Ejercicios LAB 5, 6 Y 7. Cuestiones L8, L9, y L10

MÉTODOS ESTADÍSTICOS

AARÓN HERNÁNDEZ ÁLVAREZ

CONTENIDO

CONTENIDO 1
EJERCICIOS LABORATORIO 3
Ejercicio 5.1 3
Apartado a) 3
Apartado b) 3
Apartado c) 4
Conclusiones 4
Ejercicio 5.2 5
Apartado a) 5
Apartado b) 5
Apartado c) 6
Conclusiones 7
Ejercicio 5.3 8
Apartado a) 8
Conclusiones 9
Ejercicio 5.4 10
Apartado a) 10
Apartado b) 12
Ejercicio 6.1 13
Ejercicio 6.2 15
Ejercicio 6.3 18
Ejercicio 6.4 20
Ejercicio 6.5 22
Ejercicio 7.1 27
Apartado a) 27
Apartado b) 30
Apartado e) 31
Apartado f) 32
Apartado g) 32
CUESTIONES LECTURAS 34
Cuestión 8.1 34
Apartado a) 34
Apartado b) 34
Apartado c) 35
Conclusiones 35
Cuestión 8.2 36
Apartado a) 36

Apartados b) y c) 37
Apartado d) 37
Conclusiones 37
Cuestión 8.3 38
Apartado a) 38
Apartado b) 40
Conclusiones 40
Cuestión 8.4 41
Apartado a) 41
Apartado b 41
Apartado c) 42
Conclusiones 42
Cuestión 9.1 43
Apartado a) 43
Apartado b) 44
Cuestión 9.2 46
Apartados a) y b) 46
Conclusiones 48
Cuestión 9.3 49
Apartado a) 50
Apartado b) 56
Cuestión 9.4 58
Conclusiones 59
Cuestión 10.1 60
Apartado a) 60
Apartado b) 62
Apartado c) y d) 64
Conclusiones 65
Cuestión 10.2 66
Apartado a) 67
Apartado b) 68
Apartado c) 68
Apartado d) 69
Apartado e) 70
Apartado f) 71

EJERCICIOS LABORATORIO

Ejercicio 5.1

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:

a) Ajustar una distribución normal a esos datos mediante el método de máxima verosimilitud.
b) Representar gráficamente el diagrama de barras de los datos junto con la función masa de la distribución del ajuste.
c) ¿Es la distribución resultante un buen ajuste para los datos? Razonar la respuesta.

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

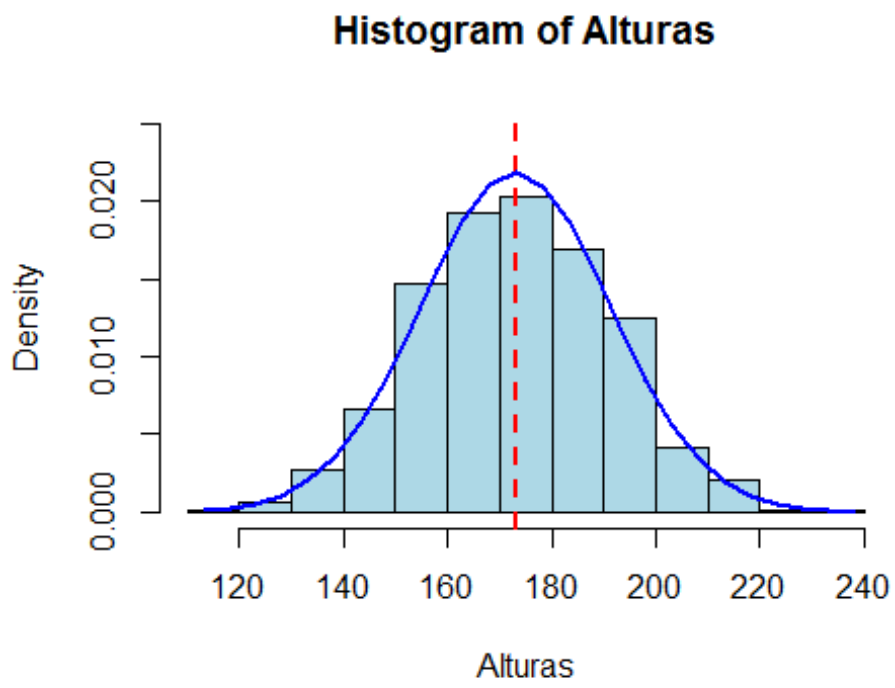
```
## Loading required package: gsubfn

## Loading required package: proto

## Loading required package: RSQLite

library(MASS)

datos<-read.table("Archivos/Alturas_Estudiantes_EII.txt", dec =
".", sep=",")
```

Apartado a)

```
attach(datos)
n<-length(Alturas)

parametro<-fitdistr(Alturas,"normal")
mu<-parametro$estimate[1]
sigma<-parametro$estimate[2]
```

Apartado b)

```
hist(Alturas, freq = F, col="lightblue", ylim=c(0,0.025))
abline(v=mu, col="red", lty=2, lwd=2)
x<-seq(min(Alturas),max(Alturas),5)
points(x,dnorm(x,mu,sigma), type="l", col="blue", lwd=2)
```



Apartado c)

```
sdmu<-parametro$sd[1]
sdmu

##     mean
## 0.724717

sdsigma<-parametro$sd[2]
sdsigma

##        sd
## 0.5124523
```

Conclusiones

Poca desviación en ambos casos, la distribución sigue una forma gaussiana

Ejercicio 5.2

Ejercicio 2. El fichero "***sueldos_hosteleria.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 antigüedad análogas.

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.
b) Encontrar el intervalo de confianza para la varianza y la desviación estándar en las condiciones del apartado anterior.
c) Visualizar los datos asumiendo que han podido obtenerse de una distribución normal de media 18510€ y desviación estándar de 850€. Explicar las conclusiones.

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

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

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

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

```
library(MASS)

sueldos<-read.table("Archivos/sueldos_hosteleria.txt", dec = "."
, sep=",")
attach(sueldos)
```

Apartado a)

```
n<-length(Sueldos)
media<-mean(Sueldos)
S<-sd(Sueldos)
t<-qt(0.95,n-1)

#Calculamos los limites intervalo de confianza para mu:
#Formula: x-t+(S/sqrt(n))

mu1<-(media-t*(S/sqrt(n))) #Limite por la izquierda
mu2<-(media+t*(S/sqrt(n))) #Limite por la derecha
```

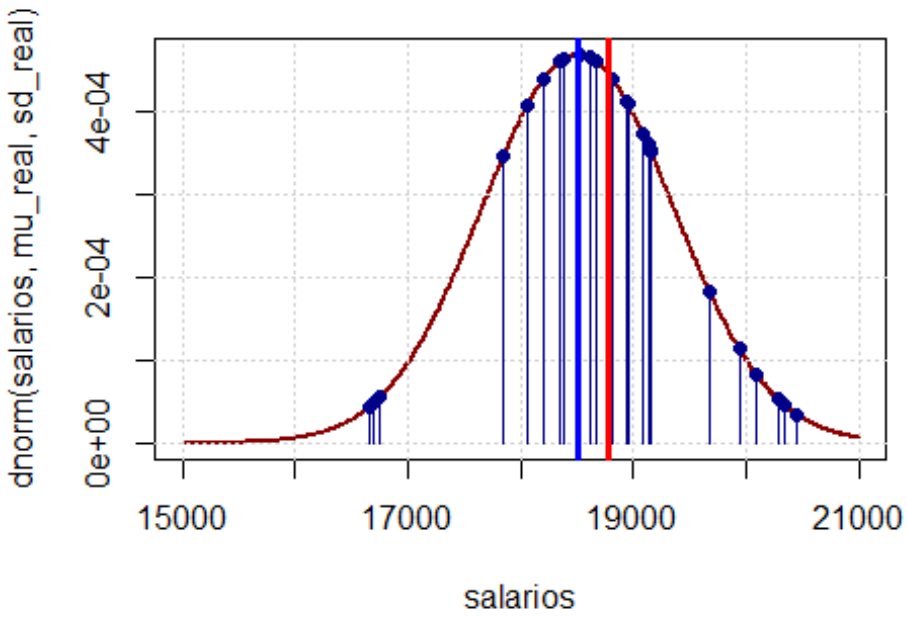
Apartado b)

```
#Calculamos los limites intervalo de confianza para sigma
#Formula: sqrt((n-1)*S^2/xilim)
```

```
xisup<-qchisq(0.95,n-1)
xiinf<-qchisq(0.05,n-1)

sigma21<-((n-1)*S^2)/xiinf
sigma22<-((n-1)*S^2)/xisup

sdsigma1<-sqrt(sigma21) #Limite por la izquierda
sdsigma2<-sqrt(sigma22) #Limite por la derecha
```

Apartado c)

```
mu_real<-18510
sd_real<-850
liminf<-15000
limsup<-21000
salarios<-seq(15000,21000,10)
plot(salarios, dnorm(salarios, mu_real, sd_real), type="l", col="darkred", lwd=2)
grid()
points(Sueldos,dnorm(Sueldos,18510,850), type="h", lwd=1, col="darkblue")
points(Sueldos,dnorm(Sueldos,18510,850), type="p", pch=19, lwd=1, col="darkblue")
abline(v=media, col="red", lwd=3)
abline(v=18510, col="blue", lwd=3)
```



Podemos calcular t y verificar que la media calculada con respecto la media de la empresa es cuasi equivalentes

```r
t.test( Sueldos, alternative=c("two.sided"), mu=18510,
        paired=FALSE, var.equal=FALSE, conf.level=0.9)

##
##  One Sample t-test
##
## data:  Sueldos
## t = 1.2478, df = 24, p-value = 0.2242
## alternative hypothesis: true mean is not equal to 18510
## 90 percent confidence interval:
##  18411.73 19137.79
## sample estimates:
## mean of x
##  18774.76
```

Conclusiones

La probabilidad es de 0,22/1, es decir, del 22% de ser la media. Esto implica que es considerable como solución. Se puede apreciar poca desviación en ambos casos, la distribución seguirá una forma Gaussiana, es decir, distribuida nominalmente.

Asimismo se puede concluir que la ejecución de diferentes test según la capa de medición nos da la posibilidad de verificar si la hipótesis es correcta. Una mayor aproximación a través de los estadísticos que forman las muestras, como la varianza o la desviación, puede además apoyar una toma de decisión aún más precisa.

Ejercicio 5.3

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, ¿tiene razón los empresarios? (Utilizar un nivel de significación del 5 %).

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

## Loading required package: gsubfn

## Loading required package: proto

## Loading required package: RSQLite

library(MASS)

sueldos<-read.table("Archivos/sueldos_hosteleria.txt", dec = "."
, sep=",")
attach(sueldos)
```

Apartado a)

Podemos, al igual que en ejercicio anterior, calcular t y verificar que tanto la media calculada como la media afirmada por la empresa son cuasi equivalentes.

```
t.test( Sueldos, alternative=c("two.sided"), mu=18510,
        paired=FALSE, var.equal=FALSE, conf.level=0.95)

##
##  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
```

Conclusiones

La probabilidad es de 0,22/1, es decir, del 22% de ser la media, Esto implica que es considerable como solución.

Asimismo, el rango de sueldos se acorta respecto al caso anterior, lo cual tiene sentido debido a que a menor sea el intervalo de confianza menor es el rango de acierto.

Ejercicio 5.4

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 considerarse que los cuatro métodos de poda producen resultados equivalentes.

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

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

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

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

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

```
library(MASS)

poda<-read.table("Archivos/plantas_poda.txt", dec = ".", sep=","
)
attach(poda)
aov(Biomasa~Tipo_Poda)
```

```
## Call:
##    aov(formula = Biomasa ~ Tipo_Poda)
##
## Terms:
##                 Tipo_Poda Residuals
## Sum of Squares   85356.47 124020.33
## Deg. of Freedom         4        25
##
## Residual standard error: 70.43304
## Estimated effects may be unbalanced
```

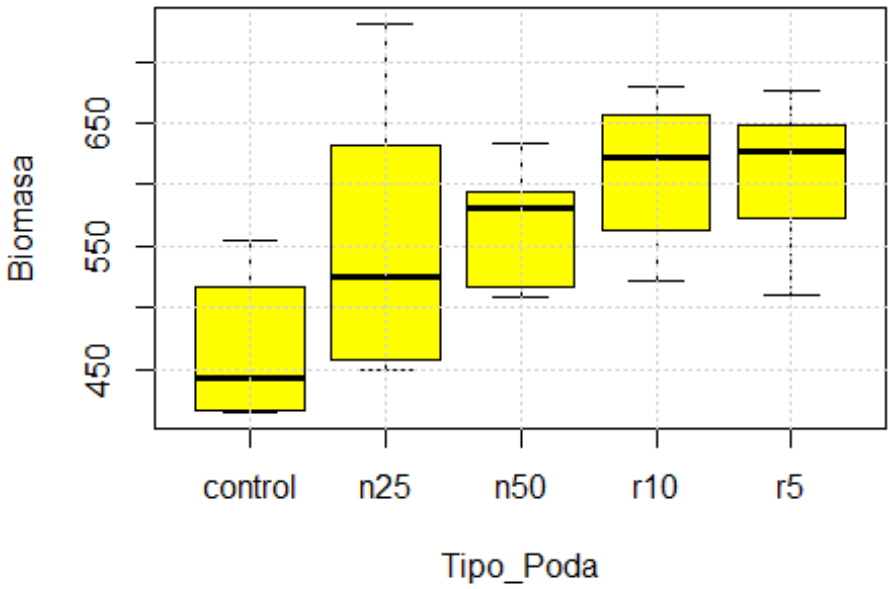
Apartado a)

```
x25<-mean(Biomasa[Tipo_Poda=="n25"])
x50<-mean(Biomasa[Tipo_Poda=="n50"])
x10<-mean(Biomasa[Tipo_Poda=="r10"])
```

```
x5<-mean(Biomasa[Tipo_Poda=="r5"])
xcont<-mean(Biomasa[Tipo_Poda=="control"])
```

Ejecutamos un boxplot para ver las medias

```
boxplot(Biomasa~Tipo_Poda, col="yellow")
grid()
```



Ejecutamos el test ANOVA

```
anova<-aov(Biomasa~Tipo_Poda, data=poda)
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
```

Ejecutamos el limite de f para verificar que la F del test anova

```
k<-5
n<-length(Biomasa)
pf<-qf(0.95,k-1,n-k)
```

Y por ultimo el inverso para verficar que coincide con el valor de probabilidad del test ANOVA.

