Análisis al dataframe present

Alejandro Zavala

2023-08-09

Contents


```
# Clear environment
rm(list = ls())
# Call libraries
library("knitr")
library("ggplot2")
library("gridExtra") #Multiple plots
library("statsr")
```

```
{FALSE} ## Loading required package: BayesFactor
```

```
{FALSE} ## Loading required package: coda
```

```
{FALSE} ## Loading required package: Matrix
```

```
{FALSE} ## ************ ## Welcome to BayesFactor 0.9.12-4.4. If you have questions,
please contact Richard Morey (richarddmorey@gmail.com). ##  ## Type BFManual() to open
the manual. ## ************
```

```
library("tidyverse")
```

```
{FALSE} ## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0
-- ## v dplyr     1.1.2     v readr     2.1.4 ## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1 ## v purrr     1.0.1     v tidyr     1.3.0
```

```
{FALSE} ## -- Conflicts ---------------------------------------- tidyverse_conflicts()
-- ## x dplyr::combine() masks gridExtra::combine() ## x tidyr::expand()  masks Matrix::expand()
## x dplyr::filter()  masks stats::filter() ## x dplyr::lag()     masks stats::lag() ##
x tidyr::pack()    masks Matrix::pack() ## x tidyr::unpack()  masks Matrix::unpack() ##
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to
become errors
```

Descripción del análisis

Se hara un pequeño analisis a la tabla "present" proveniente de la libreria **statsr**. Este dataframe contiene la cantidad de niños y niñas nacidos en Estados Unidos de América de 1940 a 2013.

Explorando el dataframe

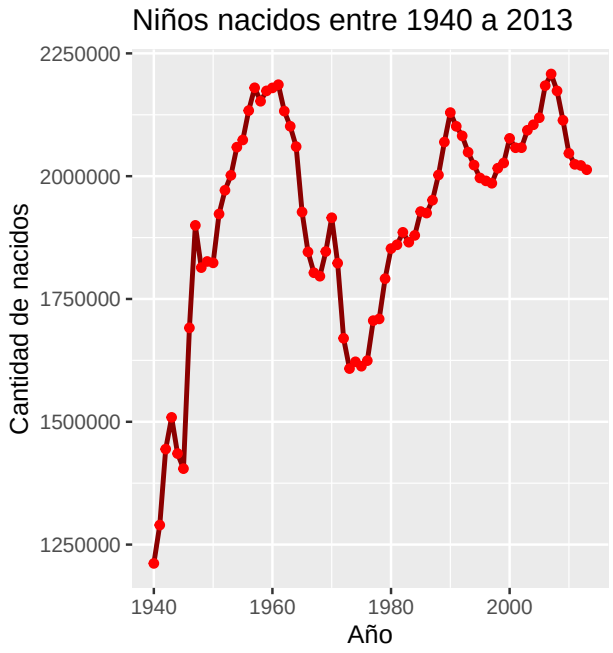
```
data(present) # Load dataframa
kable(head(present,n=10)) # Show few records
```

year	boys	girls
1940	1211684	1148715
1941	1289734	1223693
1942	1444365	1364631
1943	1508959	1427901
1944	1435301	1359499
1945	1404587	1330869
1946	1691220	1597452
1947	1899876	1800064
1948	1813852	1721216
1949	1826352	1733177

Visualización de datos

Gráfico por sexo del nacido

```
g_boys <- ggplot(data = present, aes(x = year, y = boys)) +
  geom_line(size=1,color="darkred") +
  geom_point(color="red") +
  xlab("Año") +
  ylab("Cantidad de nacidos") +
  ggtitle("Niños nacidos entre 1940 a 2013")

g_girls <- ggplot(data = present, aes(x = year, y = girls)) +
  geom_line(size=1,color="magenta4") +
  geom_point(color="magenta4") +
  xlab("Año") +
  ylab("Cantidad de nacidos") +
  ggtitle("Niñas nacidos entre 1940 a 2013")

grid.arrange(g_boys, g_girls, nrow = 1)
```


Niños nacidos entre 1940 a 2013
Cantidad de nacidos
2250000
2000000
1750000
1500000
1250000
1940
1960
1980
2000
Año

