UNIVERSIDAD DE LAS PALMAS
DE GRAN CANARIA

MÉTODOS ESTADÍSTICOS
LECTURAS 4,5,6,7
LABS 3, 4, 5
SEGUNDA ENTREGA

AITOR VENTURA DELGADO

GRADO EN INGENIERÍA INFORMÁTICA

03 DE DICIEMBRE DE 2019

ESCUELA DE INGENIERÍA INFORMÁTICA

ÍNDICE

LECTURA 4: CUESTIÓN 1

A un operador de lavado de coches se le paga en función del número de vehículos que lava. Supóngase que las probabilidades de que entre las 17:00 y 18:00 de cualquier jueves cobre una cierta cantidad ***C****i en euros vienen dadas por la siguiente tabla:*

C_i	7	9	11	13	15	17
p_i	1/12	1/12	1/4	1/4	1/6	1/6

Calcular la ganancia esperada del operador para este tramo horario y establecer una medida coherente de su variabilidad. Explicar las respuestas.

Para calcular la ganancia, teniendo en cuenta que se trata de una variable aleatoria discreta, se puede obtener multiplicando cada uno de los valores de la primera fila por su probabilidad correspondiente, y sumando los productos.

```
> prob <- c(1/12, 1/12, 1/4, 1/4, 1/6, 1/6)
> dinero <- c(7,9,11,13,15,17)
> ganancia <- sum(dinero*prob)
> ganancia
[1] 12.66667
```

Se obtendría una ganancia de 12.67€.

Para calcular la variabilidad, simplemente podemos hacer la diferencia del valor más alto con el más bajo. Ello nos daría un resultado de 0.167.

```
> variabilidad <- 1/4 - 1/12
> variabilidad
[1] 0.1666667
```

LECTURA 4: CUESTIÓN 2

Se están analizando las proporciones del presupuesto que una empresa industrial del Polígono de Arinaga destina a controles medioambientales y de contaminación. Para ello se lleva a cabo un proyecto de recopilación de datos típico de Data Science. En el desarrollo de este se determina que la distribución de tales proporciones está dada por:

$$f(y) = \begin{cases} 5(1-y)^4, & 0 \le y \le 1 \\ 0, & otro\ caso \end{cases}$$

a) Verificar que la función de densidad anterior es válida.

La función es válida porque para cada suceso X del espacio muestral, f(x) resulta ser mayor o igual a 0, y para cada suceso diferente cierto o seguro, f(x) resulta ser un valor concreto.

b) ¿Cuál es la probabilidad de que una empresa elegida al azar gaste menos del 10% de su presupuesto en controles medioambientales y de contaminación?

```
> p_m_10 <- function(x) 5*(1-x)^4
> integral_10 <- integrate(p_m_10, lower = 0, upper = 0.1)
> integral_10
0.40951 with absolute error < 4.5e-15
```

Para calcular esta probabilidad, calculamos la integral definida de la función entre 0 y 0.10. Ello nos da una probabilidad que es 0.40951.

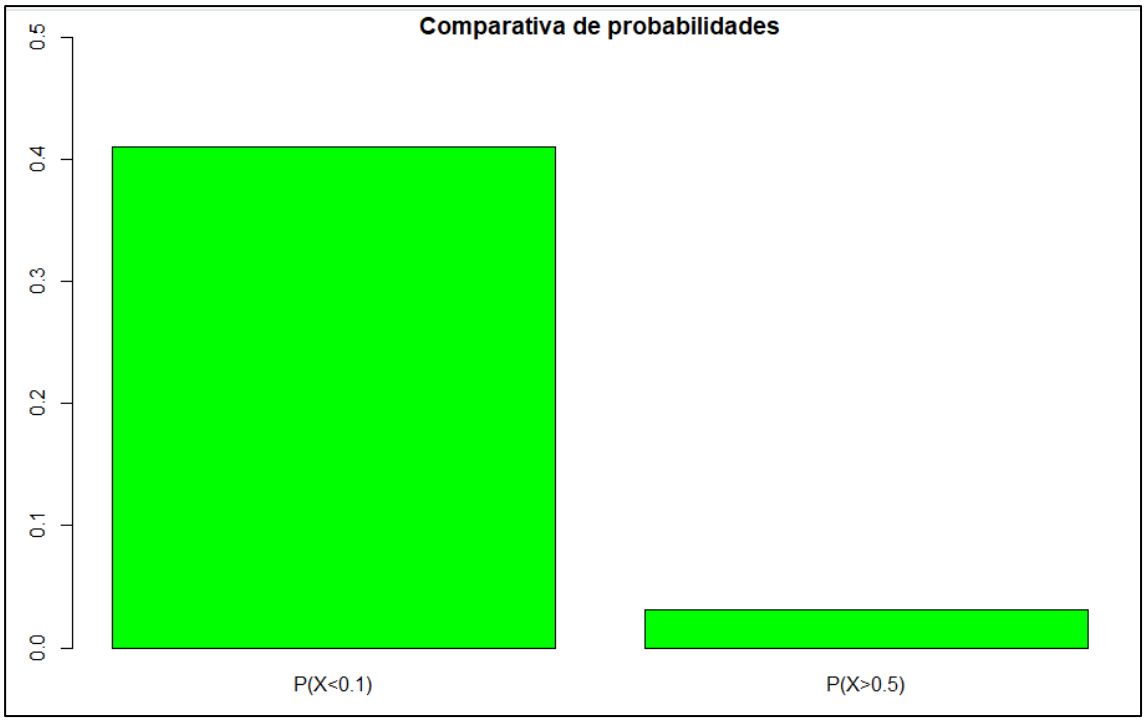
c) ¿Cuál es la probabilidad de que una empresa elegida al azar gaste más del 50% de su presupuesto en controles medioambientales y de contaminación?

```
p_M_50 <- function(x) 5*(1-x)^4
integral_50 <- integrate(p_M_50, lower = 0.5, upper = 1)
integral_50
03125 with absolute error < 3.5e-16
```

De manera parecida, para este apartado calculamos la integral definida de la función entre 0.5 y 1. Ello nos da una probabilidad que es 0.03125.

d) Visualizar gráficamente los apartados b) y c).

```
p <- c("P(X<0.1)","P(X>0.5)")
integrales <- c(0.40951, 0.03125)
barplot(integrales,width = 1,space = NULL,names.arg = p, col = "green", main =
          "Comparativa de probabilidades", ylim = c(0,0.5))
```



LECTURA 4: CUESTIÓN 3

De acuerdo con un estudio sociológico realizado por investigadores de la ULPGC, aproximadamente un 45% de los consumidores de tranquilizantes en la provincia de Las Palmas empezaron a consumirlos por problemas psicológicos. Calcular la probabilidad de que entre los siguientes 10 consumidores entrevistados de la provincia de Las Palmas:

Se trata de una distribución binomial con éxito del 0.45% y de tamaño 10.

a) Exactamente 4 comenzaron a consumir tranquilizantes por problemas psicológicos.

```
> p_4 <- dbinom(4,10,0.45)
> p_4
[1] 0.2383666
```

La probabilidad de que exactamente 4 comenzaran a consumir por problemas psicológicos es 0.2384.

b) Al menos 6 comenzaron a consumir tranquilizantes por problemas psicológicos.

Para calcular este apartado, primero debemos calcular todas las probabilidades anteriores a 6. Luego, para calcular la probabilidad que nos piden, debemos de resultar la suma de todos los anteriores con uno.

```
> p_0 <- dbinom(0,10,0.45)
> p_1 <- dbinom(1,10,0.45)
> p_2 <- dbinom(2,10,0.45)
> p_3 <- dbinom(3,10,0.45)
> p_4 <- dbinom(4,10,0.45)
> p_5 <- dbinom(5,10,0.45)
> F_6 <- p_0 + p_1 + p_2 + p_3 + p_4 + p_5
> p_6 <- 1-F_6
> p_6
[1] 0.2615627
```

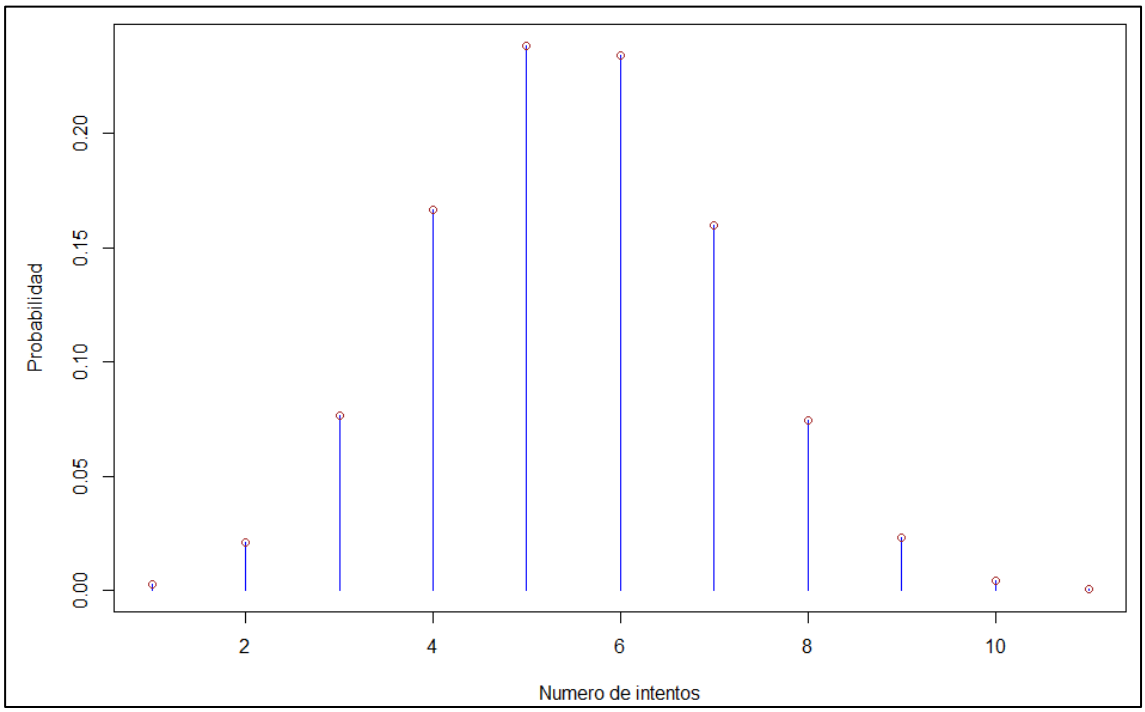
Ello nos da una probabilidad que es 0.2516.

c) Analizar la distribución de probabilidad subyacente y sus características principales.

Observamos que se trata de una distribución discreta de probabilidad. ¿Por qué? Porque f(x) = 0, y la sumatoria de todas las funciones es igual a uno. Además, P(X=x) es f(x). Más concretamente resulta ser una distribución binomial, puesto que podemos ver el proceso de Bernoulli. Este se caracteriza por diferentes aspectos, tales como que el experimento (cada caso) consta de ensayos (pruebas) repetidos, que cada ensayo produce un resultado que se puede clasificar como éxito o fracaso, como que la probabilidad de un éxito permanece constante de una prueba a otra, y que las pruebas repetidas son independientes.

d) Visualizar la función de densidad.

```
dist_bin <- dbinom(0:10,10,0.45, log = F)
plot(dist_bin, xlab = "Numero de intentos", ylab = "Probabilidad", col = "blue", type = "h")
points(1:11, dist_bin, col = "brown")
```



LECTURA 4: CUESTIÓN 4

El número de clientes que llega al departamento de reclamaciones de "MediaMark" en el centro comercial Las Terrazas es de 5 cada veinte minutos. Establecer un modelo de la posible distribución de probabilidad y explicar sus características. Así mismo, con este modelo:

Se trata de un proceso de Poisson, siendo 5 el número de clientes cada 20 minutos.

a) Calcular la probabilidad de que lleguen más de 10 clientes en un período de una hora.

```
> p_10 <- dpois(10,15)
> p_10
[1] 0.04861075
```

El resultado es una probabilidad correspondiente a 0.0486.

b) Calcular la probabilidad de que en veinte minutos lleguen menos de 5 clientes.

```
> p_x_m_5 <- dpois(0:4,5)
> p_sum <- sum(p_x_m_5)
> p_sum
[1] 0.4404933
```

En este caso se realiza una sumatoria con las probabilidades de 0 hasta 4. Ello nos da una probabilidad correspondiente a 0.4405.

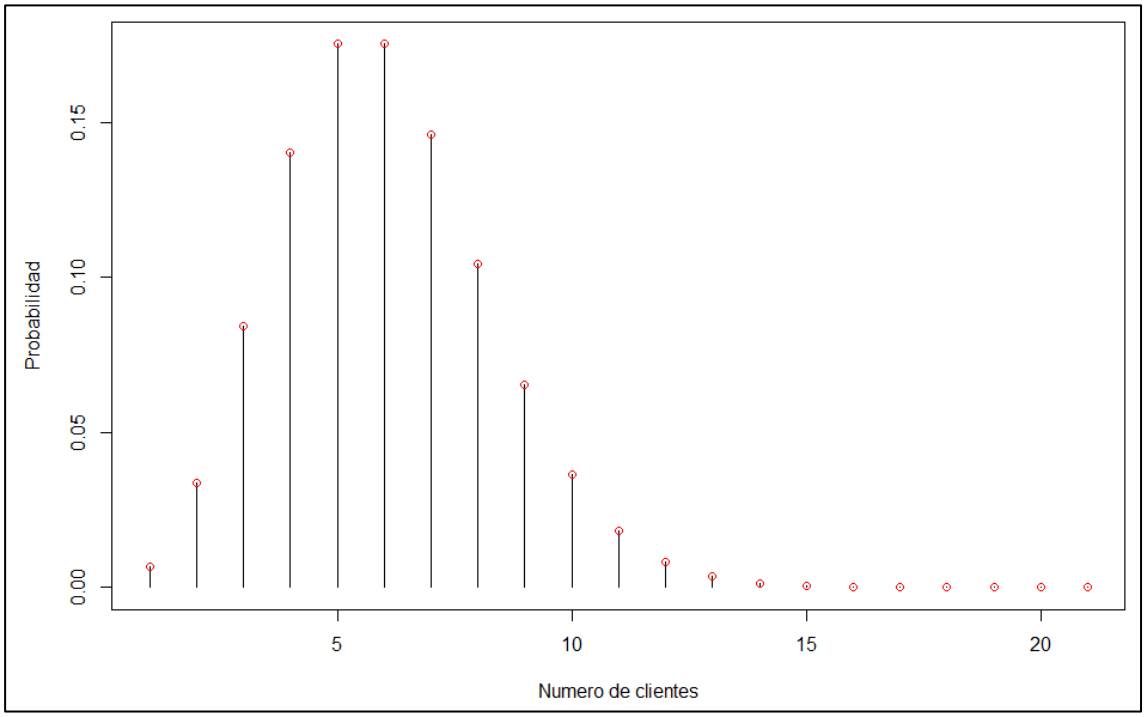
c) Cual es el número medio de llegadas en un período de dos horas.

El número medio de llegadas será la multiplicación entre el número de veces que transcurren 20 minutos en un período de dos horas. Ello es entonces 6*5 = 30.

El número medio de llegadas de clientes en dos horas será de 30 clientes.

d) Mostrar gráficamente las funciones de distribución de probabilidad correspondientes.

```
x <- dpois(0:20,5)
plot(x, col = "black", xlab = "Numero de clientes", ylab = "Probabilidad", type = "h")
points(1:21,x,col="red")
```



LECTURA 4: CUESTIÓN 5

Se sabe como resultado de análisis previos que el 3.5% de las personas que se les revisa el equipaje en el aeropuerto de Gran Canaria llevan objetos cuestionables.

Se trata de una distribución geométrica.

a) ¿Cuál es la probabilidad de que en una serie de 15 personas cruce sin problemas antes de encontrar a una que tenga un objeto no permitido para embarcar con él?

```
> r <- dgeom(15, 0.035)
> r
[1] 0.02051057
```

Es una probabilidad correspondiente a 0.0205.

b) ¿Cuál es el número esperado de personas que pasarán normalmente hasta que se pare en una por tener un objeto de estas características?

```
> num_personas <- (1-0.035)/0.035^2
> num_personas
[1] 787.7551
```

Pasarán aproximadamente 788 personas antes de que se pare a una por tener objetos cuestionables.

c) Razonar sobre la distribución de probabilidad subyacente y explicar su uso y características más significativas.

Contamos con una distribución geométrica porque al calcular P(X) la X es el número de ensayos hasta obtener el primer éxito.

