

Programación para el análisis de datos

Daniel Fraiman



Contenidos clase 4 de R: Ordenando datos, estadística descriptiva y dependencia entre variables.

1.- Bajar los datos *births2006smpl.rda* del campus, y cargarlos utilizando el comando *load("births2006smpl.rda")*.

- (a) Renombre los datos. **Hint:** `datos = births2006smpl`
- (b) ¿Qué tipo de objeto es `datos`? **Hint:** `class(★)`
- (c) ¿Cuántas filas y cuántas columnas tiene la tabla de datos? **Hint:** `dim(datos)`
- (d) Mire los primeros datos para entender qué información tiene la tabla. **Hint:** `head(datos)`.

DOB_MM= Month of date of birth

DOB_WK= Day of week of birth

MAGER= Mother's age

TBO_REC= Total birth order

WTGAIN= Weight gain by mother

SEX= a factor with levels F M, representing the sex of the child

APGAR5= APGAR score

DMEDUC= Mother's education level

UPREVIS= Number of prenatal visits

ESTGEST= Estimated weeks of gestation

DMETH_REC= Delivery Method

DPLURAL= "Plural Births;" levels include 1 Single, 2 Twin, 3 Triplet, 4 Quadruplet, and 5 Quintuplet or higher

DBWT= Birth weight, in grams

- (e) Renombre las variables. **Hint:** `names(datos)=c("mes","dia","edad_madre","num_hijos","peso_ganado_madre","sexo","apgar","educ","visitas_medicas","gestacion","tipo parto","parto_multiple","peso")`
- (f) **Observación:** Hay otra manera de hacer el item (d) usando el paquete *tidyverse*.
`> datos = datos %>% rename(mes = DOB_MM, dia = DOB_WK)`
- (g) Limpie la base. ¿Cuál es el valor más alto para la variable gestación? Cambie este valor por NA. Luego calcule el número promedio de semanas de gestación. **Hint:** `mean(datos$gestacion,na.rm=T)`. ¿Qué significa *ra.rm=T*?
- (h) Cambie los valores de la variable `dia` por domingo, lunes, martes, ..., sábado. **Hint:** `datos$dia=as.factor(datos$dia); levels(datos$dia)=c("domingo", ..., "sabado")`.
- (i) Agregarle a la tabla una nueva variable construida a partir de las variables `peso` y `gestación`. Usar el comando *mutate* paquete *tidyverse*. Abajo un ejemplo.
Hint: `> datos = mutate(datos, W=peso/gestacion)`

- (j) Seleccionar solamente los casos que tienen sexo F (o sea *filtrar por sexo F*). Usar el comando *filter* pero tenga en cuenta que *filter* trabaja sobre dataframes. Abajo como hacerlo. **Hint:**
- ```
> datos2=as.data.frame(datos) # as_tibble(datos) es otra opción
> datos_filtrados_F=filter(datos2,sexo=="F")
```
- (k) Seleccionar solamente los casos que tienen gestación menor a 30 semanas.
- (l) Seleccionar solamente la variable *visitas\_medicas*. **Hint:** Use el comando *select*( $\star_1, \star_2$ ).
- (m) Se pueden seleccionar dos o más variables, ¿cómo?
- (n) ¿Qué le parece que hace el comando *arrange*? Correr *arrange(datos2,edad\_madre)*. Correr *arrange(datos2,edad\_madre,num\_hijos)*.
- (o) ¿Qué pasa si corre lo de abajo?
- ```
> datos %>% group_by(educ)
```
- (p) ¿Qué pasa si corre lo de abajo?
- ```
> datos %>% group_by(educ) %>% summarise(promedio = mean(visitas_medicas))
```

## 2.- Tipos de variables.

- (a) ¿Qué tipo de variable es el peso? **R:** `class(datos$peso)`
- (b) ¿Qué tipo de variable es el tipo\_parto? **R:** `class(datos$tipo_parto)`
- (c) ¿Qué valores toma la variable categórica educ ? **R:** `levels(datos$educ)`
- (d) Notar que no están ordenadas las categorías como uno querría. Se pueden ordenar escribiendo el comando.
- ```
datos$educ=factor(datos$educ,levels=c("No formal educ.", "1Y elementary", "2Y elementary", "3Y elementary", "4Y elementary", "5Y elementary", "6Y elementary", "7Y elementary", "8Y elementary", "1Y high", "2Y high", "3Y high", "4Y high", "1Y college", "2Y college", "3Y college"))
```

Estudiando una única variable

3.- Variable numérica.

