Ciclo de vida de los datos - P2

Jorge Alaiza

January 2, 2019

Ciclo de vida de los datos.

1. *Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?*

```
Para esta práctica he decidido realizarla sobre uno de los datasets de Kaggle, este d
ataset esta orientado a la recolección de datos sobre registros de suicidios, estos d
atos por razones obvias vienen enmascarados para proteger la identidad de los afectad
os por lo que no aparece información certera de la edad del sujeto u otra información
que pueda afectar a su privacidad.

este conjunto de datos contiene la siguiente información:

Country: Pais donde ocurre el suicidio
Year: año en el que este suicidio ocurre
Sex: Sexo del conjunto de personas que lo han cometido
Age: Rango de edad de las personas que se suicidan
suicides_no: Volumen de suicidios cometidos por pais/anyo/sexo y rango de edad
population: Volumen de población que forma el subconjunto por pais/anyo/sexo/rango de
edad

El valor que realmente se quiere extraer sobre este dataaset es el relativo a que con
juntos sociales les afecta mas la necesidad de suicidarse, por pais, edad o sexo (a g
randes rasgos, no se tiene información sobre el trabajo u otros componentes que pudie
ran afectar a que finalmente se suicidara)
```

2. *Integración y selección de los datos de interés a analizar.*

```
library(readr)

suicide_raw_data<- read_delim("/home/alaiza/Desktop/who-suicide-statistics/who_suicid
e_statistics.csv",
                      ",", escape_double = FALSE,
                      trim_ws = TRUE)
```

```
## Parsed with column specification:
## cols(
##   country = col_character(),
##   year = col_double(),
##   sex = col_character(),
##   age = col_character(),
##   suicides_no = col_double(),
##   population = col_double()
## )
```

Para este estudio, al solo disponer de unas pocas columnas que considero de gran valor no se van a filtrar en primera instancia, se van a utilizar las 6 columnas.

3. *Limpieza de los datos*

3.1 *¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?.*

```
missingvalues <- function(array){
  for (j in 1:length(array)) {
    if(is.na(array[j])){
      return('yes')
    }
  }
  return('no')
}

missingvalues(suicide_raw_data$country)
```

```
## [1] "no"
```

```
missingvalues(suicide_raw_data$year)
```

```
## [1] "no"
```

```
missingvalues(suicide_raw_data$sex)
```

```
## [1] "no"
```

```
missingvalues(suicide_raw_data$age)
```

```
## [1] "no"
```

```
missingvalues(suicide_raw_data$suicides_no)
```

```
## [1] "yes"
```

```
missingvalues(suicide_raw_data$population)
```

```
## [1] "yes"
```

Se puede observar que para el número de suicidios y volumen de la población hay valores nulos, en este caso como solo queremos la información que nos indique la tasa de suicidios que están debidamente documentados se procederá a eliminar las filas con estos valores, por otro lado el que existan valores definidos a 0 (que son muchos los casos) indicarán que no hay suicidios (si es que los datos estan bien construidos), pero hay otros controles que debemos hacer previamente sobre los datos:

```
for (j in 1:length(suicide_raw_data$suicides_no)) {
  if((!is.na(suicide_raw_data$suicides_no[j]) & !is.na(suicide_raw_data$population
[j]))){
    if(suicide_raw_data$suicides_no[j]>=suicide_raw_data$population[j]){
      print('Hay valores sin lógica')
    }
    if(suicide_raw_data$population[j]==0){
      print('Hay valores sin lógica')
    }
  }
}
```

```
Con el código anterior, al no imprimir nada se sabe que dentro de las filas que nos v
an a quedar no hay incongruencias del tipo que haya mas suicidios que población o hay
a poblacion igual a 0 (algo posible pero muy extraño)
```

```
data_fixed <- suicide_raw_data
data_fixed <- data_fixed[complete.cases(data_fixed), ]
```

3.2 *Identificación y tratamiento de valores extremos.*

Los valores extremos que se van a buscar al estar asociados a un conjunto de población (por sexo y edad) lo mejor va a ser calcular los valores extremos del ratio entre suicidios y poblacion perteneciente a ese subconjunto de poblacion, o lo que es lo mismo, la division entre “suicides_no” y “population”