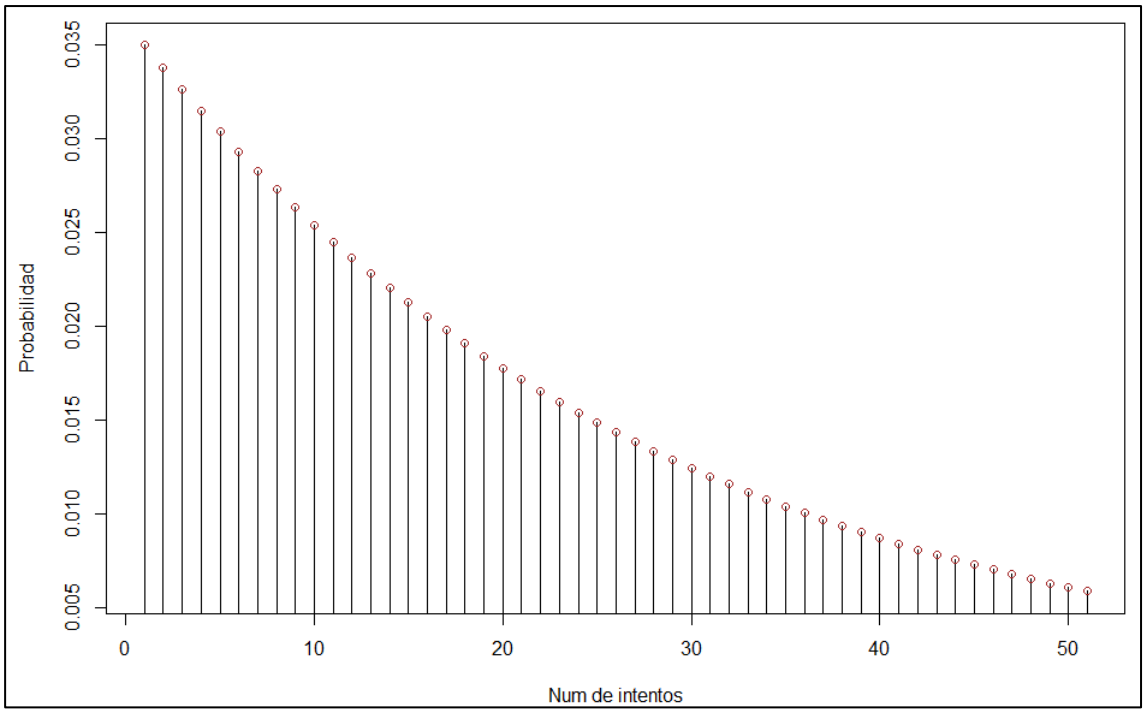
d) Si por cada caso de una persona sin problemas en el equipaje el tiempo medio es de 1 minuto y por cada caso de una persona con objetos no adecuados el tiempo medio se alarga en 5 minutos, analizar los tiempos probables medios de espera para un vuelo de 120 pasajeros.

```
> #d)
> p_c <- 120*0.035
> p_c
[1] 4.2
> p_nc <- 120-p_c
> p_nc
[1] 115.8
> tiempo <- (p_c*5)+p_nc
> tiempo
[1] 136.8
```

Habrían cerca de 4 personas con objetos cuestionables, 116 sin, y nos tomaría 137 minutos de espera.

e) Mostrar gráficamente las funciones de distribución de probabilidad correspondientes y visualizar explícitamente en caso del apartado d).

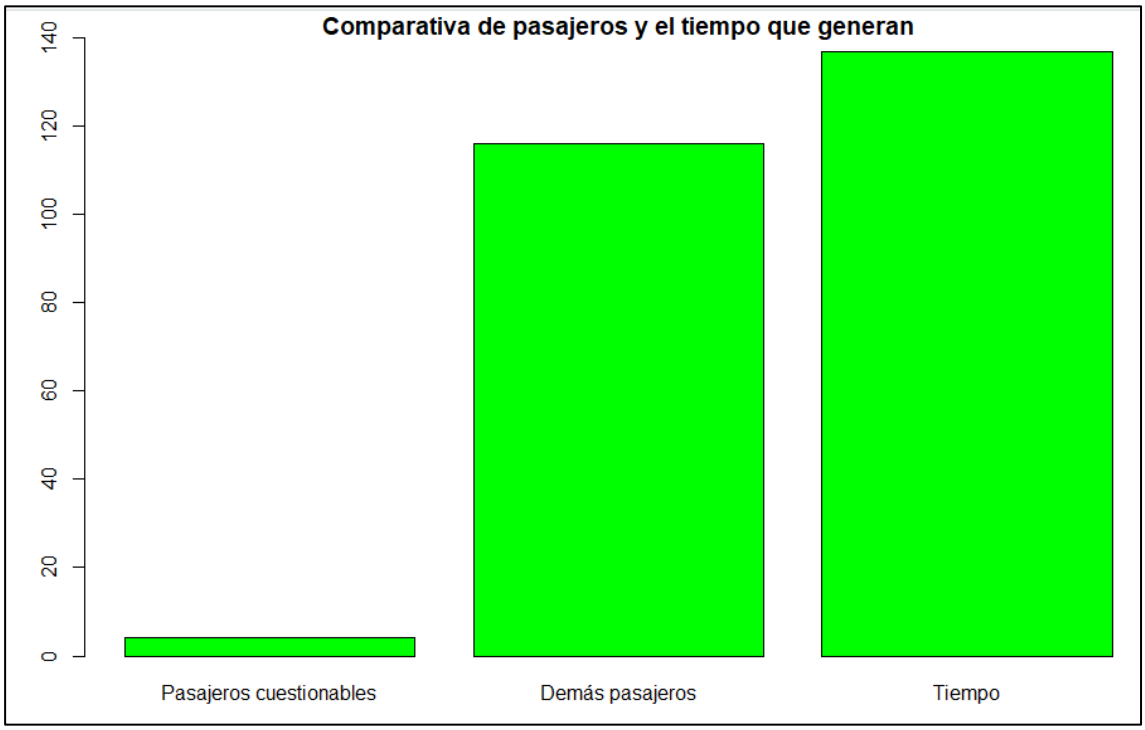
```
> dist_geom <- dgeom(0:50, 0.035, log = F)
> plot(dist_geom, xlab = "Num de intentos", ylab = "Probabilidad", col = "black", type = "h")
> points(1:51, dist_geom, col = "brown")
```



Vemos que es la gráfica correspondiente a una distribución geométrica.

A continuación procedemos a mostrar las gráficas del apartado d):

```
> nombres <- c("Pasajeros cuestionables", "Demás pasajeros", "Tiempo")
> d <- c(p_c, p_nc, tiempo)
> barplot(d, width = 1, space = NULL, names.arg = nombres, col = "green", main =
+           "Comparativa de pasajeros y el tiempo que generan", ylim = c(0,140))
```



LECTURA 5: CUESTIÓN 1

La estatura de los 835 estudiantes de la Escuela de Ingeniería Informática se distribuye según una normal de media de 176.5 centímetros y una desviación estándar de 7.1 centímetros. Encontrar cuántos de estos estudiantes se esperaría que tuvieran una estatura:

```
u <- 176.5
o <- 7.1
```

a) Menor que 160 centímetros.

```
> p_160 <- (160-u)/o
> pnorm(p_160)
[1] 0.01006426
```

Existirá una probabilidad de 0.010.

b) Entre 171.5 y 180 centímetros.

```
> p_180 <- (180-u)/o
> p_171.5 <- (171.5-u)/o
> pnorm(p_180)-pnorm(p_171.5)
[1] 0.4483326
```

Tiene una probabilidad de 0.4483.

c) Igual a 175 centímetros.

```
> p_175 <- (175-u)/o
> p_175
[1] -0.2112676
```

Probabilidad nula.

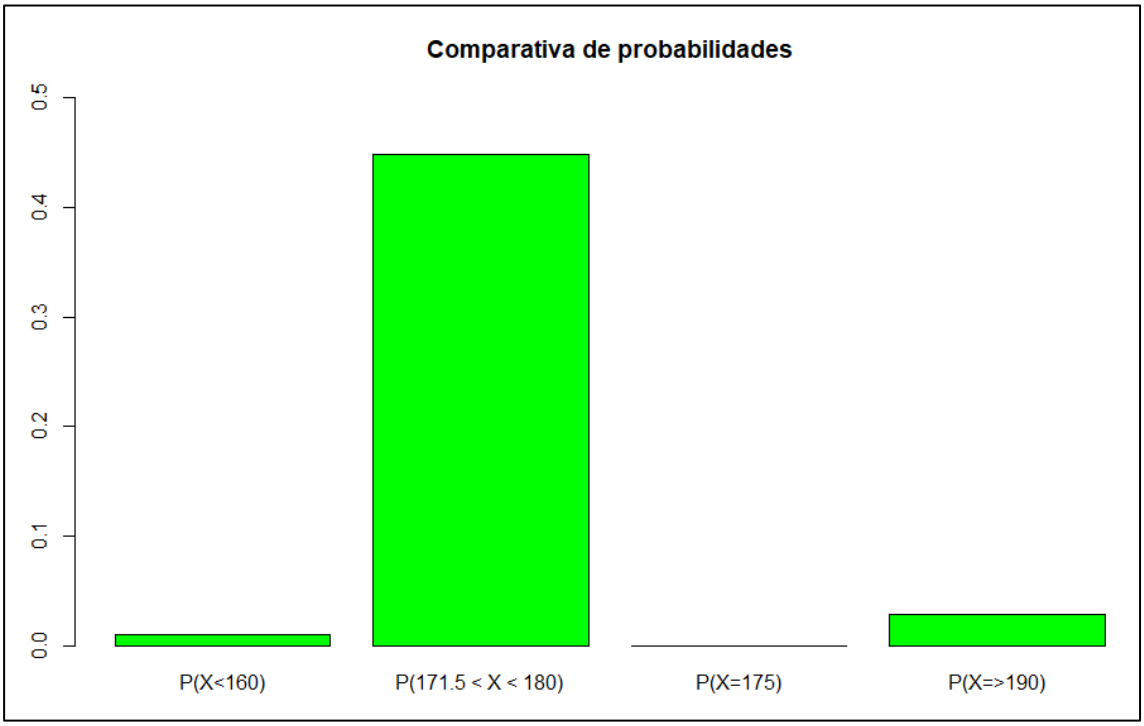
d) Mayor o igual a 190 centímetros.

```
> p_190 <- (190-u)/o
> 1-pnorm(p_190)
[1] 0.02862427
```

Existe una probabilidad de 0.0286.

e) Visualizar la distribución y las probabilidades de los grupos de estatura resultantes de los apartados anteriores.

```
> p <- c("P(X<160)","P(171.5 < X < 180)","P(X=175)","P(X=>190)")
> r <- c(0.01006426, 0.4483326, 0, 0.02862427)
> barplot(r,width=1,space=NULL,names.arg=p,col="green",main="Comparativa de probabilidades",ylim=c(0,0.5))
```



LECTURA 5: CUESTIÓN 4

Una empresa de distribución y logística de las Islas Canarias tiene una máquina especial para el empaquetado de artículos calificados como frágiles. Si un artículo se coloca de forma incorrecta en la máquina no se podría extraer su contenido e incluso se podría dañar. En este caso se dice que "falló la máquina".

a) Si la probabilidad de que falle la máquina es de 0.05. ¿Cuál es la probabilidad de que ocurra más de un fallo en un lote de 35 paquetes?

```
> p_1 <- 1-pbinom(1,35,0.05)
> p_1
[1] 0.5279735
```

La probabilidad será 0.52797.

b) Si la probabilidad de que falle la máquina es de 0.05 y se empaqueta un lote de 500 artículos. ¿Cuál es la probabilidad de que ocurran más de 10 fallos?

```
> p_10 <- 1-pbinom(10,500,0.05)
> p_10
[1] 0.99954
```

Existe una probabilidad muy alta, tan cercana como que se podría decir un 100%.

c) Analizar la distribución de probabilidad elegida para este caso, justificar su uso, y visualizar los resultados.

Se trata de una distribución de probabilidad binomial. Esta se caracteriza por tener un número de éxitos en una secuencia de n ensayos de Bernoulli independientes entre sí, con una probabilidad fija p de ocurrencia de éxito entre los ensayos.

LECTURA 6: CUESTIÓN 1

Considérese una población normal, con varianza desconocida, que tiene una media de 20.5.

a) ¿Es posible obtener una muestra aleatoria de tamaño 8 de esta población con una media de 23.75 y una desviación estándar de 4.0?

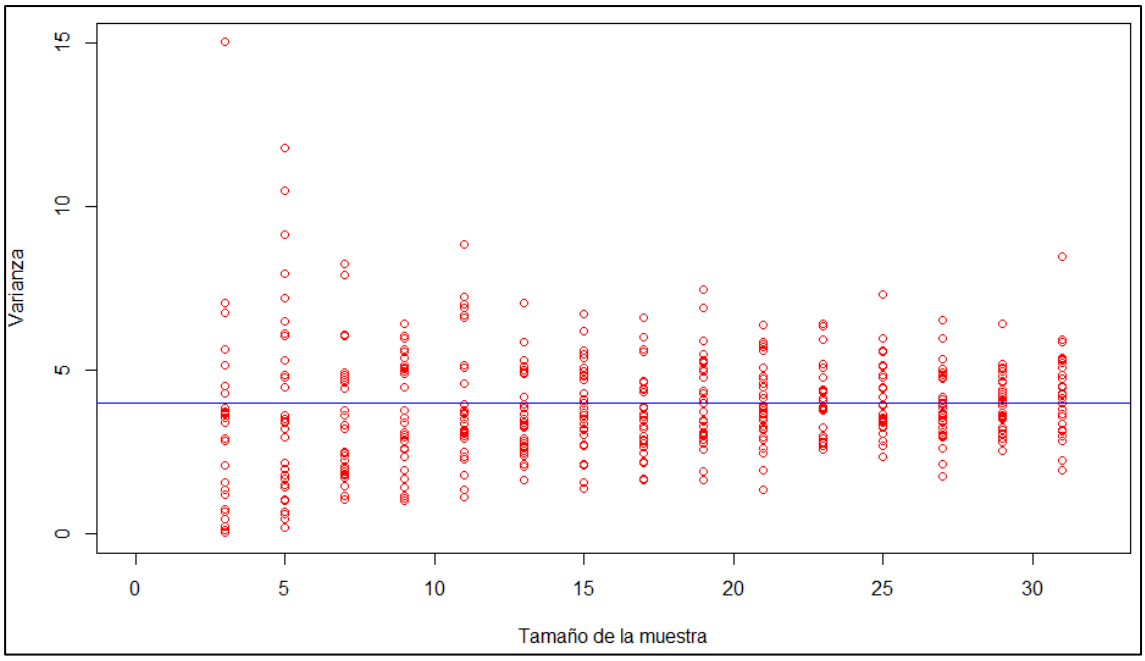
No es posible. Se trata de una distribución normal, por lo que la media de la muestra debe ser la misma que la de la población.

b) Si no fuera posible, ¿a qué conclusión llegaría?

A que cualquier muestra de longitud n tomada de la población tendrá la misma media que dicha población independientemente de la varianza.

c) Razonar sobre el tamaño de la muestra y su relación sobre el posible intervalo de confianza para la media de la misma.

```
plot(c(0,32),c(0,15),type="n",xlab="Tamaño de la muestra",ylab="Varianza")
for(n in seq(3,31,2)){
  for(i in 1:30){
    x <- rnorm(n,mean=10,sd=2)
    points(n,var(x),col="red")
  }
}
abline(a=4,b=0,col="blue")
```



LECTURA 6: CUESTIÓN 2

Las calificaciones de un examen de Métodos Estadísticos durante los últimos cinco años tienen aproximadamente una distribución normal de media μ = 7.45 y una varianza de σ^2 = 0.8. ¿Se seguiría considerando que σ^2 = 0.8 es un valor válido de la varianza si una muestra aleatoria de 20 estudiantes que se examinan obtiene un valor de s^2 = 20? Razonar y justificar teóricamente la respuesta.

Las distribuciones muestrales vienen determinadas por la siguiente fórmula:

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}$$

Por tanto, debemos de tener en cuenta los valores de la variable aleatoria x2, luego si estos valores son mayores que algunos valores específicos tendremos que calcular el valor de la varianza.

LECTURA 6: CUESTIÓN 3

La empresa "Tirma" ha puesto en marcha un proceso en el que se utiliza una máquina para llenar envases de cartón con batido de chocolate. La especificación que es estrictamente indispensable para el llenado de la máquina es 900 ± 150 gramos. El proveedor considera que cualquier envase de cartón que no cumpla con tales límites de

peso en el llenado está defectuoso. Se espera que al menos 99% de los envases de cartón cumplan con la especificación. En el caso de que μ = 900 y σ = 100,

a) ¿Qué proporción de envases de cartón del proceso están defectuosos?

```
> x <- rnorm(100,900,100)
> ux <- mean(x)
> ox <- sd(x)
> nx <- length(x)
> t.test(x,alternative="two.sided",mu=900,conf.level=0.99)

        One Sample t-test

data:  x
t = -1.3297, df = 99, p-value = 0.1867
alternative hypothesis: true mean is not equal to 900
99 percent confidence interval:
 863.1227 912.0878
sample estimates:
mean of x
 887.6053
```

Hay una proporción de envases defectuosos de un 1%.

b) Si se hacen cambios para reducir la variabilidad, ¿cuánto se tiene que reducir σ para que haya 0.99 de probabilidades de cumplir con la especificación?

```
> y <- rnorm(100,900,50)
> uy <- mean(y)
> oy <- sd(y)
> ny <- length(y)
> t.test(x,alternative="two.sided",mu=900, conf.level= 0.99)

        One Sample t-test

data:  x
t = -1.3297, df = 99, p-value = 0.1867
alternative hypothesis: true mean is not equal to 900
99 percent confidence interval:
 863.1227 912.0878
sample estimates:
mean of x
 887.6053
```

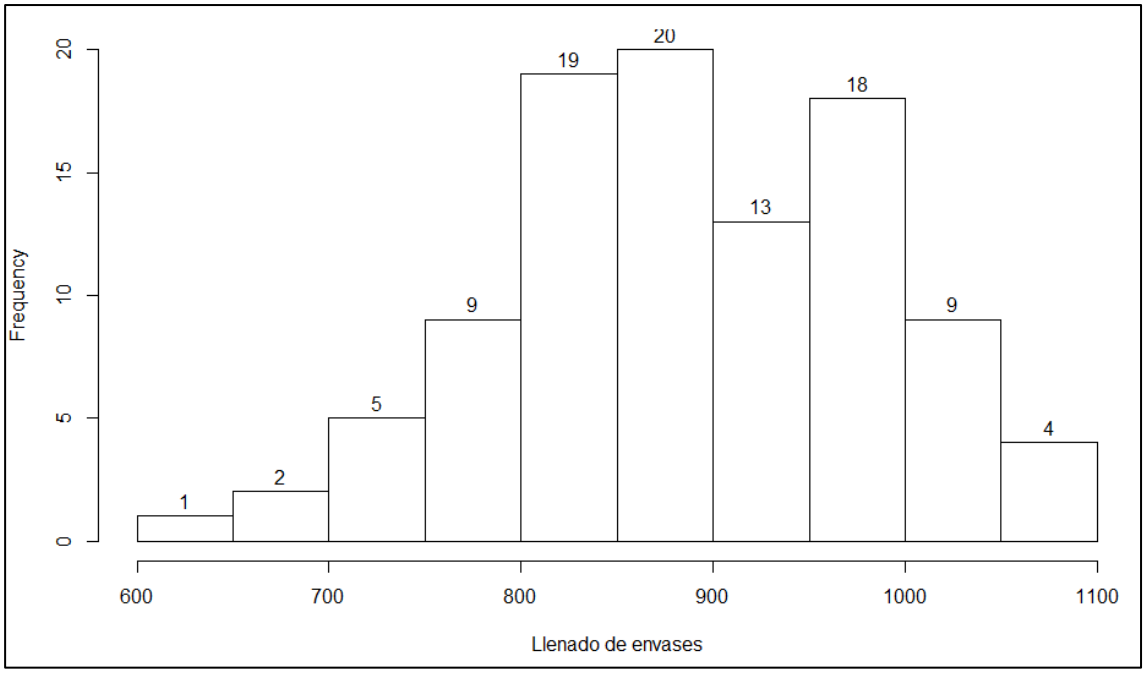
Si la variación es menor que la especificación de 900+-150, se tendría que usar cualquier tipo de variación y seguiría teniendo un 99% de probabilidad de cumplir con la especificación.

