Análisis Exploratorio de los Datos ESTADO DE SALUD

Alicia Perdices Guerra

21 de mayo, 2021

Contents

- 1.ANÁLISIS EXPLORATORIO POR PAISES.
 - 1.1 EN RELACIÓN CON EL ESTADO DE SALUD
 - * 1.1.1 Análisis Descriptivo
 - $\ast\,$ 1.1.2 Visualización y Distribución de la variable "Value"
 - · PARA HOMBRES
 - · PARA MUJERES
 - · EN TOTAL (HOMBRES Y MUJERES)
 - * 1.1.3 Normalidad de la variable "Value"
 - · PARA HOMBRES
 - · PARA MUJERES
 - · EN TOTAL (HOMBRES Y MUJERES)

1.ANÁLISIS EXPLORATORIO POR PAISES Se procede en primer lugar a cargar todos los archivos para poder realizar el análisis.

```
estado<-read.csv("C:/temp/EstadoDeSalud_Sexo_clean.csv", sep= ",")
```

1.1.- EN RELACIÓN CON EL ESTADO DE SALUD

• 1.1.1 Análisis Descriptivo

Se procede a realizar el análisis descriptivo:

summary(estado)

##	TIME	GEO	UNIT	SEX
##	Min. :2010	Length:990	Length:990	Length:990
##	1st Qu.:2012	Class :character	Class :character	Class :character
##	Median :2014	Mode :character	Mode :character	Mode :character
##	Mean :2014			
##	3rd Qu.:2017			
##	Max. :2019			
##	INDIC_HE	Value	Value_imp	
##	Length:990	Min. :50.60	Mode :logical	
##	Class :charact	er 1st Qu.:58.62	FALSE:972	

```
## Mode :character Median :62.20 TRUE :18

## Mean :62.18

## 3rd Qu.:65.00

## Max. :76.10
```

Se filta el dataframe para que la variable GEO aparezcan solo los paises objeto de estudio.(Para cada archivo relacionado con El Estado de Salud (Años de vida Sana) y unificamos la información). Además se selecciona la información relevante de la variable SEX (Males, Females).

```
#Estado de Salud (Años de Vida Sana)
#============
estado_paises<- filter(estado,
              +(GEO!="European Union - 27 countries (from 2020)")&
             +(GEO!="European Union - 28 countries (2013-2020)"))
estado_males<-filter(estado_paises, SEX=="Males")</pre>
nrow(estado_males)
## [1] 310
estado_females<-filter(estado_paises, SEX=="Females")</pre>
nrow(estado_females)
## [1] 310
head(estado_females)
##
     TIME
                                                       GEO UNIT
                                                                    SEX
## 1 2010
                                                   Belgium Year Females
## 2 2010
                                                  Bulgaria Year Females
## 3 2010
                                                   Czechia Year Females
## 4 2010
                                                   Denmark Year Females
## 5 2010 Germany (until 1990 former territory of the FRG) Year Females
                                                   Estonia Year Females
                                          INDIC HE Value Value imp
## 1 Healthy life years in absolute value at birth 62.6
                                                             FALSE
## 2 Healthy life years in absolute value at birth 67.1
                                                             FALSE
## 3 Healthy life years in absolute value at birth 64.5
                                                             FALSE
## 4 Healthy life years in absolute value at birth 61.4
                                                             FALSE
## 5 Healthy life years in absolute value at birth 58.7
                                                             FALSE
## 6 Healthy life years in absolute value at birth 58.2
                                                             FALSE
head(estado_males)
##
     TIME
                                                       GEO UNIT
                                                                  SEX
## 1 2010
                                                   Belgium Year Males
## 2 2010
                                                  Bulgaria Year Males
## 3 2010
                                                   Czechia Year Males
## 4 2010
                                                   Denmark Year Males
## 5 2010 Germany (until 1990 former territory of the FRG) Year Males
```

```
## 6 2010
                                                   Estonia Year Males
                                          INDIC_HE Value Value_imp
##
## 1 Healthy life years in absolute value at birth 64.0
                                                            FALSE
## 2 Healthy life years in absolute value at birth 63.0
                                                            FALSE
## 3 Healthy life years in absolute value at birth 62.2
                                                            FALSE
## 4 Healthy life years in absolute value at birth 62.3
                                                            FALSE
## 5 Healthy life years in absolute value at birth 57.9
                                                            FALSE
## 6 Healthy life years in absolute value at birth 54.2
                                                            FALSE
```

Se crea un Dataframe con toda la información:

[1] 74

```
max(estado_salud$Estado_females)
```

[1] 76.1

```
#Generamos el fichero filtrado para utilizarlo en el siguiente análisis.
write.csv(estado_salud, file="Estado_Salud_Analisis.csv", row.names = FALSE)
```

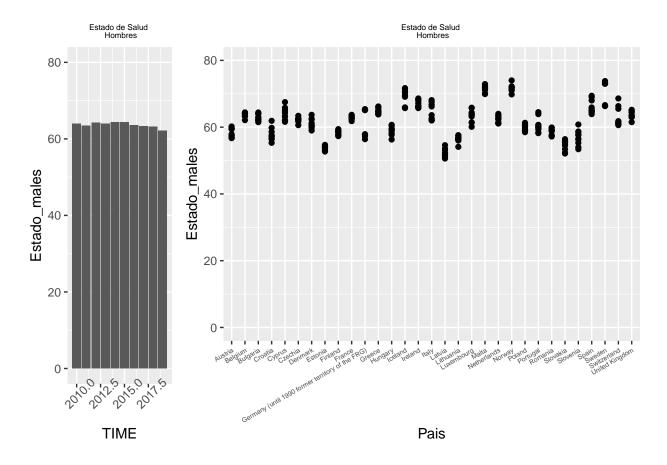
Se reescalan los datos:

```
estado_salud["Estado_males_norm"] <-
   rescale(estado_salud$Estado_males, to=c(0,1))
estado_salud["Estado_females_norm"] <-
   rescale(estado_salud$Estado_females, to=c(0,1))</pre>
```

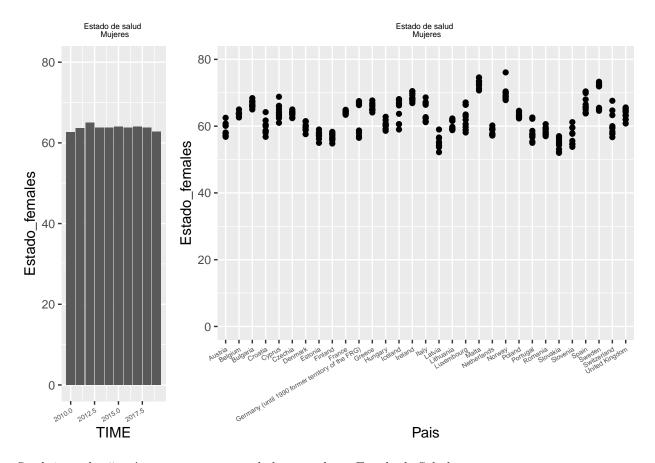
• 1.1.2 Visualización y Distribución de la información"

Se visualiza las variable que nos dan información sobre el estado de salud realizada en función de TIME, y País.

```
theme(axis.text.x = element_text(angle = 45))+
 scale_y_continuous(limit=c(0,80))+
 ggtitle("Estado de Salud \n Hombres")+
 theme (plot.title = element_text(size=rel(0.5), hjust = 0.5))
#Gráfica de barras de la información sobre el Estado de Salud en Mujeres"
plot2=ggplot(data=estado_salud)+
 geom_col(aes(x=TIME,y=Estado_females))+
 theme(axis.text.x = element_text(size= 5,angle = 30,vjust=1,hjust = 1))+
 scale_y_continuous(limit=c(0,80))+
 ggtitle("Estado de salud \n Mujeres")+
 theme (plot.title = element_text(size=rel(0.5),hjust = 0.5))
#Estado de Salud GRÁFICAS DE PUNTOS
#-----
#Gráfica de puntos de la información sobre el Estado de Salud en Hombres y Mujeres por Países"
plot3=ggplot(data=estado_salud)+
 geom_point(aes(x=Pais,y=Estado_males))+
 theme(axis.text.x = element_text(size= 5,angle = 30,vjust=1,hjust = 1))+
 scale_y_continuous(limit=c(0,80))+
 ggtitle("Estado de Salud \n Hombres")+
 theme (plot.title = element_text(size=rel(0.5),hjust=0.5))
#Gráfica de puntos de la información sobre la Detección del Cáncer de Cérvix Uterino por Países"
plot4=ggplot(data=estado salud)+
 geom_point(aes(x=Pais,y=Estado_females))+
 theme(axis.text.x = element_text(size= 5,angle = 30,vjust=1,hjust = 1))+
 scale_y_continuous(limit=c(0,80))+
 ggtitle("Estado de salud \n Mujeres")+
 theme (plot.title = element_text(size=rel(0.5),hjust=0.5))
grid.arrange(plot1,plot3,widths=c(1,3), ncol=2)
```



grid.arrange(plot2,plot4,widths=c(1,3), ncol=2)



Se obtienen los 5 países con una mayor edad con un buen Estado de Salud.

