Entrega de ejercicios Tema 1

Blanca Cano Camarero

17 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:

```r
#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
```

```r
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()
```

```r
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.

Solución

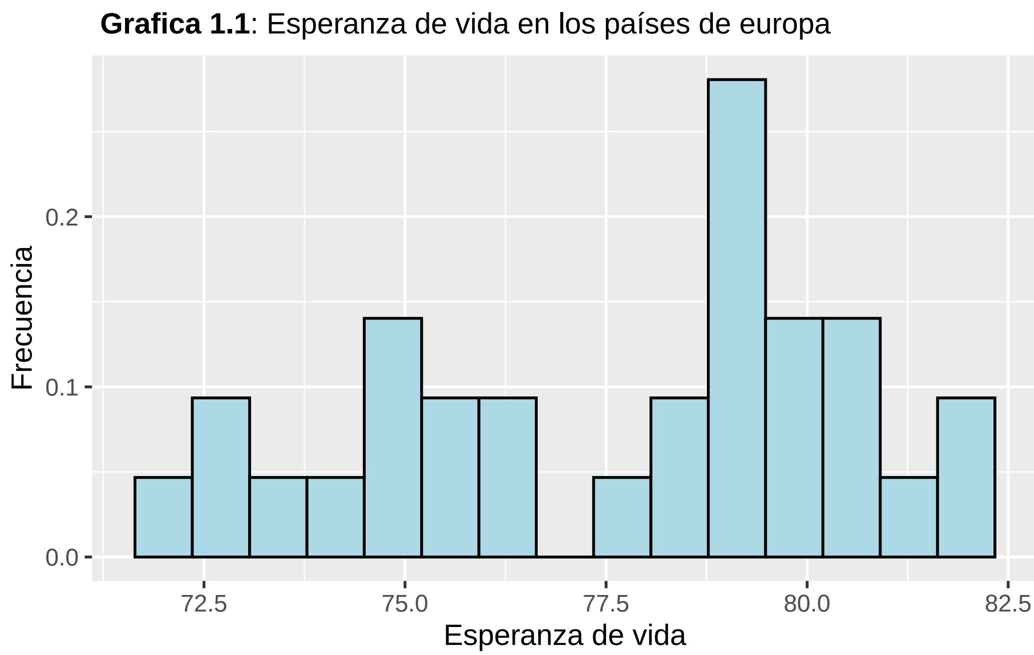
```
life_countries <- gapminder %>%
    filter(
      continent == "Europe",
      year == 2007
    ) %>%
  select(country, lifeExp)
```

Es necesario inicar cuántas barras van a ser necesaras en nuestro histograma, para ello hemos optado por una barra por año, esto es calculado con

```
# 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)
)
```

A continuación mostramos el código del histograma:

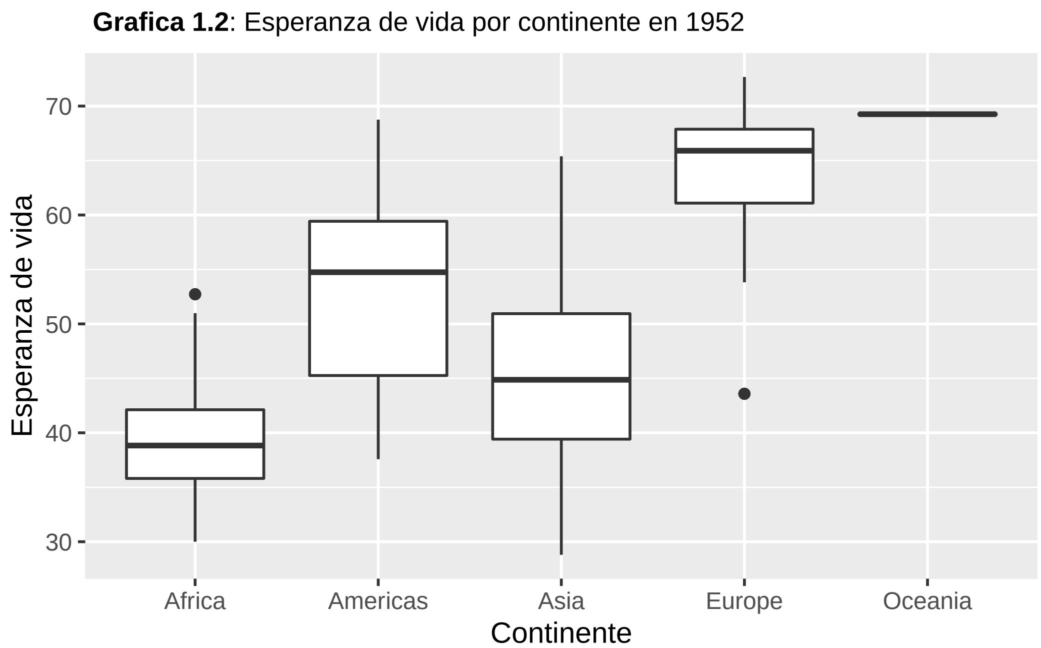
```
ggplot(data=life_countries) +
  geom_histogram(aes(x=lifeExp,  y=..density..),
                 fill='lightblue',
                 col='black',
                 bins=15) +
                 labs(x="Esperanza de vida",
                      y="Frecuencia") +
  ggtitle("<span style='font-size: 11pt;'>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



