	ALVEOLAR	NASAL

a) **Garda esta información nun obxecto chamado «fonemas». Cal é a dimensión da nosa matriz?**
b) **Proba a gardar de novo a información concatenando un vector que conteña os puntos de articulación e outro vector que conteña o modo de articulación**

Exercicio 5
Solución

b) Proba a gardar de novo a información concatenando un vector que conteña os puntos de articulación e outro vector que conteña o modo de articulación

```
punto<-c("bilabial","dental","bilabial","alveolar")
modo<-c("oclusiva","oclusiva","nasal","nasal")
fonemas2<-cbind(punto,modo)
fonemas2
      punto      modo
[1,] "bilabial"   "oclusiva"
[2,] "dental"     "oclusiva"
[3,] "bilabial"    "nasal"
[4,] "alveolar"  "nasal"
```

Matriz - Acceso

Como acceder a unha ou varias compoñentes dunha matriz

Agora teremos que indicar a fila e a columna á que queremos acceder:

matriz[nº fila, nºcolumna]

Exemplo:

```
fonemas[1,]          #accedemos á primeira fila
 PUNTO      MODO
"bilabial" "oclusiva"
fonemas[,2]          #accedemos á segunda columna
     p             t             m          n
"oclusiva" "oclusiva"    "nasal"    "nasal"
fonemas[3,2]         #accedemos á terceira fila e á segunda columna
[1] "nasal"
fonemas[c(2,3),1]    #consultar o punto de articulación do 2º e do 3º rexistro
       t        m
 "dental" "bilabial"
```

Lista

Conxunto de elementos de diferente tipo

Colección de obxectos que convén agrupar por algún tipo de razón.

Estes obxectos poden ser de **diferente tipo** de ter características distintas:

Para **definir unha lista:** ***list()***

```
clasificacion<-list(fonemas=cbind(c("bilabial","dental","bilabial","alveolar"),c("oclusiva","oclusiva",
"nasal","nasal")),letras=c("vogais","consoantes"))
clasificacion
$fonemas
     [,1]      [,2]
[1,] "bilabial" "oclusiva"
[2,] "dental"   "oclusiva"
[3,] "bilabial" "nasal"
[4,] "alveolar" "nasal"
$letras
[1] "vogais"     "consoantes"
```

R project

II) Estrutura da información

Importante!

Conxunto de datos

Ficheiro ou base de datos

Información estruturada en filas e columnas:

- As **filas** son os **rexistros**
- As **columnas** son as **variables** (as características que se miden)

Variable 1 | Variable 2 | Variable 3

	A	B	C	D	E	F	G	H	I	J
1	id_xuiz	data_intento	Idade	Sexo	Concello da vivenda familiar	NATIVO	Lingua materna	Lingua habitual	Lingua materna da nai	Lingua materna do pai
2	id_1	07/10/2014	25	Muller	ORDES	bilingüe	Máis galego c	Máis galego ca c	Máis galego ca cas	Máis galego ca c
3	id_2	08/10/2014	31	Home	VILABOA	castelán	Só castelán	só galego	Só castelán	Só castelán
4	id_3	17/10/2014	27	Home	PONTEVEDRA	castelán	Só castelán	Máis galego ca c	Só galego	Só galego
5	id_4	08/10/2014	21	Muller	GUARDA, A	bilingüe	Máis castelán	Máis galego ca c	Só castelán	só galego
6	id_5	06/10/2014	20	Home	VIGO	bilingüe	Máis castelán	Máis galego ca c	só galego	só galego
7	id_6	06/10/2014	23	Muller	ORDES	galego	Só galego	só galego	só galego	só galego
8	id_7	15/10/2014	23	Home	LEIRO	bilingüe	Máis galego c	Máis galego ca c	só galego	só galego
9	id_8	14/10/2014	34	Home	VIGO	bilingüe	Máis castelán	só galego	Máis castelán ca g	Máis galego ca c
10	id_9	08/10/2014	18	Muller	SANXENXO	castelán	Só castelán	Máis castelán ca	Só castelán	Só castelán
11	id_10	06/10/2014	23	Muller	PONTE CALDELAS	galego	Só galego	só galego	só galego	só galego
12	id_11	07/10/2014	25	Home	AMES	bilingüe	Máis castelán	Máis castelán ca	Máis galego ca cas	Máis castelán ca
13	id_12	07/10/2014	21	Home	TEO	bilingüe	Máis galego c	Máis galego ca c	Máis galego ca cas	Máis galego ca c
14	id_13	15/10/2014	25	Home	OURENSE	castelán	Máis castelán	Máis galego ca c	Só castelán	Máis castelán ca

Informante 3

Informante 5

R project

II) Estrutura da información

Importante!

Conxunto de datos - Creación

Ficheiro ou base de datos

```
idade=c(25,31,27,21,20,23,23,34,18)
nativo=c("bilingue","castelan","castelan","bilingue","bilingue","galego","bilingue","bilingue","castelan")
```

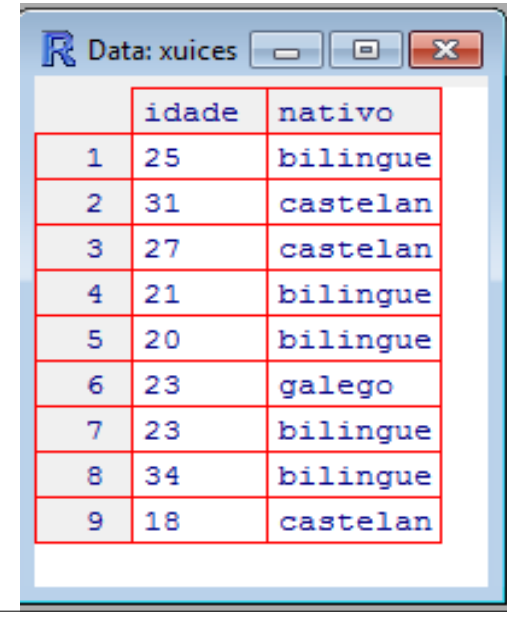
Para **definir un conxunto de datos:** ***data.frame()***

```
xuices=data.frame(idade,nativo)
xuices
```

	idade	nativo
1	25	bilingue
2	31	castelan
3	27	castelán
4	21	bilingue
5	20	bilingue
6	23	galego
7	23	bilingue
8	34	bilingue
9	18	castelan

Para visualizar os datos (só vale con obxectos data.frame)

```
View(xuices)
```



	idade	nativo
1	25	bilingue
2	31	castelan
3	27	castelan
4	21	bilingue
5	20	bilingue
6	23	galego
7	23	bilingue
8	34	bilingue
9	18	castelan

R project

II) Estrutura da información

Importante!

Conxunto de datos - Lectura

Ficheiro ou base de datos

Imos utilizar a base «xuices.csv»

```
xuices2<-read.csv("xuices.csv",header=T,sep=";")
View(xuices2)
class(xuices2)
[1] "data.frame"
```

Os obxectos creados como lectura dun arquivo de datos xa son clasificados como data.frame

Data: xuices2

	id_xuiz	Idade	Sexo	Concello_da_vivenda_familiar	NATIVO
1	id_1	25	Muller	ORDES	bilingue
2	id_2	31	Home	VILABOA	castelan
3	id_3	27	Home	PONTEVEDRA	castelan
4	id_4	21	Muller	GUARDA_A	bilingue
5	id_5	20	Home	VIGO	bilingue
6	id_6	23	Muller	ORDES	galego
7	id_7	23	Home	LEIRO	bilingue
8	id_8	34	Home	VIGO	bilingue
9	id_9	18	Muller	SANXENXO	castelan
10	id_10	23	Muller	PONTE_CALDELAS	galego
11	id_11	25	Home	AMES	bilingue
12	id_12	21	Home	TEO	bilingue
13	id_13	25	Home	OURENSE	castelan
14	id_14	19	Muller	SANTIAGO_DE_COMPOSTELA	galego
15	id_15	28	Muller	VILAGARCIA_DE_AROUSA	bilingue
16	id_16	31	Home	CORUnA_A	castelan
17	id_17	24	Muller	CERCEDA	galego
18	id_18	28	Home	LUGO	castelan
19	id_19	25	Home	TRAZO	bilingue

R project

II) Estrutura da información

Importante!

Conxunto de datos - Acceso

Consultas en data.frame()

As consultas realizaranse da mesma forma que nas matrices:

obxecto[nº fila, nº col]

Exemplos:

xuices2[19,] #xuices2[nºfila,]

```
   id_xuiz Idade Sexo Concello_da_vivenda_familiar   NATIVO Lingua_materna         Lingua_habitual Lingua_materna_da_nai
19   id_19    25 Home                          TRAZO bilingue      So_galego Mais_galego_ca_castelan             so_galego
   Lingua_materna_do_pai Concello_de_residencia_habitual Anos_residindo_nese_concello          Estudos_musicais
19             so_galego                          TRAZO                Entre_5_e_10 Non_teno_estudos_musicais
                    Estudos Area_estudo residencia_fora Meses_residindo_fora
19 graduado_en_traballo_social      NONHUM           Berna                   60
```

xuices2[,4] #xuices2[,nºcol]

```
 [1] ORDES                VILABOA              PONTEVEDRA           GUARDA_A             VIGO
 [6] ORDES                LEIRO                VIGO                 SANXENXO             PONTE_CALDELAS
[11] AMES                 TEO                  OURENSE              SANTIAGO_DE_COMPOSTELA VILAGARCIA_DE_AROUSA
[16] CORUnA_A             CERCEDA              LUGO                 TRAZO                VILARMAIOR
[21] LEIRO                CANGAS               CORUnA_A             PEREIRO_DE_AGUIAR_O  VILA_DE_CRUCES
[26] AMES                 SANTIAGO_DE_COMPOSTELA NIGRaN             LUGO                 CHANTADA
[31] XINZO_DE_LIMIA       RIBADAVIA            SANTIAGO_DE_COMPOSTELA TORDOIA
26 Levels: AMES CANGAS CERCEDA CHANTADA CORUnA A GUARDA A LEIRO LUGO NIGRaN ORDES OURENSE PEREIRO DE AGUIAR O ... XINZO DE LIMIA
```

xuices2[1,3] #xuices2[nºfila,nºcol]

```
[1] Muller
Levels: Home Muller
```

R project

II) Estrutura da información

Importante!

Conxunto de datos - Acceso

Consultas en data.frame()

As consultas baixo certas condicións: ***which()***

Exemplos:

```
posicions<-which(xuices2[,4]=="SANTIAGO_DE_COMPOSTELA");posicions
[1] 14 27 33
```

```
xuices2[posicions,]
```

	id_xuiz	Idade	Sexo	Concello_da_vivenda_familiar	NATIVO	Lingua_materna	Lingua_habitual	Lingua_materna_da_nai
14	id_14	19	Muller	SANTIAGO_DE_COMPOSTELA	galego	So_galego	so_galego	so_galego
27	id_27	21	Home	SANTIAGO_DE_COMPOSTELA	castelan	So_castelan	Mais_galego_ca_castelan	So_castelan
33	id_33	25	Home	SANTIAGO_DE_COMPOSTELA	galego	So_galego	so_galego	Mais_galego_ca_castelan

	Lingua_materna_do_pai	Concello_de_residencia_habitual	Anos_residindo_nese_concello	Estudos_musicais
14	so_galego	SANTIAGO_DE_COMPOSTELA	Mais_de_15	Estudos_medios
27	So_castelan	SANTIAGO_DE_COMPOSTELA	Entre_5_e_10	Estudos_elementais
33	so_galego	SANTIAGO_DE_COMPOSTELA	Mais_de_15	Estudos_medios

	Estudos	Area_estudo	residencia_fora	Meses_residindo_fora
14	grao_en_filoloxia_inglesa	HUM	0	0
27	grao_lingua_e_literatura_galega	HUM	Arxentina	108
33	Administracion_e_direccion_de_empresas	NONHUM	0	0

R project

II) Estrutura da información

Exercicio 6

Anteriormente traballamos coa base "tempos_compostos_galego_medieval.csv".

Imos facer algunhas consultas nela…

a) **Extrae a información do cuarto rexistro.**

b) **Consulta a variable "auxiliar". Poderías dicir automaticamente cantos tipos de verbos auxiliares temos?**

c) **Imos consultar só un tipo de verbos, os verbos de tipo "suceso".**

   i. **Extrae toda a información dos verbos deste tipo.**
   ii. **Fai unha consulta máis específica extraendo só na consola os verbos que se clasifican como de tipo "suceso".**

Exercicio 6
Solución

a) **Extrae a información do cuarto rexistro.**

tempos_compostos[4,]

```
     tipo_de_verbo   verbo auxiliar num_aparicion
4 procesos_fisicos desecar      ser             1
```

a) **Consulta a variable "auxiliar". Poderías dicir automaticamente cantos tipos de verbos auxiliares temos?**

auxiliar

```
 [1] aver ser  ser  ser  aver ser  ser  ser  ser  ser  ser  ser  aver aver aver aver aver aver
[20] aver ser
Levels: aver ser
```

Exercicio 6
Solución

c) Imos consultar só un tipo de verbos, os verbos de tipo “suceso”.

i. Extrae toda a información dos verbos deste tipo.

```
posicions_suceso<-which(tipo_de_verbo=="suceso");posicions_suceso
[1] 13 14 15 16
tempos_compostos[posicions_suceso,]
   tipo_de_verbo    verbo auxiliar num_aparicion
13        suceso  acaeçer      ser             3
14        suceso  acaeçer     aver             3
15        suceso conteçer     aver             4
16        suceso   passar     aver             3
```

ii. Fai unha consulta máis específica extraendo só na consola os verbos que se clasifican como de tipo “suceso”.

```
tempos_compostos[posicions_suceso,2]
[1] acaeçer  acaeçer  conteçer passar
```

Módulo II – A información en R

III) Escritura/Exportación de datos

R project

III) Exportación de datos

Como «escribir» ficheiros de datos desde R?

Exportar obxectos **data.frame a** un **ficheiro .txt**: ***write.table()***

R project

III) Exportación de datos

Como «escribir» ficheiros de datos desde R?

Exportar obxectos **data.frame a** un **ficheiro .txt**: ***write.table()***

Imos crear un data.frame:
Tras facer a consulta no INE observamos que...

Resultados por provincia de residencia
Apellido: CALAZA

Provincia	Apellido 1º
	Total
Total	215
Araba/Álava	5
Barcelona	12
Bizkaia	11
Córdoba	..
Coruña, A	66
Huelva	5
Lugo	66
Madrid	32
Pontevedra	9

Resultados por provincia de residencia
Apellido: BEIS

Provincia	Apellido 1º
	Total
Total	98
Barcelona	..
Bizkaia	5
Coruña, A	62
Pontevedra	17

Queremos gardar estes datos xunto coa **clasificación** dos mesmos:

«Calaza» : delexical
«Beis» : toponímico

R project

III) Exportación de datos

Como «escribir» ficheiros de datos desde R?

Exportar obxectos **data.frame a** un **ficheiro .txt**: ***write.table()***

Exercicio para pensar

Teremos que construír un obxecto que teña tres columnas, unha contendo o apelido, outra a provincia e outra a clasificación correspondente...

R project

III) Exportación de datos

Como «escribir» ficheiros de datos desde R?

Exportar obxectos **data.frame a** un **ficheiro .txt**: ***write.table()***

Exercicio para pensar

Teremos que construír un obxecto que teña tres columnas, unha contendo o apelido, outra a provincia e outra a clasificación correspondente...

Unha pequena mostra...

Data: antroponimia

	apelidos	provincia	clasificacion
200	Calaza	Pontevedra	delexical
201	Calaza	Pontevedra	delexical
202	Calaza	Pontevedra	delexical
203	Calaza	Pontevedra	delexical
204	Calaza	Pontevedra	delexical
205	Calaza	Pontevedra	delexical
206	Calaza	Pontevedra	delexical
207	Beis	Bizcaia	toponimico
208	Beis	Bizcaia	toponimico
209	Beis	Bizcaia	toponimico
210	Beis	Bizcaia	toponimico
211	Beis	Bizcaia	toponimico
212	Beis	A_Coruna	toponimico
213	Beis	A_Coruna	toponimico
214	Beis	A_Coruna	toponimico

R project

III) Exportación de datos

Como «escribir» ficheiros de datos desde R?

Exportar obxectos **data.frame a** un **ficheiro .txt**: ***write.table()***

```
apelido_calaza=c(rep("Alava",5),rep("Barcelona",12),rep("Bizkaia",11),rep("A_Coruna",66),rep("Huelva",5),rep("Lugo",66),
                 rep("Madrid",32),rep("Pontevedra",9))
apelido_beis=c(rep("Bizcaia",5),rep("A_Coruna",62),rep("Pontevedra",17))

apelidos=c(rep("Calaza",length(apelido_calaza)),rep("Beis",length(apelido_beis)))
provincia=c(apelido_calaza,apelido_beis)
clasificacion=c(rep("delexical",length(apelido_calaza)),rep("toponimico",length(apelido_beis)))

antroponimia=data.frame(apelidos,provincia,clasificacion)
```

R project

III) Exportación de datos

Como «escribir» ficheiros de datos desde R?

Exportar obxectos **data.frame a** un **ficheiro .txt**: ***write.table()***

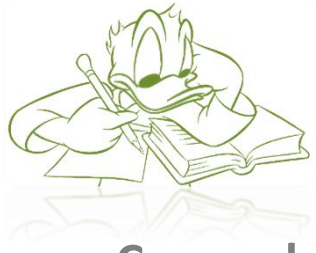
```
apelido_calaza=c(rep("Alava",5),rep("Barcelona",12),rep("Bizkaia",11),rep("A_Coruna",66),rep("Huelva",5),rep("Lugo",66),
                 rep("Madrid",32),rep("Pontevedra",9))
apelido_beis=c(rep("Bizcaia",5),rep("A_Coruna",62),rep("Pontevedra",17))

apelidos=c(rep("Calaza",length(apelido_calaza)),rep("Beis",length(apelido_beis)))
provincia=c(apelido_calaza,apelido_beis)
clasificacion=c(rep("delexical",length(apelido_calaza)),rep("toponimico",length(apelido_beis)))

antroponimia=data.frame(apelidos,provincia,clasificacion)
```

write.table(obxecto, "nome_do_arquivo")

write.table(antroponimia,"antroponimia")

R project

III) Exportación de datos



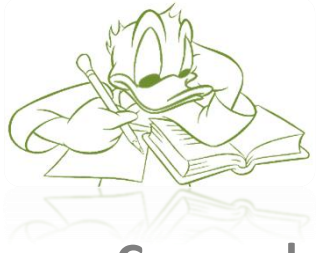
Exercicio 7

Crea unha base de datos desde R project que conteña a seguinte información, e posteriormente gárdaa como un arquivo .txt:

obra	clasificacion_palabras	frecuencia
ZAPINE953	N	662
ZAPINE953	ADV	157
ZAPINE953	DET	196
ZAPINE953	CONX	186
ZAPDIS951	N	120
ZAPDIS951	ADV	38
ZAPDIS951	DET	43
ZAPDIS951	CONX	31

R project

III) Exportación de datos



Exercicio 7
Solución

Crea unha base de datos desde R project que conteña a seguinte información, e posteriormente gárdaa como un arquivo .txt:

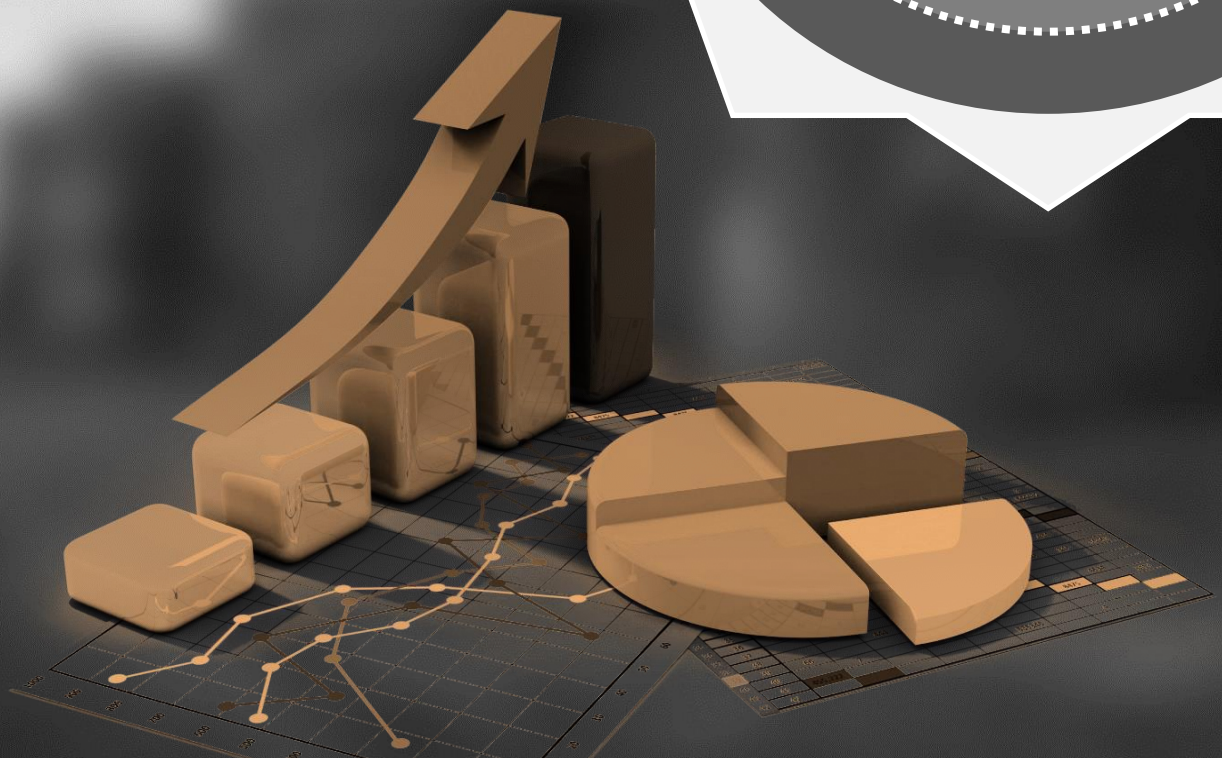
```
obra<-c(rep("ZAPINE953",4),rep("ZAPDIS951",4))
clasificacion_palabras<-c(rep(c("N","ADV","DET","CONX"),2))
frecuencias<-c(662,157,196,186,120,38,43,31)
base_obras<-data.frame(obra,clasificacion_palabras,frecuencias)
base_obras
```

```
       obra clasificacion_palabras frecuencias
1 ZAPINE953                     N         662
2 ZAPINE953                   ADV         157
3 ZAPINE953                   DET         196
4 ZAPINE953                  CONX         186
5 ZAPDIS951                     N         120
6 ZAPDIS951                   ADV          38
7 ZAPDIS951                   DET          43
8 ZAPDIS951                  CONX          31
```

```
write.table(base_obras,"base_obras.txt")
```