• Estado de Salud en Hombres

```
#Para "Estado_males"
#Se filtra por Año
y_2010<-filter(estado_salud, TIME==2010)
y_2011<-filter(estado_salud, TIME==2011)</pre>
y_2012<-filter(estado_salud, TIME==2012)</pre>
y_2013<-filter(estado_salud, TIME==2013)</pre>
y_2014<-filter(estado_salud, TIME==2014)
y_2015<-filter(estado_salud, TIME==2015)</pre>
y_2016<-filter(estado_salud, TIME==2016)</pre>
y_2017<-filter(estado_salud, TIME==2017)</pre>
y_2018<-filter(estado_salud, TIME==2018)</pre>
y_2019<-filter(estado_salud, TIME==2019)</pre>
#Se ordena por "Estado_males"
estado_salud_5paises_2010<-y_2010[with(y_2010, order(-y_2010$Estado_males)),]
estado_salud_5paises_2011<-y_2011[with(y_2011, order(-y_2011$Estado_males)),]
estado_salud_5paises_2012<-y_2012[with(y_2012, order(-y_2012$Estado_males)),]
```

Table 1: Países con mayor edad con buena salud en hombres en 2010

	País	ES Hombres
18	Malta	70.1
29	Norway	69.8
28	Iceland	69.3
27	Sweden	66.6
8	Greece	66.1

Table 2: Países con mayor edad con buena salud en hombres en $2011\,$

	País	ES Hombres
29	Norway	74.0
18	Malta	69.9
28	Iceland	69.1
27	Sweden	66.3
30	Switzerland	66.3

Table 3: Países con mayor edad con buena salud en hombres en $2012\,$

	País	ES Hombres
29	Norway	71.9
18	Malta	71.5
28	Iceland	70.4
30	Switzerland	68.6

	País	ES Hombres
27	Sweden	66.4

Table 4: Países con mayor edad con buena salud en hombres en $2013\,$

	País	ES Hombres
28	Iceland	71.7
18	Malta	71.6
29	Norway	71.0
27	Sweden	66.4
7	Ireland	65.7

Table 5: Países con mayor edad con buena salud en hombres en $2014\,$

	País	ES Hombres
27	Sweden	73.0
18	Malta	72.3
29	Norway	72.2
28	Iceland	70.8
7	Ireland	66.3

Table 6: Países con mayor edad con buena salud en hombres en $2015\,$

	País	ES Hombres
27	Sweden	73.1
18	Malta	72.6
29	Norway	71.8
28	Iceland	71.5
7	Ireland	66.5

Table 7: Países con mayor edad con buena salud en hombres en $2016\,$

	País	ES Hombres
27	Sweden	73.0
29	Norway	72.0
18	Malta	71.1
28	Iceland	69.6
12	Italy	67.6

Table 8: Países con mayor edad con buena salud en hombres en $2017\,$

	País	ES Hombres
27	Sweden	73.2
18	Malta	71.9
29	Norway	71.8
9	Spain	69.0
7	Ireland	67.9

Table 9: Países con mayor edad con buena salud en hombres en $2018\,$

	País	ES Hombres
27	Sweden	73.7
18	Malta	71.9
29	Norway	71.7
7	Ireland	68.3
9	Spain	68.0

Table 10: Países con mayor edad con buena salud en hombres en $2019\,$

	País	ES Hombres
27	Sweden	73.8
18	Malta	72.9
29	Norway	71.2
9	Spain	69.4
7	Ireland	68.6

A continuación, se aprupa toda la información por paises en una tabla:

```
a1<-group_by(estado_salud,Pais) #Se agrupa por paises

#Se selecciona las variables Pais y Estado_males

a2<-select(a1,Pais:Estado_males)

#Se muestra la información por cada país, con Value=suma de

#los valores de cada país en los 10 años.

a3<-(summarize(a2,suma=sum(Estado_males)/10))

a4<-data.frame(a3) #Se convierte la información en un dataframe.

#Se ordena el DataFrame por la variable Suma de forma descendente.

a5<-a4[with(a4,order(-a4$suma)),]

#Se crea una tabla con toda la información

kable(a5[0:5,c(1,2)],

col.names = c("País", "ES Hombres"),

caption = "Países con la mayor media en mayor Edad con buena salud en Hombres en 2010-2019")
```

Table 11: Países con la mayor media en mayor Edad con buena salud en Hombres en 2010-2019

	País	ES Hombres
22	Norway	71.74
20	Malta	71.58
29	Sweden	70.55
14	Iceland	68.99
15	Ireland	66.85

