TEMA 8 TEO

ALEJANDRO DAVID ARZOLA SAAVEDRA

Cuestión 1: Se toma una muestra aleatoria simple de una población que sigue una distribución N(μ, σ2), 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.

```{r}
# HO Media ≥ 10
# H1 Media < 10
#Muestra
n=10
Ho=10
x=c(3.58,10.03,4.77,9.71,10.4,14.66,8.45,5.4,9.75,10.1)
media=mean(x)
media
desviacion_tipica=sd(x)
desviacion_tipica
α=0.05
y=qt(α,n-1)
y
Estadistico=(media-Ho)/(desviacion_tipica/sqrt(n))
Estadistico
# Desicicion = no rechazar H0 y concluir que el número promedio de una muestra aleatoria simple de una población no es significativamente menor que 10
PT<-seq(5,20,0.001)
sigma_muestra<-desviacion_tipica
mu<-Ho
media_muestra<-media
```

```
tmedia_muestra<-(media_muestra-mu)/(sigma_muestra/sqrt(n-1))

tmu<-(mu-mu)/(sigma_muestra/sqrt(n-1))

TPT<-(PT-mu)/(sigma_muestra/sqrt(n-1))

DP0<-dt(TPT, n-1)

plot(PT,DP0, type= "l", col="brown", ylab= "Densidad de Probabilidad", xlab= "Estadístico t (Muestra aleatoria simple de una población)")

abline(v=mu, col="green")

abline(v=media_muestra, col="blue")

alfa<-α

# Intevalo de decision

Zona_critica1<-qt(alfa,n-1)*(sigma_muestra/sqrt(n-1))+mu

Zona_critica1

#b)

#No Podría afirmarse con los datos que la media de la población es inferior a 10 solo podriamos llegar a afirmar que Ho no es verdadera #c)

#Acetamos Ho , Siendo Ho verdadera = No error

#Aceptamos Ho , Siendo Ho falsa = Error tipo II (β)

#Rechazamos Ho , Siendo Ho verdadera = Error tipo I (α)

#Rechazamos Ho, Sieno Ho falsa = No error


# α= P(RECHAZO/SIENDO Ho Verdadero)

# β= P(ACIERTO/SIENDO Ho Falsa)

# Potencia= Probabilidad de rechazar Ho

# P= 1- β

#

# qt = -1.833113

# -1.833= VC-Ho/S/sqrt(n)

VC=y*(desviacion_tipica/sqrt(n))+n

VC

# Error tipo 1 es 0,05

# Error tipo 2 es
```

β= P(Acetar/Ho falsa)

 # P(Media > 8.098184)

t=(VC-media)/(desviacion_tipica/sqrt(n))

t

β=pt(t,9)

β

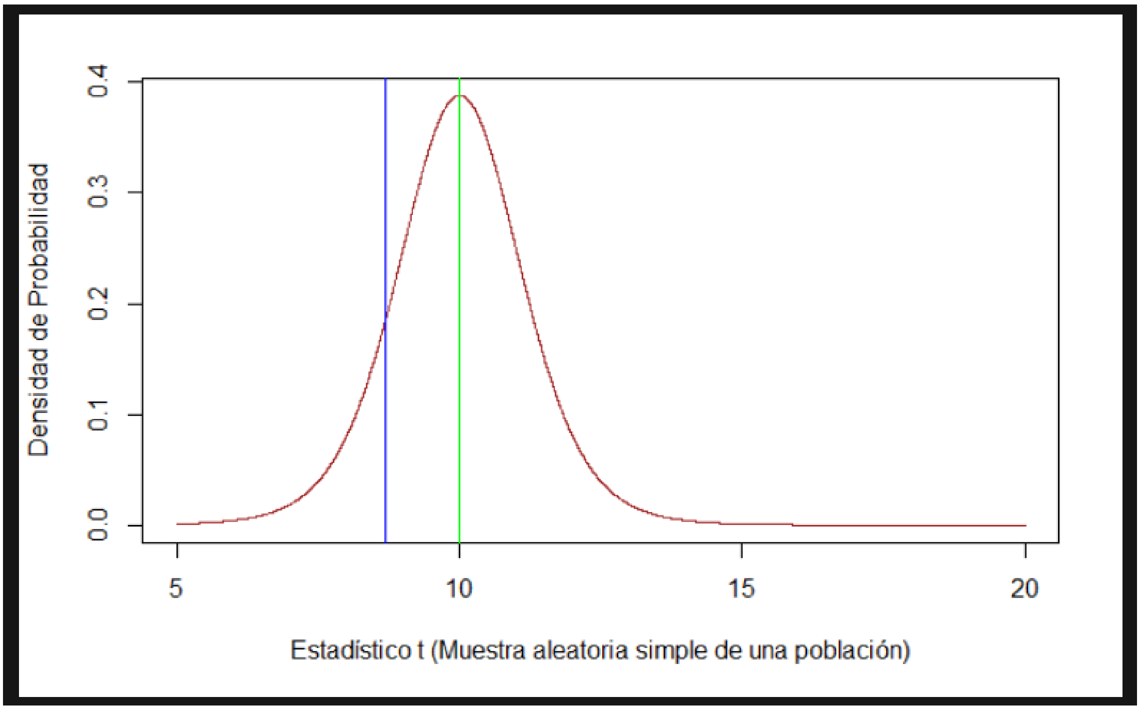
Potencia=1-β

Potencia

power.t.test(delta = 1, sd=desviacion_tipica, sig.level= α, power= Potencia)

#En este caso tomaremos n=136, que nos garantiza alcanzar una potencia del 70%

```


```

```
[1] 8.685
[1] 3.280797
[1] -1.833113
[1] -1.267495
[1] 7.99531
[1] 8.098184
[1] -0.5656175
[1] 0.2927384
[1] 0.7072616

     Two-sample t test power calculation

              n = 136.0903
          delta = 1
             sd = 3.280797
      sig.level = 0.05
          power = 0.7072616
    alternative = two.sided

NOTE: n is number in *each* group
```

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:

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

````
```{r}

# Intervalo de confianza al 95 % para la M1-M2 con V1^2,V2^2 desconcidas


# a) Contraste para la igualdad de varianzas

#Ho: V1^2=v2^2

#H1: V1^2/=V2^2

n1=16

n2=10

v1=n1-1

v1

v2=n2-1

v2

# Se trata de una fisher

qf(0.975,v1,v2)

qf(0.025,v1,v2)

# Estadistico
````
````

```
Varianza1=61500

Varianza2=90201

f=Varianza2/Varianza1

f

# Haceptamos Ho =) podemos considerar varianza1=varianza2

# Por tanto:

# Para I.C para m1-m2 con varianzaas desconocidas pero iguales

# Sp^2=(n1-1)*S1^2+(n2-1)*S2^2/n1+n2-2

# mu1-mu2(= I[(x1-x2)+-tα/2*Sp*sqrt(1-n1+1/n2)]

v=n1+n2-2 # Grados de libertad

α=0.05

α_medios=α/2

y=qt(α_medios,v)

y

#c)

#Región critica RC: t>-2.063899

sp=((Varianza1*(n1-1))+(Varianza2*(n2-2)))/n1+n2-2

sp2=sqrt(sp)

sp2

media1=1515.60

media2=1298.35

t=((media1-media2)-v)/(sp2*sqrt(1/n1+1/n2))

t

# Decisión: no rechazar H0 no podemos concluir la existencia de discriminación salarial entre hombres y mujeres en el ayuntamiento

#b)

# tv,0.05 < tv,0.025

# por tanto va a disminuir la amplitud del invervalo=) mas riesgo de error en la estimacion

#d)

# Evidentemente si aumenta las n1 y n2 sera mas fiable la estimacion, y seguirian siendo iguales no podemos concluir la existencia de discriminación salarial entre hombres y mujeres en el ayuntamiento
```

```
# Por otro lado

#         (  Error ) (   Error  )

#---------/----------/------------/-----------

#       x min     x1-x2       x max


# (x1-x2)+-Error

# al aumentar n1 y n2 (Error disminuye)

print("Error original")

E1=y*sp2*sqrt(1/n1+1/n2)

E1


nn1=45

nn2=30

sp=((Varianza1*(nn1-1))+(Varianza2*(nn2-2)))/nn1+nn2-2

sp3=sqrt(sp)

sp3

media1=1515.60

media2=1298.35

t=((media1-media2)-v)/(sp3*sqrt(1/n1+1/n2))

t

# E= Tv,α/2*Sp*sqrt(1/n1+1/n2)

print("Error aumentando la n1 y n2")

E2=y*sp3*sqrt(1/nn1+1/nn2)

E2
```

```
[1] 15
[1] 9
[1] 3.769357
[1] 0.3202345
[1] 1.466683
[1] -2.063899
[1] 320.5694
[1] 1.495446
[1] "Error original"
[1] -266.7087
[1] 341.0079
[1] 1.405816
[1] "Error aumentando la n1 y n2"
[1] -165.8886
```

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 €)

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.

```{r}

# Intervalo de confianza para la media con varianza desconocida

Ph<-
c(1500.21,1210.12,2060.01,1500.08,890.50,1800.30,2015.22,3200.00,1005.40,880.66,2010.1,810.10,2500.00,515.01,625.12,720.25,1601.79,2150.1,605.22,701.30,1012.34,917.45,820.39,1002.20,1102.45,1219.70,623.56)

mu<-mean(Ph)

mu

sigma<-sd(Ph)

sigma

n<-length(Ph)

n

Ic1<-mu-qt(0.975, df=(n-1))*sigma/sqrt(n)

Ic2<-mu+qt(0.975, df=(n-1))*sigma/sqrt(n)

cat(c("lím.Inf. IC 95%=", as.character(round(Ic1,3)), "  lím.Sup. IC 95% =", as.character(round(Ic2,3))))

#Una posible conclusión es quelas rentas mensuales de los habitantes del barrio de Vegueta es superior a 1030, pues 1030 no está en el intervalo de confianza

AB<-seq(1000,2000,0.001)

DPAB<-dt(AB,n-1, mu)

plot(AB,DPAB, type = "l", col="brown", xlab="Media", ylab="DF")

xliminf<-Ic1

xlimsup<-Ic2

xv<-AB[AB>=xliminf & AB<=xlimsup]

yv<-DPAB[AB>=xliminf & AB<=xlimsup]
```

```
xv<-c(xv,xlimsup,xliminf)

yv<-c(yv,0,0)

polygon(xv,yv,col = "green", density = 20)

grid()

t.test(Ph,conf.level = 0.95)

#b)

alfa=0.05

Ho=1000

# Ho: desviacion ≤ 1000

# H1: desviacion >1000

# var1=1.000.000

# var2=1.000.000

t=(1000-mu)/(1000000/sqrt(27))

t

n1=n-1

qchisq(0.95,n1)

s=sigma*sigma

var=Ho*Ho

chicuadrado=((n1)*s)/1000

chicuadrado

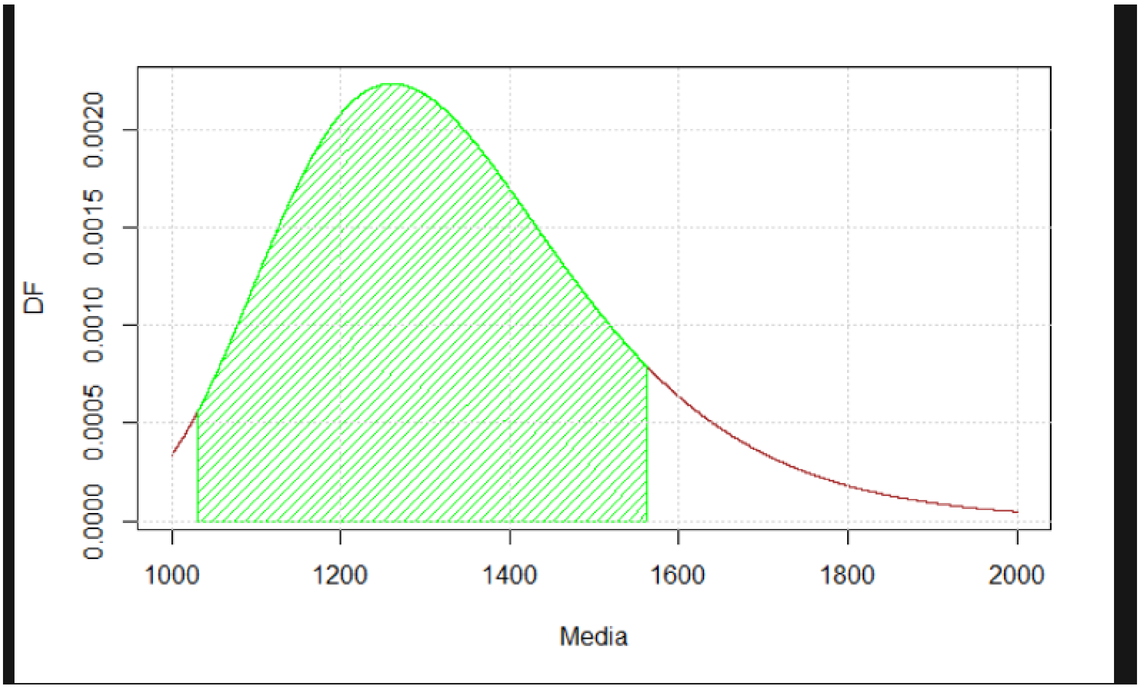
#Región crítica RC (hipótesis alternativa) X^2>X^alfa2


#Rechazamos la hipótesis nula H0 cuando χ^2>238.8851498

pchisq(11,n1)


# Hay evidencia para 1000 euros, no rechazamos la hipótesis nula H0 por el bajo valor del estadistico X^2

#Una dispersion alta sería un indicativo de un reparto desigual de la renta en este barrio.
```




```
[1] 1296.281
[1] 672.6922
[1] 27
lím.Inf. IC 95%= 1030.173   lím.Sup. IC 95% = 1562.389
        One Sample t-test

data:  Ph
t = 10.013, df = 26, p-value = 2.063e-10
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 1030.173 1562.389
sample estimates:
mean of x
 1296.281

[1] -0.00153952
[1] 38.88514
[1] 11765.38
[1] 0.004450883
```