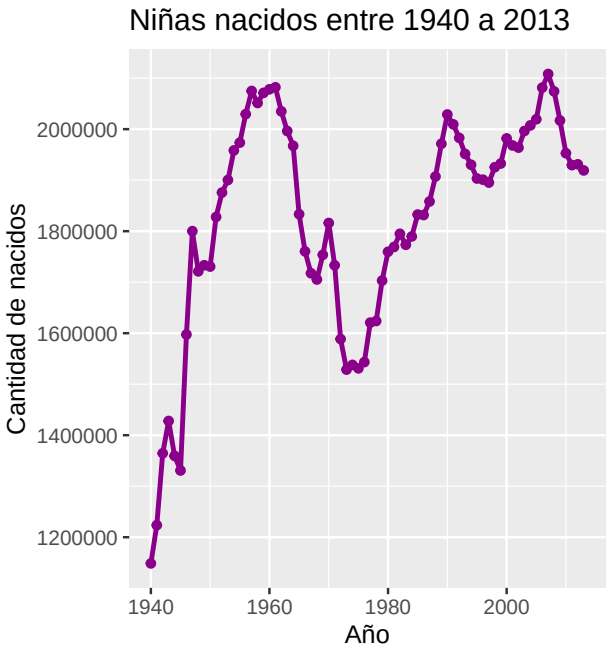
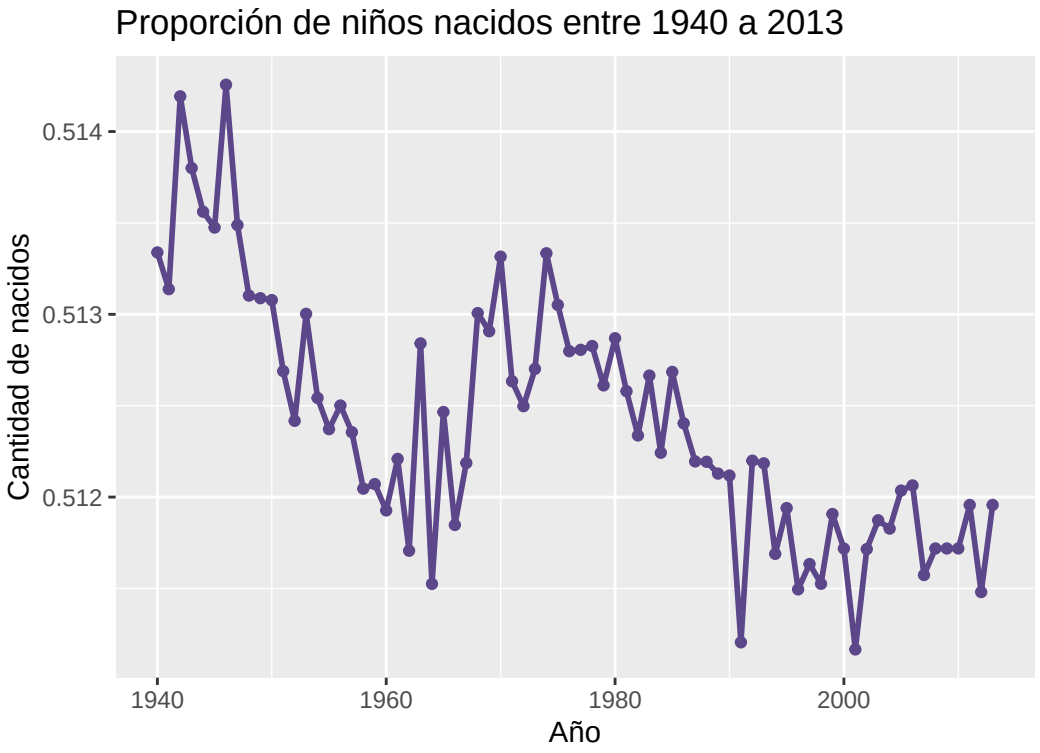

Niñas nacidos entre 1940 a 2013
Cantidad de nacidos
2000000
1800000
1600000
1400000
1200000
1940
1960
1980
2000
Año