• Estado de Salud en Mujeres

```
#Para "Estado_females"
#Se filtra por Año
y 2010<-filter(estado salud, TIME==2010)
y_2011<-filter(estado_salud, TIME==2011)</pre>
y_2012<-filter(estado_salud, TIME==2012)</pre>
y_2013<-filter(estado_salud, TIME==2013)</pre>
y_2014<-filter(estado_salud, TIME==2014)</pre>
y_2015<-filter(estado_salud, TIME==2015)</pre>
y_2016<-filter(estado_salud, TIME==2016)</pre>
y_2017<-filter(estado_salud, TIME==2017)</pre>
y_2018<-filter(estado_salud, TIME==2018)</pre>
y_2019<-filter(estado_salud, TIME==2019)</pre>
#Se ordena por Estado_females"
estado_salud_5paises_2010<-y_2010[with(y_2010, order(-y_2010$Estado_females)),]
estado_salud_5paises_2011<-y_2011[with(y_2011, order(-y_2011$Estado_females)),]
estado_salud_5paises_2012<-y_2012[with(y_2012, order(-y_2012$Estado_females)),]
estado_salud_5paises_2013<-y_2013[with(y_2013, order(-y_2013$Estado_females)),]
```

Table 12: Países con mayor edad con buena salud en Mujeres en $2010\,$

	País	ES Mujeres
18	Malta	70.1
29	Norway	69.8
28	Iceland	69.3
8	Greece	66.1
2	Bulgaria	63.0

Table 13: Países con mayor edad con buena salud en Mujeres en $2011\,$

	País	ES Mujeres
29	Norway	74.0
18	Malta	69.9
7	Ireland	66.1
28	Iceland	69.1
16	Luxembourg	65.8

Table 14: Países con mayor edad con buena salud en Mujeres en 2012

	País	ES Mujeres
18	Malta	71.5
29	Norway	71.9
7	Ireland	66.0
28	Iceland	70.4
30	Switzerland	68.6

Table 15: Países con mayor edad con buena salud en Mujeres en 2013

	País	ES Mujeres
18	Malta	71.6
29	Norway	71.0
7	Ireland	65.7
28	Iceland	71.7
2	Bulgaria	62.4

Table 16: Países con mayor edad con buena salud en Mujeres en $2014\,$

	País	ES Mujeres
18	Malta	72.3
27	Sweden	73.0
29	Norway	72.2
7	Ireland	66.3
28	Iceland	70.8

Table 17: Países con mayor edad con buena salud en Mujeres en $2015\,$

	País	ES Mujeres
18	Malta	72.6
27	Sweden	73.1
29	Norway	71.8
7	Ireland	66.5
5	Germany (until 1990 former territory of the FRG)	65.3

Table 18: Países con mayor edad con buena salud en Mujeres en $2016\,$

	País	ES Mujeres
27	Sweden	73.0
18	Malta	71.1
7	Ireland	67.2
13	Cyprus	67.5
29	Norway	72.0

Table 19: Países con mayor edad con buena salud en Mujeres en 2017

	País	ES Mujeres
18	Malta	71.9
27	Sweden	73.2
29	Norway	71.8
9	Spain	69.0
7	Ireland	67.9

Table 20: Países con mayor edad con buena salud en Mujeres en 2018

País ES Mujere 18 Malta 71.
18 Malta 71.
27 Sweden 73.
7 Ireland 68.
29 Norway 71.
9 Spain 68.

Table 21: Países con mayor edad con buena salud en Mujeres en $2019\,$

	País	ES Mujeres
18	Malta	72.9
27	Sweden	73.8
7	Ireland	68.6
9	Spain	69.4
12	Italy	68.1

A continuación, se aprupa toda la información por paises en una tabla:

Table 22: Países con la mayor media en mayor Edad con buena salud en Mujeres en 2010-2019

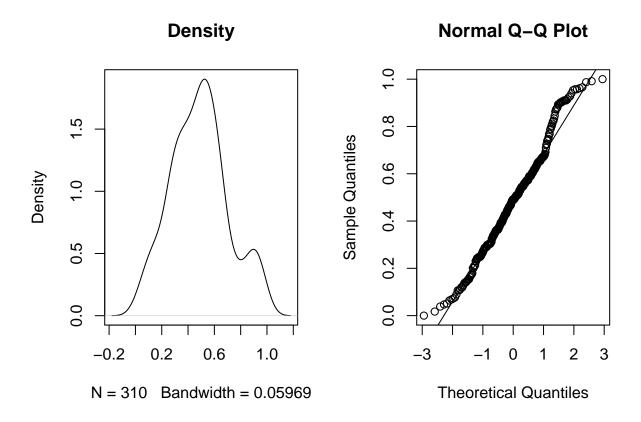
	País	ES Mujeres
20	Malta	72.87
22	Norway	69.86
29	Sweden	69.50
15	Ireland	68.72
3	Bulgaria	66.61

• 1.1.3 Normalidad de la variable "Value (Estado_males,Estado_females)"

Se comprueba con métodos visuales si la variable tiene una distribución normal.