Módulo III - Estatística

I) Introdución. Conceptos básicos

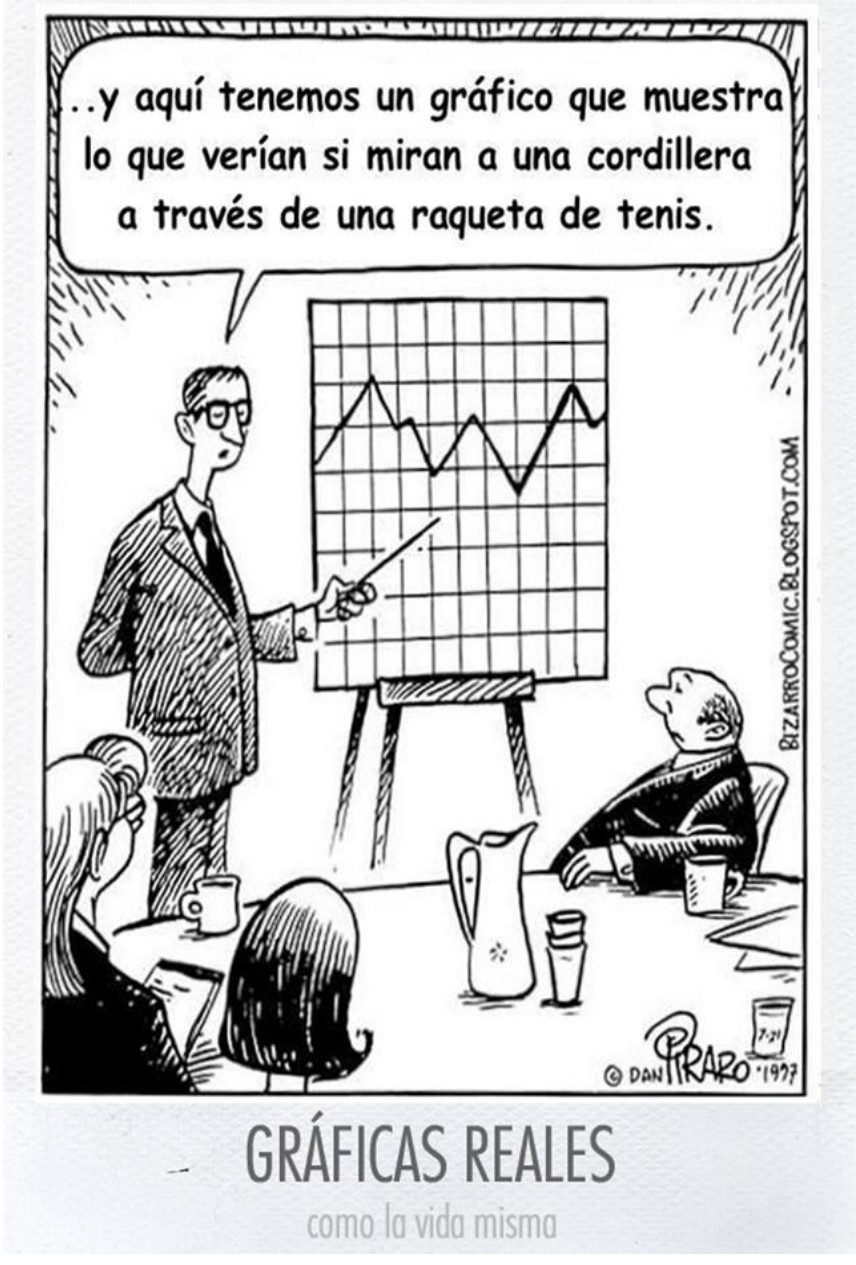


Estatística

I) Introdución

O propósito da estatística...

recompilar, organizar, presentar, resumir, analizar, interpretar e usar os datos para tomar decisións e resolver problemas



Estatística

I) Introdución

O propósito da estatística...

recompilar, organizar, presentar, resumir, analizar, interpretar e usar os datos para tomar decisións e resolver problemas

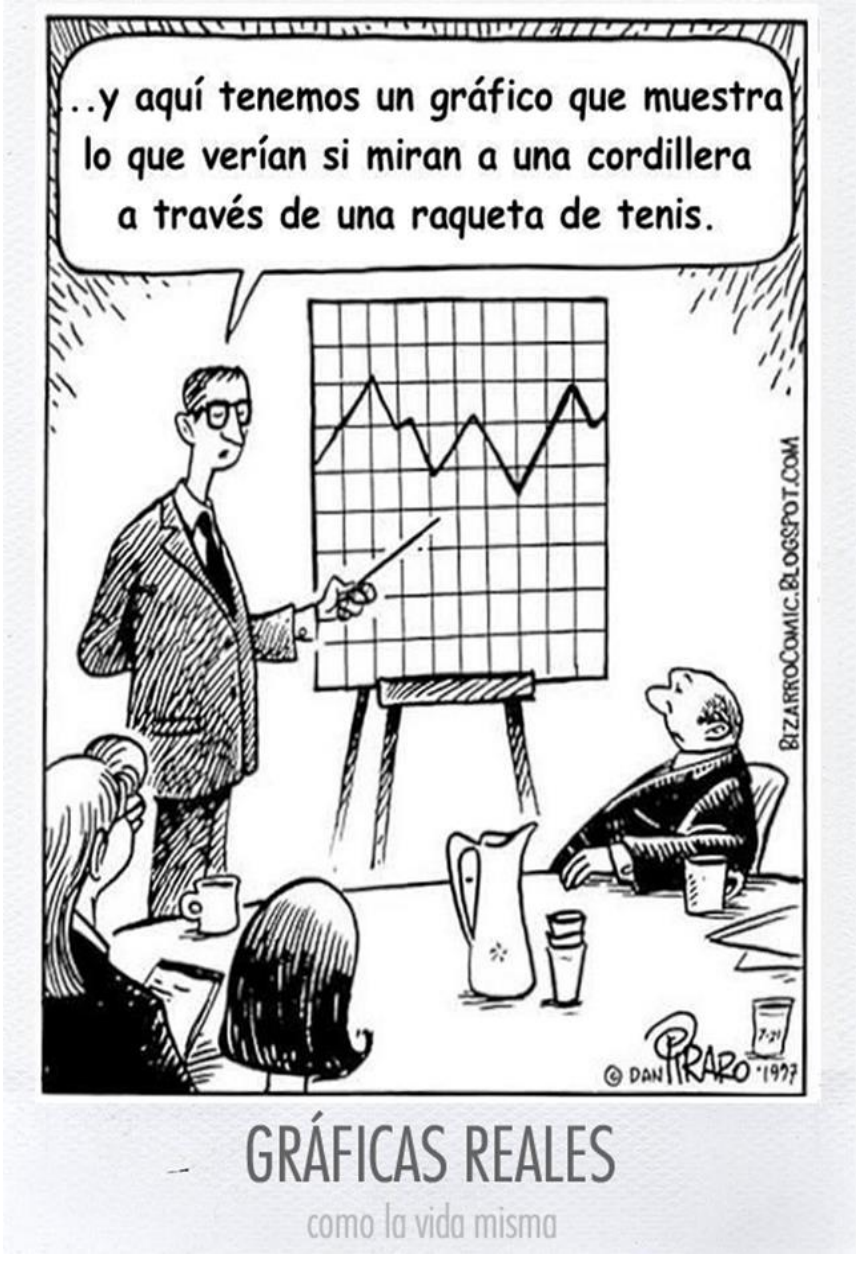
- **Estatística descritiva**

  Resumir, **describir** ou presentar a información a través de:

  - **Táboas**
  - **Gráficos**
  - **Resumos estatísticos** (medidas tendencia central, medidas dispersión, medidas de localización)

- **Estatística inferencial**

  Métodos que usan a información para facer **predicións**, tomar **decisións** ou facer **inferencias**.



Estatística

I) Conceptos básicos

Como comezar?

Poboación: Universo de individuos ó cal se refire o estudo que se pretende realizar.

Que podemos ter nós?

Mostra: Subconxunto da poboación cuxos valores da(s) variable(s) que se pretende(n) analizar son coñecidos.

Que variable?

Variable: Trazo ou característica dos elementos da poboación que se pretende analizar.

Módulo IV – Estatística descritiva

I) Tipos de variables

I) Definición

II) Variables cualitativas

III) Variables cuantitativas

Estatística

I) Definición

Importante!

Variable: Trazo ou característica dos elementos da poboación que se pretende analizar.

A diferenciación de variables estatísticas determinará o tipo de técnica que se pode utilizar (Ex.: Representacións gráficas)

TIPOS DE VARIABLES

Variables cualitativas (valores non numéricos)
Cualitativas nominais
Cualitativas ordinais
Variables cuantitativas (valores numéricos)
Cuantitativas discretas
Cuantitativas continuas

Cualitativas nominais

Miden características que non toman valores numéricos.

(«categorías sen orde»)

Exemplos:

- **Sexo**: home ou muller
- **País de orixe**: España, Arxentina ou México
- **Etnia**: asiática, africana e europea.
- **Perfil lingüístico**: monolingüe, bilingüe, multilingüe
- **Método de ensinanza**: resposta, audiolingüe ou tradución gramática
- **Palabra usada para «cheminea»:** cheminea, fumeira, troneira
- **Pronunciación de <c>**: [k], [θ]

Cualitativas ordinais

Miden características que non toman valores numéricos pero si presentan entre os seus posibles valores unha relación de orde

Exemplos:

- **Educación**: estudos universitarios, estudos secundarios ou estudos primarios («graos de nivel de estudos»)

- **«Intelixencia» - Análise actitudinal á pregunta: É intelixente a persoa que está falando?**: totalmente de acordo, de acordo, nin de acordo nin en desacordo, desacordo, totalmente desacordo («graos de acordo»)

Estatística

III) Variables cuantitativas

Cuantitativas discretas

Miden características que toman valores numéricos pero nun número discreto de valores (no conxunto dos números naturais) «resultado dun conteo»

Exemplos:

- **Número de viaxes fora do país**: 1,2,3,4,...
- **Número de linguas faladas:** 1,2,3,4,5,6...
- **Número de libros que les nun ano:** 1,2,3,4,....

Cuantitativas continuas

Miden características que toman valores numéricos pero poden tomar valores dentro dun intervalo real («xeralmente poden levar decimais»)

Exemplos:

- **Idade**: 21; 21’3; 25; 25’5; 30; 34’5;...
- **Horas que pasas escribindo:** 0’5; 0’99; 1’25; 1’5; 2;...
- **Frecuencia dun son (Hz):** 657; 500; 507; 590; 594; 463; 518 ...

Módulo IV – Estatística descritiva

II) Variables cualitativas

I) Descrición dos datos

II) Representación gráfica

III) Construción por clases

Estatística

I) Descrición dos datos

Variables cualitativas: Características non medidas numericamente

- **Cualitativas nominais - categorías sen orde**
- **Cualitativas ordinais - categorías con orde**

Os valores dunha variable cualitativa dispóñense nunha **táboa de frecuencias**:

- **Frecuencias absolutas (n_i)**: número de casos que presentan cada un dos niveis/valores da variable.
- **Frecuencias relativas (f_i)** : proporción de casos que presentan cada un dos niveis/valores da variable.

Categorías das variables

x_i	n_i	f_i
x_1	n_1	$f_1 = n_1/N$
x_2	n_2	$f_2 = n_2/N$
...	...	...
x_k	n_k	$f_k = n_k/N$
...	...	....
	$\sum_i n_i = N$	$\sum_i f_i = 1$

N: número total de casos

Importante!

Estatística

I) Descrición dos datos

Exemplo (antroponimia):

```
antroponimia<-read.table("antroponimia.txt",header=T)

#Visualizar os datos:
antroponimia
```

	apelidos	provincia	clasificacion
1	Calaza	Alava	delexical
2	Calaza	Alava	delexical
3	Calaza	Alava	delexical
4	Calaza	Alava	delexical
5	Calaza	Alava	delexical
6	Calaza	Barcelona	delexical
7	Calaza	Barcelona	delexical
8	Calaza	Barcelona	delexical
9	Calaza	Barcelona	delexical
10	Calaza	Barcelona	delexical
11	Calaza	Barcelona	delexical

```
#ou ben:
View(antroponimia)

#Nomes das variables:
names(antroponimia)
```

```
[1] "apelidos"     "provincia"    "clasificacion"
```

Importante!

Comandos de interese para un primeiro contacto coa base de datos:

View()
names()

Estatística

I) Descrición dos datos

Exemplo (antroponimia):

#Resumo das variables contidas na base de datos:

summary(antroponimia)

```
  apelidos        provincia      clasificacion
Beis   : 84   A_Coruna  :128   delexical :206
Calaza:206   Lugo       : 66   toponimico: 84
             Madrid     : 32
             Pontevedra: 26
             Barcelona  : 12
             Bizkaia    : 11
             (Other)    : 15
```

Variables cualitativas coa súa frecuencia absoluta

Importante!

Comandos de interese para un primeiro contacto coa base de datos:

View()
names()
summary()

Estatística

I) Descrición dos datos

Exemplo (antroponimia):

#Acceso a cada unha das variables

#co comando $:

antroponimia$apelidos
antroponimia$provincia
antroponimia$clasificacion

#facendo un attach

attach(antroponimia)
apelidos

#Ver de que clase son as variables

class(apelidos)

```
[1] "factor"
```

levels(apelidos)

```
[1] "Beis"   "Calaza"
```

Importante!

Comandos de interese para un primeiro contacto coa base de datos:

View()
names()
summary()
attach()

Variables cualitativas

#Resumo descriptivo de variables cualitativas:

table(apelidos) **#frecuencias absolutas**

```
apelidos
  Beis Calaza
    84    206
```

table(apelidos,clasificacion)

```
        clasificacion
apelidos delexical toponimico
  Beis           0         84
  Calaza       206          0
```

taboa=table(apelidos)

tab_cont_apel=table(apelidos,clasificacion)

addmargins(tab_cont_apel)

```
        clasificacion
apelidos delexical toponimico Sum
  Beis           0         84  84
  Calaza       206          0 206
  Sum          206         84 290
```

prop.table(taboa) **#frecuencias relativas**

```
apelidos
     Beis    Calaza
0.2896552 0.7103448
```

prop.table(taboa)*100 #frecuencias relativas en %

```
apelidos
    Beis   Calaza
28.96552 71.03448
```

Importante!

Comandos de interese para variables cualitativas:

table()

prop.table()

addmargins()

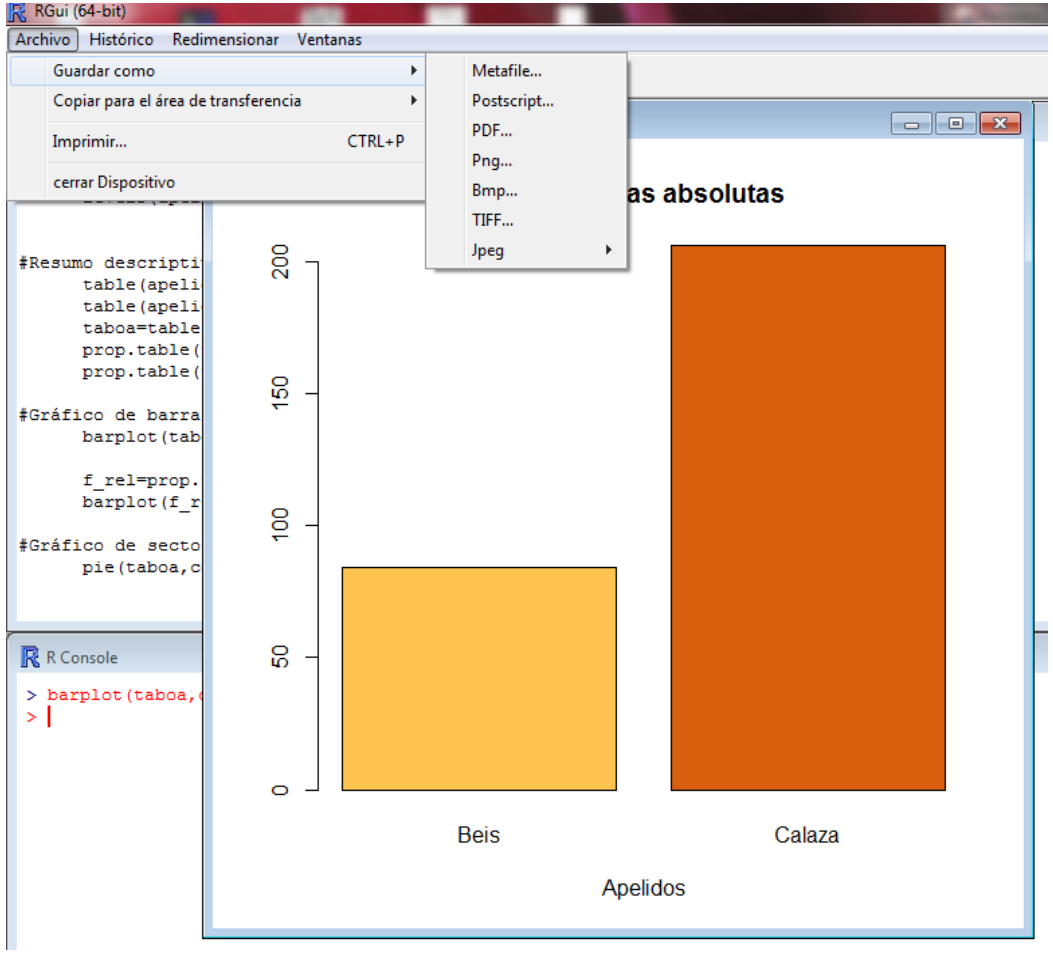
Estatística

II) Representación gráfica

Variables cualitativas

1. Gráfico de barras

barplot(taboa,col=c("#fec44f","#d95f0e"),main="Frecuencias absolutas",xlab="Apelidos")



Importante!

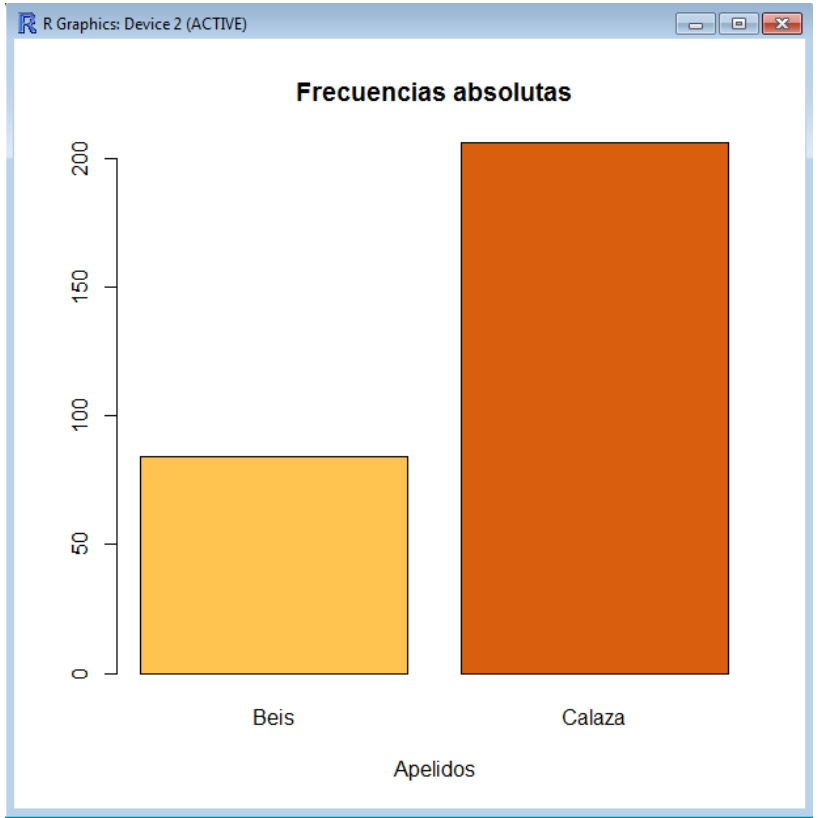
Comandos de interese para variables cualitativas:

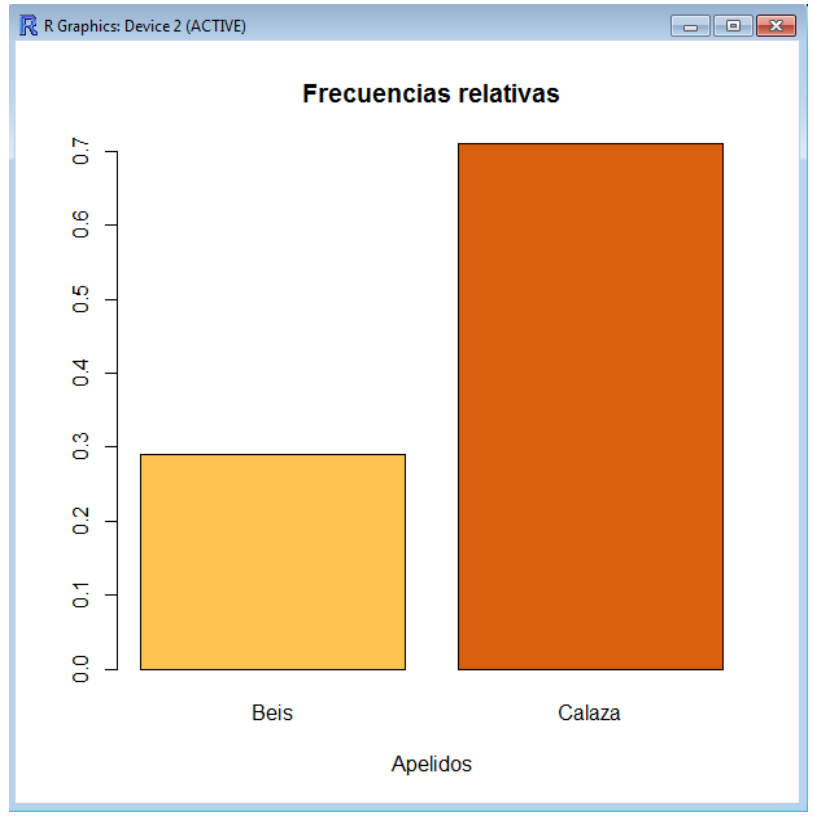
table()
prop.table()
addmargins()
barplot()

Variables cualitativas

1. Gráfico de barras

```
barplot(taboa,col=c("#fec44f","#d95f0e"),main="Frecuencias absolutas",xlab="Apelidos")

f_rel=prop.table(taboa) #frecuencias relativas
barplot(f_rel, col=c("#fec44f","#d95f0e"),main="Frecuencias relativas",xlab="Apelidos")
```

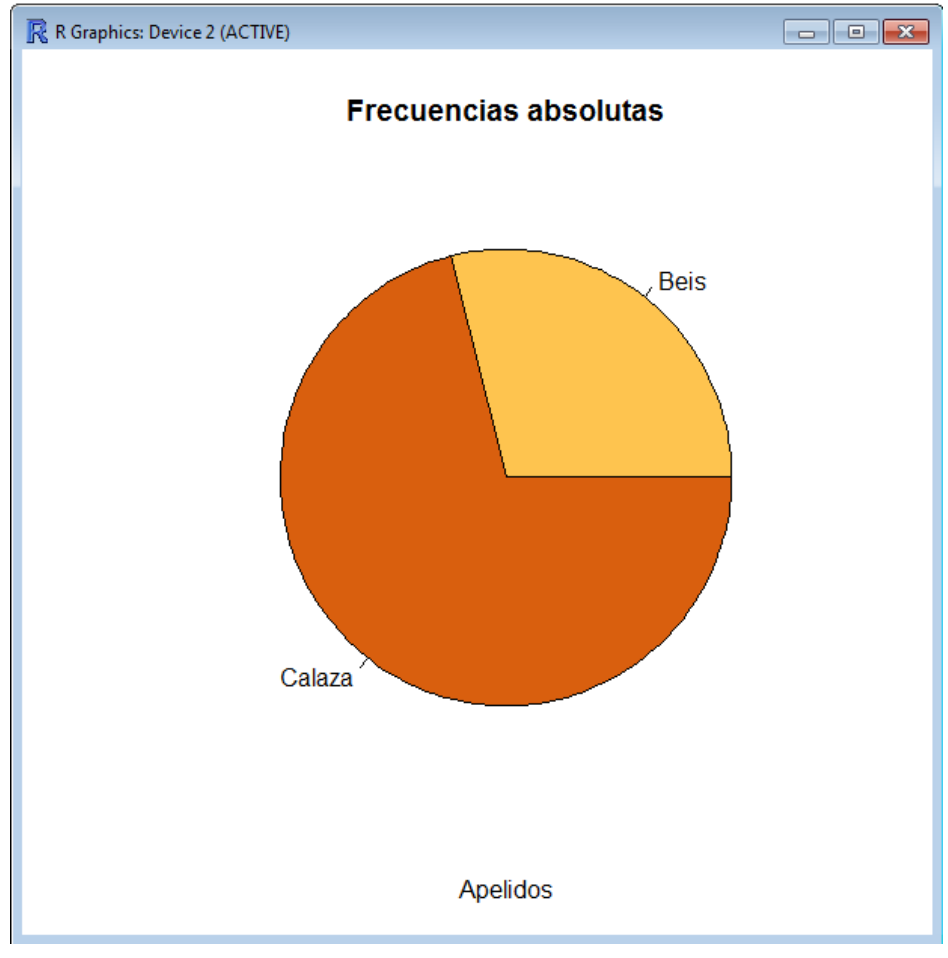




Variables cualitativas

2. Gráfico de sectores

```
pie(taboa,col=c("#fec44f","#d95f0e"),main="Frecuencias absolutas",xlab="Apelidos")
```



Importante!

