Ejercicios LAB 3 Y 4
Cuestiones L4, L5, L6 y L7

MÉTODOS ESTADÍSTICOS

AARÓN HERNÁNDEZ ÁLVAREZ

Contenido

EJERCICIOS LABORATORIO

Ejercicio 3.1

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 en la Unión Europea. Con los datos en él incluidos.

a) Ordenar el data frame alfabéticamente por países
b) Calcular la media, mediana y cuantiles de la duración del trabajo en semanas.
c) Evaluar los parámetros de dispersión de la duración.
d) Ordenar los países por las semanas de trabajo acumuladas en un año
e) Visualizar las diferencias con un diagrama de caja y distinguir los valores singulares. Explicar los campos de datos resultado del uso de la función ***boxplot***()
f) Mostrar gráficamente la situación de España en e)
g) Visualizar gráficamente las variaciones entre países de la UE (ordenados por duración y sin ordenar) y señalar en el gráfico los valores que corresponde a España.
h) (**Opcional**) El fichero "***H_T_A_UE_2017.txt***" contiene los datos para 22 países de la UE del total del número de horas medio por trabajador en un año. Con sus datos, calcular la duración media semanal de la jornada laboral en Grecia, España y Alemania. Explicar las conclusiones.

```
setwd(".")
library(knitr)
library(ggplot2)
library(sqldf)
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Loading required package: RSQLite
```

Apartado a)

Obtener el dataframe ordenado por nombre de países

```
datos<-read.table("Archivos/Empleo.txt", dec = ".", sep=",")
datos_ord_pais<-datos[order(datos$Pais),]
```

Apartado b)

Calcular la media, mediana y cuantiles de la duración

```
attach(datos_ord_pais)
media_d<-mean(Duracion)
mediana_d<-median(Duracion)
cuantiles_d<-quantile(Duracion)
```

Apartado c)

Evaluar los parámetros de dispersión de la duración.

```
var_d<-var(Duracion)
sd_d<-sd(Duracion)
```

Apartado d)

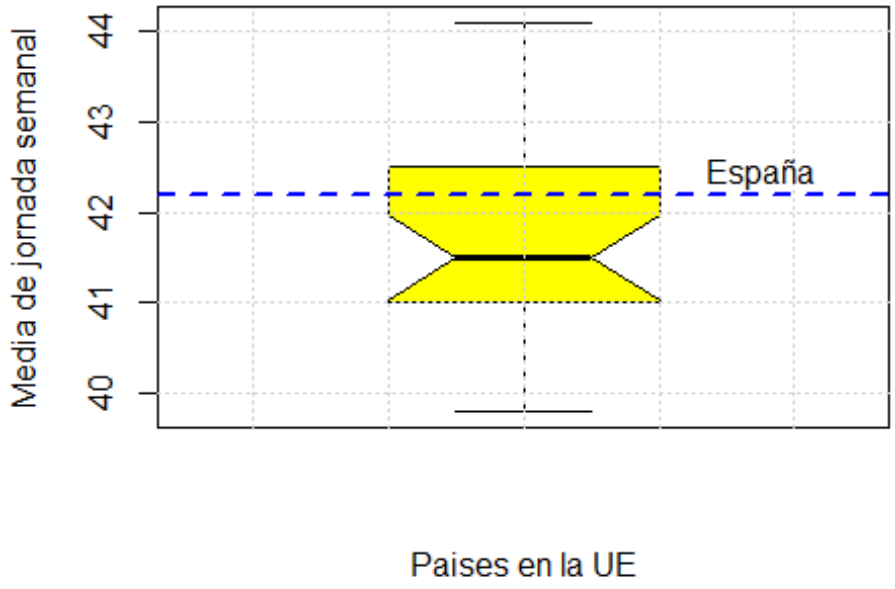
Ordenar los países por semanas trabajadas.

```
detach()
datos_ord_semana<-datos[order(datos$Duracion),]
```

Apartado e)

Visualizar las diferencias con un diagrama de caja y distinguir los valores singulares.

```
attach(datos_ord_semana)
boxplot(Duracion, notch = T,col = "Yellow",
        ylab = "Media de jornada semanal",
        xlab = "Paises en la UE")
grid()
abline(h=Duracion[Pais=="España"],col="blue",lwd=2, lty=2)
text(1.35,y=Duracion[Pais=="España"] +0.25,labels="España")
```



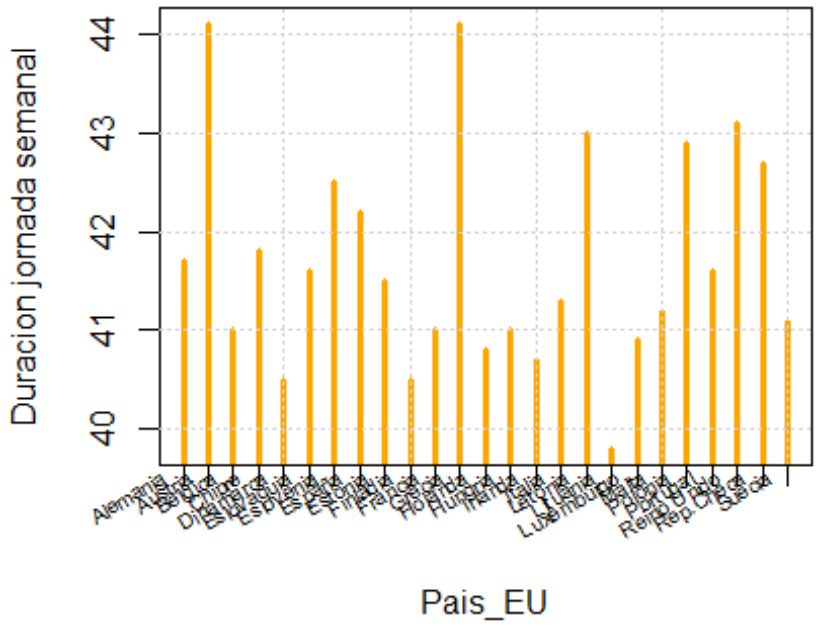

```
detach()
```

Apartado f)

```
attach(datos_ord_pais)
par(mar=c(6,6,2,2)+0.1)
plot(1:nlevels(Pais), Duracion[Pais==levels(Pais)],
    xaxt = "n", lwd =3, type="h", col="orange",
    xlab="Pais_EU",
    ylab="Duracion jornada semanal")
grid()
axis(side =1, at = 1:length(Pais), labels=F)
text(1:nlevels(Pais), par("usr")[3]-0.1,
```

```
     labels=Pais, srt=30, pos=2, cex=0.6,
     xpd = TRUE)
points(which(Pais=="España"), Duracion[Pais="España"], type="h",
       col="green",lwd=3)
points(which(Pais=="España"), Duracion[Pais="España"], type="p",
       col="green",lwd=3)
```

Apartado g)




```
detach()
```

Apartado h)

Revisar las horas trabajadas por semana en GRE, ESP Y GER

```
horas_trabajo<-read.table("Archivos/H_T_A_UE_2017.txt", dec = ".
", sep=",")
attach(horas_trabajo)
media_semanal_ESP<-as.numeric(horas_trabajo[horas_trabajo$Pais==
"ESP",])/as.numeric(datos$Duracion[datos$Pais == "España"])
media_semanal_GRE<-as.numeric(horas_trabajo[horas_trabajo$Pais==
"GRC",])/as.numeric(datos$Duracion[datos$Pais == "Grecia"])
media_semanal_GER<-as.numeric(horas_trabajo[horas_trabajo$Pais==
"DEU",])/as.numeric(datos$Duracion[datos$Pais == "Alemania"])
data.frame(media_semanal_ESP,media_semanal_GER,media_semanal_GRE
)
```

```
##   media_semanal_ESP media_semanal_GER media_semanal_GRE
## 1         0.1184834        0.07194245         0.2267574
## 2        39.9644550       32.51798561        43.2199546
```

Conclusiones

Se puede comprobar que la media de horas en España se encuentra en una intermedia respecto a los restantes países de la UE, aunque es cierto que en ese intervalo, es superior al resto de países ubicados en dicho intervalo.

Por otra parte, Alemania goza de una semana laboral correcta, aunque lejos de Finlandia o Lituania, por ejemplo.

Una mejora de las condiciones laborales, sumada a medidas como un plan nacional de empleo podría ayudar a mitigar esta cifra en el caso de España.

Ejercicio 3.2

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 pìde:

a) Calcular, las medias de la velocidad de reacción en función del empleo o no de Puromicina.
b) Evaluar los parámetros de dispersión de la velocidad de reacción.
c) Visualizar si la concentración del sustrato influye en la velocidad de reacción en los casos en que se trata o no con Puromicina.
d) Ordenar por velocidad de reacción.
e) Analizar los efectos del uso de la Puromicina y de la concentración del sustrato en la velocidad de reacción. Realizar algunas predicciones
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?

```
setwd(".")
library(knitr)
library(ggplot2)
library(sqldf)
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Loading required package: RSQLite
```

```
library(DMwR2)
```

```
## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
default<-read.table("Archivos/Puromicina.txt", dec = ".", sep=",
",header=T)
```

Apartado a)

```
attach(default)
media_VR<-aggregate(velocidad_reaccion~Puromicina,default,mean)
kable(media_VR)
```

Puromicina	velocidad_reaccion
treated	141.5833

untreated 110.7273

```
mediana_VR<-aggregate(velocidad_reaccion~Puromicina,default,median)
kable(mediana_VR)
```

Puromicina	velocidad_reaccion
treated	145.5
untreated	115.0

Apartado b)

#parámetros d e dispersión

```
varianza_VR<-aggregate(velocidad_reaccion~Puromicina,default,var)
kable(varianza_VR)
```

Puromicina	velocidad_reaccion
treated	2805.356
untreated	1334.218

```
sd_VR<-aggregate(velocidad_reaccion~Puromicina,default,sd)
kable(sd_VR)
```

Puromicina	velocidad_reaccion
treated	52.96561
untreated	36.52695

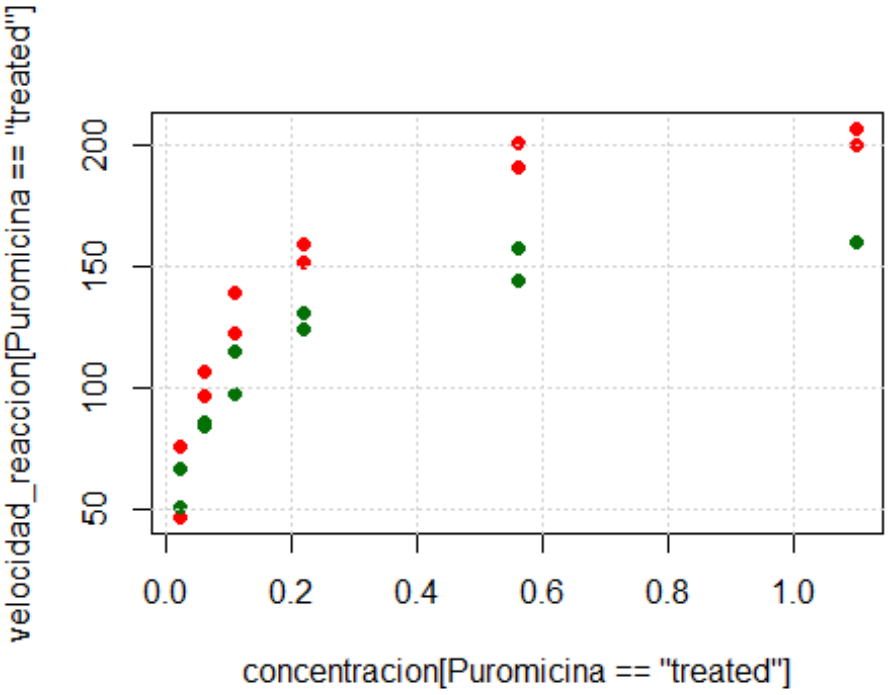
```
detach()
```

Apartados c) y d)

Visualizar si la concetraciónde sustrato influye

Datos sin ajuste.

```
attach(default)
plot(concentracion[Puromicina=="treated"], velocidad_reaccion[Puromicina=="treated"],
     col="red", pch=19)
points(concentracion[Puromicina=="untreated"], velocidad_reaccion[Puromicina=="untreated"],
     col="#037005", pch=19)
grid()
```




```
detach()
```

Datos ajustados.

```
medias_datos<-aggregate(velocidad_reaccion~Puromicina+concentrac
ion, default, mean)
kable(medias_datos)
```

Puromicina	concentracion	velocidad_reaccion
treated	0.02	61.5
untreated	0.02	59.0
treated	0.06	102.0
untreated	0.06	85.0
treated	0.11	131.0
untreated	0.11	106.5
treated	0.22	155.5
untreated	0.22	127.5
treated	0.56	196.0
untreated	0.56	151.0
treated	1.10	203.5
untreated	1.10	160.0

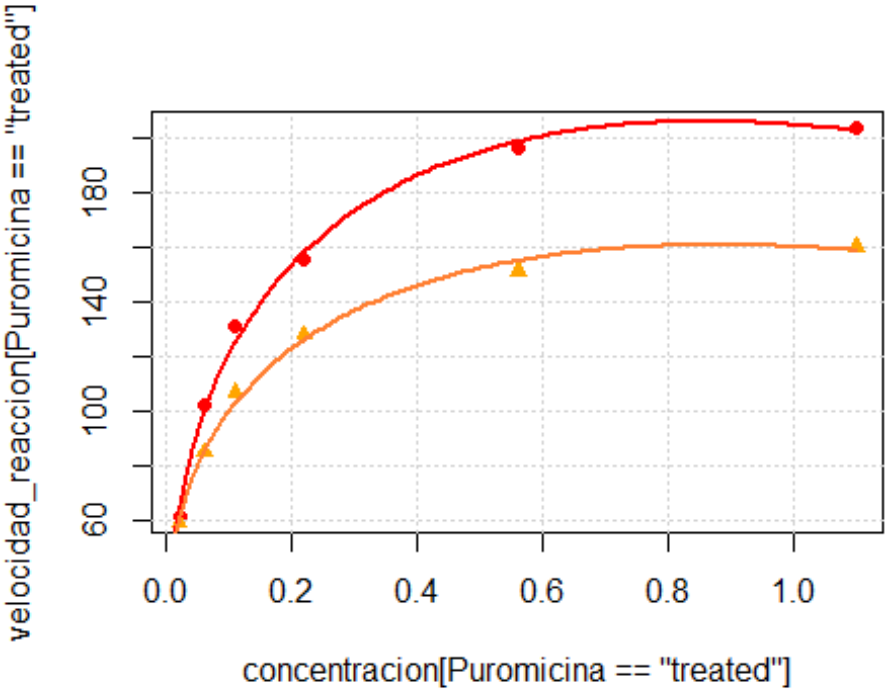
```
attach(medias_datos)
plot(concentracion[Puromicina=="treated"],
     velocidad_reaccion[Puromicina=="treated"],
     col="red", pch=19)
```

```
grid()
points(concentracion[Puromicina=="untreated"],
       velocidad_reaccion[Puromicina=="untreated"],
       col="orange", pch=17)

datos_ord_velocidad<-default[order(default$velocidad_reaccion),]
kable(datos_ord_velocidad)
```

	concentracion	velocidad_reaccion	Puromicina
2	0.02	47	treated
14	0.02	51	untreated
13	0.02	67	untreated
1	0.02	76	treated
15	0.06	84	untreated
16	0.06	86	untreated
3	0.06	97	treated
17	0.11	98	untreated
4	0.06	107	treated
18	0.11	115	untreated
5	0.11	123	treated
20	0.22	124	untreated
19	0.22	131	untreated
6	0.11	139	treated
21	0.56	144	untreated
8	0.22	152	treated
22	0.56	158	untreated
7	0.22	159	treated
23	1.10	160	untreated
9	0.56	191	treated
12	1.10	200	treated
10	0.56	201	treated
11	1.10	207	treated

```
yt<-velocidad_reaccion[Puromicina=="treated"]
xt<-concentracion[Puromicina=="treated"]
modelo1<-lm(yt~xt+I(xt^(1/2)))
xv<-seq(from=0,to=1.1,by=0.01)
yv<-predict(modelo1,list(xt=xv))
lines(xv,yv, col="red", lwd=2)

yt2<-velocidad_reaccion[Puromicina=="untreated"]
xt2<-concentracion[Puromicina=="untreated"]
modelo2<-lm(yt2~xt2+I(xt2^(1/2)))
```

```
xv2<-seq(from=0, to=1.1, by=0.01)
yv2<-predict(modelo2, list(xt2=xv2))
lines(xv2,yv2, col="#ff8033", lwd=2)
```




```
detach()
```

Apartado f)

```
datos_NA<-read.table("Archivos/Puromicina_NA.txt", dec = ".", se
p=",", header=T)
kable(datos_NA)
```

concentracion	velocidad_reaccion	Puromicina
0.02	76	treated
0.02	47	treated
0.06	97	treated
0.06	107	treated
NA	123	treated
0.11	139	treated
0.22	159	treated
0.22	152	treated
0.56	191	NA
0.56	201	treated