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:

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.

```{r}
f=c(6.2,6.6,5.8,5.4,6.15,6.68,7.0,5.8,5.6,5.85,6.2,6.4,6.75,5.3,6.3)

#a)
# Contraste para la media con varianza desconocida

# Ho: M = 5.35
# H1: M > 5.35
n=length(f)
n
x=qt(0.99,n-1)
x
media=mean(f)
Ho=5.35
desviacion_estandar=sd(f)
#Estadistico
Estadistico=(media-Ho)/(desviacion_estandar/sqrt(21))
Estadistico
# Desición: No aceptamos Ho podemos asumir que su consumo medio es mayor.

#b)Se trata de un intervalo par la media con varianza desconocida
#[media+- tn-1,α/2*S/sqrt(n)]
# tn-1,α/2*S/sqrt(n)=ERROR
#n=tn-1,α/2
error=6
n=(x*desviacion_estandar-error)^2
```

n

#Deberia observaso 21 muestras

#Por otro lado no es posible hacer un recorrido anual de 25000 kilometros dado que es mayor a 250

#Debido a que se consumen mas de 6.0 litros por cada 100 kilometros

ideal=6*250

ideal

kilometros=Estadistico*250

kilometros

#c)

y=qt(0.90,n-1)

y

H1=6

Estadistico2=(media-H1)/(desviacion_estandar/sqrt(21))

Estadistico2

#kilometros2=Estadistico2*250

#kilometros2

m=(y*desviacion_estandar-error)^2

m

En este caso si seria posible hacer un recorrido anual de 25000 kilometros

```

~~~
[1] 15
[1] 2.624494
[1] 7.121626
[1] 21.8438
[1] 1500
[1] 1780.406
[1] 1.32351
[1] 1.227241
[1] 28.42144
~~~

Chunk 1
```

TEMA 9 TEO

ALEJANDRO DAVID ARZOLA SAAVEDRA

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.

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.

```
```{r}

# Primero vamos a realizar la tabla con los datos

setwd("C:/Users/34636/Desktop/Alejandro_Arzola_ME_3_entrega/DATOS")

f=read.table("ejercicio1_tema9C.csv",header=T,sep=";")

f


N=168 # La suma de las filas y las columnas

nij1=66*82/168

nij1

nij2=66*86/168

nij2

nij3=102*82/168

nij3

nij4=102*86/168

nij4

# Aplicaremos las prueba de homogeneidad

chicuadrado=((35-32.21)^2/32.21)+((31-33.78)^2/33.78)+((47-49.78)^2/49.78)+((55-52.21)^2/52.21)

chicuadrado

# Contrastemos:
```
```

#Ho:Las variables son indepentientes

#H1:Ho no es cierta

α=0.05

m=2

n=2

v=(m-1)*(n-1)

qchisq(0.95,v)

Aceptamos Ho

```

```
[1] 32.21429
[1] 33.78571
[1] 49.78571
[1] 52.21429
[1] 0.7747967
[1] 3.841459
```

| Xi.Yi <fctr> | Hombres <int> | Mujeres <int> | ni <int> |
|---|---|---|---|
| SI (Politica) | 35 | 31 | 66 |
| NO (Politica) | 47 | 55 | 102 |
| nj | 82 | 86 | 168 |

**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:**

**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.**

```{r}

#a)

# Se presentan situaciones de dependencia entre los datos de los dos grupos. Los datos de cada muestra pertenecen a los mismos individuos, se trata de datos apareados

#Ho=  Ambas materias se consideran equivalentes

**#H1= Ho no es cierta**

**Algebra<-**
**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)**

**Programacion<-**
**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)**

**wilcox.test(Algebra,Programacion, paired=TRUE, conf.int = TRUE, conf.level= 0.95)**

**#En este caso el valor de p(0.7153) permite afirmar el sostenimiento de la hipótesis nula y que las muestras se refieren a la misma población con un nivel de significancia del 0.05.**

```
        Wilcoxon signed rank test with continuity correction

data:  Algebra and Programacion
V = 219.5, p-value = 0.7153
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -0.4500104  0.6000672
sample estimates:
(pseudo)median
     0.1000575
```

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

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}

#a)

# Vamos a a realizar el test ANOVA para saber si existen diferencias de medias entre las distintas grasas

# H0: μ1 = μ2 =μ3= μ4

# H1 : Al menos dos de las medias no son iguales.

# k=4 grados de libertad

# N=24

# Grados de libertad k -1 = 3, N-k = 20

qf(0.95,3,20)

# En este caso F debe ser mayor que 3.098391 para considerar que es falsa H0 y rechazar la igualdad de medias μ1= μ2 =μ3= μ4

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)

datos<-data.frame(variable=c(grasa_A,grasa_B,grasa_C,grasa_D))

grupo=factor(c(rep(1,length(grasa_A)),rep(2,length(grasa_B)),rep(3,length(grasa_C)),rep(4,length(grasa_D))))

attach(datos)

ANOVA<-aov(variable~grupo,data= datos)

summary(ANOVA)
```

Como 3.098391 Se rechaza H0

Por lo cual podemos aceptar que existen diferencias significativas entre ellas

```
#b)

xdatos<-data.frame(grasa_A,grasa_B,grasa_C,grasa_D)

boxplot(xdatos,

col = "green",

ylab="Cantidad de grasa absorbida de las rosquillas",

xlab="Tipo de grasa de las rosquillas",

staplewex=1,

border= "brown")

grid()

# La que

summary(grasa_A)

summary(grasa_B)

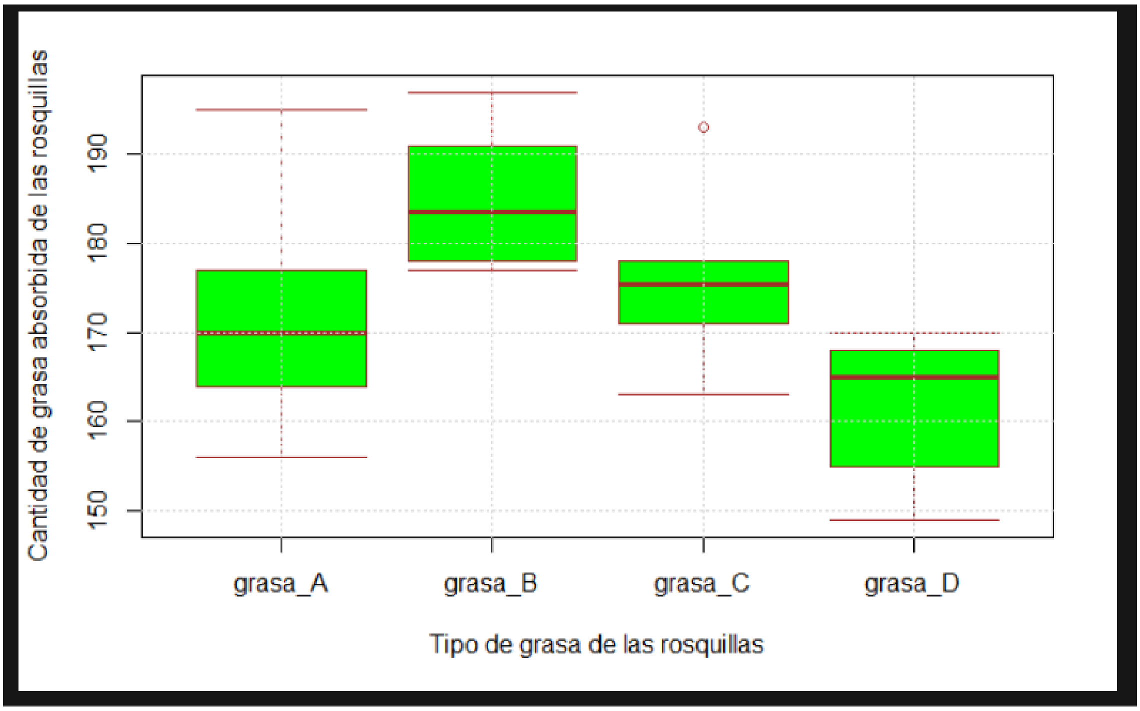
summary(grasa_C)

summary(grasa_D)

# La grasa menos perjudicial es la la grasa D por que es la que menos grasa absorve de media
```

```
[1] 3.098391
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
grupo        3   1636   545.5   5.406 0.00688 **
Residuals   20   2018   100.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  156.0   165.0   170.0   172.0   175.8   195.0
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  177.0   179.0   183.5   185.0   189.5   197.0
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  163.0   172.0   175.5   176.0   177.5   193.0
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  149.0   157.2   165.0   162.0   167.5   170.0
```



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?

```{r}

test1<-c(92,89,86,83,77,71,62,2.6,53,40)

test2<-c(88,85,93,79,70,87,52,84,41,64)

cor(test1, test2 , method = c( "pearson"))

#Lo que indica una correlación positiva alta (0.6363636) entre las notas de algebra y programacion

# Suma de cuadrados diferencia rangos

d_2<-(test1-test2)^2;

sum(d_2)

n<-length(test1)

r<-1-((6*sum(d_2)/(n*(n^2-1))))

# Coeficiente Spearman

r

# Estadistico

z<-r*sqrt(n-1); z

# RegionCritica

z_critico<-qnorm(0.975); z_critico
```

#

cor(test1,test2 , method= c("spearman"))

cor.test(test1, test2 , method = c("spearman"))

En este caso el bajo valor de p=0.05445 mayor que 0.05 confirma la aceptacion de hipótesis nula de correlación cero.,Existe relacion

El coeficiente de correlación muestral sirve para medir el coeficiente de correlación poblacional ρ, la relación lineal entre dos variables continuas en este caso los dos test

Otra forma de hacerlo

n=length(test1)

ncuadrado=n^2

di=test1-test2

di

dicuadrado=di^2

dicuadrado

sumadicuadrado=sum(dicuadrado)

sumadicuadrado

r5=1-(6/(n*(ncuadrado-1))*sumadicuadrado)

r5

Zo=(r5-0)/(1/(sqrt(n-1)))

Zo

```

```
[1] 0.2907495
[1] 7847.96
[1] -46.56339
[1] -139.6902
[1] 1.959964
[1] 0.6363636

        Spearman's rank correlation rho

data: test1 and test2
S = 60, p-value = 0.05445
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.6363636

 [1]   4.0   4.0   -7.0   4.0   7.0 -16.0 10.0 -81.4 12.0 -24.0
 [1]   16.00   16.00   49.00   16.00   49.00 256.00 100.00 6625.96 144.00 576.00
[1] 7847.96
[1] -46.56339
[1] -139.6902
```
```

TEMA 10 TEO

ALEJANDRO DAVID ARZOLA SAAVEDRA

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

Número de Muertes

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.

```
```{r}

#Al ser tan pequeño puede haber un probable problema en la medida de bondad de ajuste

dosificacion=c(50,56,62,70,80)

Numerodemuertes=c(0,4,5,6,9)

#para n pequeños se hace un pequeño ajuste

# Vamos a encontrar la mejor relación entre x e y

y=dosificacion

x=Numerodemuertes

# Es una regresion simple pues se trata de  un solo regresor

# Y= B0+b1X

#Plotemamos la dosificaicion y las variables

plot(x,y,ylim=c(0,100),pch=19,col="purple",xlab="Numero de muertes de los ratones",
ylab="Dosificacion",cex=1.5)

grid()
```
```

```
#Realizamos el modelo de regresion simple

modelo=lm(y~x)# Los valores de yse distribuyen alrededor de la recta verdadera o recta de regresión de la población y=β0+ β1x


#La verdadera recta de regresión pasa a través de las medias de la respuesta y las observaciones reales se encuentran sobre la distribución, alrededor de las medias.


