Entrega de ejercicios Tema 1

Blanca Cano Camarero

14 de octubre de 2022

Indice de contenidos

Ejercicio 1

El paquete `gapminder` contiene un fichero de datos de población, esperanza de vida y renta per cápita de los países del mundo entre 1952 y 2007. Instala el paquete y lleva a cabo los siguientes gráficos:

```
#package installation
# uncomment to install
#install.packages("gapminder")
#install.packages("dplyr")
library(gapminder)
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --

v ggplot2 3.3.6     v purrr   0.3.4
v tibble  3.1.8     v stringr 1.4.1
v tidyr   1.2.0     v forcats 0.5.2
v readr   2.1.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```
library(ggtext) #markdown titles
colnames(gapminder)
```

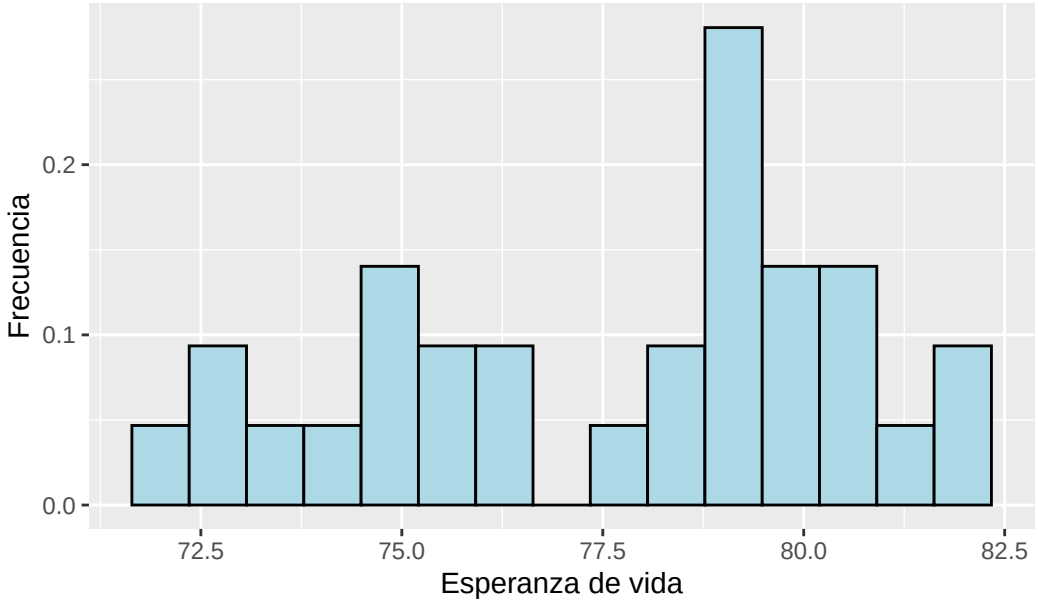
```
[1] "country"   "continent" "year"      "lifeExp"   "pop"       "gdpPercap"
```

Apartado 1.1

Un histograma de la esperanza de vida en 2007 de los países de Europa.

```
life_countries <- gapminder %>%
    filter(
      continent == "Europe",
      year == 2007
    ) %>%
  select(country, lifeExp)
# In order to know the number of bin calc maximum and minimum
# one bin per year
number_of_bin <- ceiling(
  max( life_countries$lifeExp)
  -
    min( life_countries$lifeExp)
)

ggplot(data=life_countries) +
  geom_histogram(aes(x=lifeExp,  y=..density..),
                 fill='lightblue',
                 col='black',
                 bins=15) +
                 labs(x="Esperanza de vida",
                      y="Frecuencia") +
  ggtitle("<span style='font-size: 11pt;'>
  **Grafica 1.1**:
  Esperanza de vida en los países de europa
  </font>") +
  theme(plot.title = element_markdown())
```