1.10	NA	treated
1.10	200	treated
0.02	67	untreated
0.02	51	untreated
0.06	84	untreated
0.06	86	untreated
0.11	98	untreated
0.11	115	NA
0.22	131	untreated
0.22	NA	untreated
0.56	144	untreated
0.56	NA	untreated
1.10	160	untreated

```
# Completa los NA con un valor centralizado.
datos_ajustados<-centralImputation(datos_NA)
kable(datos_ajustados)
```

concentracion	velocidad_reaccion	Puromicina
0.020	76	treated
0.020	47	treated
0.060	97	treated
0.060	107	treated
0.165	123	treated
0.110	139	treated
0.220	159	treated
0.220	152	treated
0.560	191	treated
0.560	201	treated
1.100	119	treated
1.100	200	treated
0.020	67	untreated
0.020	51	untreated
0.060	84	untreated
0.060	86	untreated
0.110	98	untreated
0.110	115	treated
0.220	131	untreated
0.220	119	untreated
0.560	144	untreated

0.560	119	untreated
1.100	160	untreated

```
# Método del vecino más cercano. Ajusta los NA al k más cercano a ellos.
datos_ajustados_vecino<-knnImputation(datos_NA)
kable(datos_ajustados_vecino)
```

concentracion	velocidad_reaccion	Puromicina
0.0200000	76.0000	treated
0.0200000	47.0000	treated
0.0600000	97.0000	treated
0.0600000	107.0000	treated
0.1506064	123.0000	treated
0.1100000	139.0000	treated
0.2200000	159.0000	treated
0.2200000	152.0000	treated
0.5600000	191.0000	treated
0.5600000	201.0000	treated
1.1000000	175.9734	treated
1.1000000	200.0000	treated
0.0200000	67.0000	untreated
0.0200000	51.0000	untreated
0.0600000	84.0000	untreated
0.0600000	86.0000	untreated
0.1100000	98.0000	untreated
0.1100000	115.0000	treated
0.2200000	131.0000	untreated
0.2200000	106.0601	untreated
0.5600000	144.0000	untreated
0.5600000	126.6490	untreated
1.1000000	160.0000	untreated

```
# Empleando NA.omit()
datos_ajustados_naomit<-na.omit(datos_NA)
kable(datos_ajustados_naomit)
```

	concentracion	velocidad_reaccion	Puromicina
1	0.02	76	treated
2	0.02	47	treated

3	0.06	97	treated
4	0.06	107	treated
6	0.11	139	treated
7	0.22	159	treated
8	0.22	152	treated
10	0.56	201	treated
12	1.10	200	treated
13	0.02	67	untreated
14	0.02	51	untreated
15	0.06	84	untreated
16	0.06	86	untreated
17	0.11	98	untreated
19	0.22	131	untreated
21	0.56	144	untreated
23	1.10	160	untreated

```
# Empleado complete.cases()
datos_ajustados_cc<-datos_NA[complete.cases(datos_NA)==TRUE,]
kable(datos_ajustados_cc)
```

	concentracion	velocidad_reaccion	Puromicina
1	0.02	76	treated
2	0.02	47	treated
3	0.06	97	treated
4	0.06	107	treated
6	0.11	139	treated
7	0.22	159	treated
8	0.22	152	treated
10	0.56	201	treated
12	1.10	200	treated
13	0.02	67	untreated
14	0.02	51	untreated
15	0.06	84	untreated
16	0.06	86	untreated
17	0.11	98	untreated
19	0.22	131	untreated
21	0.56	144	untreated
23	1.10	160	untreated

```
attach(datos_ajustados_naomit)
media_VR_naomit<-aggregate(velocidad_reaccion~Puromicina,datos_ajustados_naomit,mean)
kable(media_VR_naomit)
```

Puromicina	velocidad_reaccion
treated	130.8889
untreated	102.6250

```
mediana_VR_naomit<-aggregate(velocidad_reaccion~Puromicina,datos_ajustados_naomit,median)
kable(mediana_VR_naomit)
```

Puromicina	velocidad_reaccion
treated	139
untreated	92

```
detach()

attach(datos_ajustados_cc)
media_VR_cc<-aggregate(velocidad_reaccion~Puromicina,datos_ajustados_cc,mean)
kable(media_VR_cc)
```

Puromicina	velocidad_reaccion
treated	130.8889
untreated	102.6250

```
mediana_VR_cc<-aggregate(velocidad_reaccion~Puromicina,datos_ajustados_cc,median)
kable(mediana_VR_cc)
```

Puromicina	velocidad_reaccion
treated	139
untreated	92

```
detach()

#Creramos una tabl nueva con 0
datos_ceros<-datos_NA
datos_ceros$concentracion[is.na(datos_ceros$concentracion)]<-0
datos_ceros$velocidad_reaccion[is.na(datos_ceros$velocidad_reaccion)]<-0
kable(datos_ceros)
```

concentracion	velocidad_reaccion	Puromicina
0.02	76	treated
0.02	47	treated
0.06	97	treated
0.06	107	treated

0.00	123	treated
0.11	139	treated
0.22	159	treated
0.22	152	treated
0.56	191	NA
0.56	201	treated
1.10	0	treated
1.10	200	treated
0.02	67	untreated
0.02	51	untreated
0.06	84	untreated
0.06	86	untreated
0.11	98	untreated
0.11	115	NA
0.22	131	untreated
0.22	0	untreated
0.56	144	untreated
0.56	0	untreated
1.10	160	untreated

```
datos_ceros_knn<-knnImputation(datos_ceros)

# Resumenes de las tablas ajustadas
summary(datos_NA)
```

```
##  concentracion    velocidad_reaccion     Puromicina
##  Min.   :0.0200   Min.   : 47.0      treated  :11
##  1st Qu.:0.0600   1st Qu.: 85.5      untreated:10
##  Median :0.1650   Median :119.0      NA's     : 2
##  Mean   :0.3214   Mean   :121.4
##  3rd Qu.:0.5600   3rd Qu.:153.8
##  Max.   :1.1000   Max.   :201.0
##  NA's   :1        NA's   :3
```

```
summary(datos_ajustados)
```

```
##  concentracion    velocidad_reaccion     Puromicina
##  Min.   :0.0200   Min.   : 47.0      treated  :13
##  1st Qu.:0.0600   1st Qu.: 91.5      untreated:10
##  Median :0.1650   Median :119.0
##  Mean   :0.3146   Mean   :121.1
##  3rd Qu.:0.5600   3rd Qu.:148.0
##  Max.   :1.1000   Max.   :201.0
```

```
summary(datos_ajustados_vecino)
```

```
##  concentracion    velocidad_reaccion     Puromicina
##  Min.   :0.0200   Min.   : 47.0      treated  :13
##  1st Qu.:0.0600   1st Qu.: 91.5      untreated:10
##  Median :0.1506   Median :123.0
##  Mean   :0.3139   Mean   :123.3
##  3rd Qu.:0.5600   3rd Qu.:155.5
##  Max.   :1.1000   Max.   :201.0
```

```
summary(datos_ajustados_naomit)
```

```
##  concentracion    velocidad_reaccion     Puromicina
##  Min.   :0.0200   Min.   : 47.0      treated  :9
##  1st Qu.:0.0600   1st Qu.: 84.0      untreated:8
##  Median :0.1100   Median :107.0
##  Mean   :0.2659   Mean   :117.6
##  3rd Qu.:0.2200   3rd Qu.:152.0
##  Max.   :1.1000   Max.   :201.0
```

```
summary(datos_ajustados_cc)
```

```
##  concentracion    velocidad_reaccion     Puromicina
##  Min.   :0.0200   Min.   : 47.0      treated  :9
##  1st Qu.:0.0600   1st Qu.: 84.0      untreated:8
##  Median :0.1100   Median :107.0
##  Mean   :0.2659   Mean   :117.6
##  3rd Qu.:0.2200   3rd Qu.:152.0
##  Max.   :1.1000   Max.   :201.0
```

```
summary(datos_ceros_knn)
```

```
##  concentracion    velocidad_reaccion     Puromicina
##  Min.   :0.0000   Min.   :  0.0      treated  :13
##  1st Qu.:0.0600   1st Qu.: 71.5      untreated:10
##  Median :0.1100   Median :107.0
##  Mean   :0.3074   Mean   :105.6
##  3rd Qu.:0.5600   3rd Qu.:148.0
##  Max.   :1.1000   Max.   :201.0
```

Conclusiones

A priori, aunque las diferencias en el resultado final no inciden de manera importante, sí que es verdad que se ven cambios aparentes, sobretodo en las medias y cuartiles.

Esto se acentúa aún más cuando se aplica el trato de los NA, puesto que en el caso de las omisiones, el 3er cuartil aumenta respecto a sus homónimos, así como la media de concentración, debido a la omisión de muestras.

Si se aceptan los valores como ceros, el cambio es mucho más sutil debido a la persistencia de los valores como registros, lo cual afecta tanto a la media como a los cuartiles.

En esencia, la elección de estas opciones debe ajustarse ante todo a los usos que se vayan a dar a las diferentes muestras desde el punto de vista estadístico y su validación a la hora de realizar depreciaciones en los desbalances de los resultados finales.

Ejemplo de lo anterior son por ejemplo los estudios sobre grandes masa de población o a niveles generalistas

Ejercicio 3.3

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 ULPG decide adquirir 150 de estos equipos. Se pide:

a) Analizar el tipo de función de probabilidad subyacente y explicar sus características.
b) Calcular la probabilidad de que el número de averías sea de 5.
c) Encontrar la probabilidad de que el número de averías sea mayor o igual a 3.
d) ¿Qué valor de la variable deja por debajo de sí el 75% de la probabilidad?
e) Encontrar el número mínimo n tal que la probabilidad de que el número de averías sea superior a 0.99
f) Calcular el percentil 95% de la distribución.
g) Obtener una muestra de tamaño 1000 de esta distribución.
h) Representar gráficamente la muestra de g) mediante un diagrama de barras y comparar éste con las frecuencias esperadas según el modelo que genera los datos.

```
setwd(".")
library(knitr)
library(ggplot2)
library(sqldf)
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Loading required package: RSQLite
```

```
library(DMwR2)
```

```
## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
#Binomial básica
n<-150
p<-0.02
mu<-n*p
sigma<-sqrt(n*p*(1-p))
```

Apartado a)

Analizar el tipo de función subyacente.

Apartado b)

```r
#Calcular la prob de que el numero de averías sea 5.
prob5<-dbinom(5,n,p)
```

Apartado c)

```r
#Encontrar la prob de que el numero de averías sea 3 o mas.
#Opción A con dbinom
prob3omas_db<-1-(dbinom(0,n,p)+dbinom(1,n,p)+dbinom(2,n,p))
#Opción B con pbinom
prob3omas_pb<-1-(pbinom(2,n,p))
```

Apartado d)

```r
#Tercer cuartil.
qr3<-qbinom(0.75,n,p)
```

Apartado e)

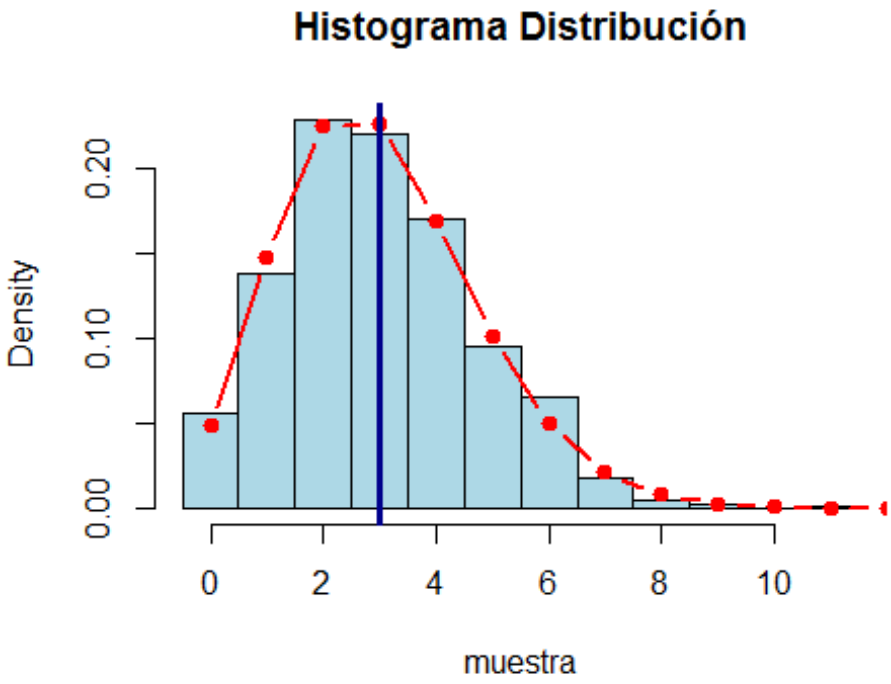
```r
#Numero minimo para que la probabilidad sea superior al 99%.
averiassup099<-log(0.99)/log(0.02)
```

Apartado f)

```r
#Calcular el percentil 95 de la distribución.
per95<-qbinom(0.95,n,p)
```

Apartado g) y h)

```r
#Obtener una muestra de tamaño 1000 de la distribución.
set.seed(35200)
muestra<-rbinom(1000,n,p)
media_muestra<-mean(muestra)
sd_muestra<-sd(muestra)
hist(muestra, breaks=seq(-0.5,max(muestra+0.5)), col="lightblue"
, freq=F,
     main="Histograma Distribución")

#Representar gráficamente el apartado G.
x<-seq(0,max(muestra)+1)
fx<-dbinom(x,n,p)
points(x,fx, type="b", col="red", lwd=2, pch=19)
abline(v=mu, col="darkblue", lwd=3)
```



Conclusiones

Se puede apreciar que, al ser una distribución normal, tanto las frecuencias esperadas como el propio diagrama de barras coinciden en su representación.

Esto tiene sentido debido a que la probabilidad de averías irá descendiendo gradualmente hacia 0 debido a que n crece hacia infinito, mientras que si mantenemos n en un intervalo realista, las posibilidades subyacentes de avería aumentan a menor sea el número de ítems testeados.

Ejercicio 3.4

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.
b) Calcular la probabilidad de sus valores mayores o iguales a 1 y menores o iguales a 6.
c) Obtener el percentil 75 de la distribución.
d) ¿Qué valor es el que deja por debajo de sí el 5% de los valores más bajos de la variable?
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.
f) Explicar la influencia del parámetro lambda en la distribución y visualizar los diferentes resultados superpuestos.

```
setwd(".")
library(knitr)
library(ggplot2)
library(sqldf)

## Loading required package: gsubfn

## Loading required package: proto

## Loading required package: RSQLite

library(DMwR2)

## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

Apartado a)

```
#Calcular la probabilidad de que sea mayor o igual a 5.5
lambda<-3
p55<-ppois(5.5,lambda)
mu<-lambda
sigma<-sqrt(lambda)
```

Apartado b)

```
px1y6_con1<-ppois(6,lambda)-ppois(0,lambda)
px1y6_sin1<-ppois(6,lambda)-ppois(1,lambda)
```

Apartado c)

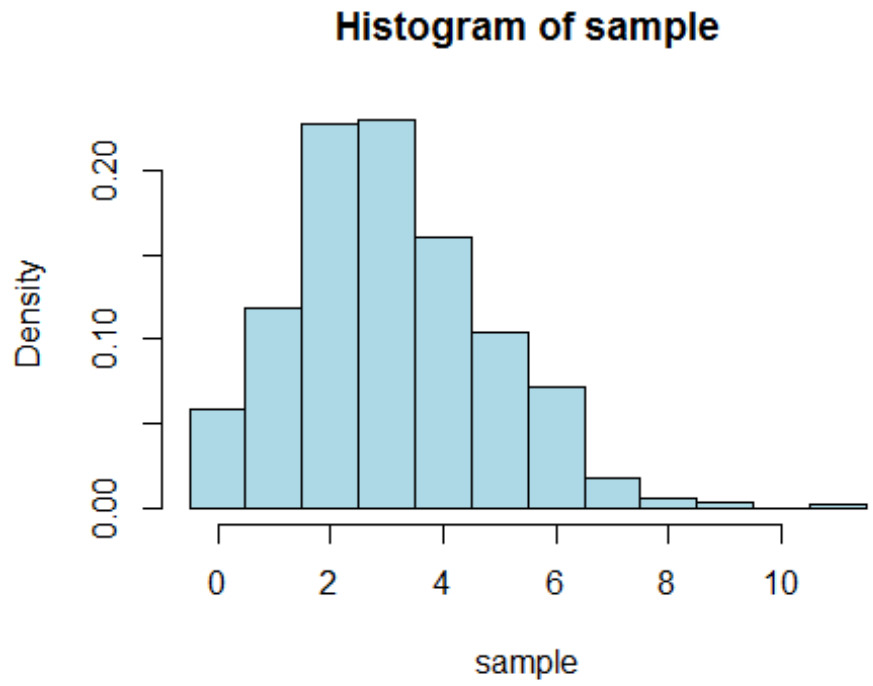
```r
#Calcular el 5% de los valores mas bajos.
per75<-qpois(0.75,lambda)
```

Apartado d)

