El conjunto de datos `168Perros.txt` (a bajar de la página web de la asignatura) contiene medidas de los niveles de glucosa (en mg/dl) y alanina aminotransferasa (ALT, U/l) en una muestra de 168 perros aparentemente sanos.

a) Dibuja un histograma de la variable ALT e indica razonadamente si los datos se podrían modelizar mediante una distribución normal. Dibuja después un histograma del logaritmo de dicha variable y responde a la misma pregunta.

b) Si quisiéramos ajustar un modelo N(μ,σ) a log(ALT), ¿cómo estimaríamos los valores de los parámetros μ y σ? Dibuja la densidad normal con los parámetros estimados sobre el histograma para comprobar la bondad del ajuste.

c) Bajo la hipótesis de normalidad de (b), ¿qué probabilidad hay de que un perro sano tenga un nivel de log(ALT) superior a 4?. Compara esta probabilidad con la frecuencia observada en los datos.

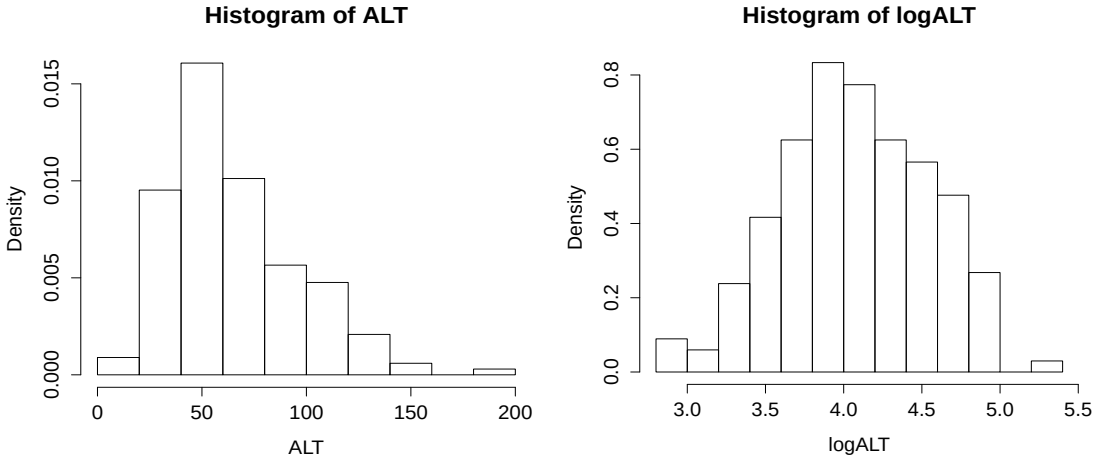
Fuente de los datos: Kaneko, Harvey y Bruss (2008). *Clinical Biochemistry of Domestic Animals*. Elsevier Academic Press.

Solución: Primero cargamos los datos y guardamos la variable ALT:

```
Datos = read.table("168Perros.txt",header=TRUE)
ALT = Datos$ALT
```

a) Dibujamos los histogramas de ALT y log(ALT) respectivamente:

```
hist(ALT,freq=FALSE)
logALT = log(ALT)
hist(logALT,freq=FALSE)
```



Es evidente que el histograma de la variable ALT muestra asimetría hacia la derecha. Sabemos que una muestra de una v.a. normal debería exhibir poca asimetría porque la densidad normal es simétrica respecto a la media. Así que la hipótesis de normalidad no sería adecuada para la variable ALT, pero sí para la variable log(ALT). El histograma de log(ALT) muestra una asimetría más atenuada que podría ser fruto de la propia aleatoriedad en el muestreo.

b) Si suponemos que log(ALT) sigue una distribución N(μ,σ), entonces los e.m.v. de μ y σ^2 serían respectivamente

$$\hat{\mu} = \text{media muestral de log(ALT)} = 4.08$$

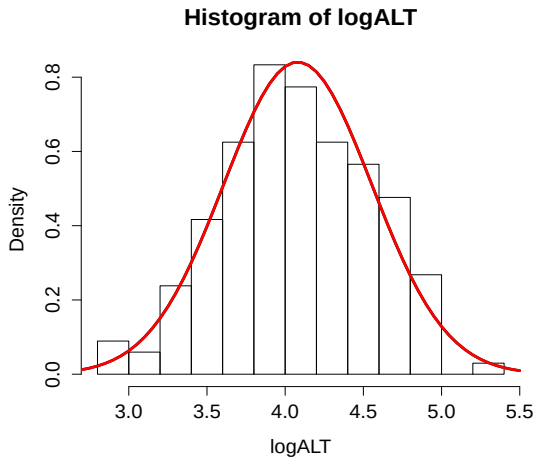
y

$$\hat{\sigma}^2 = \text{varianza muestral de log(ALT)} = 0.23.$$

```
m = mean(logALT)
n = length(ALT)
v = (n-1)*var(logALT)/n
```

Para superponer la densidad $N(\hat{\mu}, \hat{\sigma})$ al histograma de log(ALT):

```
x = seq(2,5.5,0.05)
d = dnorm(x,mean=m,sd=sqrt(v))
lines(x,d,type="l",lwd=3,col="red",xlab="",ylab="")
```



En el dibujo vemos que el ajuste de la normal a los datos de log(ALT) es bueno.

c) Si suponemos que log(ALT) $\sim$ N(4.08,$\sqrt{0.23}$) aproximadamente, entonces

$$\begin{aligned} P\{\log(\text{ALT}) > 4\} \quad &\simeq P\left\{Z > \frac{4-4.08}{\sqrt{0.23}}\right\} = P\{Z > -0.17\} \\ &= 1 - P\{Z > 0.17\} = 1 - 0.4325 = 0.5675, \end{aligned}$$

siendo Z una v.a. N(0,1). La frecuencia relativa observada de los datos mayores que 4 es:

```
sum(logALT>4)/n
[1] 0.547619
```

muy cercana a la probabilidad teórica.