Grafica 1.1: Esperanza de vida en los países de europa

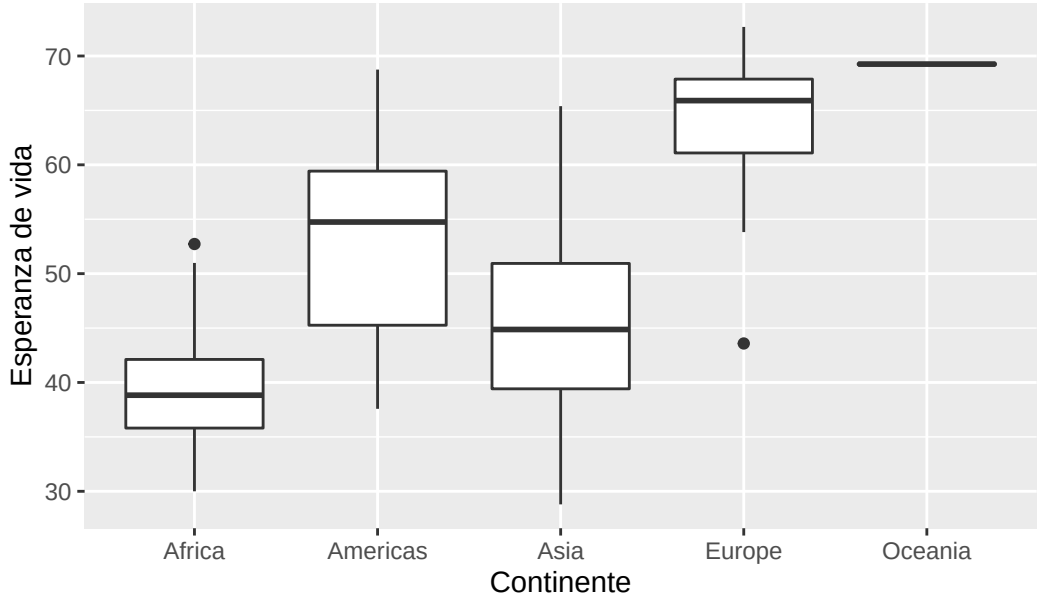


Apartado 1.2

Diagramas de cajas con las esperanzas de vidas de cada continente en el año 1952.

```
filtered <- gapminder %>%
    filter(
      year == 1952
    ) %>%
  select(continent, lifeExp)

ggplot(filtered, aes(x = continent, y = lifeExp )) +
  geom_boxplot() +
  labs(y="Esperanza de vida",
       x= "Continente") +
  ggtitle("<span style='font-size: 10pt;'>
  **Grafica 1.2**:
  Esperanza de vida por continente en 1952
  </font>") +
  theme(plot.title = element_markdown())
```