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())
```

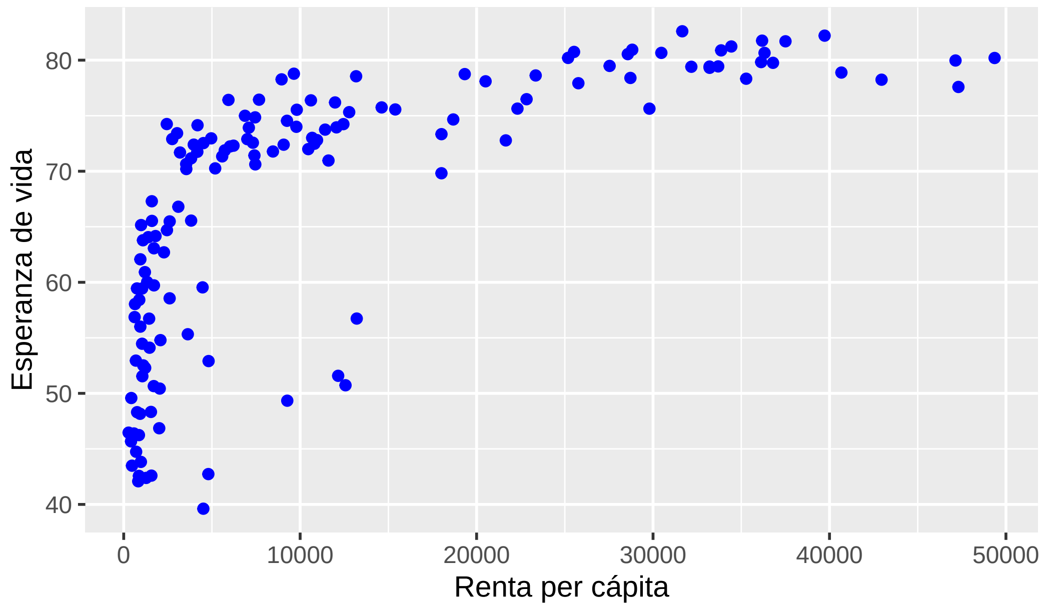


Apartado 1.3

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

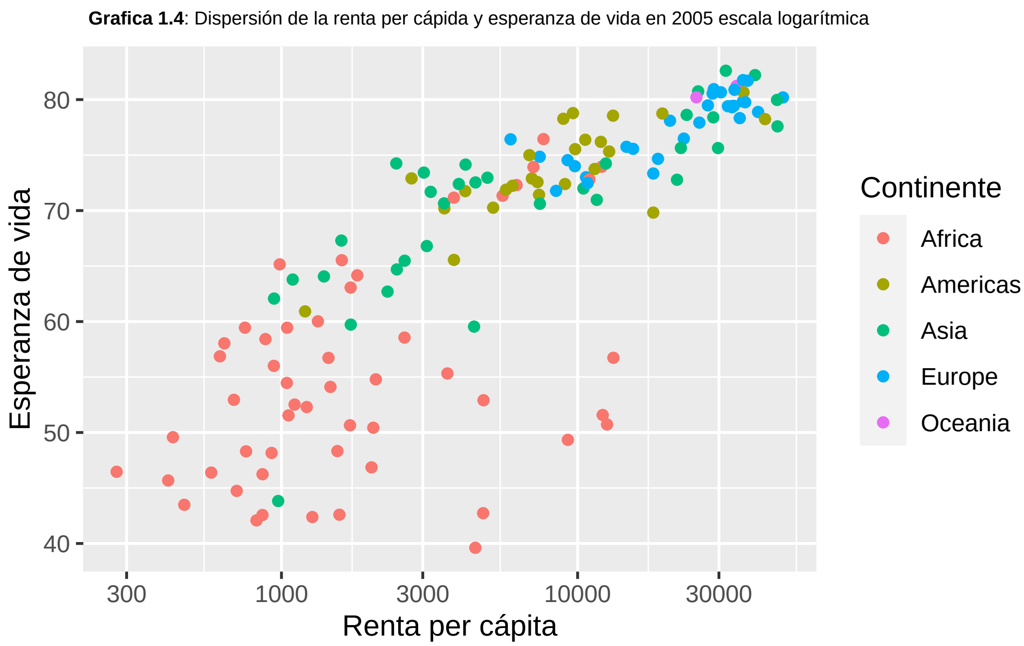


Apartado 1.4

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())
```