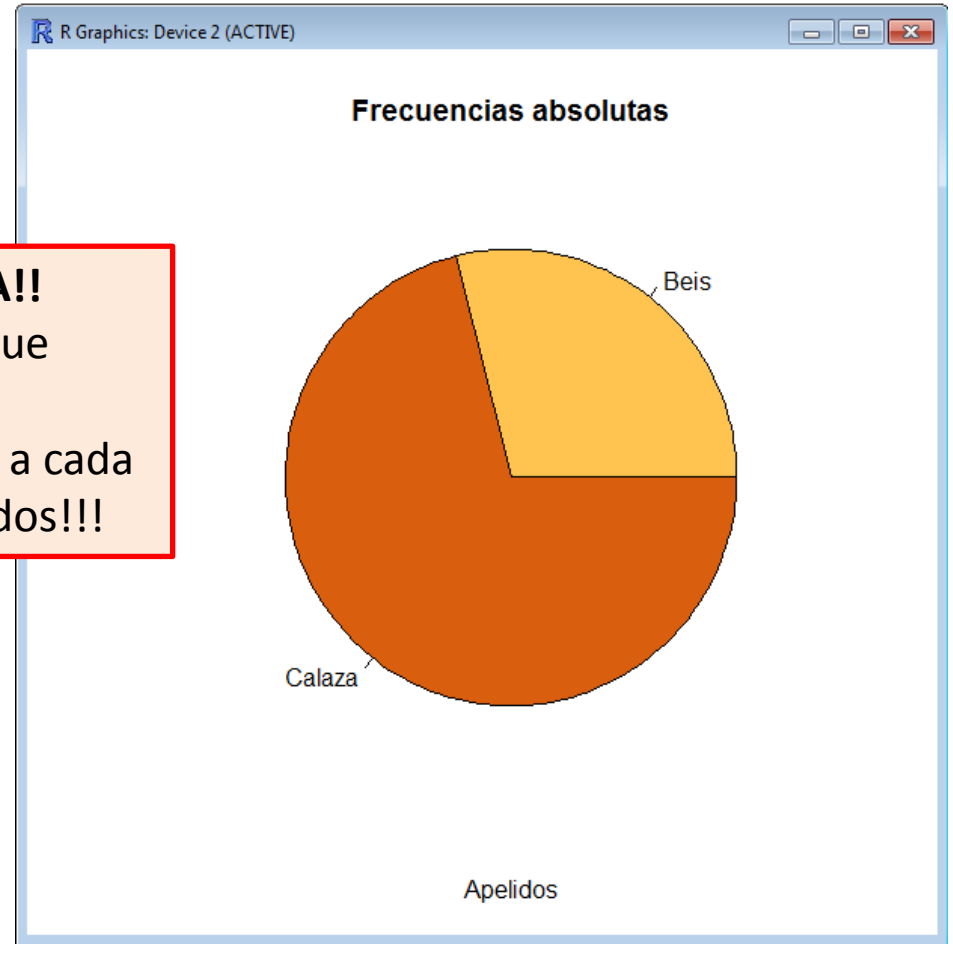
Comandos de interese para variables cualitativas:

table()
prop.table()
addmargins()
barplot()
pie()

Variables cualitativas

2. Gráfico de sectores

```
pie(taboa,col=c("#fec44f","#d95f0e"),main="Frecuencias absolutas",xlab="Apelidos")
```



ALERTA!!
Falta saber que porcentaxe corresponde a cada un dos apelidos!!!

Importante!

Comandos de interese para variables cualitativas:

table()
prop.table()
addmargins()
barplot()
pie()

Estatística

II) Representación gráfica

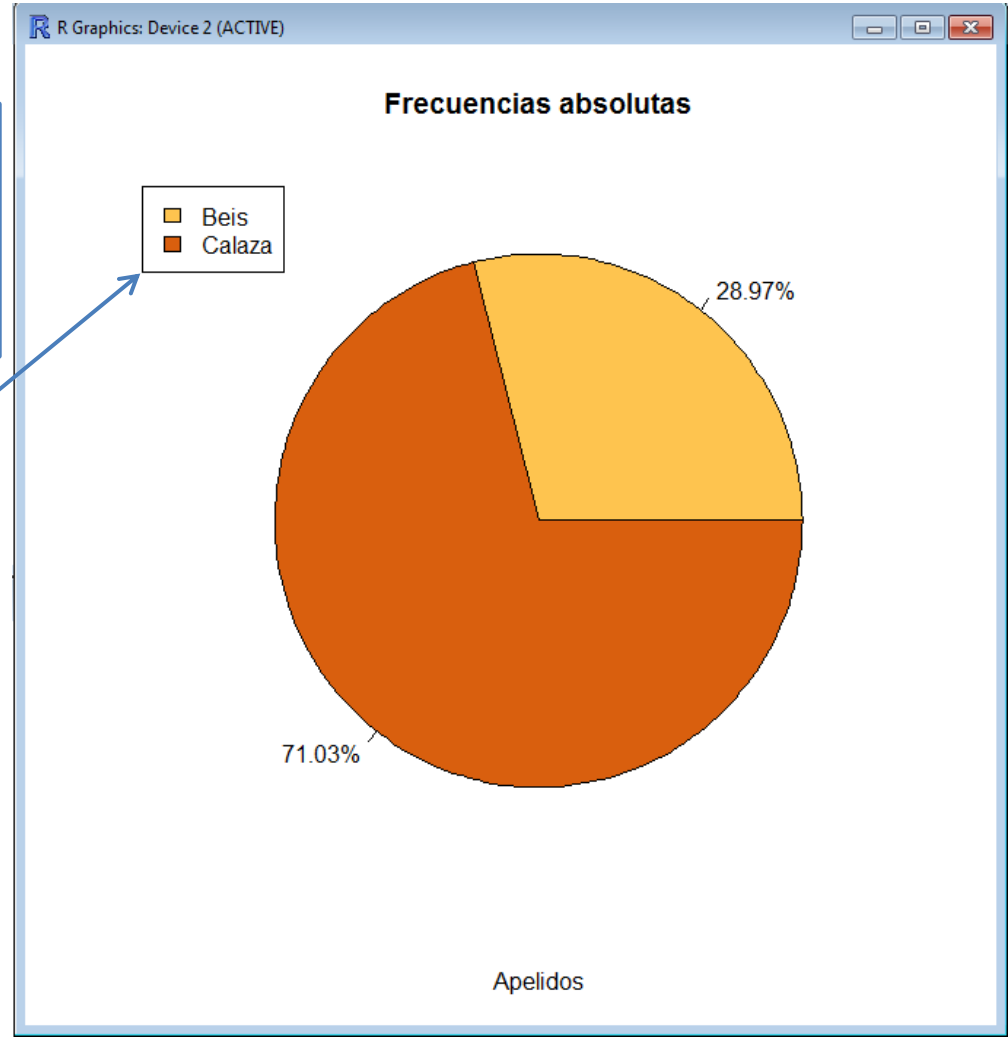
Variables cualitativas

2. Gráfico de sectores

```
etiquetas<-prop.table(taboa)*100
etiquetas=round(etiquetas,2)
etiquetas
apelidos
  Beis Calaza
 28.97  71.03
etiquetas<-paste(etiquetas,"%",sep="")
etiquetas
[1] "28.97%" "71.03% "

 pie(taboa,labels=etiquetas,col=c("#fec44f","#d95f0e"),
              main="Frecuencias absolutas",xlab="Apelidos")
 legend(-1.2,1,legend=levels(apelidos),
              fill=c("#fec44f","#d95f0e"))
```

Imos incluír como «etiquetas» as frecuencias relativas de cada un dos apelidos

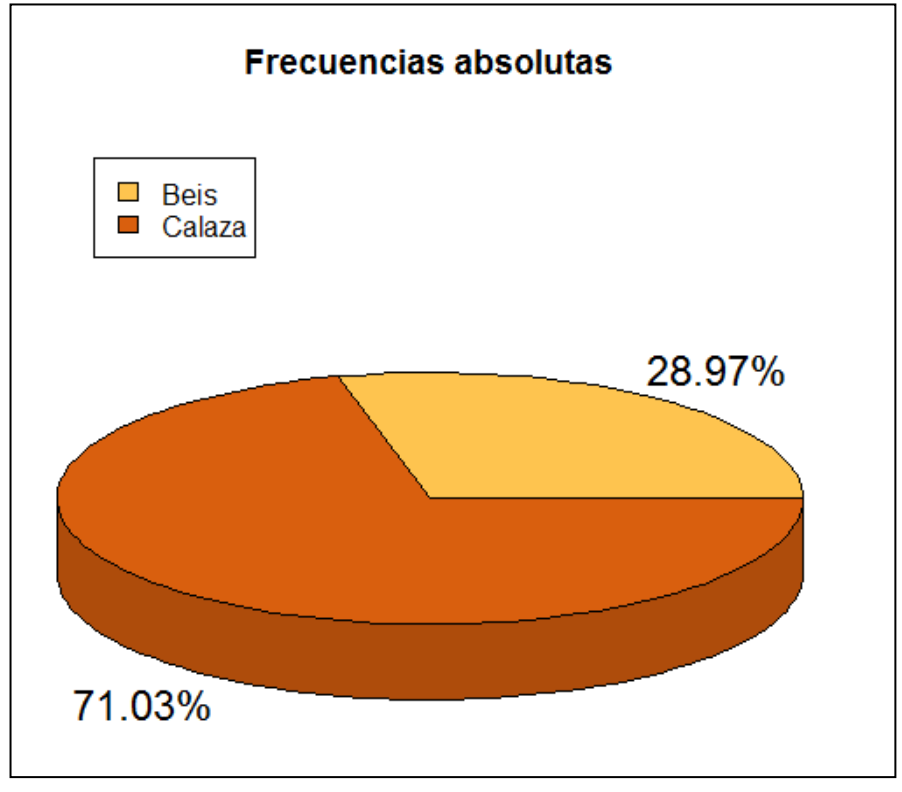
Coordenadas na gráfica onde inserimos a lenda



Variables cualitativas

2. Gráfico de sectores (3D)

```
library(plotrix)
pie3D(taboa,labels=etiquetas,col=c("#fec44f","#d95f0e"),main="Frecuencias absolutas")
legend(-0.9,1,legend=levels(apelidos),   fill=c("#fec44f","#d95f0e"))
```



Importante!

Comandos de interese para variables cualitativas:

table()
prop.table()
addmargins()
barplot()
pie()/pie3D()

Estatística

III) Construción por clases

Variables cualitativas

Casos nos que é interesante **construír agrupacións**:

A pesar de que a variable é numérica (por exemplo: conteos) nós estamos interesados en consideralos como grupos.

Exemplo:
Estamos a facer unha auditoría sobre o número de estudantes que figuran durante este ano en 25 materias de Filoloxía Galega, e atopámonos co seguinte:

25 50 48 40 15 16 5 10 6 31 56 55 2
3 5 15 6 5 3 2 8 49 4 14 6

Para darlle un pouco de orde a estes datos debemos **agrupar os nosos datos en categorías.**

Variables cualitativas

Casos nos que é interesante **construír agrupacións**:

Exemplo:

Primeiro temos que **decidir o número de clases (categorías)** que queremos ou que necesitamos. Como temos 25 casos parece razoable dividilos en 5 clases (clase moi pequena, pequena, normal, bastante numerosa, ou moi numerosa).

Nota

Depende do noso criterio a elección do número de clases.

Un criterio bastante extendido consiste en tomar como número de clases o enteiro máis próximo a $\sqrt{n}$.

Variables cualitativas

Casos nos que é interesante **construír agrupacións**:

Exemplo:

Para **determinar o tamaño de cada clase**, usaremos:

$$tamaño\ da\ clase = \frac{\{valor\ máis\ grande\ dos\ nosos\ datos - valor\ máis\ pequeno\}}{número\ de\ clases} = \frac{56\ - 2}{5} = 10.8$$

Para asegurarnos de que se inclúan os extremos dos nosos datos (e dado que non existe 0.8 estudantes), tomaremos clases de tamaño 11.

Estatística

III) Construción por clases

Variables cualitativas

Casos nos que é importante **construír agrupacións**:

Exemplo:
Entonces consideraremos as seguintes categorías:

Tamaño	Intervalos de clase	
moi pequena	[2,13)	2-12 estudantes
pequena	[13,24)	13-23 estudantes
normal	[24,35)	24-34 estudantes
bastante numerosa	[35,46)	35-45 estudantes
moi numerosa	[46,57)	46-56 estudantes

Entón temos que asociar os valores que correspondan a cada clase, e observar cantas materias temos en Filoloxía de cada un dos tamaños prefixados.

Variables cualitativas

Casos nos que é importante **construír agrupacións**:

Exemplo: **Como facer isto en R?**

```
#Número de materias
    num_alum=c(25,50, 48, 40, 15, 16, 5, 10, 6, 31, 56, 55, 2, 3, 5, 15, 6,5, 3, 2,8, 49, 4, 14, 6)
    length(num_alum)
    [1] 25

#Número de intervalos
    num_int_clase=sqrt(length(num_alum))
    num_int_clase
    [1] 5
```

Estatística

III) Construción por clases

Variables cualitativas

Casos nos que é importante **construír agrupacións**:

Exemplo:

Como facer isto en R?

Función para dividir en rangos

Número de intervalos

```
rangos<- cut(num_alum, breaks=5,include.lowest=T)
rangos
 [1] (23.6,34.4] (45.2,56.1] (45.2,56.1] (34.4,45.2] (12.8,23.6] (12.8,23.6] [1.95,12.8] [1.95,12.8] [1.95,12.8] (23.6,34.4]
[11] (45.2,56.1] (45.2,56.1] [1.95,12.8] [1.95,12.8] [1.95,12.8] (12.8,23.6] [1.95,12.8] [1.95,12.8] [1.95,12.8] [1.95,12.8]
[21] [1.95,12.8] (45.2,56.1] [1.95,12.8] (12.8,23.6] [1.95,12.8]
Levels: [1.95,12.8] (12.8,23.6] (23.6,34.4] (34.4,45.2] (45.2,56.1]

rangos_con_nome_categorias<- cut(num_alum, breaks=5,include.lowest=T,labels=c("moi
pequena", "pequena", "normal", "bastante numerosa",  "moi numerosa"))

table(rangos_con_nome_categorias) #frecuencias absolutas
rangos_con_nome_categorias
        moi pequena            pequena             normal  bastante numerosa     moi numerosa
                 13                   4                  2                  1                5
```

Etiquetas para as categorías

Estatística

III) Construción por clases

Variables cualitativas

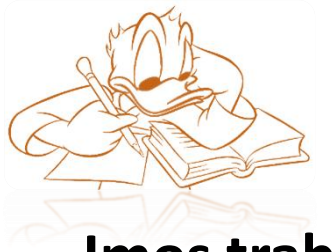
Casos nos que é importante **construír agrupacións**:

Exemplo: **Como facer isto en R?**

```
taboa=table(rangos_con_nome_categorias)
prop.table(taboa)          #frecuencias relativas
```

```
rangos_con_nome_categorias
    moi pequena        pequena         normal  bastante numerosa     moi numerosa
           0.52           0.16           0.08               0.04             0.20
```

Estatística



Exercicio 8

Imos traballar cos datos «tempos_compostos_galego_medieval.csv» que utilizaramos no exercicio 1.

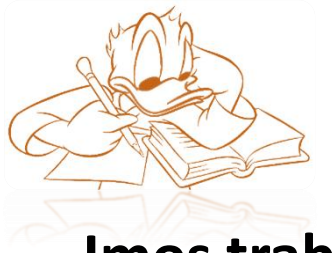
1º) Cargamos os datos

2º) Visualizámolos

Data: tempos_compostos

	tipo_de_verbo	verbo	auxiliar	num_aparicion
1	paso_de_tempo	durar	aver	3
2	paso_de_tempo	passar	ser	13
3	procesos_fisicos	(de)mudar	ser	5
4	procesos_fisicos	desecar	ser	1
5	procesos_fisicos	dormir	aver	3
6	procesos_fisicos	enloqueçer	ser	1
7	procesos_fisicos	escalentar	ser	1
8	procesos_fisicos	finar	ser	9
9	procesos_fisicos	guarir,guaresçer	ser	5
10	procesos_fisicos	morir	ser	82
11	procesos_fisicos	nasçer	ser	7
12	procesos_fisicos	parir	ser	1
13	suceso	acaeçer	ser	3
14	suceso	acaeçer	aver	3
15	suceso	conteçer	aver	4
16	suceso	passar	aver	3
17	permanencia	albergar	aver	1
18	permanencia	fincar	aver	9
19	permanencia	folgar	aver	1
20	permanencia	morar	aver	7
21	permanencia	posar	ser	1

Estatística



Exercicio 8

Imos traballar cos datos «tempos_compostos_galego_medieval.csv» que utilizaramos no exercicio 1.