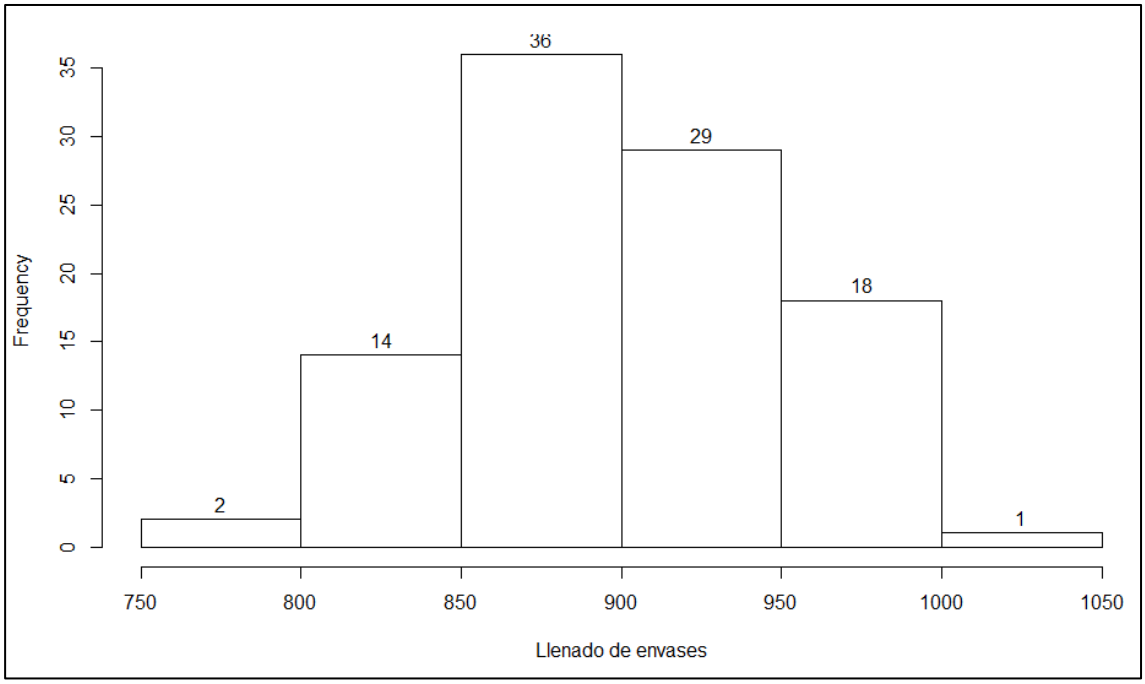
c) ¿Cuál es el tamaño de muestra para que en este segundo caso se garanticen las especificaciones?

Cualquiera superior a 30 para evitar la hipótesis de normalidad de datos.

d) Visualizar gráficamente los casos a) y b).

Supóngase una distribución normal para el peso.

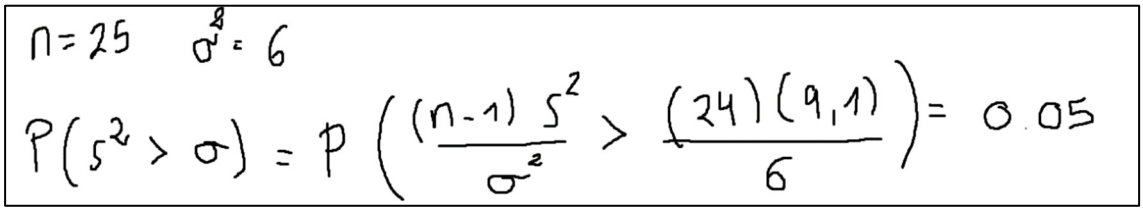




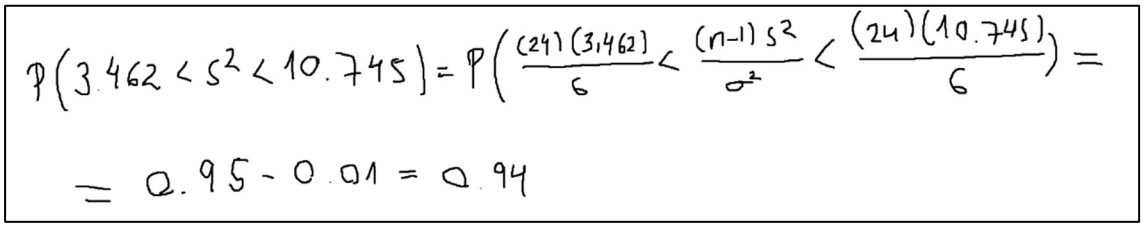
LECTURA 6: CUESTIÓN 4

Supóngase que las varianzas muestrales son mediciones continuas. Calcular la probabilidad de que una muestra aleatoria de 25 observaciones, de una población normal con varianza $\sigma^2 = 6$, tenga una varianza muestral s^2

a) Mayor que 9.1.



b) Comprendida entre 3.462 y 10.745.



LECTURA 7: CUESTIÓN 1

Una empresa de material eléctrico del polígono industrial de Arinaga fabrica para el mercado europeo bombillas que tienen una duración distribuida de forma aproximadamente normal, con una desviación estándar de 40 horas. Si una muestra de 30 bombillas tiene una duración promedio de 780 horas, se pide:

a) Calcular un intervalo de confianza del 96% para la media de la población de todas las bombillas producidas por esta empresa.

```
> linf <- 780-qt(0.96, df=(29))*40/sqrt(30)
> lsup <- 780+qt(0.96, df=(29))*40/sqrt(30)
> cat(c("lim.Inf (96%) =", as.character(round(linf,3)), " lim.Sup.(96%) =", as.character(round(lsup,3))))
lim.Inf (96%) = 766.751  lim.Sup.(96%) = 793.249
```

El intervalo de confianza al 96% es [766,793].

b) ¿A qué conclusiones se llegan a partir de la información suministrada por muestra? Razonar la respuesta y justificar teóricamente la misma.

Como la distribución es normal, deducimos basándonos en la teoría sobre dicho tipo de distribuciones que la media de la muestra y la de la población son la misma, así como ocurre con la desviación.

c) ¿Cuál sería el tamaño de la muestra para garantizar en un 99% la duración promedio resultante?

Para poder garantizar un 99% la duración promedio, la muestra debería de ser n=32.

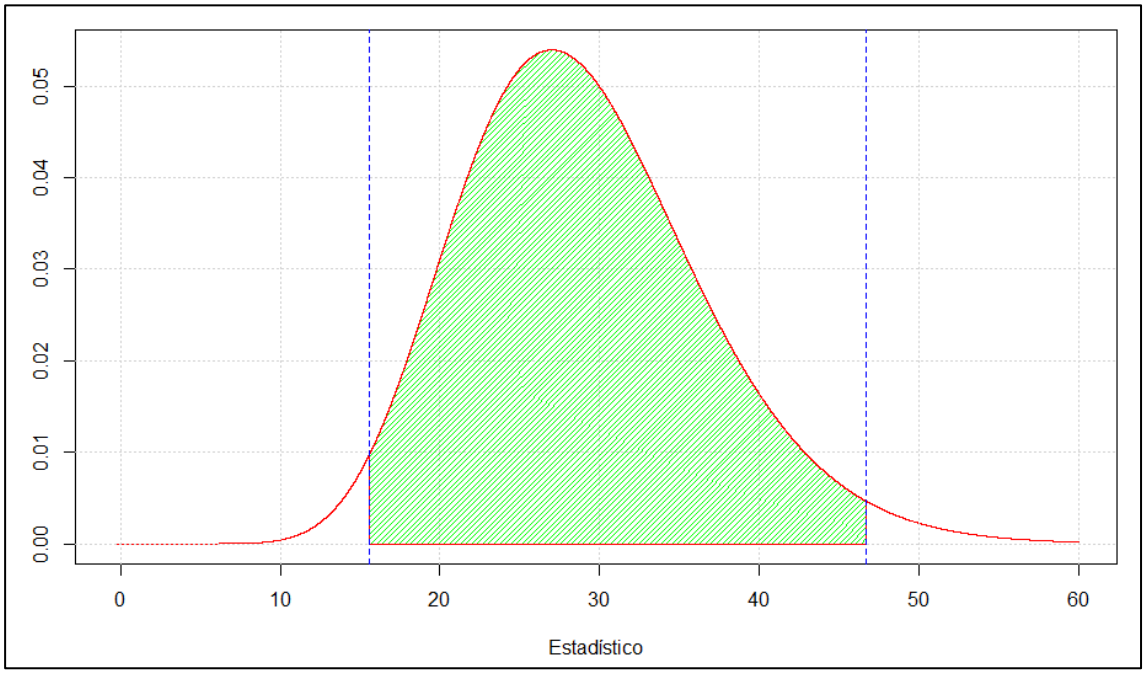
d) ¿Se podría con estos datos calcular un intervalo de tolerancia del 99%?

No, porque estas muestras poseen una distribución de tipo normal, la que suele tener una desviación mayor una vez se le pasan las pruebas y por tanto

necesitaríamos una muestra mayor, la descrita en el anterior apartado, para garantizar ese 99%.

e) Mostrar gráficamente los intervalos para las hipótesis establecidas y visualizar las conclusiones.

```
u <- 780
o <- 40
n <- 30
x <- rnorm(n,u,o)
qchisq(0.02,n-1)
qchisq(0.96,n-1)

pt <- seq(-0.4,60,0.001)
alpha <- 0.04
s2 <- 0.376
dp <- dchisq(pt, n-1)
plot(pt,dp,type="l",col="red",ylab="Densidad de probabilidad",xlab="Estadístico")
xliminf <- (n-1)*s2/qchisq((1-alpha/2),n-1)
xlimsup <- (n-1)*s2/qchisq(alpha/2,n-1)
chixliminf <- (n-1)*s2/xlimsup
chixlimsup <- (n-1)*s2/xliminf
grid()
xv <- pt[pt>=chixliminf&pt<=chixlimsup]
yv <- dp[pt>=chixliminf&pt<=chixlimsup]
xv <- c(xv,chixlimsup,chixliminf)
yv <- c(yv,dp[1],dp[1])
polygon(xv,yv,col="green",density=30,border="red")+abline(v=chixlimsup,col="blue",lty=2)+
  abline(v=chixliminf,col="blue",lty=2)
```



LECTURA 7: CUESTIÓN 2

Una máquina para un taller de la zona industrial del Cebadal produce piezas metálicas de forma cilíndrica para aparatos de aire acondicionados. Se toma una muestra de las piezas y los diámetros de las mismas son 1.01, 0.97, 1.03, 1.04, 0.99, 0.98, 0.99, 1.01, y 1.03 centímetros.

a) Calcular un intervalo de confianza del 99% para la media del diámetro de las piezas que se manufacturan con esta máquina, establézcase las acotaciones necesarias y razónense las respuestas.

```
> muestra <- c(1.01, 0.97, 1.03, 1.04, 0.99, 0.98, 0.99, 1.01, 1.03)
> n <- length(muestra)
> sigma <- sd(muestra)
> u <- mean(muestra)
> t.test(x = muestra,y = NULL,alternative=c("two.sided"),mu=u,paired=F,var.equal=F,conf.level= 0.99)

        One Sample t-test

data:  muestra
t = 0, df = 8, p-value = 1
alternative hypothesis: true mean is not equal to 1.005556
99 percent confidence interval:
 0.9780956 1.0330155
sample estimates:
mean of x
 1.005556
```

b) ¿Se podría realizar alguna inferencia sobre la varianza poblacional?

```
> o <- sd(muestra)
> o
[1] 0.02455153
```

La varianza es tan mínima que la diferencia entre una pieza y otra es inapreciable.

LECTURA 7: CUESTIÓN 3

Para un control rutinario de la Consejería de Sanidad se ha tomado una muestra aleatoria de 25 tabletas de aspirina con antiácido de una cierta marca, y se ha comprobado que contiene, en promedio, 325.05 mg de aspirina en cada tableta, con una desviación estándar de 0.5 mg. Calcular los límites de tolerancia del 95% que contendrán el 90% del contenido de aspirina para esta marca. Justificar teóricamente la respuesta.

```
> n <- 25
> u <- 325.05
> o <- 0.5
> porc <- (90*u)/100
> liminf <- porc-qt(0.975,df=(n-1))*o/sqrt(n)
> limsup <- porc+qt(0.975,df=(n-1))*o/sqrt(n)
> cat(c("lim.inf (95%) =",as.character(round(liminf,3)), "lim.sup (95%) =", as.character(round(limsup,3))))
lim.inf (95%) = 292.339 lim.sup (95%) = 292.751
```

Este resultado sería tolerable. La variación entre los límites de tolerancia calculados es de 0.4, es 0.1 menor a lo esperado.

LECTURA 7: CUESTIÓN 4

Se realiza un estudio para determinar si cierto tratamiento tiene algún efecto sobre la cantidad de metal que se elimina en una operación de encurtido. Una muestra aleatoria de 100 piezas se sumerge en un baño por 24 horas sin el tratamiento, lo que produce un promedio de 12.2 milímetros de metal eliminados y una desviación estándar muestral de 1.1 milímetros. Una segunda muestra de 200 piezas se somete al tratamiento, seguido de 24 horas de inmersión en el baño, lo que da como resultado una

eliminación promedio de 9.1 milímetros de metal, con una desviación estándar muestral de 0.9 milímetros.

a) Calcular un estimado del intervalo de confianza del 98% para la diferencia entre las medias de las poblaciones.

```
> n1 <- 100
> u1 <- 12.2
> o1 <- 1.1
> n2 <- 200
> u2 <- 9.1
> o2 <- 0.9
> x <- rnorm(n1,u1)
> y <- rnorm(n2,u2)
> t.test(x,y,alternative=c("two.sided"),mu=u1-u2,)

        Welch Two Sample t-test

data:  x and y
t = -0.049645, df = 201.64, p-value = 0.9605
alternative hypothesis: true difference in means is not equal to 3.1
95 percent confidence interval:
 2.852516 3.335328
sample estimates:
mean of x mean of y
12.151473  9.057551
```

El intervalo es [2.85,3.33].

b) Analizar según los datos si el tratamiento reduce o no la cantidad media del metal eliminado. Razonar adecuadamente la respuesta.

El tratamiento sí que reduce la cantidad de metal eliminado porque vemos como la media de la prueba con el tratamiento es menor que la de la prueba sin tratamiento. Además, la varianza de la segunda prueba también es menor, por tanto, la posibilidad de fallo de dicha varianza no sobrepasaría la media de la prueba sin tratamiento. Así podemos concluir que el tratamiento es efectivo.

LECTURA 7: CUESTIÓN 5

Una cooperativa de taxis de Las Palmas trata de decidir si comprará neumáticos de la marca A o de la marca B para su flotilla de taxis. Para estimar la diferencia entre las dos marcas realiza un experimento utilizando 12 neumáticos de cada marca, lo cual se utilizan hasta que se desgastan. Los resultados son:

	Media	Desviación Estándar
Marca A	**36300 kms.**	**5000 kms.**
Marca B	**38100 kms.**	**6100 kms.**

```
n <- 12
muA <- 36300
muB <- 38100
oA <- 5000
oB <- 6100
```

a) Calcular un intervalo de confianza del 95% para $\mu_A - \mu_B$, suponiendo que las poblaciones se distribuyen de forma aproximadamente normal.

```
> muestra <- uA-uB
> o <- oA-oB
> liminf <- muestra-qt(0.975, df=(n-1))*o/sqrt(n)
> limsup <- muestra+qt(0.975, df=(n-1))*o/sqrt(n)
> cat(c("lim.Inf. (95%) =", as.character(round(liminf,3)), " lim.Sup. (95%) =", as.character(round(limsup,3))))
lim.Inf. (95%) = -1345.942  lim.Sup. (95%) = -2254.058
```

Obtenemos un intervalo de confianza de [-1345,-2254].

b) Analice los resultados bajo las suposiciones de que las varianzas poblacionales sean o no iguales y explicar los mismos. Justificar las respuestas.

```
> x <- rnorm(n, uA)
> y <- rnorm(n, uB)
> t.test(x,y,alternative=c("two.sided"),mu=uA-uB,paired=T,var.equal=F,conf.level=0.95)

        Paired t-test

data:  x and y
t = 0.80706, df = 24, p-value = 0.4276
alternative hypothesis: true difference in means is not equal to -1800
95 percent confidence interval:
 -1800.373 -1799.149
sample estimates:
mean of the differences
              -1799.761

> t.test(x,y,alternative=c("two.sided"),mu=uA-uB,paired=T,var.equal=T,conf.level=0.95)

        Paired t-test

data:  x and y
t = 0.80706, df = 24, p-value = 0.4276
alternative hypothesis: true difference in means is not equal to -1800
95 percent confidence interval:
 -1800.373 -1799.149
sample estimates:
mean of the differences
              -1799.761
```

Como observamos, no hay diferencia, sin embargo, teniendo ambos tests en cuenta, podemos concluir que la marca A es mejor que la B, debido a que la desviación en las medidas de las medias de la marca B nos deja ver que a la larga los neumáticos de la marca A son más duraderos.