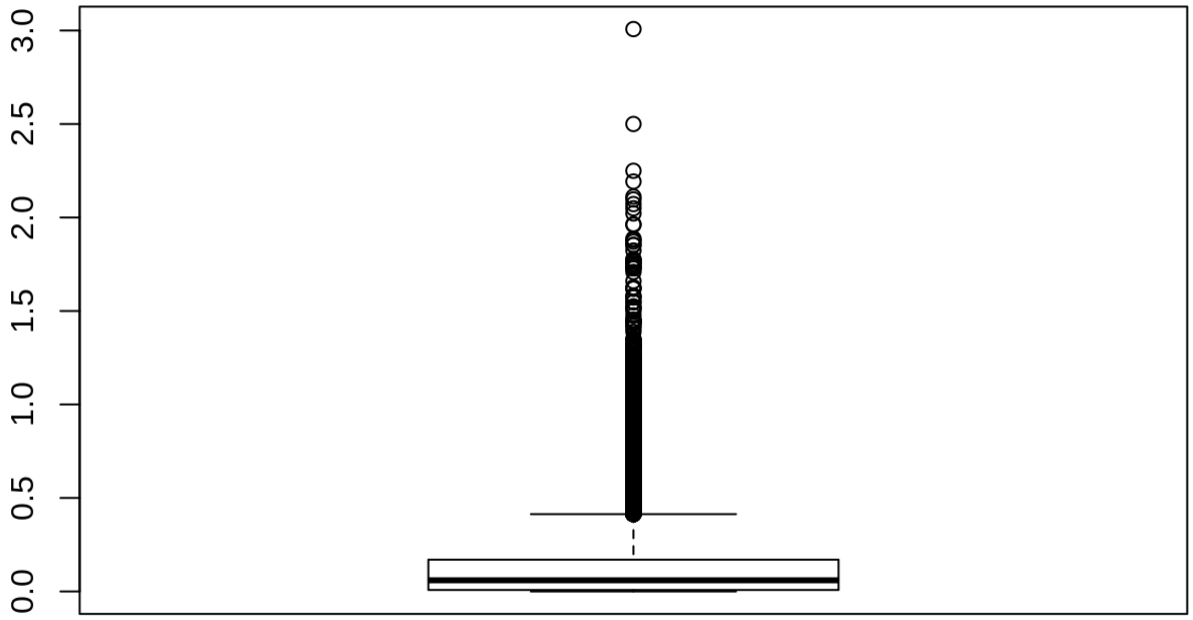
```
ratio <- data_fixed$suicides_no

for (j in 1:length(ratio)) {
  if(ratio[j]!=0){
    ratio[j]<- data_fixed$suicides_no[j]/data_fixed$population[j]*1000
  }
}

data_fixed <- invisible(cbind(data_fixed, ratio))

head(data_fixed, n=10)
```

```
##    country year    sex         age suicides_no population      ratio
## 1  Albania 1987 female 15-24 years          14     289700 0.04832585
## 2  Albania 1987 female 25-34 years           4     257200 0.01555210
## 3  Albania 1987 female 35-54 years           6     278800 0.02152080
## 4  Albania 1987 female  5-14 years           0     311000 0.00000000
## 5  Albania 1987 female 55-74 years           0     144600 0.00000000
## 6  Albania 1987 female   75+ years           1      35600 0.02808989
## 7  Albania 1987   male 15-24 years          21     312900 0.06711409
## 8  Albania 1987   male 25-34 years           9     274300 0.03281079
## 9  Albania 1987   male 35-54 years          16     308000 0.05194805
## 10 Albania 1987   male  5-14 years           0     338200 0.00000000
```

```
summary(data_fixed$ratio)
```