1º) Cargamos os datos

```
tempos_compostos=read.csv("tempos_compostos_galego_medieval.csv",header=T,sep=";")
```

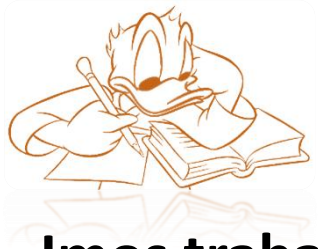
2º) Visualizámolos

```
View(tempos_compostos)
```

Data: tempos_compostos

	tipo_de_verbo	verbo	auxiliar	num_aparicion
1	paso_de_tempo	durar	aver	3
2	paso_de_tempo	passar	ser	13
3	procesos_fisicos	(de)mudar	ser	5
4	procesos_fisicos	desecar	ser	1
5	procesos_fisicos	dormir	aver	3
6	procesos_fisicos	enloqueçer	ser	1
7	procesos_fisicos	escalentar	ser	1
8	procesos_fisicos	finar	ser	9
9	procesos_fisicos	guarir,guaresçer	ser	5
10	procesos_fisicos	morir	ser	82
11	procesos_fisicos	nasçer	ser	7
12	procesos_fisicos	parir	ser	1
13	suceso	acaeçer	ser	3
14	suceso	acaeçer	aver	3
15	suceso	conteçer	aver	4
16	suceso	passar	aver	3
17	permanencia	albergar	aver	1
18	permanencia	fincar	aver	9
19	permanencia	folgar	aver	1
20	permanencia	morar	aver	7
21	permanencia	posar	ser	1

Estatística



Exercicio 8

Imos traballar cos datos «tempos_compostos_galego_medieval.csv» que utilizaramos no exercicio 1.

a) Ver de que clase é cada variable

b) Nas variables cualitativas (variables categóricas): «tipo_de_verbo» e «auxiliar»

   i. Facer unha táboa resumo de cada variable coas frecuencias relativas e outra coas absolutas

   ii. Representación gráfica de cada unha delas

Estatística

Exercicio 8: Solución

«tempos_compostos_galego_medieval.csv»

a) Ver de que clase é cada variable

```
attach(tempos_compostos)
class(tipo_de_verbo)
class(num_aparicion)
```

b) Nas variables cualitativas (variables categóricas): «tipo_de_verbo» e «auxiliar»

   i. Facer unha táboa resumo de cada variable coas frecuencias relativas e outra coas absolutas

Estatística

Exercicio 8: Solución

«tempos_compostos_galego_medieval.csv»

a) Ver de que clase é cada variable

```
attach(tempos_compostos)
class(tipo_de_verbo)
class(num_aparicion)
```

b) Nas variables cualitativas (variables categóricas): «tipo_de_verbo» e «auxiliar»

i. Facer unha táboa resumo de cada variable coas frecuencias relativas e outra coas absolutas

```
tab_auxiliar=table(auxiliar) ; tab_auxiliar
auxiliar
aver  ser
   9   12
prop.table(tab_auxiliar)
auxiliar
     aver       ser
0.4285714 0.5714286
```

Estatística

Exercicio 8: Solución

«tempos_compostos_galego_medieval.csv»

b) Nas variables cualitativas (variables categóricas) : «tipo_de_verbo» e «auxiliar»

ii. Representación gráfica de cada unha delas (diagrama de barras)

Estatística

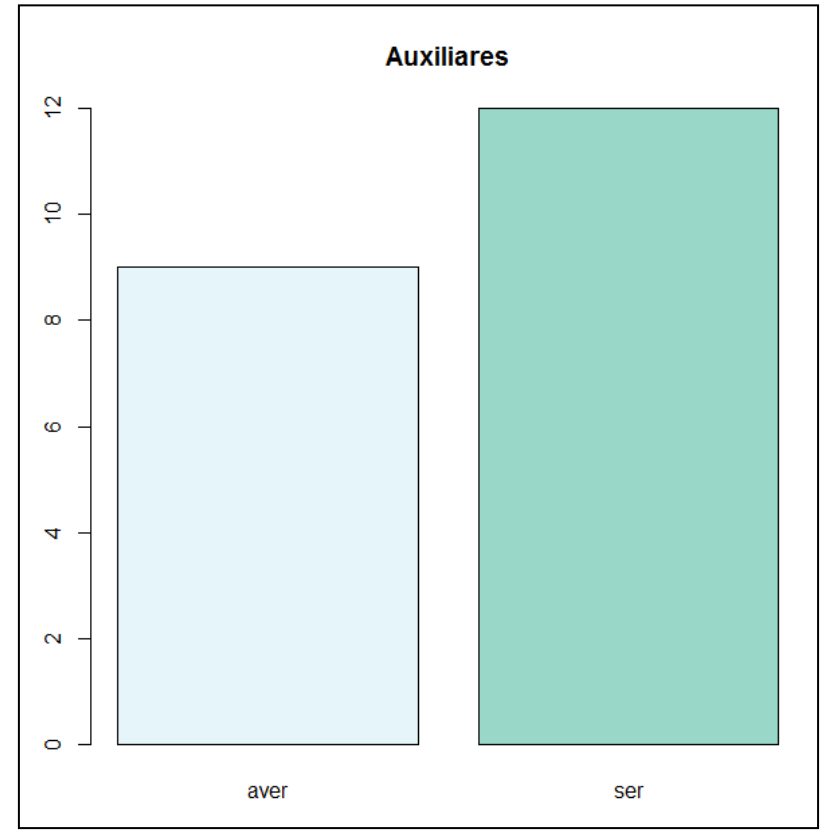
Exercicio 8: Solución

«tempos_compostos_galego_medieval.csv»

b) Nas variables cualitativas (variables categóricas) : «tipo_de_verbo» e «auxiliar»

ii. Representación gráfica de cada unha delas (diagrama de barras)

```
barplot(tab_auxiliar,col=c("#e5f5f9","#99d8c9"),main="Auxiliares")
```



Estatística

Exercicio 8: Solución

«tempos_compostos_galego_medieval.csv»

b) Nas variables cualitativas (variables categóricas): «tipo_de_verbo» e «auxiliar»

ii. Representación gráfica de cada unha delas (diagrama de sectores)

Estatística

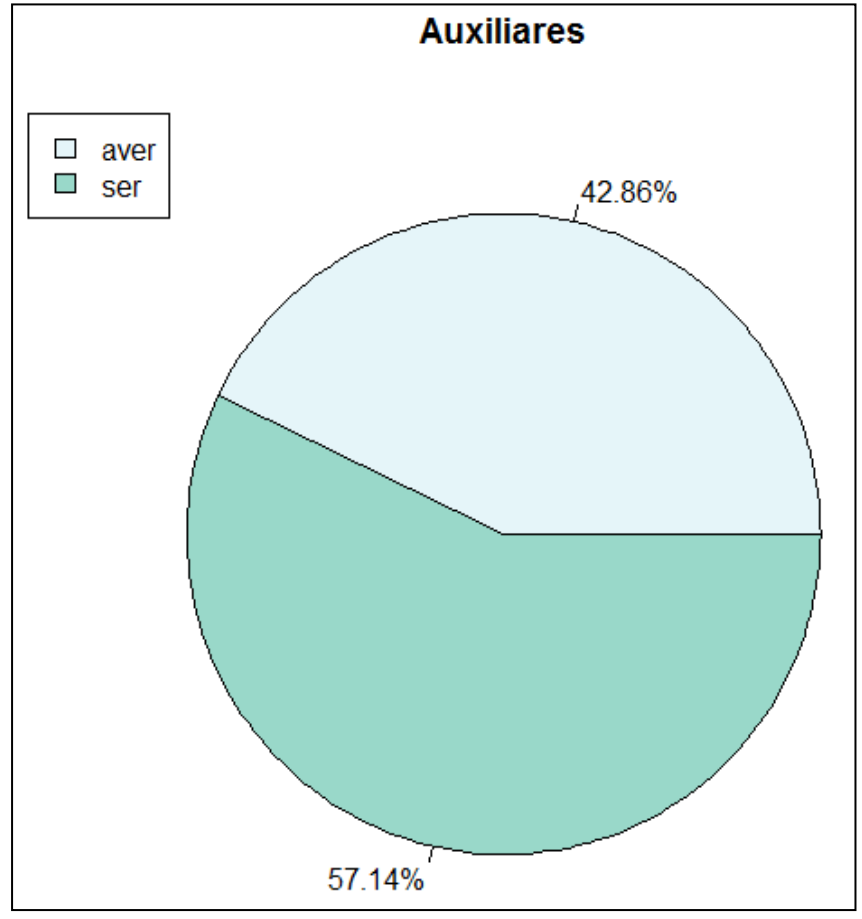
Exercicio 8: Solución

«tempos_compostos_galego_medieval.csv»

b) Nas variables cualitativas (variables categóricas): «tipo_de_verbo» e «auxiliar»

ii. Representación gráfica de cada unha delas (diagrama de sectores)

```
etiquetas<-prop.table(tab_auxiliar)*100
etiquetas<-round(etiquetas,2)
etiquetas<-paste(etiquetas,"%",sep="")

pie(tab_auxiliar,labels=etiquetas,col=c("#e5f5f9","#99d8c9"),
            main="Auxiliaries")

legend(-1.2,1.05,legend=levels(auxiliar),fill=c("#e5f5f9","#99d8c9"))
```



Estatística

Exercicio 8: Solución

«tempos_compostos_galego_medieval.csv»

b) Nas variables cualitativas (variables categóricas): «tipo_de_verbo» e «auxiliar»

ii. Representación gráfica de cada unha delas (diagrama de sectores,3D)

Estatística

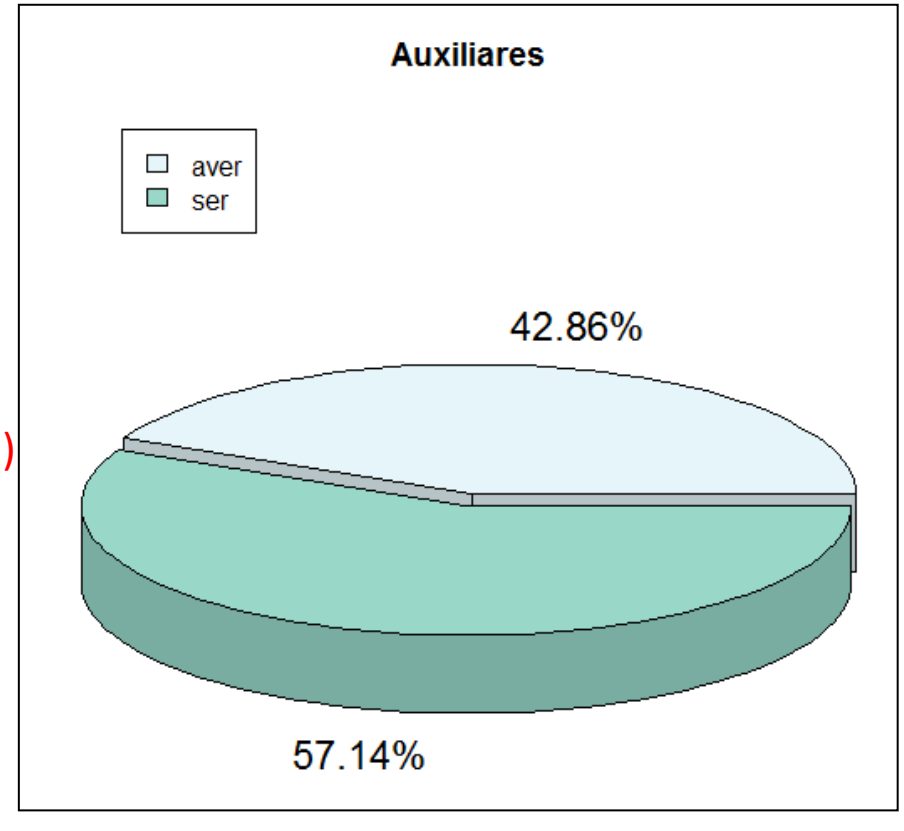
Exercicio 8: Solución

«tempos_compostos_galego_medieval.csv»

b) Nas variables cualitativas (variables categóricas): «tipo_de_verbo» e «auxiliar»

ii. Representación gráfica de cada unha delas (diagrama de sectores,3D)

```
pie3D(tab_auxiliar,labels=etiquetas,explode=0.03,
      col=c("#e5f5f9","#99d8c9"),main="Auxiliares")

legend(-0.9,1.05,legend=levels(auxiliar),fill=c("#e5f5f9","#99d8c9"))
```



Módulo IV – Estatística descritiva

III) Variables cuantitativas

I) Descrición dos datos

II) Representación gráfica

Estatística

Variables cuantitativas

Cantidades numéricas

- **Cuantitativas discretas: número finito discreto de valores (ex.: número de linguas faladas, número de libros que les nun ano,...)**
  - I) Descrición dos datos
  - II) Representación gráfica
- **Cuantitativas continuas: infinitos valores nun intervalo real (ex.: idade, frecuencia dun son,...)**
  - I) Descrición dos datos
  - II) Representación gráfica

Estatística

Variables cuantitativas discretas

Miden características que toman valores numéricos pero nun número discreto de valores (no conxunto dos números naturais) «resultado dun conteo»

Exemplos:

- **Número de viaxes fora do país**: 1,2,3,4,...
- **Número de linguas faladas:** 1,2,3,4,5,6...
- **Número de libros que les nun ano:** 1,2,3,4,....

Variables cuantitativas discretas

O **tratamento** é similar ás variables cualitativas (ten sentido falar de **frecuencias** e das representacións das mesmas en **diagramas de barras** e **de sectores**)

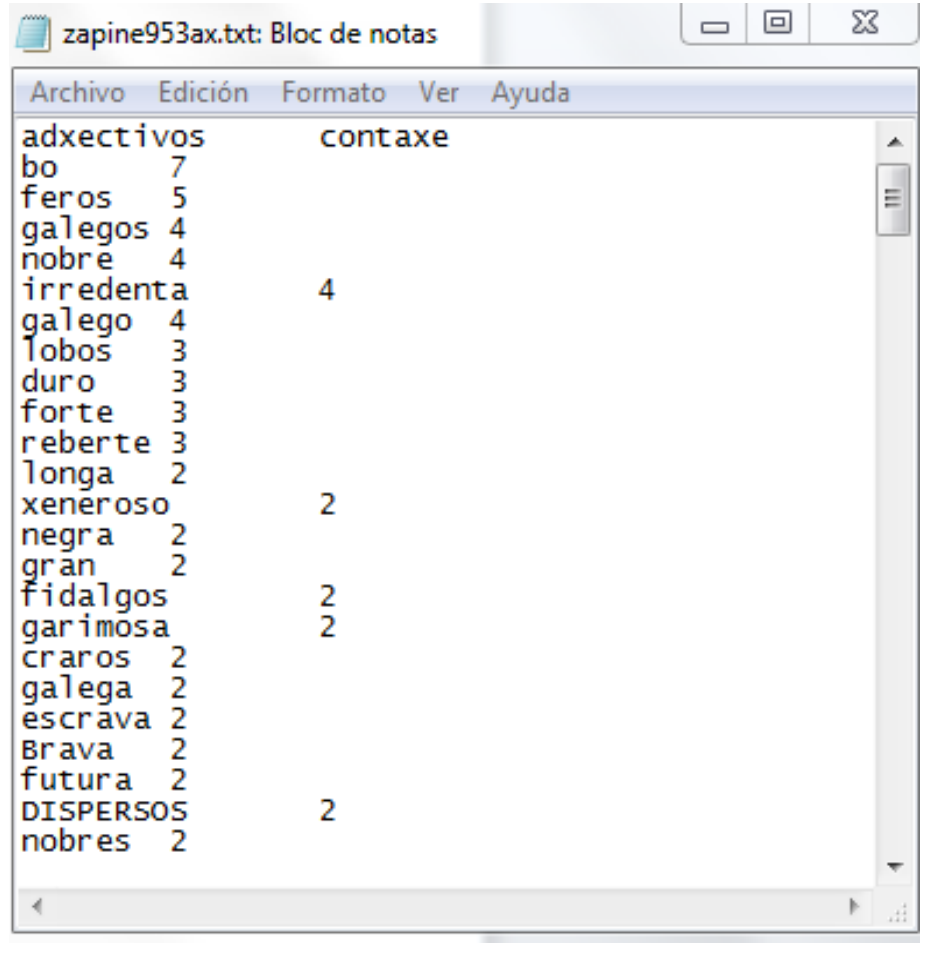
Exemplo:
Para poder reproducir o estudo das variables discretas partimos dunha base de datos extraída do TILG. Imos facer unha descrición dos adxectivos da obra ZAPINE953, á cal tedes gardada no arquivo:

«zapine953ax.txt»

Variables cuantitativas discretas

Exemplo:

«zapine953ax.txt»



Variables cuantitativas discretas

Exemplo:

«zapine953ax.txt»

Cargamos os datos e visualizámolos en R:

```
adxectivos<-read.table("zapine953ax.txt",header=T)

#Visualizámolos:

View(adxectivos)
```

Data: adxectivos

	adxectivo	contaxe
1	bo	7
2	feros	5
3	galegos	4
4	nobre	4
5	irredenta	4
6	galego	4
7	lobos	3
8	duro	3
9	forte	3
10	reberte	3
11	longa	2
12	xeneroso	2
13	negra	2
14	gran	2
15	fidalgos	2
16	garimosa	2
17	craros	2
18	galega	2
19	escrava	2

Variables cuantitativas discretas

Exemplo:

«zapine953ax.txt»

Contaxe: número de veces que se repite o adxectivo na obra

Como se comporta o número de repeticións dun adxectivo?

Adóitanse repetir 4 veces na obra? Ou 5? Ou 6?...

Variables cuantitativas discretas

Exemplo:

«zapine953ax.txt»

Contaxe: número de veces que se repite o adxectivo na obra

```
attach(adxectivos)
contaxe
 [1] 7 5 4 4 4 4 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [54] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[107] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[160] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
class(contaxe)
[1] "integer"
```

Nós queremos considerar cada número como un nivel (que represente o grao de repetición do adxectivo na obra).
Os niveis serían: 1,2,3,4,5,7

Variables cuantitativas discretas

Exemplo:

«zapine953ax.txt»

Contaxe: número de veces que se repite o adxectivo na obra

```
contaxe<-as.factor(contaxe)
class(contaxe)
[1] "factor"
levels(contaxe)
[1] "1" "2" "3" "4" "5" "7“
table(contaxe)
contaxe
  1   2   3   4   5   7
178  21   4   4   1   1
```

Importante!

Comandos de interese para variables cuantitativas discretas:

***as.factor()*: para codificar un vector como un factor**

***levels()*: coñecer os niveis dunha variable**

***table()*: frecuencias asociadas a cada un dos niveis**

Estatística II) Representación gráfica

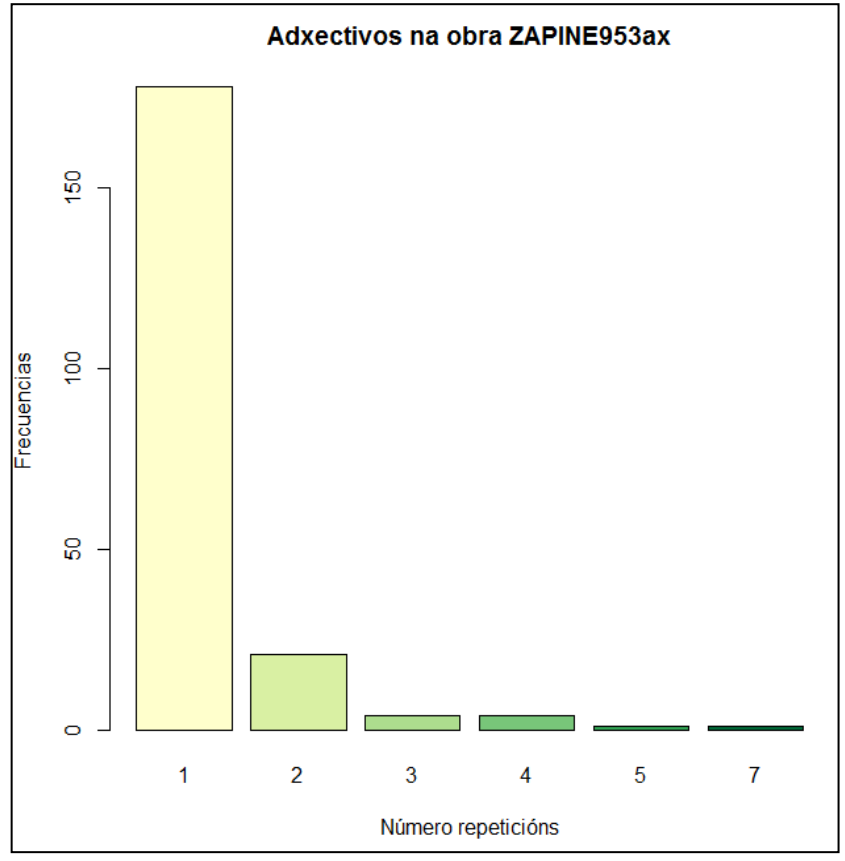
Variables cuantitativas discretas

1. Gráfico de barras

Cores para cada nivel da variable

```
plot(contaxe,col=c("#ffffcc","#d9f0a3","#addd8e","#78c679","#31a354","#006837"),
     ylab="Frecuencias",xlab="Número repeticións",main="Adxectivos na obra ZAPINE953ax")
```

A variable



Importante!

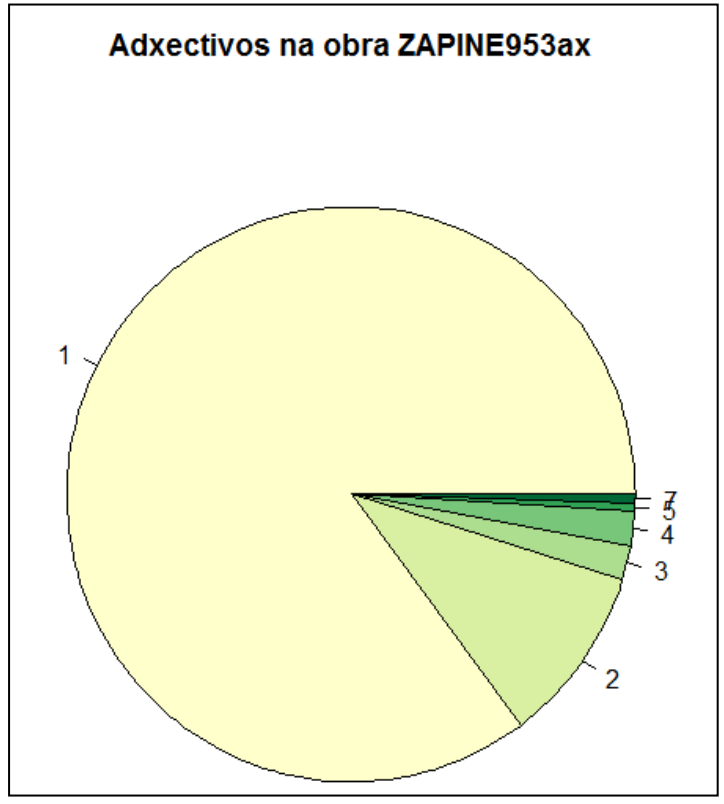
Comandos de interese para variables cuantitativas discretas:

plot()

Variables cuantitativas discretas

2. Gráfico de sectores

```
tab_contaxe<-table(contaxe)
pie(tab_contaxe,col=c("#ffffcc","#d9f0a3","#addd8e","#78c679","#31a354","#006837"),
        main="Adxectivos na obra ZAPINE953ax")
```



Importante!

Comandos de interese para variables cuantitativas discretas:

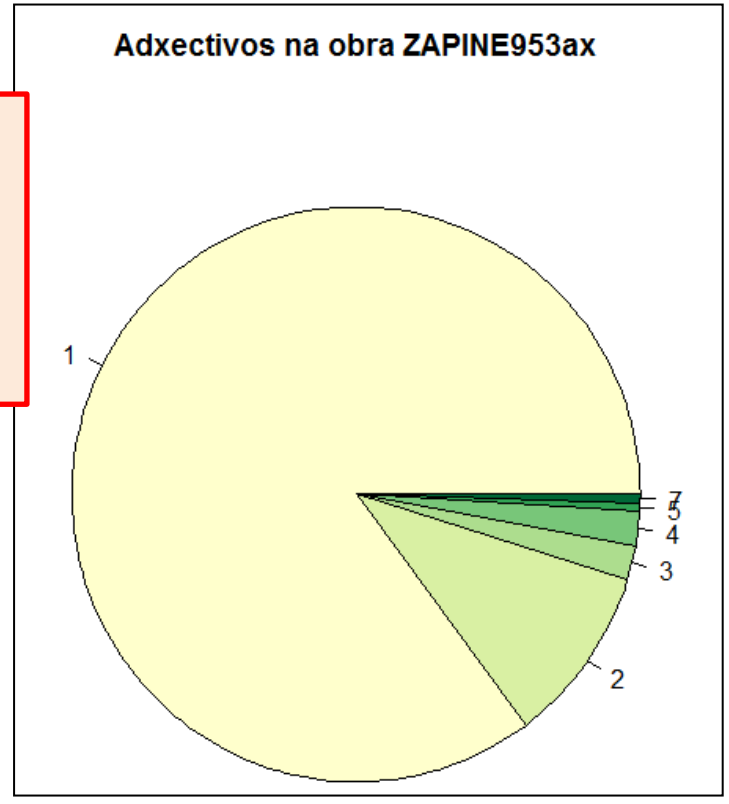
plot()
pie()/pie3D()

Estatística II) Representación gráfica

Variables cuantitativas discretas

2. Gráfico de sectores

```
tab_contaxe<-table(contaxe)
pie(tab_contaxe,col=c("#ffffcc","#d9f0a3","#addd8e","#78c679","#31a354","#006837"),
        main="Adxectivos na obra ZAPINE953ax")
```



ALERTA!!
Falta saber que porcentaxe corresponde a cada un dos niveis!!!