Grafica 1.2: Esperanza de vida por continente en 1952

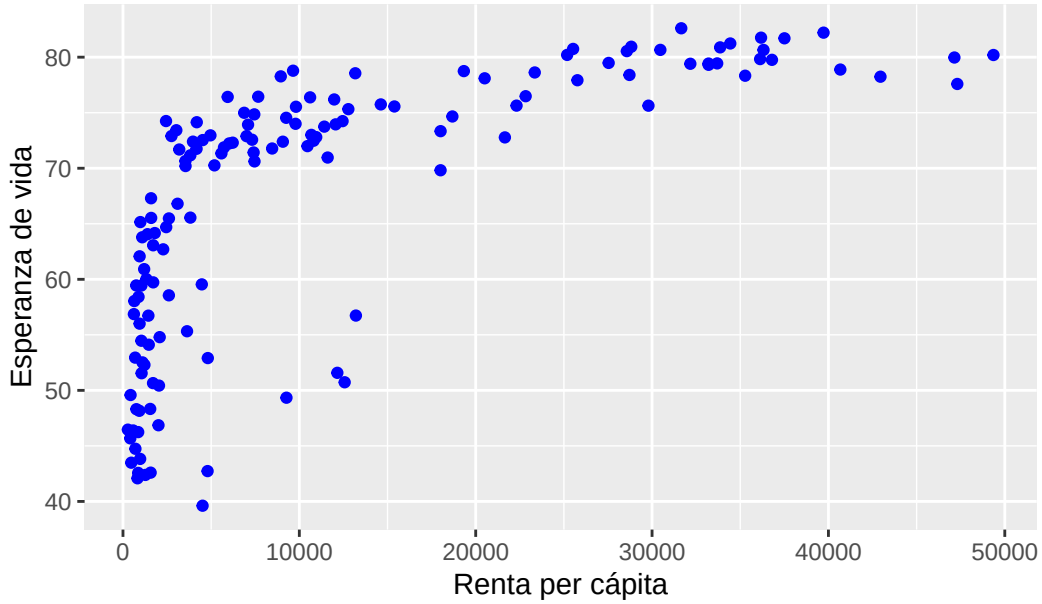


Apartado 1.3

Un diagrama de dispersión de la renta per cápita y la esperanza de vida de cada país en el año 2007.

```
filtered <- gapminder %>%
    filter(
      year == 2007
    ) %>%
  select(gdpPercap, lifeExp)

ggplot(filtered) +
  geom_point(aes(x = gdpPercap, y = lifeExp), col = 'blue')+
  labs(y="Esperanza de vida",
       x= "Renta per cápita") +
  ggtitle("<span style='font-size: 9pt;'>
  **Grafica 1.3**:
   Dispersión de la renta per cápida y esperanza de vida en 2005
  </font>") +
  theme(plot.title = element_markdown())
```

Grafica 1.3: Dispersión de la renta per cápida y esperanza de vida en 2005



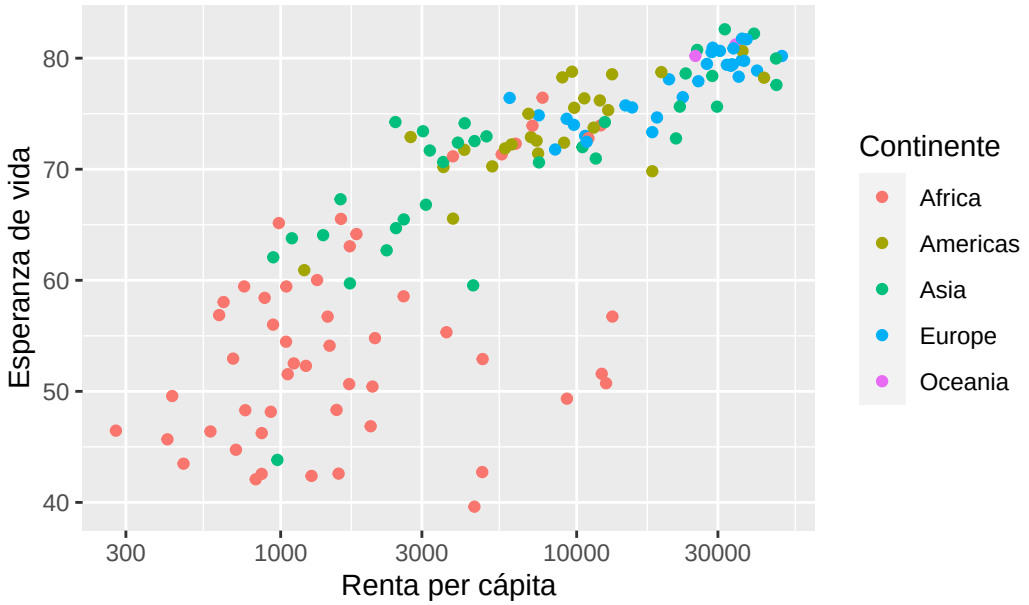
Apartado 1.4

Mejora el gráfico anterior representando cada punto con un color diferente en función del continente al que pertenece cada país y representando la renta per cápita en una escala logarítmica.

```
filtered <- gapminder %>%
    filter(
      year == 2007
    ) %>%
  select(gdpPercap, lifeExp, continent)

ggplot(filtered) +
  geom_point(aes(x = gdpPercap, y = lifeExp, col = continent))+
  scale_x_log10() +
  labs(y="Esperanza de vida",
       x= "Renta per cápita",
       col="Continente") +
  ggtitle("<span style='font-size: 7pt;'>
  **Grafica 1.4**:
   Dispersión de la renta per cápida y esperanza de vida en 2005
   escala logarítmica
```

```
</font>") +
theme(plot.title = element_markdown())
```