#--------------------------------------------------------------------------

#Diferencia entre la estima y el valor que vamos obteniendo

coefficients(modelo)

#S=3.9

summary(modelo)

#El summary nos indica la estima de los coeficientes

# Nuestra

b0b1=coefficients(modelo)

#Asimetria

confint(modelo)

abline(modelo,col="red")

 # el B0= esta entre  36.317177 57.77628

 # el B1= esta entre  1.539896  5.35730

e=residuals(modelo)

# Todos los posibles errores siguen una normal


plot(e,type="h",lwd=2,col="red")

abline(h=0)

grid()

#Normalidad aceptable si salemas de 0.05

shapiro.test(e)

# p value suma de cuadrados totales menos la suma de los cuadrados del error

ks.test(e,"pnorm")
```

```
library(tseries)

#Mide la simetria

jarque.bera.test(e)


#Podemos  ver que no es horizontal

#Lo representamos un histograma de los residuos

hist(e,freq =F,col="green",density=25,border="brown")

valores=seq(min(e),max(e),0.1)

points(valores,dnorm(valores,mean=mean(e),sd=sd(e)),type="l",col="blue")

confint(modelo)

#Aunque si que aparecen una seria de asimetrias

coefficients(modelo)

#-----------------------------------------------------------------------------------------

# Si es posible hacer predicciones

# Sigue una normal  y hay una suma de unas variables

#Se trata ahora obtener un intervalo de predicción para cualquier valor único y0 de la variable y0.

#x=5 es la mejor estima

 # En el intervalo de confianza nos quedamos con la muestra de abajo por que matamos a las 5 ratas es la dosis letal minima.

prediccion=predict(modelo,newdata=data.frame(x=5),inverval="pred")

# Predicción de intervalo de confianza para la media de múltiples observaciones o respuesta media

#Lo ploteamos

# 64 es la prediccion

# el maximo es sobre sobre 77

# el minimo sobre 50

plot(x,y,ylim=c(0,100))

abline(modelo,col="blue")

lines(c(5,5),c(prediccion[2],prediccion[3]), col="brown")

points(c(5,5),c(prediccion[2],prediccion[3]), col="brown")

dato=data.frame(x,y)

abl=coefficients(modelo)
```

```

```
(Intercept)   x
  47.046729   3.448598

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

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

```
	Shapiro-Wilk normality test

data: e
W = 0.85361, p-value = 0.2062

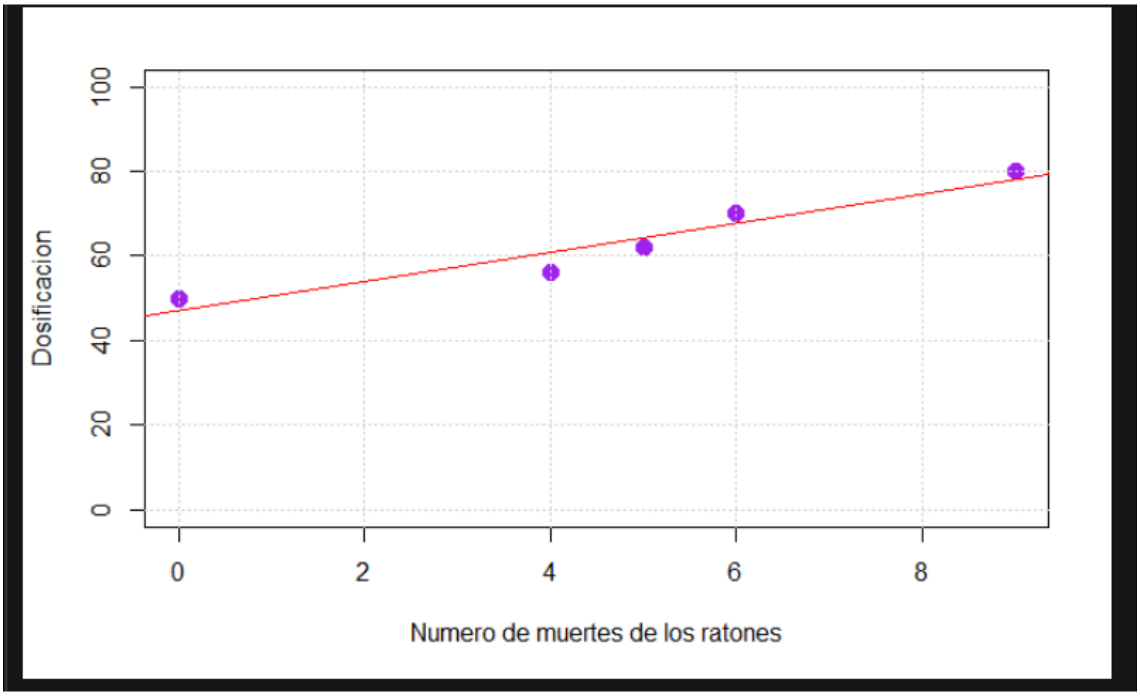
	One-sample Kolmogorov-Smirnov test

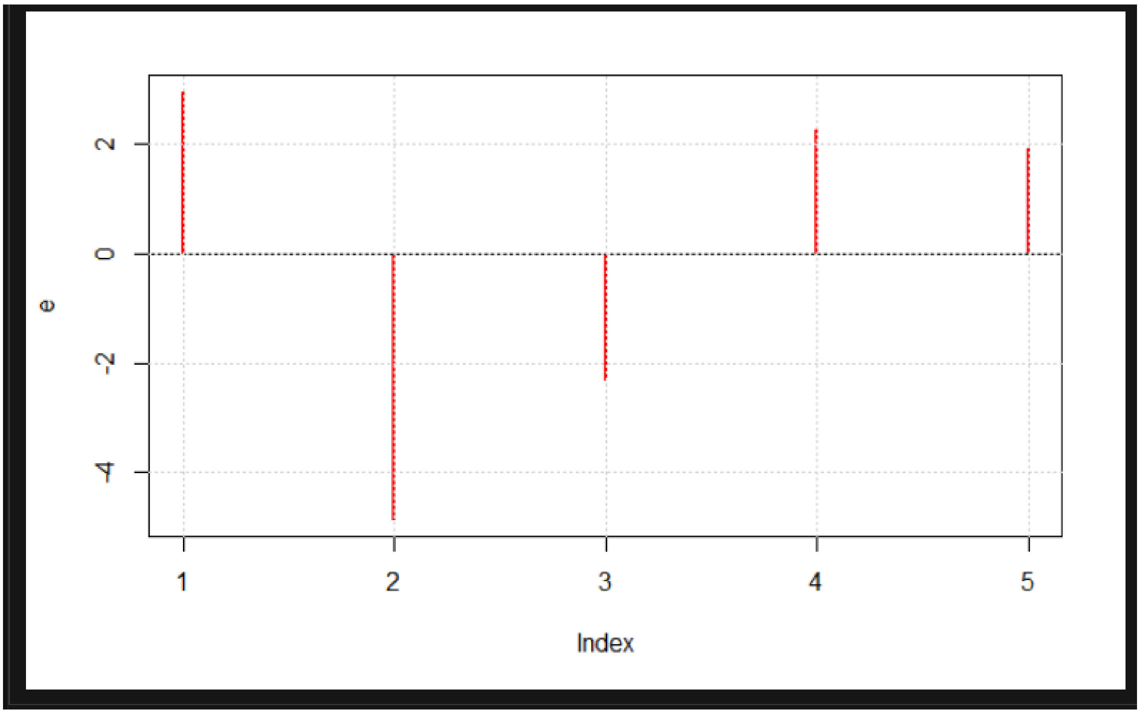
data: e
D = 0.57231, p-value = 0.04424
alternative hypothesis: two-sided

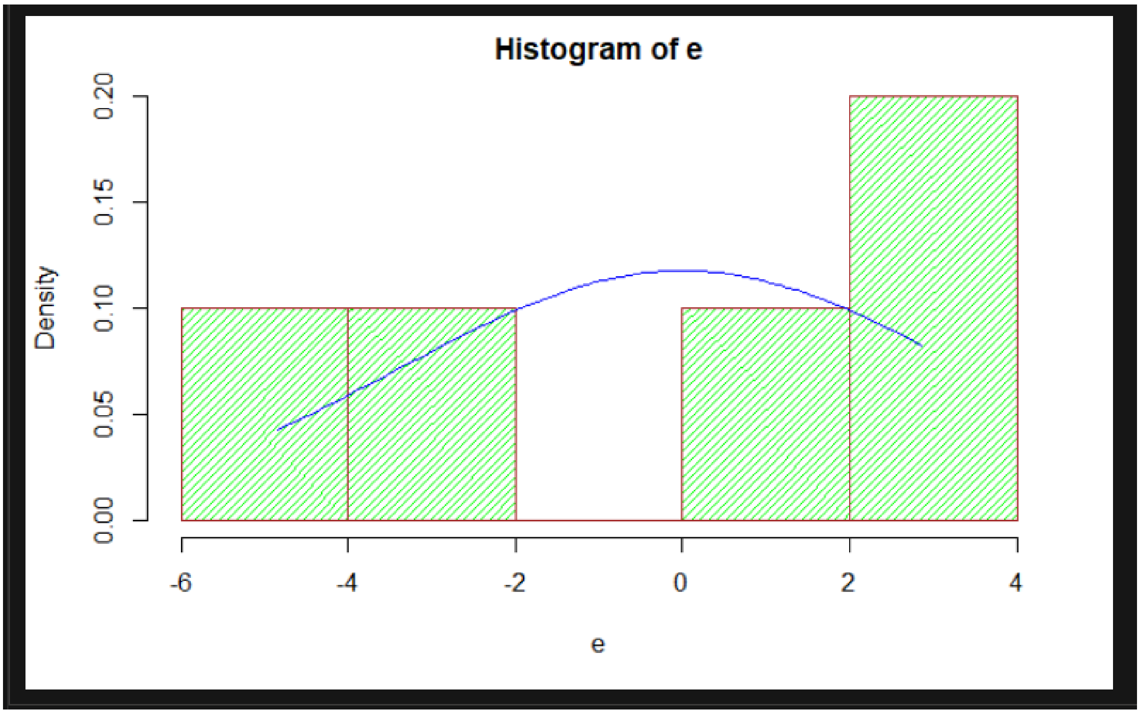
	Jarque Bera Test

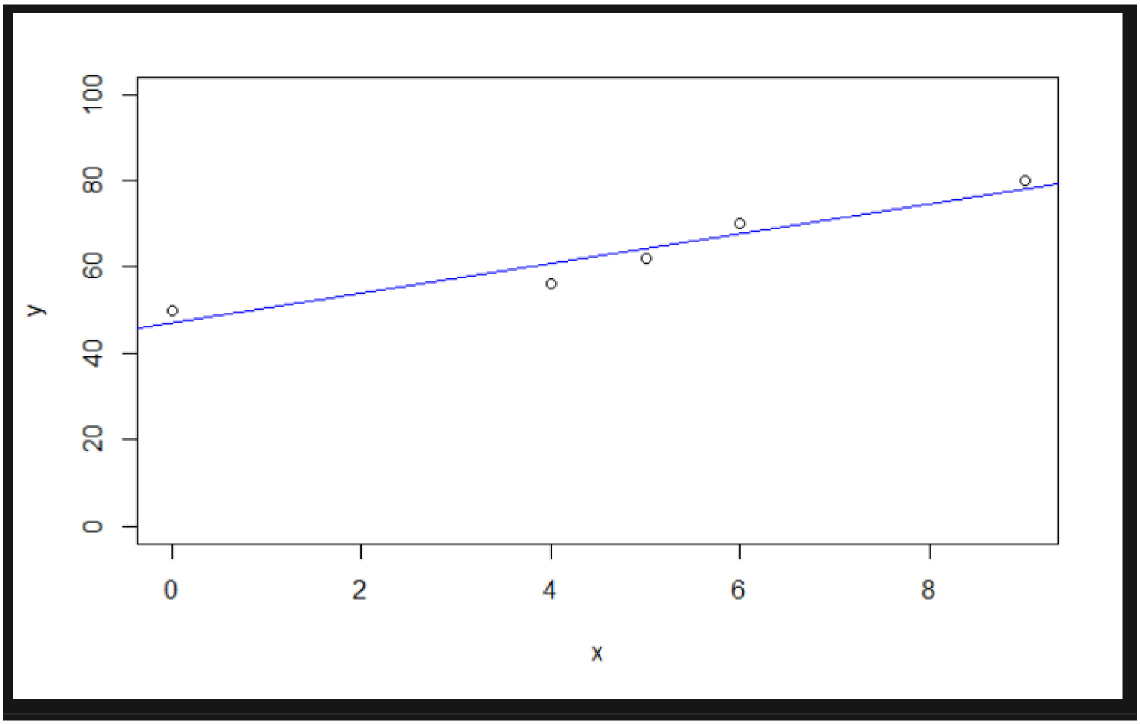
data: e
X-squared = 0.67318, df = 2, p-value = 0.7142

                2.5 %   97.5 %
(Intercept) 36.317177 57.77628
x   1.539896 5.35730
(Intercept)   x
  47.046729   3.448598
```
```


Dosificacion
0
20
40
60
80
100
0
2
4
6
8
Numero de muertes de los ratones


e
-4
-2
0
2
1
2
3
4
5
Index


Histogram of e
Density
0.00
0.05
0.10
0.15
0.20
-6
-4
-2
0
2
4
e


y
0
20
40
60
80
100
0
2
4
6
8
x

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:

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.

```
```{r}

# Ensayos

x=c(8.1,7.8,8.5,9.8,9.5,8.9,8.6,10.2,9.3,9.2,10.5)

#Se obtienen unos resultados

y=c(1.0,1.1,1.2,1.3,1.4,1.5,1.6,1.7,1.8,1.9,2.0)

n=length(x)

#Lo ploteamos

plot(x,y,pch=19,col="red")


 #Calculamos el modelo de regresion lineal simple

modelo=lm(y~x)

summary(modelo)

b1=coefficients(modelo)

#Si el coeficientes es 0 no hay relacion entre las variables

# Error estandar es 1/3

# el r cuadrado es 0.4.4999

# el r cuadrado ajustado nos salio  0.4443

  # el p value es del 1 % podemos  admitir que x e y se relacionan pero el ajuste  no es muy bueno
```
```

```
abline(modelo,col="blue",lwd=2)

# Relacion indirecta

# la bondad de ajuste la calculamos con el residual

e=residuals(modelo)

plot(e,type="h",lwd=2,col="red")

abline(h=0)

grid()

shapiro.test(e)

# p value bajo

ks.test(e,"pnorm")

# intenta ajustar la acumulada de los errores y el test de shapiro lo ajuste a una normal

# no sale proximo a 0.5

library(tseries)

# tiene en cuenta la simetria

# la distribucion es asimetrica el error.

jarque.bera.test(e)

# lo visualizamos

hist(e,freq=F,col="green",density=25,border="brown")

valores=seq(min(e),max(e),0.1)

#la media es 0

# la sigma es  0.06791268

# calculamos la media y la desviacion estandar

points(valores,dnorm(valores,mean=mean(e),sd=sd(e)),type="l",col="blue")

confint(modelo)

coefficients(modelo)

#Si es posible hacer una prediccion

# Hacemos la prediccion  para el valor 1.75

prediccion=predict(modelo,newdata=data.frame(x=1.75),interval = "pred")

#ploteamos la linea del modelo

plot(x,y,pch=19,col="red")
```

```
grid()

abline(modelo,col="blue",lwd=2)

# la respuesta media es por donde pasa la verdadera recta de regresion

 # a partir de ese valor, como se mueve la respuesta media

# nos sale la estima es 9 y picola prediccion es mas ajustadada , la media es mas pequeña

# Para cada valor de la variable donde puede pasar la recta real

lines(c(1.75,1.75),c(prediccion[2],prediccion[3]),col="brown")

points(c(1.75,1.75),c(prediccion[2],prediccion[3]),col="brown")

#generamos el data frame

datos=data.frame(x,y)

#Usamos el ggplot

library(ggplot2)

g=ggplot(data=datos, aes(x=x,y=y))

g

g+geom_point(colour="red")+

  geom_smooth(method="lm")+

  geom_linerange(aes(x=5,ymin=prediccion[2],ymax=prediccion[3]))
```

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.38589 -0.16483 -0.03325  0.18319  0.37990 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) -1.02204    0.84417  -1.211    0.257  
x            0.27632    0.09213   2.999    0.015 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2472 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
```

```
        Shapiro-Wilk normality test

data:  e
W = 0.97428, p-value = 0.9261

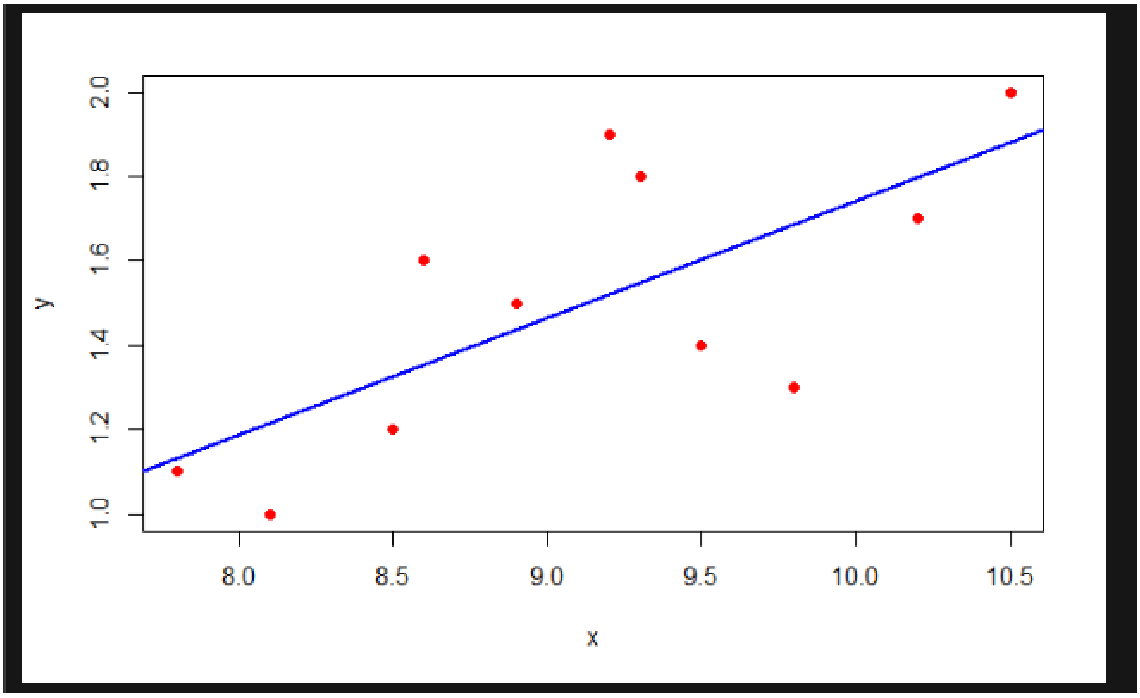

        One-sample Kolmogorov-Smirnov test

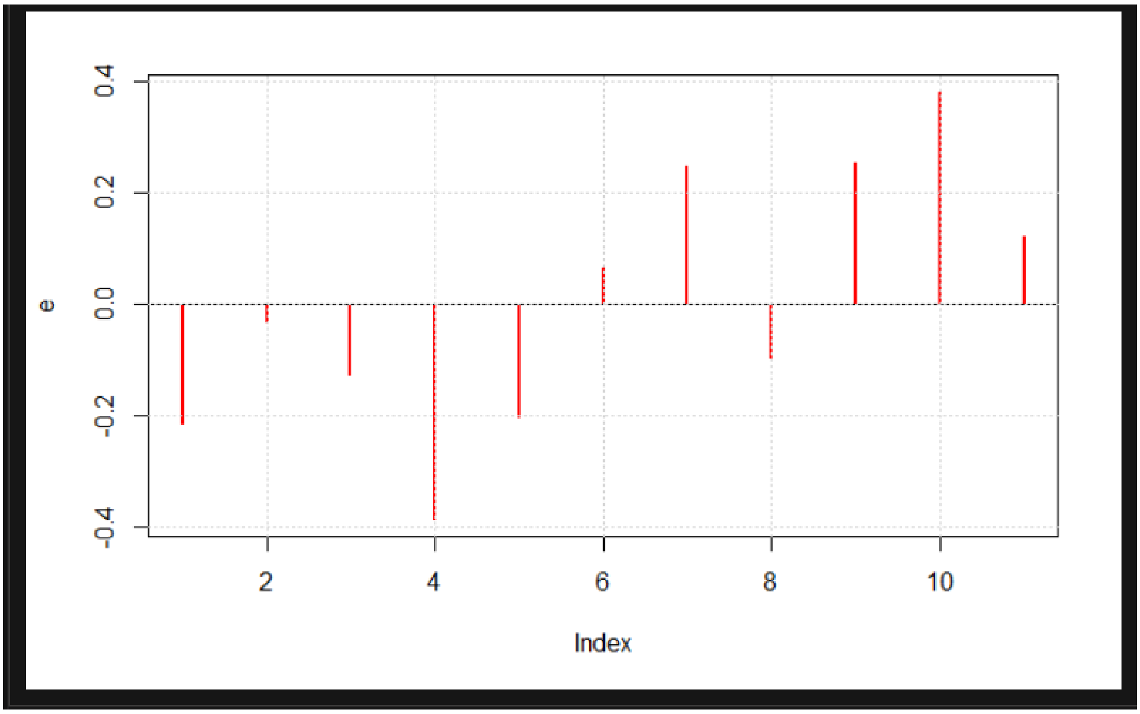
data:  e
D = 0.35201, p-value = 0.1007
alternative hypothesis: two-sided

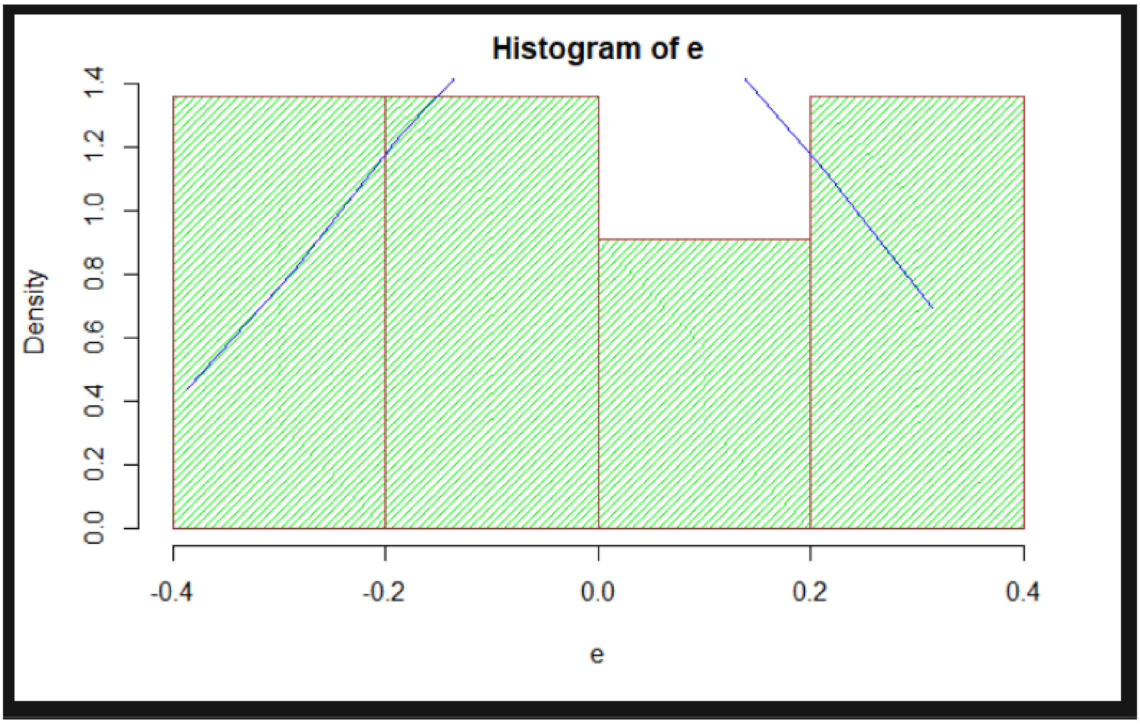

        Jarque Bera Test

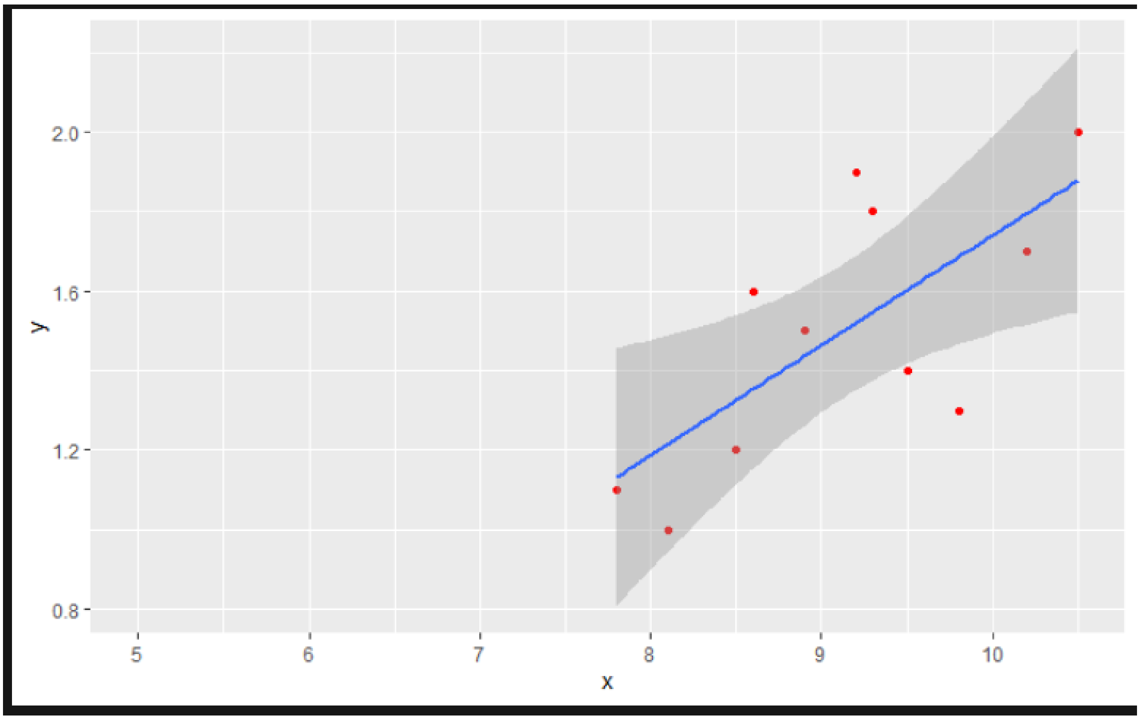
data:  e
X-squared = 0.46357, df = 2, p-value = 0.7931