```
Pr<-1-pf(4.302,k-1,n-k)
```

Apartado b)

Estudio

No se pueden considera que son equivalentes porque al 0,05 de diferencia da un resultado mucho menor, luego la probabilidad de que ocurra el hecho de que todas de la o mismo es muy baja, por no decir nula.

A continuación nos fijamos en las medias correspondientes y observamos que la media en r10 y r5 es muy similar, luego verificamos mediante un test cual de ambas es cualitativamente mejor

- Hipótesis 0: r5 y r10 tienen medias similares
- Hipótesis 1: no la tienen a una confianza al 0.05

t.test(x10, x5):

```
 Welch Two Sample t-test

 data:  x10 and x5
 t = 0.0048625, df = 9.9962, p-value = 0.9962
 alternative hypothesis: true difference in means is not equal to 0
 95 percent confidence interval:
  -76.20846  76.54179
 sample estimates:
 mean of x mean of y
  610.6667  610.5000
```

Como el test T da casi igual en probabilidad, ambas son prácticamente iguales, luego se puede afirmar, en este caso, que la ejecución de los procedimientos o sistemas de poda r10 y r5 es irrelevante en para el resultado final. Esto coincide con el gráfico mostrado al principio, donde las varianzas de ambos son muy parecidas y sus medias similares.

Si se diera el caso de que las varianzas oscilan demasiado, el test ANOVA no se podría utilizar y sería necesario emplear el test alternativo para varianzas distintas.

Ejercicio 6.1

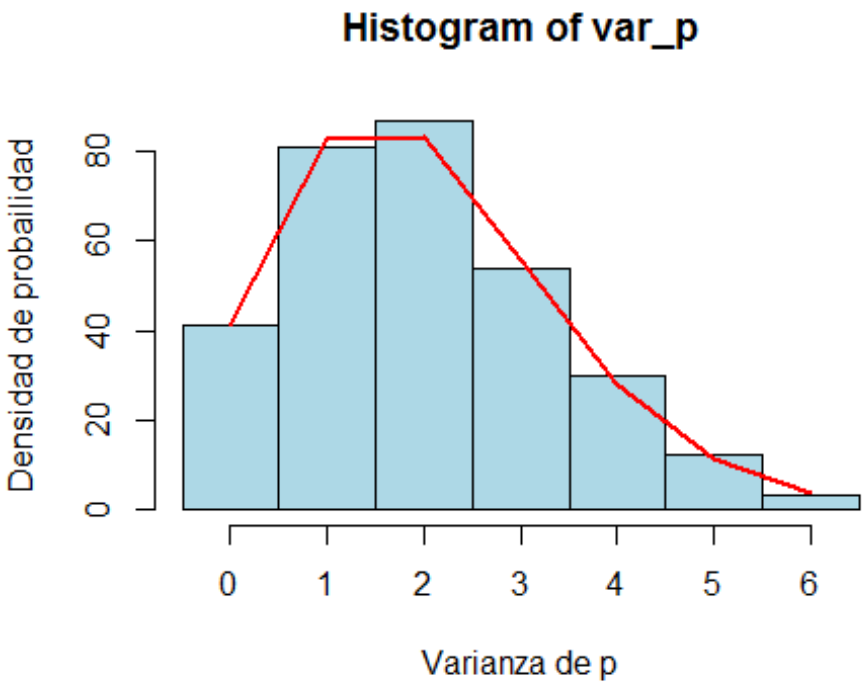
Ejercicio 1. Se desea contrastar si la distribución que muestra las solicitudes de crédito recibidas en una sucursal bancaria en 308 días sigue o no una distribución de Poisson. Utilizar para el contraste un nivel de significación del 5%.

Número de Solicitudes	Número de Días
0	41
1	81
2	87
3	54
4	30
5	12
6	3

```
library(sqldf)

## Loading required package: gsubfn

## Loading required package: proto

## Loading required package: RSQLite

library(MASS)
library(vcd)

## Loading required package: grid

library(knitr)

#se desea contrastar
dias<-c(0,1,2,3,4,5,6)
solicitudes<-c(41,81,87,54,30,12,3)
var_poisson<-data.frame(dias,solicitudes);
kable(var_poisson)
```

dias	solicitudes
0	41
1	81
2	87
3	54
4	30
5	12
6	3

```
var_p<-c(rep(dias,solicitudes))

ajuste_pois<-goodfit(var_p, type="poisson", method = "MinChisq")
summary(ajuste_pois)

##
##   Goodness-of-fit test for poisson distribution
##
##               X^2 df P(> X^2)
## Pearson 1.385323  5 0.925912

ajuste_pois$par

## $lambda
## [1] 2.007465

#Graficar

hist(var_p, breaks=(-0.5:6.5), col="lightblue", freq=T,
     xlab="Varianza de p",
     ylab ="Densidad de probailidad")
points(0:6, ajuste_pois$fitted, col="red", type="l", lwd=2)
```



Ejercicio 6.2

Ejercicio 2: Se realiza un muestreo de plantas que han sido tratadas con tres tipos de fertilizantes diferentes y se analiza si han florecido, obteniéndose los resultados que refleja la siguiente tabla:

	Fertilizante A	Fertilizante B	Fertilizante C
Han Florecido	34	73	63
No Han Florecido	16	12	12

Contrastar si existe o no relación entre el tipo de fertilizante empleado y la presencia o ausencia de floración. Utilizar para el contraste un nivel de significación del 5%. Razonar la respuesta.

```
library(sqldf)

## Loading required package: gsubfn

## Loading required package: proto

## Loading required package: RSQLite

library(MASS)
library(vcd)

## Loading required package: grid

library(knitr)

XY<-matrix(c(34,73,63,16,12,12),ncol=3, nrow=2, byrow=T)
colnames(XY)<-c("Fert. A", "Fert. B", "Fert. C")
rownames(XY)<-c("Han florecido", "No han florecido")
tabla<-as.table(XY)
kable(tabla)
```

	Fert. A	Fert. B	Fert. C
Han florecido	34	73	63
No han florecido	16	12	12

```
ampliada<-addmargins(tabla)
kable(ampliada)
```

	Fert. A	Fert. B	Fert. C	Sum
Han florecido	34	73	63	170
No han florecido	16	12	12	40

Sum		50	85	75	210

Aplicacion del metodo del constraste de inpedendencia de datos.

```
ni<-ampliada[3,]
nj<-ampliada[,4]
N<-as.numeric(ampliada[3,4])

pXY<-tabla^2
suma<-0;
for (i in 1:3) {
  for (j in 1:2) {
    suma<-suma+as.numeric(pXY[j,i]/(ni[i]*nj[j]))
  }
}
chi2<-N*(suma-1)

chi2
```

```
## [1] 7.231557
```

Calculamos de la q chisq para verificar la independencia

```
gl<-((nrow(tabla)-1)*(ncol(tabla)-1))

qchisq(0.95,gl)
```

```
## [1] 5.991465
```

```
#ejecuamos el test chi para vaerficar lo anterior

resultado1<-chisq.test(tabla,correct = T)
resultado1
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabla
## X-squared = 7.2316, df = 2, p-value = 0.0269
```

```
#Como el valor de p es inferior a 0,5, y el valor de chi2 es may
or que el qchisq
#Asumimos que son dependientes. Entonces verificamos cual de tod
os da mejor rendimiento

prop_f<-100*(ampliada[1,]/ampliada[3,])
kable(prop_f)
```

	x
Fert. A	68.00000
Fert. B	85.88235

Fert. C 84.00000

Sum 80.95238

```
#El B da mejor rendimiento, luego nos quedamos con Ã©ste
```

Ejercicio 6.3

Ejercicio 3: El Cuadro siguiente contiene una tabla de contingencia basada en los datos de una muestra de estudiantes de Ingeniería Informática y de otras titulaciones de la ULPGC clasificados según el tiempo de uso de más de dos horas al día en redes sociales. ¿Se puede decir, a la luz de esos datos, que existe una relación significativa entre el uso de redes sociales y que sean o no estudiantes de Ingeniería Informática?

	Estudiantes II	Otros Títulos
Uso de más de dos horas	75	73
Uso de menos de dos horas	15	32

```
library(sqldf)

## Loading required package: gsubfn

## Loading required package: proto

## Loading required package: RSQLite

library(MASS)
library(vcd)

## Loading required package: grid

library(knitr)

XY<-matrix(c(75,73,15,32),ncol=2, nrow=2, byrow=T)
colnames(XY)<-c("Estudiantes II", "Otros tÃ慎猛tulos")
rownames(XY)<-c("MÃ¡s de dos horas", "Menos de dos horas")
tabla<-as.table(XY)
kable(tabla)
```

	Estudiantes II	Otros tÃ<U+00AD>tulos
MÃ¡s de dos horas	75	73
Menos de dos horas	15	32

```
ampliada<-addmargins(tabla)
kable(ampliada)
```

	Estudiantes II	Otros tÃ<U+00AD>tulos	Sum
MÃ¡s de dos horas	75	73	148
Menos de dos horas	15	32	47
Sum	90	105	195

Aplicacion del metodo del constraste de inpedendencia de datos.

```
ni<-ampliada[3,]
nj<-ampliada[,3]
```

Aplicamos la corrección de Yates, debido a que solo es un grado de libertad

```
N<-as.numeric(ampliada[3,3])

esperada<-tabla^2
suma<-0
for (i in 1:2) {
  for (j in 1:2) {
    esperada[i,j]<-((ni[j]*nj[i])/N)
    suma<-suma+((abs(tabla[i,j]-esperada[i,j])-0.5)^2)/esperada[
i,j]
  }
}
chi2<-suma
chi2

## [1] 4.325313
```

Calculamos de la qchisq para verificar la independencia

```
gl<-((nrow(tabla)-1)*(ncol(tabla)-1))

qchisq(0.95,gl)

## [1] 3.841459
```

Como el valor de chi2 es mayor que qchisq, descartamos la hipótesis de independencia de ambos

Ejecutamos el test chi para re-verificar lo anterior

```
resultado1<-chisq.test(tabla,correct = T)
resultado1

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tabla
## X-squared = 4.3253, df = 1, p-value = 0.03755
```

Es decir, existe dependencia ente la carrera y el uso del móvil.

Ejercicio 6.4

Ejercicio 4: El cuadro siguiente contiene una tabla donde se reflejan los resultados de dos radiólogos que analizan las mismas radiografías para determinar si un paciente se ha fracturado un brazo o no.

		Jefe de Servicio	
		Brazo Fracturado	Brazo Normal
Internista	Brazo Fracturado	103	12
	Brazo Normal	18	35

a) Explicar la aplicación del test de McNemar (***mcnemar.test***) para tablas de contingencia que tengan que ver con los resultados de dos pruebas sobre los mismos individuos (datos apareados).
b) ¿Se puede decir, a la luz de esos datos, que existe dependencia entre el médico que ha realizado el diagnóstico y el resultado del mismo?

```
library(sqldf)

## Loading required package: gsubfn

## Loading required package: proto

## Loading required package: RSQLite

library(MASS)
library(vcd)

## Loading required package: grid

library(knitr)

#EJERCICIO 4: Test de Mcnemar

XY<-matrix(c(103,12,18,35),ncol=2, nrow=2, byrow=T)
colnames(XY)<-c("Fracturado (Jefe)", "Brazo normal (Jefe)")
rownames(XY)<-c("Fracturado (Internista)", "Brazo normal (Intern
ista)")
tabla<-as.table(XY)
kable(tabla)
```

	Fracturado (Jefe)	Brazo normal (Jefe)
Fracturado (Internista)	103	12
Brazo normal (Internista)	18	35

Aplicacion del metodo de Mcnemar. Mide la simetria de la matriz

```
resultado_mcn<-mcnemar.test(tabla)
resultado_mcn
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  tabla
## McNemar's chi-squared = 0.83333, df = 1, p-value = 0.3613
```

Como la probabilidad es baja, se puede considerar que no existe simetría entre ambos Por ende, admitimos la hipótesis de simetría

```
resultado_chi2<-chisq.test(tabla)
resultado_chi2
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tabla
## X-squared = 52.941, df = 1, p-value = 3.437e-13
```

Al salir la probabilidad de independencia muy cercana al cero, se puede constatar que son dependientes ambos del tratamiento.

Ejercicio 6.5

Ejercicio 5. Se llevaron a cabo las pruebas con tres tratamientos (**A**, **B** y **C**) para una enfermedad infecciosa leve sobre tres grupos de pacientes. Además, se incluyó un grupo adicional, al cual se le suministró una medicación placebo (**P**). Estos tratamientos se valoran en función del tiempo de recuperación en días. Los resultados se indican en la tabla. Se pide estudiar si existen diferencias significativas entre los diferentes tratamientos utilizando el test de Kruskal-Wallis.

P	15	12	10	8	11	9	6	10		
A	7	8	9	8	7	10	9	8	7	10
B	8	9	8	6	7	8	9	8	7	6
C	10	12	10	8	9	11	10	9	8	

```
library(sqldf)
```

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

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

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

```
library(MASS)
library(vcd)
```

```
## Loading required package: grid
```

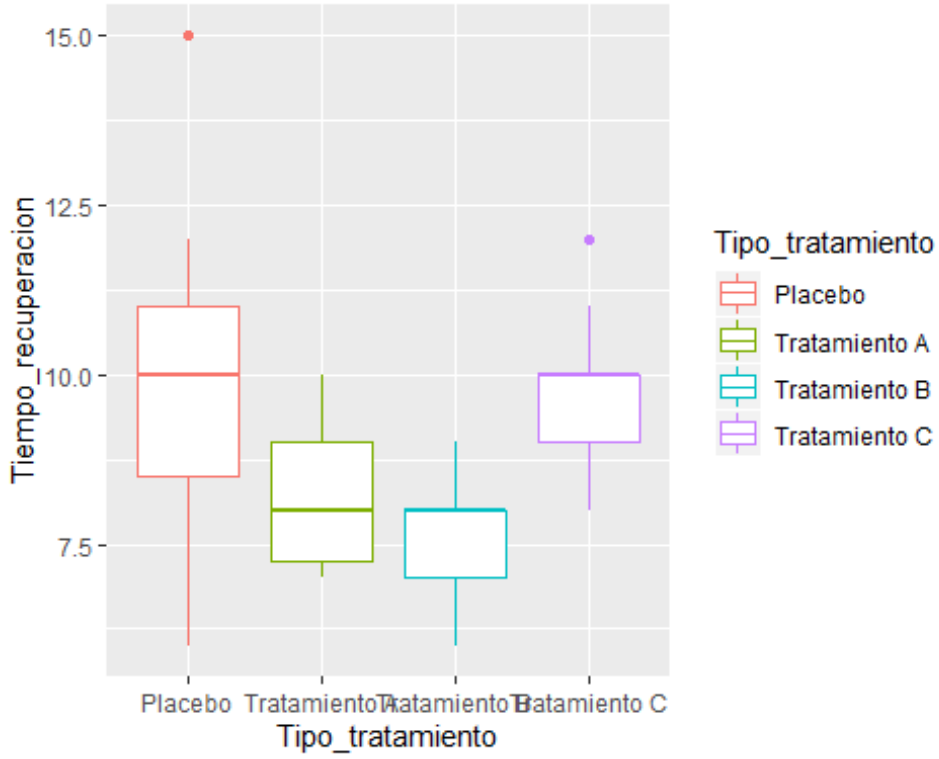
```
library(knitr)
library(ggplot2)