Estado males

```
par(mfrow=c(1,2))
plot(density(estado_salud$Estado_males_norm) ,main="Density")
qqnorm(estado_salud$Estado_males_norm)
qqline(estado_salud$Estado_males_norm)
```



Para estudiar si una muestra proviene de una población con distribución normal, se disponen de tres herramientas:

- Histograma o Densidad
- Gráficos cuantil cuantil (QQplot)
- Pruebas de hipótesis.

Si en la prueba de Densidad se observa sesgo hacia uno de los lados de la gráfica, sería indicio de que la muestra no proviene de una población normal. Si por otra parte, sí se observa simetría, **NO** se garantiza que la muestra provenga de una población normal. En estos casos sería necesario utilizar otras herramientas como **QQplot y pruebas de hipótesis**.

En la gráfica Densidad de la variable "Estado_males_norm", se observa cierto sesgo hacia la derecha, aunque con una figura bastante aproximada a una curva normal. Se puede confirmar observando la gráfica QQplot en la que la línea que grafica qqline sirve de referencia para interpretar el gráfico. Si se tuviese una muestra distribuída normalmente, se esperaría que los puntos del gráfico quantil quantil estuviesen perfectamente alineados con la línea de referencia, y observamos que para este caso, "Estado_males_norm" se alinea bastante exceptuando los extremos.

Para confirmar, ya que en este caso existen dudas sobre la normalidad, se realizan las pruebas de hipótesis:

- \$h 0: La muestra proviene de una población normal.
- \$h_1: La muestra NO proviene de una población normal.

Se aplica la prueba Shapiro-Wilk:

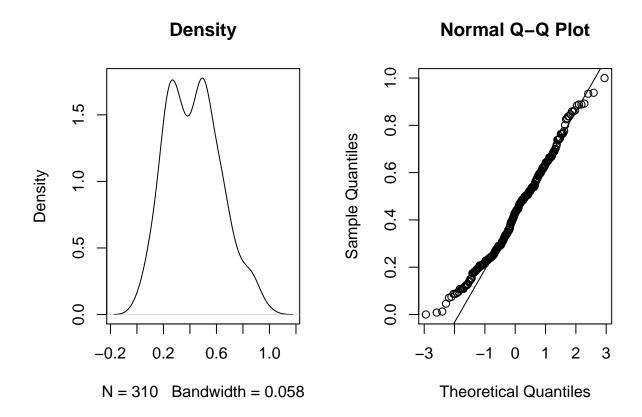
shapiro.test(estado_salud\$Estado_males_norm)

```
##
## Shapiro-Wilk normality test
##
## data: estado_salud$Estado_males_norm
## W = 0.98481, p-value = 0.002326
```

Se observa un p-value pequeño, más pequeño que algunos niveles de significación (como por ejemplo alpha=0.5) pero mayor que alpha=0.001 por lo que en este último caso no se rechaza la hipótesis nula y asumimos **Normalidad** en la muestra.

${\bf Estado_females}$

```
par(mfrow=c(1,2))
plot(density(estado_salud$Estado_females_norm) ,main="Density")
qqnorm(estado_salud$Estado_females_norm)
qqline(estado_salud$Estado_females_norm)
```



En la gráfica Densidad de la variable "Estado_females_norm", se observa una gráfica parecida a la obtenida anteriormente, existe cierto sesgo pero con una forma aproximada a una curva normal. Se puede confirmar observando la gráfica QQplot en la que la línea que grafica qqline sirve de referencia para interpretar el gráfico, se se alinea con los puntos de los valores de la variable "Estado_females_norm" exceptuando en los extremos.

Tras aplicar la prueba Shapiro-Wilk se comprueba:

shapiro.test(estado_salud\$Estado_females_norm)

```
##
## Shapiro-Wilk normality test
##
## data: estado_salud$Estado_females_norm
## W = 0.9835, p-value = 0.001247
```

Se observa un p-value pequeño, más pequeño que algunos niveles de significación (como por ejemplo alpha=0.5) pero mayor que alpha=0.001 por lo que en este último caso no se rechaza la hipótesis nula y asumimos **Normalidad** en la muestra.