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 $n_1 = 60$ del estrato (1) y otra de $n_2 = 90$ 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

Sea W_{ij} la variable aleatoria que indica si el sujeto del estrato $i \in \{1, 2\}$ y $j \in \{1, \dots, n_i\}$ Se tiene que la variable aleatoria sigue una distribución de Bernouilli de parámetro p_i esto es $W_{ij} \sim \text{Bernouilli}(p_i)$.

Notemos que para ambos estratos estamos ante una distribución binomial, donde:

Denotaremos como $X = \sum_{j=1}^{n_1} W_{1j} \sim Bin(60, p_1)$ a la variable aleatorio que indica el número de individuos enfermos dentro del estrato (1) y como $Y = \sum_{j=1}^{n_2} W_{2j} \sim Bin(90, p_2)$ 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 $\hat{p}_i = n_i^{-1} \sum_{j=1}^{n_i} W_{ij} = \bar{W}$

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

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 la prevalencia en la población total puede calcularse a partir de la prevalencia en cada uno de los estratos como:

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

Es natural por tanto proponer como estimador de p 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.

De la independencia de X e Y se deduce que:

$$\begin{aligned}
Var(T) &= Var(0.4\hat{p}_1 + 0.6\hat{p}_2) \\
&= Var(0.4n_1^{-1}X_1 + 0.6n_2^{-1}X_2) \\
&= \frac{0.4^2}{60^2}Var(X_1) + \frac{0.6^2}{90^2}Var(X_2)
\end{aligned} \tag{3.4}$$

Que por tratarse de una binomial será de la forma

$$Var(X_i) = n_i p_i(1-p_i). \tag{3.5}$$

sustituyendo (3.5) en (3.4) resulta:

$$\begin{aligned} Var(T) &= \frac{0.4^2}{n_1^2} n_1 p_1(1-p_1) + \frac{0.6^2}{n_2^2} n_2 p_2(1-p_2) \\ &= \frac{0.4^2}{60} p_1(1-p_1) + \frac{0.6^2}{90} n_2 p_2(1-p_2) \\ &= \frac{1}{375} p_1(1-p_1) + \frac{1}{250} p_2(1-p_2). \end{aligned} \tag{3.6}$$

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

Se dice que un estimador $\hat{\theta}_1$ es más eficiente que otro estimador $\hat{\theta}_2$ si

$$Var(\hat{\theta}_1) < Var(\hat{\theta}_2).$$

Si $p_1 = p_2$ entonces ambos tendrían la misma distribución, es decir $X \sim B(n_1, p_1)$ y $Y \sim B(n_2, p_1)$. Además la variable aleatoria que no tiene encuentra los estratos es la suma de las dos, esto es:

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

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

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

Por lo que Z la variable aleatoria que contiene el número de enfermos en toda la población vendría dada como:

$$Z = X + Y \sim B(n, p_1) = B(n, p).$$

Si el estimador sigue siendo el promedio entonces

$$\hat{p} = n^{-1} \sum_{i=1,j=1}^{2,n_i} W_{ij} = n^{-1}(X+Y)$$

Y su varianza vendría determinada por

$$\begin{aligned} Var(n^{-1}(X+Y)) &= n^{-2}np(1-p) \\ &= n^{-1}p(1-p) \\ &= \frac{1}{150}p(1-p). \end{aligned} \tag{3.7}$$

Si hacemos $p_1 = p_2 = p$ en la varianza del estimador estratificada $Var(T)$ obtenida en (3.6) resulta

$$\begin{aligned} Var(T) &= \frac{1}{375}p_1(1-p_1) + \frac{1}{250}p_2(1-p_2) \\ &= \frac{1}{375}p(1-p) + \frac{1}{250}p(1-p) \\ &= \left(\frac{1}{375} + \frac{1}{250}\right)p(1-p) \\ &= \frac{1}{150}p(1-p). \end{aligned} \tag{3.8}$$

Como podemos observar (3.7) y (3.8) dan lugar a la misma varianza, luego podemos afirmar que una no es más eficiente que otra en el caso $p_1 = p_2$.

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

La eficiencia es un indicador de precisión, cuando menor sea la varianza menor será los errores medios cometidos. Planteamos por tanto la función diferencia de varianzas:

Teniendo presente (3.2) y (3.7) tenemos que

$$\begin{aligned} Var(\hat{p}) &= \frac{1}{n}(0.4p_1 + 0.6p_2)(1 - (0.4p_1 + 0.6p_2)) \\ &= \frac{1}{150}(0.4p_1 + 0.6 \times 0.1)(1 - (0.4p_1 + 0.6 \times 0.1)) \\ &= \frac{1}{150}(0.4p_1 + 0.06)(1 - (0.4p_1 + 0.06)) \end{aligned} \tag{3.9}$$

donde (3.9) puede verse como una función dependiente de p_1.

Sustituyendo $p_2 = 0.1$ en (3.6) obtenemos la siguiente función dependiente de p_1:

$$\begin{aligned} Var(T) &= \frac{1}{375}p_1(1 - p_1) + \frac{1}{250}p_2(1 - p_2) \\ &= \frac{1}{375}p_1(1 - p_1) + \frac{1}{250}0.1(1 - 0.1) \\ &= \frac{1}{375}p_1(1 - p_1) + \frac{9}{25000}. \end{aligned} \tag{3.10}$$

Definimos ahora la función $diferencia : [0, 1] \longrightarrow \mathbb{R}$

$$\begin{aligned} diferencia(p_1) &= [Var(\hat{p})](p_1) - [Var(T)](p_1) \\ &= \frac{1}{150}(0.4p_1 + 0.06)(1 - (0.4p_1 + 0.06)) - \frac{1}{375}p_1(1 - p_1) - \frac{9}{25000} \\ &= 0.0016p_1^2 - 0.00032p_1 + 0.000016. \end{aligned} \tag{3.11}$$

