Preparación

Carlos Javier Rincón

versión:2019

Contents

Preparación de datos **1**
Re-nombrar . . . 2
Filtros. . . . 2
Ordenar . . . 3
Seleccionar . . . 4
Nuevas variables . . . 5
Resumen . . . 7
Combinación de "verbos" . . . 7

Ejercicio: **8**

Preparación de datos

En esta guía se trabajará con el paquete `tidyverse` para la preparación de una base de datos.

```
#install.packages("tidyverse")
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.1
```

Utilizaremos el archivo RN (recién nacidos) que contiene las siguientes variables: peso, talla, cuidador (madres, padre, ambos), número de controles, lactancia (Si=1 , No = 0), sexo del recién nacido (H=0, M=1), horas de sueño y estrato socio económico (1 a 6).

Carguemos la base de datos:

```
library(rio)
```

```
## Warning: package 'rio' was built under R version 3.6.1
```

```
RN<-import("RN.xlsx")
```

Listado de variables y su estructura:

```
names(RN) # Nombre variables
```

```
## [1] "id"        "peso"      "talla"     "cuidador"  "controles" "lactancia"
## [7] "sexo"      "sueno"     "estrato"
```

```
str(RN)  # Estructura
```

```
## 'data.frame':    369 obs. of  9 variables:
##  $ id       : num  1 2 3 4 5 6 7 9 11 12 ...
##  $ peso     : num  0.27 1.48 0.83 0.52 0.25 2.41 1.13 2.08 0.81 1.28 ...
##  $ talla    : num  46.1 48.2 40.7 56.3 51.5 44.5 45 37.6 43.5 45.2 ...
##  $ cuidador : chr  "madre" "madre" "madre" "madre" ...
##  $ controles: num  5 8 6 6 4 7 5 12 3 3 ...
##  $ lactancia: chr  "No" "Si" "Si" "No" ...
```

```
##  $ sexo     : chr  "M" "M" "M" "M" ...
##  $ sueno    : num  9.4 7.7 6.2 6.9 6.7 8.9 8.7 4.5 8.4 4.6 ...
##  $ estrato  : chr  "E5" "E1" "E4" "E1" ...
```

- Principales funciones (verbos) del paquete `dplyr` incluido en `tidyverse`:

Re-nombrar

Cambiar el nombre de las variables

```
RN2<-rename(RN,EST=estrato,identificador=id)
```

Filtros.

Extraer una submuestra por los valores específicos de una o varias variable.

```
F1<-filter(RN,sexo=="H")
F1[1:5,]
```

```
##   id peso talla cuidador controles lactancia sexo sueno estrato
## 1  6 2.41  44.5    madre         7        No    H   8.9      E5
## 2  9 2.08  37.6    madre        12        No    H   4.5      E1
## 3 11 0.81  43.5    madre         3        Si    H   8.4      E4
## 4 14 1.62  45.4    madre         3        No    H   6.3      E5
## 5 17 1.74  54.6    madre         1        Si    H   4.8      E2
```

```
F2<-filter(RN,sexo=="H" & cuidador=="madre")
F2[1:5,]
```

```
##   id peso talla cuidador controles lactancia sexo sueno estrato
## 1  6 2.41  44.5    madre         7        No    H   8.9      E5
## 2  9 2.08  37.6    madre        12        No    H   4.5      E1
## 3 11 0.81  43.5    madre         3        Si    H   8.4      E4
## 4 14 1.62  45.4    madre         3        No    H   6.3      E5
## 5 17 1.74  54.6    madre         1        Si    H   4.8      E2
```

```
F3<-filter(RN,cuidador=="madre"|cuidador=="ambos")
F3[1:5,]
```

```
##   id peso talla cuidador controles lactancia sexo sueno estrato
## 1  1 0.27  46.1    madre         5        No    M   9.4      E5
## 2  2 1.48  48.2    madre         8        Si    M   7.7      E1
## 3  3 0.83  40.7    madre         6        Si    M   6.2      E4
## 4  4 0.52  56.3    madre         6        No    M   6.9      E1
## 5  5 0.25  51.5    madre         4        Si    M   6.7      E1
```

```
F4<-filter(RN,peso<=2.5)
F4[1:5,]
```

```
##   id peso talla cuidador controles lactancia sexo sueno estrato
## 1  1 0.27  46.1    madre         5        No    M   9.4      E5
## 2  2 1.48  48.2    madre         8        Si    M   7.7      E1
## 3  3 0.83  40.7    madre         6        Si    M   6.2      E4
## 4  4 0.52  56.3    madre         6        No    M   6.9      E1
## 5  5 0.25  51.5    madre         4        Si    M   6.7      E1
```

