(a) Utilizando el fichero `Datos-lipidos.txt`, estima, mediante un intervalo de confianza de nivel 0.95, la proporción de pacientes que tienen una concentración de colesterol superior o igual a 220 mg/dl. ¿Qué tamaño muestral habría que usar para tener una probabilidad aproximada de 0.95 de no cometer un error mayor que 0.01 en la estimación de esta proporción?.

(b) Suponiendo que la distribución de la variable "concentración de colesterol" fuese aproximadamente normal, obtener un estimador puntual para la proporción indicada en el apartado anterior.

(a) Denotamos por C la concentración de colesterol en un paciente elegido al azar y por X la v.a. $\mathbb{1}_{\{C\geq 220\}}$, que vale 1 si el colesterol del paciente es superior o igual a 220 mg/dl y 0 si no. Entonces X sigue una distribución Bernoulli(p), con $p = \mathbb{P}\{C \geq 220\}$. Un intervalo de confianza al nivel $1-\alpha$ para p es

$$\mathrm{IC}_{1-\alpha}(p) = \left(\bar{x} \mp z_{\alpha/2} \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \right). \tag{1}$$

Calculemos con R el intervalo de confianza de interés $\mathrm{IC}_{0.95}(p) = (0.423 \mp 0.050) = (0.3730.473)$:

```
Datos = read.table('Datos-lipidos.txt')
C = Datos$V1
X = C>=220
mX = mean(X)
z = qnorm((1-0.95)/2,0,1,lower.tail=FALSE)
n = length(X)
dt = sqrt(mX*(1-mX)/n)
IC = c(mX-z*dt,mX+z*dt)
```

Para determinar el tamaño muestral n tal que el error sea menor o igual que 0.01 con una confianza aproximada del 95 %:

$$z_{0.025} \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \leq 0.01.$$

Podemos sustituir $\bar{x}$ por la estimación 0.423, proporcionada por los datos que tenemos en el fichero `Datos-lipidos.txt`:

$$1.96 \sqrt{\frac{0.423(1-0.423)}{n}} \leq 0.01 \Rightarrow n \geq 1.96^2\ 0.423\ (1-0.423)\ 10^4 = 9376.2.$$

Si acotamos $\bar{x}(1-\bar{x})$ por 0.25, el máximo de la función $x(1-x)$ para $x \in [0,1]$,

$$1.96 \sqrt{\frac{0.25}{n}} \leq 0.01 \Rightarrow n \geq 1.96^2\ 0.25\ 10^4 = 9604.$$

(b) Si suponemos que $C \sim N(\mu, \sigma)$ (lo cual no parece descabellado observando la figura), entonces $p = \mathbb{P}\{Z \geq \frac{220-\mu}{\sigma}\}$, siendo Z una N(0,1). Como μ y σ son desconocidos los estimamos mediante $\bar{c} = 213.84$ y $s_c = 43.73$ respectivamente, con lo que obtenemos la siguiente estimación de p:

$$\hat{p} = \mathbb{P}\left\{Z \geq \frac{220 - 213.84}{43.73}\right\} = 0.44.$$