                  2.5 %    97.5 %
(Intercept) -2.93168224 0.8876025
x            0.06791268 0.4847255
(Intercept)           x
 -1.0220399   0.2763191
```




0.4
0.2
0.0
-0.2
-0.4
e
2
4
6
8
10
Index


Histogram of e
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
-0.4
-0.2
0.0
0.2
0.4
e



2.0
1.6
1.2
0.8
y
5
6
7
8
9
10
x

TEMA 5 LAB

ALEJANDRO DAVID ARZOLA SAAVEDRA

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

```{r}
#)
getwd()
library(MASS)
setwd("C:/Users/34636/Desktop/Alejandro_Arzola_ME_3_entrega/DATOS")
p=read.csv("Alturas_Estudiantes_EII.txt",sep=",",head=T)
p
```

	Alturas <dbl>
1	187.06
2	174.97
3	191.87
4	174.13
5	185.93
6	160.09
7	175.24
8	187.29
9	168.24
10	169.94

1-10 of 635 rows Previous 1 2 3 4 5 6 ... 64 Next

```{r}
#a)
```

 # Estimación de los parámetros μy σde la distribución normal por máxima verosimilitud:

ajuste.normal<-fitdistr(p$Alturas,"normal")

ajuste.normal

 # El ajuste mediante el método de máxima verosimilitud viene dado por una N (173, 18). Para ver y quedarnos con μ y σ

```
      mean          sd     
  173.0663622    18.2622938 
 (  0.7247170) (  0.5124523)
```{r}

#b)

mu<-ajuste.normal$estimate[1]

sigma<-ajuste.normal$estimate[2]

mu

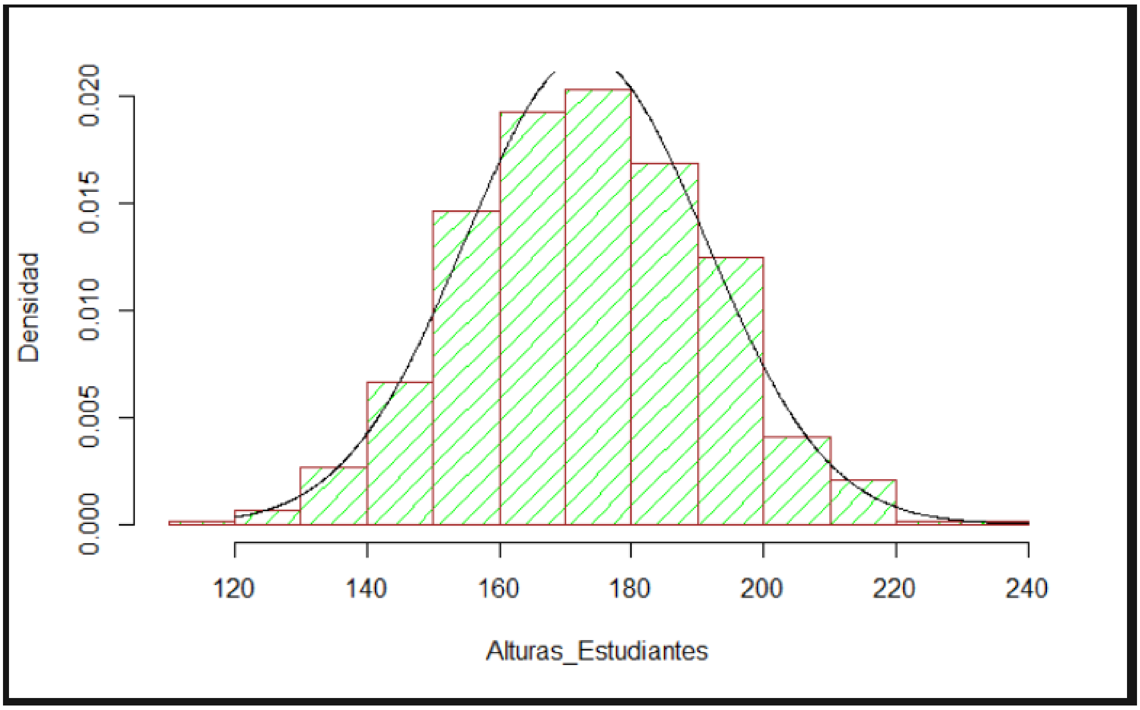
sigma

hist(p$Alturas,breaks=13,freq=FALSE,main="",xlab="Alturas_Estudiantes",ylab="Densidad",col="green",density=10, angle= 45, border= "brown")

lines(1200:2400/10,dnorm(1200:2400/10,mu,sigma,))

```

```
    mean 
173.0664 
      sd 
18.26229 
```
```




```{r}

#c)

#Si lo es, por que como  podemos visualizar  las media donde se encuentra la mayoria de los datos coincide donde nos encuentra con nuestro histograma realizado y que las desviaciones sufridas coinciden con las nuestras

```

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.

```
```{r}

#a)

setwd("C:/Users/34636/Desktop/Alejandro_Arzola_ME_3_entrega/DATOS")

suel2= read.table("sueldos_hosteleria.txt", dec=".", sep = ",", header = T)

suel2

attach(suel2)

x=mean(Sueldos)# devuelve la media muestralde datos.

x

y<-sd(Sueldos) # devuelve la desviación típica muestral de datos.

y

w=sum(sqrt(Sueldos))# devuelve  la suma de las √x

w

z<-length(Sueldos)

z

t.test(suel2, mu=18510, conf.level = 0.95)

alfa=0.1


ci2= x-qt(1-alfa/2, df=z-1)*y/sqrt(z)

cs2= x+qt(1-alfa/2,df=z-1)*y/sqrt(z)
```
```

cm <- c(ci2,cs2)

cm

#Nos aseguramos con un 90 % que el intervalo de confianza se encuentra entre 18411.73 19137.79

```

```
[1] 18774.76
[1] 1060.941
[1] 3424.195
[1] 25

        One Sample t-test

data: suel2
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

[1] 18411.73 19137.79
```

| | Sueldos <dbl> |
|---|---|
| 1 | 19676.77 |
| 2 | 20341.72 |
| 3 | 16695.23 |
| 4 | 16755.38 |
| 5 | 18196.91 |
| 6 | 18355.07 |
| 7 | 20283.26 |
| 8 | 20445.26 |
| 9 | 19145.95 |
| 10 | 18527.83 |

1-10 of 25 rows     Previous 1 2 3 Next

| | Sueldos <dbl> |
|---|---|
| 11 | 18610.32 |
| 12 | 18611.34 |
| 13 | 18673.79 |
| 14 | 18810.31 |
| 15 | 18954.51 |
| 16 | 18387.05 |
| 17 | 19127.75 |
| 18 | 16662.35 |
| 19 | 18935.37 |
| 20 | 17845.27 |

11-20 of 25 rows     Previous 1 2 3 Next
```

	Sueldos <dbl>
21	18057.19
22	19939.68
23	19087.54
24	20090.37
25	19152.85

21-25 of 25 rows Previous 1 2 3 Next

```{r}

#b) # Aplicamos la formula del intervalo de confianza para la varianza

# hallamos los salarios  con α= 0.05, los límites inferior y superior (ci,cs) son:

cvi = (z-1)*y^2/qchisq(1-alfa/2,(z-1))

cvs=(z-1)*y^2/qchisq(alfa/2,(z-1))

cv=c(cvi,cvs)

cv

sqrt(cv)

# Como conclusion verificamos con con un  90 % que el intervlao se encuentra entre 861,304 y 1396.79

```

```
[1]  741844.6 1950712.3
[1]  861.304 1396.679
```

```{r}

#c)

su<-15500:22000

plot(su, dnorm(su, 18510.0,850 ), col="black", type = "l", lty=3,lwd=0.5, xlab = "salarios anuales", ylab = "densidad",col.axis = "blue")

grid()

points(Sueldos, dnorm(Sueldos,18510,0.850), col  = rainbow(25), pch=19)

points(Sueldos, dnorm(Sueldos,18510,0.850), col  = rainbow(25), type = "h", lwd=0.75, lty=3)

# linea que representa la media de los sueldos

abline(v=x, col="green", lwd=2)
```

Intervalo de confianza menor linea que lo representa

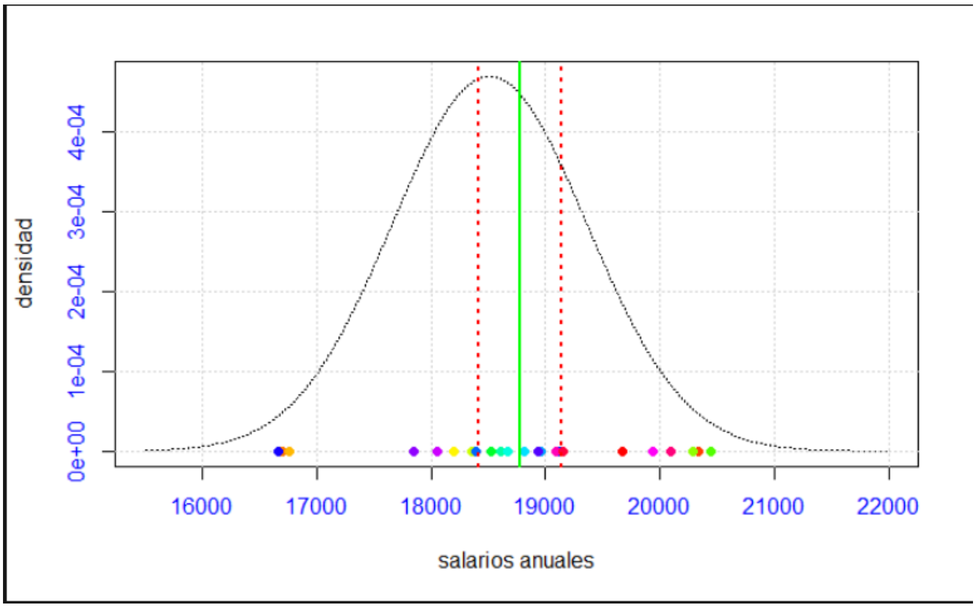
abline(v=ci2, col="red", lty=3, lwd=2)

Intervalo de confianza mayor linea que lo representa

abline(v=cs2, col="red", lty=3, lwd=2)

Como conlcusion tenemos que la mayoria se encuentra entre nuestro intervalo de confianza y los que lo superan o son menores sufren una mayor dispersion y son menos cantidad tambien se observa que se encuentra un poco ladeado de nuestra media

```


```

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 %).

```{r}

setwd("C:/Users/34636/Desktop/Alejandro_Arzola_ME_3_entrega/DATOS")

Datos<-read.table("sueldos_hosteleria.txt", header=TRUE, dec=".",sep=",")

Datos

suel2<-Datos$Sueldos

suel2

hist(suel2,freq=TRUE,main="",labels=TRUE, xlab="sueldos")


#Nos piden que confirmemos la afirmación: el salario medio establecido sea 18510 euros. Hay que a plasmar dicha afirmación en la hipótesis alternativa, que es la única a la que podemos asignar la carga del nivel de confianza.