Importante!

Comandos de interese para variables cuantitativas discretas:

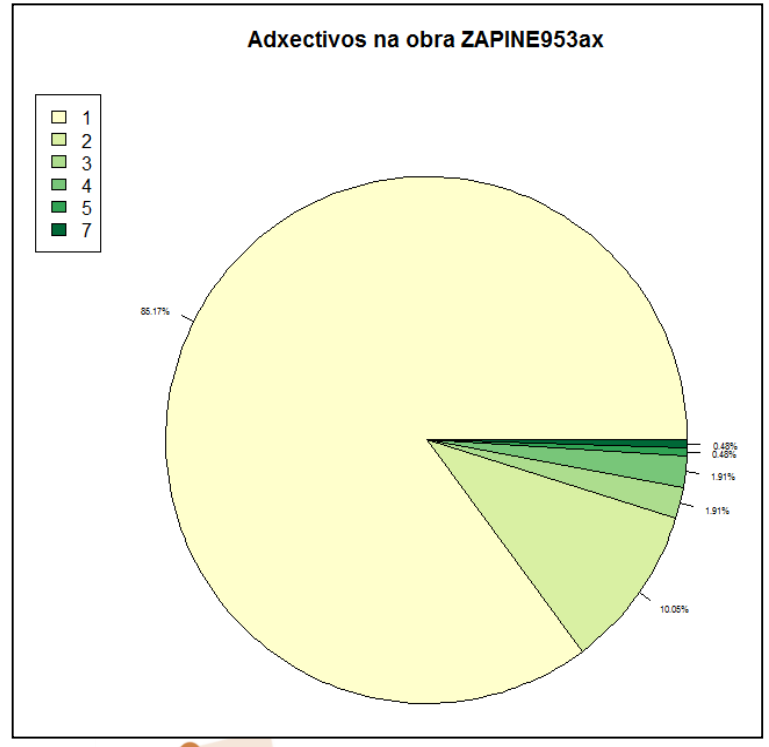
plot()
pie()/pie3D()

Estatística II) Representación gráfica

Variables cuantitativas discretas

2. Gráfico de sectores

```
tab_contaxe<-table(contaxe)

etiquetas<-prop.table(tab_contaxe)*100
etiquetas<-round(etiquetas,2)
etiquetas<-paste(etiquetas,"%",sep="")

pie(tab_contaxe,labels=etiquetas,cex=0.5,
col=c("#ffffcc","#d9f0a3","#addd8e","#78c679","#31a354",
"#006837"),main="Adxectivos na obra ZAPINE953ax")

legend(-1.2,1.05,legend=levels(contaxe),
fill=c("#ffffcc","#d9f0a3","#addd8e","#78c679",
"#31a354","#006837"))
```



Importante!

Comandos de interese para variables cuantitativas discretas:

plot()
pie()/pie3D()

Estatística

Exercicio 9

Imos traballar cos datos «1NT004916.txt» que utilizaramos no exercicio 1.

a) De que clase é a variable «contaxe»

b) Na variable «contaxe»:

   i. Facer unha táboa resumo de cada variable coas frecuencias relativas e outra coas absolutas segundo os diferentes «niveis de repetición»

   ii. Representación gráfica de cada unha delas

Estatística

Exercicio 9: Solución

Imos traballar cos datos «1NT004916.txt» que utilizaramos no exercicio 1.

a) De que clase é a variable «contaxe»

```
obra_demos=read.table("1NT004916.txt",header=T)
View(obra_demos)
attach(obra_demos)
names(obra_demos)
[1] "demostrativo" "contaxe"

class(contaxe)
[1] "integer"
```

Estatística

Exercicio 9: Solución

Imos traballar cos datos «1NT004916.txt» que utilizaramos no exercicio 1.

b) Na variable «contaxe»:

i. Facer unha táboa resumo de cada variable coas frecuencias relativas e outra coas absolutas segundo os diferentes «niveis de repetición»

```
contaxe_dem<-as.factor(contaxe)
table(contaxe_dem)
contaxe_dem
 1  2  3  6
10  5  3  1
taboa_dem<-table(contaxe_dem)
prop.table(taboa_dem)
contaxe_dem
        1         2         3         6
0.52631579 0.26315789 0.15789474 0.05263158
```

Para poñer en tanto por cento:

```
round(prop.table(taboa_dem),3)*100
contaxe_dem
  1    2    3    6
52.6 26.3 15.8  5.3
```

Estatística

Exercicio 9: Solución

Imos traballar cos datos «1NT004916.txt» que utilizaramos no exercicio 1.

b) Na variable «contaxe»:

ii. Representación gráfica de cada unha delas

Estatística

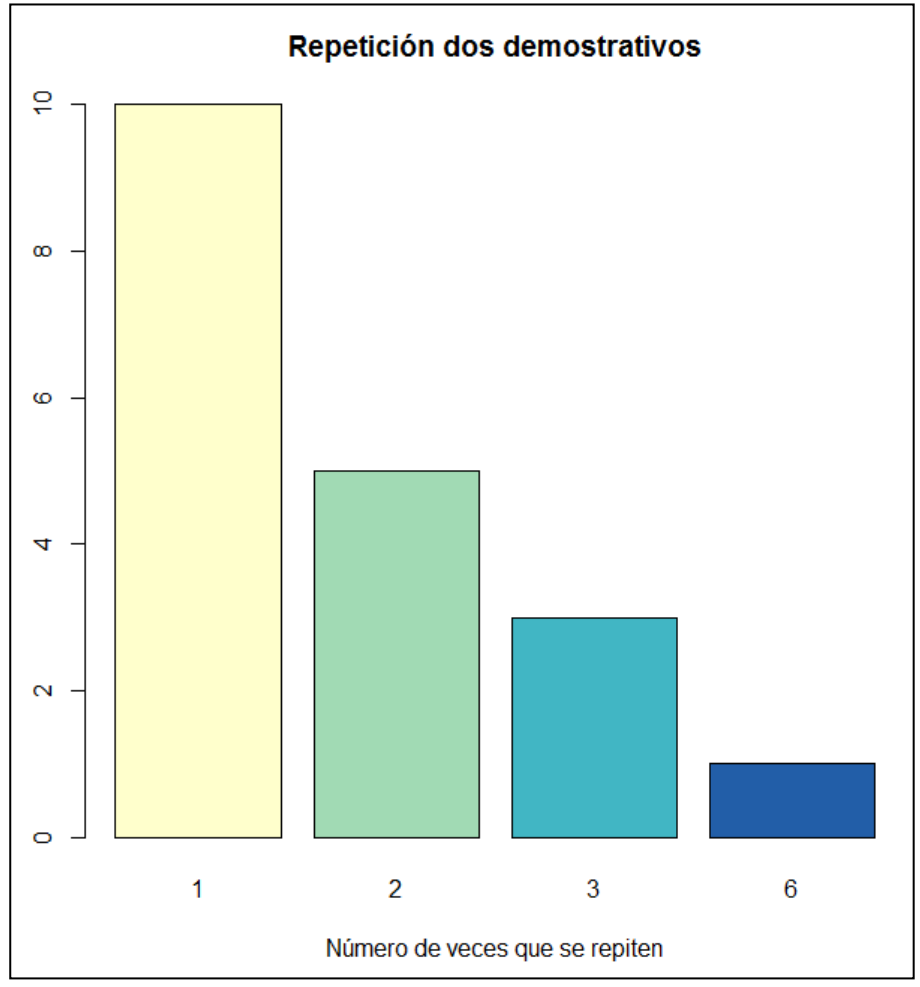
Exercicio 9: Solución

Imos traballar cos datos «1NT004916.txt» que utilizaramos no exercicio 1.

b) Na variable «contaxe»:

ii. Representación gráfica de cada unha delas

```
plot(contaxe_dem,
        col=c("#ffffcc","#a1dab4","#41b6c4","#225ea8"),
        main="Repetición dos demostrativos",
        xlab="Número de veces que se repiten")
```



Estatística

Exercicio 9: Solución

Imos traballar cos datos «1NT004916.txt» que utilizaramos no exercicio 1.

b) Na variable «contaxe»:

ii. Representación gráfica de cada unha delas

Estatística

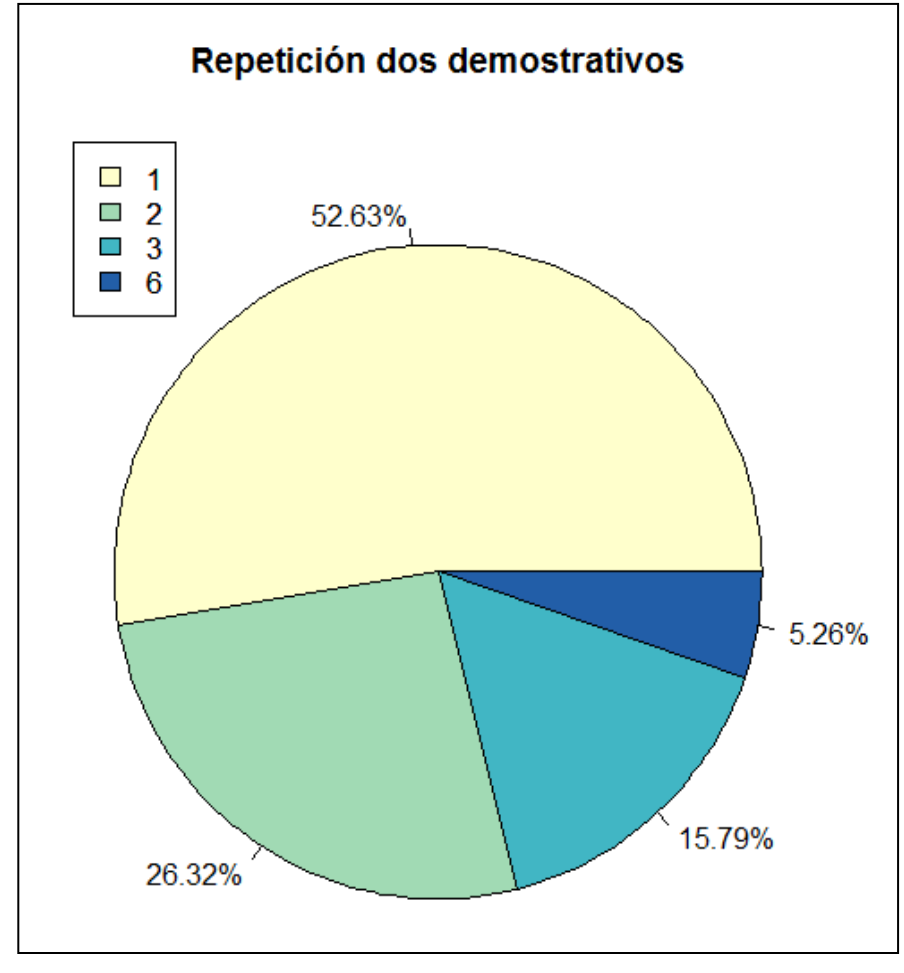
Exercicio 9: Solución

Imos traballar cos datos «1NT004916.txt» que utilizaramos no exercicio 1.

b) Na variable «contaxe»:

ii. Representación gráfica de cada unha delas

```
etiquetas<-prop.table(taboa_dem)*100
etiquetas<-round(etiquetas,2)
etiquetas<-paste(etiquetas,"%",sep="")
etiquetas
[1] "52.63%" "26.32%" "15.79%" "5.26%"
pie(taboa_dem,labels=etiquetas,
            col=c("#ffffcc","#a1dab4","#41b6c4","#225ea8"),
            main="Repetición dos demostrativos")

legend(-0.9,1.05,legend=levels(contaxe_dem),
            fill=c("#ffffcc","#a1dab4","#41b6c4","#225ea8"))
```



Estatística

Exercicio 9: Solución

Imos traballar cos datos «1NT004916.txt» que utilizaramos no exercicio 1.

b) Na variable «contaxe»:

ii. Representación gráfica de cada unha delas

Estatística

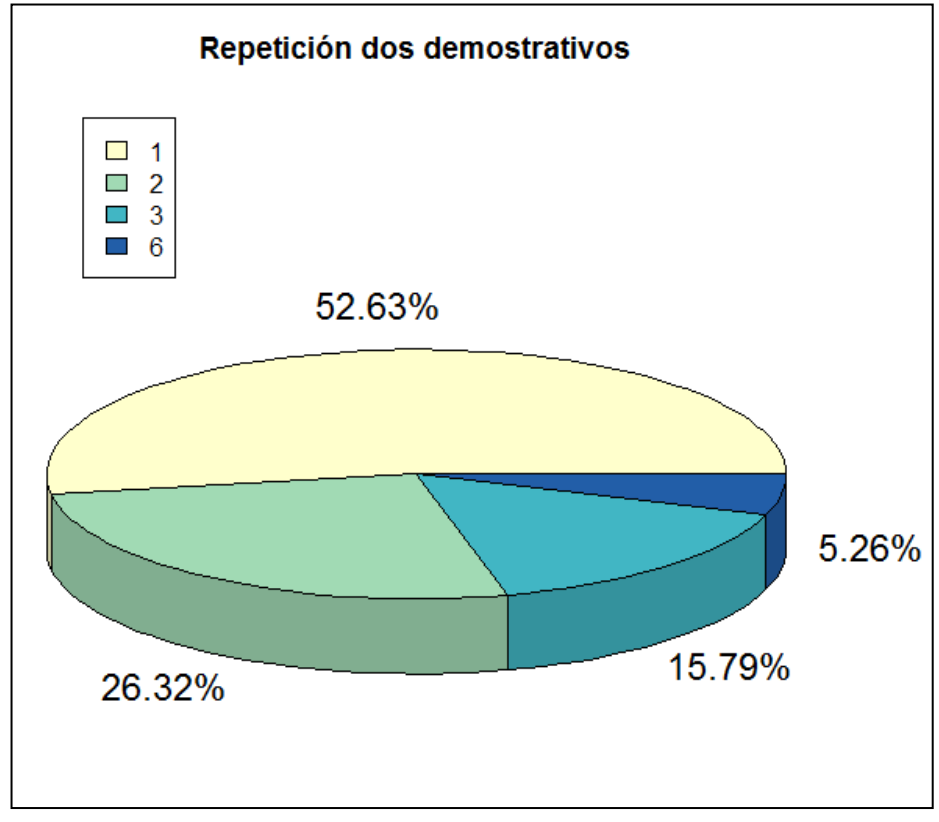
Exercicio 9: Solución

Imos traballar cos datos «1NT004916.txt» que utilizaramos no exercicio 1.

b) Na variable «contaxe»:

ii. Representación gráfica de cada unha delas

```
library(plotrix)
pie3D(taboa_dem,labels=etiquetas,
        col=c("#ffffcc","#a1dab4","#41b6c4","#225ea8"),
        main="Repetición dos demostrativos")

legend(-0.9,1.05,legend=levels(contaxe_dem),
        fill=c("#ffffcc","#a1dab4","#41b6c4","#225ea8"))
```



Estatística

Variables cuantitativas

Cantidades numéricas

- **Cuantitativas discretas: número finito discreto de valores (Número de linguas faladas, número de libros que les nun ano)**

I) Descrición dos datos

II) Representación gráfica

- **Cuantitativas continuas: infinitos valores nun intervalo real (idade, frecuencia dun son…)**

I) Descrición dos datos

II) Representación gráfica

Estatística

I) Descrición dos datos

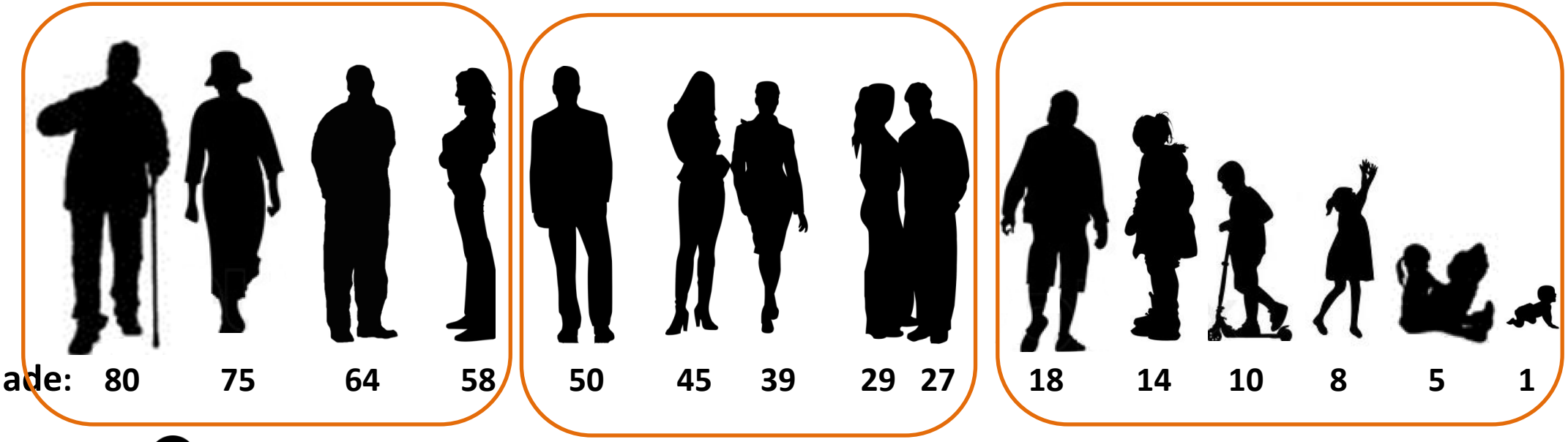
Variables cuantitativas continuas

Idade: 80 75 64 58 50 45 39 29 27 18 14 10 8 5 1

Estatística

I) Descrición dos datos

Variables cuantitativas continuas



> 50

50 - 20

< 20

Variables cuantitativas continuas

Unha **variable cuantitativa** pódese **describir mediante unha táboa de frecuencia agrupando por intervalos**. Ós intervalos chamarémolos **intervalos de clase**.

Consideracións:

- *Número de intervalos a considerar*
- *Amplitude de cada intervalo*
- *Posición dos intervalos*: os intervalos serán contiguos e deberán situarse alí onde se atopen as observacións.

Variables cuantitativas continuas

En R project....

Exemplo:

As idades dos nosos informantes son as seguintes:

20, 21, 22, 51,55, 23, 24, 24, 22, 26, 30, 32, 31, 40, 45

```
idades<-c(20, 21, 22, 51,55, 23, 24, 24, 22, 26, 30, 32, 31, 40, 45)
```

Variables cuantitativas continuas

En R project....

Exemplo:

As idades dos nosos informantes son as seguintes:

20, 21, 22, 51,55, 23, 24, 24, 22, 26, 30, 32, 31, 40, 45

```
idades<-c(20, 21, 22, 51,55, 23, 24, 24, 22, 26, 30, 32, 31, 40, 45)
```

Variable

Extremos dos intervalos

```
rangos<- cut(idades, breaks=c(20,35,56) , include.lowest=T, right = F)
rangos
 [1] [20,35) [20,35) [20,35) [35,56] [35,56] [20,35) [20,35) [20,35) [20,35)
[10] [20,35) [20,35) [20,35) [20,35) [35,56] [35,56]
Levels: [20,35) [35,56]
```

Para que nos intervalos inclúa o extremo inferior pero non o superior

Variables cuantitativas continuas

En R project....

Exemplo:

- Construímos a táboa de frecuencias unha vez que temos os intervalos construídos

Frecuencias absolutas

```
taboa_idades<-table(rangos); taboa_idades
rangos
[20,35) [35,56]
     11       4
```

Frecuencias relativas

```
prop.table(taboa_idades)
rangos
   [20,35)          [35,56]
 0.7333333        0.2666667
```

Variables cuantitativas continuas

Unha **variable cuantitativa** pódese **describir tamén mediante as seguintes medidas estatísticas:**

Medidas de centralización:
- Media
- Mediana
- Moda

Medidas de localización:
- Cuantís (cuartís, decís, percentís...)

Medidas de posición

Medidas de dispersión:
- Rango
- Varianza
- Desviación típica
- Coef. Variación

Variables cuantitativas continuas

Medidas de centralización:

- Media
- Mediana
- Moda

Media mostral: Sexa n o tamaño da nosa mostra e $x_1, x_2, \dots, x_n$ son os valores que toma a nosa variable. A media virá dada por:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Exemplo: *As idades dos nosos informantes son as seguintes 20, 21, 22 e 23. A media das idades virá dada por:*

$$\frac{20+21+22+23}{4} = 21.5$$

Variables cuantitativas continuas

Medidas de centralización:

- Media
- Mediana
- Moda

Mediana: Unha vez ordenados os valores que toma a nosa variable de menor a maior, podemos definir a mediana como aquel valor que nos **deixa a súa esquerda o mesmo número de datos que á súa dereita**. Exemplos:

Se n (tamaño da mostra) par:

Idades dos nosos informantes:

20 21 22 23

Mediana:

$\frac{21+22}{2}$=21,5

Se n impar:

Idades dos nosos informantes:

20 21 22 23 24

Mediana:

22

Variables cuantitativas continuas

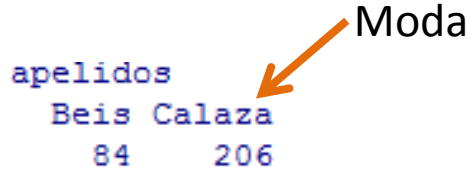
Medidas de centralización:

- Media
- Mediana
- Moda

Moda: Valor da variable que presenta **maior frecuencia**. A diferenza das outras medidas, a moda pode calcularse tamén para variables cualitativas. Pero ao mesmo tempo, non pode calcularse para variables continuas sen agrupación de intervalos por clases. Exemplos:

Variable cualitativa:

Retomamos o exemplo dos apelidos no que tiñamos que:



Variable cuantitativa discreta:

Retomamos o exemplo da aparición do demostrativo na obra:

Número de aparicións	Frecuencia
1	10
2	5
3	3
6	1

Moda

Variable cuantitativa continua:

Poñamos que temos clasificados os informantes en dous grupos de idade:

Moda

Intervalos de idades	Frecuencia
[20,35)	11
[35,56)	4

Variables cuantitativas continuas

Medidas de centralización:

- Media
- Mediana
- Moda

En R project....

- Media

```
mean(idades)
[1] 31.06667
```

- Mediana

```
median(idades)
[1] 26
```

- Moda

```
sort(taboa_idades)
rangos
[35,56]        [20,35)
 4              11      <-- Moda
```

Variables cuantitativas continuas

Unha **variable cuantitativa** pódese **describir tamén mediante as seguintes medidas estatísticas:**

Medidas de centralización:

- Media
- Mediana
- Moda

Medidas de localización:

- Cuantís (cuartís, decís, percentís...)