Gráfico de proporción de niños

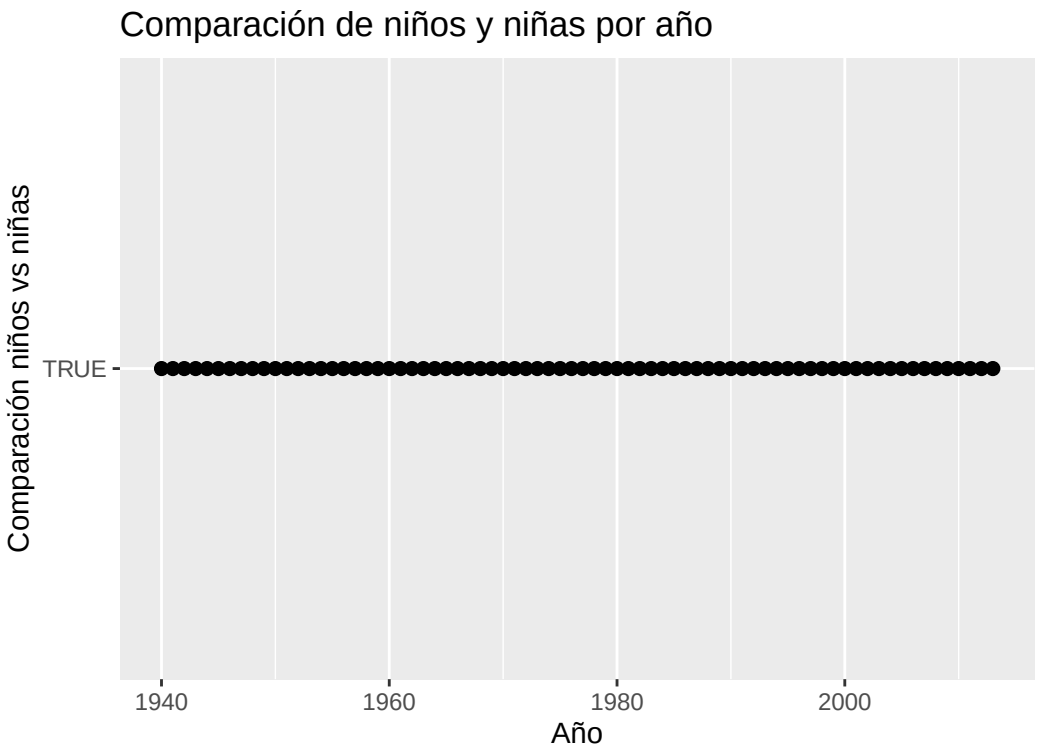
Creemos dos variables, una de ellas para tener el numero total de recién nacidos y la proporción de niños respecto a la misma, asimismo hagamos un grafico que ilustre la tendencia de la serie (si es que tiene)

```
present$total <- present$boys + present$girls # Total by year
present$prop_boys <- present$boys/present$total # Proportion de boys borned in every year


ggplot(data = present, aes(x = year, y = prop_boys)) +
  geom_line(size=1,color="mediumpurple4") +
  geom_point(color="mediumpurple4") +
  xlab("Año") +
  ylab("Cantidad de nacidos") +
  ggtitle("Proporción de niños nacidos entre 1940 a 2013")
```



Veamos si nacen mas niños que niñas por año, **True** si se cumple y **False** caso contrario

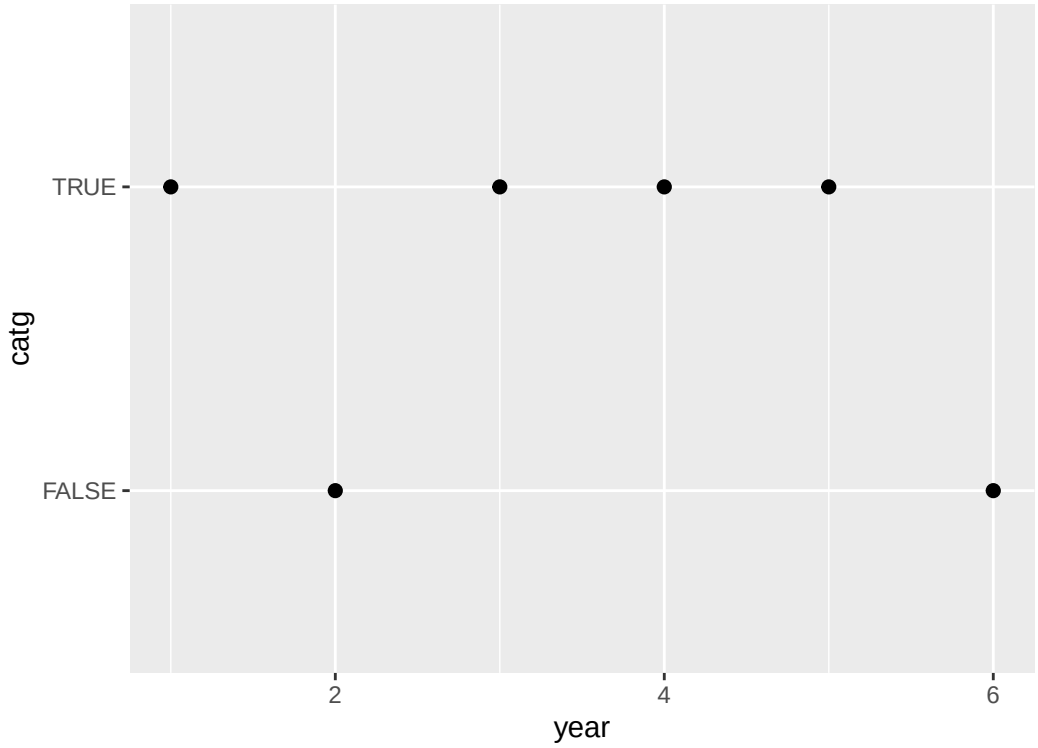
```
present$more_boys <- present$boys > present$girls

ggplot(data = present, aes(x = year, y = more_boys)) +
  geom_point(size = 2) +
  xlab("Año") +
  ylab("Comparación niños vs niñas") +
  ggtitle("Comparación de niños y niñas por año")
```