P<-c(15,12,10,8,9,6,10)
A<-c(7,8,9,8,7,10,9,8,7,10)
B<-c(8,9,8,6,7,8,9,8,7,6)
C<-c(10,12,10,8,9,11,10,9,8)
tratamientos_f<-factor(rep(1:4, c(length(P),length(A),length(B),
length(C))),
                       labels = c("Placebo","Tratamiento A","Tra
tamiento B","Tratamiento C"))
tratamientos_v<-c(P,A,B,C)
datos_t<-as.data.frame(tratamientos_v)
datos_t[,2]<-tratamientos_f
names(datos_t)<-c("Tiempo_recuperacion", "Tipo_tratamiento")
kable(datos_t)
```

Tiempo_recuperacion	Tipo_tratamiento
15	Placebo
12	Placebo
10	Placebo
8	Placebo
9	Placebo
6	Placebo
10	Placebo
7	Tratamiento A
8	Tratamiento A
9	Tratamiento A
8	Tratamiento A
7	Tratamiento A
10	Tratamiento A
9	Tratamiento A
8	Tratamiento A
7	Tratamiento A
10	Tratamiento A
8	Tratamiento B
9	Tratamiento B
8	Tratamiento B
6	Tratamiento B
7	Tratamiento B
8	Tratamiento B
9	Tratamiento B
8	Tratamiento B
7	Tratamiento B
6	Tratamiento B
10	Tratamiento C
12	Tratamiento C
10	Tratamiento C
8	Tratamiento C
9	Tratamiento C
11	Tratamiento C
10	Tratamiento C
9	Tratamiento C
8	Tratamiento C

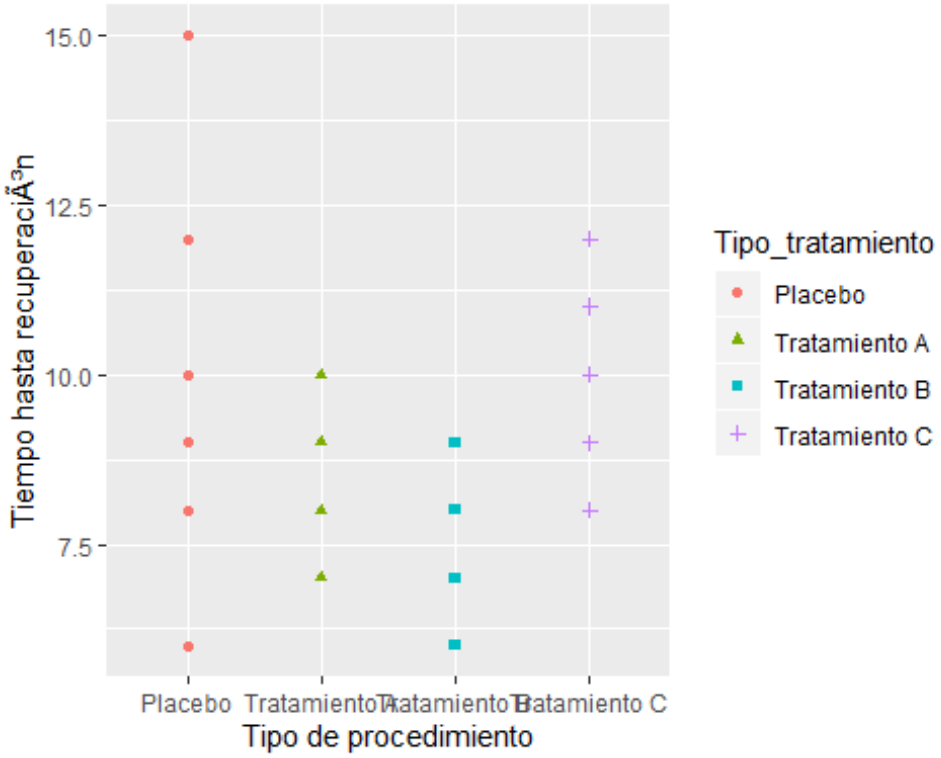
```
tabla<-table(datos_t)
kable(tabla)
```

	Placebo	Tratamiento A	Tratamiento B	Tratamiento C
6	1	0	2	0
7	0	3	2	0
8	1	3	4	2
9	1	2	2	2
10	2	2	0	3
11	0	0	0	1
12	1	0	0	1
15	1	0	0	0

```
g<-ggplot(data=datos_t, aes(x=Tipo_tratamiento,
                            y=Tiempo_recuperacion,
                            color=Tipo_tratamiento))
g+geom_boxplot()
```




```
g2<-g+xlab("Tipo de procedimiento")
g2<-g2+ylab("Tiempo hasta recuperaciÃ³n")
g2<-g2+geom_point(aes(shape=Tipo_tratamiento))
g2
```




```
attach(datos_t)
kruskal.test(Tiempo_recuperacion,Tipo_tratamiento, datos_tratami
ento)

##
##  Kruskal-Wallis rank sum test
##
## data:  Tiempo_recuperacion and Tipo_tratamiento
## Kruskal-Wallis chi-squared = 10.697, df = 3, p-value = 0.0134
8
```

Al salir la probabilidad de igualdad muy baja, concluimos definitivamente que son diferentes entre sí.

Verificamos cuál de las dos más bajas es mejor, si la A o la B, aplicaremos el test de Wilcoxon para verificar cambios entre ambas

```
wilcox.test(Tiempo_recuperacion[Tipo_tratamiento=="Tratamiento A
"],
            Tiempo_recuperacion[Tipo_tratamiento=="Tratamiento B
"],
            alternative = "two.sided")

## Warning in wilcox.test.default(Tiempo_recuperacion[Tipo_trata
miento ==
## "Tratamiento A"], : cannot compute exact p-value with ties

##
##  Wilcoxon rank sum test with continuity correction
##
```

```
## data:  Tiempo_recuperacion[Tipo_tratamiento == "Tratamiento A
"] and Tiempo_recuperacion[Tipo_tratamiento == "Tratamiento B"]
## W = 65, p-value = 0.2567
## alternative hypothesis: true location shift is not equal to 0
```

Ante una probabilidad de 0,25 según el test de Wilcoxon, se puede asumir que ambos tratamientos para las plantas producen resultados iguales.

Ejercicio 7.1

Ejercicio 1: El fichero "***Aloe_Vera.txt***" contiene datos de cuatro variedades de plantas de Aloe obtenidas de una plantación experimental.

a) Estudiar las variedades que dan más rendimiento desde el punto de vista de su masa y masa seca.

b) Analizar las dependencias entre la masa y la altura de la variedad "*barbadensis*"

c) Estimar el modelo de regresión con la función ***lm***.

d) Analizar el modelo estimado con la función ***summary*** y obtener un posible intervalo de confianza para las conclusiones de los distintos parámetros.

e) Evaluar una predicción para una masa de x_0=5.1 gramos y encontrar un intervalo de confianza para la misma.

f) Encontrar el coeficiente de determinación R^2

g) Realizar un análisis de varianza para estudiar la bondad del ajuste y la linealidad de la regresión. Explicar los resultados obtenidos.

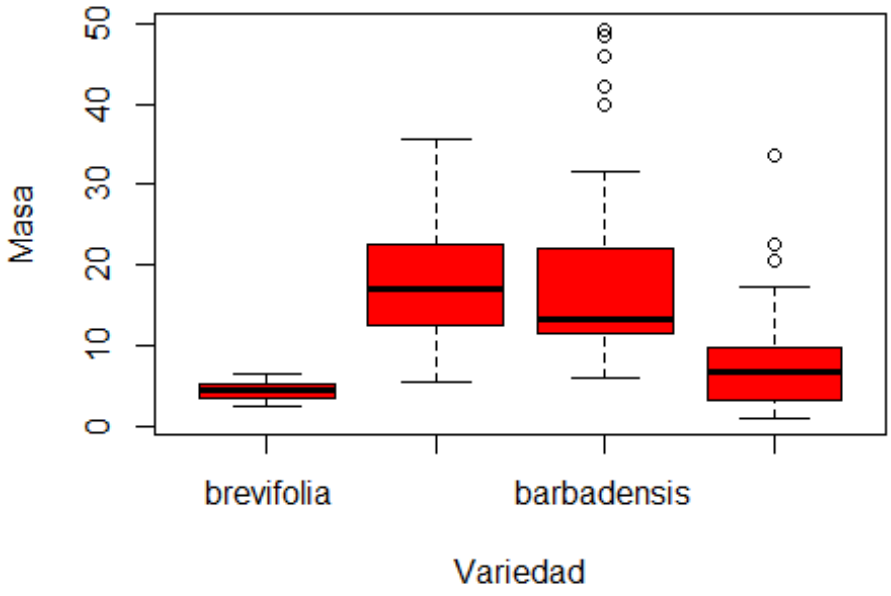
h) **(Opcional)** Analizar si fuera posible aplicar el estudio anterior y la suposición de homocedasticidad (varianza constante a lo largo de las observaciones) al caso de analizar las relaciones entre la masa y la masa seca para la variedad "*saponaria*". Utilizar el test de White (variedad del test de Breusch-Pagan ***bptest***, del paquete ***lmtest***). Explicar las conclusiones.

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

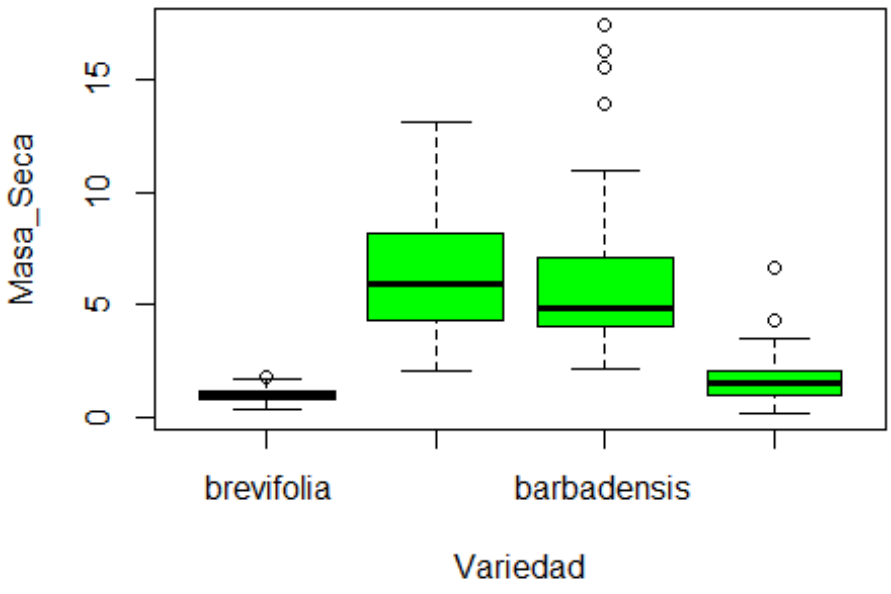
```
## Loading required package: gsubfn

## Loading required package: proto

## Loading required package: RSQLite

library(MASS)

datos<-read.table("Archivos/Aloe_Vera.txt", sep=",", dec=".")
attach(datos)
```

Apartado a)

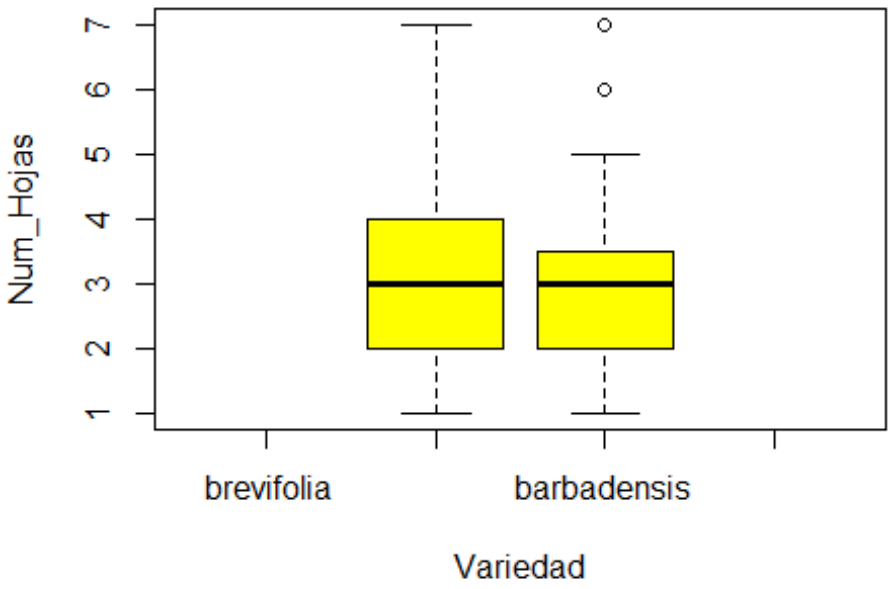
```
boxplot(Masa~Variedad, Variedad, col="red")
```



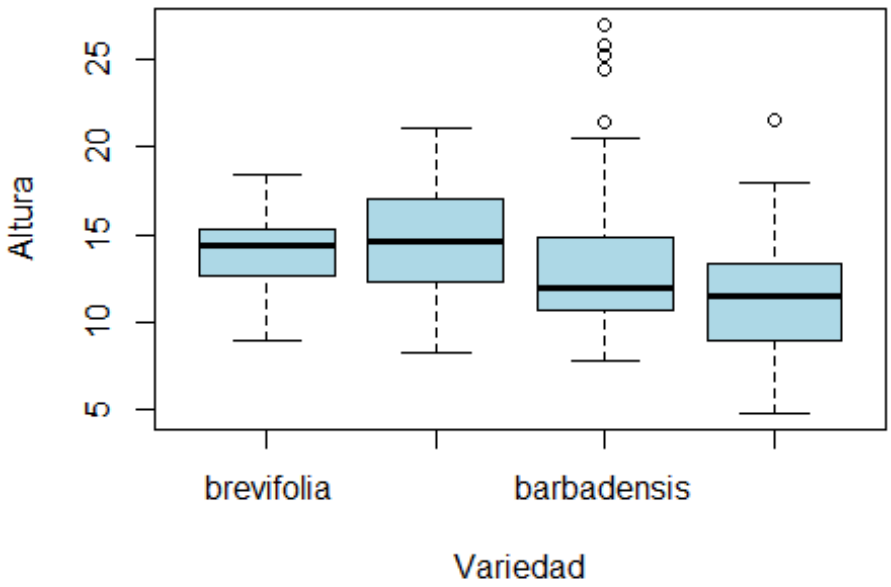