Grafica 1.4: Dispersión de la renta per cápida y esperanza de vida en 2005 escala logarítmica



Ejercicio 3

Se desea estimar la prevalencia p de cierto trastorno gástrico. Está relacionada con la edad y por tanto se divide la población en dos estratos:

1. menores de 30 años que son un 40% de la población.
2. mayores de 60 años que son un 60% de la población.

Se toma una muestra de 60 del estrato (1) y otra de 90 del (2). Teniendo entonces una muestra estratificada de tamaño $n = 150$ individuos.

Para cada uno de ellos se observa si tienen o no la enfermedad.

Abstracción del problema 3

Notemos que para ambos estratos estamos ante una distribución binomial, donde Sea $X \sim Bin(60, p_1)$ la variable aleatorio que indica el número de individuos enfermos dentro del estrato (1) y $Y \sim Bin(90, p_2)$ la variable aleatorio que indica el número de individuos enfermos dentro del estrato (2).

Ambas variables son independientes.

Apartado 3.1

A partir de $\hat{p_i}$ la proporción muestral de inividuos enfermos en estrato $i \in \{1, 2\}$ formula un estimador insesgado de la prevalencia de p en la población.

Solución propuesta apartado 3.1

Comenzaremos aprenciando que $\hat{p_i}$ para todo $i \in \{1, 2\}$ es un estimador insesgado, ya que la media lo es:

Sea $W \sim Bernoulli(p_i)$ donde $\hat{p}_i = n_i^{-1} \sum W_i = \bar{W}$

$$E\bar{W} = E\left[n^{-1} \sum_{i=1}^{n} W_i\right] = n^{-1} \sum_{i=1}^{n} EW_i = p_i.$$

Es decir que para todo $i \in \{1, 2\}$

$$E\hat{p}_1 = p_1 \text{ y } E\hat{p}_2 = p_2. \tag{3.1}$$

Además por cómo está distribuida la población se tiene que

$$p = 0.4p_1 + 0.6p_2. \tag{3.2}$$

Proponemos por tanto como estimador a T, definido como:

$$T(X, Y) = 0.4\hat{p}_1 + 0.6\hat{p}_2. \tag{3.3}$$

veamos que (3.3) es insesgado:

$$\begin{aligned} E_{X,Y}T &= E_{X,Y}[0.4\hat{p}_1 + 0.6\hat{p}_2] \\ &= 0.4E_{X,Y}[\hat{p}_1] + 0.6E_{X,Y}[\hat{p}_2] \\ &= 0.4p_1 + 0.6p_2 \\ &= p. \end{aligned}$$

Donde la última igualdad se debe a (3.2).

Acabamos de probar por tanto que T es insesgado.

Apartado 3.2

En función de p_1 y p_2 calcula la varianza del estimador T.

Solución propuesta apartado 3.

Tengamos presente que X e Y son dos variables aleatorias independientes.

Y que además por ser

Por tanto

$$\begin{aligned} Var(T) &= Var(0.4\hat{p}_1 + 0.6\hat{p}_2) \\ &= (0.4)^2 Var(\hat{p}_1) + (0.6)^2 Var(\hat{p}_2) \end{aligned} \tag{3.4}$$