A la vista de (3.11) podemos ver que

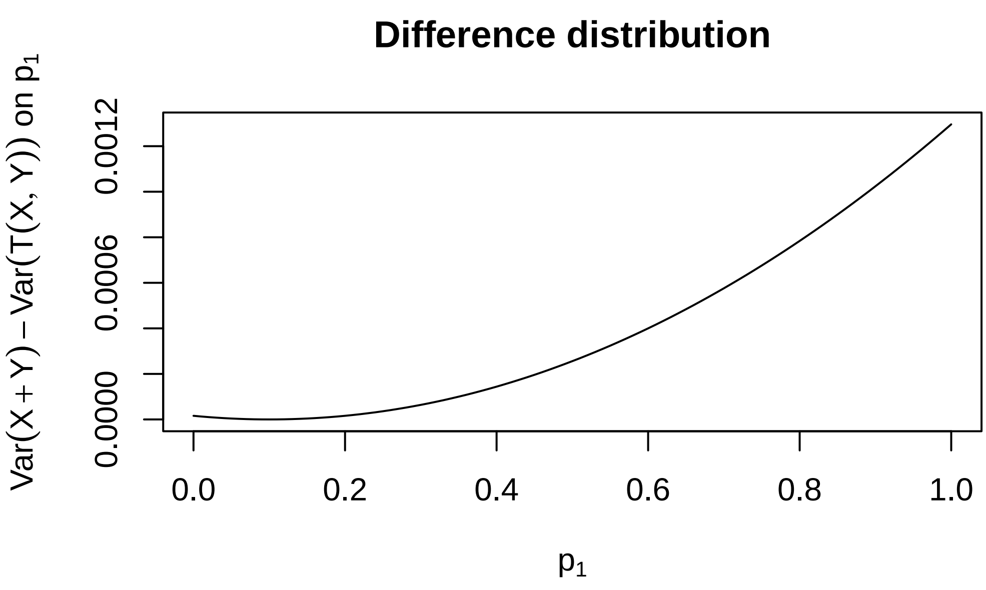
$$\begin{aligned} \frac{\partial}{\partial p_1} diferencia(p_1) &= \frac{\partial}{\partial p_1} 0.0016p_1^2 - 0.00032p_1 + 0.000016 \\ &= \frac{\partial}{\partial p_1} 0.0032p_1 - 0.00032 \end{aligned} \tag{3.12}$$

alcanza un mínimo en $p_1 = 0.1$ cuyo valor es

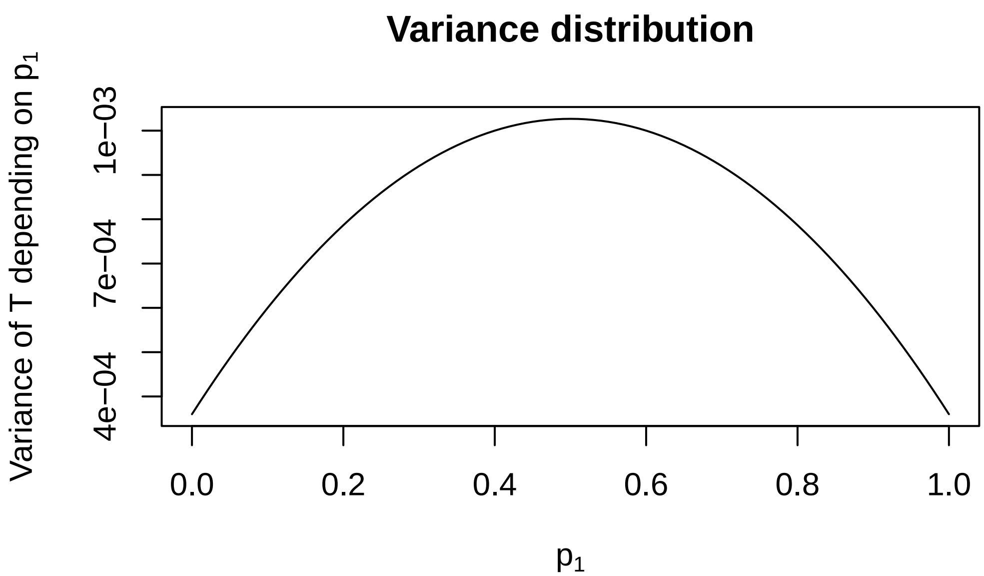
$$
\begin{aligned}
diferencia(p_1 = 0.1) &= 0.0016p_1^2 - 0.00032p_1 + 0.000016 \\
&= \frac{16}{10^6} - \frac{32}{10^6} + \frac{16}{10^6} \\
&= 0.
\end{aligned}
$$

Es decir, salvo en $p_1 = 0.1$ que sería indiferente, para el resto de casos es mejor usar el estimador estratificado. (Notemos además que este es el caso en que $p_1 = p_2$ del apartado anterior).