```r
#Calcular el 5% de los valores mas bajos.
per5<-qpois(0.05,lambda)
```

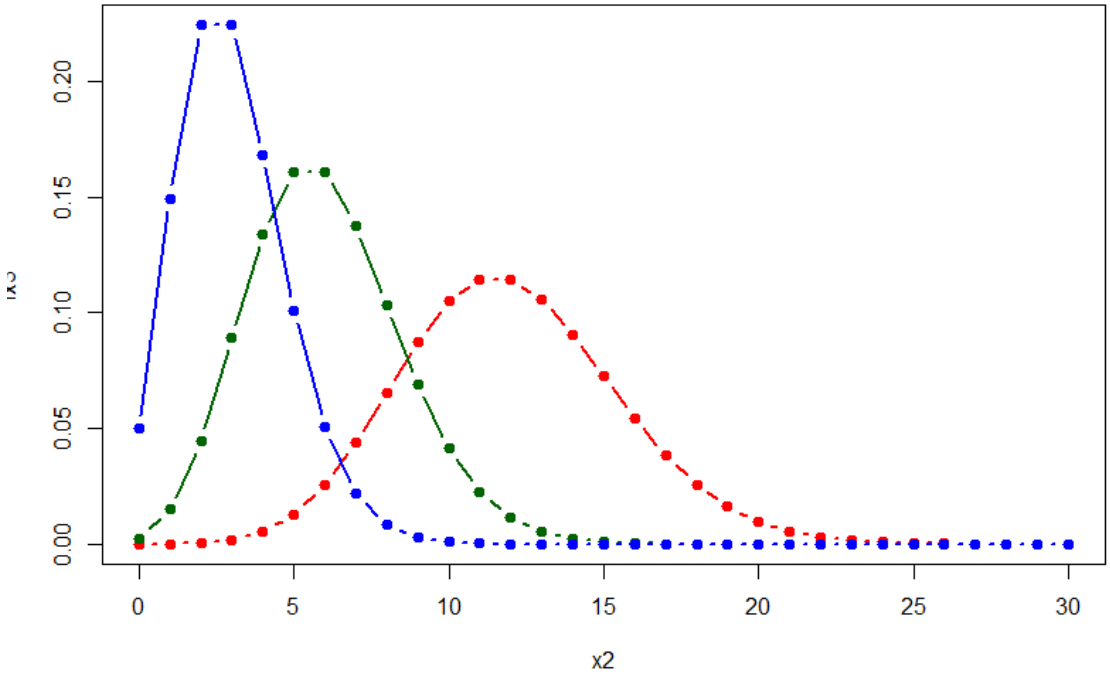
Apartado e)

```r
set.seed(35200)
sample<-rpois(500,3)
media_sample<-mean(sample)
sd_sample<-sd(sample)
hist(sample, breaks=seq(-0.5,max(sample+0.5)), col="lightblue",
freq=F, add=F)
```



Apartado f)

```r
#Pillamos lambda por defect (3).
x<-seq(0,max(sample)+1)

x2<-seq(0,30)
fx<-dpois(x2,lambda)
fx2<-dpois(x2,2*lambda)
fx3<-dpois(x2,4*lambda)
plot(x2,fx3, col="red",type="b", lwd=2, pch=19)
points(x2,fx2, type="b", col="darkgreen", lwd=2, pch=19)
points(x2,fx, type="b", col="blue", lwd=2, pch=19)
```



Conclusiones

La influencia del parámetro lambda es clara, y se puede verificar en la gráfica como su modificación aumenta o disminuye la distribución de Poisson resultante, aumentando su pico y reduciendo su anchura.

Como se observa, un incremento o multiplicación de lambda por un escalar genera una distribución mucho más cerrada y ajustada hacia el origen mientras que una lamba menor genera una distribución más achatada y amplia, luego se corrobora un relación de interdependencia, explicada en la forma de Poisson.

Ejercicio 3.5

Ejercicio 5: Consideremos una variable aleatoria ***W*** con distribución ***N*** (200, 25). Se pide:

a) ***P*** [150 < ***W*** ≤ 250]
b) ***P*** [***W*** ≥ 255].
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?
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.
e) Obtener y visualizar la función de distribución acumulada y situar sobre ella los resultados de a) y b)
f) Calcular los coeficientes que definen los factores de forma de la distribución (Curtosis y Asimetría). Razonar las respuestas.

```
setwd(".")
library(knitr)
library(ggplot2)
library(sqldf)
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Loading required package: RSQLite
```

```
library(DMwR2)
```

```
## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
library(e1071)
```

Apartado a)

```
mu<-200
sigma<-250
Wdcha<-pnorm(250,200,25)
Wizq<-pnorm(200,200,25)
direfer<-Wdcha-Wizq
```

Apartado b)

```
Px255<-1-pnorm(255,200,25)
```

Apartado c)

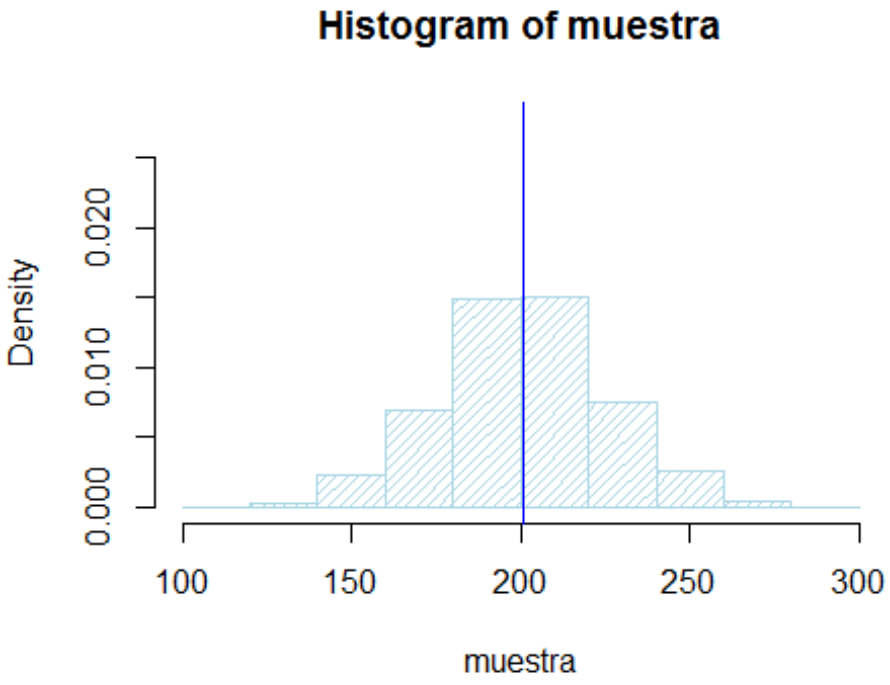
Percentil 95*percentil 5

```
p05<-qnorm(0.05,200,25)
p95<-qnorm(0.95,200,25)
```

Apartado d)

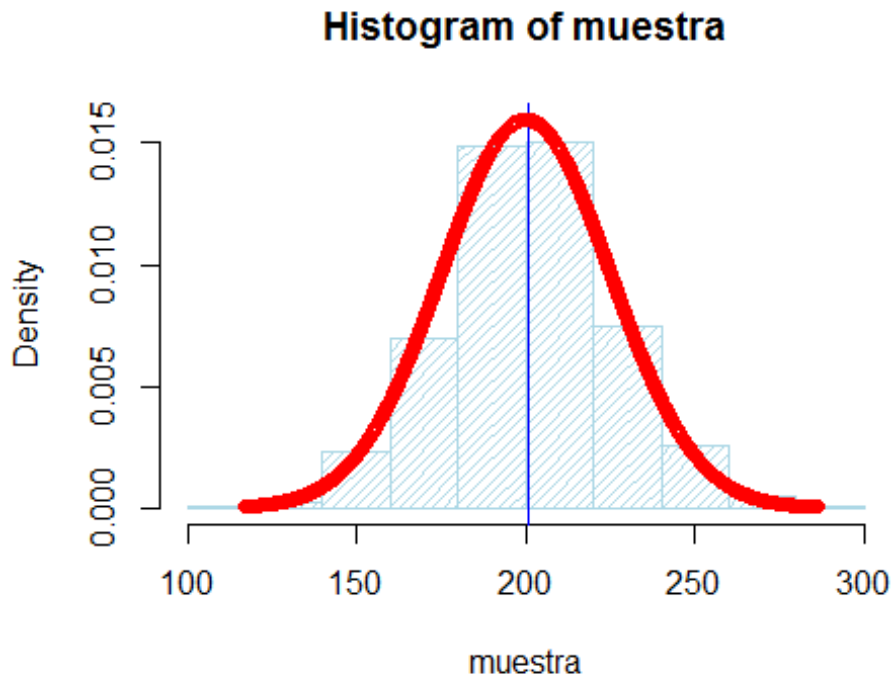
Generamos una muestra de 34600

```
set.seed(34600)
muestra<-rnorm(1000,200,25)
hist(muestra, freq=F, col="lightblue", density=25, ylim=c(0,2*ma
x(Px255)))
abline(v=mean(muestra), col="blue",lwd=1)
```

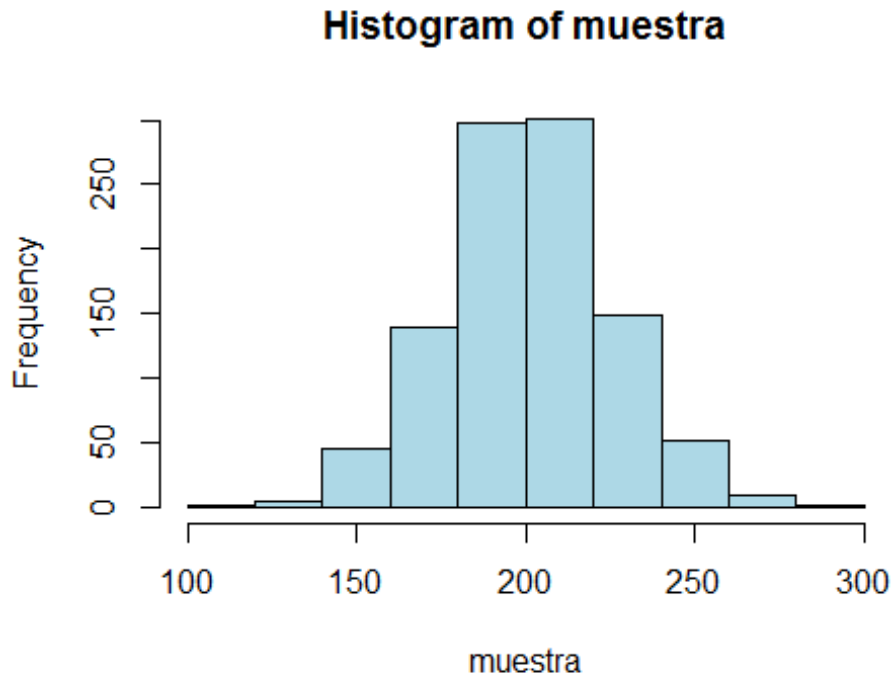


Apartado e)

```
x<-seq(min(muestra),max(muestra),0.1)
fx<-dnorm(x,200,25)
hist(muestra, freq=F, col="lightblue", density=25, ylim=c(0,max(
fx)))
abline(v=mean(muestra), col="blue",lwd=1)
points(x,fx,col="red",lwd=0.5)
```

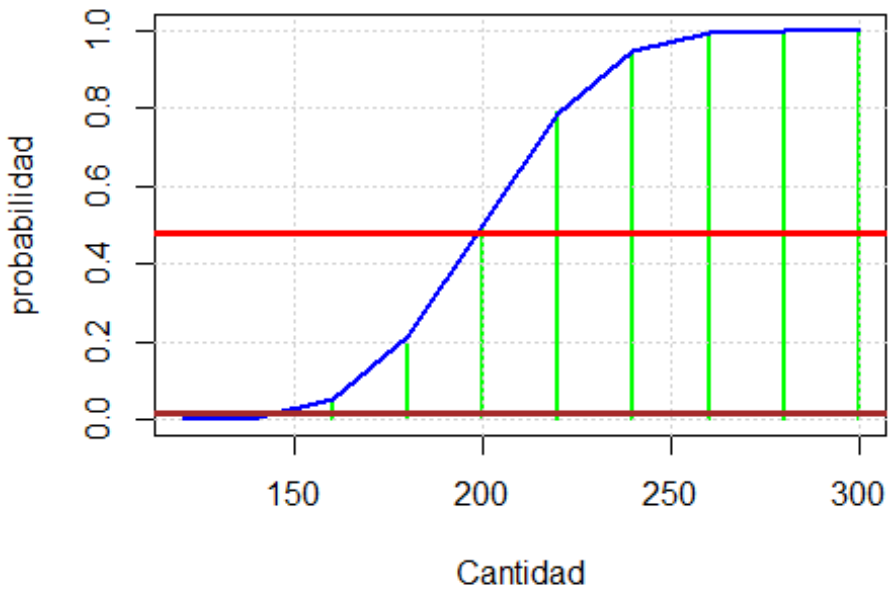



```
fhist<-hist(muestra, col="lightblue")
```




```
f_x<-cumsum(fhist$counts)/sum(fhist$counts)
x_ac<-fhist$breaks[2:length(fhist$breaks)]
plot(x_ac,f_x, type="h", lwd=2, col="green",
    xlab="Cantidad", ylab="probabilidad")
grid()
```

```
f_xt<-pnorm(x_ac,200,25)
points(x_ac,f_xt,type="l",col="blue", lwd=2)
abline(h=direfer,col="red",lwd=3)
abline(h=Px255,col="brown", lwd=3)
```



Apartado f)

```
skewness(muestra)

## [1] 0.08306719

kurtosis(muestra)

## [1] 0.1653612
```

Conclusiones

Se puede observar que la distribución tiene una asimetría muy ligeramente positiva y un leve aplastamiento, pudiendo clasificarse como mesocúrtica.

Esto tiene sentido, debido a que durante el graficado y desarrollo de los cálculos realizados hemos ido desarrollando una función que toma la forma de una binomial cuasi normal, luego la dispersión y aplastamiento iban a ser, por lógica, muy ligeros o incluso nulos.

Ejercicio 4.1

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 (en formato GIF o similar) de la evolución de las pirámides de población en esos años.

Para obtener más fácilmente los datos se puede ir directamente a la dirección: (link) y consultar los datos de "población según sexos y grupos de edad grandes y quinquenales".

(Animación final adjuntada en la entrega)

```
setwd(".")
library(knitr)
library(ggplot2)
library(sqldf)
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Loading required package: RSQLite
```

```
library(DMwR2)
```

```
## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo
```

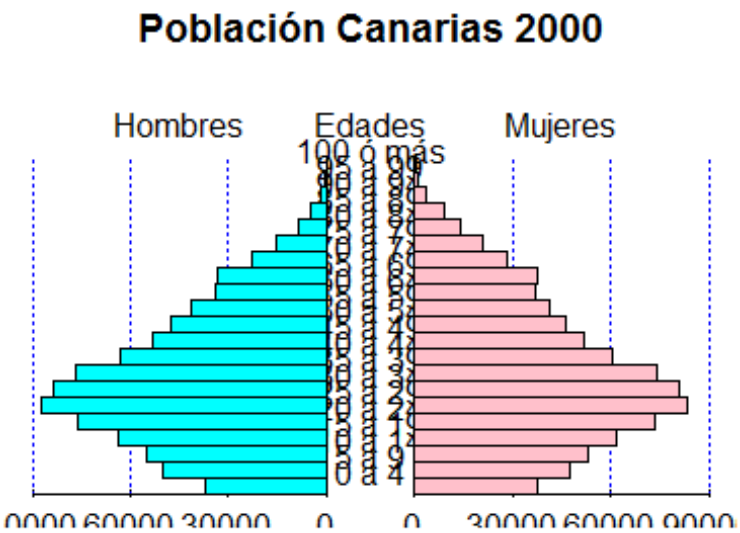
```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

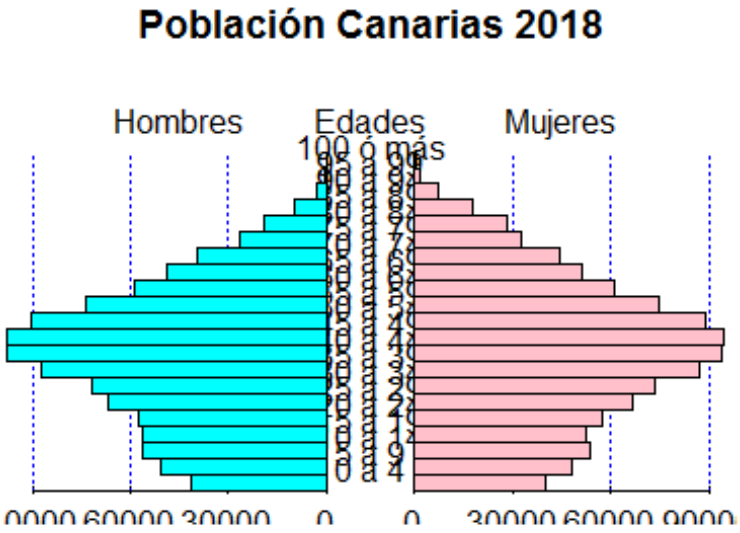
```
library(pyramid)
library(readxl)
```

Apartado a)

```
poblacion <- read_excel("Archivos/E30260A_0023.xls", col_types =
c("numeric", "numeric", "text", "text"))
attach(poblacion)

for(i in 2000:2018){
  pyramid(as.data.frame(poblacion[Year==as.character((i)),1:3]),
          Llab = "Hombres",
          Rlab="Mujeres",
          Clab = "Edades",
          AxisFM = "d",
          Laxis = seq(0,100000,30000),
          Raxis = seq(0,100000,30000),
```

```
          main= paste("Población Canarias",as.character((i))))
}
```





Ejercicio 4.3

Ejercicio 3. El fichero "***germinacion.csv***" contiene los datos de germinación de semillas de dos genotipos de la planta parásita Orobanche y dos extractos de plantas huésped (judia 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:

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.
b) Calcular los parámetros de centralización y dispersión del conjunto de muestras para cada genotipo y tipo de planta huésped y analizar gráficamente el efecto del genotipo en la germinación. Explicar las conclusiones.
c) Utilizar la función ***lm()*** para ver la tendencia e influencia de los genotipos en la germinación. ¿Son estadísticamente independientes las variables de genotipo ("*Orobanche*") y de tipo de planta huésped ("*extract*")? Razonar y justificar las respuestas.

```
setwd(".")
library(knitr)
library(ggplot2)
germinacion<-read.table("Archivos/germination.csv", header=T, se
p=",")
attach(germinacion)
```