Donde para cada $i \in \{1, 2\}$

$$Var(\hat{p}_i) = E[(\hat{p}_i - Ep_i)^2] = E[(\hat{p}_i - p)^2]$$

Que por tratarse de una binomial será de la forma

$$Var(\hat{p}_i) = n_i p_i (1 - p_i). \tag{3.5}$$

Si lo pensamos como el promedio de la suma de Bernuillis:

$$Var\left(n_1^{-1} \sum_{i=1}^{n_1} X_i \right) = n_1^{-2} \sum_{i=1}^{n_1} Var(X_i) = n_1^{-2} n_1 p_1 (1 - p_1)$$

sustituyendo (3.5) en (3.4) resulta:

$$\begin{aligned} Var(T) &= 0.16 n_1 p_1 (1 - p_1) + 0.36 n_2 p_2 (1 - p_2) \\ &= 9.6 p_1 (1 - p_1) + 32.4 p_2 (1 - p_2) \end{aligned}$$

Apartado 3.3

Si $p_1 = p_2$ ¿Se incrementa la eficiencia por el hecho de usar una muestra estratificada en lugar de una muestra de vaiid de tamaño 150, extraída sin tener en cuenta los estratos.

Solución propuesta apartado 3.3

La clave está en cómo se indique, si se quiere saber la probabilidad de que X jóvenes y Y mayores estén enfermos.

Si $p_1 = p_2$ entonces ambos tendrían la misma distribución $X \sim B(n_1, p_1)$ y $Y \sim B(n_2, p_1)$ y entonces

$$X + Y \sim B(n_1 + n_2, p_1) = B(n, p_1)$$

Por otra parte El parte si tenemos presente la igualdad (3.2) entonces se satiszace que:

$$p = 0.4 p_1 + 0.6 p_2 = 0.4 p_1 + 0.6 p_1 = p_1$$

Por los que el modelo sin estratos sería de la forma $$ Z B(n, p) = B(n,p) $$

Por ser ambos modelos iguales no habría diferencia en estimar al estrato X con Z.

```
9.6+32.5
```