```
boxplot(Masa_Seca~Variedad, Variedad, col="green")
```




```
boxplot(Num_Hojas~Variedad, Variedad, col="yellow")
```




```
boxplot(Altura~Variedad, Variedad, col="lightblue")
```




```
masa_seca<-aggregate(Masa_Seca~Variedad, Variedad, mean)
masa_seca

##       Variedad Masa_Seca
## 1  brevifolia  1.010145
```

```
## 2 arborescens  6.350000
## 3 barbadensis  6.110714
## 4   saponaria  1.650000

masa<-aggregate(Masa~Variedad, Variedad, mean)
masa

##      Variedad      Masa
## 1  brevifolia  4.412500
## 2 arborescens 18.081481
## 3 barbadensis 17.576786
## 4   saponaria  7.521429
```

Apartado b)

```
Barbadensis<-subset(datos, subset=(Variedad == "barbadensis"))
detach()
attach(Barbadensis)

x<-Masa
y<-Altura
n<-length(x)
plot(Altura~Masa, data = Barbadensis, ylim=c(0,length(y)))
```

Apartado c)

```
modelobar<-lm(y~x)
abline(modelobar, col="red", lwd=2)
```

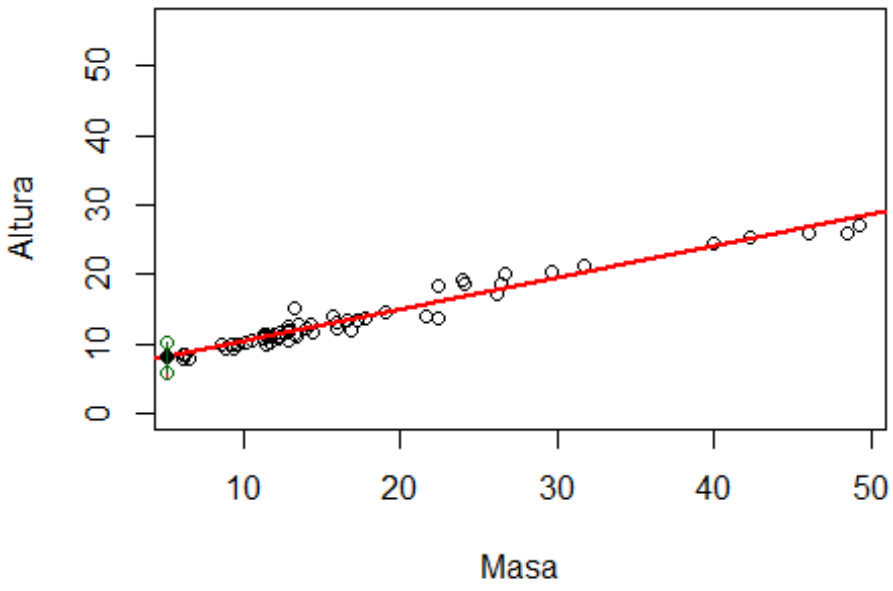
Apartado d)

```
summary(modelobar)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1730 -0.7323 -0.1087  0.4301  3.3263
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.74854    0.27944   20.57   <2e-16 ***
## x            0.45645    0.01366   33.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1
##
```

```
## Residual standard error: 1.07 on 54 degrees of freedom
## Multiple R-squared:  0.9539, Adjusted R-squared:  0.953
## F-statistic:  1117 on 1 and 54 DF,  p-value: < 2.2e-16

confint(modelobar)

##                 2.5 %   97.5 %
## (Intercept) 5.1883011 6.308774
## x           0.4290643 0.483832

#Calculo manual de los errores
e<-residuals(modelobar)
SCE<-sum(e^2)
SCE

## [1] 61.83836

s2<-SCE/(n-2)
s2

## [1] 1.145155

var<-sqrt(SCE/(n-2))
var

## [1] 1.070119
```

Apartado e)

```
x0<-5.1
prediccion_masa<-predict(modelobar,newdata = data.frame(x=x0))
prediccion_masa

##        1
## 8.076423

points(x0,prediccion_masa, pch=16,col="black")
lines(c(x0,x0), c(x0,prediccion_masa),col="red")

inter_prediccion<-predict(modelobar,level=0.95, newdata =
                            data.frame(x=x0), interval = "pred")
inter_prediccion

##        fit      lwr      upr
## 1 8.076423 5.885093 10.26775

lines(c(x0,x0), c(inter_prediccion[2],inter_prediccion[3]),col="
darkgreen")
points(c(x0,x0), c(inter_prediccion[2],inter_prediccion[3]),col=
"darkgreen")
```



Apartado f)

```
#Apartado f

STCC<-sum((y-mean(y))^2)
STCC

## [1] 1340.734

cdd<-1-(SCE/STCC)
cdd

## [1] 0.9538772

#Muy próximo a la unidad. muy buen ajuste
```

Apartado g)

```
x_factor<-as.factor(x)
datos2<-data.frame(x_factor,y)
detach(Barbadensis)

attach(datos2)

## The following objects are masked _by_ .GlobalEnv:
##
##     x_factor, y

Y_M_F<-rep(0,length(x_factor))
for (i in 1:length(x_factor)) {
```

```
  Y_M_F[i]<-mean(y[x_factor==x[i]])
}

SCEpuro <- sum((y-Y_M_F)^2)

#Error falta de ajuste
SC <- SCE-SCEpuro
SC

## [1] 48.32169

#Calculo del s2puro
k<-nlevels(x_factor)
s2puro<-SCEpuro/(n-k)
s2puro

## [1] 1.501852

fSC<-SC/(s2puro*(k-2))
fSC

## [1] 0.7149942

1-pf(fSC,1,k-2)

## [1] 0.4022647
```

CUESTIONES LECTURAS

Cuestión 8.1

Cuestión 1: Se toma una muestra aleatoria simple de una población que sigue una distribución $N(\mu, \sigma^2)$, donde μ y σ son desconocidas. Los valores obtenidos son:

3.58, 10.03, 4.77, 9.71, 10.4, 14.66, 8.45, 5.4, 9.75, 10.1

Utilizando α= 0.05:

a) ¿Hay evidencias para pensar que la media de la población sea mayor o igual que 10? Razonar la respuesta.
b) ¿Podría afirmarse con los datos que la media de la población es inferior a 10?
c) Calcular los errores tipo I, tipo II y la potencia de la prueba en su caso.

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

## Warning: package 'pwr' was built under R version 3.6.2

#Mu y sigma son desconocidas, distribucion N(mu,sigma^2)
datos <- c(3.58,10.03,4.77,9.71,10.4,14.66,8.45,5.4,9.75,10.1)
alfa <- 0.05
n <- 10
```

Apartado a)

```
# - H0 = mu >= 10
# - H1 = mu < 10
xm <- mean(datos)
sigma <- var(datos)
mu <- 10
t <- (xm-mu)/sigma/sqrt(n)
t

## [1] -0.03863377

qt(0.05,9) # Se rechaza la hipótesis, pues qt es menor que t

## [1] -1.833113
```

Apartado b)

```
# - H0 = mu < 10,
# - H1 = mu >= 10
qt(0.95,9) # Como qt por el otro extremo es mayor que t se acept
a la hipotesis nula

## [1] 1.833113
```

Apartado c)

Calculo Errores Tipo I y II y la Potencia de la prueba

```
# El Error tipo I es alfa = 0.05
t1<-alfa

# La potencia viene dada por 1-beta. Se puede extraer con pwr.t.
test
potencia<-pwr.t.test(mu,1,0.05,type="one.sample")
potencia

##
##      One-sample t test power calculation
##
##               n = 10
##               d = 1
##       sig.level = 0.05
##           power = 0.8030969
##     alternative = two.sided

# beta sale de la potencia t = 1-beta
beta <- 1-t
beta + potencia$power

## [1] 1.841731
```

Conclusiones

a) Debido a que la *t* de Student calculada es mayor que el resultado de la prueba del que a un intervalo de confianza al 90%, se rechaza por conclusión la hipótesis nula, luego la población podría ser menor de 10 elementos.

b) En el segundo caso, la t calculada corrobora la teoría indicada en el punto anterior, puesto que el resultante 1,83 es mayor que el -0.03 resultado del cálculo de *t*.

Cuestión 8.2

Cuestión 2: En un ayuntamiento de la Isla de Gran Canaria se sospecha que se está produciendo una discriminación salarial de sus empleadas dentro de una determinada categoría y antigüedad laboral. Para analizar el hecho se ha decidido tomar muestras simples e independientes una de 16 empleados públicos varones y otra de 10 empleadas, y se les preguntó sobre su salario percibido en euros. Los datos se recogen en la siguiente tabla:

Estadístico	Empleados	Empleadas
Media ($\bar{X}$)	1.515,60	1.298,35
Varianza (S^2)	61.500	90.201

a) Establecer un intervalo de confianza al 95% para la diferencia de los salarios entre empleados y empleadas públicas en este ayuntamiento
b) ¿Cuáles serían las diferencias de los límites si se establece al 90%? Razonar la respuesta.
c) A partir del resultado de a) razonar sobre la existencia de discriminación salarial entre hombres y mujeres en el ayuntamiento de referencia
d) ¿A qué conclusiones se llegaría si los tamaños de las muestras fueran de 45 para los empleados y de 30 para las empleadas? ¿sería diferentes? Razonar la respuesta

```
setwd(".")
library(knitr)
n1 <- 16
x1 <- 1515.60
s1 <- sqrt(61.500)
n2 <- 10
x2 <- 1298.35
s2 <- sqrt(90.201)
```

Apartado a),

Intervalo confianza del 95%

```
a<- ((s1^2/n1)+(s2^2/n2))^2
b<- (s1^2/n1)^2/(n1-1)
c<- (s2^2/n2)^2/(n2-1)
v<-a/(b+c)

izquierda005<-qt(0.025,v)
derecha005<-qt(0.975,v)

t<-(x1-x2)/sqrt((s1^2/n1)+(s2^2/n2))
```

Apartados b) y c)

Intervalo de confianza del 90%

```
izquierda010<-qt(0.05,v)
derecha010<-qt(0.95,v)

t

## [1] 60.57233

izquierda005

## [1] -2.114633

derecha005

## [1] 2.114633
```

Apartado d)

```
n1 <- 45
n2 <- 30

a2<- (s1^2/n1+s2^2/n2)^2
b2<- (s1^2/n1)^2/(n1-1)
c2<- (s2^2/n2)^2/(n2-1)
v2<-a2/(b2+c2)

izquierda005_2<-qt(0.025,v2)
derecha005_2<-qt(0.975,v2)

t2<-(x1-x2)/sqrt((s1^2/n1)+(s2^2/n2))

t2

## [1] 103.8848

izquierda005_2

## [1] -2.004878

derecha005_2

## [1] 2.004878
```

Conclusiones

a) Según las pruebas realizadas en los apartados a y b se pude afirmar que con la población indicada y los datos de salario, no existe, a priori, discriminación salarial.
b) Ídem respecto al apartado a)

Cuestión 8.3

Cuestión 3: Se desea conocer la media y la dispersión de las rentas mensuales de los habitantes del barrio de Vegueta en la ciudad de Las Palmas de Gran Canaria con un nivel de significación del 5%. Para ello se realizó una muestra aleatoria simple en la que se observaron las rentas mensuales en euros de los vecinos que se detallan en la siguiente tabla:

Rentas Mensuales (en €)		
1500,21	880,66	605,22
1210,12	2010,1	701,30
2060,01	810,10	1012,34
1500,08	2500,00	917,45
890,50	515,01	820,39
1800,30	625,12	1002,20
2015,22	720,25	1102,45
3200,00	1601,79	1219,70
1005,40	2150,1	623,56

a) Encontrar el correspondiente intervalo de confianza de dos colas para la media de rentas. Razonar la respuesta.
b) ¿Supera, con el nivel de significancia referido, los 1000 euros la desviación típica de las rentas mensuales de los habitantes del barrio? Justificar adecuadamente la respuesta y fundamentarla desde un punto de vista teórico.

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

Apartado a)

```
datos<-c(1500.21, 880.66, 605.22, 1210.12, 2010.1, 701.30, 2060.
01, 810.10,
         1012.34, 1500.08, 2500.00, 917.45, 890.50, 515.01, 820.
39, 1800.30,
         625.12, 1002.20, 2015.22, 720.25, 1102.45, 3200.00, 160
1.79, 1219.70,
         1005.40, 2150.1, 623.56)
kable(datos)
```

x
1500.21
880.66
605.22
1210.12
2010.10
701.30
2060.01
810.10

```
 1012.34
 1500.08
 2500.00
  917.45
  890.50
  515.01
  820.39
 1800.30
  625.12
 1002.20
 2015.22
  720.25
 1102.45
 3200.00
 1601.79
 1219.70
 1005.40
 2150.10
  623.56
```

```r
# Media
mu<-mean(datos)
mu
```

```
## [1] 1296.281
```

```r
# Desviación
sigma<-sd(datos)
sigma
```

```
## [1] 672.6922
```

```r
# Tamaño datos
n<-length(datos)
n
```

```
## [1] 27
```

```r
# Intervalo de confianza inferior y superior
confInf<-mu-qt(0.95, df=(n-1))*sigma/sqrt(n)
confInf
```

```
## [1] 1075.472
```

```r
confSup<-mu+qt(0.95, df=(n-1))*sigma/sqrt(n)
confSup
```

```
## [1] 1517.089
```

Apartado b)

```
a<-qchisq(0.95,26)
b<-a*a
c<-b*26
d<-1000*1000
res<-c/d
res*100

## [1] 3.93134
```

Conclusiones

Se trata en primera instancia de una prueba tipo Xi cuadrado. Si el resultado de los cálculos es menor que 1 se considerará correcto, esto es,

Si, por el contrario, el resultado de los cálculos es mayor que 1 se considerará erróneo o inverosímil, ergo la hipótesis enunciada se rechaza.

La fórmula de la xi cuadrado se ha empleado en el ejercicio queda, por tanto, tal que:

$$(\mathrm{n}-1)*\left(\frac{s^2}{\sigma^2}\right)$$

Sabiendo esto, al ser mayor de 1 el resultado (concretamente 3,93), se rechaza la hipótesis nula, es decir, la desviación típica no supera los 1000 euros en el barrio.