- (a) Estudie gráficamente la variable *peso* de los recién nacidos.
- Realice un histograma del peso de los recién nacidos. **Hint:** `hist(\star)`
 - ¿Para qué sirven los argumentos `xlab`, `main`, `col`, `breaks` en el histograma?
`hist(datos$peso,xlab="Peso",main="Histograma del peso",col=2)`
 - Calcule medidas de resumen del peso. **Hint:** `mean(\star ,na.rm=T)`, `median(\star)`, `sd(\star)`, `mad(\star)`, `IQR(\star)`
 - Realice un boxplot de la variable peso. **Hint:** `boxplot(\star)`
 - Cambie el color del boxplot anterior.


- vi. Grafique la acumulada empírica. **Hint:** `plot(ecdf(★))`
- vii. Realice un `qqplot` usando alguna distribución “razonable”. **Hint:** vaya a la teórica.
- viii. Realice un `qqplot` usando la distribución exponencial con $\lambda = 1$. **Hint:** goto Hint 3 (a) vii.
- (b) Calcule medidas de resumen de la *edad* de la madre. **Hint:** `mean(★)`, `median(★)`, `summary(★)`, `max(★)`, `min(★)`, `sd(★)`, `mad(★)`, `IQR(★)`

4.- Variable categórica.

- (a) Estudie la variable *tipo_parto*.
 - i. Construya una tabla de frecuencia con la variable *tipo_parto*. **Hint** `table(★)`
 - ii. ¿Cuántos partos fueron por cesárea (C-section)?
 - iii. Utilizando la tabla del item (a) realice un diagrama de barra. **Hint** `barplot(★)`
 - iv. A partir de la tabla del item (a) realice un diagrama de torta. **Hint** `pie(★)`
- (b) Estudie la variable *dia* del parto.
 - i. Realice un diagrama de barra.
 - ii. ¿Quedaron ordenados los días? Vuelva a realizar un diagrama de barra pero ahora ordene los días haciendo algo similar a lo realizado en el ejercicio 2 (d).

Estudiando la relación entre dos variables

5.- Relación entre una variable numérica y una categórica.

- (a) Estudie la relación entre peso del recién nacido y la multiplicidad del parto.
 - i. Grafique esta relación
 `plot(datos$parto_multiple, datos$peso)`
 - ii. ¿Cómo se interpreta el gráfico anterior?
- (b) Estudie la relación entre la edad de la madre y la educación.
 - i. Grafique esta relación. Verifique que tiene bien ordenadas la categoría educación.
 - ii. ¿Se puede observar alguna relación entre las dos variables?
 - iii. Realice el gráfico del item i pero ahora haciendo que la cajas de la primaria tengan el mismo color, lo mismo para el secundario y para la universidad.
Hint: `col=c(★1,★2,...,★n)`

6.- Relación entre una variable categórica y otra categórica.

- (a) Estudie la relación entre el tipo de parto y el día del parto.
 - i. ¿Qué pasa si escribe `table(datos$tipo_parto, datos$dia)`?

Programación para el análisis de datos

Daniel Fraiman



- ii. Guarde la información de la tabla en alguna variable.

Hint: `★= table(datos$tipo_parto,datos$dia)`

- iii. Haga un barplot de la tabla, utilice `beside=T` dentro del barplot. Y represente con los colores rojo y verde las barras.

7.- Relación entre una variable numérica y otra numérica.

- (a) Estudie la relación entre el tiempo de gestación y el índice Apgar.

- i. Grafique los datos. **R:** `plot(datos$gestacion,datos$apgar)`

- ii. ¿Puede observar alguna relación?

- iii. Grafique el índice Apgar promedio para cada uno de los valores de gestación.

R: `tabla3=aggregate(datos$apgar,list(datos$gestacion), mean,na.rm=T)`
`tabla3= datos %>% select(gestacion,apgar) %>% group_by(gestacion) %>%`
`summarize(APGAR=mean(apgar,na.rm=T))`
`plot(tabla3,xlab="gestacion",ylab="apgar")`

- iv. ¿Ahora puede visualizar alguna relación?

8.- (a) Cargue los archivos `datos_indec_2022.csv` y `datos_indec_2010.csv` . Hint: `datos=read.csv("datos_in`

- (b) Preste atención a ambas planillas de datos. ¿Están las mismas provincias representadas?

- (c) Junte la información de ambas planillas en un único data frame usando las herramientas de las guías 1 y 2. (for, if, etc)

- (d) Consolide una planilla con todas las provincias en las que hay información completa.

- (e) ¿Qué diferencia hay entre los comandos de abajo?

`left_join(datos1, datos2, by = c("AREA"="Jurisdiccion"))`

`right_join(datos1, datos2, by = c("AREA"="Jurisdiccion"))`

`inner_join(datos1, datos2, by = c("AREA"="Jurisdiccion"))`

`full_join(datos1, datos2, by = c("AREA"="Jurisdiccion"))`