Apartado a)

```
n_germinadas<- sample - count
p_n_germinadas<- 100*(n_germinadas/sample)
germinacion<-cbind(germinacion,n_germinadas,p_n_germinadas)
attach(germinacion)

## The following objects are masked _by_ .GlobalEnv:
##
##     n_germinadas, p_n_germinadas

## The following objects are masked from germinacion (pos = 3):
##
##     count, extract, Orobanche, sample

kable(germinacion[1:10,])
```

count	sample	Orobanche	extract	n_germinadas	p_n_germinadas
10	39	a75	judia	29	74.35897
23	62	a75	judia	39	62.90323
23	81	a75	judia	58	71.60494
26	51	a75	judia	25	49.01961
17	39	a75	judia	22	56.41026
5	6	a75	pepino	1	16.66667

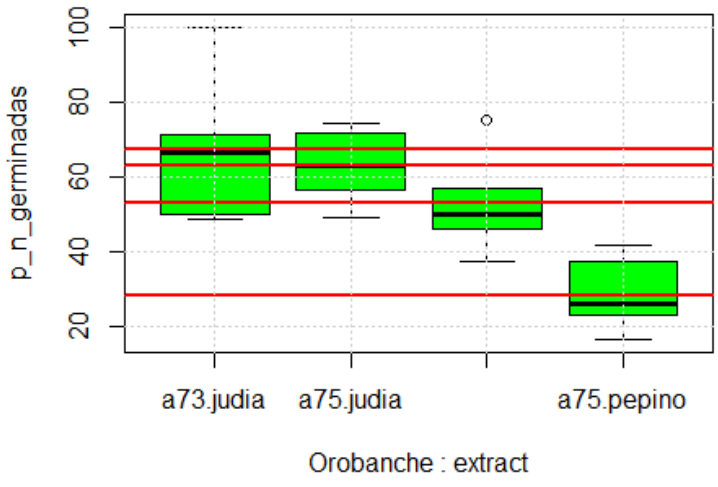
53	74	a75	pepino	21	28.37838
55	72	a75	pepino	17	23.61111
32	51	a75	pepino	19	37.25490
46	79	a75	pepino	33	41.77215

Apartado b)

```
medias_n_g<-aggregate(p_n_germinadas~Orobanche+extract, germinac
ion, mean)
kable(medias_n_g)
```

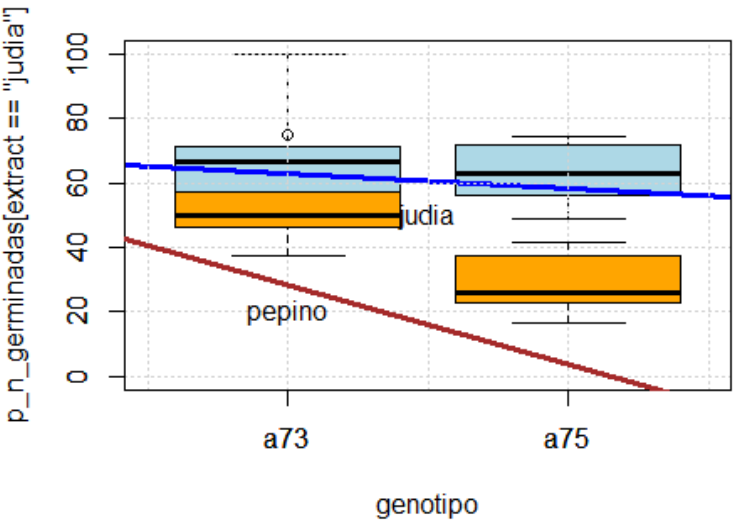
Orobanche	extract	p_n_germinadas
a73	judia	67.39683
a75	judia	62.85940
a73	pepino	53.14784
a75	pepino	28.46002

```
boxplot(p_n_germinadas~Orobanche+extract,col="green")

#A73 judia
abline(h = medias_n_g[1,3], col="red", lwd="2")

#A75 judia
abline(h = medias_n_g[2,3], col="red", lwd="2")

#A73 pepino
abline(h = medias_n_g[3,3], col="red", lwd="2")

#A75 pepino
abline(h = medias_n_g[4,3], col="red", lwd="2")
grid()
```



Apartado c)

```
#Judias
boxplot(p_n_germinadas[extract=="judia"]~Orobanche[extract=="judia"],
        col="lightblue", ylim=c(0,100), xlab="genotipo")

modelo1<-lm(p_n_germinadas[extract=="judia"]~Orobanche[extract=="judia"])
abline(modelo1, col="blue", lwd=3)
text(1.5,50,labels="judia")
grid()

#Pepinos
boxplot(p_n_germinadas[extract=="pepino"]~Orobanche[extract=="pepino"],
        col="orange", ylim=c(0,100), xlab="genotipo", add=T)
modelo1<-lm(p_n_germinadas[extract=="pepino"]~Orobanche[extract=="pepino"])
abline(modelo1, col="brown", lwd=3)
text(1,20,labels="pepino")
```



Ejercicio 4.4

Ejercicio 4. Obtener del Instituto Canario de Estadística (ISTAC) el fichero con los datos de empleo en actividades relacionadas con la Ingeniería Informática (códigos de rama de actividad CNAE-09: 62,63 y 95) en Canarias por trimestres en el periodo 2009 a 2018. (ojo con el formato de los datos y los separadores de decimales). Se pide

a) Analizar gráficamente la variación de cada tipología de empleo en las Islas Canarias (por islas y totales) en el periodo considerado e intentar explicar sus valores singulares.
b) Utilizando la librería ***mgcv***(), encontrar un modelo de seguimiento del empleo representar gráficamente su evolución y predicciones (efectuar una posible a diciembre de 2018).
c) Encontrar la isla donde hay más empleo y en qué etapa.
d) Analizar comparativamente la evolución durante dos años del empleo en dos islas diferentes y explicar sus variaciones y sus aspectos comunes.

```
setwd(".")
library(knitr)
library(ggplot2)
library(mgcv)
```

```
## Loading required package: nlme
```

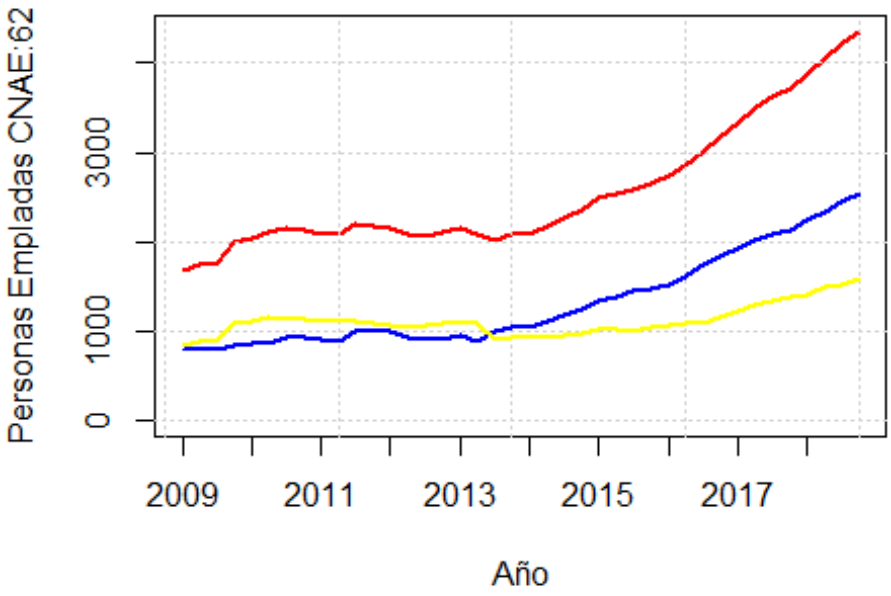
```
## This is mgcv 1.8-28. For overview type 'help("mgcv-package")'
.
```

```
empleo<-read.csv("Archivos/Empleos_Informatica_Canarias_2009-18.
csv", sep=";")
attach(empleo)
```

Apartado a)

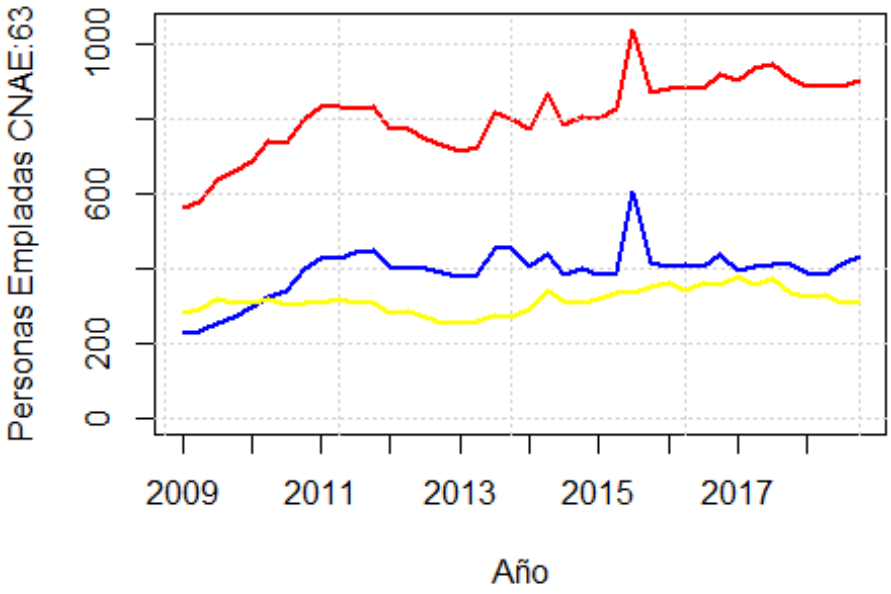
```
#------------CAN 62------------#
CAN_62<-empleo[TRIMESTRES=="CNAE_62",2]
index<-seq(length(CAN_62),1,-1)
CAN_62_t<-CAN_62[index]

plot(1:length(CAN_62_t), CAN_62_t, xaxt="n", type="l",
     xlab="Año", ylab="Personas Empladas CNAE:62",
     ylim = c(0,max(CAN_62_t)),col="red", lwd=2)
years<-c("2009","2010","2011","2012","2013","2014","2015",
         "2016","2017","2018")
axis(side=1, at = seq(1,length(CAN_62),4),
     labels = years)
grid()

#GRAFICAS DE EMPLEO PARA GC Y TF
TFE_62<-TENERIFE[TRIMESTRES=="CNAE_62"]
GCA_62<-GRAN.CANARIA[TRIMESTRES=="CNAE_62"]
```

```
#PUNTOS PARA CADA UNA DE LAS ANTERIORES
points(1:length(CAN_62),TFE_62[index], type="l", col="blue",lwd=
2)
points(1:length(CAN_62),GCA_62[index], type="l", col="yellow",lw
d=2)
```

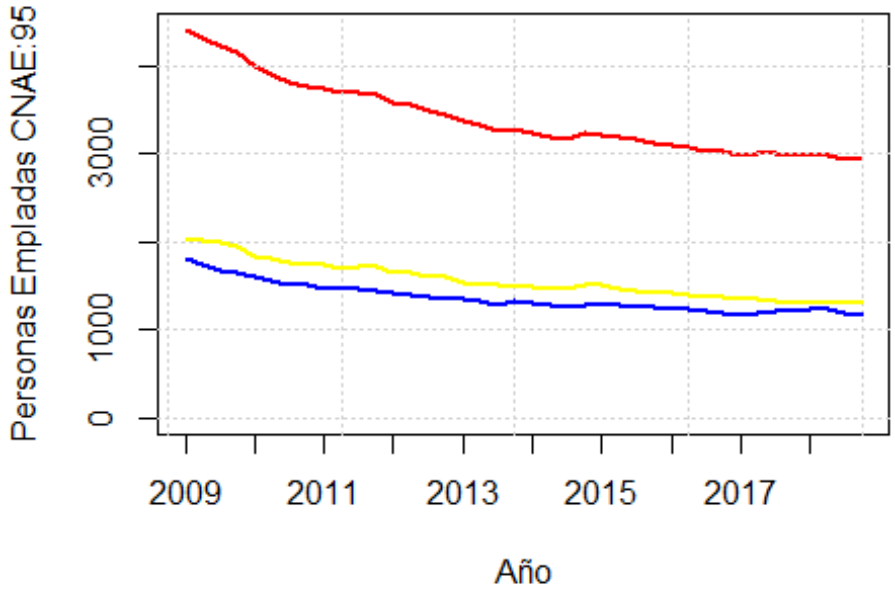



```
#------------CAN 63------------#
CAN_63<-empleo[TRIMESTRES=="CNAE_63",2]
index<-seq(length(CAN_63),1,-1)
CAN_63_t<-CAN_63[index]

plot(1:length(CAN_63_t), CAN_63_t, xaxt="n", type="l",
     xlab="Año", ylab="Personas Empladas CNAE:63",
     ylim = c(0,max(CAN_63_t)),col="red", lwd=2)
years<-c("2009","2010","2011","2012","2013","2014","2015",
         "2016","2017","2018")
axis(side=1, at = seq(1,length(CAN_63),4),
     labels = years)
grid()

#GRAFICAS DE EMPLEO PARA GC Y TF
TFE_63<-TENERIFE[TRIMESTRES=="CNAE_63"]
GCA_63<-GRAN.CANARIA[TRIMESTRES=="CNAE_63"]

#PUNTOS PARA CADA UNA DE LAS ANTERIORES
points(1:length(CAN_63),TFE_63[index], type="l", col="blue",lwd=
2)
```

```
points(1:length(CAN_63),GCA_63[index], type="l", col="yellow",lw
d=2)
```



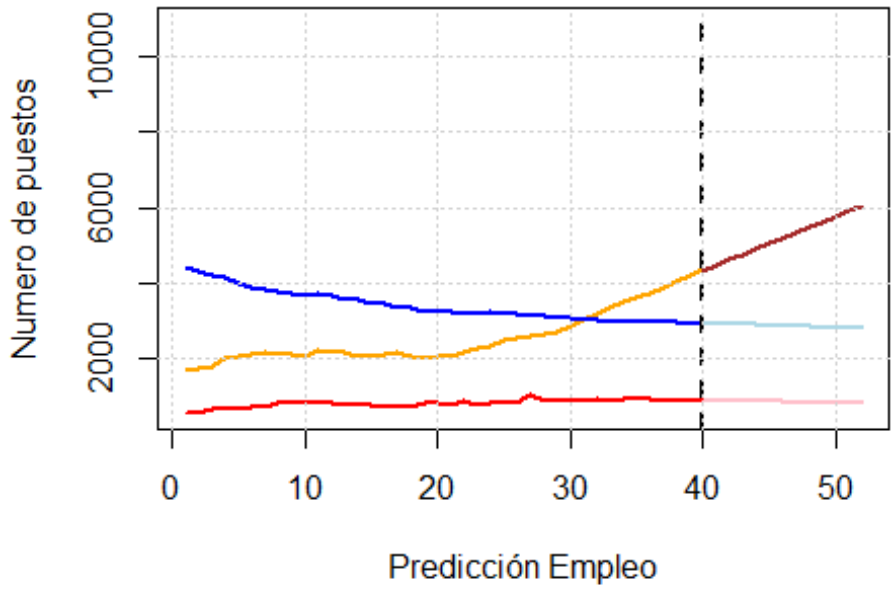

```
#------------CAN 95------------#
CAN_95<-empleo[TRIMESTRES=="CNAE_95",2]
index<-seq(length(CAN_95),1,-1)
CAN_95_t<-CAN_95[index]

plot(1:length(CAN_95_t), CAN_95_t, xaxt="n", type="l",
     xlab="Año", ylab="Personas Empladas CNAE:95",
     ylim = c(0,max(CAN_95_t)),col="red", lwd=2)
years<-c("2009","2010","2011","2012","2013","2014","2015",
         "2016","2017","2018")
axis(side=1, at = seq(1,length(CAN_95),4),
     labels = years)
grid()

#GRAFICAS DE EMPLEO PARA GC Y TF
TFE_95<-TENERIFE[TRIMESTRES=="CNAE_95"]
GCA_95<-GRAN.CANARIA[TRIMESTRES=="CNAE_95"]

#PUNTOS PARA CADA UNA DE LAS ANTERIORES
points(1:length(CAN_95),TFE_95[index], type="l", col="blue",lwd=
2)
points(1:length(CAN_95),GCA_95[index], type="l", col="yellow",lw
d=2)
```



Apartado b)

```
#----------------- Grafica inicial ----------------#
y63<-CAN_63_t
x63<-seq(1,40)
y62<-CAN_62_t
x62<-seq(1,40)
y95<-CAN_95_t
x95<-seq(1,40)

plot(x63,y63, xlim=c(1,52), ylim=c(min(y63),2.5*max(y62)),
     xlab="Predicción Empleo", ylab ="Numero de puestos", col="r
ed", type="l", lwd=2)
points(x62,y62, type="l", col="orange", lwd=2)
points(x95,y95, type="l", col="blue", lwd=2)


#-----------------CAN 62-----------------#
x<-x62
y<-y62
modelogam62<-gam(y~s(x))
xv62<-(40:52)
yv62<-predict(modelogam62, list(x=xv62))
points(xv62,yv62, type="l", col="brown", lwd=2)


#-----------------CAN 63-----------------#
x<-x63
```

```
y<-y63
modelogam63<-gam(y~s(x))
xv63<-(40:52)
yv63<-predict(modelogam63,list(x=xv63))
points(xv63,yv63, type="l", col="pink", lwd=2)


#-----------------CAN 95-----------------#
x<-x95
y<-y95
modelogam95<-gam(y~s(x))
xv95<-(40:52)
yv95<-predict(modelogam95,list(x=xv95))
points(xv95,yv95, type="l", col="lightblue", lwd=2)

#-----------------FINAL -----------------#

abline(v=40, col="black", lwd=2, lty=2)
grid()
```



Conclusiones

Se pueden esclarecer las siguientes conclusiones:

- Se ha notado un aumento de empleo en las dos islas mayores a lo largo de los años, si bien el incremento ha sido variable en función al tipo de sector involucrado.