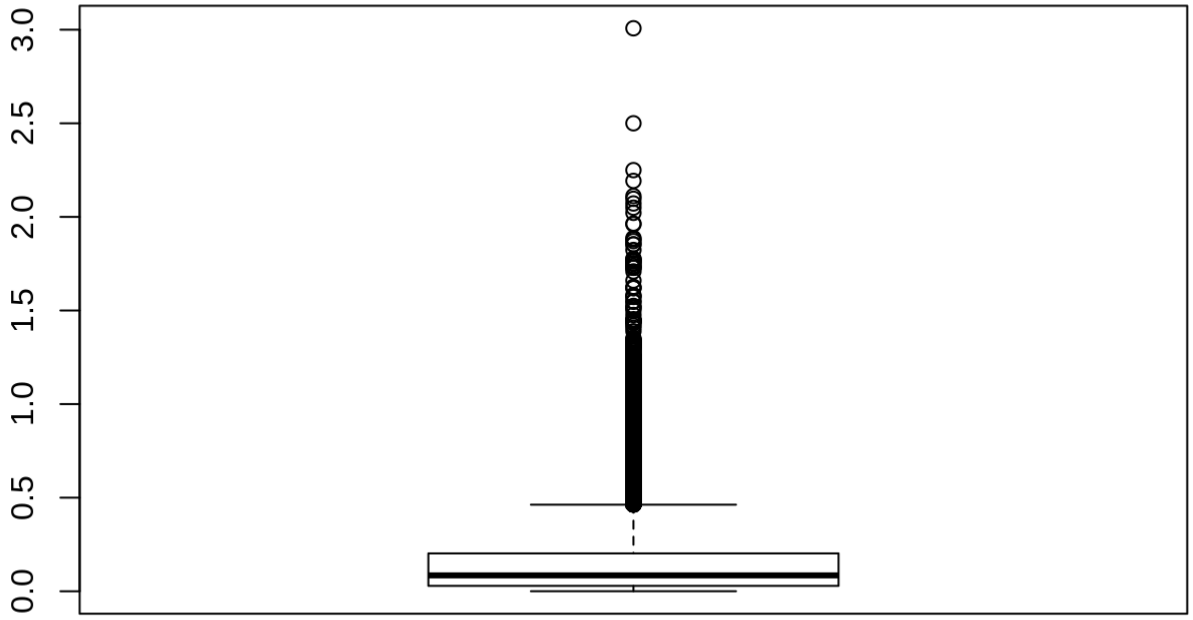
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000000 0.007818 0.059756 0.131851 0.170101 3.007519
```

```
boxplot(data_fixed$ratio)
```



se puede observar que hay muchos valores considerables “atípicos” o por lo menos que merecen ser revisados

```
vector2 <-  data_fixed$ratio[ data_fixed$ratio != 0 ]
boxplot(vector2)
```



se puede ver que el “ruido” que pueden meter los paises pequeños y sin casos de suicidio no cambia notablemente el diagrama, por alguna razón son anormales estos valores, se procede a averiguar mas información de esos valores.

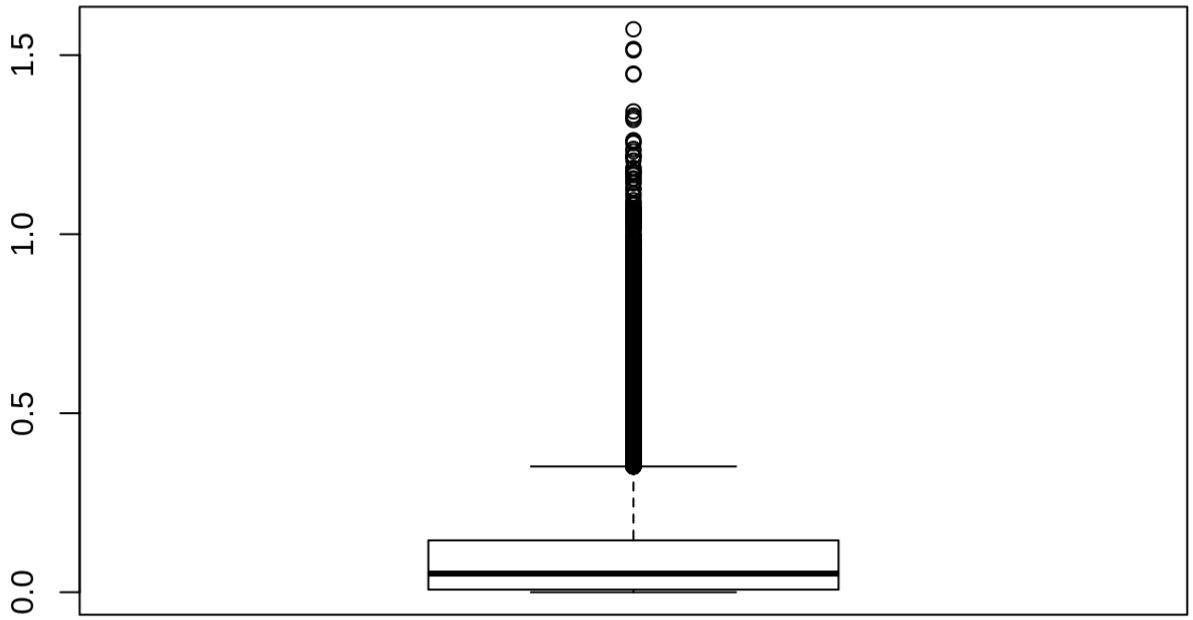
```
country <-  data_fixed$country [data_fixed$ratio>= 1.5 ]
year <-  data_fixed$year [data_fixed$ratio>= 1.5 ]
sex <-  data_fixed$sex [data_fixed$ratio>= 1.5 ]
age <-  data_fixed$age [data_fixed$ratio>= 1.5 ]
ratio <-  data_fixed$ratio [data_fixed$ratio>= 1.5 ]