Cuestión 8.4

Cuestión 4: El propietario de un vehículo híbrido de la marca Toyota piensa que el consumo medio de gasolina, en circuito combinado de carretera-ciudad, es superior a los 5,35 litros cada 100 km que es lo que los distribuidores de la marca publicitaban y que le impulsaron a decidir su compra. Para analizar su decisión ha realizado las siguientes medidas aleatorias de consumos medios cada 100 km durante el año 2018:

6.2, 6.6, 5.8, 5.4, 5.3, 6.15, 6.68, 7.0, 5.8, 5.6, 5.85, 6.2, 6.4, 6.75, 5.3, 6.3

a) Con un nivel de significancia del 1% analizar si fue una decisión correcta y fundada la adquisición del vehículo por tener un consumo medio de 5.35 l/100km.
b) ¿Cuántas ocasiones debería observarse a lo largo del año 2019 el consumo medio para que, con una probabilidad de 0.99, se detectase un consumo medio de 6.0 litros por cada 100 km?, ¿Sería posible hacer el análisis en condiciones con un recorrido anual de 25.000 km? Explicar y documentar teóricamente las respuestas.
c) Responder al apartado b) pero con una probabilidad del 0.90. Razonar la respuesta.

```
setwd(".")
library(knitr)
datos <- c(6.2,6.6,5.8,5.4,5.3,6.15,6.68,7.0,5.8,5.6,5.85,6.2,6.
4,6.75,5.3,6.3)
n <- length(datos)
pro <- 5.35
media <- mean(datos)
s <- var(datos)
sd_c <-sd(datos)
```

Apartado a)

- alfa a 0.01
- H0 -> mu >= 5.35
- H1 -> mu != 5.35

```
alfa <- 0.01
t <- (media-pro)/s/sqrt(n)
t

## [1] 0.6500344

qt <- qt(0.01,15)
```

Apartado b)

prob = 0.99.

```
pro2 <- 6.0
d_c <- (pro2-pro)/s
```

```
b <- 0.99
power.t.test(n = NULL, delta = d_c,sd = sd_c,sig.level = 0.01,po
wer = b,alternative = "one.sided")

##
##      Two-sample t test power calculation
##
##               n = 4.049214
##           delta = 2.305322
##              sd = 0.5309955
##       sig.level = 0.01
##           power = 0.99
##     alternative = one.sided
##
## NOTE: n is number in *each* group
```

Apartado c)

```
b_2 <- 0.90
power.t.test(n = NULL, delta = d_c,sd = sd_c,sig.level = 0.01,po
wer = b_2,alternative = "one.sided")

##
##      Two-sample t test power calculation
##
##               n = 3.14055
##           delta = 2.305322
##              sd = 0.5309955
##       sig.level = 0.01
##           power = 0.9
##     alternative = one.sided
##
## NOTE: n is number in *each* group
```

Conclusiones

- Se acepta h0 puesto que la proximidad a 0 es mayor que qt.
- A una probabilidad el 99%, el power test da un tamaño de muestra de aprox 4 miembros.
- A una probabilidad el 90%, el power test da un tamaño de muestra de aprox 3 miembros.

Cuestión 9.1

Cuestión 1: El cuadro siguiente contiene una tabla de contingencia basada en los datos de una encuesta de una muestra de hombres y mujeres de clasificados por su interés en participar activamente en la vida política.

	Hombres	Mujeres
Interesadas/os	35	31
No Interesadas/os	47	55

a) ¿Se puede decir, a la luz de esos datos, que existe una relación significativa entre el género y esa clasificación? Justificar experimental y teóricamente las respuestas.

b) Desarrollar en ***R*** una función propia (con opciones según los casos) para realizar las pruebas de verificación de este tipo de hipótesis y contrastar su efectividad con las funciones que ya incorpora ***R*** para las mismas.

```
library(knitr)
Tabla<-matrix(c(35,31,47,55),2,2,byrow=TRUE)
colnames(Tabla)<-c("Hombre","Mujer"); rownames(Tabla)<-c("Intere
sados/as","No Interesadas/os")
Tabla<-as.table(Tabla)
kable(Tabla)
```

	Hombre	Mujer
Interesados/as	35	31
No Interesadas/os	47	55

```
#Se distribuyen con una distribucion:
#v = (2-1)(2-1) = 1 grado de libertad
qchisq(0.95,1)

## [1] 3.841459
```

Apartado a)

Pruebas de homogeneidad de forma teórica

```
XY<-matrix(c(35,47,31,55),
           ncol=2,nrow = 2)
colnames(XY)<-c("Hombre", "Mujer")
rownames(XY)<-c("Interesados/as","No Interesados/as")
XY

##                    Hombre Mujer
## Interesados/as         35    31
## No Interesados/as      47    55

pXY<-matrix(c(35/66*35/82,47/102*47/82, 31/66*31/86,55/102*55/86
),
```

```
            ncol=2,nrow = 2)
sum(pXY)

## [1] 1.004614

CHI2<-sum(XY)*(sum(pXY)-1)
CHI2

## [1] 0.7750765
```

Región Crítica

```
gl<-(nrow(XY)-1)*(ncol(XY)-1)
qchisq(0.95,gl)

## [1] 3.841459

CHI2<-sum(XY)*(sum(pXY)-1)
CHI2

## [1] 0.7750765
```

Región Crítica

```
gl<-(nrow(XY)-1)*(ncol(XY)-1)
qchisq(0.95,gl)

## [1] 3.841459

test<-chisq.test(XY)
test

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  XY
## X-squared = 0.52181, df = 1, p-value = 0.4701
```

Mediante la prueba teórica se podría asumir que la hipótesis de homogeneidad es aceptable debido a que el valor del estadístico de contraste es de casi 0,78, respecto a un grado de libertad (matriz de 2x2).

Asimismo, el valor de probabilidad es inferior al valor de confiabilidad (0,05), luego se puede admitir la hipótesis nula de que existe homogeneidades decir, los porcentajes serán equivalentes en cierta medida independientemente de si son hombres o mujeres

Apartado b)

Repetición de la prueba a partir del test Macnemar, aplicando en su caso la corrección de Yates para evitar el problema de las matrices cuadradas de N=2

```
result <- mcnemar.test(Tabla, correct = T)
result
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  Tabla
## McNemar's chi-squared = 2.8846, df = 1, p-value = 0.08943
```

Puesto que el valor de probabilidad excede al de (confianza 0.08943 > 0.05) la hipótesis nula es admisible se admite la hipótesis nula de existencia de simetría

El valor para 1-a de la distribución x2 con 1 grados de libertad es:

```
qchisq(0.95, 1)
```

```
## [1] 3.841459
```

```
x2 <- (47-32)^2/(47+31)
x2
```

```
## [1] 2.884615
```

Y para x2= 2.884615 con 1 grado de libertad.

```
res <- 1-pchisq(2.884615,1)
res
```

```
## [1] 0.08942938
```

Cómo 0.08942938 > 0.05 concluimos que la hipótesis nula se cumple, en consecuencia también la simetría

```
result2 <- chisq.test(Tabla, correct = TRUE)
result2
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  Tabla
## X-squared = 0.52181, df = 1, p-value = 0.4701
```

Como 0.4701 > 0.05

Cuestión 9.2

Cuestión 2: Las calificaciones que un grupo de 30 estudiantes de Ingeniería Informática han obtenido en las asignaturas de Álgebra y Programación en el curso 2017-18 se recogen a siguiente tabla:

N.º Estudiante	1	2	3	4	5	6	7	8	9	10
Álgebra	5.7	8.6	3.6	1.5	8.8	5.9	4.9	8.6	7.6	5.0
Programación	5.0	7.0	5.2	1.3	7.2	6.6	3.1	8.6	6.0	6.1

N.º Estudiante	11	12	13	14	15	16	17	18	19	20
Álgebra	7.7	2.6	8.6	7.5	5.8	6.2	9.9	7.1	5.6	6.2
Programación	8.0	5.0	9.2	7.3	4.2	6.6	9.1	7.6	4.0	5.1

N.º Estudiante	21	22	23	24	25	26	27	28	29	30
Álgebra	7.6	6.5	6.7	4.5	4.8	6.9	8.9	2.6	5.5	7.0
Programación	8.0	8.1	9.1	4.5	3.2	7.6	7.1	4.6	6.0	5.8

a) Analizar si los resultados, como medida de progreso, con ambas materias pueden considerarse equivalentes y tienen las mismas calificaciones medias. Utilizar un nivel de significancia de 0.05. Razonar y fundamentar teóricamente las respuestas.

b) Realizar un programa en ***R*** para llevar a cabo la prueba estadística necesaria y contrastarlo con las funciones que permite ***R*** para este tipo de pruebas.

Apartados a) y b)

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

notas_al <- c(5.7,8.6,3.6,1.5,8.8,5.9,4.9,8.6,7.6,5.0,
              7.7,2.6,8.6,7.5,5.8,6.2,9.9,7.1,5.6,6.2,
              7.6,6.5,6.7,4.5,4.8,6.9,8.9,2.6,5.5,7.0)
n1 <- length(notas_al)

notas_pr <- c(5.0,7.0,5.2,1.3,7.2,6.6,3.1,8.6,6.0,6.1,
                  8.0,5.0,9.2,7.3,4.2,6.6,9.1,7.6,4.0,5.1,
                  8.0,8.1,9.1,4.5,3.2,7.6,7.1,4.6,6.0,5.8)
n2 <- length(notas_pr)
alfa <- 0.05
```

Cálculo manual de las notas

```
notas_al_ord <- sort(notas_al)
notas_pr_ord <- sort(notas_pr)
notas_al_ord
```

```
##  [1] 1.5 2.6 2.6 3.6 4.5 4.8 4.9 5.0 5.5 5.6 5.7 5.8 5.9 6.2
6.2 6.5 6.7
## [18] 6.9 7.0 7.1 7.5 7.6 7.6 7.7 8.6 8.6 8.6 8.8 8.9 9.9
```

```
notas_pr_ord
```

```
##  [1] 1.3 3.1 3.2 4.0 4.2 4.5 4.6 5.0 5.0 5.1 5.2 5.8 6.0 6.0
6.1 6.6 6.6
## [18] 7.0 7.1 7.2 7.3 7.6 7.6 8.0 8.0 8.1 8.6 9.1 9.1 9.2
```

Valores de los rangos calculados a mano

```
R1 <- 808.5
R2 <- 849.5

U1 <- n1*n2+((n1*(n1+1))/2)-R1
U2 <- n1*n2+((n2*(n2+1))/2)-R2
U1

## [1] 556.5

U2

## [1] 515.5

mu_U <- (n1*n2)/2
sigma_U_2 <- (n1*n2*(n1+n2+1))/12
Z <- (U2-mu_U)/sqrt(sigma_U_2)
Z

## [1] 0.9683799

qnorm(0.025)

## [1] -1.959964

qnorm(0.975)

## [1] 1.959964
```

Usando wilcox.test Quitamos los 0 de X-Y

```
resta <- notas_al - notas_pr
resta

##  [1]  0.7  1.6 -1.6  0.2  1.6 -0.7  1.8  0.0  1.6 -1.1 -0.3 -
2.4 -0.6  0.2
## [15]  1.6 -0.4  0.8 -0.5  1.6  1.1 -0.4 -1.6 -2.4  0.0  1.6 -
0.7  1.8 -2.0
## [29] -0.5  1.2

notas_al_new <- c(5.7,8.6,3.6,1.5,8.8,5.9,4.9,7.6,5.0,
             7.7,2.6,8.6,7.5,5.8,6.2,9.9,7.1,5.6,6.2,
             7.6,6.5,6.7,4.8,6.9,8.9,2.6,5.5,7.0)

notas_pr_new <- c(5.0,7.0,5.2,1.3,7.2,6.6,3.1,6.0,6.1,
                  8.0,5.0,9.2,7.3,4.2,6.6,9.1,7.6,4.0,5.1,
                  8.0,8.1,9.1,3.2,7.6,7.1,4.6,6.0,5.8)

wilcox.test(notas_al_new,notas_pr_new,paired = T,exact = F)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  notas_al_new and notas_pr_new
## V = 219.5, p-value = 0.7153
## alternative hypothesis: true location shift is not equal to 0
```

Conclusiones

La prueba de los rangos con signo de Wilcoxon es una prueba no paramétrica para comparar el rango medio de dos muestras relacionadas y determinar si existen diferencias entre ellas. Se utiliza como alternativa a la prueba t de Student cuando no se puede suponer la normalidad de dichas muestras.

En el apartado del cálculo manual se considera que, al estar el valor de z (0.96837) en el intervalo de confianza (-1.95, 1.95), la hipótesis nula es susceptible de aceptación.

Sin embargo, esto sería completamente irrevocable mediante una segunda prueba de comprobación, la cual ejecutamos mediante la prueba de Wilcoxon, antes mencionada.

Como el valor de probabilidad calculado en el test de Wilcoxon se dispara a casi el 72%, hay motivos suficientes para decir que se acepta de forma definitiva la hipótesis nula.

Luego si, los resultados pueden considerarse equivalentes, ergo sus medias serán equivalentes y pueden ser empleado como media de progreso de la clase.

Cuestión 9.3

Cuestión 3: La tabla siguiente muestra las cantidades (en gramos) de cuatro tipos diferentes de grasa absorbidas por rosquillas desde un análisis experimental en un laboratorio de la ULPGC. Cada medida se corresponde con seis lotes de rosquillas.

Tipos de Grasa			
A	**B**	**C**	**D**
164	178	175	155
172	191	193	166
168	197	178	149
177	182	171	164
195	177	176	168
156	185	163	170

Una empresa del sector de la alimentación del polígono de Arinaga quiere utilizar estos tipos de grasa.

a) ¿Existen diferencias significativas ente ellas? Justificar la respuesta y el tipo de prueba estadística empleada.
b) En su caso, y si tuvieran costes similares, ¿Qué tipo de grasa se recomendaría utilizar? Utilizar ***R*** en los cálculos.