- Dicho crecimiento se experimenta de forma notoria en el caso de las actividades de servicios y primera categoría.
- Las actividades de bajo nivel formativo, como las asistenciales han visto reducido su número de personas contratadas, o bien porque esos servicios ya no tienen demanda o porque han sido derivados a otros sectores más prominentes.

Entre islas, es importante destacar la cuasi total absorción de la mayoría de empleados por parte de Gran Canaria y Tenerife.

- En el caso de Tenerife resalta el hecho de que es la isla con mayor empleabilidad por en el sector, debido en gran parte al traslado de competencias gubernamentales y a la fuerte inversión en IT realizada por el Cabildo en la última década.
- Gran Canaria, pese a ir por detrás, ha experimentado también una mejoría en los últimos años, aumentando su plantel informático de forma considerable.

CUESTIONES LECTURAS

Cuestión 4.1

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.

En primer lugar extremos μ

μ = E(x) = 7 * 1/12 + 9 * 1/12 + 11 * 1/4 + 13 * 1/4 +15 * 1/6 + 17 * 1/6 = 12.666€

Podemos concluir que, gracias a que la mayor aportación a la ganancia esperada, que proviene del rango entre los programas de 11 y 13 € (para ser exactos, una aportación del 50%), se gana un total de 12.666 €

b) Establecer una medida coherente de su variabilidad

Por definición, podemos establecer dos medidas distintas: podemos aplicar una varianza (σ^2= E(x^2) - μ^2) o una desviación típica (S = σ).

Partiendo de μ = 12.6 €, proponemos ambas fórmulas:

E(x^2) = 49/12 + 81/12 + 121/4 + 169/4 + 225/6 + 289/6 = 65/6 + 145/2 + 257/3 = 169 €

σ^2 = 169 - (12.6)^2 = 169 - 158.76 = 10.24

S = σ = 3.2

Para la elección escogeremos en este caso, la desviación típica, debido a que trabaja en una dimensión equivalente a la de las medidas de los elementos de la tabla (en euros)

Teniendo en cuenta que la varianza es un valor estadístico que indica **cómo puede variar o no una muestra** escogeremos la desviación estándar o desviación típica (σ), pues establece un resultado general que, si bien puede ser positivo o negativo, se encuentra en la misma unidad de media que el cómputo general.

Por ello se puede decir que la ganancia esperada por el operario en este tramo sería de una 3,2€

Cuestión 4.2

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:

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

a) Verificar que la función de densidad anterior es válida
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?
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?
d) Contrastar con ***R*** los resultados y visualizar gráficamente los apartados b) y c).

```
setwd(".")
library(knitr)
```

Apartado a)

```
interval<-0.01
x<-seq(from=-1,to=2,by=interval)
f_X<-function(x1){
  f<-rep(0,length(x1))
  for(i in 1:length(x1)) {
    if((x1[i]<=0)|(x1[i]>=1)) {
      fX=0
    } else {
      fX=5*((1-x1[i])*(1-x1[i])*(1-x1[i])*(1-x1[i]))
    }
    f[i]=fX
  }
  return(f)
}
plot(x,f_X(x), col="blue", type="l", lwd = 3,
     ylab = "función de densidad f(x)")
grid()
min(f_X(x))
```

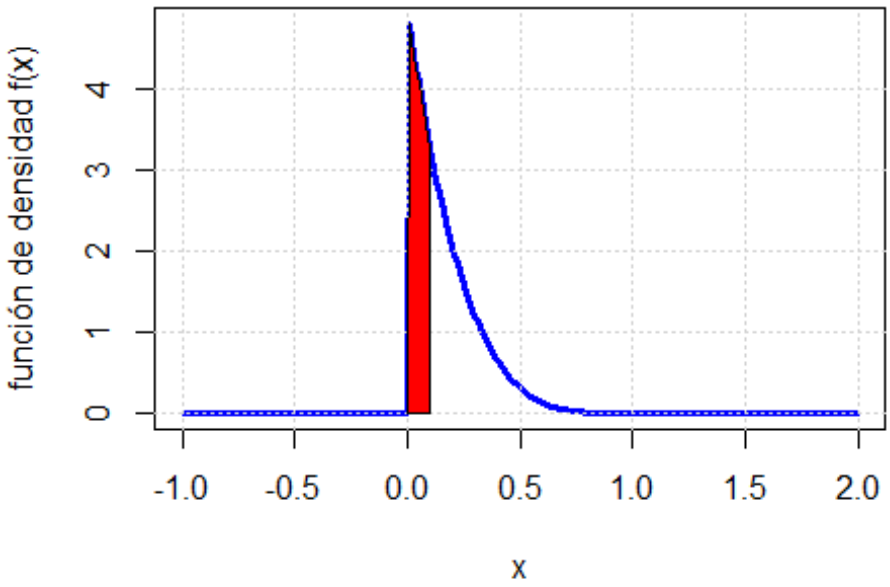
```
## [1] 0
```

```
prob_1<-integrate(f_X,-1,2)
prob_1
```

```
## 1 with absolute error < 6.6e-05
```

Apartado b)

```
x_1 <- 0
x_2 <- 0.1
p <- f_X(seq(x_1,x_2,interval))
z <- c(x_1,seq(x_1,x_2,interval),x_2)
p <- c(0,p,0)
polygon(z,p,col = "red")
```

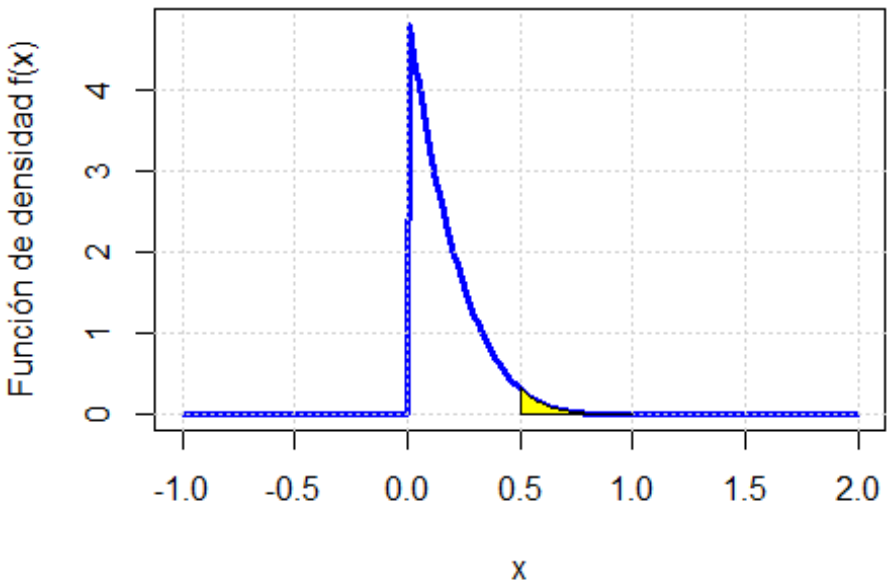



```
integrate(f_X,x_1,x_2)
```

```
## 0.40951 with absolute error < 4.5e-15
```

Apartado c)

```
plot(x,f_X(x), col="blue", type="l", lwd = 3,
     ylab = "Función de densidad f(x)")
grid()
x_3 <- 0.5
x_4 <- 1
p <- f_X(seq(x_3,x_4,interval))
z <- c(x_3,seq(x_3,x_4,interval),x_4)
p <- c(0,p,0)
polygon(z,p,col = "yellow")
```




```
integrate(f_X,x_3,x_4)
```

```
## 0.03125 with absolute error < 3.5e-16
```

Cuestión 4.3

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:

a) Exactamente 4 comenzaron a consumir tranquilizantes por problemas psicológicos.
b) Al menos 6 comenzaron a consumir tranquilizantes por problemas psicológicos
c) Analizar la distribución de probabilidad subyacente y sus características principales.

```
setwd(".")
library(knitr)

n <- 10
p <- 0.45
```

Apartado a)

```
p4 <- dbinom(4,n,p)
```

Apartado b)

```
p6<- 1-pbinom(5,n,p)
```

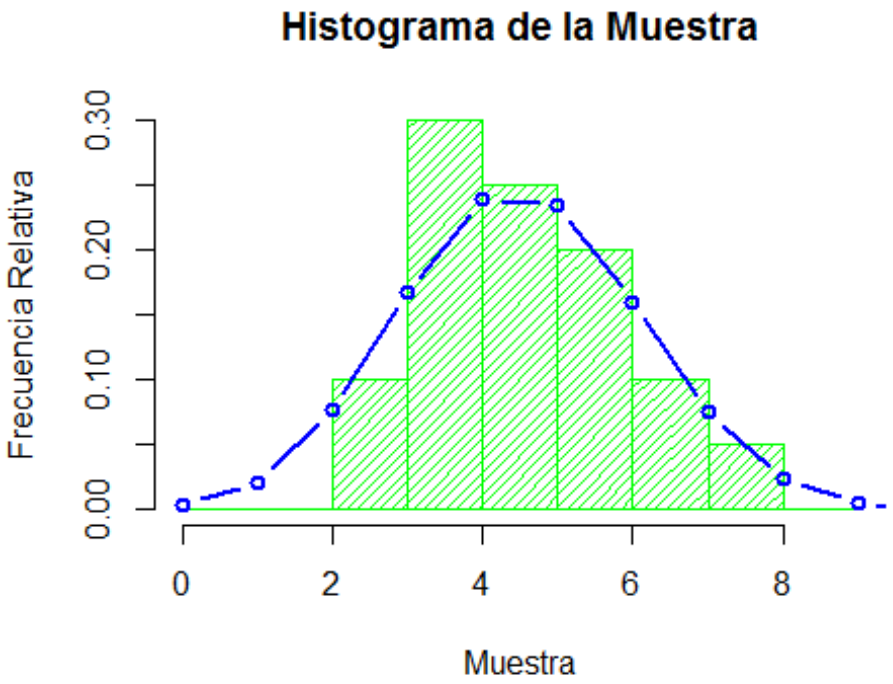
Apartado c)

Se puede asumir que la distribución que puede tomar esta forma es de tipo binomial debido a la aleatoriedad de la muestra, el conocimiento de las probabilidades de pertenecía al subgrupo de medición y el hecho de que la probabilidad de partencia desciende hacia los extremos de la distribución.

Apartado d)

Simulado a 15000 estudiantes.

```
set.seed(15000)
muestra <- rbinom(20,n,p)
hist(muestra,breaks=seq(0,max(muestra)+1),col = "green",density
= 25,freq = FALSE,
     main = "Histograma de la Muestra",xlab = "Muestra",ylab = "
Frecuencia Relativa")
x <- seq(0,10,1)
fx <- dbinom(x,n,p)
points(x,fx,type = "b",col = "blue",lwd = 2)
```


Histograma de la Muestra
Frecuencia Relativa
0.00
0.10
0.20
0.30
0
2
4
6
8
Muestra

Cuestión 4.4

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:

a) Calcular la probabilidad de que lleguen mas de 10 clientes en un periodo de una hora.
b) Calcular la probabilidad de que en veinte minutos lleguen menos de 5 clientes.
c) Cual es el número medio de llegadas en un periodo de dos horas.
d) Comprobar con ***R*** los resultados anteriores y mostrar gráficamente las funciones de distribución de probabilidad correspondientes.

```
setwd(".")
library(knitr)
```

Distribución de Poisson ya que el número de resultados que ocurren en un intervalo o región específica es independiente del número que ocurre en cualquier otro intervalo de tiempo.

Del enunciado vemos que el promedio es de 5 clientes cada 20 minutos,por tanto, ??t=5

Apartado a)

```
# P(x > 10) = 1 - P(x=<10)
# x=10
# ??t=15 ya que 60/20=3
ppois(10,15)
```

```
## [1] 0.1184644
```

```
1-ppois(10,15)
```

```
## [1] 0.8815356
```

Apartado b)

```
# Como solo nos dice menor que 5 incluimos el valor nulo (ningún cliente)
# P(x<5) = P(0)+P(1)+P(2)+P(3)+P(4)
sum(dpois(0:4,5))
```

```
## [1] 0.4404933
```

```
ppois(0:4,5,lower.tail = TRUE)
```

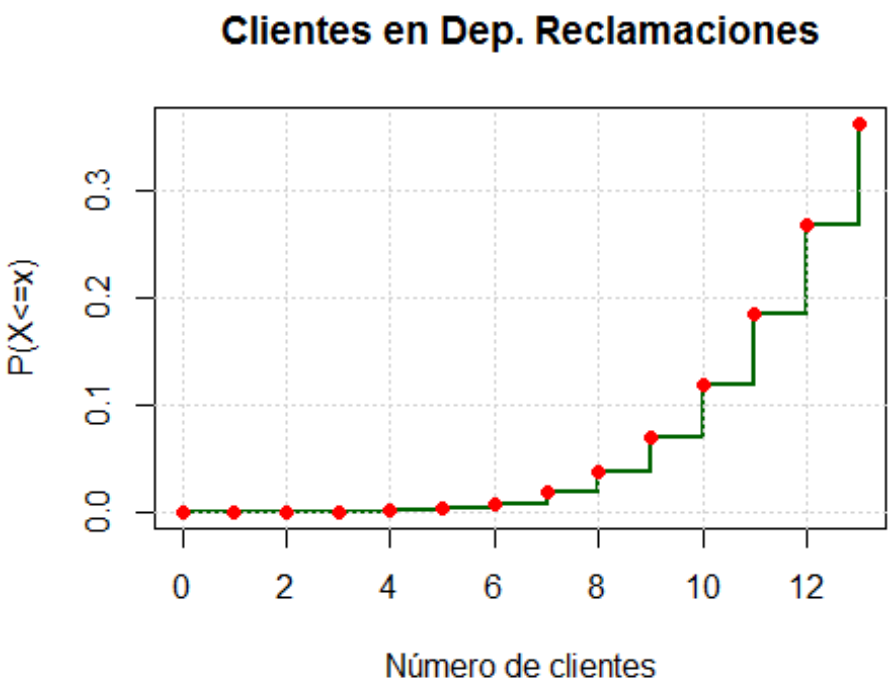
```
## [1] 0.006737947 0.040427682 0.124652019 0.265025915 0.4404932
85
```

Apartado c)

```
# Sabemos que dos horas es lo mismo que dos periodos de 60 min,
o lo que es lo mismo
# 120 minutos. Así, como nuestro promedio inicial era cada 20 mi
n, 120/20= 6 y 5*6=30.
# La media de llegadas de clientes en dos horas es 30.
```

Apartado d)

```
# Gráfico distribución del primer apartado
x<-0:13
plot(x,ppois(x,15), xlab="Número de clientes",
     ylab="P(X<=x)",type= "s", col="darkgreen", lwd=2,
     main="Clientes en Dep. Reclamaciones")
grid()
points(x,ppois(x,15), pch=19, col="red")
```




```
# Gráfico distribución del segundo apartado
plot(x,ppois(x,5), xlab="Número de clientes",
     ylab="P(X<=x)",type= "s", col="darkred", lwd=2,
     main="Clientes en Dep. Reclamaciones")
grid()
points(x,ppois(x,5), pch=19, lwd=2, col="blue")
```

Cuestión 4.5

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.

a) ¿Cuál es la probabilidad de que una serie de 15 personas cruce sin problemas antes de encontrar a una que tenga un objeto no permitido para embarcar con él?
b) ¿Cuál es el número esperado de personas que pasarán normalmente hasta que se pare en una por tener un objeto de estas características?
c) Razonar sobre la distribución de probabilidad subyacente y explicar su uso y características más significativas.
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 a 5 minutos, analizar los tiempos probables medios de espera para un vuelo de 120 pasajeros.
e) Comprobar con ***R*** los resultados anteriores, mostrar gráficamente las funciones de distribución de probabilidad correspondientes y visualizar explícitamente en caso del apartado d).

```
setwd(".")
library(knitr)
```

Apartado a)

```
Prob_A <- dnbinom(14,1,0.035)
Prob_A

## [1] 0.02125448
```

Apartado b)

```
pctg_mal <- 3.5
pctg_bien <- 100-3.5
pctg_bien

## [1] 96.5

x1<-seq(0,10,1)
```

Apartado c)

Apartado c -> se pide explicar la distribución subyacente. Vemos que es una binomial negativa, porque se repiten las pruebas hasta que ocurre un número fijo de éxitos. Tiene como elementos notables, el total n, la probabilidad de éxito p y la probabilidad de fracaso q = 1-p.