Colocar un data.frame de bajo de otro cuando tienen las mismas variables en el mismo orden

```r
F_ambos<-filter(RN,cuidador=="ambos")
F_madre<-filter(RN,cuidador=="madre")
F_otro<-filter(RN,cuidador=="Otro")

F_todos<-bind_rows(F_ambos,F_madre,F_otro)
```

Ordenar

Organiza la base de datos por los valores de un grupo de variables.

```r
O1<-arrange(RN,peso)
O1[1:5,]
```

```
##   id peso talla cuidador controles lactancia sexo sueno estrato
## 1 31 0.22  49.6    madre         8        Si    M   5.8      E5
## 2  5 0.25  51.5    madre         4        Si    M   6.7      E1
## 3  1 0.27  46.1    madre         5        No    M   9.4      E5
## 4 19 0.45  48.1    madre         8        No    H   9.8      E2
## 5  4 0.52  56.3    madre         6        No    M   6.9      E1
```

```r
O2<-arrange(RN,desc(peso))
O2[1:5,]
```

```
##    id  peso talla cuidador controles lactancia sexo sueno estrato
## 1 343 14.85  48.0     Otro         5        No    M   5.7      E4
## 2 330 13.10  52.9    ambos         1        No    M   4.2      E3
## 3 355 13.03  46.2    ambos         3        No    M   7.1      E3
## 4 361 13.00  47.8    ambos         2        Si    H   8.9      E1
## 5 351 12.50  61.5    ambos         4        Si    M   9.0      E1
```

```r
O3<-arrange(RN,cuidador)
O3[1:5,]
```

```
##    id peso talla cuidador controles lactancia sexo sueno estrato
## 1 117 4.97  51.7    ambos         2        No    M   9.6      E2
## 2 118 3.11  53.4    ambos         2        No    H   4.1      E3
## 3 119 3.57  47.2    ambos         4        Si    H   8.2      E5
## 4 121 4.55  53.4    ambos         3        Si    M   7.5      E3
## 5 122 2.53  43.1    ambos         3        No    H   9.9      E3
```

```r
O4<-arrange(RN,sexo,cuidador)
O4[1:5,]
```

```
##    id peso talla cuidador controles lactancia sexo sueno estrato
## 1 118 3.11  53.4    ambos         2        No    H   4.1      E3
## 2 119 3.57  47.2    ambos         4        Si    H   8.2      E5
## 3 122 2.53  43.1    ambos         3        No    H   9.9      E3
## 4 123 2.98  45.2    ambos         5        No    H   4.8      E3
## 5 128 5.56  50.5    ambos         2        Si    H   4.9      E1
```

```r
O5<-arrange(RN,sexo,cuidador,talla)
O5[1:5,]
```

```
##    id peso talla cuidador controles lactancia sexo sueno estrato
## 1 299 8.44  38.9    ambos         1        Si    H   6.7      E3
## 2 250 9.22  40.6    ambos         0        Si    H  10.0      E3
```

```
## 3 178 5.61  41.5    ambos         5        Si    H   9.2      E2
## 4 195 5.70  41.7    ambos         2        Si    H   9.5      E1
## 5 279 8.16  42.4    ambos         3        Si    H   5.5      E4
```

Seleccionar

Construir una nueva base de datos seleccionando un listado específico de variables:

```
S1<-select(RN,id,peso,talla)
S1[1:5,]
```

```
##   id peso talla
## 1  1 0.27  46.1
## 2  2 1.48  48.2
## 3  3 0.83  40.7
## 4  4 0.52  56.3
## 5  5 0.25  51.5
```

```
S2<-select(RN,id,peso:cuidador,sueno)
S2[1:5,]
```

```
##   id peso talla cuidador sueno
## 1  1 0.27  46.1    madre   9.4
## 2  2 1.48  48.2    madre   7.7
## 3  3 0.83  40.7    madre   6.2
## 4  4 0.52  56.3    madre   6.9
## 5  5 0.25  51.5    madre   6.7
```

```
S3<-select(RN,-controles) # excluir la variable
S3[1:5,]
```

```
##   id peso talla cuidador lactancia sexo sueno estrato
## 1  1 0.27  46.1    madre        No    M   9.4      E5
## 2  2 1.48  48.2    madre        Si    M   7.7      E1
## 3  3 0.83  40.7    madre        Si    M   6.2      E4
## 4  4 0.52  56.3    madre        No    M   6.9      E1
## 5  5 0.25  51.5    madre        Si    M   6.7      E1
```

```
S4<-select(RN,starts_with("c")) # comienza con
S4[1:5,]
```

```
##   cuidador controles
## 1    madre         5
## 2    madre         8
## 3    madre         6
## 4    madre         6
## 5    madre         4
```

```
S5<-select(RN,ends_with("o")) # termina con
S5[1:5,]
```

```
##   peso sexo sueno estrato
## 1 0.27    M   9.4      E5
## 2 1.48    M   7.7      E1
## 3 0.83    M   6.2      E4
## 4 0.52    M   6.9      E1
## 5 0.25    M   6.7      E1
```

```r
S6<-select(RN,contains("lac")) # contiene
S6[1:5,]
```

```
## [1] "No" "Si" "Si" "No" "Si"
```