```r
library(resample)
#DATOS
grasa_A<-c(164,172,168,177,195,156)
grasa_B<-c(178,191,197,182,177,185)
grasa_C<-c(175,193,178,171,176,163)
grasa_D<-c(155,166,149,164,168,170)

##CREAMOS LAS POBLACIONES
#MEDIAS
mu_A<-mean(grasa_A)
mu_B<-mean(grasa_B)
mu_C<-mean(grasa_C)
mu_D<-mean(grasa_D)

#DESVIACIONES TIPICAS
sigma_A<-sqrt(var(grasa_A))
sigma_B<-sqrt(var(grasa_B))
sigma_C<-sqrt(var(grasa_C))
sigma_D<-sqrt(var(grasa_D))

#POBLACIONES
pobA<-rnorm(grasa_A,mean=mu_A,sd=sigma_A)
pobB<-rnorm(grasa_B,mean=mu_B,sd=sigma_B)
```

```
pobC<-rnorm(grasa_C,mean=mu_C,sd=sigma_C)
pobD<-rnorm(grasa_D,mean=mu_D,sd=sigma_D)

#HACEMOS LAS PRUEBAS Y LAS VISUALIZAMOS
size<-length(grasa_A)
A<-sample(pobA,size)
B<-sample(pobB,size)
C<-sample(pobC,size)
D<-sample(pobD,size)

#HALLAMOS LAS DIFERENCIAS DE MEDIAS ENTRE GRUPOS
AB<-mean(A)-mean(B)
AC<-mean(A)-mean(C)
AD<-mean(A)-mean(D)
BC<-mean(B)-mean(C)
BD<-mean(B)-mean(D)
CD<-mean(C)-mean(D)

permutaciones<-choose(4*size,size)
permutaciones
```

```
## [1] 134596
```

Apartado a)

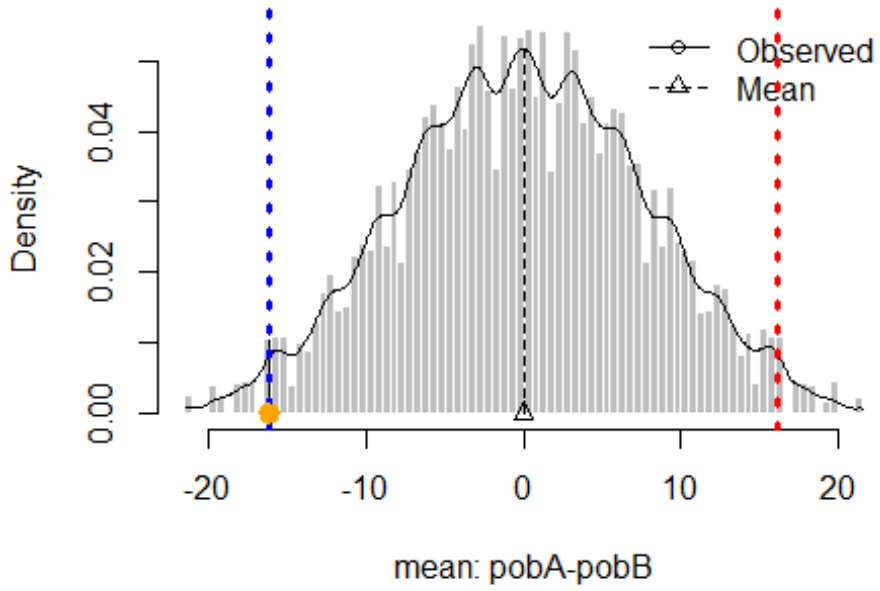
Emparejando las muestras con el test de permutación empleado en R, aplicamos las permutaciones para cada pareja de sustancias grasas y vemos la distribución resultante entre sus medias a ver si existen diferencias radicales.

Media A con B

```
TEST1<-permutationTest2(data = pobA,mean, data2 = pobB, R=permut
aciones, seed =35200)
TEST1
```

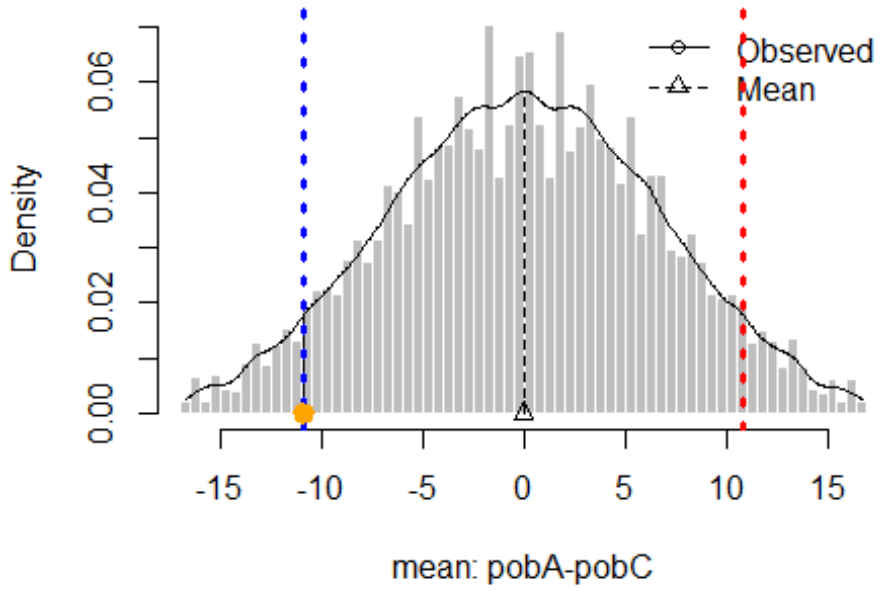
```
## Call:
## permutationTest2(data = pobA, statistic = mean, data2 = pobB,
##     R = permutaciones, seed = 35200)
## Replications: 134596
## Two samples, sample sizes are 6 6
##
## Summary Statistics for the difference between samples 1 and 2
:
##                   Observed        Mean Alternative     PValue
## mean: pobA-pobB -16.16903 -0.02415416   two.sided 0.03133799
```

```
plot(TEST1, col="grey")
abline(v=AB,lty=3,col="blue",lwd=3)
abline(v=-AB,lty=3,col="red",lwd=3)
points(AB,0,lwd=5, col="orange", pch=19)
```



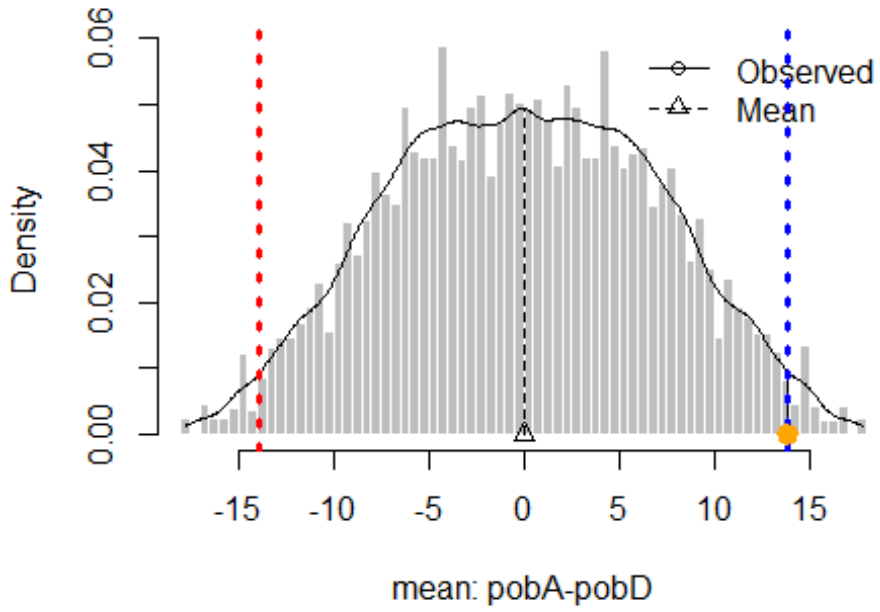
Media A con C

```
TEST2<-permutationTest2(data = pobA,mean, data2 = pobC, R=permut
aciones, seed =35200)
TEST2

## Call:
## permutationTest2(data = pobA, statistic = mean, data2 = pobC,
##     R = permutaciones, seed = 35200)
## Replications: 134596
## Two samples, sample sizes are 6 6
##
## Summary Statistics for the difference between samples 1 and 2
:
##                   Observed        Mean Alternative    PValue
## mean: pobA-pobC -10.85388 -0.01302978   two.sided 0.1043262

plot(TEST2, col="grey")
abline(v=AC,lty=3,col="blue",lwd=3)
abline(v=-AC,lty=3,col="red",lwd=3)
points(AC,0,lwd=5, col="orange", pch=19)
```



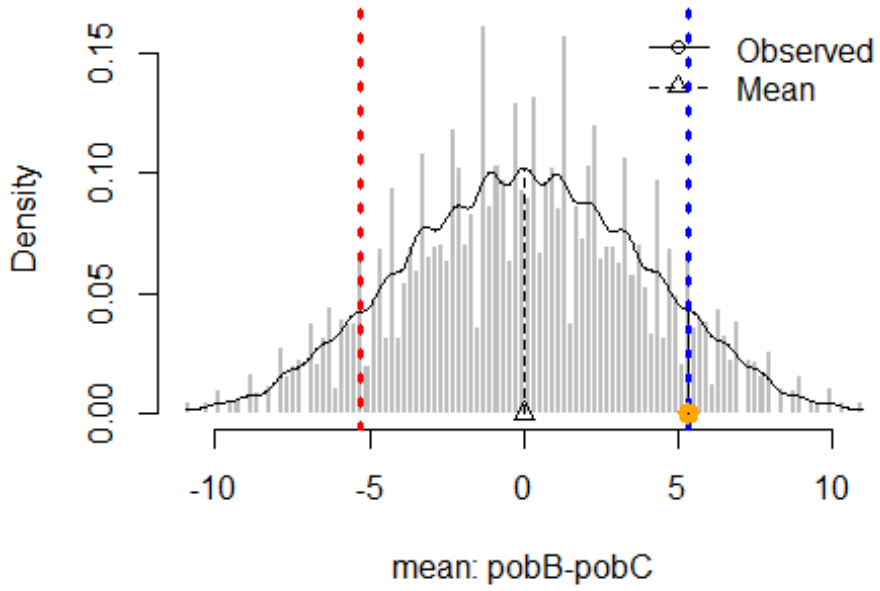
Media A con D

```
TEST3<-permutationTest2(data = pobA,mean, data2 = pobD, R=permut
aciones, seed =35200)
TEST3

## Call:
## permutationTest2(data = pobA, statistic = mean, data2 = pobD,
##     R = permutaciones, seed = 35200)
## Replications: 134596
## Two samples, sample sizes are 6 6
##
## Summary Statistics for the difference between samples 1 and 2
:
##                  Observed       Mean Alternative     PValue
## mean: pobA-pobD 13.87433 0.02027019   two.sided 0.03731138

plot(TEST3, col="grey")
abline(v=AD,lty=3,col="blue",lwd=3)
abline(v=-AD,lty=3,col="red",lwd=3)
points(AD,0,lwd=5, col="orange",pch=19)
```



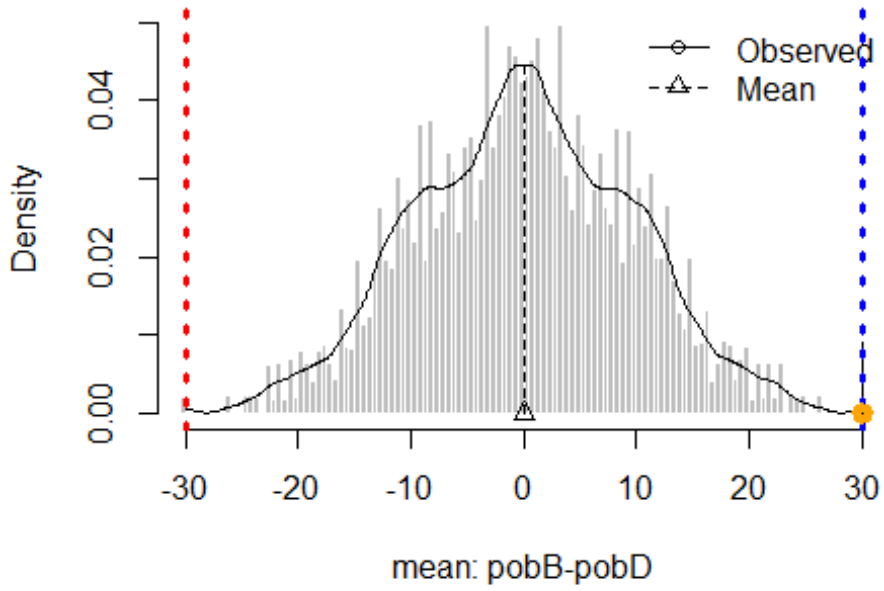
Media B con C

```
TEST4<-permutationTest2(data = pobB,mean, data2 = pobC, R=permut
aciones, seed =35200)
TEST4

## Call:
## permutationTest2(data = pobB, statistic = mean, data2 = pobC,
##     R = permutaciones, seed = 35200)
## Replications: 134596
## Two samples, sample sizes are 6 6
##
## Summary Statistics for the difference between samples 1 and 2
:
##                  Observed        Mean Alternative    PValue
## mean: pobB-pobC 5.315153 0.008766352   two.sided 0.1904054

plot(TEST4, col="grey")
abline(v=BC,lty=3,col="blue",lwd=3)
abline(v=-BC,lty=3,col="red",lwd=3)
points(BC,0,lwd=5, col="orange")
```



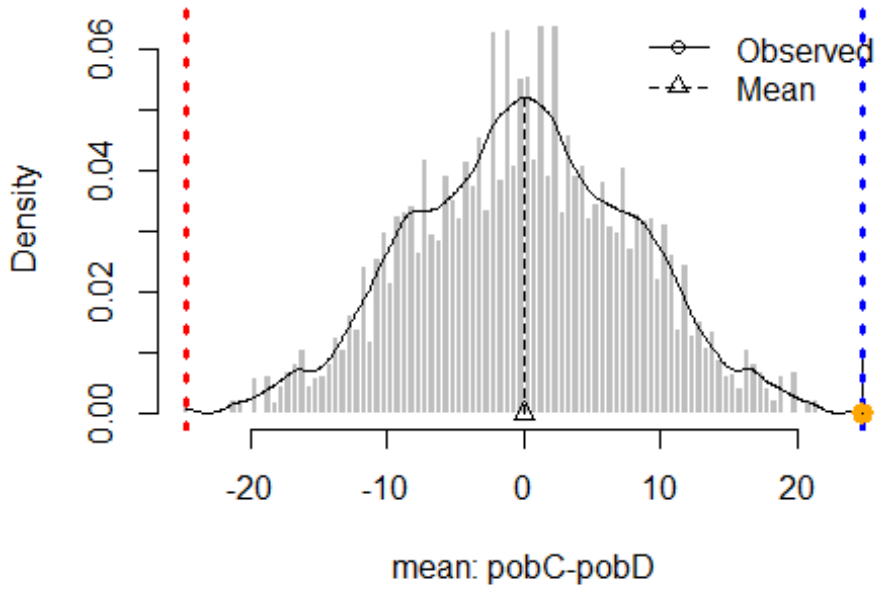
Media B con D

```
TEST5<-permutationTest2(data = pobB,mean, data2 = pobD, R=permut
aciones, seed =35200)
TEST5

## Call:
## permutationTest2(data = pobB, statistic = mean, data2 = pobD,
##     R = permutaciones, seed = 35200)
## Replications: 134596
## Two samples, sample sizes are 6 6
##
## Summary Statistics for the difference between samples 1 and 2
:
##                  Observed       Mean Alternative      PValue
## mean: pobB-pobD 30.04336 0.04206632   two.sided 0.002005988

plot(TEST5, col="grey")
abline(v=BD,lty=3,col="blue",lwd=3)
abline(v=-BD,lty=3,col="red",lwd=3)
points(BD,0,lwd=5, col="orange")
```



Media C con D

```
TEST6<-permutationTest2(data = pobC,mean, data2 = pobD, R=permut
aciones, seed =35200)
TEST6