Como se observa en el grafico, en todos los años la cantidad de niños nacidos fue mayor que la de niñas

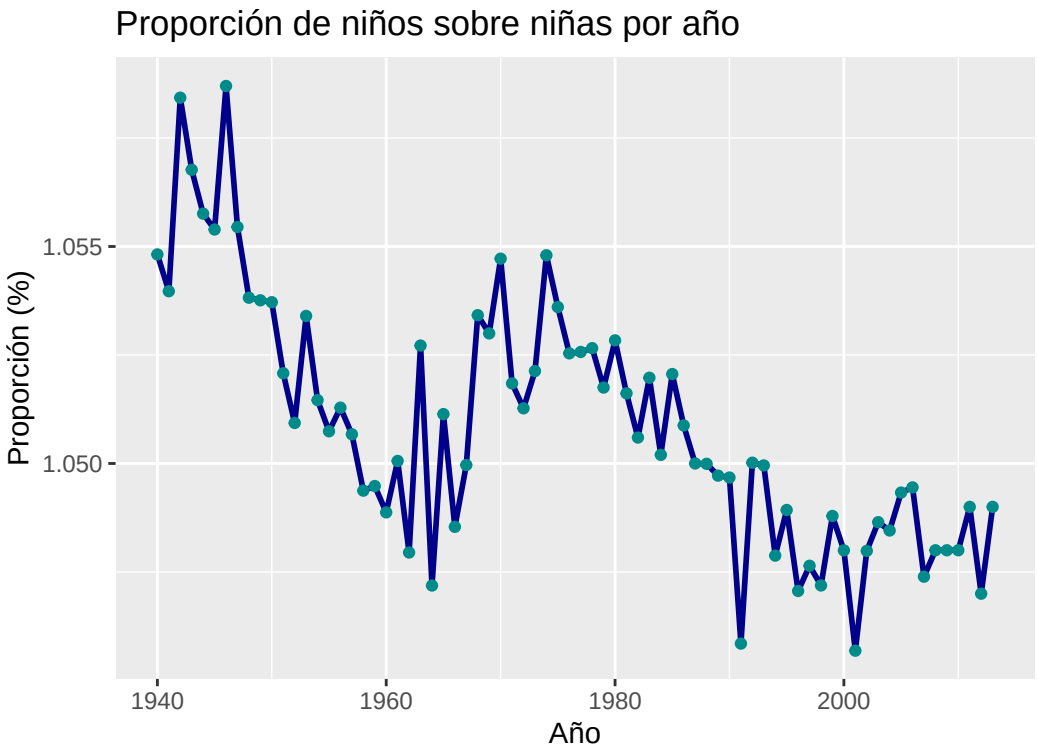
Adicional veamos con un ejemplo si hubiera otra variable

```
# type your code for Question 6 here, and Knit
catg <- c(TRUE,FALSE,TRUE,TRUE,TRUE,FALSE)
year <- c(1, 2, 3, 4, 5,6)

df <- data.frame(year, catg)


ggplot(data = df, aes(x = year, y = catg)) +
  geom_point(size=2)
```