Medidas de posición

Medidas de dispersión:

- Rango
- Varianza
- Desviación típica
- Coef. Variación

Variables cuantitativas continuas

Medidas de localización:

- Cuantís (cuartís, decís, percentís...)

Cuantís: vimos que a mediana divide os datos en dúas partes iguais. Pero tamén pode ser de interese outros parámetros, os cuantís, que **dividan os datos da distribución en partes iguais**, é dicir, en intervalos que comprendan o mesmo número de valores. Sexa $p \in (0,1)$, defínese o **cuantil p como o número que deixa á súa esquerda unha frecuencia relativa p**.

Algúns teñen nomes específicos:

- Así os **cuartís** son os cuantís de orde (0.25, 0.5, 0.75) e represéntanse por Q1, Q2, Q3. Os cuartís dividen a distribución en catro partes iguais.
- Os **decís** son os cuantís de orde (0.1, 0.2,..., 0.9).
- Os **percentís** son os cuantís de orde j/100, onde j =1,2,...,99.

Variables cuantitativas continuas

Medidas de localización:

- Cuantís (cuartís, decís, percentís...)

En R project....

```
summary(idades)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.00   22.50   26.00   31.07   36.00   55.00
```

Q1 Q2 Q3

Variables cuantitativas continuas

Unha **variable cuantitativa** pódese **describir tamén mediante as seguintes medidas estatísticas:**

Medidas de centralización:
- Media
- Mediana
- Moda

Medidas de localización:
- Cuantís (cuartís, decís, percentís...)

Medidas de posición

Medidas de dispersión:
- Rango
- Varianza
- Desviación típica
- Coef. Variación

Variables cuantitativas continuas

Medidas de dispersión:

- Rango
- Varianza
- Desviación típica
- Coef. Variación

Rango: Sexan $x_1, x_2, \dots, x_n$ son os valores que toma a nosa variable. O rango (ou recorrido) defínese como:

$$\text{Rango} = \max(x_i) - \min(x_i)$$

Exemplo: *As idades dos nosos informantes son as seguintes 20 21 22 23. Logo o rango virá dado por:*

$$Rango = 23 - 20 = 3$$

Variables cuantitativas continuas

Medidas de dispersión:

- Rango
- Varianza
- Desviación típica
- Coef. Variación

Varianza mostral: Unha medida de dispersión que nos permite cuantificar a discrepancia dos datos respecto da media.

Sexan $x_1, x_2, \dots, x_n$ son os valores que toma a nosa variable. Defínese a varianza mostral como:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Exemplo: *As idades dos nosos informantes son as seguintes 20 21 22 23. Logo o rango virá dado por:*

$$s^2 = \frac{1}{4-1}[(20-21.5)^2 + (21-21.5)^2 + (22-21.5)^2 + (23-21.5)^2] = \frac{1}{3}\cdot 5 = 1.66$$

Variables cuantitativas continuas

Medidas de dispersión:

- Rango
- Varianza
- Desviación típica
- Coef. Variación

Desviación típica mostral: medida de dispersión que se expresa na mesma escala que a variable.

$$s = \sqrt{s^2}$$

Exemplo: *As idades dos nosos informantes son as seguintes 20 21 22 23. Logo o rango virá dado por:*

$$s = \sqrt{1.667} = 1.29$$

Variables cuantitativas continuas

Medidas de dispersión:

- Rango
- Varianza
- Desviación típica
- Coef. Variación

Coeficiente de variación: medida de dispersión que non depende da escala (medida relativa) e que, por tanto, pode ser de utilidade cando queremos comparar as dispersións relativas a varias mostras (que non teñen por que estar na mesma escala, como é o caso da desv. típica ou da varianza). Defínese como:

$$\mathrm{CV} = \frac{s}{\bar{x}}$$

Exemplo: *As idades dos nosos informantes son as seguintes 20 21 22 23. Logo o rango virá dado por:*

$$CV = 1.29/21{,}5{=}0.06$$

Variables cuantitativas continuas

Medidas de dispersión:

- Rango
- Varianza
- Desviación típica
- Coef. Variación

En R project....

- Rango

```
max(idades)-min(idades)
[1] 35
```

- Varianza

```
var(idades)
[1] 130.3524
```

- Desviación típica

```
sd(idades)      , ou,   sqrt(var(idades))
 [1] 11.4172            [1] 11.4172
```

- Coeficiente de variación

```
cv<-sd(idades)/mean(idades);cv
[1] 0.3675063
```

Variables cuantitativas continuas

1) Histograma

Gráfico que representa frecuencias mediante áreas. O histograma constrúese colocando no eixe de abscisas os intervalos de clase, como fragmentos da recta real, levantando sobre eles rectángulos con **área proporcional á frecuencia**.

- A diferenza do diagrama de barras, os rectángulos represéntanse contiguos.
- O aspecto do histograma cambia variando o número de clases e o punto onde empeza a primeira clase.
- Canto maior é a área dunha clase, maior é a súa frecuencia.
- O histograma axuda a describir cómo é a distribución da variable, se é simétrica (cun eixe de simetría), bimodal (con dous máximos) etc.

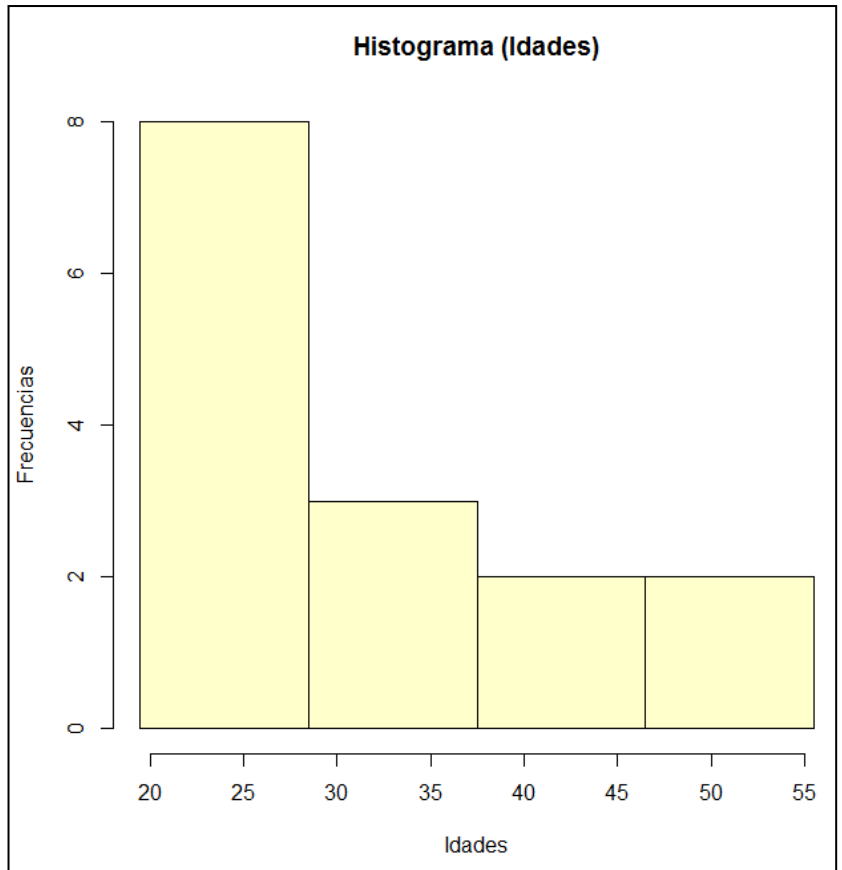
Estatística

II) Representación gráfica

Variables cuantitativas continuas

1) Histograma

hist(idades,breaks=c(19.5,28.5,37.5,46.5,55.5),include.lowest=T,right = F,col=c("#ffffcc"), main="Histograma (Idades)", xlab="Idades",ylab="Frecuencias")



Importante!

Comandos de interese para variables cuantitativas continuas:

hist()

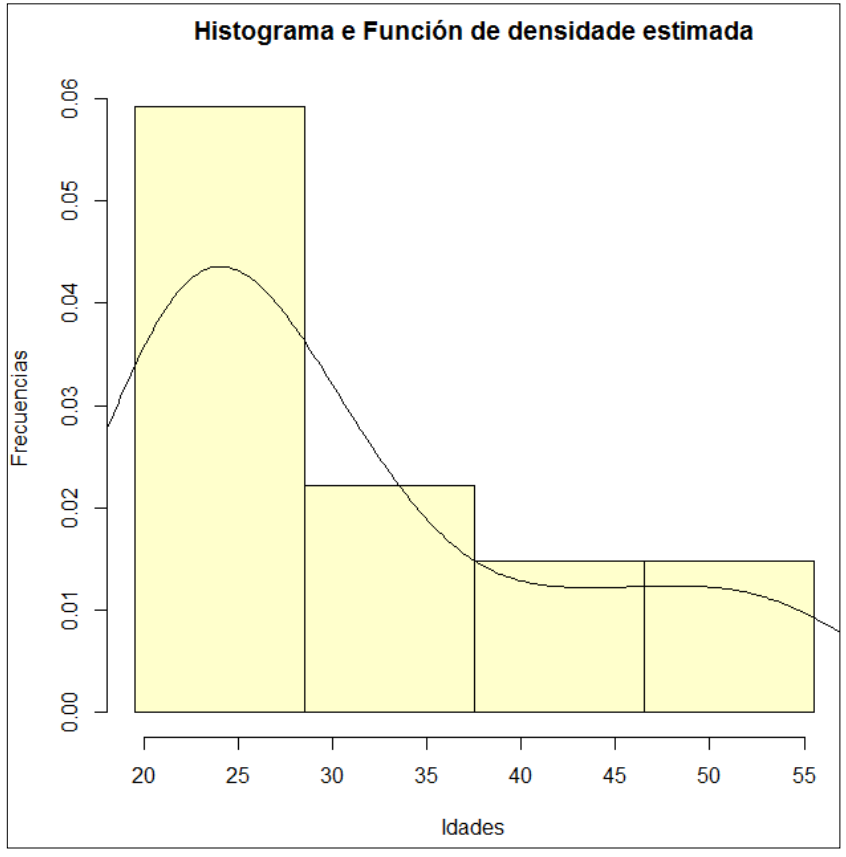
Estatística

II) Representación gráfica

Variables cuantitativas continuas

1) Histograma

hist(idades**,probability=T**,breaks=c(19.5,28.5,37.5,46.5,55.5),include.lowest=T,right = F,col=c("#ffffcc"), main="Histograma e Función de densidade estimada", xlab="Idades“, ylab="Frecuencias")



lines(density(idades))

Importante!

Comandos de interese para variables cuantitativas continuas:

hist()

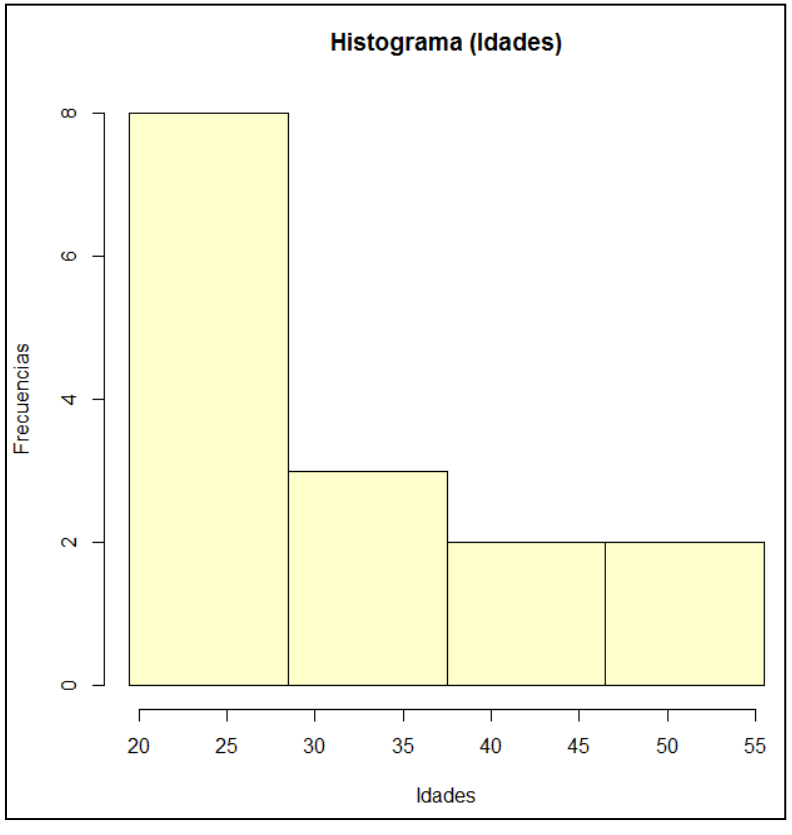
Estatística

II) Representación gráfica

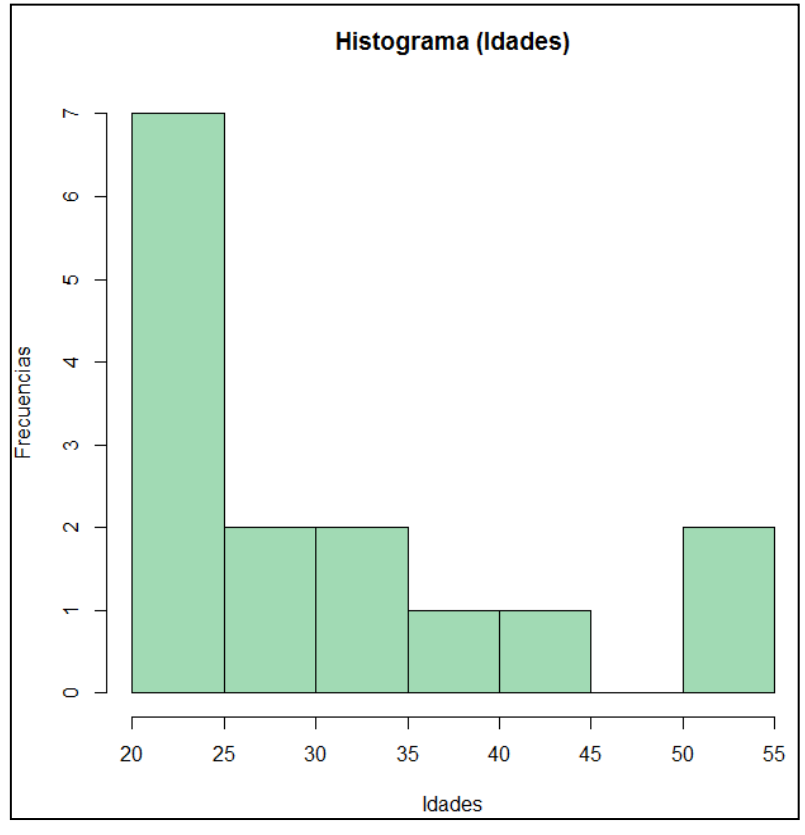
Variables cuantitativas continuas

1) Histograma

```
hist(idades,breaks=c(19.5,28.5,37.5,46.5,55.5),
include.lowest=T,right = F,col=c("#ffffcc"),main="Histograma
(Idades)", xlab="Idades",ylab="Frecuencias")
```




```
hist(idades,col="#a1dab4",main="Histograma (Idades)",
xlab="Idades",ylab="Frecuencias")
```



Variables cuantitativas continuas

2) Diagrama de caixa (Boxplot)

Os diagramas de caixa (boxplots) dannos información visual sobre como están distribuídos os datos. Constan de:

- Unha **caixa central** delimitada polos cuartís **Q1** e **Q3**. Dentro desa caixa debúxase unha **liña** que representa a mediana (cuartil **Q2**).

- Dos extremos da caixa salen unhas liñas (denominadas **bigotes**) que se estenden ata os puntos LI = max{min(xi),Q1 +1.5RI} y LS = mín{max(xi);Q3 + 1:5RI} que representarían o rango razoable ata o cal se poden atopar datos.

- Os datos que caen fóra dos bigotes represéntanse mediante un asterisco, e denomínanse **datos atípicos**.

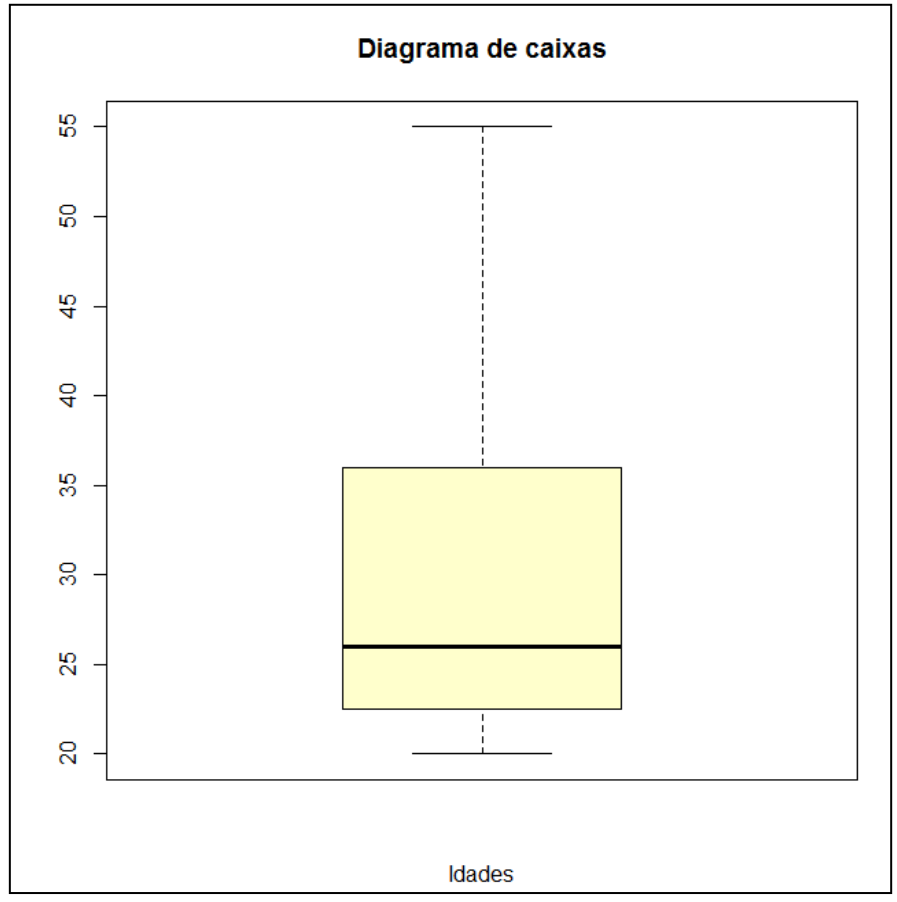
Estatística

II) Representación gráfica

Variables cuantitativas continuas

2) Diagrama de caixa (Boxplot)

```
boxplot(idades,col="#ffffcc",main="Diagrama de caixas",xlab="Idades")
```



Importante!

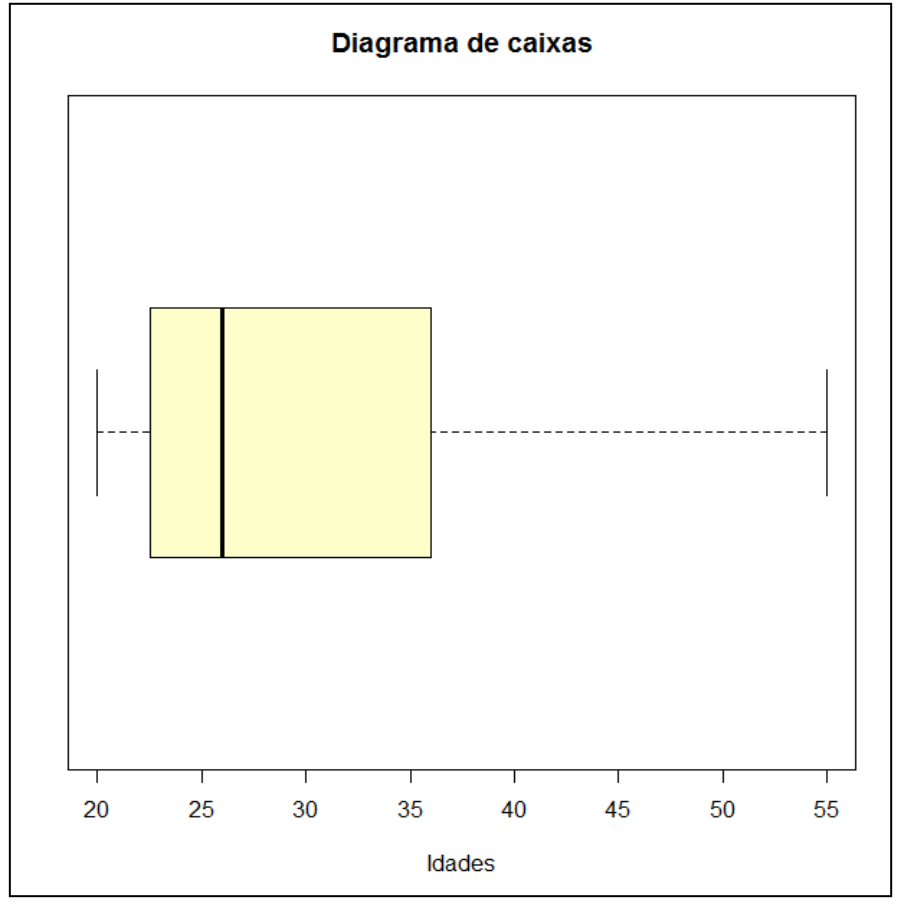
Comandos de interese para variables cuantitativas continuas:

hist()
boxplot()

Variables cuantitativas continuas

2) Diagrama de caixa (Boxplot)

boxplot(idades,col="#ffffcc",main="Diagrama de caixas",xlab="Idades", **horizontal=T)**



Importante!

Comandos de interese para variables cuantitativas continuas:

hist()
boxplot()

Estatística

Exercicio 10

Imos traballar cos datos «frecuencias.csv»

Na base de datos imos ver:

a) De que clase son cada unha das variables?

b) Comprobar a frecuencia media de «f1» en función da vogal «/o/» e da vogal «/e/».

c) Para estes dous casos extraer:
   i. As medidas estatísticas: media, mediana, desviación típica, coef. de variación.
   ii. Representacións gráficas adecuadas.