#Si μ es el salario medio  de los trabajadores


#H0: μ= 18510

#H1: μ≠ 18510


#Como n<30 se deben testear la normalidad de las poblaciones para poder aplicar #los diferentes test que la necesitan

t.test(suel2,alternative="two.sided",mu=18510)

# Dado que el p-valorno es inferior al 5%, no tenemos suficientes evidencias en los datos para rechazar la hipótesis nula (μ= 18510) en favor de la alternativa (μ≠ 18510),


α= 0.05

αa=α/2

t=(18774.763-18510)/(1039.506/sqrt(25))

t


ajuste.normal<-fitdistr(suel2,"normal")
```

mu<-ajuste.normal$estimate[1]

sigma<-ajuste.normal$estimate[2]

c(mu,sigma)

 # Región critica RC: t<-1.710882

 # No Rechazamos H0 un nivel de significancia α=5% cuando el estadístico t calculado excede a tα/2,n -1 o es menor que –tα/2,n -1

#Es decir, que el valor hipotético que hemos considerado para la media, 18510, está dentro de este intervalo, luego éste es un valor de confianza para μ.

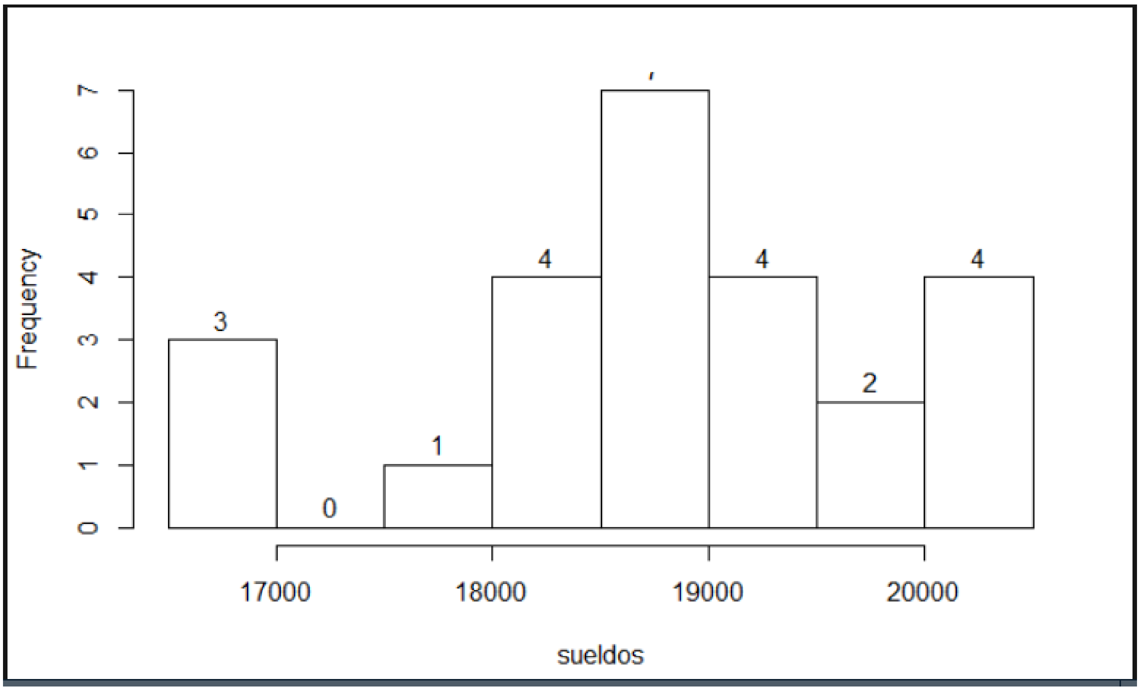
#Si los datos fueran tales que el intervalo de confianza para μdejara fuera al valor 18510, tendríamos razones para pensar que el valor de μes significativamente distinto de 18510

```
 [1] 19676.77 20341.72 16695.23 16755.38 18196.91 18355.07 20283.26 20445.26 19145.95
[10] 18527.83 18610.32 18611.34 18673.79 18810.31 18954.51 18387.05 19127.75 16662.35
[19] 18935.37 17845.27 18057.19 19939.68 19087.54 20090.37 19152.85

        One Sample t-test

data:  suel2
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 

[1] 1.273504
     mean        sd 
18774.763  1039.506 
```



	Sueldos <dbl>
1	19676.77
2	20341.72
3	16695.23
4	16755.38
5	18196.91
6	18355.07
7	20283.26
8	20445.26
9	19145.95
10	18527.83

1-10 of 25 rows Previous 1 2 3 Next

	Sueldos <dbl>
11	18610.32
12	18611.34
13	18673.79
14	18810.31
15	18954.51
16	18387.05
17	19127.75
18	16662.35
19	18935.37
20	17845.27

11-20 of 25 rows Previous 1 2 3 Next

	Sueldos
	<dbl>
21	18057.19
22	19939.68
23	19087.54
24	20090.37
25	19152.85

21-25 of 25 rows

Previous 1 2 3 Next

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

```
```{r}

setwd("C:/Users/34636/Desktop/Alejandro_Arzola_ME_3_entrega/DATOS")

h=read.table("plantas_poda.txt", dec=".", sep = ",", header = T)

h

attach(h)


#--------------------------------------------------------------------------------

#Vamos a visualizar los datos

control<-Biomasa[Tipo_Poda=="control"]

n25<-Biomasa[Tipo_Poda=="n25"]

n50<-Biomasa[Tipo_Poda=="n50"]

r10<-Biomasa[Tipo_Poda=="r10"]

r5<-Biomasa[Tipo_Poda=="r5"]


n<-length(control)

yij<-c(n25,n50,r10,r5,control)

Y_m_T<-mean(yij)

STC<-sum((yij-Y_m_T)^2);STC
```
```

```
SCT<-n*((mean(control)-Y_m_T)^2+(mean(n25)-Y_m_T)^2+ (mean(n50)-Y_m_T)^2+(mean(r10)-
Y_m_T)^2+(mean(r5)-Y_m_T)^2);SCT

SCE<-STC-SCT;SCE

s12<-SCT/4;s12

s2<-SCE/(25);s2

FA<-s12/s2;FA
#a)
#HO= las medias son iguales
#H1= 2 o mas medias no son iguales
# Como F> 2.975154 Se rechaza H0
datos<-data.frame(variable=c(n25,control,n50,r10,r5))
grupo=factor(c(rep(1,n),rep(2,n),rep(3,n),rep(4,n),rep(5,n)))
attach(datos)
ANOVA<-aov(variable~grupo,data= datos)
summary(ANOVA)
qf(0.95,3,26)
# Por lo tanto no podemos  considerar que los 4 metodos producen resultados equivalentes
#b)
#
xdatos<-data.frame(control,n25,n50,r10,r5)
boxplot(xdatos,
col = c("red","green","purple","yellow","blue"),
ylab="Biomasa",
xlab="Tipo_Poda",
staplewex=1,
border= "brown")
grid()
```

 # El metodo superior de poda es el r5 por que es el que mas biomasa de media acaba produciendo al final sera la que mas rendimiento de crecimiento tendra.

```

~~~
[1] 209376.8
[1] 85356.47
[1] 124020.3
[1] 21339.12
[1] 4960.813
[1] 4.301536
            Df Sum Sq Mean Sq F value  Pr(>F)
grupo        4  85356   21339   4.302 0.00875 **
Residuals   25 124020    4961
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] 2.975154
~~~

| | Biomasa <int> | Tipo_Poda <fctr> |
|---|---|---|
| 1 | 551 | n25 |
| 2 | 457 | n25 |
| 3 | 450 | n25 |
| 4 | 731 | n25 |
| 5 | 499 | n25 |
| 6 | 632 | n25 |
| 7 | 595 | n50 |
| 8 | 580 | n50 |
| 9 | 508 | n50 |
| 10 | 583 | n50 |

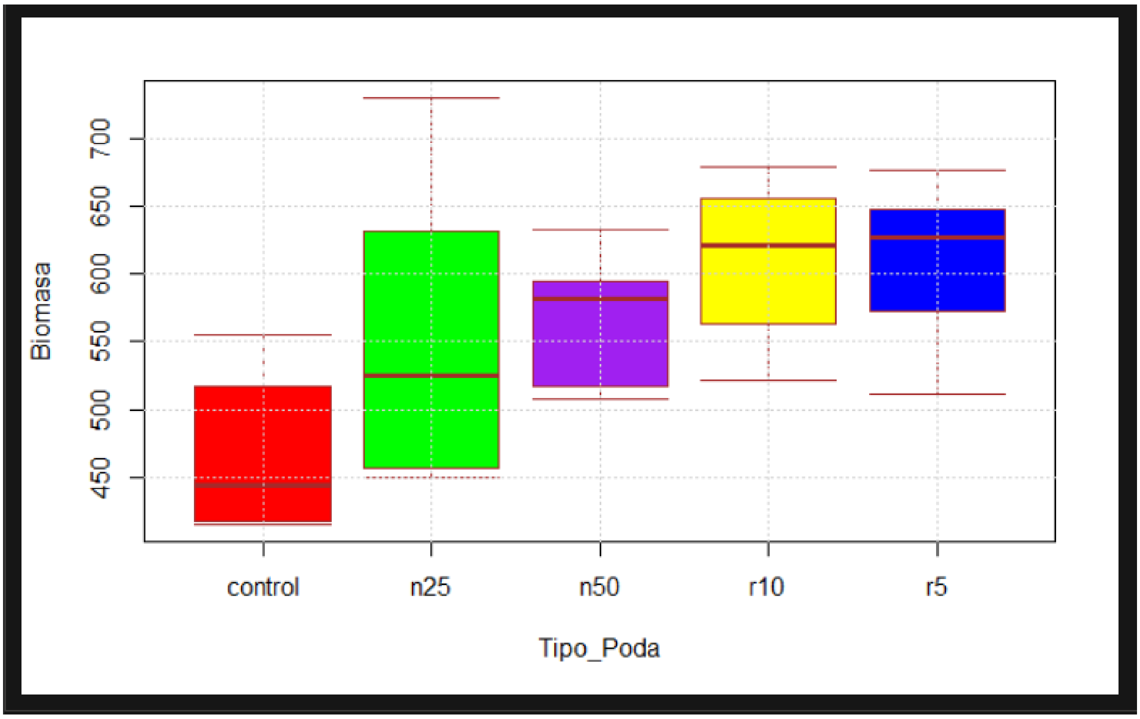
1-10 of 30 rows Previous 1 2 3 Next

| | Biomasa <int> | Tipo_Poda <fctr> |
|---|---|---|
| 11 | 633 | n50 |
| 12 | 517 | n50 |
| 13 | 639 | r5 |
| 14 | 615 | r5 |
| 15 | 511 | r5 |
| 16 | 573 | r5 |
| 17 | 648 | r5 |
| 18 | 677 | r5 |
| 19 | 417 | control |
| 20 | 449 | control |

11-20 of 30 rows Previous 1 2 3 Next
```

	Biomasa <int>	Tipo_Poda <fctr>
21	517	control
22	438	control
23	415	control
24	555	control
25	563	r10
26	631	r10
27	522	r10
28	613	r10
29	656	r10
30	679	r10

21-30 of 30 rows Previous 1 2 3 Next



TEMA 6 LAB

ALEJANDRO DAVID ARZOLA SAAVEDRA

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%.

```{r}

setwd("C:/Users/34636/Desktop/Alejandro_Arzola_ME_3_entrega/DATOS")

library("readxl")

ejercicio1=read.table("ejercicio1_tema6_lab.csv",header=T,sep=";")

names(ejercicio1)

ejercicio1

#Ho: Sigue una distribucion Poisson

#H1: Rechazamos Ho

library(vcd)

attach(ejercicio1)

ajuste<-goodfit(ejercicio1, type="poisson", method="MinChisq")

ajuste

summary(ajuste)

#Podemos rechazar Ho el valor de p es bajo practicamente no se puede aplicar por que se trata de valores pequeños.

chisq.test(ejercicio1, simulate.p.value = TRUE)

```

```
Observed and fitted values for poisson distribution
with parameters estimated by `MinChisq'

 count observed       fitted pearson residual
     0        0 5.707088e-15    -7.554527e-08
     1        0 2.045511e-13    -4.522733e-07
     2        0 3.665719e-12    -1.914607e-06
     3        6 4.379506e-11     9.066480e+05
     4        0 3.924211e-10    -1.980962e-05
     5        0 2.813000e-09    -5.303772e-05
     6        0 1.680373e-08    -1.296292e-04
     7        0 8.603891e-08    -2.933239e-04
     8        0 3.854714e-07    -6.208634e-04
     9        0 1.535101e-06    -1.238992e-03
    10        0 5.502046e-06    -2.345644e-03
    11        0 1.792746e-05    -4.234083e-03
    12        5 5.354573e-05     6.832867e+02
    13        0 1.476280e-04    -1.215023e-02
    14        0 3.779444e-04    -1.944079e-02
    15        0 9.030753e-04    -3.005121e-02
    16        0 2.022978e-03    -4.497753e-02
    17        0 4.265104e-03    -6.530776e-02
    18        0 8.492672e-03    -9.215569e-02
    19        0 1.602057e-02    -1.265724e-01
    20        0 2.871014e-02    -1.694407e-01
    21        0 4.900081e-02    -2.213613e-01
    22        0 7.983032e-02    -2.825426e-01
```

22	0	7.983032e-02	-2.825426e-01
23	0	1.244020e-01	-3.527066e-01
24	0	1.857819e-01	-4.310242e-01
25	0	2.663487e-01	-5.160898e-01
26	0	3.671677e-01	-6.059436e-01
27	0	4.874028e-01	-6.981424e-01
28	0	6.239032e-01	-7.898754e-01
29	0	7.710925e-01	-8.781187e-01
30	4	9.212394e-01	3.207670e+00
31	0	1.065119e+00	-1.032046e+00
32	0	1.192986e+00	-1.092239e+00
33	0	1.295713e+00	-1.138294e+00
34	0	1.365894e+00	-1.168715e+00
35	0	1.398738e+00	-1.182682e+00
36	0	1.392583e+00	-1.180077e+00
37	0	1.348983e+00	-1.161457e+00
38	0	1.272361e+00	-1.127990e+00
39	0	1.169319e+00	-1.081350e+00
40	0	1.047756e+00	-1.023600e+00
41	0	9.159327e-01	-9.570437e-01
42	0	7.816306e-01	-8.840987e-01
43	0	6.515089e-01	-8.071610e-01
44	0	5.307072e-01	-7.284965e-01
45	0	4.226975e-01	-6.501519e-01
46	0	3.293511e-01	-5.738912e-01
47	0	2.511589e-01	-5.011575e-01
48	0	1.875403e-01	-4.330592e-01
49	0	1.371784e-01	-3.703760e-01