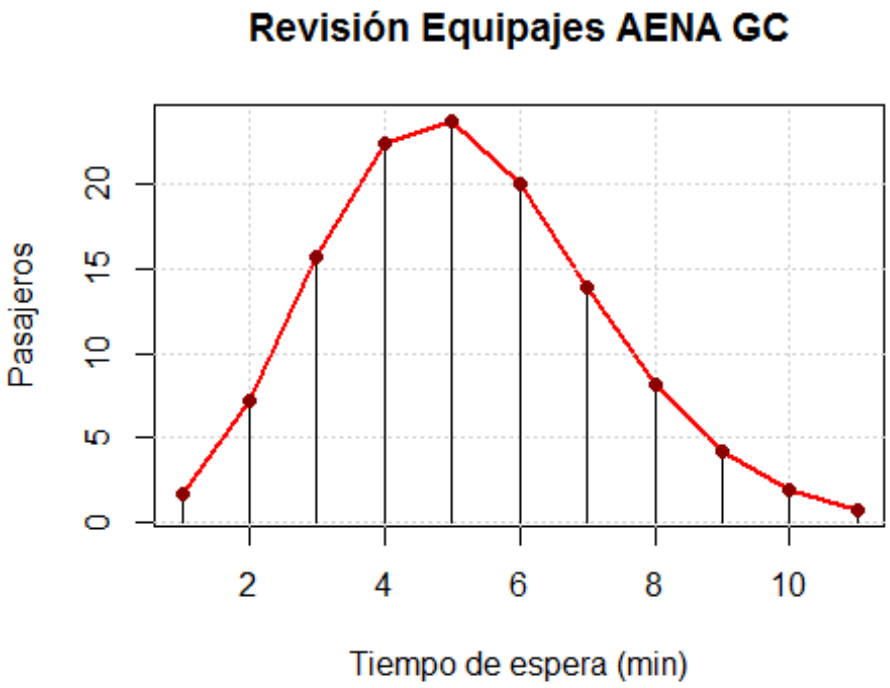
Apartado d)

Se analizan los tiempos medios de espera para un vuelo de 120 pasajeros. Si el 96.5% de los pasajeros pasa sin problema eso es que 116 solo tardarán un minuto (116 minutos).

Los cuatro restantes, que son los que tienen problemas en el acceso, tardarán 5 minutos cada uno (20 minutos).En conclusión el tiempo medio que tendrán que esperar los pasajeros es de 134 minutos.

Apartado e)

```
x <- dbinom(x1,120,0.035)*120
plot(x,type = "l",col = "red",lwd = 2,
     xlab = "Tiempo de espera (min)",
     ylab = "Pasajeros",
     main = "Revisión Equipajes AENA GC")
grid()
points(x, type="h")
points(x, type="p", col="darkred", pch=19)
```



Cuestión 5.1

Cuestión 1: La estatura de los 835 estudiantes de la Escuela de Ingeniería Informática se distribuye según una normal de media 176.5 centímetros y una desviación estándar de 7.1 centímetros. Encontrar cuántos de estos estudiantes se esperaría que tuvieran una estatura:

a) Menor que 160 centímetros.
b) Entre 171.5 y 180 centímetros.
c) Igual a 175 centímetros.
d) Mayor o igual a 190 centímetros.
e) Analizar los resultados con ***R*** y visualizar la distribución y las probabilidades de los grupos de estatura resultantes de los apartados anteriores.

Apartado a)

Primero convertimos la distribución normal en estándar n(x; 0, 1) desde μ = 176.5 y σ = 7.1 calculamos

```
mu <- 176.5
sigma <- 7.1
Z_160 <- (160-mu)/sigma
Z_160
```

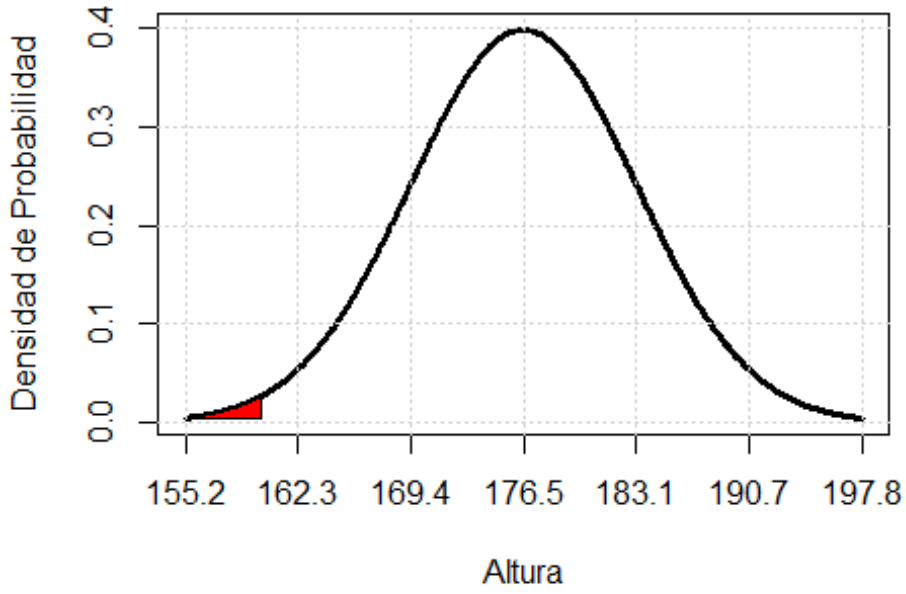
```
## [1] -2.323944
```

Para encontrar la probabilidad de que la altura sea inferior a 160 (Z_160) utilizamos pnorm(), esto nos brindará la opción de habilitar el límite izquierdo

```
pnorm(Z_160)
```

```
## [1] 0.01006426
```

```
x<-seq(-3,3,0.01)
z<-seq(-3,Z_160,0.01)
p<-dnorm(z)
z<-c(z,Z_160,-3)
p<-c(p,min(p),min(p))
plot(x,dnorm(x), type="l",xaxt="n",
     ylab = "Densidad de Probabilidad", xlab="Altura", lwd=3)

axis(1,at=-3:3, labels = c("155.2","162.3","169.4","176.5","183.
1","190.7","197.8"))
polygon(z,p,col="red")
grid()
```



Apartado b)

Recalculamos, al igual que en el apartado anterior, la distribución normal y a continuación sacamos la probabilidad en dicha distribución de que se dé una altura de 180.

Por último, restamos y vemos la probabilidad de entrada

```
mu <- 176.5
sigma <- 7.1
Z_171.5 <- (171.5-mu)/sigma
Z_171.5
```

```
## [1] -0.7042254
```

```
mu <- 176.5
sigma <- 7.1
Z_180 <- (180-mu)/sigma
Z_180
```

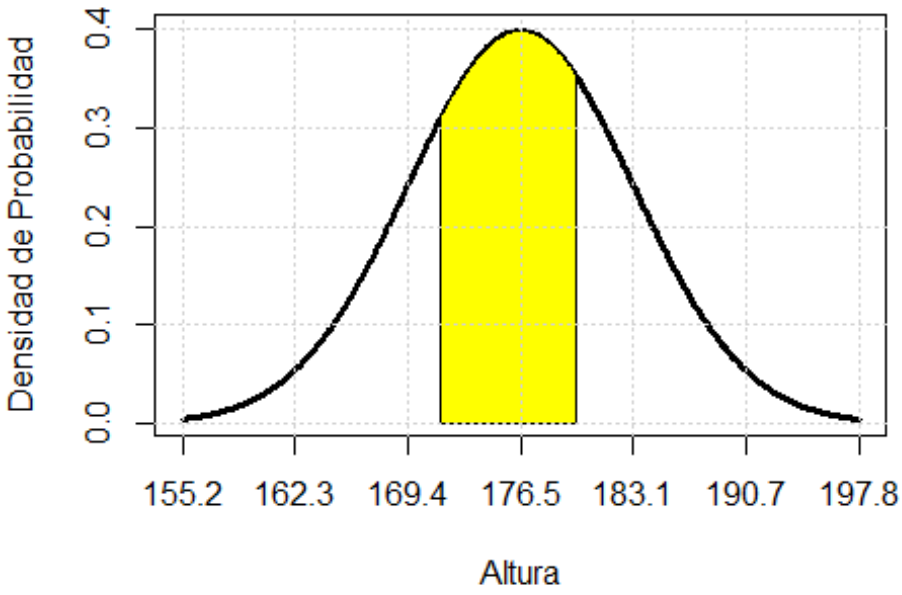
```
## [1] 0.4929577
```

```
pnorm(Z_180)-pnorm(Z_171.5)
```

```
## [1] 0.4483326
```

```
x<-seq(-3,3,0.01)
z<-seq(Z_171.5, Z_180,0.01)
p<-dnorm(z)
z<-c(z,Z_180,Z_171.5)
```

```
p<-c(p,0,0)
plot(x,dnorm(x), type="l",xaxt="n",
     ylab = "Densidad de Probabilidad", xlab="Altura", lwd=3)
axis(1,at=-3:3, labels = c("155.2","162.3","169.4","176.5","183.
1","190.7","197.8"))
polygon(z,p,col="yellow")
grid()
```

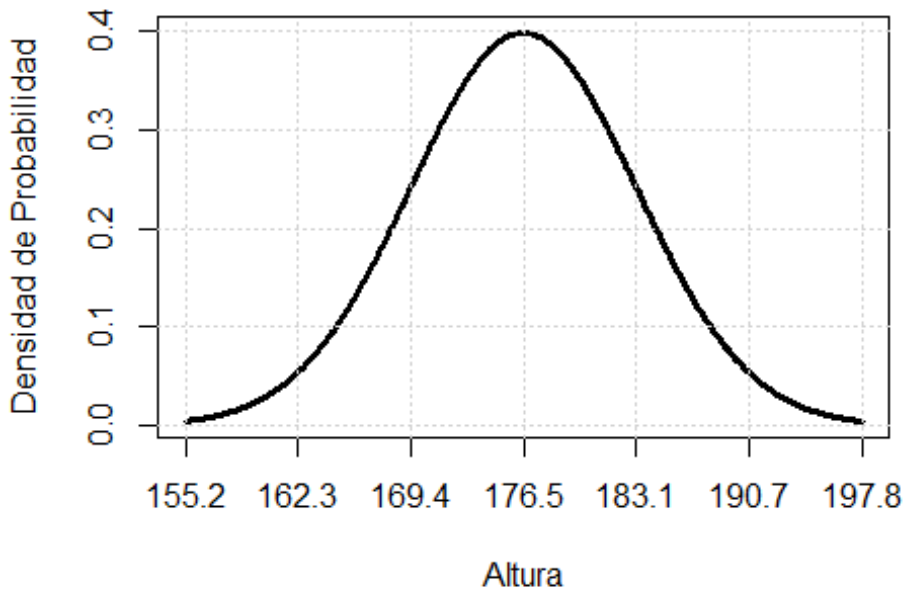


Apartado c)

Realizamos los mismos cálculos, pero para 175

```
mu <- 176.5
sigma <- 7.1
Z_175 <- (175-mu)/sigma
Z_175

## [1] -0.2112676

pnorm(Z_175)

## [1] 0.4163392

x<-seq(-3,3,0.01)
z<-seq(Z_175,Z_175,0.00)
p<-dnorm(z)
z<-c(z,Z_175,Z_175)
p<-c(p,min(p),min(p))
plot(x,dnorm(x), type="l",xaxt="n",
     ylab = "Densidad de Probabilidad", xlab="Altura", lwd=3)
```

```
axis(1,at=-3:3, labels = c("155.2","162.3","169.4","176.5","183.
1","190.7","197.8"))
polygon(z,p,col="red")
grid()
```



Apartado d)

```
mu <- 176.5
sigma <- 7.1
Z_190 <- (190-mu)/sigma
Z_190
```

```
## [1] 1.901408
```

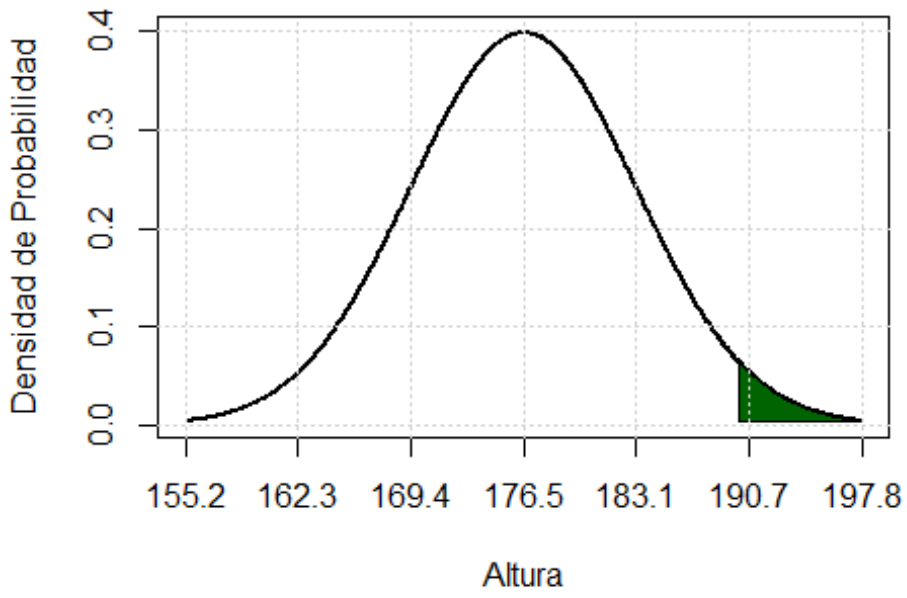
Calculamos la probabilidad de que sea menor que 190 y éste valor se los restamos a la unidad. Con ello obtenemos el valor final.

```
1-pnorm(Z_190)
```

```
## [1] 0.02862427
```

```
x<-seq(-3,3,0.01)
z<-seq(Z_190,3,0.01)
p<-dnorm(z)
z<-c(z,3,Z_190)
p<-c(p,min(p),min(p))
plot(x,dnorm(x), type="l",xaxt="n",ylab = "Densidad de Probabili
dad",
```

```
     xlab="Altura", lwd=2)
axis(1,at=-3:3, labels = c("155.2","162.3","169.4","176.5","183.
1","190.7","197.8"))
polygon(z,p,col="darkgreen")
grid()
```



Apartado e)

Se puede verificar que un gran porcentaje de los alumnos (un 44,83%) se pueden encontrar comedidos en el intervalo de los 175 a los 180 centímetros.

Del restante, muy pocos alumnos pueden ser susceptibles de tener una altura menor de 160cm (tan sólo un 1%), así como un porcentaje muy cercano aunque mayor, gozarían de una altura por encima de la media, (cerca del 3%).

Con esto se puede concluir que en las muestra se describe una altura promedia estándar, por lo que se podría concluir, a priori, que los estudiantes de la EII gozan de una altura, en general, en el promedio poblacional del país.

Cuestión 5.2

Cuestión 2: Una persona viaja diariamente en automóvil de su casa al trabajo y tarda, con atascos diarios, de media unos 25.5 minutos con una desviación de 5.1 minutos. Si sale de casa a la 8:10 y debe estar en su trabajo a las 9:00 y trabaja 240 días anuales:

a) ¿Cuántos días se espera que llegue tarde? Razonar la respuesta y explicar claramente que elementos se han utilizado.
b) ¿Cuál es la probabilidad de que un viaje le tome al menos media hora?
c) Evaluar una hora posible de salida para que el porcentaje de días que llega tarde sea inferior al 5%.

```
setwd(".")
library(knitr)
library(sqldf)
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Loading required package: RSQLite
```

Apartado a)

```
# A buena hora
pnorm(50,25.5,5.1)
```

```
## [1] 0.9999992
```

```
pnorm(50,25.5,5.1)*240
```

```
## [1] 239.9998
```

```
# Tarde
1-pnorm(50,25.5,5.1)
```

```
## [1] 7.77939e-07
```

```
1-pnorm(50,25.5,5.1)*240
```

```
## [1] -238.9998
```

Apartado b)

La probabilidad de que tarde al menos 30 min se calcularía con la distribución correspondiente.

```
pnorm(30,25.5,5.1)*240
```

```
## [1] 194.6897
```

Apartado c)

```
# Si sale a las 8:26, la probabibilidad de que llegue tarde será
1-pnorm(34,25.5,5.1)
```

```
## [1] 0.04779035
```

Gráfico de la distribución normal

```
mean<-25.5
sd<-5.1
x <- seq(-5, 5, length=100)*sd+mean
hx <- dnorm(x, mean, sd)

plot(x, hx, type="l", xlab="Tiempo (min)", col="red",
     main="Tiempo de llegada al trabajo", lwd = 2,
     ylab="Densidad")
grid()

lb<-20.4
ub<-30.6
filtro<-x>=lb&x<=ub
lines(x, hx)
polygon(c(lb,x[filtro],ub), c(0,hx[filtro],0), col="red")
```


```
zonaprob<- (pnorm(ub, mean, sd) - pnorm(lb, mean, sd))*100
zonaprob
```

```
## [1] 68.26895
```

Cuestión 5.3

Cuestión 3: Una práctica común de las aerolíneas de "*bajo coste*" es vender más billetes (pasajes de ida sólo) que el número total de plazas disponibles en un vuelo específico, pues los clientes que compran los billetes no siempre se presentan al vuelo. Supóngase que el porcentaje de pasajeros que no se presentan a la hora de salida del vuelo es del 1.9 %. Para un vuelo determinado con 185 plazas, se vendieron un total de 196 billetes.

a) ¿Cuál es la probabilidad de que la compañía aérea haya sobrevendido el vuelo?
b) Razonar sobre la distribución de probabilidad subyacente y, en su caso, sobre las posibles simplificaciones.