Estatística

Exercicio 10
Solución

Imos traballar cos datos «frecuencias.csv»

a) De que clase son cada unha das variables?

```
frec<-read.csv2("frecuencias.csv",header=T,sep=";")
View(frec)

attach(frec)
class(SEXO)
[1] "factor"
class(PALABRA)
[1] "factor"
class(vowel)
[1] "factor"

class(stress)
[1] "factor"
class(f1)
[1] "integer"
```

Data: frec

	SUJETO	SEXO	PALABRA	vowel	stress	f1
1	1	woman	comité	/o/	pretonic	446
2	1	woman	emitir	/E/	pretonic	657
3	1	woman	hoxe	/e/	final non st	467
4	1	woman	omitir	/O/	pretonic	749
5	1	woman	pomo	/o/	stressed	485
6	1	woman	remitir	/e/	pretonic	417
7	1	woman	semia	/E/	stressed	449
8	1	woman	toma	/O/	stressed	727
9	1	woman	toma	/O/	stressed	693
10	1	woman	toma	/O/	stressed	747
11	1	woman	toxo	/o/	final non st	461
12	1	woman	xema	/e/	stressed	639
13	2	woman	comité	/o/	pretonic	405
14	2	woman	emitir	/E/	pretonic	544
15	2	woman	hoxe	/e/	final non st	474
16	2	woman	hoxe	/e/	final non st	470
17	2	woman	omitir	/O/	pretonic	466
18	2	woman	pomo	/o/	stressed	446
19	2	woman	remitir	/e/	pretonic	481
20	2	woman	remitir	/e/	pretonic	377
21	2	woman	semia	/E/	stressed	646
22	2	woman	toma	/O/	stressed	686

Estatística

Exercicio 10
Solución

Imos traballar cos datos «frecuencias.csv»

b) Comprobar a frecuencia media de «f1» en función da vogal «/o/» e da vogal «/e/».

Recordatorio

which()

Para a vogal /o/:

```
which(vowel=="/o/")
 [1]  1  5 11 13 18 23 25 29 33 35 39 43 45 49 53 55 59 63 65 69 73 75 79 83
f1[which(vowel=="/o/")]
 [1] 446 485 461 405 446 445 412 421 389 475 467 439 498 463 480 390 500 378 420
[20] 460 392 453 438 419

media<-mean(f1[which(vowel=="/o/")]);media
[1] 440.9167
```

Estatística

Exercicio 10
Solución

Imos traballar cos datos «frecuencias.csv»

b) Comprobar a frecuencia media de «f1» en función da vogal «/o/» e da vogal «/e/».

Recordatorio

which()

Para a vogal /e/:

```
which(vowel=="/e/")
 [1]  3  6 12 15 16 19 20 24 27 30 37 40 44 47 50 54 57 60 64 67 70 74 77 80 84
f1[which(vowel=="/e/")]
 [1] 467 417 639 474 470 481 377 557 480 422 437 435 482 438 461 431 376 424 586
[20] 360 433 485 369 417 455
mean(f1[which(vowel=="/e/")])
[1] 454.92
```

Estatística

Exercicio 10
Solución

Imos traballar cos datos «frecuencias.csv»

c) Para estes dous casos extraer:
   i. As medidas estatísticas: media, mediana, desviación típica, coef. de variación.

Para a vogal /o/:

```
summary(f1[which(vowel=="/o/")])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  378.0   417.2   445.5   440.9   464.0   500.0
```

Q1 Q2 Mediana Media Q3

```
des<-sd(f1[which(vowel=="/o/")]);des
[1] 35.27768
```

```
cv<-des/media;cv
[1] 0.07754699
```

Estatística

Exercicio 10
Solución

Imos traballar cos datos «frecuencias.csv»

c) **Para estes dous casos extraer:**
   i. **As medidas estatísticas: media, mediana, desviación típica, coef. de variación.**

Para a vogal /e/:

```
summary(f1[which(vowel=="/e/")])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  360.0   422.0   438.0   454.9   480.0   639.0

des<-sd(f1[which(vowel=="/e/")]);des
[1] 64.83114

cv<-des/media;cv
[1] 0.1425111
```

Estatística

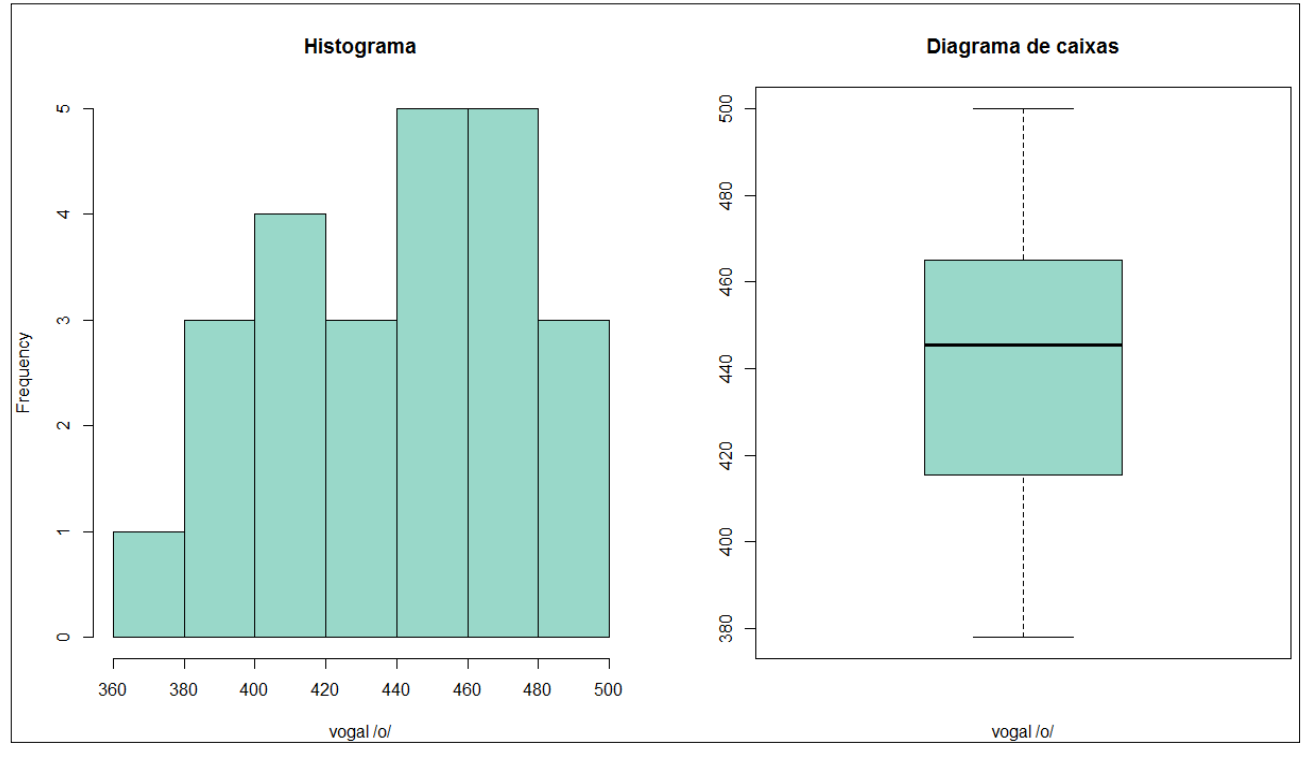
Exercicio 10
Solución

Imos traballar cos datos «frecuencias.csv»

c) Para estes dous casos extraer:
   ii. Representacións gráficas adecuadas

Para a vogal /o/:

```
par(mfrow=c(1,2))
hist(f1[which(vowel=="/o/")],
        col="#99d8c9",main="Histograma",
        xlab="vogal /o/")
boxplot(f1[which(vowel=="/o/")],
        col="#99d8c9",main="Diagrama de
        caixas",xlab="vogal /o/")
```



Estatística

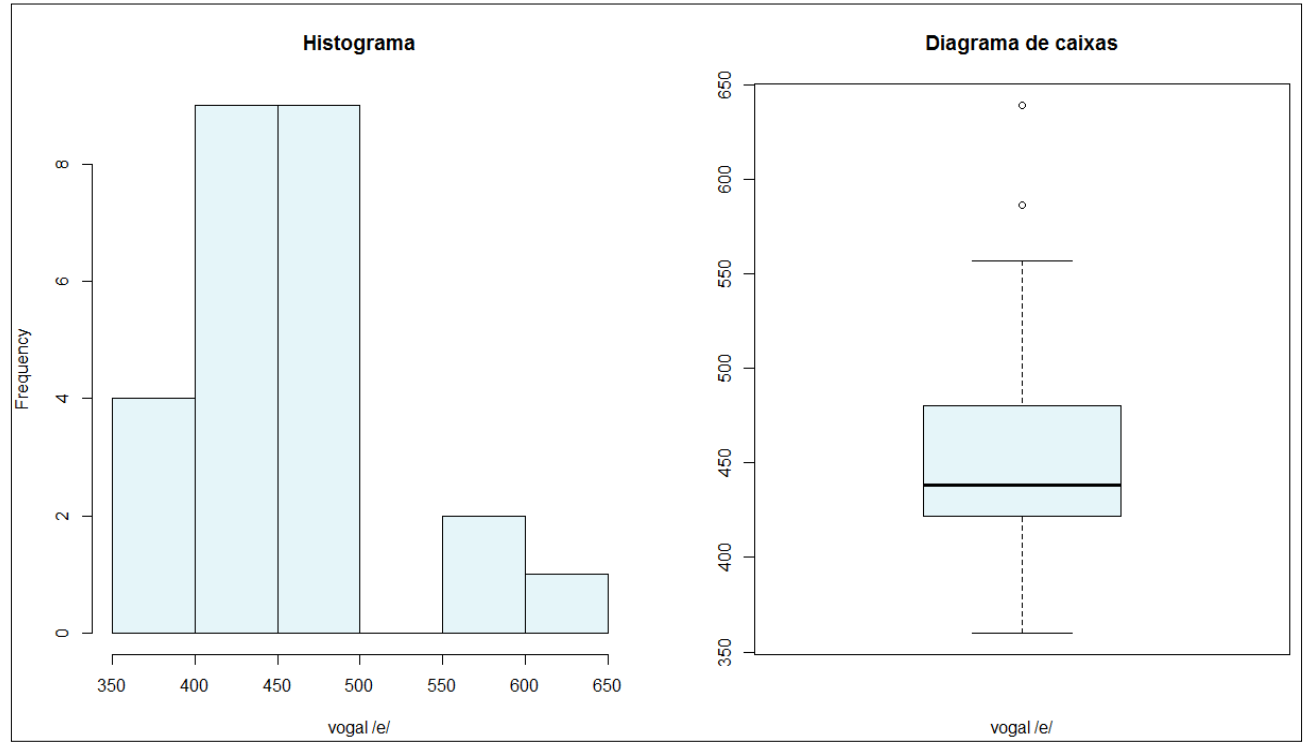
Exercicio 10
Solución

Imos traballar cos datos «frecuencias.csv»

c) Para estes dous casos extraer:
   ii. Representacións gráficas adecuadas

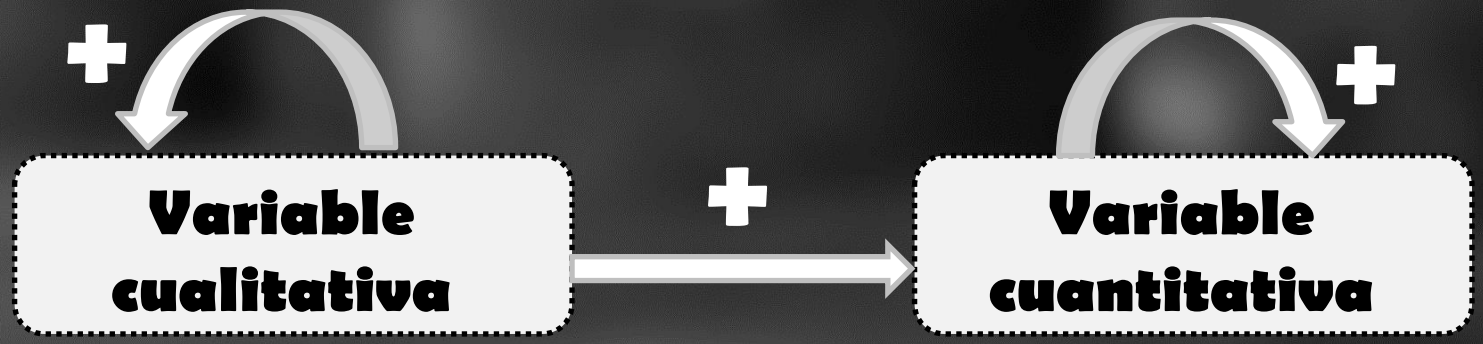
Para a vogal /e/:

```
par(mfrow=c(1,2))
hist(f1[which(vowel=="/e/")],
        col="# e5f5f9",main="Histograma",
        xlab="vogal /e/")
boxplot(f1[which(vowel=="/e/")],
        col="# e5f5f9",main="Diagrama de
        caixas",xlab="vogal /o/")
```



Módulo IV – Estatística descritiva

IV) Descritiva bivariante



Estatística

Descritiva bivariante

Imos estudar conxuntamente pares de variables, que poden ser:

- **Cualitativa + cualitativa**
  - *Táboas de continxencia, barras agrupadas*
- **Cualitativa + cuantitativa**
  - *Boxplots segregados polas categorías da variable cualitativa*
- **Cuantitativa + cuantitativa**
  - *Diagramas de dispersión*

Estatística

Descritiva bivariante

Imos estudar conxuntamente pares de variables, que poden ser:

- Cualitativa + cualitativa
  - *Táboas de continxencia, barras agrupadas*
- Cualitativa + cuantitativa
  - *Boxplots segregados polas categorías da variable cualitativa*
- Cuantitativa + cuantitativa
  - *Diagramas de dispersión*

Estatística

- **Cualitativa + cualitativa**
  - ***Táboas de continxencia**, onde se recollan as distribución de frecuencias das variables.*

Exemplo

Lembramos os datos: **«tempos_compostos_galego_medieval.csv»**

A táboa de continxencia no que se recollan as frecuencias conxuntas dos verbos segregados por «tipo_de_verbo» e «auxiliar».

```
tab_continxencia=table(tipo_de_verbo,auxiliar)
addmargins(tab_continxencia)
```

	auxiliar		
tipo_de_verbo	aver	ser	**Sum**
paso_de_tempo	1	1	2
permanencia	4	1	5
procesos_fisicos	1	9	10
suceso	3	1	4
Sum	9	12	21

Importante!

Comandos de interese:

table() *: table(variable1,variable2)*

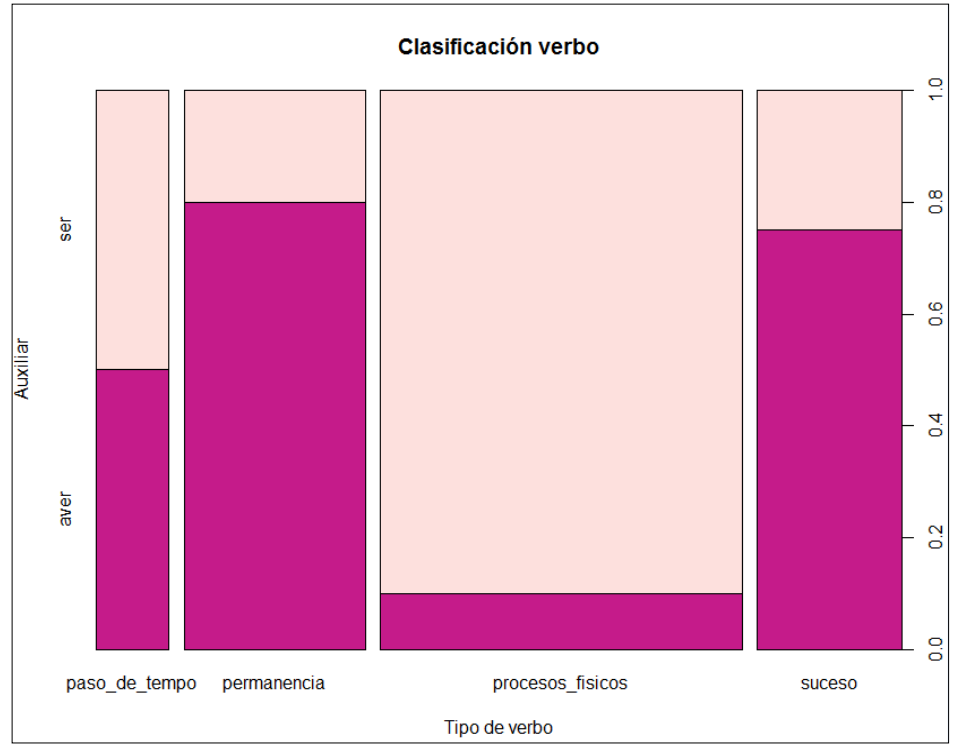
Estatística

- **Cualitativa + cualitativa**
  - ***Diagrama de barras**, onde se recollan as distribución de frecuencias das variables.*

Exemplo

O diagrama de barras no que se recollan as frecuencias conxuntas dos verbos segregados por «tipo_de_verbo» e «auxiliar».

```
plot(tipo_de_verbo,auxiliar,main="Clasificación
verbo",xlab="Tipo de verbo",ylab="Auxiliar",
col=c("#c51b8a","#fde0dd"))
```



Importante!

Comandos de interese:

***table()** : table(variable1,variable2)*
***plot()** : plot(variable1,variable2)*

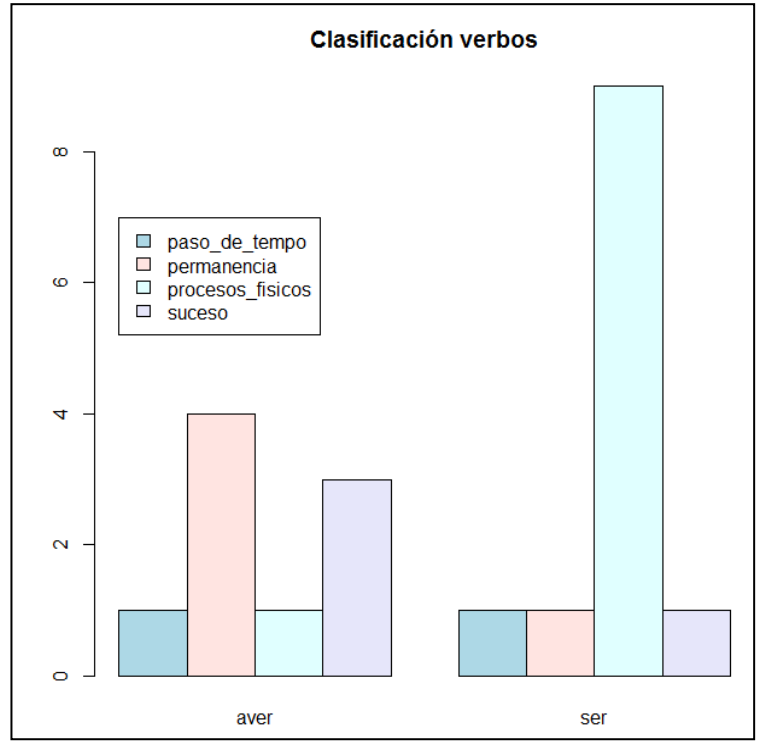
Estatística

- **Cualitativa + cualitativa**
  - ***Diagrama de barras**, onde se recollan as distribución de frecuencias das variables.*

Exemplo

O diagrama de barras no que se recollan as frecuencias conxuntas dos verbos segregados por «tipo_de_verbo» e «auxiliar».

```
barplot(tab_continxencia, beside = TRUE,
    col = c("lightblue", "mistyrose", "lightcyan","lavender"))
legend(1,7,rownames(tab_continxencia), fill=c("lightblue",
            "mistyrose", "lightcyan", "lavender"))
title(main = "Clasificación verbos")
```



Importante!

Comandos de interese:

***table()** : table(variable1,variable2)*
***plot()** : plot(variable1,variable2)*
***barplot()**: barplot(taboacontinxencia)*

Estatística

Descritiva bivariante

Imos estudar conxuntamente pares de variables, que poden ser:

- **Cualitativa + cualitativa**
  - *Táboas de continxencia, barras agrupadas*
- **Cualitativa + cuantitativa**
  - *Boxplots segregados polas categorías da variable cualitativa*
- **Cuantitativa + cuantitativa**
  - *Diagramas de dispersión*

Estatística

Descritiva bivariante

Imos estudar conxuntamente pares de variables, que poden ser:

- **Cualitativa + cualitativa**
  - *Táboas de continxencia, barras agrupadas*
- **Cualitativa + cuantitativa**
  - *Boxplots segregados polas categorías da variable cualitativa*
- **Cuantitativa + cuantitativa**
  - *Diagramas de dispersión*

Estatística

- **Cualitativa + cuantitativa**
  - *Boxplot (diagrama de caixa) segundo cada categoría*

Exemplo
Lembramos os datos: **«frecuencias.csv»**

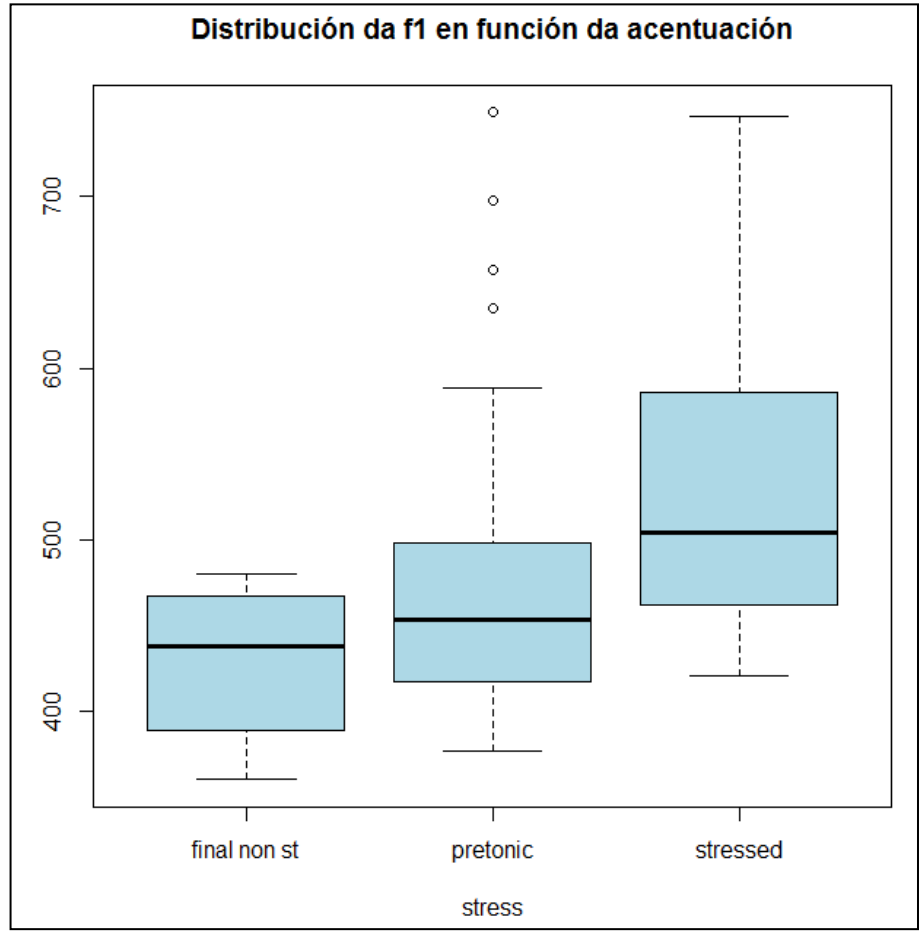
Data: frec

	SUJETO	SEXO	PALABRA	vowel	stress	f1
1	1	woman	comité	/o/	pretonic	446
2	1	woman	emitir	/E/	pretonic	657
3	1	woman	hoxe	/e/	final non st	467
4	1	woman	omitir	/O/	pretonic	749
5	1	woman	pomo	/o/	stressed	485
6	1	woman	remitir	/e/	pretonic	417
7	1	woman	semia	/E/	stressed	449
8	1	woman	toma	/O/	stressed	727
9	1	woman	toma	/O/	stressed	693
10	1	woman	toma	/O/	stressed	747
11	1	woman	toxo	/o/	final non st	461
12	1	woman	xema	/e/	stressed	639
13	2	woman	comité	/o/	pretonic	405
14	2	woman	emitir	/E/	pretonic	544
15	2	woman	hoxe	/e/	final non st	474
16	2	woman	hoxe	/e/	final non st	470
17	2	woman	omitir	/O/	pretonic	466
18	2	woman	pomo	/o/	stressed	446
19	2	woman	remitir	/e/	pretonic	481
20	2	woman	remitir	/e/	pretonic	377
21	2	woman	semia	/E/	stressed	646
22	2	woman	toma	/O/	stressed	686

Estatística

- **Cualitativa + cuantitativa**
  - *Boxplot (diagrama de caixa) segundo cada categoría*

Exemplo
Lembramos os datos: **«frecuencias.csv»**



boxplot(f1~stress,xlab="stress",main="Distribución da f1 en función da acentuación",col="lightblue")

Importante!