## Call:
## permutationTest2(data = pobC, statistic = mean, data2 = pobD,
##     R = permutaciones, seed = 35200)
## Replications: 134596
## Two samples, sample sizes are 6 6
##
## Summary Statistics for the difference between samples 1 and 2
:
##                  Observed      Mean Alternative      PValue
## mean: pobC-pobD 24.72821 0.0319768   two.sided 0.002005988

plot(TEST6, col="grey")
abline(v=CD,lty=3,col="blue",lwd=3)
abline(v=-CD,lty=3,col="red",lwd=3)
points(CD,0,lwd=5, col="orange")
```



Apartado b)

A partir de los test de permutación realizados sobre cada una de las poblaciones contrapuestas a sus homólogas, hemos verificado la diferencia existente entre cada una de las posibles combinaciones existentes entre las 4 muestras

El Valor de probabilidad (p-value), es calculado como la proporción correspondiente de las permutaciones de la muestra en las que el valor de la diferencia calculada en le procedimiento de la prueba (diff) es mayor que el de la diferencia observada entre ambas muestras

A partir de las pruebas de permutación realizadas, se puede verificar que las distribuciones que más se aproximan a una normal, y por tanto difieren en menor medida son las combinaciones BC y AC, mientras que aquellas en las que existen mayores diferencias son las combinaciones CD y BD.

Esto se confirma por el valor de probabilidad en las pruebas de permutación, cuyo VP permite determinar, en pocas palabras si un resultado o efecto es estadísticamente significativo cuando es improbable que haya sido debido al azar. Una "diferencia estadísticamente significativa" solamente significa que hay evidencias estadísticas de que hay una diferencia.

Debido a que el VP en la combinación AD es el mayor de todos, ello implica una menor posibilidad de diferencia, por lo cual hemos de observar la diferencia en medias.

Esta sigue siendo la menor de todas en la misma combinación (AD), luego ¿podemos determinar que cualquiera de las dos grasas es la mejor?

Una respuesta convencional sería que sí, pero realmente habría que realizar test de variabilidad y contrastarlos con las medias muestrales para ver si las varianzas se corresponden con el comportamiento de las medias muestrales y no se tratas de muestras que ofrecen resultados irregulares (varianza alta) enmascarados sobre medias aparentemente similares.

Cuestión 9.4

Cuestión 4: La puntuación de 10 estudiantes en dos pruebas psicológicas se detallan en la tabla siguiente. Calcular el coeficiente de correlación de Pearson y el coeficiente de Spearman para los rangos. Explicar los cálculos y contrastar con las funciones de ***R*** los resultados. ¿Qué conclusiones pueden extraerse de los resultados de ambos coeficientes?

Estudiante	A	B	C	D	E	F	G	H	I	J
Test 1	92	89	86	83	77	71	62	2.6	53	40
Test 2	88	85	93	79	70	87	52	84	41	64

```
setwd(".")
library(knitr)
d1 <- c(92,89,86,83,77,71,62,2.6,53,40)
d2 <- c(88,85,93,79,70,87,52,84,41,64)
n <- length(d1)
```

Coeficiente de Correlación de Pearson

Manualmente, Aplicamos formula rp = (n*(sum) x*y - (sum)x * (sum)y)/sqrt((n*(sum)x*$^{2-((suma)x)2)}$(n*(suma)y$^{2-((suma)y)2}$))

```
suma_xy <- sum(d1*d2)
suma_x <- sum(d1)
suma_y <- sum(d2)

rp <- (n*suma_xy-suma_x*suma_y)/sqrt((n*49899.76-429811.36)*(n*57865-552049))
rp
```

```
## [1] 0.2907495
```

Usando la función cor

```
cor(d1,d2,method = c("pearson")) # Dependencia positiva/proporcional
```

```
## [1] 0.2907495
```

Coeficiente de Spearman

Manualmente, ordenando por rangos de menor a mayor valor en cada conjunto

```
di_cuadrado <- (10-9)^2+(9-7)^2+(8-10)^2+(7-5)^2+(6-4)^2+
               (5-8)^2 + (4-2)^2 + (1-6)^2 + (3-1)^2+ (2-3)^2
r <- 1-((6*di_cuadrado/(n*(n^2-1))))
r
```

```
## [1] 0.6363636
```

Usando la función cor

```
cor(d1,d2,method = c("spearman"))
```

```
## [1] 0.6363636
```

Conclusiones

En estadística:

- El **coeficiente de correlación de Spearman (CCSp)**, ρ (rho) es una medida de la correlación (la asociación o interdependencia) entre dos variables aleatorias (tanto continuas como discretas).
- El **coeficiente de correlación de Pearson (CCPe)** es una medida lineal entre dos variables aleatorias cuantitativas. A diferencia de la covarianza, la correlación de Pearson es independiente de la escala de medida de las variables.

En base a esto, se puede concluir que, según el CCPe, existe una correlación positiva entre ambas muestras, situado en la dentro del intervalo de aproximación [0,1].

Esto es refutado por la prueba del CCSp, que da como resultado 0.63 sobre el intervalo de [-1,1], luego como conclusión final determinamos que ambas muestras están relacionadas por algún o algunos factores

Cuestión 10.1

Cuestión 1: Se determinó la mortalidad, en grupos de diez, de ratones que mueren con dosis de un determinado tipo de droga según se refleja en la siguiente tabla:

Dosificación	50	56	62	70	80
Número de Muertes	0	4	5	6	9

a) Realizar un análisis de regresión simple entre ambas variables.
b) Calcular la suma de cuadrados del error y realizar una prueba para la falta de ajuste. Evaluar y analizar gráficamente las relaciones y los errores residuales correspondientes.
c) Encontrar los intervalos de confianza para los coeficientes de regresión.
d) ¿Es posible realizar predicciones con este modelo lineal?, en caso afirmativo estimar la dosis letal mínima (DLM), esto es, la dosis que matará a la mitad de los ratones.

Explicar las respuestas. Utilizar ***R*** en los cálculos donde sea necesario.

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

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

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

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

```
library(MASS)
library(tseries)
```

```
## Warning: package 'tseries' was built under R version 3.6.2
```

```
## 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)

```
y<-c(50,56,62,70,80)
x<-c(0,4,5,6,9)
n<- length(x)
plot(x,y,ylim = c(0,max(y)+2), pch = 18, col ="Blue", xlab ="Mue
rtes",  ylab = "Dosis de sustancia", cex = 1.5)
grid()
```

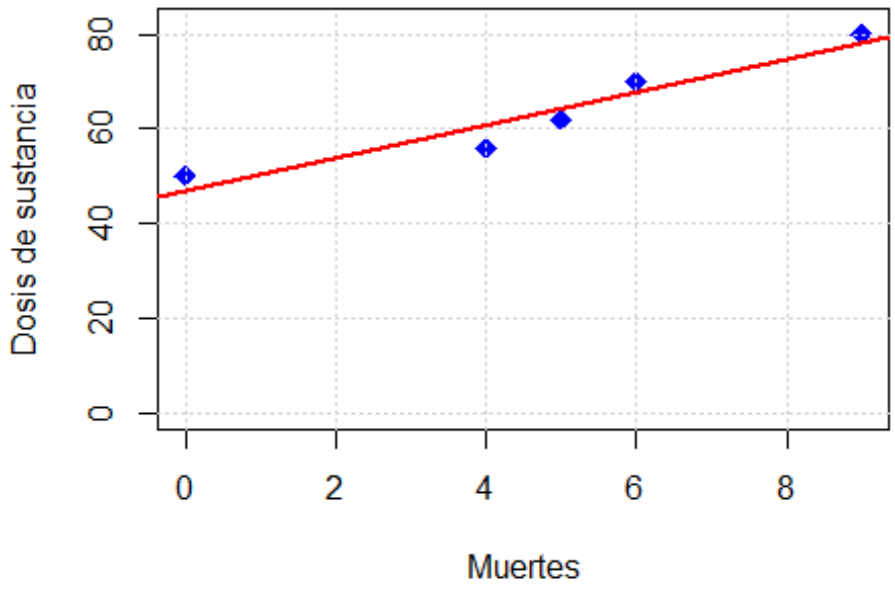
```r
#Medias
x_m<-mean(x)
y_m<-mean(y)
x2<-(x*x)
x_var<-sum(x2)/length(x)-(x_m^2)
y2<-(y*y)
y_var<-sum(y2)/length(y)-(y_m^2)
covar<-sum(x*y)/length(y)-(x_m * y_m)
pend<-covar/x_var
ord<-y_m-(pend*x_m)

# Creamos el modelo lineal resultante
modelo<-lm(y~x)
summary(modelo)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      1      2      3      4      5
##  2.953 -4.841 -2.290  2.262  1.916
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  47.0467     3.3715   13.95 0.000797 ***
## x             3.4486     0.5998    5.75 0.010450 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1
##
## Residual standard error: 3.924 on 3 degrees of freedom
## Multiple R-squared:  0.9168, Adjusted R-squared:  0.8891
## F-statistic: 33.06 on 1 and 3 DF,  p-value: 0.01045
```

```r
# Asignamos los coeficientes y verificamos el intervalo de confi
anza para los coeficientes
coef<-coefficients(modelo)

confint(modelo)
```

```
##                 2.5 %   97.5 %
## (Intercept) 36.317177 57.77628
## x            1.539896  5.35730
```

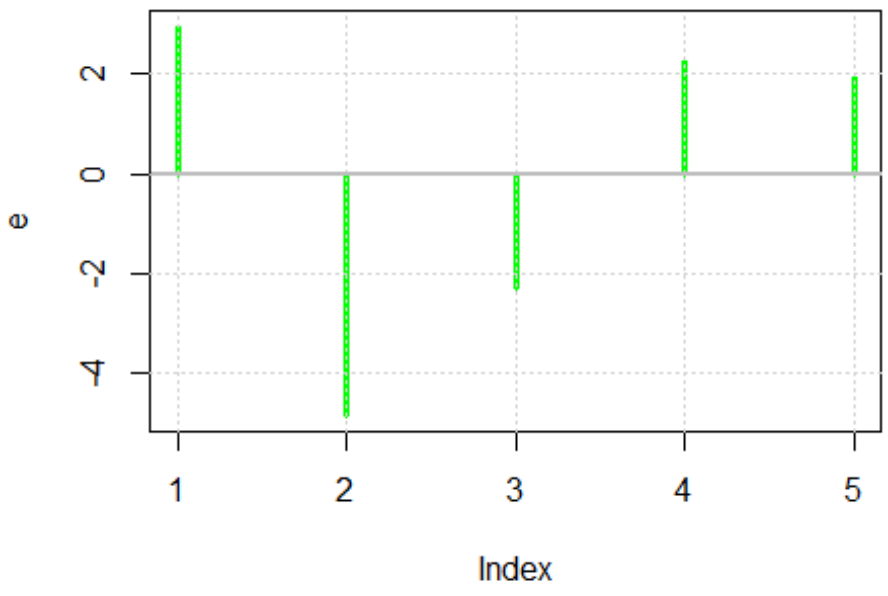
```r
abline(modelo, col = "red", lwd = 2)
```



Apartado b)

```
SCE<-sum((y-(ord + pend*x))^2)
SCE

## [1] 46.18692

summary.aov(modelo)

##             Df Sum Sq Mean Sq F value Pr(>F)
## x            1  509.0   509.0   33.06 0.0104 *
## Residuals    3   46.2    15.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1

x_factor <- as.factor(x)
x_factor

## [1] 0 4 5 6 9
## Levels: 0 4 5 6 9

datos<-data.frame(x_factor, y)
datos

##   x_factor  y
## 1        0 50
## 2        4 56
## 3        5 62
```

```
## 4         6 70
## 5         9 80

y_m_f <- rep(0,n)
for(i in 1:n) {y_m_f[i] <- mean(y[x_factor == x[i]])}
SCEpuro <- sum((y-y_m_f)^2)
SCEpuro

## [1] 0

SCE

## [1] 46.18692

SC_ajuste <- SCE-(SCEpuro)
SC_ajuste

## [1] 46.18692

##Como solo existe una observación par cada x de la muestra, el
SCEpuro será cero debido a la nulidad de la resta de diferencias
.

qf(0.95, 1, n-2)

## [1] 10.12796

e<-residuals(modelo)
plot(e, type= "h", lwd = 3, col = "green")
grid()
abline(h= 0, col = "grey", lwd = 2)
```

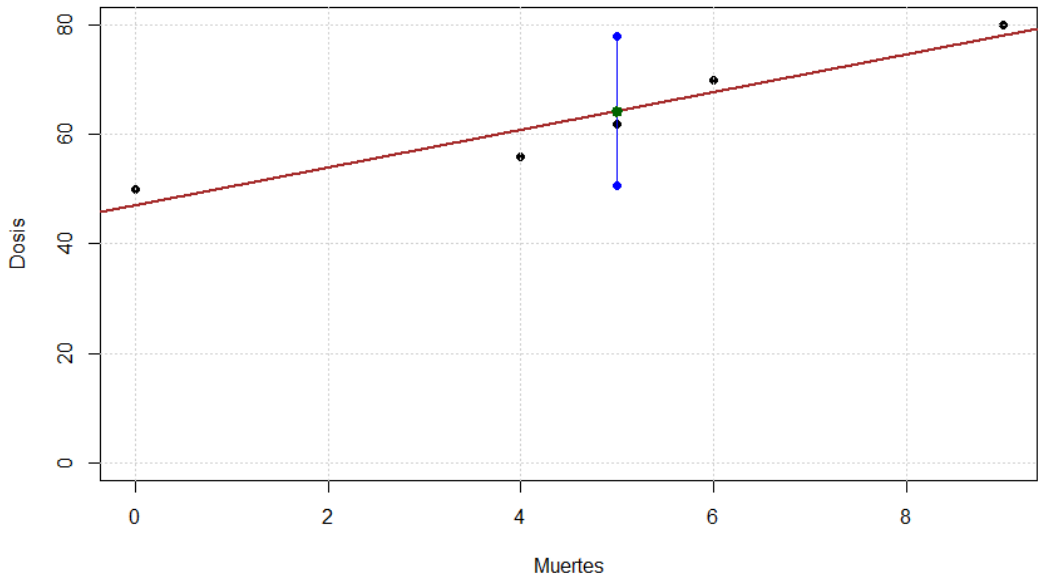


```
#Test Shappiro para verificar la normalidad
shapiro.test(e)

##
##  Shapiro-Wilk normality test
##
## data:  e
## W = 0.85361, p-value = 0.2062

#Test Kologorov Smirnov
ks.test(e,"pnorm")

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  e
## D = 0.57231, p-value = 0.04424
## alternative hypothesis: two-sided
```

Apartado c) y d)

```
confint(modelo)

##                 2.5 %   97.5 %
## (Intercept) 36.317177 57.77628
## x            1.539896  5.35730

coefficients(modelo)

## (Intercept)           x
##   47.046729    3.448598

coef<-coefficients(modelo)
predicion_DLM<- predict(modelo, newdata= data.frame(x=5), interv
al = "pred")
predicion_DLM

##        fit      lwr      upr
## 1 64.28972 50.60551 77.97393

lwr<-predicion_DLM[2]
upr<-predicion_DLM[3]
puntomedio<-5*coef[2]+coef[1]
puntomedio

##        x
## 64.28972

plot(x, y, ylim = c(0,max(y)), col = "black", pch=19,
     xlab="Dosis",
     ylab="Muertes")
grid()
abline(modelo, col = "brown", lwd = 2)
```

```
lines(c(5,5), c(lwr, upr), col="blue")
points(c(5,5), c(lwr, upr), col="blue",pch=19)
points(5, puntomedio, col="darkgreen",pch=19,lwd=3)
```



Conclusiones

La muestra analizada muestra normalidad, como bien indican las pruebas de Shapiro-Wilk y Kolmogorov. Asimismo, el error de ajuste de la propia recta es poco en el valor de R2, aproximándose a la unida, por lo que se puede decir que es fiable, aunque la pendiente sea poca.