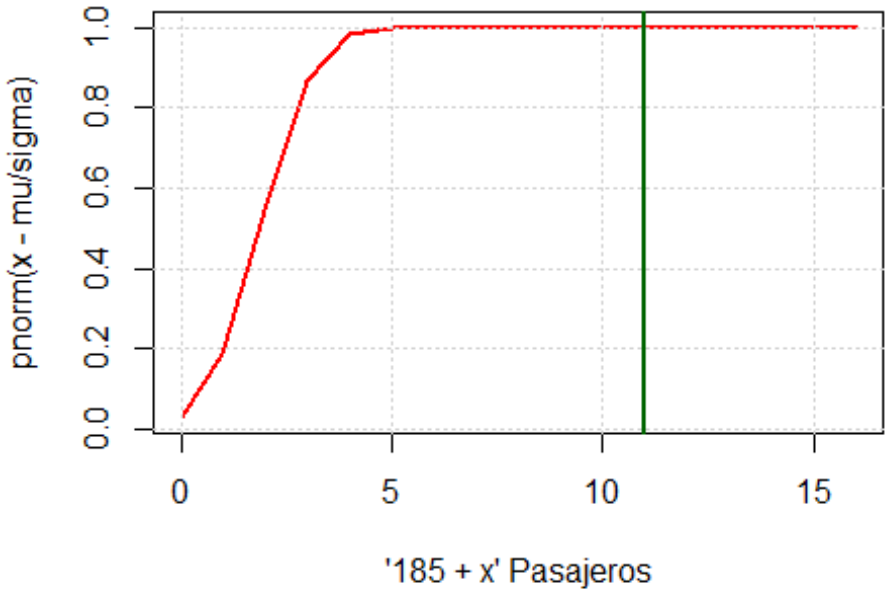
```
setwd(".")
library(knitr)
```

Apartado a)

```
n<-185
#Probabilidad de no presentarse al vuelo por parte de un pasajer
o
p<-0.019
#Probabilidad de presentarse al vuelo por parte de un pasajero
q<-0.981
mu<-(n*p)
sigma<-sqrt(n*p*q)
z<-(186-mu)/sigma
pnorm(z)
```

```
## [1] 1
```

```
x<-seq(0,16,1)
plot(x, pnorm(x-mu/sigma),type="l", col="red", lwd=2,
     xlab="'185 + x' Pasajeros")
abline(v=11, col="darkgreen", lwd=2, xlab="196")
grid()
```




```
pnorm(z)
```

```
## [1] 1
```

Apartado b)

Se ha escogido una distribución de probabilidad normal, al igual que en el ejercicio anterior.

La probabilidad de que se presente el total de pasajeros es elevada y sobrepasa las posibilidades que existen de que el número de pasajeros que no se presenten sea el necesario para no haber sobrevendido el vuelo.

Esto quiere decir que el vuelo ha sido sobrevendido, luego es posible con cuasi total seguridad que haya pasajeros que no tengan asiento. Esto se verifica a través de la gráfica, la cual indica como a partir de 185 la probabilidad de que haya asientos sobrevendidos crece rápidamente hasta llegar a 190, siendo del 100% a partir de ahí.

Cuestión 5.4

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?
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?
c) Analizar la distribución de probabilidad elegida para este caso, justificar su uso, analizar los resultados con ***R*** y visualizar los mismos.

```
setwd(".")
library(knitr)
```

Apartado a)

Cálculo sobre la ocurrencia de más de 1 fallo en un lote de 500 artículos

```
#P(X > 1)
n <- 35
p <- 0.05
PA <- 1-pbinom(1,n,p)
x <- seq(1,50,1)
y35 <-1-pbinom(x,n,p)
```

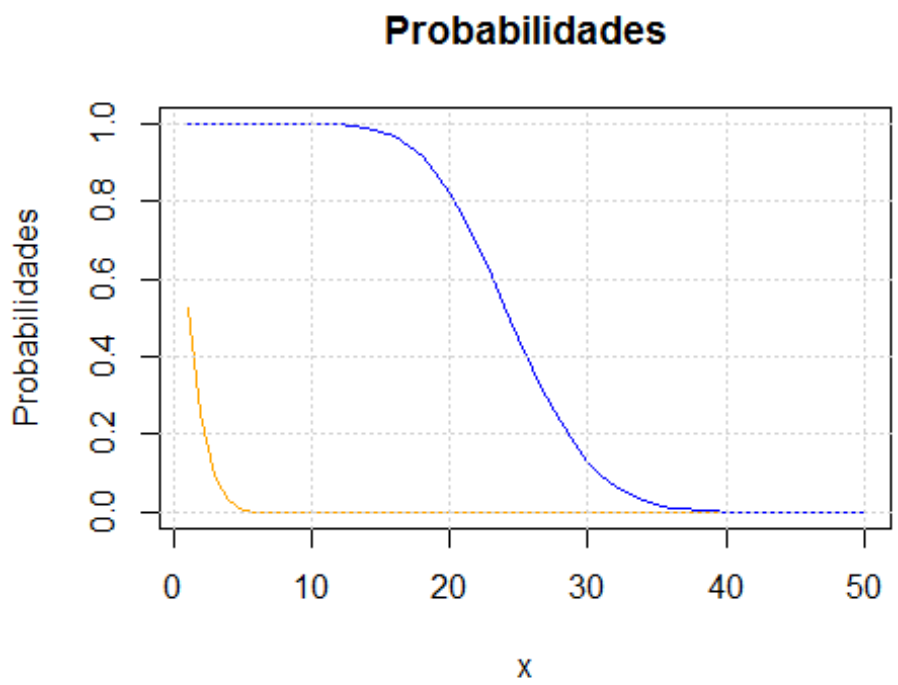
Apartado b)

Cálculo sobre la ocurrencia de más de 10 fallos en un lote de 500 artículos

```
#P(X > 10)
n2 <- 500
p2 <- 0.05
PB <-1-pbinom(9,n2,p2)
y500 <-1-pbinom(x,n2,p2)
```

Visualizacion

```
plot(x,y35, type="l", col="orange",main = "Probabilidades",ylab
= "Probabilidades", ylim = c(0, max(y500)))
points(x,y500,type = "l",col="blue")
grid()
```



Conclusiones

Al ser una cuestión en la cual se plantea el cálculo del número de éxitos en una secuencia de *n* ensayos, se ha optado la distribución binomial expresada en su forma inversa, la cual nos permitirá visualizar el comportamiento de la probabilidad respecto a una serie de muestra teóricas.

Se pueden observar en función a éste pequeño estudio, dos factores clave en el estudio de las binomiales:

- Ante una muestra pequeña la probabilidad general de no ocurrencia es alta durante las primeras fases, reduciéndose rápidamente una vez llegado a un sub-n considerable de la muestra
- Antes muestras mayores el descenso de dicha probabilidad es mucho más paulatino.

Esto nos indica que existe una relación de proporcionalidad, la cual tiene lógica: A mayor sea el número de elementos n, más tarda en aparecer una ocurrencia, mientras que a menor sea dicho número n, dicho retardo será menor, y por tanto la probabilidad ocurrencia será mayor.

Cuestión 5.5

Cuestión 5. Cierto proceso clave para el funcionamiento de un CPD, tiene una tasa de errores de 0.01 cada 10 horas. Modelar el comportamiento de este proceso asignándole la función de distribución más conveniente y justificar el uso de la misma.

a) ¿Cuál es el tiempo medio que transcurre antes del error?
b) ¿Cuál es la probabilidad de que pasen 10 días antes de que se observe un error?

No es posible realizar un modelado de la función haciendo uso de una distribución normal debido a que no son mencionadas ni las medias ni las desviaciones típicas en el caso indicado.

Hay que descartar la distribución gamma, debido a que lo que interesa hallar es la probabilidad de ocurrencia o tasa de fallos probable del sistema.

Debido a que el fallo crece de forma exponencial conforme avanza el tiempo y la probabilidad aumenta como consecuencia, emplearemos la distribución de tipo exponencial para el caso, debido a que el total de ocurrencias en el tiempo (alfa), debe de ser explícitamente 1. El tiempo por tanto lo indicará la variable beta, que es la que hay que calcular.

Apartado a)

¿Cuál es el tiempo medio que transcurre antes del error?

Sabiendo el modelo, adaptamos los datos de ocurrencia y se nos queda tal que:

$$\mu = 1/\beta \dashrightarrow 1/\mu=\beta$$
$$\beta = 1/0.01 = 100 \text{ uds. de tiempo.}$$

Apartado b)

¿Cuál es la probabilidad de que pasen 10 días antes de que se observe un error?

Calculamos la probabilidad en una distribución exponencial de un ocurrencia. En R se puede calcular empleando el comando R*: pexp(t, μ)*, siendo t el tiempo de ocurrencia de errores que se quiera aplicar y μ es la tasa de error previamente calculada. Ajustamos los días a horas (240 horas) y aplicamos el cálculo.

t->10*24=240 horas

Y sacamos la probabilidad exponencial (*pexp(240,0.01)*). Usando R para calcularla da como resultado 0.909282, es decir, cerca del 91% de posibilidades de que pasen 10 días antes de que se observe un error.

Cuestión 6.1

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?
b) Si no fuera posible, ¿a qué conclusión llegaría?
c) Razonar sobre el tamaño de la muestra y su relación sobre el posible intervalo de confianza para la media de la misma. Documentar y explicar las conclusiones

```
setwd(".")
library(knitr)
```

Apartado a)

```
mu<-20.5
xm<-23.75
s<-4
n<-8
porc<-(xm-mu)/(s/sqrt(n))
porc
```

```
## [1] 2.298097
```

```
probabilidad<-1-pt(porc,n-1)
probabilidad*100
```

```
## [1] 2.757276
```

Porcentaje de la población muestreada es 2,7%, luego la muestra es pequeña respecto al total.

Apartados b) Y c)

Calculamos si la muestra, una vez verificado su tamaño, entra en el intervalo de confianza, del 90%

```
alfa<-0.05
df1<-mu-(dt(1-(alfa/2),7)*s/sqrt(n))
df1
```

```
## [1] 20.17285
```

```
alfa<-0.95
df2<-mu+(dt(1-(alfa/2),7)*s/sqrt(n))
df2
```

```
## [1] 20.96653
```

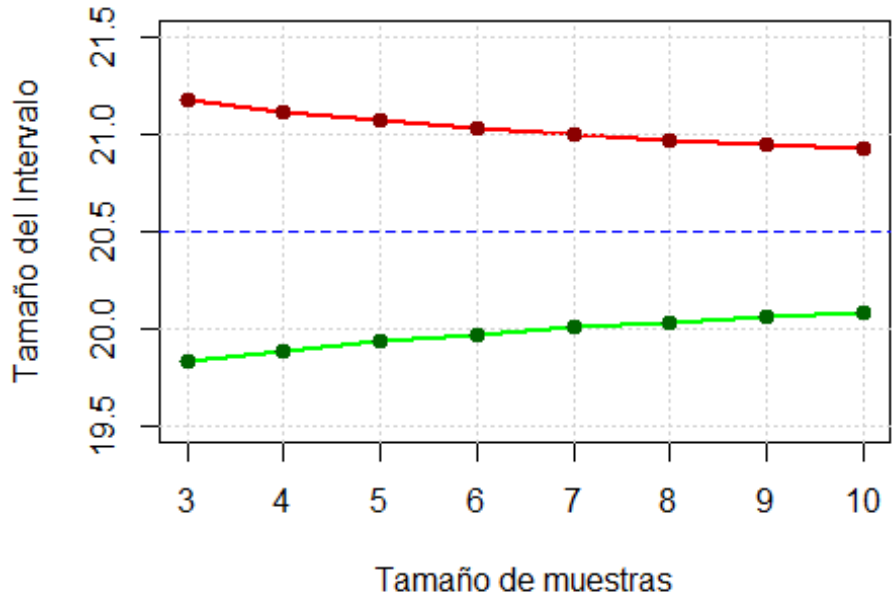
Al ser los valores derecho e izquierdo muy próximos al valor del porcentaje, entra en el intervalo de confianza, aunque por muy poco. Se podría considera como válida, pero es

imprudente debido a que la holgura respecto al IC es mínima. Viendo el comportamiento se puede concluir:

La varianza general es muy poca luego cualquier submuestra medianamente grande de *n* entrará sí o sí en el intervalo. La submuestra tomada respecto al tamaño de la población es muy grande, luego habría de reducir el tamaño de muestreo

DEMOSTRACIÓN DE LO ANTERIOR: comprobamos si el número de muestras influye a través de la gráfica

```
x<-seq(3,10,1)
plot(x, mu+(dt(1-(alfa/2),x-1)*s/sqrt(x)), type="l", col="red",
lwd=2, ylim =c(19.5,21.5),
     xlab="Tamaño de muestras", ylab="Tamaño del Intervalo")
grid()
points(x, mu-(dt(1-(alfa/2),x-1)*s/sqrt(x)), type="l", col="gree
n", lwd=2)
points(x, mu+(dt(1-(alfa/2),x-1)*s/sqrt(x)), pch=19, col="darkre
d", lwd=2)
points(x, mu-(dt(1-(alfa/2),x-1)*s/sqrt(x)), pch=19, col="darkgr
een", lwd=2)
abline(h=mu, col="blue", lwd=1.75, lty=2)
```



Conclusiones

Se verifica por tanto, que ante una población de tamaño n, es prudente revisar bien el tamaño de las muestras, puesto que un excesivamente grande puede provocar errores en el estudio final.

Cuestión 6.2

Cuestión 2: Las calificaciones de un examen de Métodos Estadísticos durante los últimos cinco años tienen aproximadamente una distribución normal de media $\mu = 7.45$ y una varianza de $\sigma^2 = 0.8$. ¿Se seguiría considerando que $\sigma^2 = 0.8$ 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 adecuadamente la respuesta y justificar teóricamente la misma.

```
setwd(".")
library(knitr)
```

Apartado a)

```
mu <- 7.45
sigma_cuad <- 0.8
n <- 20
s2 <- 20
X2<- (n-1)*s2/sigma_cuad
xlim <- qchisq(0.95,19)
xlim2 <- 1-pchisq((20*19)/0.8,19)

s2

## [1] 20

X2

## [1] 475

xlim

## [1] 30.14353

xlim2

## [1] 0
```

Conclusiones

Se pide comprobar si el valor de la varianza continua siendo válido para una varianza poblacional de 20, de una muestra de 20 estudiantes.

Debido a que el problema, por los datos indicados, corresponde a una función tipo Xi cuadrada con 19 grados de libertad. Calculándolo con la fórmula. Una vez hecho, calculamos los límites y da como resultado 0 y 30. Por lo tanto el valor de la varianza es un valor válido.

Cuestión 6.3

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 maquina es 900 ± 150 gramos. El proveedor considera que cualquier envase de cartón que no cumpla con tales límites de peso en el llenado esta defectuoso. Se espera que al menos 99% de los envases de cartón cumplan con la especificación. En el caso de que $\mu = 900$ y $\sigma = 100$,

a) ¿Qué proporción de envases de cartón del proceso están defectuosos?
b) Si se hacen cambios para reducir la variabilidad, ¿cuanto se tiene que reducir σ para que haya 0.99 de probabilidades de cumplir con la especificación?
c) ¿Cuál será el tamaño de muestra para que en este segundo caso se garanticen las especificaciones?
d) Visualizar gráficamente con R los casos a) y b)

Supóngase una distribución normal para el peso. Razonar convenientemente y justificar las respuestas.

```
setwd(".")
library(knitr)
```

Apartado a)

```
setwd(".")#700 1100 10 900
library(knitr)
limsup<-900+150
liminf<-900-150
prob<-0.99
mu<-900
sd<-100
sigma<-10000
#Apartado A
2*pnorm(750,900,100) # Se multiplica por 2
```

```
## [1] 0.1336144
```

```
#Apartado B
sigma_new <- (1050-900)/qnorm(0.995)
z <- (1050-900)/sigma_new
```

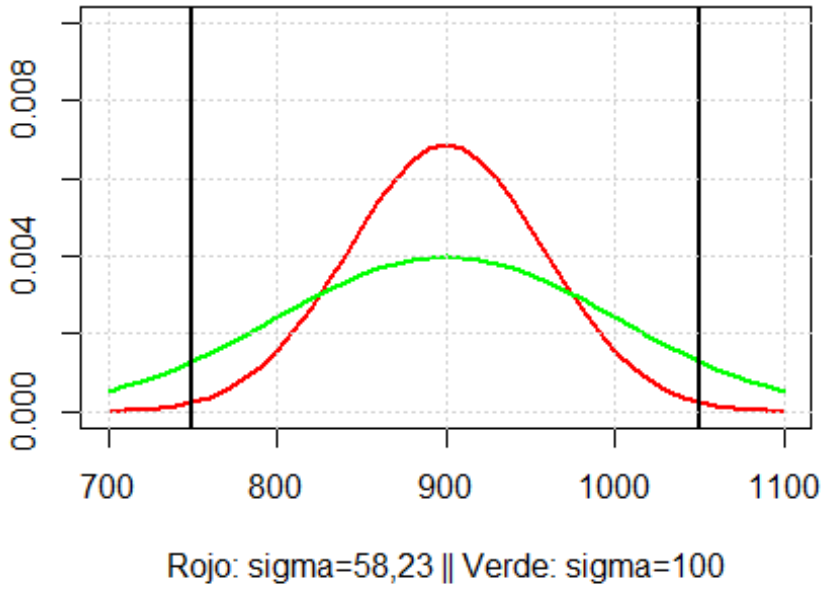
Apartado c)

El tamaño de la muestra es 1 mismamente, debido a la sigma. Al ser 0,99 con un elemento bastaría para que se cumpliese.