Comandos de interese:

boxplot() *: boxplot(var.cuantitativa ~ var. cualitativa)*

Estatística

Descritiva bivariante

Imos estudar conxuntamente pares de variables, que poden ser:

- **Cualitativa + cualitativa**
  - *Táboas de continxencia, barras agrupadas*
- **Cualitativa + cuantitativa**
  - *Boxplots segregados polas categorías da variable cualitativa*
- **Cuantitativa + cuantitativa**
  - *Diagramas de dispersión*

Estatística

Descritiva bivariante

Imos estudar conxuntamente pares de variables, que poden ser:

- **Cualitativa + cualitativa**
  - *Táboas de continxencia, barras agrupadas*
- **Cualitativa + cuantitativa**
  - *Boxplots segregados polas categorías da variable cualitativa*
- **Cuantitativa + cuantitativa**
  - *Diagramas de dispersión*

Estatística

- **Cuantitativa + cuantitativa**
  - *Diagramas de dispersión (coñecer a relación existente entre dúas variables)*

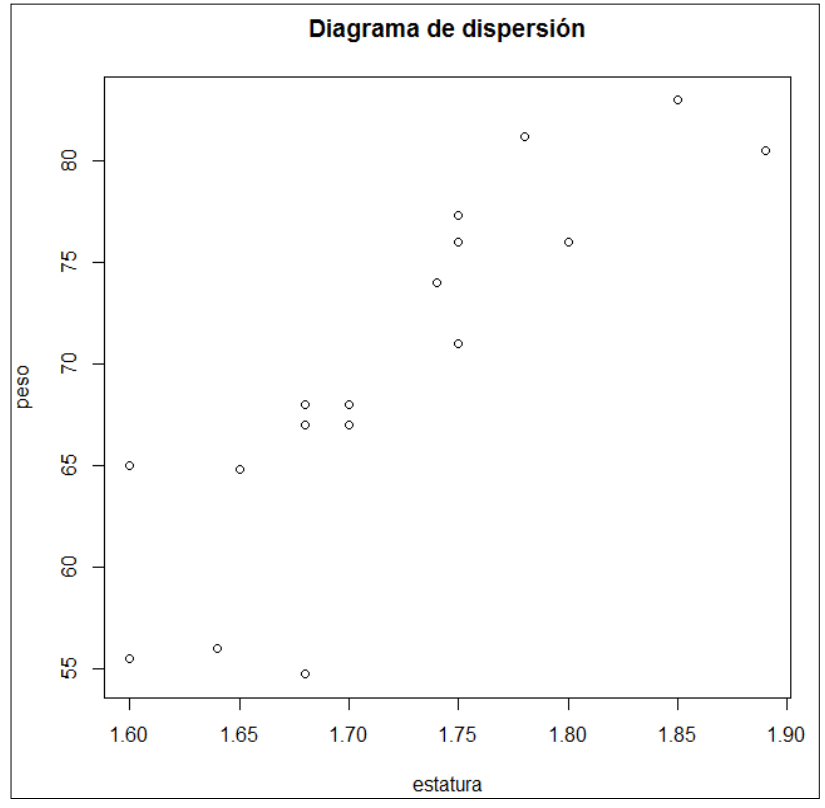
Exemplo

Poñamos que temos a seguinte información na nosa mostra:

```
peso<-c(55.5,65,54.7,64.8,81.2,76,77.3,68,80.5,68,56,76,83,71,67,67,74)
estatura<-c(1.60,1.60,1.68,1.65,1.78,1.75,1.75,1.70,1.89,1.68,1.64,1.8,1.85,1.75,1.70,1.68,1.74)
```

Representación gráfica:

```
plot(peso~estatura,main="Diagrama de dispersión")
```



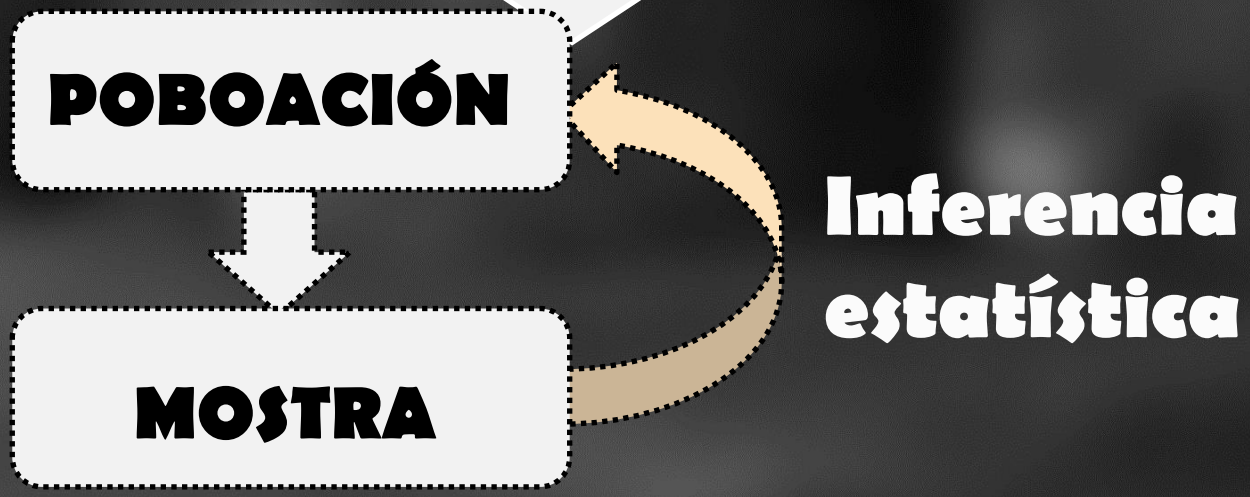
Importante!

Comandos de interese:

plot() *: plot(var.continua ~ var.continua)*

Módulo V – Estatística inferencial

I) Introdución



Estatística

I) Introdución

Ata agora…

- Describimos
- Visualizamos

vimos o que sucede na MOSTRA

Pero que sucede na POBOACIÓN?

- **Analizar**
- **Interpretar**
- **Tomar decisións**
- **Resolver problemas**
- **Extrapolar resultados**

Estatística

I) Introdución

Ata agora…

- Describimos
- Visualizamos

vimos o que sucede na MOSTRA

Pero que sucede na POBOACIÓN?

- **Analizar**
- **Interpretar**
- **Tomar decisións**
- **Resolver problemas**
- **Extrapolar resultados**

Estatística inferencial

Estatística

I) Introdución

Estatística inferencial

PARA QUE?

- **Cal é o valor dun certo "parámetro" ou característica da poboación?** ***(ESTIMACIÓN,*** *puntual ou por intervalos****)***
- **É "tal hipótese" certa á vista dos datos?** ***(CONTRASTES)***

A partir da estimación e dos contrastes, o investigador pode tratar de construír modelos (distribución ou modelos de regresión) que permitan **explicar o comportamento da poboación e facer predicións.**

Módulo V – Estatística inferencial

II) Inferencia

I) Estimación puntual

II) Intervalos de confianza

III) Contrastes de hipóteses

Módulo V – Estatística inferencial

II) Inferencia

I) Estimación puntual

II) Intervalos de confianza

III) Contrastes de hipóteses

Estatística

I) Estimación puntual

Cal é o valor dunha certa característica da POBOACIÓN?

Un **parámetro** (θ) representa unha característica que nos interesa coñecer da poboación.

A **estimación puntual** dun parámetro descoñecido, θ, consiste en aproximar o seu valor, $\hat{\theta}$, a partir dunha mostra.

Exemplos de estimación puntual:

- Da proporción : *Cal é a proporción de falantes de galego en Galicia?*
- Da media: *Cal é o promedio da idade dos galegos?*
- Da varianza: *Que dispersión teñen...?*

Estimación puntual:

- Da proporción

  Dada unha mostra formada por unha variable X, na que se recolleron un total de n rexistros, definimos a proporción mostral como:

$$\hat{p} = \frac{\text{número de individuos que cumplen unha determinada característica } X}{n}$$

Estimación puntual:

- Da media

  Dada unha mostra formada por unha variable X, e sexan $x_1, x_2, \dots, x_n$ os valores que toma a nosa variable, definimos a media mostral como:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Estimación puntual:

- Da varianza

  Estimaremos a varianza mediante a cuasivarianza mostral. Dada unha mostra formada por unha variable X, e sexan $x_1, x_2, \dots, x_n$ os valores que toma a nosa variable, definimos a cuasivarianza mostral como:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Módulo V – Estatística inferencial

II) Inferencia

I) Estimación puntual

II) Intervalos de confianza

III) Contrastes de hipóteses

Motivación:

A estimación puntual dun parámetro vainos dar un *valor aproximado* do verdadeiro valor do parámetro poboacional.

Pero pode resultar de interese obter un **rango de valores no que se sitúe ese parámetro cunha certa «probabilidade de acerto»**.

Intervalos de confianza

Exemplo:

Poñamos que tras obter o peso do alumnado dunha clase universitaria

$$n = 100\ alumnos\ e\ alumnas$$

observamos que:

$$\bar{x} = 71\ \text{kg}$$

$$s = 15\ \text{kg}\ (\ s^2 = 225)$$

$$\text{rango} = 96\ \text{kg} - 54\ \text{kg}$$

Cal é o intervalo de confianza para a media?

Exemplo:

Poñamos que tras obter o peso do alumnado dunha clase universitaria

$$n = 100\ alumnos\ e\ alumnas$$

observamos que:

$$\bar{x} = 71\ \text{kg}$$

$$s = 15\ \text{kg}\ (\ s^2 = 225)$$

$$\text{rango} = 96\ \text{kg} - 54\ \text{kg}$$

Cal é o intervalo de confianza para a media?

$$\left[71 - confianza \cdot \frac{15}{\sqrt{100}}, 71 + confianza \cdot \frac{15}{\sqrt{100}}\right]$$

Estatística

II) Intervalos de confianza

Exemplo:

Supoñamos que queremos obter o Intervalo de Confianza (IC) para a media:

$$\bar{x} \pm confianza \cdot ET\ (\bar{x})$$

onde ,

$$ET\ (\bar{x}) = \frac{s}{\sqrt{n}}$$

$confianza$ - cantidade que representará a probabilidade de acerto (xeralmente esa probabilidade tomarase dun 90%, 95%, ou 99%)

Estatística

II) Intervalos de confianza

Un intervalo de confianza vai vir dado por:

$$Estatístico \pm confianza \cdot ET(Estatístico)$$

onde,

$confianza$ - cantidade que representará a probabilidade de acerto (xeralmente esa probabilidade tomarase dun 90%, 95%, ou 99%)

Módulo V – Estatística inferencial

II) Inferencia

I) Estimación puntual

II) Intervalos de confianza

III) Contrastes de hipóteses

Motivación:

Cando queremos **comprobar** se, á vista dos datos, se cumpre **unha hipótese** que se emite acerca dun parámetro ou outra característica da poboación.

Contrastes de hipóteses

Contrastes de hipóteses

- **Hipótese nula (H_0),** a que se dá por certa. Goza de presunción de inocencia.

- **Hipótese alternativa (H_1),** a que sucede cando non é certa a hipótese nula. Por gozar a hipótese nula de presunción de inocencia, é na hipótese alternativa onde recae a carga da proba.

Nota

Rexeitamos H_0 a favor de H_1 se atopamos probas significativas nos datos a favor de H_1.

Estatística

III) Contrastes de hipóteses

- **Hipótese (H_0):** o peso medio dos alumnos universitarios é de 89 kg
- **Mostra:** x_1 , …, x_n , n alumnos/as

  Obtemos:

  a) $\bar{x} = 50\ kg$

Estatística

III) Contrastes de hipóteses

- **Hipótese (H_0):** o peso medio dos alumnos universitarios é de 89 kg
- **Mostra:** x_1 , …, x_n , n alumnos/as

  Obtemos:

  a) $\bar{x} = 50\ kg$ **Teño razón na miña hipótese?**

Estatística

III) Contrastes de hipóteses

- **Hipótese (H_0):** o peso medio dos alumnos universitarios é de 89 kg
- **Mostra:** x_1 , …, x_n , n alumnos/as

  Obtemos:

  a) $\bar{x} = 50\ kg$

  b) $\bar{x} = 70\ kg$ **Teño razón na miña hipótese?**

Estatística

III) Contrastes de hipóteses

- **Hipótese (H_0):** o peso medio dos alumnos universitarios é de 89 kg
- **Mostra**: x_1, …, x_n, n alumnos/as

  Obtemos:

  a) $\bar{x} = 50\ kg$

  b) $\bar{x} = 70\ kg$

  c) $\bar{x} = 85\ kg$, $\bar{x} = 90\ kg$ **Teño razón na miña hipótese?**

Estatística

III) Contrastes de hipóteses

- **Hipótese (H_0):** o peso medio dos alumnos universitarios é de 89 kg
- **Mostra**: x_1, …, x_n, n alumnos/as

  Obtemos:

  a) $\bar{x} = 50\ kg$

  b) $\bar{x} = 70\ kg$

  c) $\bar{x} = 85\ kg$, $\bar{x} = 90\ kg$ **Teño razón na miña hipótese?**

Que é o que está influíndo na comprobación da hipótese:

1) Como de preto estou de $\bar{x}$ para decidir se teño razón ou non?
2) O tamaño da mostra

Estatística

III) Contrastes de hipóteses

Por iso imos definir…

- **Estatístico de contraste:**

  Medida de discrepancia entre a miña hipótese e o que observamos na mostra:

$H_0: \mu_0 = 89\ kg$ (peso medio universitarios é de 89 kg)

$T = \mu_0 - \overline{x}$, **ou ben,** $T = \overline{x} - \mu_0$ (estatístico de contraste)

Estatística

III) Contrastes de hipóteses

Por iso imos definir…

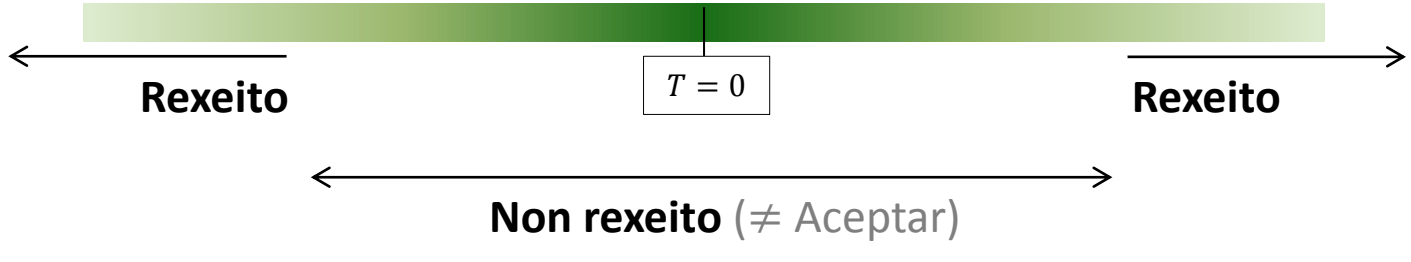
- **Estatístico de contraste:**

  Medida de discrepancia entre a miña hipótese e o que observamos na mostra:

$H_0: \mu_0 = 89\ kg$ (peso medio universitarios é de 89 kg)

$T = \mu_0 - \bar{x}$, **ou ben,** $T = \bar{x} - \mu_0$ (estatístico de contraste)

p.valor



Estatística

III) Contrastes de hipóteses

Procedemento do contraste:

1. **H_0 (H_1) : que é o que quero comprobar?**

   Conclusión: rexeito / non hai evidencias para rexeitar

2. **Definir T (discrepancia)**

3. **T grande/pequeno:**

   T grande ⟶ «Rexeitas»

   T pequeno ⟶ «Non rexeitas»

Estatística

III) Contrastes de hipóteses

Procedemento do contraste:

1. **H_0 (H_1) : que é o que quero comprobar?**

   Conclusión: rexeito / non hai evidencias para rexeitar

2. **Definir T (discrepancia)**
3. **T grande/pequeno:**

   T grande ⟶ «Rexeitas»

   T pequeno ⟶ «Non rexeitas»

4. **Como definimos T grande/pequeno?**

   Pau Gasol é alto?

Estatística

III) Contrastes de hipóteses

Procedemento do contraste:

1. **H_0 (H_1) : que é o que quero comprobar?**

   Conclusión: rexeito / non hai evidencias para rexeitar

2. **Definir T (discrepancia)**
3. **T grande/pequeno:**

   T grande ⟶ «Rexeitas»

   T pequeno ⟶ «Non rexeitas»

4. **Como definimos T grande/pequeno?**

   Pau Gasol é alto? - Si, porque a probabilidade de atopar alguén máis alto é pequena

Estatística

III) Contrastes de hipóteses

Procedemento do contraste:

1. $\mathbf{H_0}$ **(**$\mathbf{H_1}$**) : que é o que quero comprobar?**

   Conclusión: rexeito / non hai evidencias para rexeitar

2. **Definir T (discrepancia)**
3. **T grande/pequeno:**

   T grande ⟶ «Rexeitas»

   T pequeno ⟶ «Non rexeitas»

4. **Como definimos T grande/pequeno?**

   Pau Gasol é alto? - Si, porque a probabilidade de atopar alguén máis alto é pequena

5. **Regra de decisión:**

   p.valor < α ⟶ **Rexeito**

   p.valor > α ⟶ **Non rexeito**

Estatística

III) Contrastes de hipóteses

Procedemento do contraste:

1. $\mathbf{H_0}$ **(**$\mathbf{H_1}$**) : que é o que quero comprobar?**

   Conclusión: rexeito / non hai evidencias para rexeitar

2. **Definir T (discrepancia)**
3. **T grande/pequeno:**

   T grande → «Rexeitas»

   T pequeno → «Non rexeitas»

4. **Como definimos T grande/pequeno?**

   Pau Gasol é alto? - Si, porque a probabilidade de atopar alguén máis alto é pequena

5. **Regra de decisión:**

   p.valor < α **→ Rexeito**

   p.valor > α **→ Non rexeito**

Como definir α ?

	«non rexeitar»	«rexeitar»
H_0 certa		Erro (II)
H_0 falsa	Erro (I)	

Como definir α ?

Exemplo: Xuízo

H_0: O acusado é inocente

			Realidade	
			Inocente	*Culpable*
			«non rexeitar»	«rexeitar»
Veredicto	*Inocente- Liberdade*	H_0 certa	👍	Erro (II)
	Culpable - Cárcere	H_0 falsa	Erro (I)	👍

Como definir α ?

Exemplo: Xuízo

H_0: O acusado é inocente

			Realidade	
			Inocente	*Culpable*
			«non rexeitar»	«rexeitar»
Veredicto	*Inocente- Liberdade*	H_0 certa	👍	Erro (II)
	Culpable - Cárcere	H_0 falsa	Erro (I)	👍

Que é máis grave?

Erro I $\longrightarrow$ Inocente – Cárcere (condenar un inocente)

Erro II $\longrightarrow$ Culpable – Liberdade (absolver un culpable)

Como definir α ?

Exemplo: Xuízo

H_0: O acusado é inocente

			Realidade	
			Inocente	*Culpable*
			«non rexeitar»	«rexeitar»
Veredicto	*Inocente- Liberdade*	H_0 certa	👍	Erro (II)
	Culpable - Cárcere	H_0 falsa	Erro (I)	👍

Que é máis grave?

Erro I ⟶ Inocente – Cárcere (condenar un inocente)

Erro II ⟶ Culpable – Liberdade (absolver un culpable)

Como definir α ?

Exemplo: Xuízo

H_0: O acusado é inocente

			Realidade	
			Inocente	*Culpable*
			«non rexeitar»	«rexeitar»
Veredicto	*Inocente- Liberdade*	H_0 certa	👍	Erro (II)
	Culpable - Cárcere	H_0 falsa	Erro (I)	👍

Que é máis grave?

Erro I ⟶ Inocente – Cárcere (condenar un inocente)

Erro II ⟶ Culpable – Liberdade (absolver un culpable)

$\alpha = P$ **(Cárcere / Inocente)**

$\alpha = P$ (Erro I)

Procedemento do contraste:

1. $\mathbf{H_0}$ **(**$\mathbf{H_1}$**) : que é o que quero comprobar?**

   Conclusión: rexeito / non hai evidencias para rexeitar

2. **Definir T (discrepancia)**
3. **T grande/pequeno:**

   T grande ⟶ **«Rexeitas»**

   T pequeno ⟶ **«Non rexeitas»**

5. **Regra de decisión:**

   p.valor < α ⟶ **Rexeito**

   p.valor > α ⟶ **Non rexeito**

   con α **= P(error I), normalmente 1%, 5%, ou 10%.**

Bibliografía

Rasinger, S.M. (2008). *Quantitative Research in Linguistics. An introduction.* Research Methods in Linguistics

Gries, S.Th. (2009). *Statistics for Linguistics with R. A practical introduction.* De gruyter

Levshina, N. (2015). *How to do Linguistics with R. Data exploration and statistical analysis.* John benjamins Publishing Company.


GRAZAS!
Estatística
Filoloxía