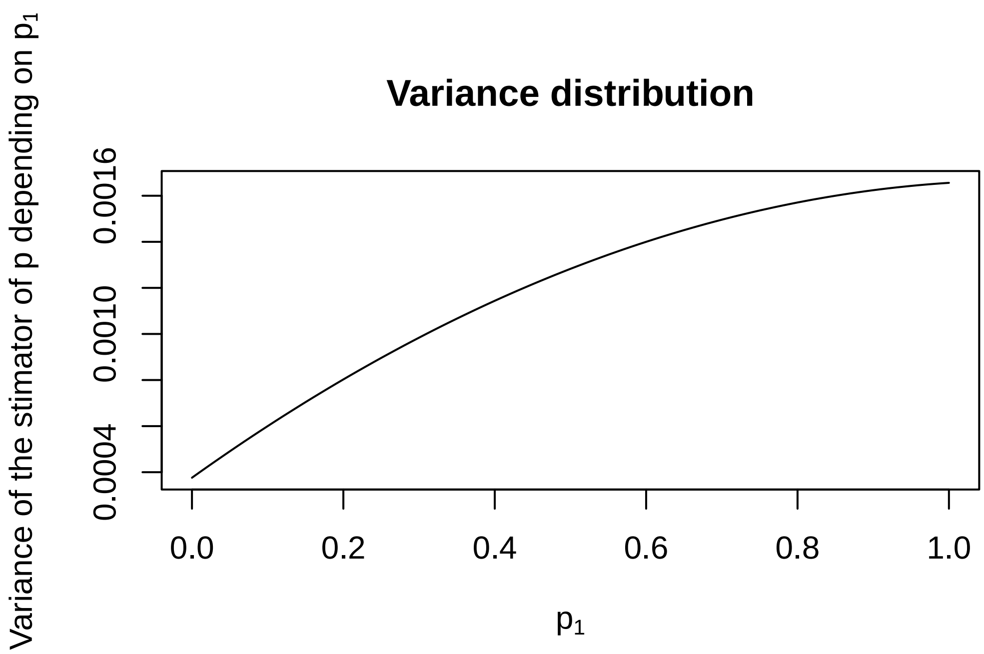
```
library(latex2exp)
diferencia<- function (p_1){
  return (0.0016*p_1^2-0.00032*p_1+0.000016)
}

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



```r
library(latex2exp)
Var_T <- function (p_1, p_2=0.1){
  return (
            (1/375)*p_1*(1-p_1)
          +
            (1/250)* p_2 *(1-p_2)
          )
}
Var_p <-
  function (p_1, p_2=0.1){
  return (
            (1/150)*(0.4*p_1+0.6*p_2)*(1-(0.4*p_1+0.6*p_2))
          )
}
# Plotting
x <- seq(0,1,0.01)
plot(
  x,
  Var_T(x),
  type='l',
  main="Variance distribution",
  ylab = TeX(r'(Variance of T depending on $p_1$)'),
  xlab = TeX(r"($p_1$)")
)
```

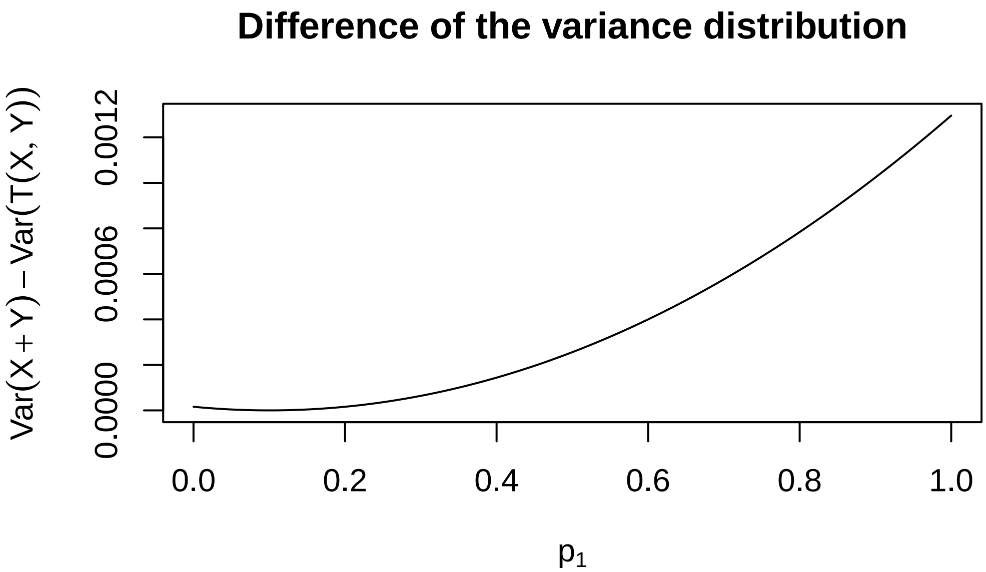