El DLM estimado, aplicando el modelo lineal obtenido en el valor intermedio de muertes es de *62,3*. Luego se podría asumir que a este valor, balo un intervalo de confianza, es el idóneo para provocar la muerte al 50% de los individuos.

Mediante el uso de la función *predict* se ven claramente los dos puntos extremos del intervalo de confianza. Debido a la ambigüedad de ajuste, se observa un rango de confianza amplio, de cerca de 30 puntos en la escala (sobre 100)

Cuestión 10.2

Cuestión 2: Se realizó un estudio sobre la cantidad de azúcar convertida en cierto proceso bioquímico a distintas temperaturas. Se toma la base de temperaturas en 25º C y las cantidades de azúcar en miligramos. Los datos se codificaron y registraron como se indica en la siguiente tabla:

Num. Ensayo	Temperatura codificada *x* (base 25 ºC)	Azúcar convertida *y* (mg)
1	1.0	8.1
2	1.1	7.8
3	1.2	8.5
4	1.3	9.8
5	1.4	9.5
6	1.5	8.9
7	1.6	8.6
8	1.7	10.2
9	1.8	9.3
10	1.9	9.2
11	2.0	10.5

a) Realizar un análisis de regresión lineal simple de ***y*** con ***x***.
b) Calcular la suma de cuadrados del error y realizar una prueba para la falta de ajuste. Evaluar gráficamente las relaciones y los errores residuales correspondientes.
c) Encontrar los intervalos de confianza para los coeficientes de regresión.
d) ¿Es posible realizar predicciones con este modelo lineal? En caso afirmativo determinar la cantidad media de azúcar convertida que se produce cuando se registra una temperatura codificada de 1.75 y el intervalo de confianza de la predicción correspondiente.
e) Definir el concepto de respuesta media y encontrar los intervalos de confianza para la misma en el apartado anterior.
f) Visualizar los resultados de los apartados a), c), d) y e) utilizando las funciones gráficas básicas de ***R*** y las de la librería *ggplot2*

Explicar las todas respuestas. Utilizar ***R*** en los cálculos donde sea necesario.

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

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

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

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

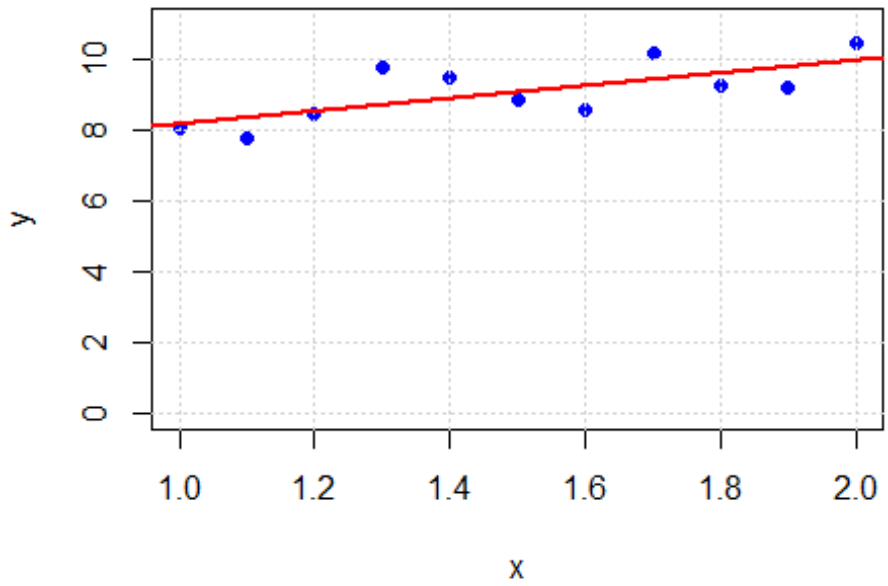
```
library(MASS)
```

Apartado a)

```
## Generamos la tabla de elementos
datos<-read.csv("Archivos/azucar.csv", sep =";", dec=",")
names(datos)<-c("Temperatura", "Conversion")
kable(datos)
```

Temperatura	Conversion
1.0	8.1
1.1	7.8
1.2	8.5
1.3	9.8
1.4	9.5
1.5	8.9
1.6	8.6
1.7	10.2
1.8	9.3
1.9	9.2
2.0	10.5

```
## Generamos el modelo de regresión lineal a mano
# Primero el plot
attach(datos)
x<-Temperatura
y<-Conversion
n<-length(x)
plot(y~x, data = datos, ylim=c(0,length(y)), col="blue", pch=19)
grid()

# Segundo el modelo lineal (el abline)
linealbasico<-lm(y~x)
abline(linealbasico, col="red", lwd=2)
```



Apartado b)

```
## Calculamos la SCE y realizamos la prueba de ajuste. Obtenemos
también los residuales
residuales<-residuals(linealbasico)
qres<-quantile(residuales)
SCE<-sum(residuales^2)
s2<-SCE/(n-2)
var<-sqrt(SCE/(n-2))
STCC<-sum((y-mean(y))^2)
SCR<-STCC-SCE

## Calculamos el STCC y luego determinamos el ajuste mediante el
CdD.
CdD <- 1-(SCE/STCC)
CdD

## [1] 0.4998864
```

Apartado c)

```
## Verificamos confint los intervalos.
confint(linealbasico)

##                 2.5 %   97.5 %
## (Intercept) 4.3219598 8.505313
## x           0.4446316 3.173550

sup<-qt(0.975, n-2)
sup
```

```
## [1] 2.262157

inf<-1-qt(0.975, n-2)
inf

## [1] -1.262157
```

La prueba de *confint* verifica un intervalo de intercepción de entre 4,32 y 8,50 en el eje de las y, valor bastante amplio.

Apartado d)

Verificamos con summary la prueba de ajuste general calculada manualmente.

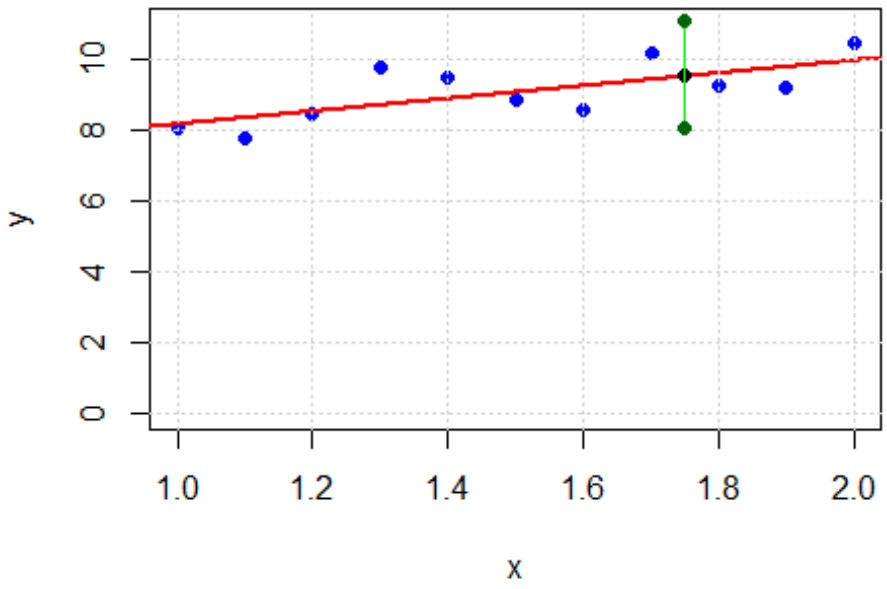
```
summary(linealbasico)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7082 -0.4868 -0.1227  0.5109  1.0346
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.4136     0.9246   6.936 6.79e-05 ***
## x             1.8091     0.6032   2.999    0.015 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1
##
## Residual standard error: 0.6326 on 9 degrees of freedom
## Multiple R-squared:  0.4999, Adjusted R-squared:  0.4443
## F-statistic: 8.996 on 1 and 9 DF,  p-value: 0.01497
```

Los cuartiles coinciden al 100% con los residuales, el error también coincide con la varianza calculada bajo 9 grados de libertad y el MRS determina un cerca del 50% respecto a la unidad, equivalente al coeficiente de determinación calculado manualmente. Tiene sentido hasta cierto punto, puesto que la pendiente es pequeña luego la dependencia es objeto de contraste El error de ajuste es del 1.47%, luego es posible realizar una prueba muy fiable del modelo.

Realizamos la predicción

```
x0<-1.75
prediccion<-predict(linealbasico, newdata=data.frame(x=x0))
prediccion

##        1
## 9.579545
```

```
plot(y~x, data = datos, ylim=c(0,length(y)), col="blue", pch=19)
abline(linealbasico, col="red", lwd=2)
grid()
points(x0,prediccion, pch=16,col="black")

inter_prediccion<-predict(linealbasico,level=0.95, newdata = dat
a.frame(x=x0), interval = "pred")
inter_prediccion

##        fit      lwr      upr
## 1 9.579545 8.046425 11.11267

lwr<-inter_prediccion[2]
upr<-inter_prediccion[3]
lines(c(x0,x0), c(lwr,upr),col="green")
points(x0,lwr, pch=16,col="darkgreen")
points(x0,upr, pch=16,col="darkgreen")
```



Mediante el uso de la función predict se ven claramente los dos puntos extremos del intervalo de confianza. Debido a la ambigüedad de ajuste, se observa un rango de confianza amplio, de cerca de 3.5 puntos en la escala

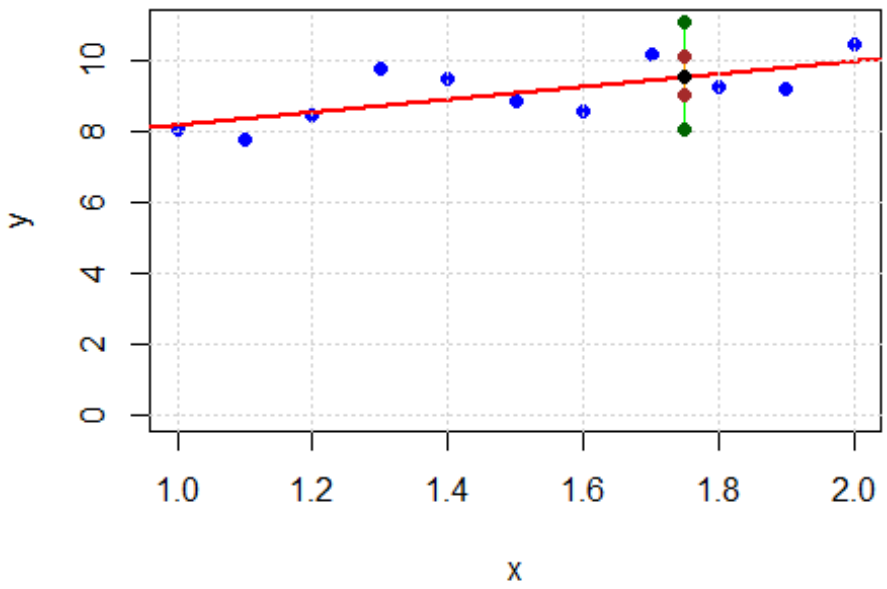
Apartado e)

Defición teórica

Realización del apartado (sacado de apuntes)

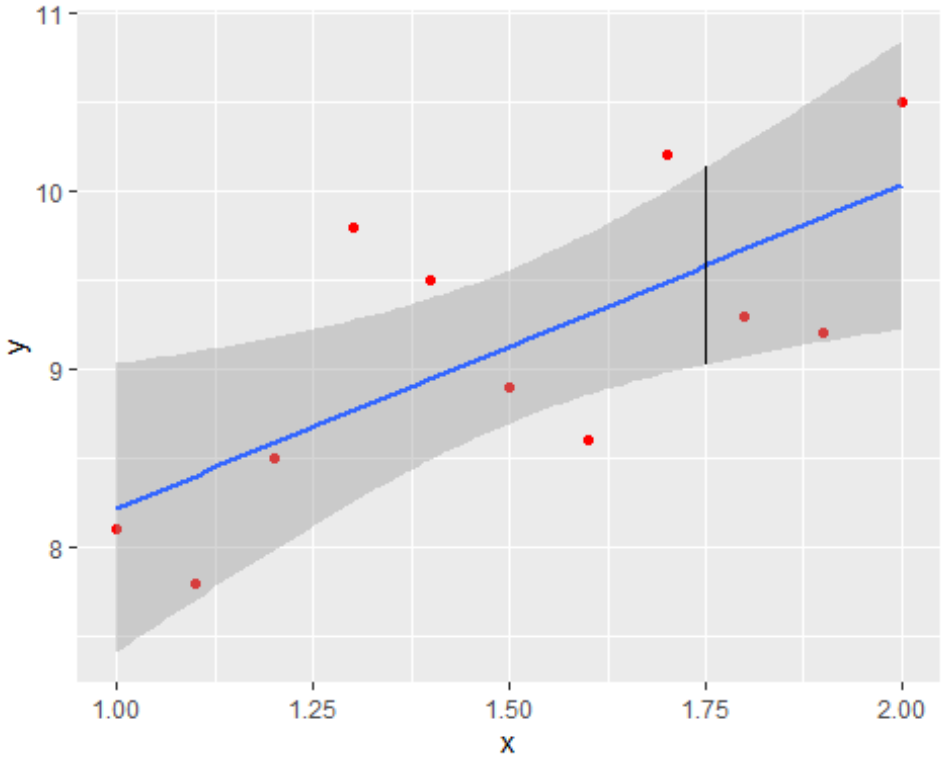
```
plot(y~x, data = datos, ylim=c(0,length(y)), col="blue", pch=19)
abline(linealbasico, col="red", lwd=2)
```

```
grid()
lines(c(x0,x0), c(lwr,upr),col="green")
points(x0,lwr, pch=16,col="darkgreen")
points(x0,upr, pch=16,col="darkgreen")

confi_prediccion2<-predict(linealbasico,level=0.95, newdata = da
ta.frame(x=x0), interval = "confidence")
confi_prediccion2

##        fit      lwr      upr
## 1 9.579545 9.029514 10.12958

lwr2<-confi_prediccion2[2]
upr2<-confi_prediccion2[3]
lines(c(x0,x0), c(lwr2,upr2),col="orange")
points(x0,lwr2, pch=16,col="brown")
points(x0,upr2, pch=16,col="brown")
points(x0,prediccion, pch=16,col="black")
```



Apartado f)

```
datos<- data.frame(x,y)
g<-ggplot(data=datos, aes(x=x , y=y))
g+geom_point(colour= "red")+ geom_smooth(method= "lm")+ geom_lin
erange(aes(x=1.75, ymin=lwr2, ymax=upr2))
```


11
10
9
8
y
1.00
1.25
1.50
1.75
2.00
x