Calculemos ahora la proporción de niño a niña cada año y guardemos estos valores en una variable llamada prop_boy_girl en el conjunto de datos actual. Trace estos valores a lo largo del tiempo para asi analizar su tendencia

```
present$prop_boy_girl <- present$boys/present$girls

ggplot(present, aes(x=year,y=prop_boy_girl)) +
  geom_line(size = 1,color="darkblue") +
  geom_point(color="darkcyan") +
  xlab("Año") +
  ylab("Proporción (%)") +
  ggtitle("Proporción de niños sobre niñas por año")
```



Como podemos observar hay una tendencia a la baja, salvo en el intervalo de 1960 a 1970 donde hubo un incremento.

Gráfico total de nacidos.

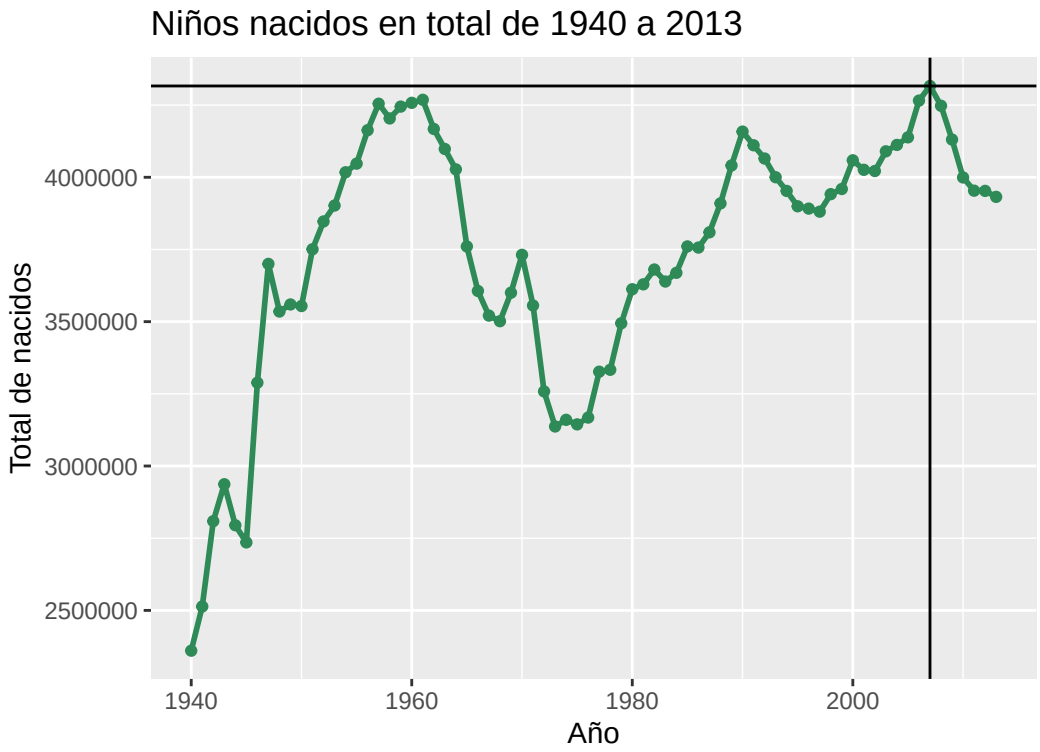
Ahora encontremos en que año se presento el maximo numero de nacimientos registrado.

```
df_max_yr <- present[present$total == max(present$total),]
df_max_yr
```

```
## # A tibble: 1 x 7
##    year    boys   girls   total prop_boys more_boys prop_boy_girl
##   <dbl>   <dbl>   <dbl>   <dbl>     <dbl> <lgl>             <dbl>
## 1  2007 2208071 2108162 4316233     0.512 TRUE               1.05
```

Gráficamente

```
ggplot(present, aes(x=year,y=total)) +
  geom_line(size = 1,color="seagreen4") +
  geom_point(color="seagreen4") +
  geom_vline(xintercept = df_max_yr$year) +
  geom_hline(yintercept = df_max_yr$total) +
  xlab("Año") +
  ylab("Total de nacidos") +
  ggtitle("Niños nacidos en total de 1940 a 2013")
```


Niños nacidos en total de 1940 a 2013
Total de nacidos
4000000
3500000
3000000
2500000
1940
1960
1980
2000
Año