dataframe_atipicos <- invisible(cbind(country, year,sex,age,ratio))
extremecases <- dataframe_atipicos[order(ratio),]

tail(extremecases, n=10)
```

```
##       country         year   sex    age         ratio
## [33,] "Hungary"       "1990" "male" "75+ years" "1.96441657845452"
## [34,] "Hungary"       "1980" "male" "75+ years" "2.02151755379388"
## [35,] "Seychelles"    "2006" "male" "75+ years" "2.04918032786885"
## [36,] "Hungary"       "1985" "male" "75+ years" "2.07169214549288"
## [37,] "Hungary"       "1982" "male" "75+ years" "2.09650582362729"
## [38,] "Hungary"       "1979" "male" "75+ years" "2.11136890951276"
## [39,] "Hungary"       "1981" "male" "75+ years" "2.19224283305228"
## [40,] "Aruba"         "1995" "male" "75+ years" "2.24971878515186"
## [41,] "French Guiana" "1979" "male" "75+ years" "2.5"
## [42,] "San Marino"    "1997" "male" "75+ years" "3.00751879699248"
```

por lo que parece, los casos mas extremos son gente de +75 años, que son un numero relevante de casos sobre una poblacion muy pequeña (la de mayores de 75 años) lo cual me hace pensar que se estan considerando los casos de eutanasia como suicidio, para poder hacer un estudio algo mas interesante (y de alguna forma poder disipar del estudio casos voluntarios de suicidio) para tratar casos de suicidio voluntario no relacionados con muertes naturales se van a eliminar el subconjunto de personas mayores a 75 años

**Aclaracion: he investigado los paises que aparecian y sin pararme en todos, parece que permiten la eutanasia en distintas formas y con distintas regulaciones.