```
   49         0 1.371784e-01     -3.703760e-01
   50         0 9.833385e-02     -3.135823e-01
   51         0 6.910669e-02     -2.628815e-01
   52         0 4.763257e-02     -2.182489e-01
   53         0 3.221183e-02     -1.794766e-01
   54         3 2.138006e-02      2.037092e+01
   55         0 1.393264e-02     -1.180366e-01
   56         0 8.917287e-03     -9.443139e-02
   57         0 5.607189e-03     -7.488116e-02
   58         0 3.465010e-03     -5.886433e-02
   59         0 2.104940e-03     -4.587962e-02
   60         0 1.257407e-03     -3.545993e-02
   61         0 7.388106e-04     -2.718107e-02
   62         0 4.270992e-04     -2.066638e-02
   63         0 2.429827e-04     -1.558790e-02
   64         0 1.360764e-04     -1.166518e-02
   65         0 7.503375e-05     -8.662202e-03
   66         0 4.074741e-05     -6.383369e-03
   67         0 2.179779e-05     -4.668810e-03
   68         0 1.148923e-05     -3.389576e-03
   69         0 5.968002e-06     -2.442949e-03
   70         0 3.055753e-06     -1.748071e-03
   71         0 1.542578e-06     -1.242006e-03
   72         0 7.678951e-07     -8.762962e-04
   73         0 3.770216e-07     -6.140209e-04
   74         0 1.826089e-07     -4.273276e-04
   75         0 8.726655e-08     -2.954091e-04
   76         0 4.115490e-08     -2.028667e-04
```

```
   70         0 3.055753e-06     -1.748071e-03
   71         0 1.542578e-06     -1.242006e-03
   72         0 7.678951e-07     -8.762962e-04
   73         0 3.770216e-07     -6.140209e-04
   74         0 1.826089e-07     -4.273276e-04
   75         0 8.726655e-08     -2.954091e-04
   76         0 4.115490e-08     -2.028667e-04
   77         0 1.915658e-08     -1.384073e-04
   78         0 8.802595e-09     -9.382214e-05
   79         0 3.993658e-09     -6.319540e-05
   80         0 1.789238e-09     -4.229939e-05
   81         1 7.917178e-10      3.553979e+04
   82         0 3.460540e-10     -1.860250e-05
   83         0 1.494352e-10     -1.222438e-05
   84         0 6.376186e-11     -7.985194e-06
   85         0 2.688619e-11     -5.184994e-06
   86         0 1.120516e-11     -3.347393e-06
   87         2 4.616216e-12      7.184304e+05
Chi-squared approximation may be incorrect
         Goodness-of-fit test for poisson distribution

                 X^2 df P(> X^2)
Pearson 1.339416e+12 86        0

        Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  ejercicio1
X-squared = 80.591, df = NA, p-value = 0.0004998
```

R Console

data.frame
7 x 2

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

7 rows

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:

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.

```{r}

setwd("C:/Users/34636/Desktop/Alejandro_Arzola_ME_3_entrega/DATOS")

res=read.table("ejercicio2_tema6_laba.csv",header=T,sep=";", row.names = 1,col.names = c(1,2,3))

res

#Ho:  existe realcion entre el tipo de fertilizante empleado

# H1: Rechazamos Ho

attach(res)

test<-chisq.test(res,correct=F)

test

test$observed

test$expected


CHI2<-210*(sum(res)-1)

CHI2

# Región Crítica

gl<-(nrow(res)-1)*(ncol(res)-1)

qchisq(0.95,gl)

library(knitr)


res2<-matrix(c(34,16,73,12,63,12),ncol=3,nrow=2,byrow=F)

colnames(res2)<-c("Fertilizante A","Fertilizante B","Fertilizante C")

rownames(res2)<-c("Han Florecido","No Han Florecido")

res2<-as.table(res2)
```

```
kable(res2)

res_ampliada<-addmargins(res2)

kable(res_ampliada)


ni<-res_ampliada[3,]

nj<-res_ampliada[,4]

N<-as.numeric(res_ampliada[3,4])

pres<-res^2

suma<-0

for(i in 1:3){

for(j in 1:2){

suma<-suma+as.numeric(pres[j,i]/(ni[i]*nj[j]))

}

}

CHI2<-N*(suma-1)

CHI2

gl<-(nrow(res)-1)*(ncol(res)-1);gl

qchisq(0.95,gl)


#Como el valor  2 es menor que el valor límite 5.991465 el estadístico  no está dentro de la RC y no  se
rechaza la hipótesis de independencia


```

```
        Pearson's Chi-squared test

data:  res
X-squared = 7.2316, df = 2, p-value = 0.0269

                 X1 X2 X3
Han Florecido    34 73 63
No Han Florecido 16 12 12
                       X1       X2       X3
Han Florecido    40.47619 68.80952 60.71429
No Han Florecido  9.52381 16.19048 14.28571
[1] 43890
[1] 5.991465
Registered S3 methods overwritten by 'htmltools':
  method               from
  print.html           tools:rstudio
  print.shiny.tag      tools:rstudio
  print.shiny.tag.list tools:rstudio
[1] 7.231557
[1] 2
[1] 5.991465
```

	X1 <int>	X2 <int>	X3 <int>
Han Florecido	34	73	63
No Han Florecido	16	12	12

2 rows

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

	Fertilizante A	Fertilizante B	Fertilizante C	Sum
Han Florecido	34	73	63	170
No Han Florecido	16	12	12	40
Sum	50	85	75	210

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?

```
```{r}

#Ho:Relacion de redes sociales y estuadiantes de ingenieria informatica

#H1:Rechazamos Ho

setwd("C:/Users/34636/Desktop/Alejandro_Arzola_ME_3_entrega/DATOS")

j1=read.table("ejercicio3_tema6_lab.csv",header=T,sep=";",row.names = 1,col.names = c(0,1))

j1

CHI2<-195*(sum(j1)-1)

CHI2

# Región Crítica

gl<-(nrow(j1)-1)*(ncol(j1)-1)

qchisq(0.95,gl)


test<-chisq.test(j1)

test

test$observed

test$expected


library(knitr)

#

j<-matrix(c(75,15,73,32),ncol=2,nrow=2,byrow=F)

colnames(j)<-c("Estudiantes II","Otros Titulos")

rownames(j)<-c("Uso de mas de dos horas","Uso de menos de dos horas")

j<-as.table(j)

kable(j)

j_ampliada<-addmargins(j)

kable(j_ampliada)
```
```

```
ni<-j_ampliada[3,]

nj<-j_ampliada[,3]

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

pj<-j^2

suma<-0

for(i in 1:2){

for(j in 1:2){

suma<-suma+as.numeric(pj[j,i]/(ni[i]*nj[j]))

}

}

CHI2<-N*(suma-1)

CHI2

gl<-(nrow(j)-1)*(ncol(j)-1);gl

qchisq(0.95,gl)

# En una tabla de contingencia de 2 ×2, donde sólo tenemos 1 grado de libertad, se aplica una corrección llamada corrección de Yates para continuidad.

Test2 <- chisq.test(j1, correct=TRUE) # Se hace el test chi-cuadrado y se asigna a la variable .Test (correct=TRUE indica que se aplique la corrección de Yates).

Test2

Test2$expected#  las frecuencias esperadas

round(Test2$residuals^2, 2)

# Por tanto, rechazamos Ho con un nivel de significación p>0.01 (p = 0.0269). No Existe una relación altamente significativa entre la asistencia al curso y el resultado en la prueba, por lo que habría que considerar que se ha mejorado, al menos, en la identificación de rapaces.

# Por lo que para un nivel de significaci´on α = 0,05  se rechaza, aunque con poca evidencia, la hipotesis de independencia.


# Para el caso de tablas 2×2 se aplica el test exacto de Fisher, aunque existe la alternativa de aplicar el test Chi-cuadrado con la correccio´n de Yates.

fisher.test(j1)
```

```
[1] 3.841459

        Pearson's Chi-squared test with Yates' continuity correction

data:  j1
X-squared = 4.3253, df = 1, p-value = 0.03755

                           X0 X1
Uso de mas de dos horas    75 73
Uso de menos de dos horas  15 32
                                 X0       X1
Uso de mas de dos horas    68.30769 79.69231
Uso de menos de dos horas  21.69231 25.30769
[1] 5.052011
numeric(0)
numeric(0)
```

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  j1
X-squared = 4.3253, df = 1, p-value = 0.03755

                                 X0       X1
Uso de mas de dos horas    68.30769 79.69231
Uso de menos de dos horas  21.69231 25.30769
                             X0   X1
Uso de mas de dos horas    0.66 0.56
Uso de menos de dos horas  2.06 1.77

        Fisher's Exact Test for Count Data

data:  j1
p-value = 0.02918
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.046495 4.723462
sample estimates:
odds ratio
  2.183104
```

	X0 <int>	X1 <int>
Uso de mas de dos horas	75	73
Uso de menos de dos horas	15	32

2 rows

	Estudiantes II	Otros Titulos
Uso de mas de dos horas	75	73
Uso de menos de dos horas	15	32

	Estudiantes II	Otros Titulos	Sum
Uso de mas de dos horas	75	73	148
Uso de menos de dos horas	15	32	47
Sum	90	105	195

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.

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?

```{r}

#Ho: n1.2=n2.1 #Dependientes los datos

#H1:Rechazamos Ho

#H0 es que los totales marginales de cada respuesta sean los mismos para cada una de la consultas(pruebas)


setwd("C:/Users/34636/Desktop/Alejandro_Arzola_ME_3_entrega/DATOS")

example=read.table("ejercicio4_tema6_labtotal.csv",header=T,sep=";")

example

# El Estadístico de contraste (concorrección de continuidad) es:

chicuadrado=((12-18)-1)^2/(12+18)

chicuadrado

# El test de Mcnemar se aplica a tablas de contingencia de 2x2 que son resultados de dos consultas (pruebas) obtenidas sobre los mismos individuos (datos apareados).

Respuesta <-matrix(c(103,18,12,35),nrow = 2, dimnames = list("Jefe de Servicio" = c("Brazo Fracturado", "Brazo Normal"),"Internista" = c("Brazo Fracturado", "Brazo Normal")))

Respuesta

# Condiciones de aplicación: Es necesario que el muestreo esté constituido por realizaciones independientes de pares aleatorios (X,Y) acoplados o apareados. Las condiciones de utilización de la aproximación por una distribución χ2 con 1grados de libertad.


#Que n1.2 y n2.1 ean bastante grandes o mayores que 20

mcnemar.test(Respuesta, y = NULL, correct = TRUE)
```

 # Permite comparar dos proporciones sobre dos poblaciones en el caso donde los dos muestreos estén apareados

#Como el valor p es 0.3613 > 0.05 No podemos rechazar Ho

```
[1] 1.633333
                    Internista
Jefe de Servicio   Brazo Fracturado Brazo Normal
  Brazo Fracturado              103           12
  Brazo Normal                   18           35

        McNemar's Chi-squared test with continuity correction

data:  Respuesta
McNemar's chi-squared = 0.83333, df = 1, p-value = 0.3613
```

X <fctr>	Brazo.fracturado <int>	Brazo.Normal <int>
Brazo Fracturado	103	12
Brazo Normal	18	35

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.

```{r}

datos <- data.frame( condicion = c(rep("P", 8), rep("A", 10), rep("B", 10), rep("C", 9)), dias =
c(15,12,10,8,11,9,6,10,7,8,9,8,7,10,9,8,7,10,8,9,8,6,7,8,9,8,7,6,10,12,10,8,9,11,10,9,8))

#Ho= Son iguales

#H1=Rechazamos Ho

aggregate(dias ~ condicion, data = datos, FUN = median)

aggregate(dias ~ condicion, data = datos, FUN = sd)

require(ggplot2)

ggplot(data = datos, mapping = aes(x = dias, colour = condicion)) +

  geom_histogram() +

  theme_bw() +

  facet_grid(. ~ condicion) +

  theme(legend.position = "none")


 kruskal.test(dias ~ condicion, data = datos)


 # El test encuentra significancia en la diferencia de al menos tres grupos

 # debido a que p value es menor que 5 entonces rechazamos Ho

pairwise.wilcox.test(x = datos$dias, g = datos$condicion, p.adjust.method = "holm" )


```

condicion <fctr>	dias <dbl>
A	8
B	8
C	10
P	10

4 rows

condicion <fctr>	dias <dbl>
A	1.159502
B	1.074968
C	1.322876
P	2.695896

4 rows

```
        Kruskal-Wallis rank sum test

data:  dias by condicion
Kruskal-Wallis chi-squared = 11.68, df = 3, p-value = 0.008564
```

```
        Pairwise comparisons using Wilcoxon rank sum test

data:  datos$dias and datos$condicion

  A     B     C
B 0.513 -     -
C 0.162 0.024 -
P 0.260 0.107 0.731

P value adjustment method: holm
```

TEMA 7 LAB

ALEJANDRO DAVID ARZOLA SAAVEDRA

Ejercicio

El fichero "Aloe_Vera.txt" contiene datos de cuatro variedades de plantas de Aloe obtenidas de una plantación experimental. Con estos datos:

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 x0=15.5 gramos y encontrar un intervalo de confianza para la misma.

f) Encontrar el coeficiente de determinación R2.