Volver a unir dos bases de datos con el mismo orden de los individuos

```r
S7<-bind_cols(S1,S6)
S7[1:5,]
```

```
##   id peso talla lactancia
## 1  1 0.27  46.1        No
## 2  2 1.48  48.2        Si
## 3  3 0.83  40.7        Si
## 4  4 0.52  56.3        No
## 5  5 0.25  51.5        Si
```

Nuevas variables

Crear nuevas variables en la base de datos:

```r
N1<-mutate(RN, razon=peso/talla)
N1[1:5,]
```

```
##   id peso talla cuidador controles lactancia sexo sueno estrato
## 1  1 0.27  46.1    madre         5        No    M   9.4      E5
## 2  2 1.48  48.2    madre         8        Si    M   7.7      E1
## 3  3 0.83  40.7    madre         6        Si    M   6.2      E4
## 4  4 0.52  56.3    madre         6        No    M   6.9      E1
## 5  5 0.25  51.5    madre         4        Si    M   6.7      E1
##         razon
## 1 0.005856833
## 2 0.030705394
## 3 0.020393120
## 4 0.009236234
## 5 0.004854369
```

```r
N2<-transmute(RN, razon=peso/talla)
N2[1:5,]
```

```
## [1] 0.005856833 0.030705394 0.020393120 0.009236234 0.004854369
```

```r
N3<-mutate(RN, peso_cat=ifelse(peso<=2.5, "bajo", "normal"))
N3[1:5,]
```

```
##   id peso talla cuidador controles lactancia sexo sueno estrato peso_cat
## 1  1 0.27  46.1    madre         5        No    M   9.4      E5     bajo
## 2  2 1.48  48.2    madre         8        Si    M   7.7      E1     bajo
## 3  3 0.83  40.7    madre         6        Si    M   6.2      E4     bajo
## 4  4 0.52  56.3    madre         6        No    M   6.9      E1     bajo
## 5  5 0.25  51.5    madre         4        Si    M   6.7      E1     bajo
```

```r
N4<-mutate(RN, peso_cat=ifelse(peso<=2.5, "bajo",ifelse(peso<=4, "normal","mucho")))
N4[1:5,]
```

```
##   id peso talla cuidador controles lactancia sexo sueno estrato peso_cat
## 1  1 0.27  46.1    madre         5        No    M   9.4      E5     bajo
```

```
## 2  2 1.48  48.2    madre         8        Si    M   7.7      E1     bajo
## 3  3 0.83  40.7    madre         6        Si    M   6.2      E4     bajo
## 4  4 0.52  56.3    madre         6        No    M   6.9      E1     bajo
## 5  5 0.25  51.5    madre         4        Si    M   6.7      E1     bajo
```

```
# Utilizando la función cut

min(RN$peso);max(RN$peso)
```

```
## [1] 0.22
```

```
## [1] 14.85
```

```
N5<-mutate(RN, peso_cat=cut(peso,c(0.2,2.5,4,14.85)))
N5[1:5,]
```

```
##   id peso talla cuidador controles lactancia sexo sueno estrato  peso_cat
## 1  1 0.27  46.1    madre         5        No    M   9.4      E5 (0.2,2.5]
## 2  2 1.48  48.2    madre         8        Si    M   7.7      E1 (0.2,2.5]
## 3  3 0.83  40.7    madre         6        Si    M   6.2      E4 (0.2,2.5]
## 4  4 0.52  56.3    madre         6        No    M   6.9      E1 (0.2,2.5]
## 5  5 0.25  51.5    madre         4        Si    M   6.7      E1 (0.2,2.5]
```

```
str(N5$peso_cat)
```

```
##  Factor w/ 3 levels "(0.2,2.5]","(2.5,4]",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
# utilizando la función case_when:

N5$peso_cat2<-case_when(N5$peso<=2.5 ~ "bajo",
              N5$peso>2.5 & N5$peso<=4 ~ "normal",
              N5$peso>4 ~ "mucho")
N5[1:5,]
```

```
##   id peso talla cuidador controles lactancia sexo sueno estrato  peso_cat
## 1  1 0.27  46.1    madre         5        No    M   9.4      E5 (0.2,2.5]
## 2  2 1.48  48.2    madre         8        Si    M   7.7      E1 (0.2,2.5]
## 3  3 0.83  40.7    madre         6        Si    M   6.2      E4 (0.2,2.5]
## 4  4 0.52  56.3    madre         6        No    M   6.9      E1 (0.2,2.5]
## 5  5 0.25  51.5    madre         4        Si    M   6.7      E1 (0.2,2.5]
##   peso_cat2
## 1      bajo
## 2      bajo
## 3      bajo
## 4      bajo
## 5      bajo
```

```
N5$peso_cat2<-NULL # Borrar variable.

# Combinar categorias de una variable

N5$estrato2<-recode(N5$estrato,E1="E_bajo",E2="E_bajo",
                    E3="E_medio",E4="E_medio",E5="E_alto",E6="E_alto")

# crear variables "dummy"

#install.packages("fastDummies")
```

```r
library(fastDummies)
N6<-dummy_cols(RN,select_columns = "estrato")

# crear una interacción

N7<-mutate(RN,Int2=paste(sexo,lactancia))
str(N7$Int2)
```

```
##  chr [1:369] "M No" "M Si" "M Si" "M No" "M Si" "H No" "M Si" "H No" ...
```

```r
N7<-mutate(RN,Int3=paste(sexo,lactancia,cuidador))
str(N7$Int3)
```

```
##  chr [1:369] "M No madre" "M Si madre" "M Si madre" "M No madre" ...
```