LAB 3: EJERCICIO 1

Leer el Data Frame que se encuentra en el fichero "Empleo.txt". El fichero contiene datos de un estudio sobre la duración media en semanas de los contratos de empleo de la Unión Europea. Con los datos en él incluidos.

```
empleo <- read.table("Empleo.txt", header = T, dec = ".", sep = ",")
```

a) Ordenar alfabéticamente por países el data frame.

Para ordenarlo alfabéticamente primero tenemos que poder manejar los datos dentro de la tabla. Ello lo conseguimos con la función <attach()>.

```
attach(empleo)
empleo_a <- empleo[order(Pais),]
empleo_a
```

Y con ello conseguimos la tabla siguiente, que, como podemos observar, se encuentra ordenada alfabéticamente.

	Pais	Duracion
1	Alemania	41.7
2	Austria	44.1
3	Belgica	41.0
4	Chipre	41.8
5	Dinamarca	40.5
23	Eslovaquia	41.6
24	Eslovenia	42.5
6	España	42.2
7	Estonia	41.5
8	Finladia	40.5
9	Francia	41.0
10	Grecia	44.1
18	Holanda	40.8
11	Hungria	41.0
12	Irlanda	40.7
13	Italia	41.3
14	Letonia	43.0
15	Lituania	39.8
16	Luxemburgo	40.9
17	Malta	41.2
19	Polonia	42.9
20	Portugal	41.6
22	Reino Unido	43.1
21	Rep.Checa	42.7
25	Suecia	41.1

b) Calcular la media, mediana y cuantiles de la duración del trabajo en semanas.

Como ya podemos manejar los datos que se refieren a la duración, simplemente es aplicar las diferentes funciones que calculan lo que se nos pide.

```
> mean(Duracion)
[1] 41.704
> median(Duracion)
[1] 41.5
> quantile(Duracion)
  0%  25%  50%  75% 100%
39.8 41.0 41.5 42.5 44.1
```

c) Evaluar los parámetros de dispersión de la duración.

Para ello calculamos la desviación estándar.

```
> sd(Duracion)
[1] 1.113358
```

d) Ordenar los países por las semanas de trabajo acumuladas en un año.

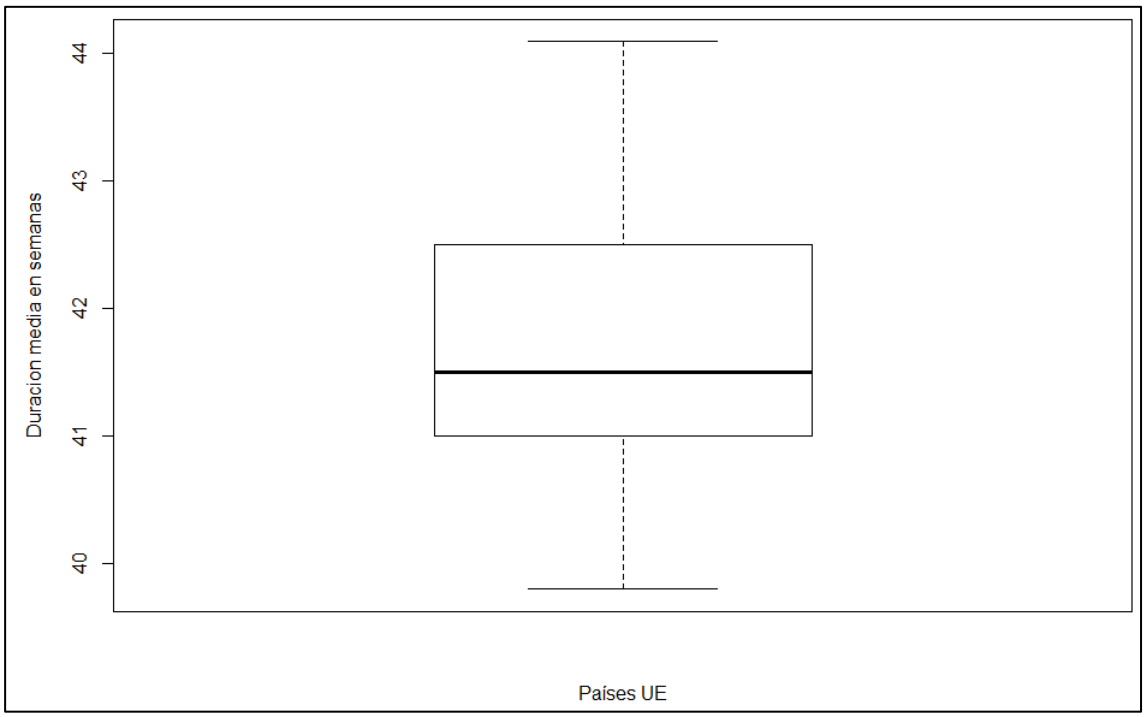
A diferencia del caso a), en vez de ordenarlos por países, los ordenamos por la duración.

```
empleo_b <- empleo[order(Duracion),]
empleo_b
```

	Pais	Duracion
15	Lituania	39.8
5	Dinamarca	40.5
8	Finladia	40.5
12	Irlanda	40.7
18	Holanda	40.8
16	Luxemburgo	40.9
3	Belgica	41.0
9	Francia	41.0
11	Hungria	41.0
25	Suecia	41.1
17	Malta	41.2
13	Italia	41.3
7	Estonia	41.5
20	Portugal	41.6
23	Eslovaquia	41.6
1	Alemania	41.7
4	Chipre	41.8
6	España	42.2
24	Eslovenia	42.5
21	Rep.Checa	42.7
19	Polonia	42.9
14	Letonia	43.0
22	Reino Unido	43.1
2	Austria	44.1
10	Grecia	44.1

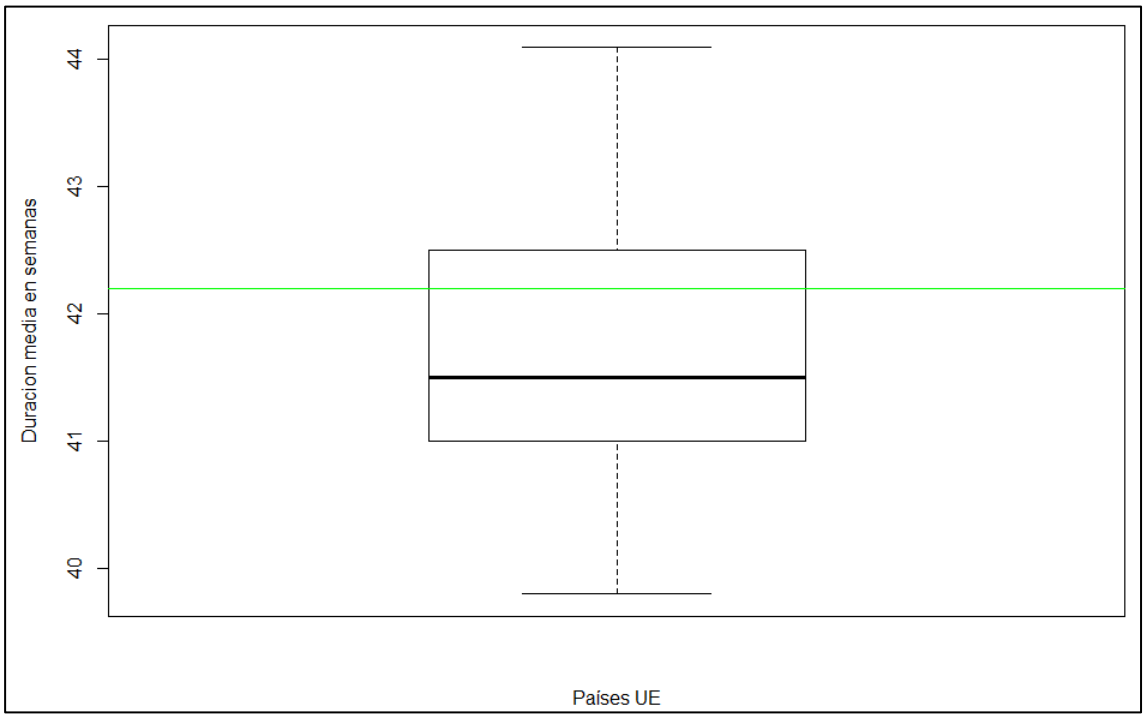
e) Visualizar las diferencias con un diagrama de caja y distinguir los valores singulares. Explicar los campos de datos resultados del uso de la función boxplot().

```
boxplot(Duracion, xlab = "Países UE", ylab = "Duracion media en semanas", col = "white")
```



f) Mostrar gráficamente la situación de España en e).

```
abline(h = Duracion[Pais == "España"], col = "green")
```

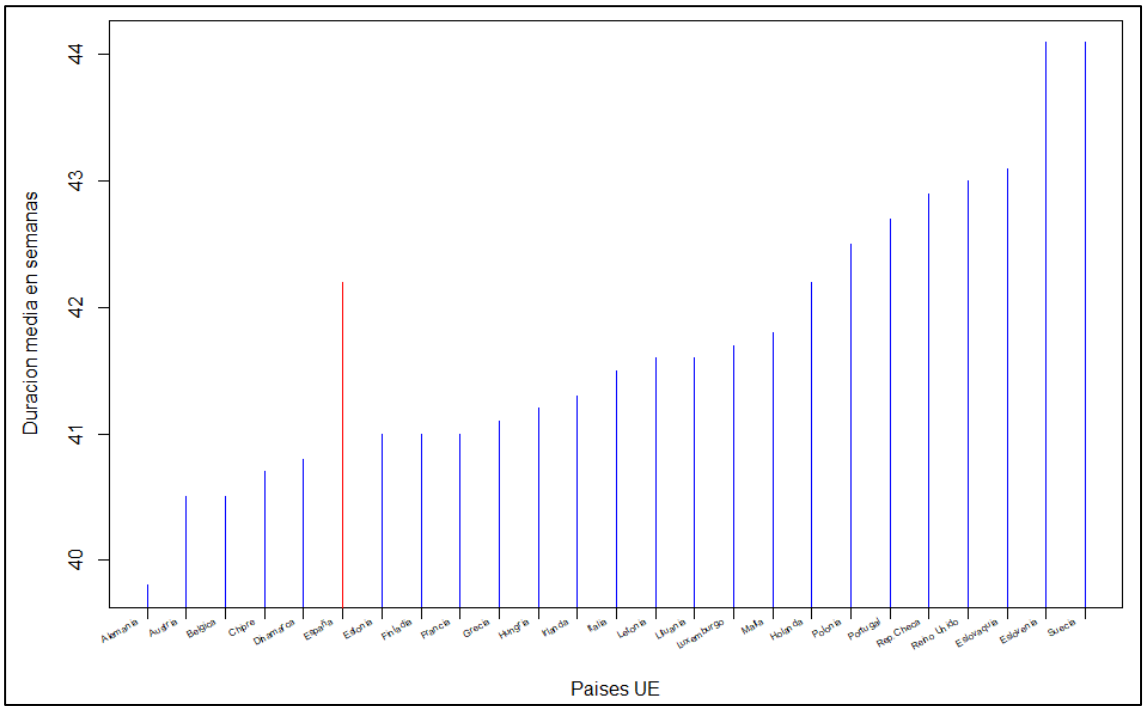


Observamos como claramente España se encuentra en la media alta de las duraciones de empleo.

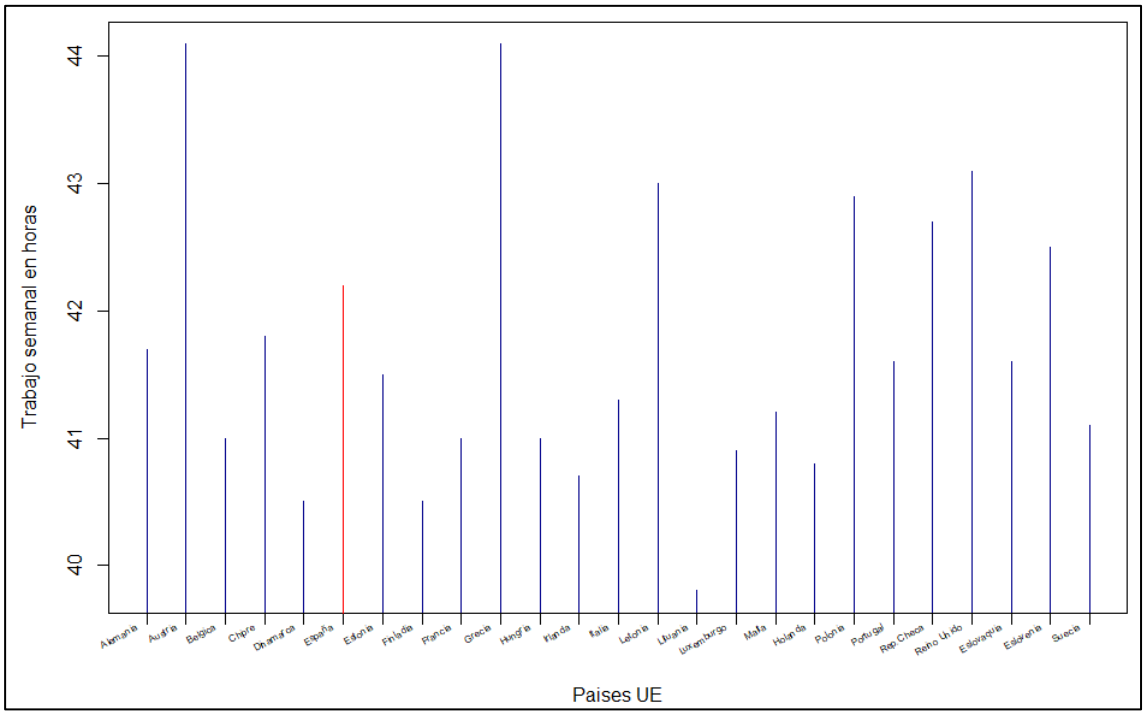
g) Visualizar gráficamente las variaciones entre países de la UE (ordenador por duración y sin ordenar) y señalar en el gráfico los valores que corresponde a España.

Primero los ordenaremos por la duración y haremos el gráfico.

```
par(mar = c(5,5,1,1) + 0.1)
plot(1: nlevels(Pais), empleo_b$Duracion, xaxt = "n", type = "h", col = "blue", xlab =
       "Paises UE", ylab = "Duracion media en semanas")
axis(side = 1, at = 1 : length(Pais), labels = F)
points(which(Pais == "España"), Duracion[Pais == "España"], type = "h", col = "red")
text(1: nlevels(Pais), par("usr")[3] - 0.1, labels = Pais, srt = 30, pos = 2, cex = 0.5, xpd = T)
```



Siendo la línea roja la perteneciente a España. Sin ordenar, el código se diferenciaría en que no usamos la tabla de empleo alfabetizada, y el gráfico nos quedaría como:



LAB 3: EJERCICIO 2

Leer el data frame que se encuentra en el fichero "Puromicina.txt". El fichero contiene datos de un estudio sobre la velocidad de reacción enzimática (en número de cuentas por minuto) en función de la concentración de sustrato (en partes por millón - ppm) en experimentos donde se trataba la enzima con puromicina ("treated") o no se trataba con esta ("untreated"). Se pide:

```
puromicina <- read.table("Puromicina.txt", sep = ",", dec = ".", header = T)
attach(puromicina)
```

a) Calcular las medias de la velocidad de reacción en función del empleo o no de puromicina.