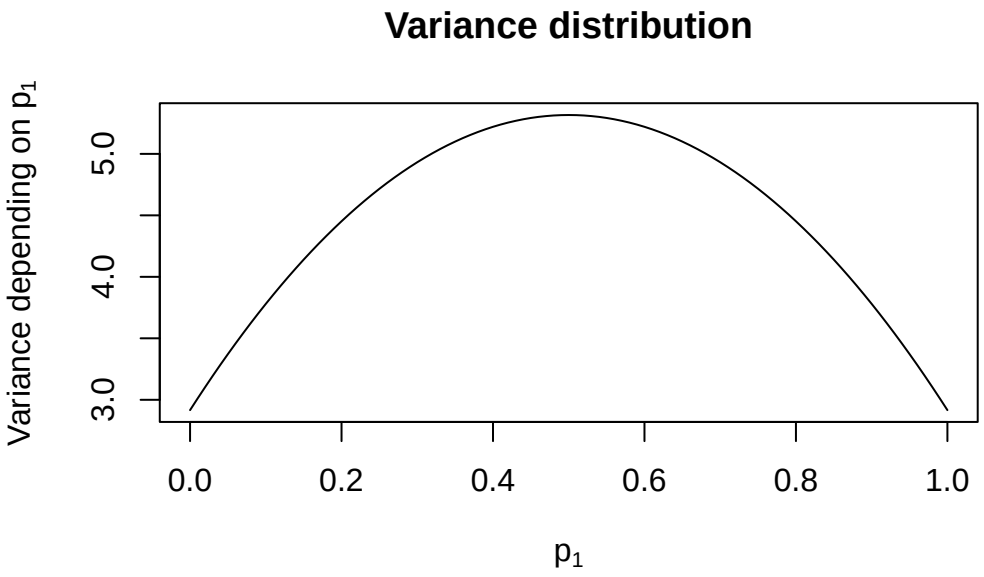
```
[1] 42.1
```

Apartado 3.4

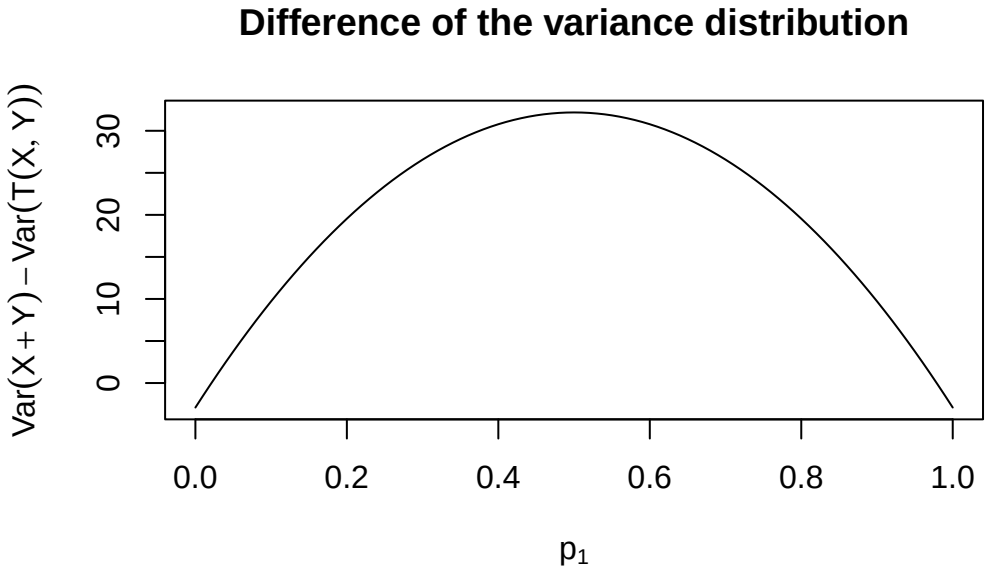
Supongamos que diez de cada cien personas mayores de 30 años tiene la enfermedad ($p_2 = 0.1$). Representa gráficamente las varianzas de los estimadores correspondientes a la muestra n estratificada como función de p_1. ¿Para qué valores de p_1 es mejor utilizar muestreo estratificado en lugar de muestreo aleatorio simple?

Solución propuesta apartado 3.

```r
library(latex2exp)
f <- function (p_1, p_2=0.1){
  return (9.6 *p_1*(1-p_1)+32.4 * p_2 * (1-p_2))
}

# Plotting
x <- seq(0,1,0.01)
plot(
  x,
  f(x),
  type='l',
  main="Variance distribution",
  ylab = TeX(r'(Variance depending on $p_1$)'),
  xlab = TeX(r"($p_1$)")
)
```



```r
# Diferencia del modelo

differences <- function (p_1, p_2=0.1, n = 150){
  return (-(9.6 *p_1*(1-p_1)+32.4 * p_2 * (1-p_2)) + n*p_1*(1-p_1) )
}

# Plotting
x <- seq(0,1,0.01)
plot(
  x,
  differences(x),
  type='l',
  main="Difference of the variance distribution",
  ylab = TeX(r'($Var(X+Y)-Var(T(X,Y))$)'),
  xlab = TeX(r"($p_1$)")
)
```




```r
#Cuanto mayor ser sea |p1 - o.5| más renta

# when the minimum is found
library(purrr)
optimize(f, c(0,1))
```

```
$minimum
```

```
[1] 6.610696e-05

$objective
[1] 2.916635
```

Ejercicio 7

El siguiente código genera una muestra de 100 datos de una distribución de Cauchy con parámetro de posición:

```
set.seed(123)
theta <- 10
n <- 100
muestra <- rt(n, 1) + theta
```

Apartado 7.1

Calcula el estimador de máxima verosimilitud de θ. ¿Se parece al valor verdadero?