```
data_fixed_v2<-subset(data_fixed, data_fixed$age!="75+ years")
boxplot(data_fixed_v2$ratio)
```



Se puede observar una mejoria notable en los resultados, sin embargo estos valores tan por encima es posible que respondan a una realidad que debamos investigar, conocimiento que debamos de realizar tras un analisis mas profundo.

4. *Análisis de los datos.*

4.1 *Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).*

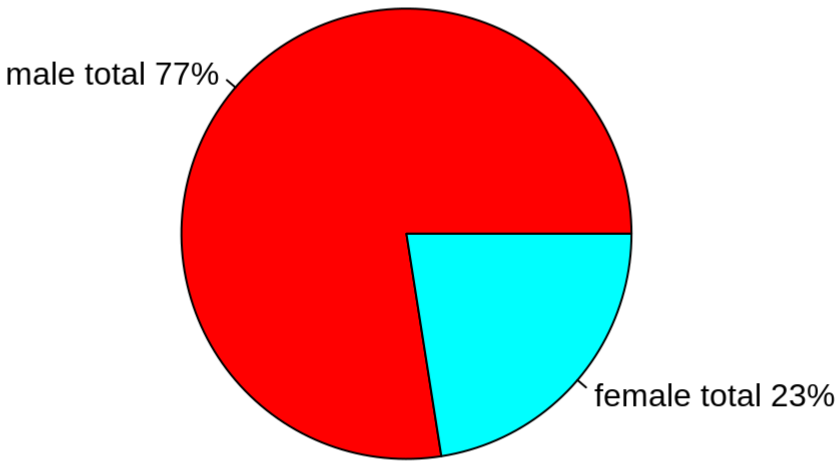
PUNTO 1 En este analisis se van a comprobar las diferencias entre hombres y mujeres comprobando que la media de los hombres es claramente superior a la de las mujeres con un % de confianza del 97% (un nivel bastante alto de confianza)

PUNTO 2 Un segundo punto a analizar sera comprobar si existen diferencias entre paises con gran volumen de poblacion y los de menor poblacion, dividiendo estos conjuntos por la mitad (divididos por la mediana, para dividir al 50% los volumenes de poblaciones y sus paises asociados).

PUNTO 1

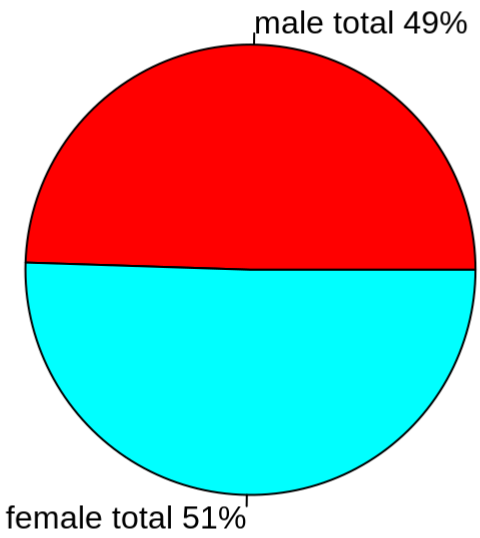
primeramente a nivel exploratorio, se comprueba si para exactamente el mismo numero de paises y rangos de edad existen diferencias en el total de suicidios

```
data_male<-data_fixed_v2[data_fixed_v2$sex=="male", ]
data_female<-data_fixed_v2[data_fixed_v2$sex=="female", ]


slices <- c(sum(data_male$suicides_no), sum(data_female$suicides_no))
lbls <- c("male total", "female total")
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=rainbow(length(lbls)),
    main="Pie Chart comparisson number of suicides per sex")
```



para comenzar, como primera aproximación, es bastante característico el que el volumen de hombres que se suicidan ocupe el 77%, a continuación, es esperable que el volumen de poblacion entre hombres y mujeres sea el mismo, por lo que se va a comprobar.

```
slices <- c(sum(data_male$population), sum(data_female$population))
lbls <- c("male total", "female total")
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=rainbow(length(lbls)),
    main="Pie Chart comparisson number of suicides per sex")
```



Esta claro que los resultados en primera instancia muestran clarisimamente que los hombres se suicidan mas de un 250% mas que las mujeres, pero hay que comprobar que este resultado extraido visualmente se puede asegurar con un nivel de significancia alto, como se ha pedido anteriormente, del 97%.

Para este estudio se va a utilizar el ratio, ya que comprende la division entre el numero de suicidios y el volumen de poblacion.

Recapitulando, se va a comprobar si : ***H0: las mujeres tienen la misma tasa de suicidios inferior a los hombres, por el contrario, H1: las mujeres tienen una tasa igual o superior.***

```
#97% ---> alfa=1-0.97 --> alfa= 0.03
#P(Z<z) = 1-alfa/2 = 1-0.03/2 = 0.985
#Según las tablas P(Z<z)=0.985 --> z=2.17
var1 <- data_male$ratio
##por un lado:
extremosuperior <- mean(var1) + 2.17 * sd(var1)/sqrt(length(var1))
extremoinferior <- mean(var1) - 2.17 * sd(var1)/sqrt(length(var1))
cat("rango de aceptacion: [", extremoinferior,",",extremosuperior,"]")
```

```
## rango de aceptacion: [ 0.1658331 , 0.1726431 ]
```

por lo que la media de ratio de las mujeres deberia comprenderse entre esos dos valores para aceptar H0, de lo contrario, por descarte seria la segunda hipotesis la ganadora.