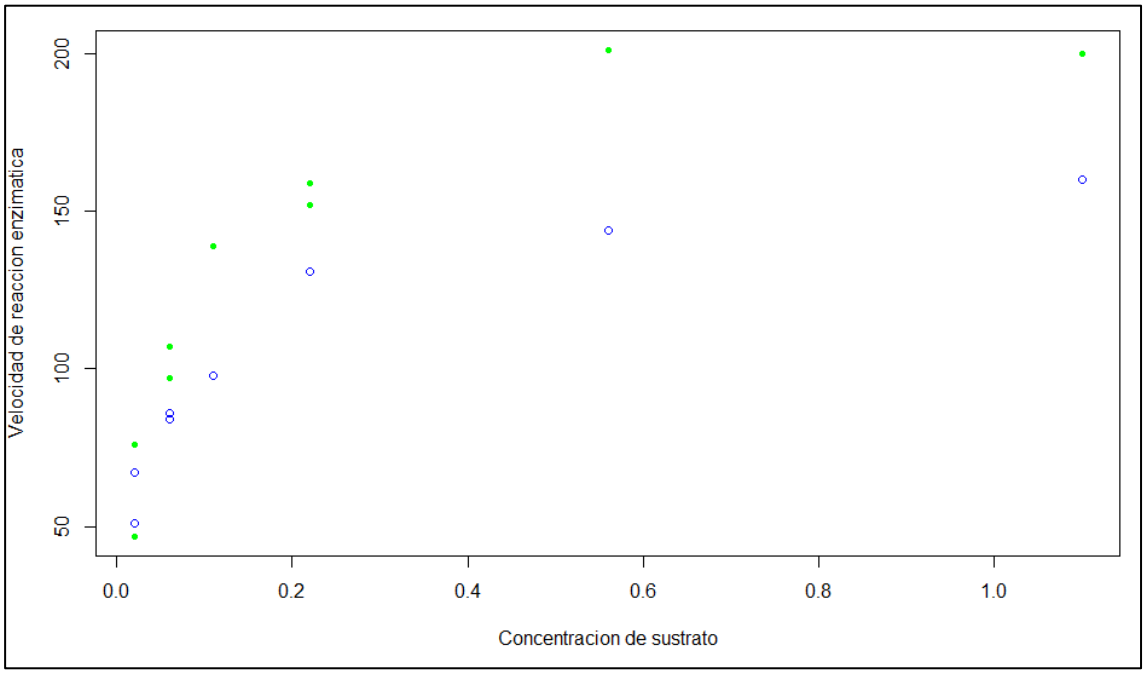
```
> media <- aggregate(velocidad_reaccion~Puromicina, puromicina, mean)
> media
  Puromicina velocidad_reaccion
1    treated           141.5833
2  untreated           110.7273
```

b) Evaluar los parámetros de dispersión de la velocidad de reacción.

```
> sd(velocidad_reaccion)
[1] 47.78475
```

c) Visualizar si la concentración del sustrato influye en la velocidad de reacción en los casos en que se trata o no con puromicina.

```
plot(concentracion[Puromicina=="treated"], velocidad_reaccion[Puromicina=="treated"], col="green",
     pch=20, xlab="Concentracion de sustrato", ylab="Velocidad de reaccion enzimatica")
points(concentracion[Puromicina=="untreated"], velocidad_reaccion[Puromicina=="untreated"], col="blue")
```



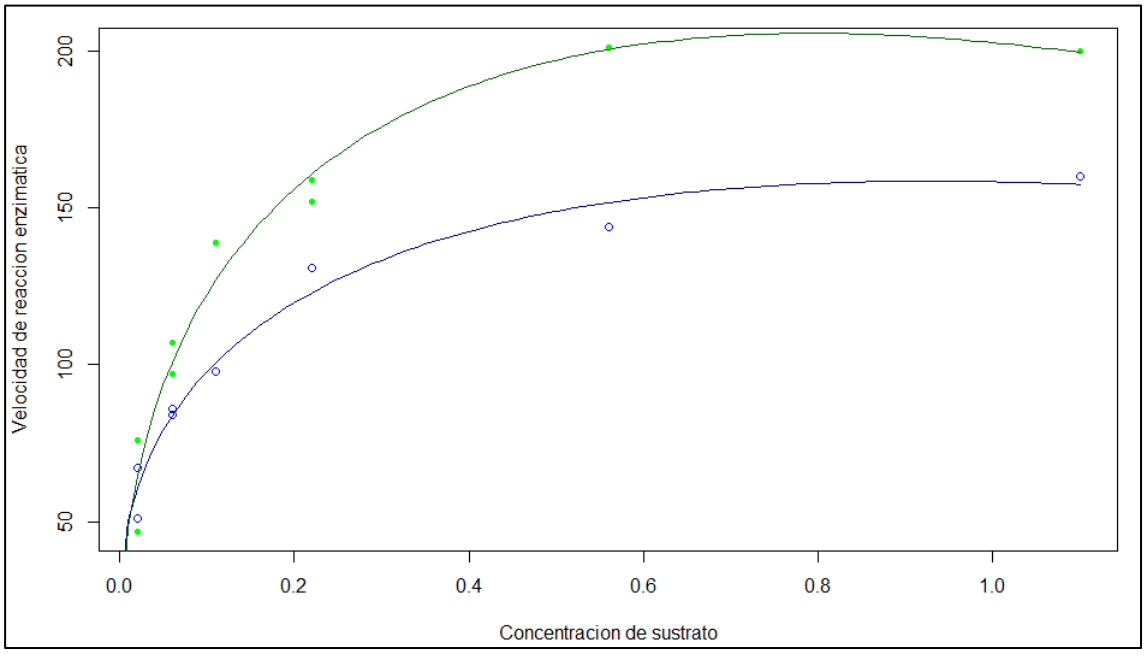
Siendo los puntos verdes donde se utiliza la puromicina, y los azules donde no.

d) Ordenar por velocidad de reacción.

```
> ord_vr <- puromicina[order(velocidad_reaccion),]
> ord_vr
   concentracion velocidad_reaccion Puromicina
2           0.02                 47    treated
11          1.10                207    treated
10          0.56                201    treated
1           0.02                 76    treated
12          1.10                200    treated
13          0.02                 67  untreated
3           0.06                 97    treated
14          0.02                 51  untreated
4           0.06                107    treated
15          0.06                 84  untreated
5           0.11                123    treated
16          0.06                 86  untreated
7           0.22                159    treated
6           0.11                139    treated
17          0.11                 98  untreated
9           0.56                191    treated
8           0.22                152    treated
>
```

e) Analizar los efectos del uso de la puromicina y de la concentración del sustrato en la velocidad de reacción. Realizar algunas predicciones.

```
> yt <- velocidad_reaccion[Puromicina=="treated"]
> xt <- concentracion[Puromicina=="treated"]
> model <- lm(yt~xt + I(xt^(1/2)))
> xv <- seq(0,1.1, 0.01)
> yv <- predict(model, list(xt = xv))
> lines(xv,yv,col="dark green")
> ynt <- velocidad_reaccion[Puromicina=="untreated"]
> xnt <- concentracion[Puromicina=="untreated"]
> model_n <- lm(ynt~xnt + I(xnt^(1/2)))
> ynv <- predict(model_n,list(xnt=xv))
> lines(xv,ynv,col="dark blue")
```



Como vemos, es un análisis acertado y una predicción posible.

f) Analizar el fichero "Puromicina_NA.txt" que contiene NAs y utilizar las funciones na.omit() o complete.cases() para evaluar el apartado a). Estudiar, y en su caso aplicar, las funciones de la librería DMwR2 para estos casos. ¿Qué conclusiones se pueden sacar? ¿Cómo afectaría el resultado si se sustituyen los NAs por ceros?

```
> library(DMwR2)
> puromicina_NA <- read.table("Puromicina_NA.txt", sep = ",", dec =".", header=T)
> puromicina_sinNA <- na.omit(puromicina_NA)
> attach(puromicina_sinNA)
The following objects are masked from puromicina_sinNA (pos = 3):

    concentracion, Puromicina, velocidad_reaccion

The following objects are masked from puromicina (pos = 5):

    concentracion, Puromicina, velocidad_reaccion

The following objects are masked from puromicina (pos = 6):

    concentracion, Puromicina, velocidad_reaccion

> media2 <- aggregate(velocidad_reaccion~Puromicina, puromicina_sinNA, mean)
> media2
  Puromicina velocidad_reaccion
1    treated           130.8889
2  untreated           102.6250
```

```
> CI <- centralImputation(puromicina_NA)
> CI
   concentracion velocidad_reaccion Puromicina
1          0.020                 76    treated
2          0.020                 47    treated
3          0.060                 97    treated
4          0.060                107    treated
5          0.165                123    treated
6          0.110                139    treated
7          0.220                159    treated
8          0.220                152    treated
9          0.560                191    treated
10         0.560                201    treated
11         1.100                119    treated
12         1.100                200    treated
13         0.020                 67  untreated
14         0.020                 51  untreated
15         0.060                 84  untreated
16         0.060                 86  untreated
17         0.110                 98  untreated
18         0.110                115    treated
19         0.220                131  untreated
20         0.220                119  untreated
21         0.560                144  untreated
22         0.560                119  untreated
23         1.100                160  untreated
```

```
> kI <- knnImputation(puromicina_NA)
> kI
   concentracion velocidad_reaccion Puromicina
1      0.0200000             76.0000    treated
2      0.0200000             47.0000    treated
3      0.0600000             97.0000    treated
4      0.0600000            107.0000    treated
5      0.1506064            123.0000    treated
6      0.1100000            139.0000    treated
7      0.2200000            159.0000    treated
8      0.2200000            152.0000    treated
9      0.5600000            191.0000    treated
10     0.5600000            201.0000    treated
11     1.1000000            175.9734    treated
12     1.1000000            200.0000    treated
13     0.0200000             67.0000  untreated
14     0.0200000             51.0000  untreated
15     0.0600000             84.0000  untreated
16     0.0600000             86.0000  untreated
17     0.1100000             98.0000  untreated
18     0.1100000            115.0000    treated
19     0.2200000            131.0000  untreated
20     0.2200000            106.0601  untreated
21     0.5600000            144.0000  untreated
22     0.5600000            126.6490  untreated
23     1.1000000            160.0000  untreated
```

```
> media_cI <- aggregate(cI$velocidad_reaccion~cI$Puromicina, cI, mean)
> media_cI
  cI$Puromicina cI$velocidad_reaccion
1       treated              132.7692
2     untreated              105.9000
> media_kI <- aggregate(kI$velocidad_reaccion~kI$Puromicina, kI, mean)
> media_kI
  kI$Puromicina kI$velocidad_reaccion
1       treated              137.1518
2     untreated              105.3709
```

LAB 3: EJERCICIO 3

El 2% de los equipos de un cierto fabricante de ordenadores tienen un fallo por mes de utilización y ningún ordenador tiene más de una avería por mes. El Departamento de Informática de la ULPGC decide adquirir 150 de estos equipos. Se pide:

a) Analizar el tipo de función de probabilidad subyacente y explicar sus características.

Se trata de una distribución de probabilidad binomial. Esta se caracteriza por tener un número de éxitos en una secuencia de n ensayos de Bernoulli independientes entre sí, con una probabilidad fija p de ocurrencia de éxito entre los ensayos.

```
n <- 150
p <- 0.02
```

b) Calcular la probabilidad de que el número de averías sea de 5.

```
> dbinom(5,n,p)
[1] 0.1011484
```

Existe una probabilidad de 0.1011.

c) Encontrar la probabilidad de que el número de averías sea mayor o igual a 3.

```
> p0 <- dbinom(0,n,p)
> p1 <- dbinom(1,n,p)
> p2 <- dbinom(2,n,p)
> 1-(p0+p1+p2)
[1] 0.5790743
```

La probabilidad corresponde a 0.5791.

d) ¿Qué valor de la variable deja por debajo de sí el 75% de la probabilidad?

```
> qbinom(0.75,n,p)
[1] 4
```

e) Encontrar el número mínimo n tal que la probabilidad de que el número de averías sea superior a 0.99.

```
> log(0.99)/log(0.02)
[1] 0.002569089
```

N tendrá que tomar el valor 0.00257.

f) Calcular el percentil 95% de la distribución.

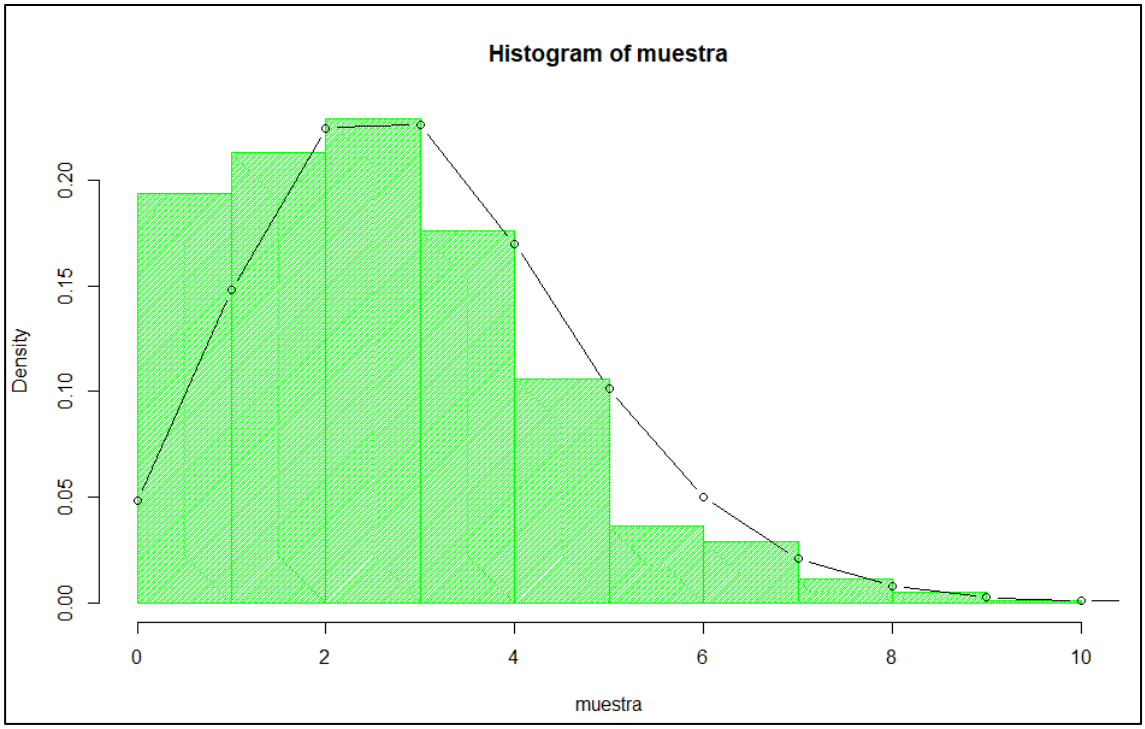
```
> qbinom(0.95,n,p)
[1] 6
```

g) Obtener una muestra de tamaño 1000 de esta distribución.

```
> muestra <- rbinom(1000,n,p)
> mean(muestra)
[1] 3.06
> sd(muestra)
[1] 1.759402
```

h) Representar gráficamente la muestra de g) mediante un diagrama de barras y comparar éste con las frecuencias esperadas según el modelo de datos.

```
> hist(muestra, freq = F, col = "green", density = 55)
> x <- seq(0,(max(muestra)+1))
> fx <- dbinom(x,n,p)
> points(x,fx,type="b")
```



Como podemos observar, lo esperado con lo obtenido tiene una similitud bastante precisa.

LAB 3: EJERCICIO 4

Consideremos una variable aleatoria que sigue una distribución P(x; 3). Se pide:

a) Calcular la probabilidad de que sea mayor o igual que 5.5.

```
> 1-ppois(5.5,3)
[1] 0.08391794
```

La probabilidad es 0.0839.

b) Calcular la probabilidad de sus valores mayores o iguales a 1 y menores o iguales a 6.

```
> F1 <- ppois(1,3)
> F6 <- ppois(6,3)
> F6-F1
[1] 0.7673432
```

La probabilidad es 0.7673.

c) Obtener el percentil 75 de la distribución.

```
> qpois(0.75,3)
[1] 4
```

El percentil es 4.

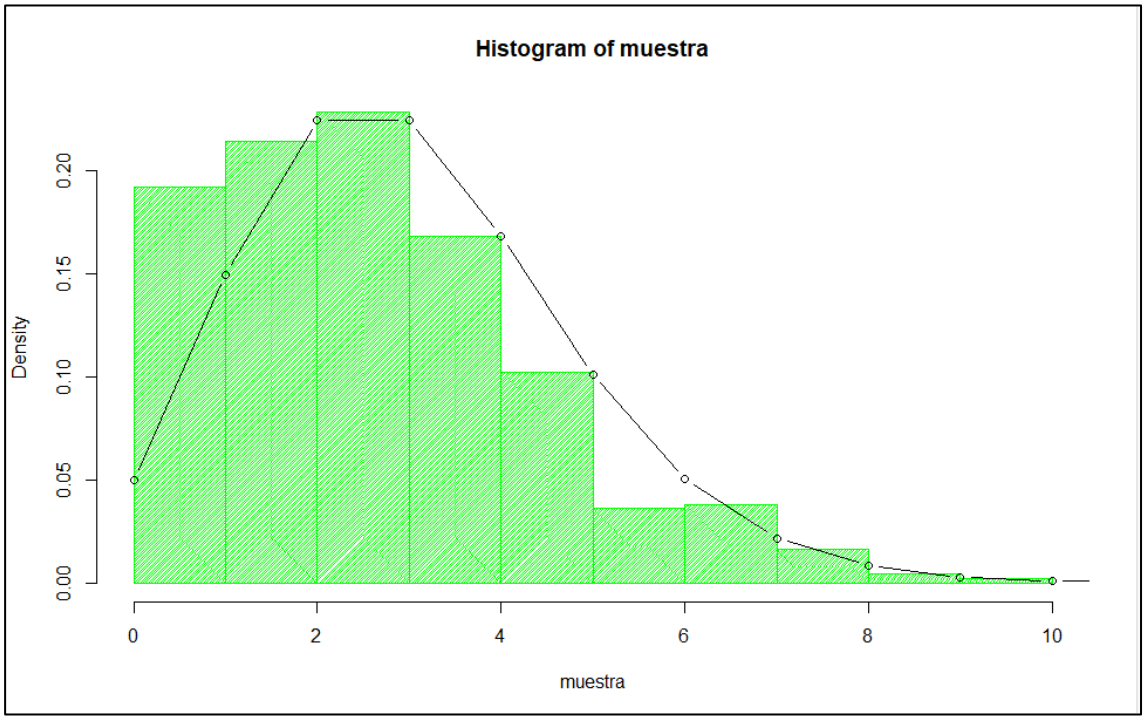
d) ¿Qué valor es el que deja por debajo de sí el 5% de los valores más bajos de la variable?

```
> qpois(0.05,3)
[1] 1
```

El primer percentil.

e) Obtener una muestra de tamaño 500 de la distribución, representarla gráficamente mediante un diagrama de barras y comparar éste con las frecuencias esperadas según el modelo que genera los datos.

```
> muestra <- rpois(500,3)
> hist(muestra, freq = F, col = "green", density = 55)
> mean(muestra)
[1] 3.098
> sd(muestra)
[1] 1.84369
> x <- seq(0,(max(muestra)+1))
> fx <- dpois(x,3)
> points(x,fx,type="b")
```

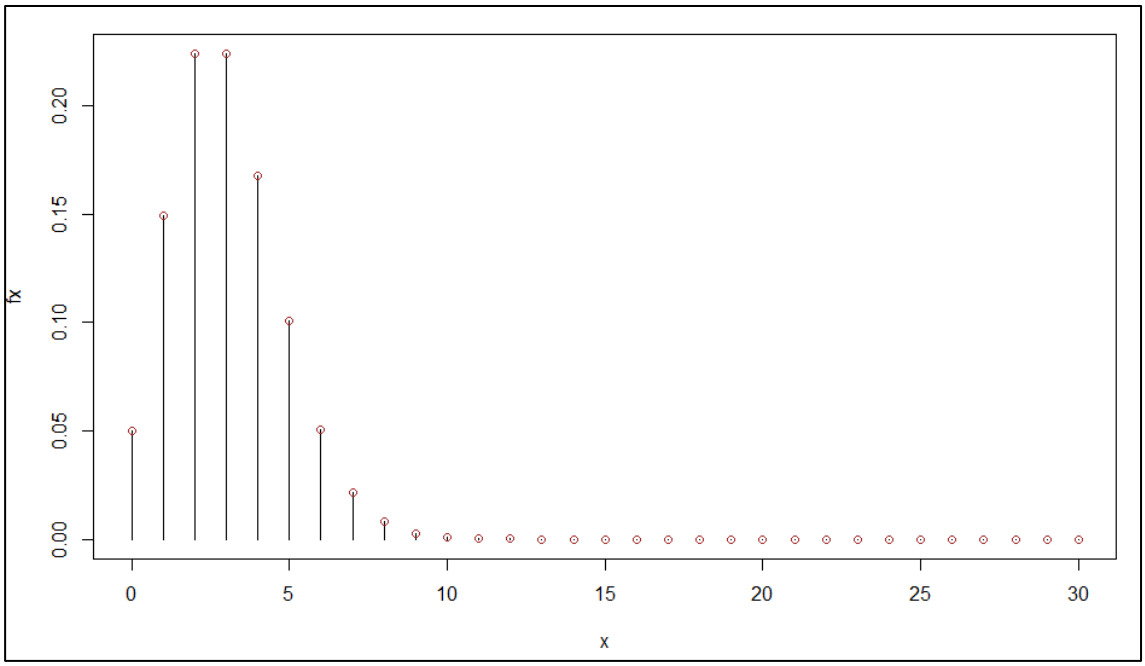


Observamos que se cumple perfectamente.