Solución propuesta apartado 7.1

Definimos la función a minimizar L como

$$L(\theta) = -\sum_{i=1}^{n} \log(1 + (x_i - \theta)^2)$$

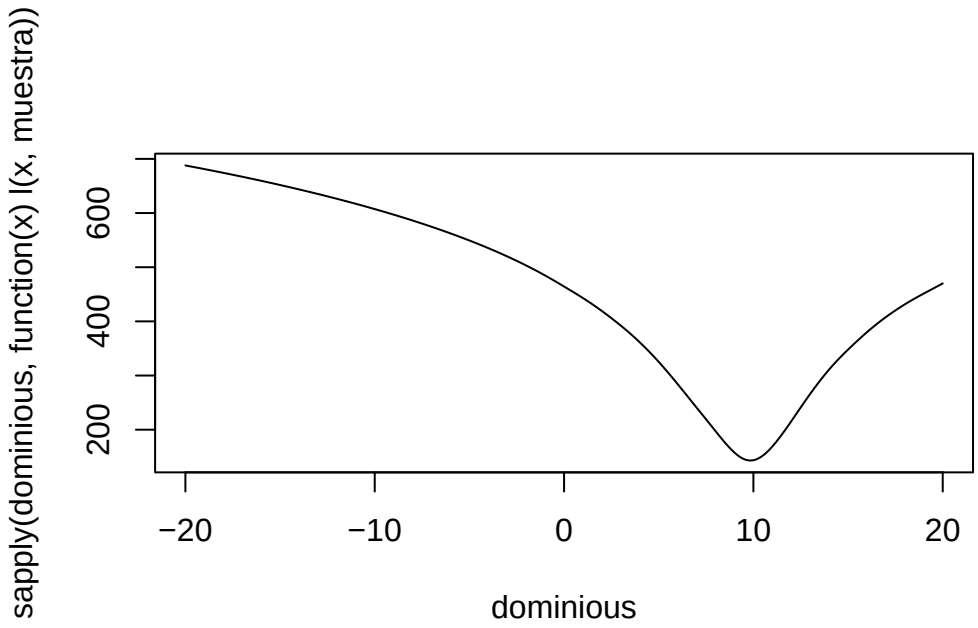
y la minimizaremos numéricamente con R:

```
# Cambiamos el signo porque optimize solo encuentra mínimos
# esta función es proporcional a la opuesta de de máximo similitud, su máximo será el esti
l <- function (theta, sample){
  return (sum(
              sapply(
                sample,
                function(x) log(1 + (x-theta)^2)
              )
          )
  )
}

# Calcula el mínimo de la función anterior dentro de un intervalo lo suficientemente grand
get_stimator <- function(sample) {
  stimator <-optimize(
      function(theta) l(theta, sample),
      c(-100,100)
  )
```

```r
  return (
    stimator$minimum
  )
}
estimador <- get_stimator(muestra)
cat('El estimador máximo verosimil encontrado es: ', estimador)
```

```
El estimador máximo verosimil encontrado es:  9.842954
```

```r
dominious<-seq(-20, 20, 0.2)
plot(
  dominious,
  sapply(dominious, function(x) l(x,muestra)),
  type='l'
)
```



Como podemos observar se ha encontrado un mínimo en $\theta^* = 9.842954$ relativamente próximo al valor real que es $\theta = 10$ que está próximo.

Apartado 7.2

Lleva a cabo algún experimento de simulación para aproximar la varianza del estimador de máxima verosimilitud.

Solución al apartado 7.2

El diseño del experimento consistirá en generar una matriz $n \times m$ de muestras, calcular el estimador verosimil para cada fila $\theta^{(i)}$ para cada $i \in \{1, \dots, n\}$y con ellos se calculará la varianza estimada

$$Var(\theta^*) = \sum_{i=1}^{n} \left(\theta^{(i)} - \theta^*\right)^2$$

```r
set.seed(123)
m = 1000
matriz_muestras <- rt(n, m) + theta
estimador_por_filas <- sapply(matriz_muestras, get_stimator)
varianza <- sum(
  sapply(
    estimador_por_filas,
    function(x) (x-estimador)^2
  )
)
cat("La varianza de nuestro estimador es ", varianza)
```

```
La varianza de nuestro estimador es  103.4079
```