```
mean(data_female$ratio)
```

```
## [1] 0.04713949
```

no se puede aceptar H0, por lo que las mujeres definitivamente tienen una tasa de suicidio muy inferior a la de los hombres.

PUNTO 2

primeramente se van a dividir en las dos categorias

```
data_agregated <- aggregate(cbind(data_fixed_v2$population, data_fixed_v2$suicides_n
o, data_fixed_v2$ratio), by=list(Category=data_fixed_v2$country), FUN=sum)
mediana<-median(data_agregated$V1)
data_big_countries<-data_agregated[data_agregated$V1 >= mediana, ]
data_small_countries<-data_agregated[data_agregated$V1 < mediana, ]
data_big_ordered <- data_big_countries[order(data_big_countries[2]),]
data_small_ordered <- data_small_countries[order(data_small_countries[2]),]
```

```
Paises con menor poblacion
```

```
head(cbind(data_small_ordered$Category, data_small_ordered$V1),n=10)
```

```
##       [,1]                    [,2]
##  [1,] "Cayman Islands"        "28400"
##  [2,] "Bermuda"               "98500"
##  [3,] "Saint Kitts and Nevis" "112800"
##  [4,] "San Marino"            "163440"
##  [5,] "Sao Tome and Principe" "260100"
##  [6,] "Macau"                 "336814"
##  [7,] "Dominica"              "374200"
##  [8,] "Rodrigues"             "408345"
##  [9,] "Cabo Verde"            "439643"
## [10,] "Kiribati"              "732927"
```

```
Paises con mayor poblacion
```

```
tail(cbind(data_big_ordered$Category, data_big_ordered$V1), n=10)
```

```
##       [,1]                      [,2]
## [50,] "Thailand"                "1801002288"
## [51,] "France"                  "1816722978"
## [52,] "Germany"                 "1856678452"
## [53,] "Italy"                   "1873477997"
## [54,] "United Kingdom"          "1898631249"
## [55,] "Mexico"                  "3066771042"
## [56,] "Japan"                   "4014646102"
## [57,] "Russian Federation"      "4384117710"
## [58,] "Brazil"                  "5395577849"
## [59,] "United States of America" "8771678750"
```

**Aclaracion: hay que tener en cuenta que la poblacion es la suma de varios años distintos, por eso los valores a la derecha no corresponden con la poblacion de los paises a dia de hoy, si solo se utilizase un año seria correcto.

H0: los paises de mayor poblacion tiene una mayor tasa de suicidios, H1: las tasas son iguales o superior la de los paises menores

```
#97% ---> alfa=1-0.97 --> alfa= 0.03
#P(Z<z) = 1-alfa/2 = 1-0.03/2 = 0.985
#Según las tablas P(Z<z)=0.985 --> z=2.17
var1 <- data_big_ordered$V3
##por un lado:
extremosuperior <- mean(var1) + 2.17 * sd(var1)/sqrt(length(var1))
extremoinferior <- mean(var1) - 2.17 * sd(var1)/sqrt(length(var1))
cat("rango de aceptacion: [", extremoinferior,",",extremosuperior,"]")
```

```
## rango de aceptacion: [ 29.25453 , 42.82235 ]
```

```
mean( data_small_ordered$V3)
```

```
## [1] 19.0645
```

Por lo tanto se tiene que aceptar que el 50% de paises con mayor poblacion cometen unas tasas de suicidios superiores al otro 50% con una tasa de confianza del 97%

5. *Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?*

En ambos casos ha sido posible comprobar que las hipotesis iniciales son ciertas, los hombres cometen suicidio con una diferencia abismal en comparacion con el subconjunto de mujeres, por otro lado, tambien ha sido posible comprobar que los paises, a cuanto mayor poblacion, mayor tasa de suicidios por cada 100 habitantes (este estudio se ha hecho sobre el ratio de suicidios por poblacion del subconjunto poblacional)