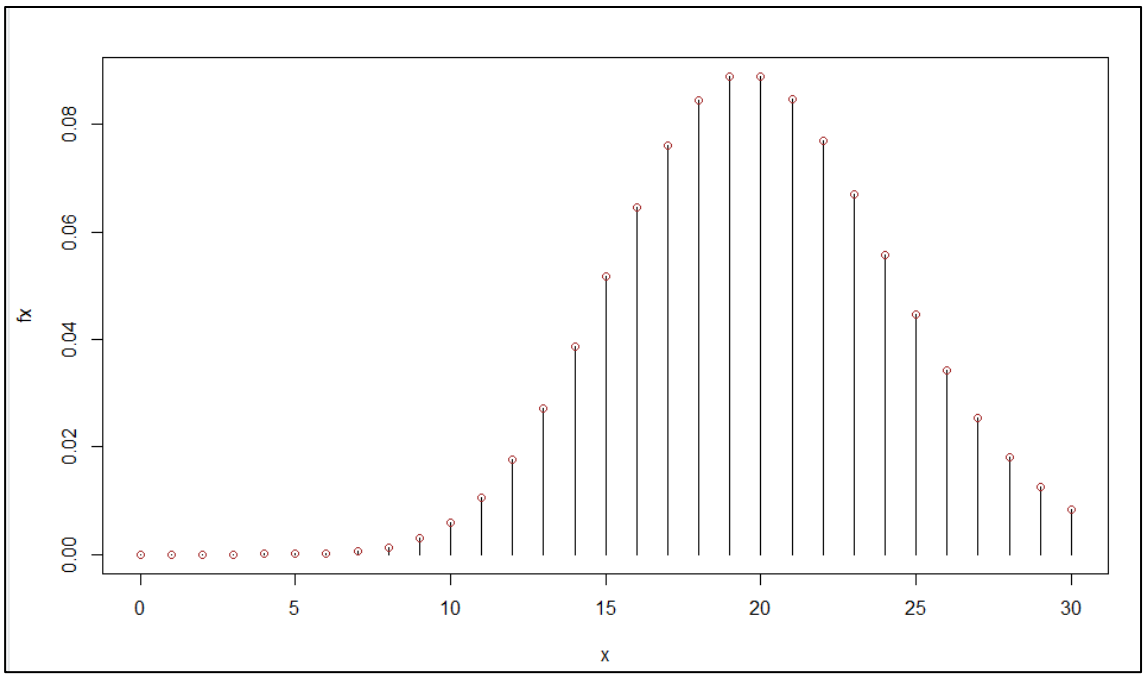
f) Explica la influencia del parámetro lambda en la distribución y visualizar los diferentes resultados superpuestos.

Dependiendo de los valores que le demos a lambda, la media irá cambiando.

```
> lambda <- 3
> x <- seq(0,30)
> fx <- dpois(x, lambda)
> plot(x,fx,type="h",col="black")
> points(x,fx,col="brown")
```




```
> lambda <- 20
> x <- seq(0,30)
> fx <- dpois(x, lambda)
> plot(x,fx,type="h",col="black")
> points(x,fx,col="brown")
```



LAB 3: EJERCICIO 5

Consideramos una variable aleatoria W con distribución N(200,25). Se pide:

```
#N(200,25)
u <- 200
o <- 25
```

a) P[150 < W <= 250]

```
> pnorm(250,u,o) - pnorm(150,u,o)
[1] 0.9544997
```

b) P[W => 255]

```
> 1-pnorm(255,u,o)
[1] 0.01390345
```

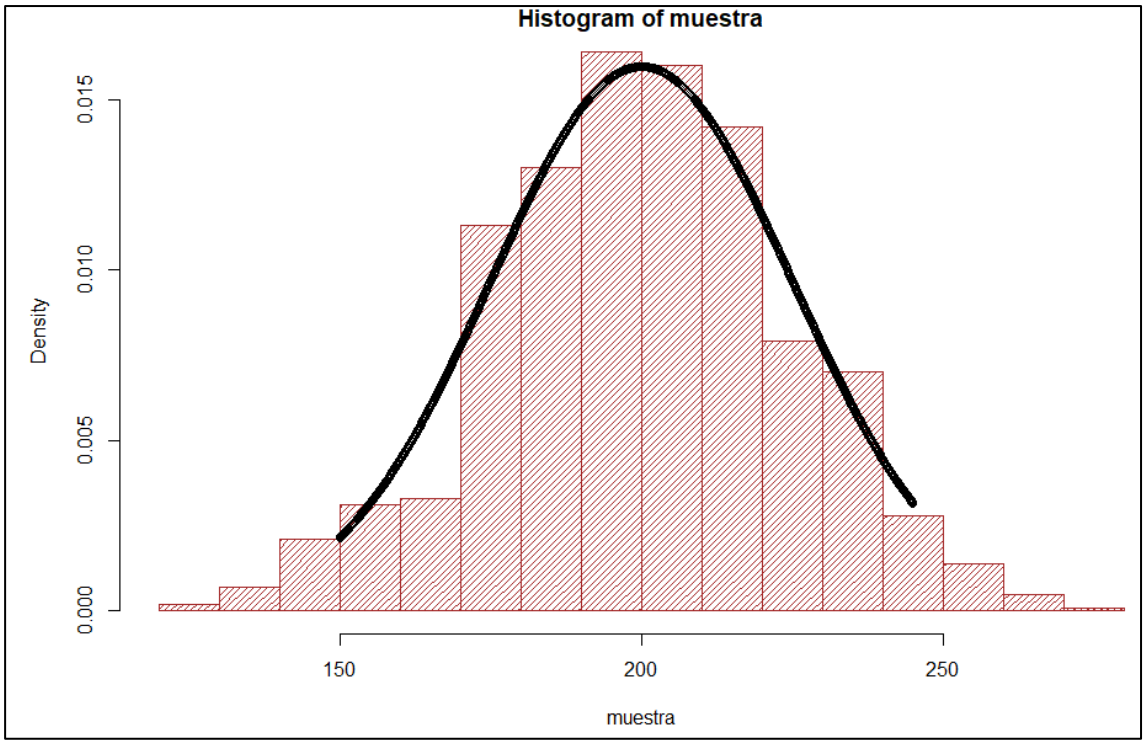
c) Si queremos desechar el 5% de valores más altos de la distribución y el 5% de valores más bajos, ¿con qué intervalo de valores nos quedaremos?

```
> qnorm(0.95,u,o)
[1] 241.1213
> qnorm(0.05,u,o)
[1] 158.8787
```

Nos encontraremos en el intervalo [158, 241].

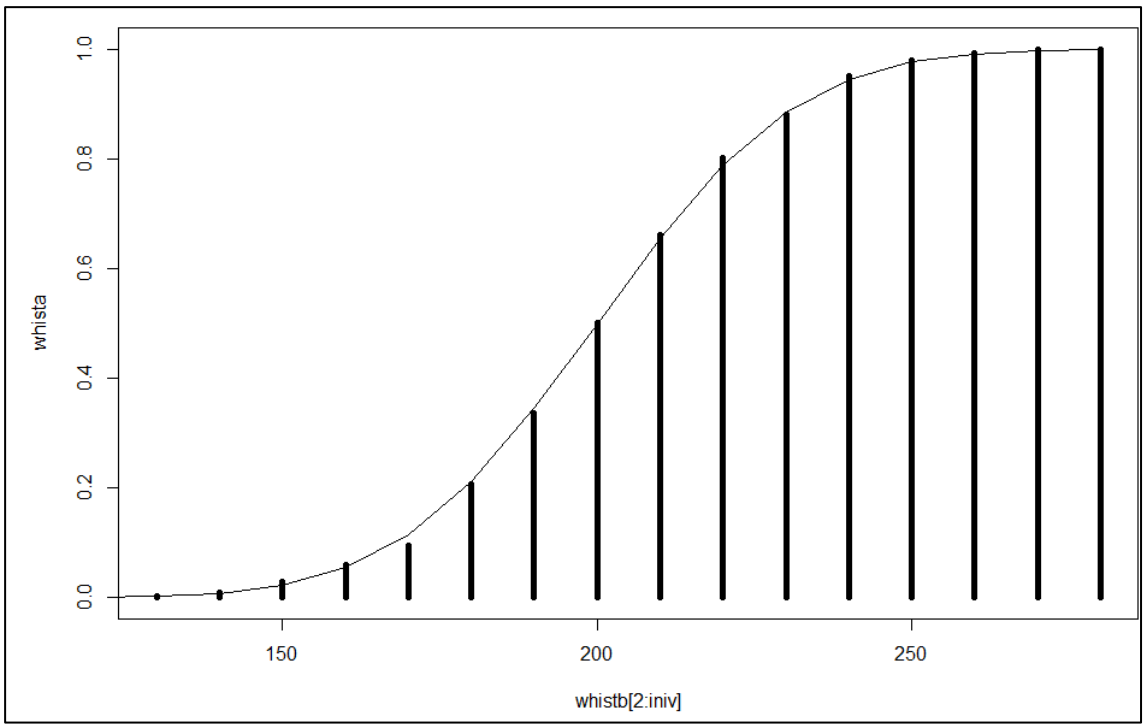
d) Obtener una muestra de tamaño 1000 de la distribución, representar la función de densidad de esta distribución y compararla con el histograma de la muestra obtenida.

```
> muestra <- rnorm(1000,u,o)
> whist <- hist(muestra, freq = F, density = 25,
+               col = "brown")
> x <- seq(150,245,0.1)
> fx <- dnorm(x,u,o)
> points(x,fx,col="black")
```



e) Obtener y visualizar la función de distribución acumulada y situar sobre ella los resultados de a) y b).

```
> whist2 <- hist(muestra)
> whista <- cumsum(whist2$counts)/sum(whist2$counts)
> whistb <- whist2$breaks
> iniv <- length(whistb)
> plot(whistb[2:iniv],whista,type="h",lwd=5)
> fx <- pnorm(whistb,u,o)
> points(whistb,fx,type="l")
```



f) Calcular los coeficientes que definen los factores de forma de la distribución (Curtosis y Asimetría). Razonar las respuestas

```
> library(e1071)
> skewness(whista)
[1] -0.1009001
> kurtosis(whista)
[1] -1.862259
```

Se trata por tanto de una geometría aplanada.

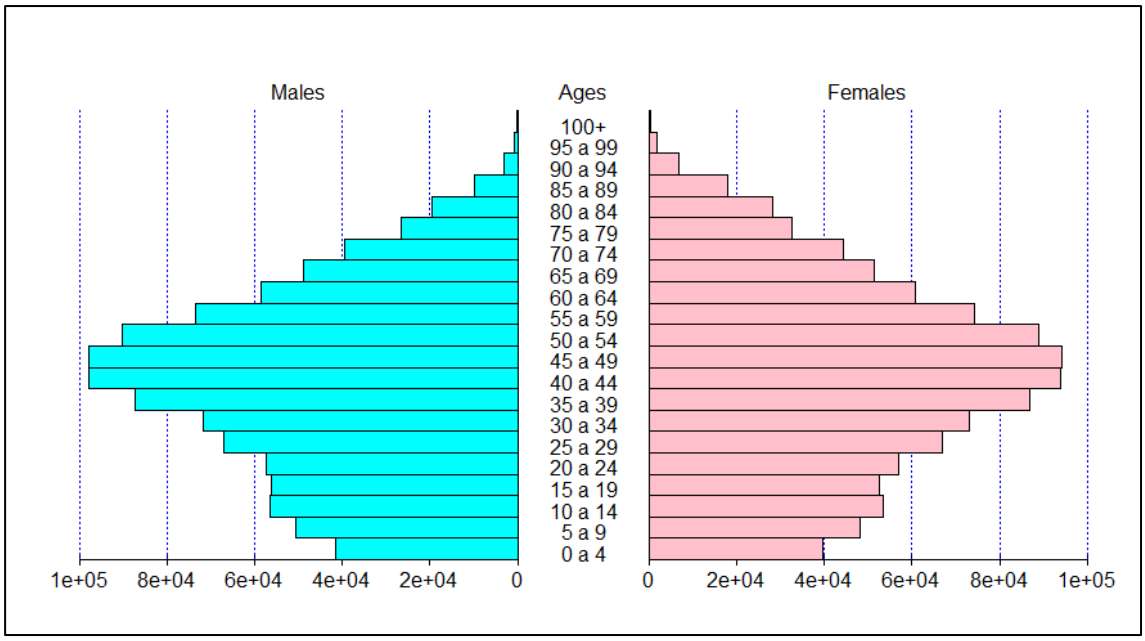
LAB 4: EJERCICIO 1

Obtener del Instituto Canario de Estadística (ISTAC) la distribución por edades de la población entre los años 2000 a 2017. Representar las pirámides de población correspondientes con la librería pyramid, y analizar la evolución anual. Razonar las conclusiones y realizar una pequeña animación de la evolución de las pirámides de población en esos años.

```
library(pyramid)
poblacion <- read.table("piramdies.txt", header=T, sep=",")
attach(poblacion)
```

En lo que se basa este ejercicio es en repetir el siguiente código que vamos a mostrar para cada año, guardar su imagen y convertir todas en formato GIF.

```
poblacion_2018 <- poblacion[YEAR=="2018",]
poblacion_2018_c <- poblacion_2018[1:3]
pyramid(poblacion_2018_c)

poblacion_2017 <- poblacion[YEAR=="2017",]
poblacion_2017_c <- poblacion_2017[1:3]
pyramid(poblacion_2017_c)

poblacion_2016 <- poblacion[YEAR=="2016",]
poblacion_2016_c <- poblacion_2016[1:3]
pyramid(poblacion_2016_c)

poblacion_2015 <- poblacion[YEAR=="2015",]
poblacion_2015_c <- poblacion_2015[1:3]
pyramid(poblacion_2015_c)

poblacion_2014 <- poblacion[YEAR=="2014",]
poblacion_2014_c <- poblacion_2014[1:3]
pyramid(poblacion_2014_c)
```



https://media.giphy.com/media/ilRrURptEWr3puzk1q/giphy.gif

LAB 4: EJERCICIO 3

El fichero "germinación.csv" contiene los datos de germinación de semillas de dos genotipos de la planta parásita orobanche y dos extractos de plantas huésped (judía y pepino) que se utilizaron para estimular la germinación. La variable "count" representa el número de semillas que germinaron de un lote de tamaño "sample". Con estos datos se pide:

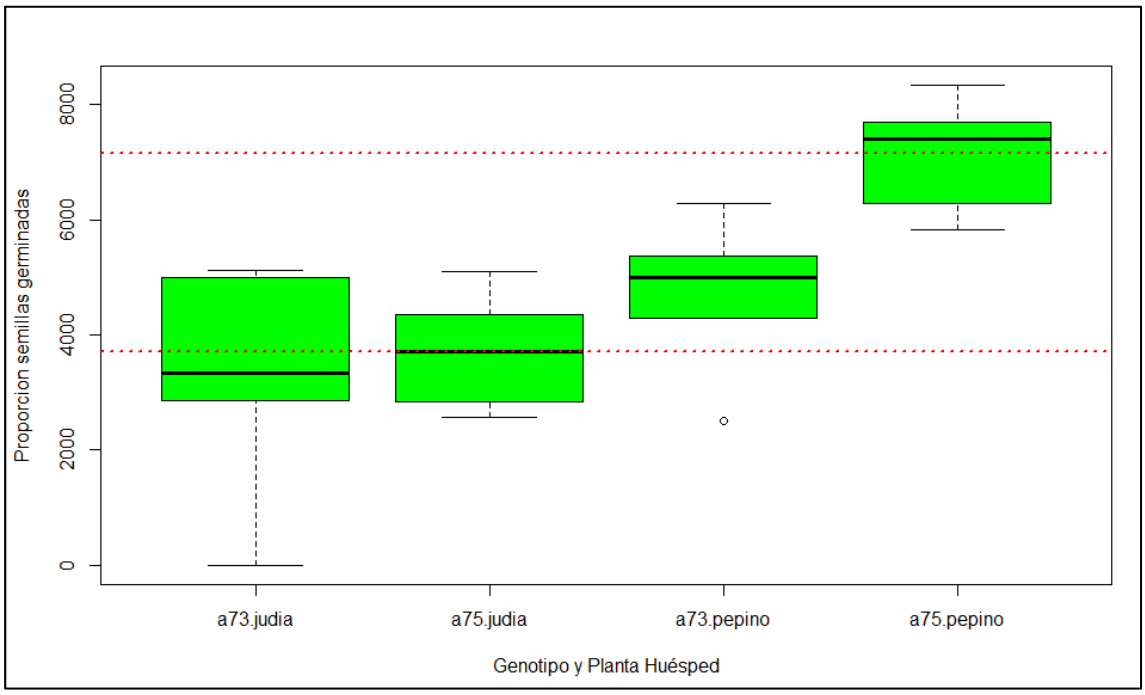
```
library(knitr)
germinacion <- read.table("germination.csv", header = T, sep = ",")
attach(germinacion)
```

a) Crear un data frame con los datos de la variable "count" y una columna adicional que incluya en número de semillas que no germinó.

```
proporcion <- 100*count/sample
no_germinados <- sample-count
germinados <- cbind(germinacion, proporcion, no_germinados)
attach(germinados)
```