Resumen

Calcular medidas de resumen:

```r
summarise(RN,media=mean(peso,na.rm = T),Desv.Esta.=sd(peso,na.rm = T))
```

```
##      media Desv.Esta.
## 1 6.438997   2.712548
```

```r
group_by(RN,cuidador,sexo) %>%
  summarise(media=mean(peso,na.rm = T),d.e.=sd(peso,na.rm=T))
```

```
## # A tibble: 6 x 4
## # Groups:   cuidador [3]
##   cuidador sexo  media  d.e.
##   <chr>    <chr> <dbl> <dbl>
## 1 ambos    H      7.03  2.29
## 2 ambos    M      7.38  2.36
## 3 madre    H      4.52  2.50
## 4 madre    M      4.53  2.52
## 5 Otro     H      7.16  2.62
## 6 Otro     M      8.36  3.46
```

Combinación de "verbos"

El conector %>% se puede utilizar para aplicar paso a paso las funciones anteriores:

```r
BD<-filter(RN,cuidador=="madre")%>%
  arrange(id)%>%
  select(id,peso,talla,sexo)%>%
  mutate(razon=peso/talla)

BD[1:5,]
```

```
##   id peso talla sexo       razon
## 1  1 0.27  46.1    M 0.005856833
## 2  2 1.48  48.2    M 0.030705394
## 3  3 0.83  40.7    M 0.020393120
## 4  4 0.52  56.3    M 0.009236234
## 5  5 0.25  51.5    M 0.004854369
```

```r
summarise(BD,mediana=median(peso,na.rm = T),RI=IQR(peso,na.rm = T))
```

```
##   mediana    RI
## 1    5.51 4.925
```

Ejercicio:

1. Cambien el nombre de tres variables en la base de datos original.
2. Construir una base de datos con lactancia = "Si" y más de 5 horas de sueño.
3. Construir una base de datos con sexo del recién nacido = "M" y estratos 5 o 6, o sexo del recién nacido = "H" y estratos 1 o 2.
4. Una las dos bases creadas en los ítem 2 y 3.
5. Organizar la base de datos por horas de sueño (de mayor a menor).
6. Organizar la base de datos por estrato y peso del recién nacido.
7. Construir una base de datos excluyendo la variable cuidador. Cree una base de datos únicamente con la variable cuidador. Una las dos bases anteriores.
8. Crear una base de datos incluyendo una variable con la raíz cuadrada de la talla.
9. Crear una base de datos incluyendo una variable que una las categorías "madre" y "ambos" de la variable cuidador y llame esta categoría "padres".
10. Crear una variable que clasifique la talla en tres categorías:
    - menor o igual a 40 = "bajos"
    - mayor 40 y menor o igual a 55 = "normal"
    - mayor a 55 = "altos"

! utilice distintas formas de hacer esta clasificación ¡

11. Cree una interacción entre la variable talla categorizada en la sección anterior y lactancia. Construya variable "dummy" de la interacción anterior.
12. Calcular la media y desviación estándar de la talla.
13. Calcular la media y desviación estándar de la talla por sexo y tipo de cuidador.