```r
# Plotting
x <- seq(0,1,0.01)
plot(
  x,
  Var_p(x),
  type='l',
  main="Variance distribution",
  ylab = TeX(r'(Variance of the stimator of $p$ depending on $p_1$)'),
  xlab = TeX(r"($p_1$)")
)
```




```r
# Diferencia del modelo

differences <- function (p_1, p_2=0.1, n = 150){
  return (Var_p(p_1) - Var_T(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"($p_1$)")
)
```



Ejercicio 7

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

```r
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 maximizar L como

$$L(\theta) = -\sum_{i=1}^{n} \log(1 + (x_i - \theta)^2) \tag{7.1}$$

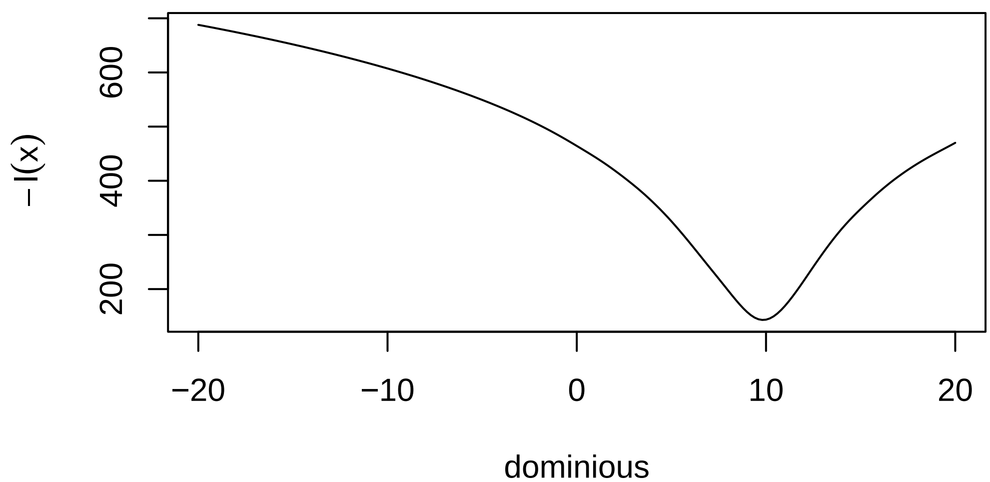
y minimizaremos numéricamente con R el opuesto de (7.1):

```r
# El mínimo de esta función es el máximo de la función de verosimilitud,
# Su máximo será el estimador buscado
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 grande
get_stimator <- function(sample) {
  stimator <-optimize(
      function(theta) l(theta, sample),
```

```
        c(-100,100)
    )
    return (
      stimator$minimum
    )
  }
  estimador <- get_stimator(muestra)
  cat('El estimador máximo verosimil encontrado es: ', estimador)
```

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

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



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

Cabe destacar que deberíamos maximizar L en todo $\mathbb{R}$, para poder encontrar el EMV, algo computacionalmente imposible con este procedimiento. Por lo que apriori si somos estrictos estariamos tan solo frente a un máximo local.

¿Cómo se podría demostrar entonces que se trata de un máximo global?

1. De manera analítica, por la monotomía del logaritmo es fácil ver que para valores menores de -100 y mayores de 100 (valores concretos de nuestro intervalo de búsqueda) la función que estamos minimizando va a seguir creciendo por los lados.
2. Se podría plantear un intervalo de confianza. Puesto que estamos con una distribución de Cauchy y no existe su media habría que hacerlo con la mediana.

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

$$Var(\hat{\theta}) = n^{-1} \sum_{i=1}^{n} \left(\theta^{(i)} - E[\hat{\theta}] \right)^2$$

```
set.seed(123)
m = 100
matriz_muestras <- matrix(rt(n*m, 1), n) + theta
estimador_por_filas <- apply(matriz_muestras,2, get_stimator)
varianza <- var(estimador_por_filas)
cat("La varianza de nuestro estimador es ", varianza)
```

```
La varianza de nuestro estimador es  0.01879276
```