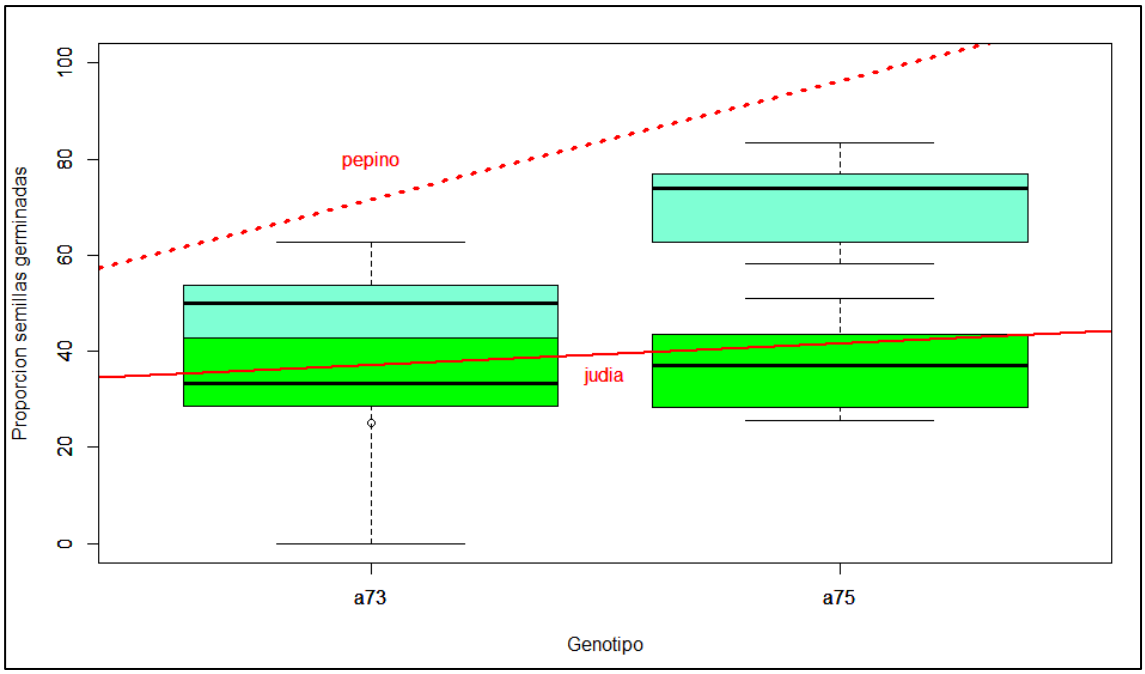
b) Calcular los parámetros de centralización y dispersión de conjunto de muestras para cada genotipo y tipo de planta huésped y analizar gráficamente el efecto del genotipo en la germinación. Explicar las conclusiones.

```
media_proporciones <- aggregate(proporcion~Orobanche+extract,germinacion,mean)
boxplot(100*proporcion~Orobanche+extract, xlab = "Genotipo y Planta Huésped",
        ylab = "Proporcion semillas germinadas", col = "green")
abline(h=100*media_proporciones[4,3], col="red", lwd=2, lty=3)
abline(h=100*media_proporciones[2,3], col="red", lwd=2, lty=3)
```



c) Utilizar la función lm() para ver la tendencia e influencia de los genotipos en la germinación. ¿Son estadísticamente independientes las variables de genotipo y de tipo de planta huésped? Razonar y justificar las respuestas.

```
boxplot(proporcion[extract == "judia"]~Orobanche[extract == "judia"],
        xlab = "Genotipo", ylab = "Proporcion semillas germinadas", col = "green",
        ylim = c(0,100))
modelo1 <- lm(proporcion[extract=="judia"]~Orobanche[extract == "judia"])
abline(modelo1, col = "red", lwd = 2)
modelo2 <- lm(proporcion[extract=="pepino"]~Orobanche[extract=="pepino"])
abline(modelo2, col="red", lwd = 3, lty=3)
boxplot(proporcion[extract=="pepino"]~Orobanche[extract=="pepino"], col = "aquamarine",
        ylim=c(0,100), add = T)
text(1,80,labels="pepino",col="red")
text(1.5,35,labels="judia",col="red")
```

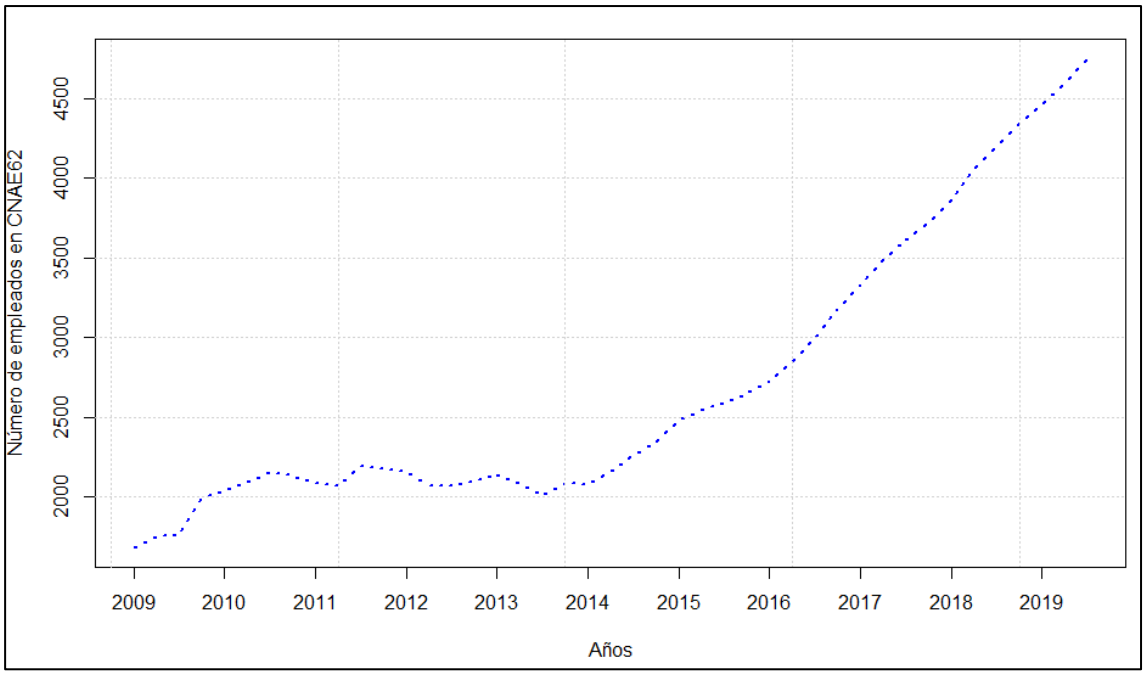


LAB 4: EJERCICIO 4

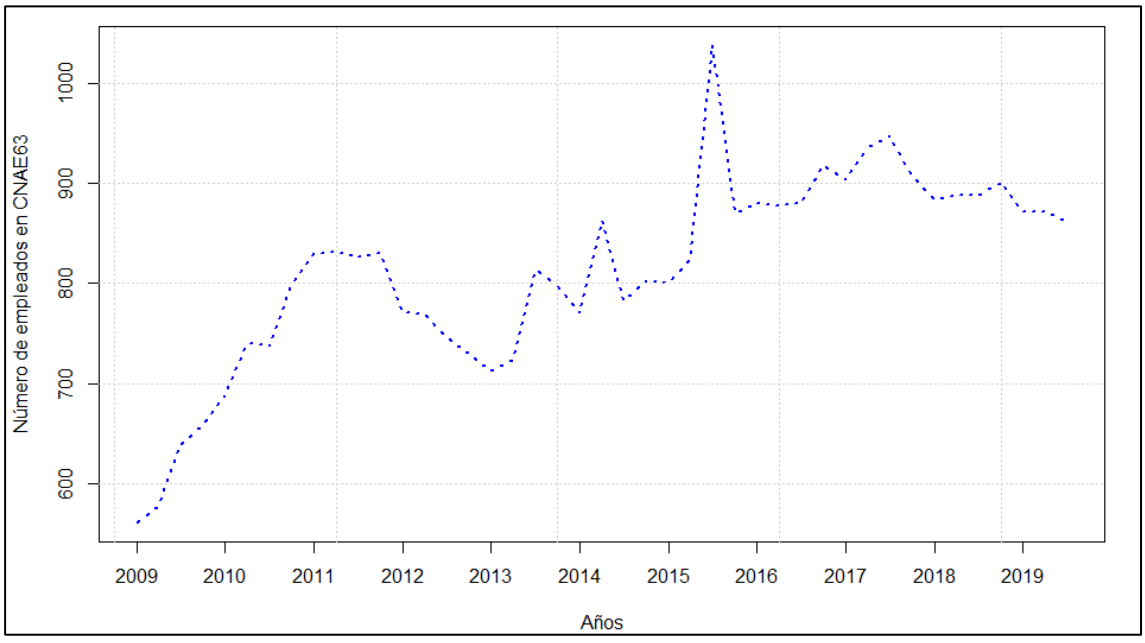
Obtener del ISTAC el fichero con los datos de empleo en actividades relacionadas con la Ingeniería Informática en Canarias por trimestres en el período 2009 a 2018. Se pide:

a) Analizar gráficamente la variación de cada tipología de empleo en las Islas Canarias (por islas y totales) en el período considerado e intentar explicar sus valores singulares.

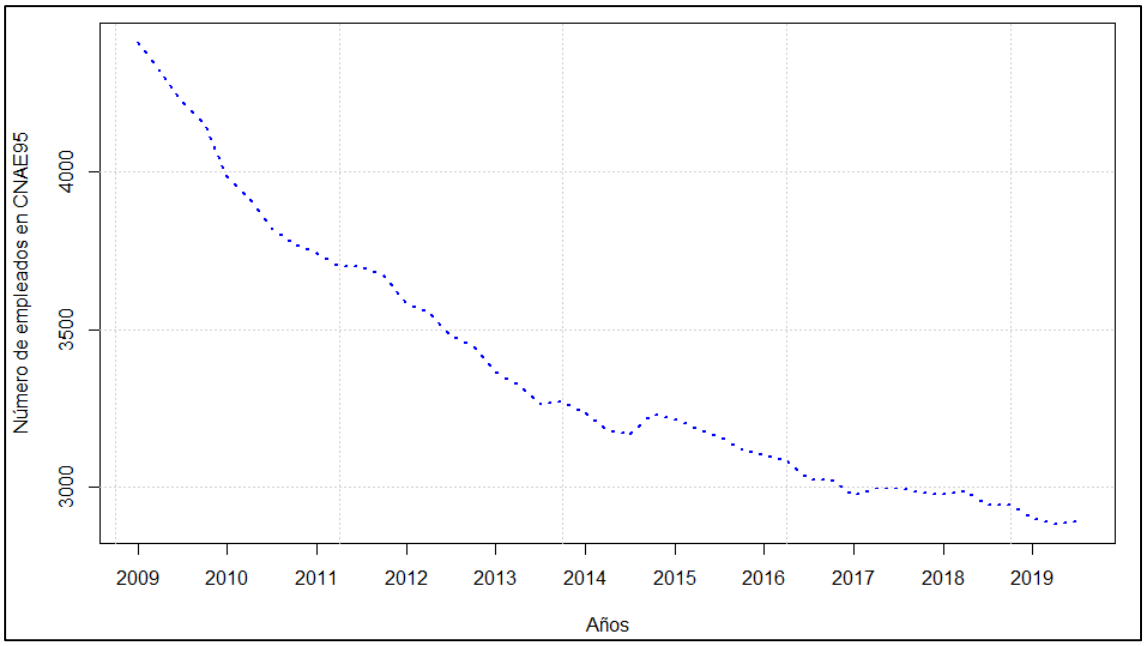
```
> canarias_62 <- datos_empleo[TRIMESTRES == "CNAE_62", 2]
> index <- seq(length(canarias_62),1,-1)
> canarias_ord_62 <- canarias_62 [index]
> tenerife_62 <- TENERIFE [TRIMESTRES == "CNAE_62"]
> granca_62 <- GRAN.CANARIA [TRIMESTRES == "CNAE_62"]
> plot(1:length(canarias_ord_62),canarias_ord_62,type="l", col="blue",
+      lwd=2,lty=3,xlab="Años",ylab="Número de empleados en CNAE62",xaxt = "n")
> grid()
> years <- c ("2009", "2010", "2011", "2012", "2013", "2014", "2015", "2016", "2017", "2018", "2019")
> axis (side=1, at=seq(1,length(canarias_ord_62), 4),labels=years)
```




```
> canarias_63 <- datos_empleo[TRIMESTRES == "CNAE_63", 2]
> index <- seq(length(canarias_63),1,-1)
> canarias_ord_63 <- canarias_63 [index]
> plot(1:length(canarias_ord_63),canarias_ord_63,type="l", col="blue",
+      lwd=2,lty=3,xlab="Años",ylab="Número de empleados en CNAE63",xaxt = "n")
> years <- c ("2009", "2010", "2011", "2012", "2013", "2014", "2015", "2016", "2017", "2018", "2019")
> axis (side=1, at=seq(1,length(canarias_ord_63), 4),labels=years)
> grid()
```



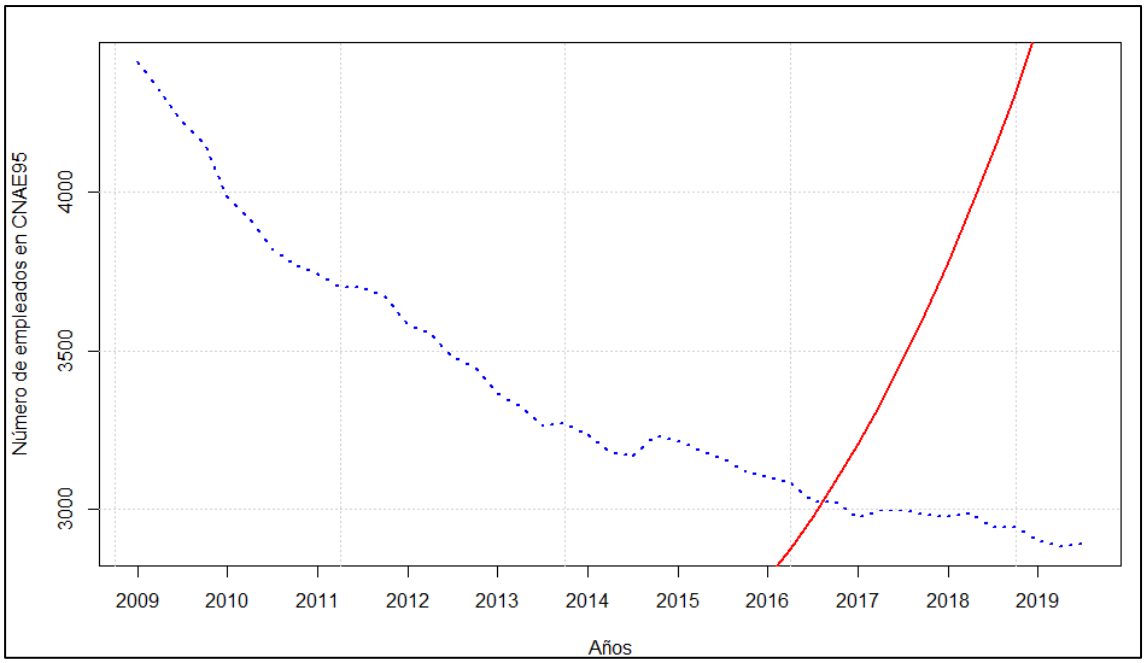
```
> canarias_95 <- datos_empleo[TRIMESTRES == "CNAE_95", 2]
> index <- seq(length(canarias_95),1,-1)
> canarias_ord_95 <- canarias_95 [index]
> plot(1:length(canarias_ord_95),canarias_ord_95,type="l", col="blue",
+      lwd=2,lty=3,xlab="Años",ylab="Número de empleados en CNAE95",xaxt = "n")
> years <- c ("2009", "2010", "2011", "2012", "2013", "2014", "2015", "2016", "2017", "2018", "2019")
> axis (side=1, at=seq(1,length(canarias_ord_95), 4),labels=years)
> grid()
```



Como observamos, el sector del CNAE_62 ha crecido según ha pasado el tiempo, CNAE_63 se ha mantenido estable, y CNAE_95 ha ido decreciendo cada vez más en toda Canarias.

b) Utilizando la librería mgcv(), encontrar un modelo de seguimiento del empleo, representar gráficamente su evolución y predicciones.

```
x <- 1:length(canarias_ord_62)
y <- canarias_ord_62
modelo1 <- gam(y~x+I(x^2)+I(x^3))
xv <- x
yv <- predict(modelo1, list(x=xv))
lines(xv, yv, col="red", lwd=2)
yv2018 <- predict(modelo1, list(x=x[2]))
points(x[2], yv2018, pch=16, col="green")
lines(c(x[2], x[2]), c(0, yv2018), col="red", lty=3, lwd=3)
```



c) Encontrar la isla donde hay más empleo y en qué etapa.

```
> max(TENERIFE)
[1] 2603
> max(GRAN.CANARIA)
[1] 2038
> max(LA.PALMA)
[1] 145
> max(EL.HIERRO)
[1] 7
> max(FUERTEVENTURA)
[1] 154
> max(LANZAROTE)
[1] 231
> max(LA.GOMERA)
[1] 36
```

Por tanto, concluimos que Tenerife es la isla donde ha habido y donde hay más empleo, pues la etapa que indica es el último trimestre de 2019.

d) Analizar comparativamente la evolución durante dos años del empleo en dos islas diferentes y explicar sus variaciones y sus aspectos comunes.