Apartado d)

```
x<-seq(700,1100,10)
plot(x, dnorm(x,900,sigma_new), type = "l", col="red",ylim = c(0
,0.01), lwd=2,
     xlab="Rojo: sigma=58,23 || Verde: sigma=100", ylab="")
points(x, dnorm(x,900,100), type = "l", col="green", lwd=2)
```

```
abline(v=750, col="black", lwd=2)
abline(v=1050, col="black", lwd=2)
grid()
```



Conclusiones

Se puede concluir que a menor sea sigma (σ), más centrada se queda la representación final de la tabla, mientras que a mayor sea la sigma menor será el pico de la distribución en su representación final debido a que z se ve influenciada de forma proporcionalmente inversa por sigma.

Cuestión 6.4

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

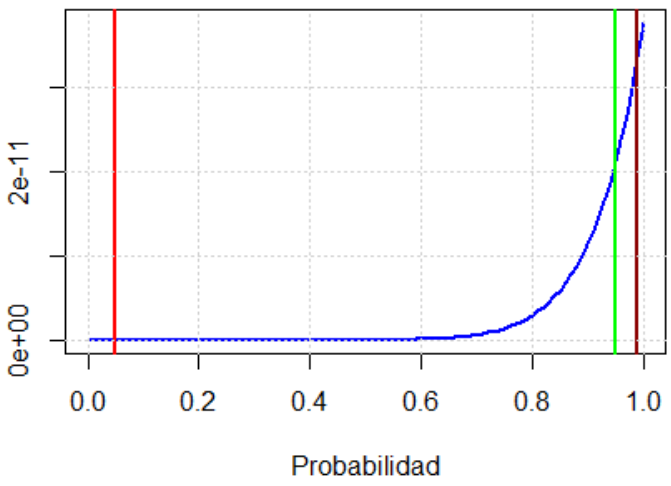
```
setwd(".")
library(knitr)
```

Apartado a)

```
PA <- 1-pchisq(24*9.1/6,24)
PA

## [1] 0.0501701
```

Apartado b)

```
PBmenos<-pchisq(24*10.745/6,24)
PBmas<-pchisq(24*3.462/6,24)
PB <- PBmenos-PBmas
PB

## [1] 0.9400097

x <- seq(0,1,0.01)
plot(x, pchisq(x*(9.1/6),24), col="blue", type="l",
     xlab="Probabilidad", ylab="", lwd=2)
abline(v=pchisq(24*9.1/6,24), col="green", lwd=2)
abline(v=pchisq(24*10.745/6,24), col="darkred", lwd=2)
abline(v=pchisq(24*3.462/6,24), col="red", lwd=2)
grid()
```



Conclusiones

Es posible anticipar, con un poco de lógica os resultados de esta prueba. Hay que tener en cuenta que al tener el intervalo 3,642 y 10,745, el margen de maniobra que se nos da para estimarlo es altísimo, luego es muy probable que la varianza se encuentre entre esos valores.

Esto lo confirmamos con un casi 95% de posibilidades.

Respecto a la posibilidad de que sea mayor 9,1, pues es todo lo contrario, debido a que el margen es muy corto, de casi 1,4 puntos en la escala, luego es lógico que las posibilidades sean pequeñas, de tan solo el 5%

Cuestión 7.1

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.
b) ¿A qué conclusiones se llegan a partir de la información suministrada por muestra? Razonar la respuesta y justificar teóricamente la misma.
c) ¿Cuál sería el tamaño de la muestra para garantizar en un 99% la duración promedio resultante?
d) ¿Se podría con estos datos calcular un intervalo de tolerancia del 99%? Razónense las respuestas.
a) Mostrar gráficamente con ***R*** los intervalos para las hipótesis establecidas y visualizar las conclusiones.

```
setwd(".")
library(knitr)

n<-30
x<-780
sigma<-40
z<- -1.96
```

Apartado a)

```
Liminf<- x+z*sigma/sqrt(n)
Liminf

## [1] 765.6862

Limsup<- x-z*sigma/sqrt(n)
Limsup

## [1] 794.3138
```

Apartado b)

Se puede considerar que la promoción de las bombillas por parte de la empresa es correcta debido a que la diferencia de la media intervaluada en una región al 96% de confianza respecto a la producción da un margen bastante favorable respecto a lo indicado, de unas 15 horas, contra 40 mencionadas en el enunciado.

Apartado c)

```
qnorm(0.005)

## [1] -2.575829
```

```
z99<--2.5758
mu<--1*(sigma/sqrt(n))*z+x
mu
```

```
## [1] 794.3138
```

```
k<-x-mu
k
```

```
## [1] -14.31382
```

```
Muestra<-(z99*sigma)/k
Muestra<-Muestra^2
Muestra #La muestra debe ser de tamaño 51
```

```
## [1] 51.81236
```

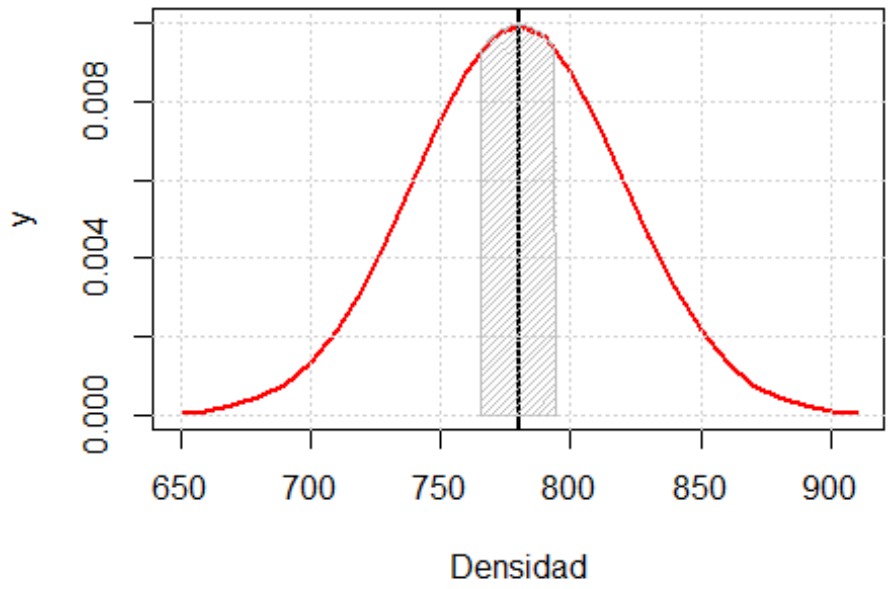
Apartado d)

Si, sería posible reutilizando el valor de z99, aunque nos queda un intervalo más amplio respecto al de confianza al 96%.

```
liminf99 <-x+z99*sigma/sqrt(n)
liminf99
## [1] 761.189

limsup99 <-x-z99*sigma/sqrt(n)
limsup99
## [1] 798.811
```

Apartado e)

```
xm<-seq(650,910,10)
xpol<-seq(Liminf,Limsup,1)
y<-dnorm(xm, x, sigma)
ypol<-dnorm(xpol, x, sigma)
xpol<-c(xpol,Limsup,Liminf)
ypol<-c(ypol,0,0)

plot(xm, y, type="l", col="red", lwd=2, xlab="Densidad")
abline(v=780, col="black", lwd=2)
polygon(xpol, ypol, col="grey", density=35)
grid()
```


y
0.000
0.004
0.008
650
700
750
800
850
900
Densidad

Cuestión 7.2

Cuestión 2: Una máquina para un taller de la zona industrial del Cebadal produce piezas metálicas de forma cilíndrica para aparatos de aire acondicionado. 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.
b) ¿Se podría realizar alguna inferencia sobre la varianza poblacional?

```
setwd(".")
library(knitr)

diametros <- c(1.01,0.97,1.03,1.04,0.99,0.98,0.99,
           1.01,1.03)
```

Apartado a)

intervalo de confianza del 99% para la media

```
mu <- mean(diametros)
sigma <- sd(diametros)
n <- length(diametros)
liminf <- mu - qt(0.99,df=(n-1))*sigma/sqrt(n)
limisup <- mu + qt(0.99,df = (n-1))*sigma/sqrt(n)

liminf

## [1] 0.9818514

limisup

## [1] 1.02926
```

Comprobación usando t.test

```
t.test(diametros,conf.level = 0.99)

##
##  One Sample t-test
##
## data:  diametros
## t = 122.87, df = 8, p-value = 2.152e-14
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  0.9780956 1.0330155
## sample estimates:
## mean of x
##  1.005556
```

El valor de los limites calculado manualmente es muy aproximado al del test, luego se podría afirmar que efectivamente, los limites calculados son correctos

Apartado b)

Es posible plantear una hipótesis acerca de si la varianza es o no susceptible de ser considerada correcta en este caso. Para ello, se calcula un intervalo de confianza del 99 % de la varianza y el resultado es 0.0065-0.0066.

Con esto, se puede teorizar que las piezas no tendrán mucha variación en la medida de sus diámetros, con respecto a las medidas en la muestra.

```
s2<- var(diametros)
liminf_var <-(n-1)*s2/qchisq(1-(0.99/2),n-1)
limsup_var <-(n-1)*s2/qchisq((0.99/2),n-1)

pinfvar<-liminf_var*100
psupvar<-limsup_var*100
```

Cuestión 7.3

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

```
setwd(".")
library(knitr)
```

Apartado a)

```
n <- 25
mu <- 325.05
sigma <- 0.5
alfa <- 0.1
k <- 1- alfa
#Calcular limites de tolerancia del 95% del 90% que contiene la aspirina
#X+-k*s
liminf <- mu-k*sigma
liminf
```

```
## [1] 324.6
```

```
limsup <- mu+k*sigma
limsup
```

```
## [1] 325.5
```

Conclusiones

Calculamos lo límites empleando la fórmula del intervalo k (mu- k * sigma) y extraemos tanto el superior como el inferior. Debido a que el resultado da una distancia por ambos lados de cerca de 0.5%, se puede asumir que la hipótesis de que al 95% de tolerancia contendrán el 90% del contenido, es cierta.

Cuestión 7.4

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,

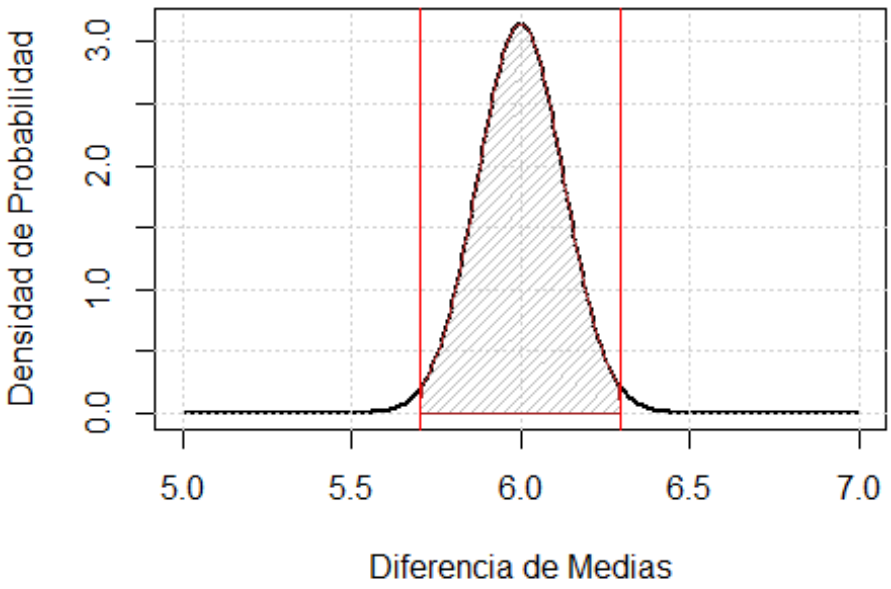
lo que produce un promedio de 12.2 milímetros de metal eliminados y una desviación estándar muestral de 1.1 milímetros. Una segunda muestra de 200 piezas se somete al tratamiento, seguido de 24 horas de inmersión en el baño, lo que da como resultado una eliminación promedio de 9.1 milímetros de metal, 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.
b) Analizar según los datos si el tratamiento reduce o no la cantidad media del metal eliminado. Razonar adecuadamente la respuesta

```
setwd(".")
library(knitr)
```

Apartado a)

```
x<-seq(5,7,0.01)
s1<-1.1
s2<-0.9
n1<-100
n2<-200
Alfa<-0.02
DPx<-dnorm(x,6,sqrt((s1^2/n1)+(s2^2/n2)))
plot(x,DPx,type="l",col="black", lwd=2,ylab="Densidad de Probabi
lidad",xlab="Diferencia de Medias")
grid()
dcha<-(Alfa/2)
xliminf<-qnorm(dcha,6,sqrt((s1^2/n1)+(s2^2/n2)))
izqda<-(1-Alfa/2)
xlimsup<-qnorm(izqda,6,sqrt((s1^2/n1)+(s2^2/n2)))
xv<-x[x>=xliminf & x <=xlimsup]
yv<-DPx[x>=xliminf & x <=xlimsup]

xv<-c(xv,xlimsup,xliminf)
yv<-c(yv,DPx[1],DPx[1])
polygon(xv,yv,col="grey",density=25,border="brown")
abline(v=xliminf, col="red")
abline(v=xlimsup, col="red")
```




```
xliminf
```

```
## [1] 5.704362
```

```
xlimsup
```

```
## [1] 6.295638
```

Conclusiones

Se verifica la diferencia de eliminación del componente metálico a través de los resultados

Al calcular el estimado del intervalo de confianza al 98 %, vemos que nos da un intervalos de entre 5.7 y 6.3. Por ello, de promedio, los valores están fuera del intervalo, por lo cual la reducción del metal eliminado no es muy alta y por ende se puede despreciar la hipótesis.

Cuestión 7.5

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, los cuales se utilizan hasta que se desgastan. Los resultados son:

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

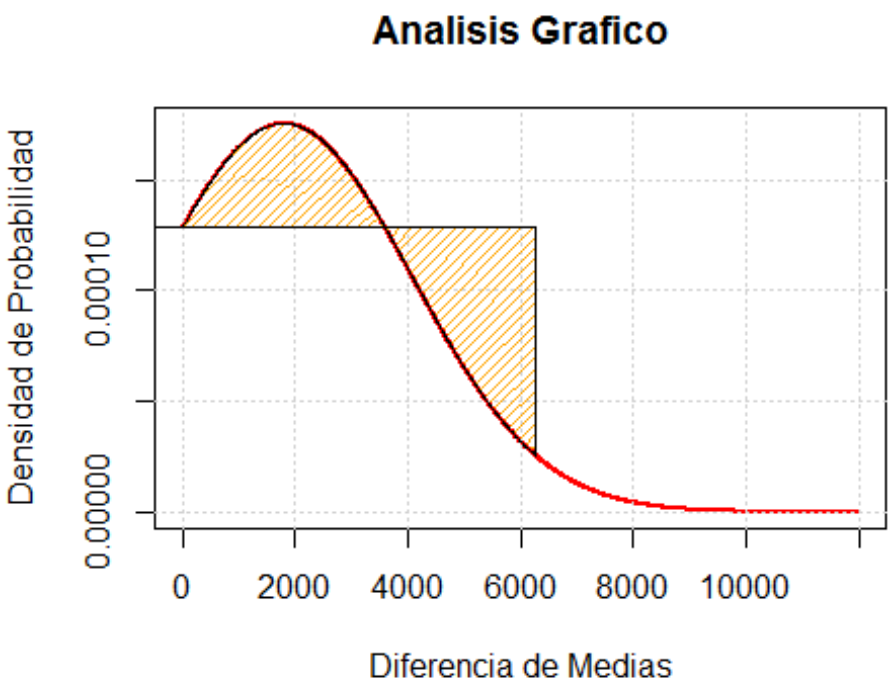
.

a) Calcular un intervalo de confianza del 95% para $\mu_A - \mu_B$, suponiendo que las poblaciones se distribuyen de forma aproximadamente normal.
b) Analice los resultados bajo las suposiciones de que las varianzas poblacionales sean o no iguales y explicar los mismos. Justificar las respuestas.

```
setwd(".")
library(knitr)

x<-seq(0,12000,10)
n<-12
x1<-36300
x2<-38100
s1<-5000
s2<-6100
Alfa<-1-0.95
```

Apartados a) y b)

```
y<-dnorm(x, x2-x1, sqrt((s1^2/n)+(s2^2/n)))
plot(x,y, type = "l", col="red", lwd=2, ylab = "Densidad de Prob
abilidad", xlab =
       "Diferencia de Medias", main ="Analisis Grafico");
grid()
xliminf<-qnorm(Alfa/2, x2-x1, sqrt((s1^2/n)+(s2^2/n)))
xlimsup<-qnorm(1-Alfa/2, x2-x1, sqrt((s1^2/n)+(s2^2/n)))
xliminf

## [1] -2662.596

xlimsup

## [1] 6262.596

xv<-x[x>=xliminf & x<=xlimsup]
yv<-y[x>=xliminf & x<=xlimsup]
xv<-c(xv,xlimsup,xliminf)
yv<-c(yv,y[1],y[1])
polygon(xv,yv,col="orange", density=25, border="black")
```



Conclusiones

Si tuviésemos en cuenta que las varianza poblaciones fuesen iguales, la diferencia entre μA-μB ser haría nula y por ende se dependería de las medias para calcular el valor de z. Esto facilita las cosas pero implica suponer que las dos poblaciones:

a) Son la misma muestra.
b) Son muy similares luego al caso habría que seguir testeando con diferentes muestras.

Si las μA-μB son diferentes, la clave queda en la diferencia de ambas y cómo esa diferencia afecta al cómputo. Si la diferencia es muy amplia pero la media es más o menos constante implica que las muestras pueden estar definidas bajo un espectro concreto. Por ende si ambas son muy distintas, no se podría sacar mas que un valor aproximado y, según la situación, descartable.