Un año tiene cuatro trimestres, y tenemos tres apartados por trimestre. Por tanto, observaremos 24 filas de la tabla, que será lo correspondiente a los dos últimos años. Nos fijaremos en Gran Canaria y Tenerife.

TRIMESTRES	CANARIAS	LANZAROTE	FUERTEVENTURA	GRAN.CANARIA	TENERIFE	LA.GOMERA	LA.PALMA	EL.HIERRO
CNAE_62	4750	93	83	1894	2603	3	48	4
CNAE_63	861	56	54	285	431	7	22	1
CNAE_95	2890	202	130	1274	1172	27	74	3
CNAE_62	4602	93	84	1808	2543	3	44	6
CNAE_63	872	58	57	298	421	6	25	2
CNAE_95	2884	205	136	1278	1148	32	74	3
CNAE_62	4459	91	74	1692	2531	3	43	6
CNAE_63	872	63	60	303	409	6	23	2
CNAE_95	2900	200	131	1285	1164	34	74	3
CNAE_62	4351	88	71	1574	2545	2	48	5
CNAE_63	902	64	61	306	433	10	21	2
CNAE_95	2943	203	127	1312	1178	35	74	3
CNAE_62	4203	88	74	1517	2450	1	47	5
CNAE_63	888	59	74	307	412	9	20	2
CNAE_95	2943	197	128	1306	1194	32	73	3
CNAE_62	4060	89	75	1485	2338	1	47	4
CNAE_63	888	61	75	327	385	11	23	2
CNAE_95	2990	196	127	1306	1237	35	76	3
CNAE_62	3866	91	71	1404	2232	2	45	3
CNAE_63	884	60	71	324	391	8	26	1
CNAE_95	2977	195	128	1302	1227	34	79	3
CNAE_62	3721	89	72	1380	2118	2	43	2
CNAE_63	907	54	72	336	408	9	25	0
CNAE_95	2982	194	126	1313	1219	34	84	3

Observamos que de manera general, Tenerife tiene más trabajadores en los sectores CNAE_62 y CNAE_63 frente a Gran Canaria. Por el otro lado, Gran Canaria tiene más trabajadores en el sector CNAE_95. Ello indica que en la provincia de Santa Cruz de Tenerife, este tipo de trabajo es mucho más popular comparado con la otra provincia. Sin embargo, se podría explicar por el número de habitantes.

LAB 5: EJERCICIO 1

El fichero "Alturas_Estudiantes_EII.txt" contiene un conjunto de datos de valores de medidas de la altura (en centímetros) de 635 estudiantes de la EII. Se pide:

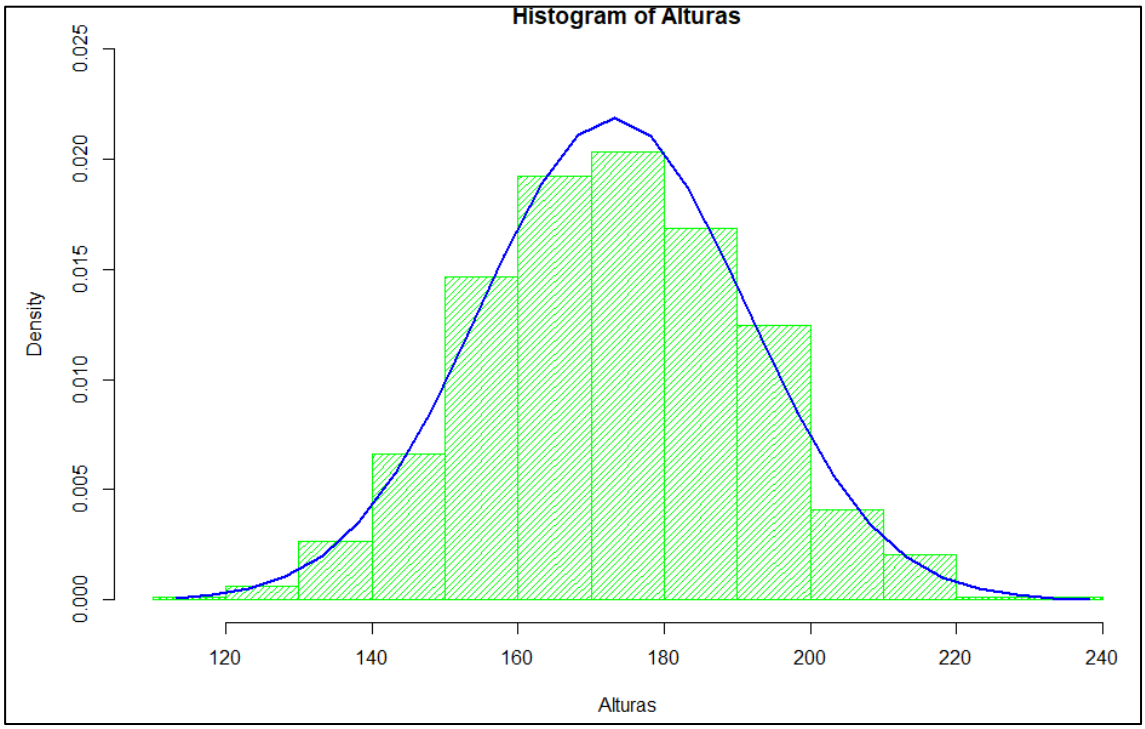
```
library(knitr)
library(MASS)

alturas <- read.table("Alturas_Estudiantes_EII.txt", header = T, sep =",")
attach(alturas)
```

a) Ajustar una distribución normal a esos datos mediante el método de máxima verosimilitud.

```
> parametro <- fitdistr(Alturas, "normal")
> parametro
      mean          sd
  173.0663622    18.2622938
 (  0.7247170) (  0.5124523)
```

b) Representar gráficamente el diagrama de barras de los datos junto con la función masa de la distribución del ajuste.

```
> hist(Alturas, col ="green", density = 25, freq = F, ylim = c(0, 0.025))
> M <- parametro$estimate[1]
> O <- parametro$estimate[2]
> X <- seq(min(Alturas),max(Alturas),5)
> points(X,dnorm(X,M,O),col="blue",type="l",lwd=2)
```



c) ¿Es la distribución resultante un buen ajuste para los datos? Razonar la respuesta.

Como vemos por la gráfica, efectivamente es un buen ajuste para los datos dados.

LAB 5: EJERCICIO 2

El fichero "sueldos_hostelería.txt" contiene una muestra obtenida en el sur de la isla en empresas del sector de la hostelería sobre el salario anual neto que percibían los trabajadores de categorías y antigüedades análogas.

```
library(knitr)
sueldos <- read.table("sueldos_hosteleria.txt", header = T, sep = ",")
attach(sueldos)
```

a) Si se supone que el salario neto anual de estos trabajadores sigue una distribución normal, obtener un intervalo de confianza al 90% para el salario medio neto anual correspondiente.

```
> xmedia <- mean(Sueldos)
> s <- sd(Sueldos)
> n <- length(Sueldos)
> t <- qt(0.95, n-1)
> xmedia - t*s/sqrt(n)
[1] 18411.73
> xmedia + t*s/sqrt(n)
[1] 19137.79
```

La media del salario se encuentra entre los 18411.73 y los 19137.79€.

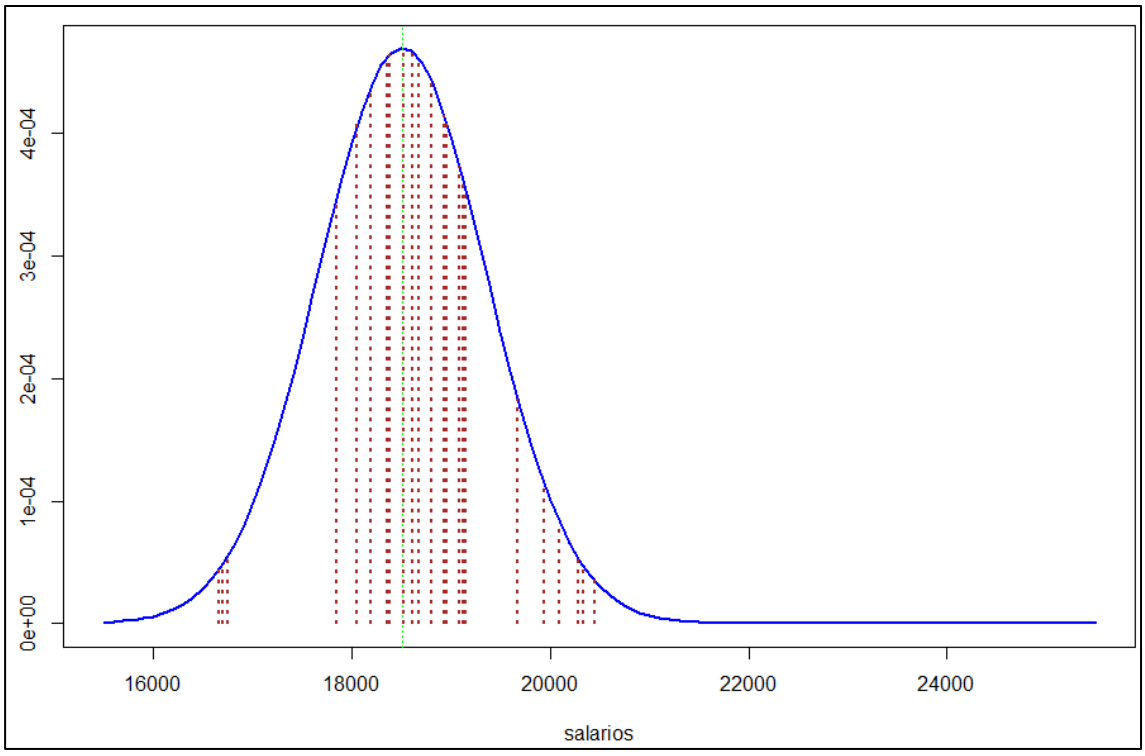
b) Encontrar el intervalo de confianza para la varianza y la desviación estándar en las condiciones del apartado anterior.

```
> x2inf <- qchisq(0.05, n-1)
> x2sup <- qchisq(0.95, n-1)
> sqrt(((n-1)*(s*s))/x2sup)
[1] 861.304
> sqrt(((n-1)*(s*s))/x2inf)
[1] 1396.679
```

La desviación se encuentra entre los 861.304 y los 1396.679€.

c) Visualizar los datos asumiendo que han podido obtenerse de una distribución normal de media 18510€ y desviación estándar 850€. Explicar las conclusiones.

```
salarios <- seq(15500,25500,100)
plot(salarios,dnorm(salarios,18510,850), type = "l", col = "blue", lwd = 2)
abline(v=18510, col = "green", lty=3)

Sueldos2 <- sort(Sueldos)
points(Sueldos2,dnorm(Sueldos2,18510,850), col = "brown", type ="h", lwd = 2, lty = 3)
```



Observamos que los salarios más frecuentes son los que se encuentran en el rango cercano de los 18510€, que es nuestro límite identificado por la línea verde.

LAB 5: EJERCICIO 3

Tras una entrevista con los empresarios del sector estos afirman que el salario medio está establecido en 18510€ netos anuales. Para verificarlo se hizo el muestreo que refleja el fichero "sueldos_hosteleria.txt", que contiene una muestra obtenida en el sur de la isla en empresas del sector de la hostelería sobre el salario anual neto que percibían los trabajadores de categorías y antigüedad análogas. Con esta información, ¿tienen razón los empresarios? (Utilizar un nivel de significación del 5%).

U es el valor hipotético con el que se compara la media o la diferencia de medias en el contraste. Si es mayor a 0.05, se acepta, por tanto, los empresarios tendrán razón. Si es menor a 0.05, los empresarios no tendrán razón.

Como vemos en el test, en este caso en concreto los empresarios tienen razón. El salario medio está establecido en 18510€, y se aceptará como válida la afirmación, por ende, la hipótesis es nula.

18336.83 es el límite inferior, mientras que 19212.70 es el límite superior. Al tratarse p-value menor al valor de significación 0.5, entonces se toma como válida la hipótesis de los empresarios.

```
> library(knitr)
> datos<-read.table("sueldos_hosteleria.txt",header = TRUE, sep = ",", dec = ".")
> str(datos)
'data.frame':   25 obs. of  1 variable:
 $ Sueldos: num  19677 20342 16695 16755 18197 ...
> kable(datos[1:7,])


|        x|
|--------:|
| 19676.77|
| 20341.72|
| 16695.23|
| 16755.38|
| 18196.91|
| 18355.07|
| 20283.26|
> attach(datos)
```

```
> t.test(Sueldos, alternative = "two.sided", mu = 18510)

        One Sample t-test

data:  Sueldos
t = 1.2478, df = 24, p-value = 0.2242
alternative hypothesis: true mean is not equal to 18510
95 percent confidence interval:
 18336.83 19212.70
sample estimates:
mean of x
 18774.76
```

LAB 5: EJERCICIO 4

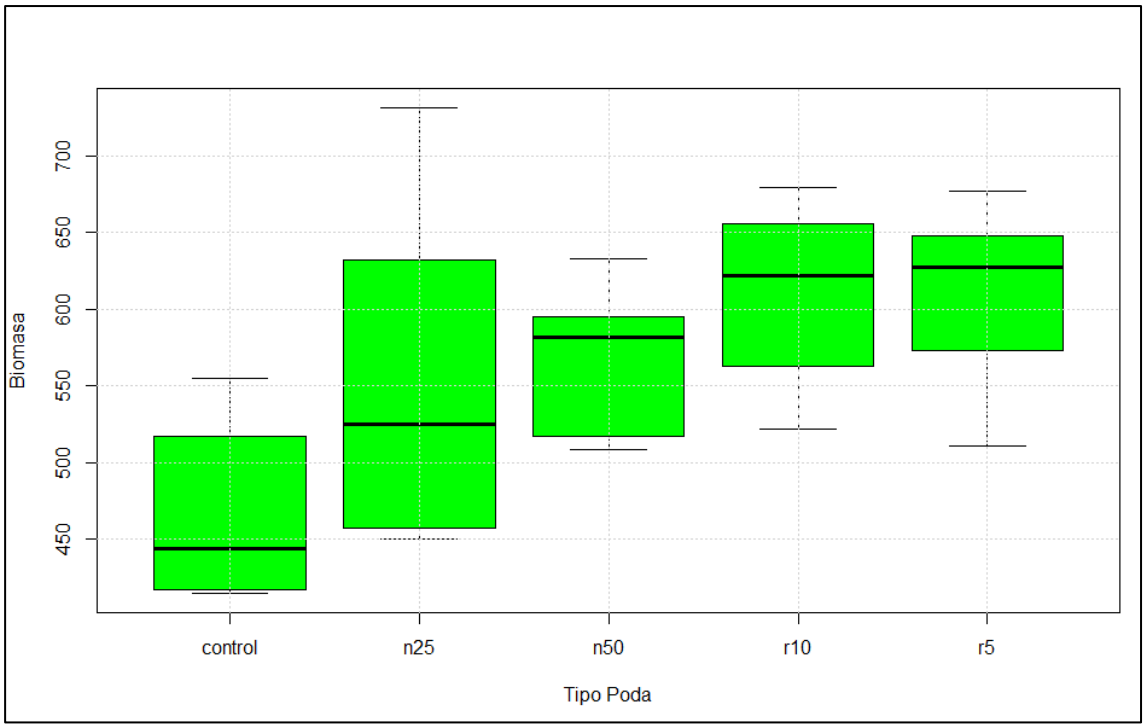
Se quiere estudiar el efecto de la poda en el rendimiento del crecimiento en un tipo de plantas. Para ello se mide la biomasa resultante de varios experimentos de poda, los datos están en el fichero "plantas_poda.txt". Se disponen datos de un grupo de plantas de control, donde no se hace ninguna poda (denominado control) y de datos de plantas relativos a dos tipos de poda, un primer tipo denominado poda ligera y rápida (con dos formas de hacerla, n25 y n50) y otro tipo denominado poda de raíz (r10 y r5).

A un nivel de confianza del 95%:

a) Analizar si puede considerar que los cuatro métodos de poda producen resultados equivalentes.

```
> library(knitr)
> datos <- read.table("plantas_poda.txt", header = TRUE,dec = ".", sep = ",")
> str(datos)
'data.frame':   30 obs. of  2 variables:
 $ Biomasa  : int  551 457 450 731 499 632 595 580 508 583 ...
 $ Tipo_Poda: Factor w/ 5 levels "control","n25",..: 2 2 2 2 2 2 3 3 3 3 ...
> kable(datos[1:7, ])


| Biomasa|Tipo_Poda |
|-------:|:---------|
|     551|n25       |
|     457|n25       |
|     450|n25       |
|     731|n25       |
|     499|n25       |
|     632|n25       |
|     595|n50       |
```

```
> attach(datos)
The following objects are masked from datos (pos = 3):

    Biomasa, Tipo_Poda

The following objects are masked from datos (pos = 4):

    Biomasa, Tipo_Poda

The following objects are masked from datos (pos = 6):

    Biomasa, Tipo_Poda

> boxplot(Biomasa~Tipo_Poda, col = "green", ylab = "Biomasa", xlab = "Tipo Poda")
> grid()
```



b) ¿Hay algún método superior a los demás? Razonar las respuestas.

```
> ANOVA <- aov(Biomasa~Tipo_Poda, data = datos)
> summary(ANOVA)
            Df Sum Sq Mean Sq F value  Pr(>F)
Tipo_Poda    4  85356   21339   4.302 0.00875 **
Residuals   25 124020    4961
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Observamos que el método superior sería el referente al n25, puesto que cubre más biomasa.