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

h)Analizar si sería 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

````
```{r}

#Leemos el documento

setwd("C:/Users/34636/Desktop/Alejandro_Arzola_ME_3_entrega/DATOS")

Aloe<-read.table("Aloe_Vera.txt", sep=",", dec=".", header = T)

Aloe

```
````

	Masa <dbl>	Altura <dbl>	Num_Hojas <int>	Masa_Seca <dbl>	Variedad <fctr>
1	28.6	19.1	4	9.3	arborescens
2	20.6	14.8	3	7.7	arborescens
3	29.2	19.7	5	10.4	arborescens
4	32.0	21.1	7	11.5	arborescens
5	24.5	19.4	4	8.4	arborescens
6	29.0	19.5	4	10.3	arborescens
7	28.9	18.9	4	10.1	arborescens
8	18.2	14.6	2	6.3	arborescens
9	7.9	10.2	1	2.7	arborescens
10	15.5	14.6	2	5.5	arborescens

1-10 of 252 rows Previous 1 2 3 4 5 6 ... 26 Next

```{r}

#a)

# Representamos a traves de los boxplots las distintas masas,hojas,altura en funcion del tipo de aloe  y luego aremos una media en funcion de su masa paara por estudiar su rendimiento

boxplot(Aloe$Masa~Aloe$Variedad, col="blue", xlab="Variedades Aloe", ylab="Masa (grms)")

grid()

#Desde el punto de vista de la masa la mejor variedad de aloe es la arborescens

boxplot(Aloe$Masa_Seca~Aloe$Variedad, col="pink", xlab="Variedades Aloe", ylab="Masa Seca (grms)")

grid()

#Y en cuanto a su masa y su altura la mejor sigue siendo la arborescens y podemos ver varios outliers en la barbadensis

boxplot(Aloe$Altura~Aloe$Variedad, col="yellow",xlab="Variedades Aloe", ylab="Altura (cms)")

grid()

# En cuanto a la numero de hojas las mejores son la arborescens y la barbadensis

boxplot(Aloe$Num_Hojas~Aloe$Variedad,col="purple", xlab="Variedades Aloe", ylab="Num. Hojas (uds)",border="blue")

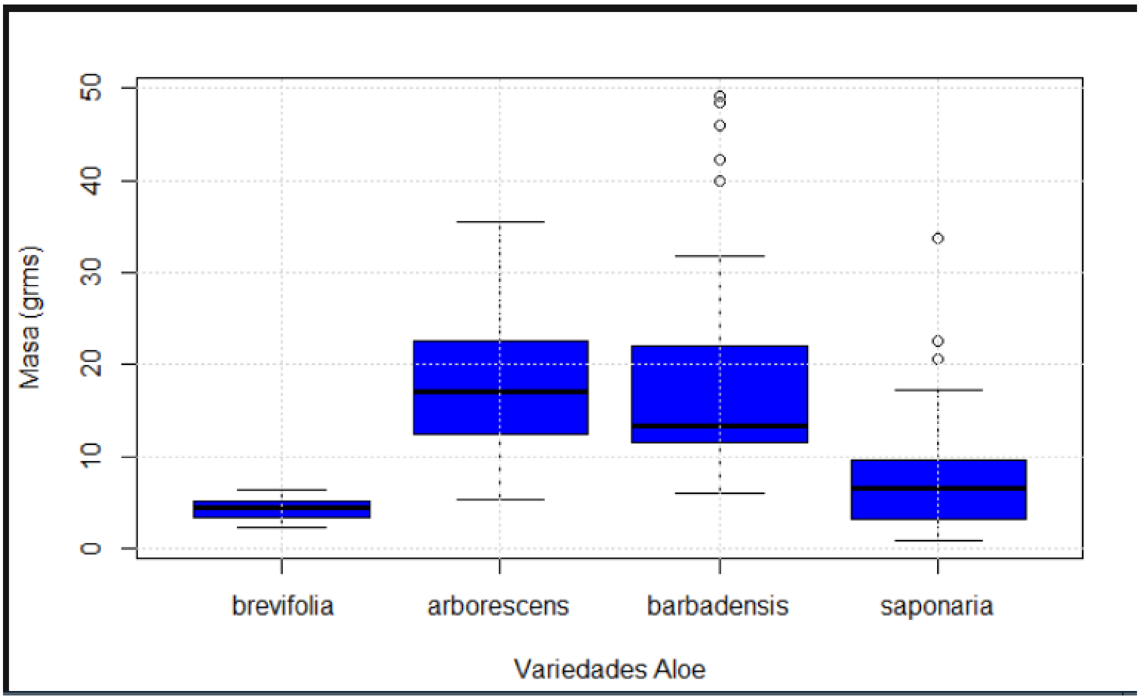
grid()

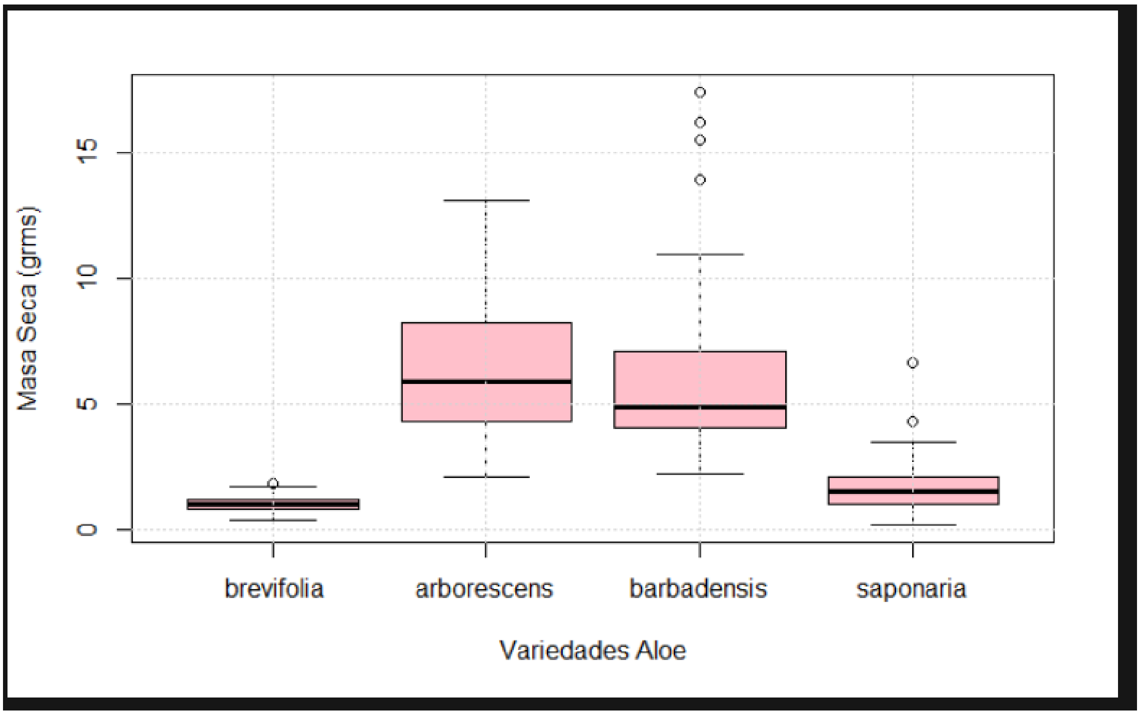
#En general la mejor variedad de aloe es la arborescens

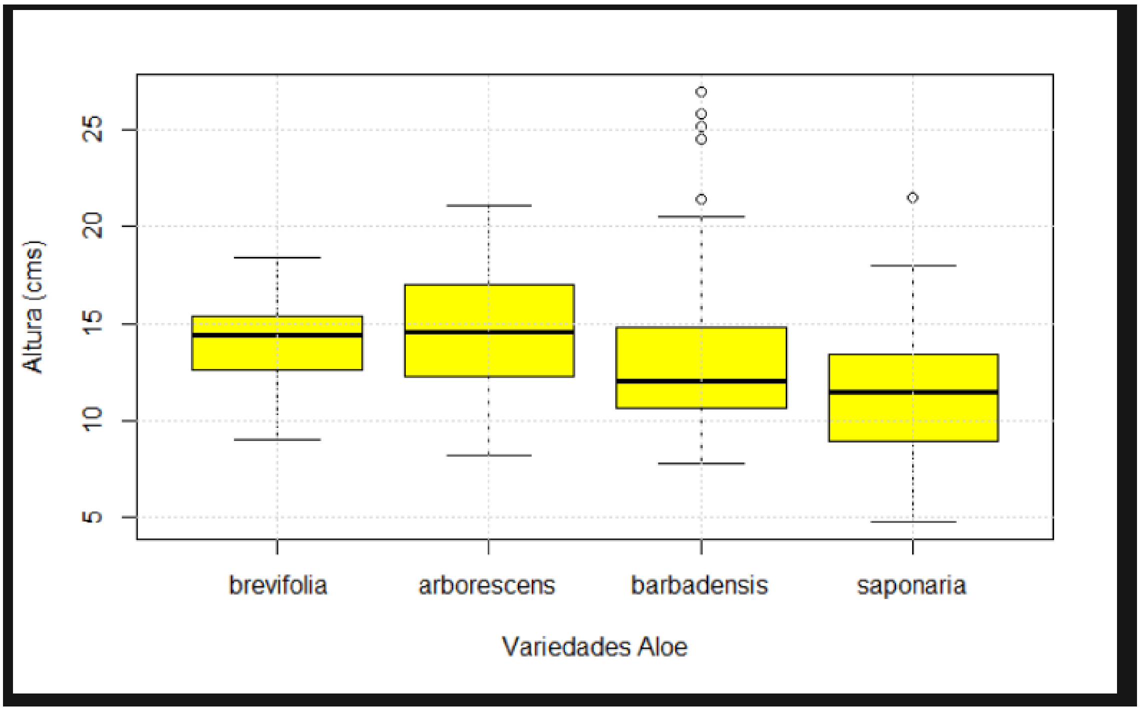
aggregate(Aloe$Masa~Aloe$Variedad, Aloe, mean)
```

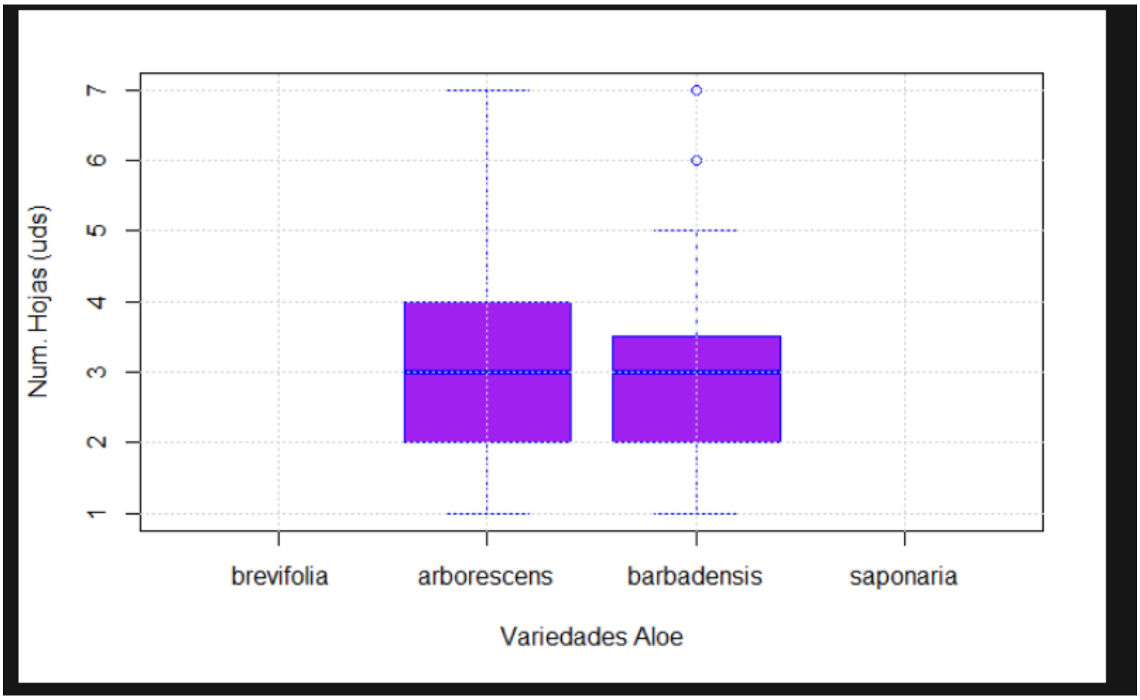
#Podemos ver que el aloe arborescens es el que mas masa produce por lo tanto es el que mas rendimiento produce. y la que menos produce es la brevifolia

```





```





Aloe$Variedad <fctr>	Aloe$Masa <dbl>
brevifolia	4.412500
arborescens	18.081481
barbadensis	17.576786
saponaria	7.521429

4 rows

```{r}
#b)

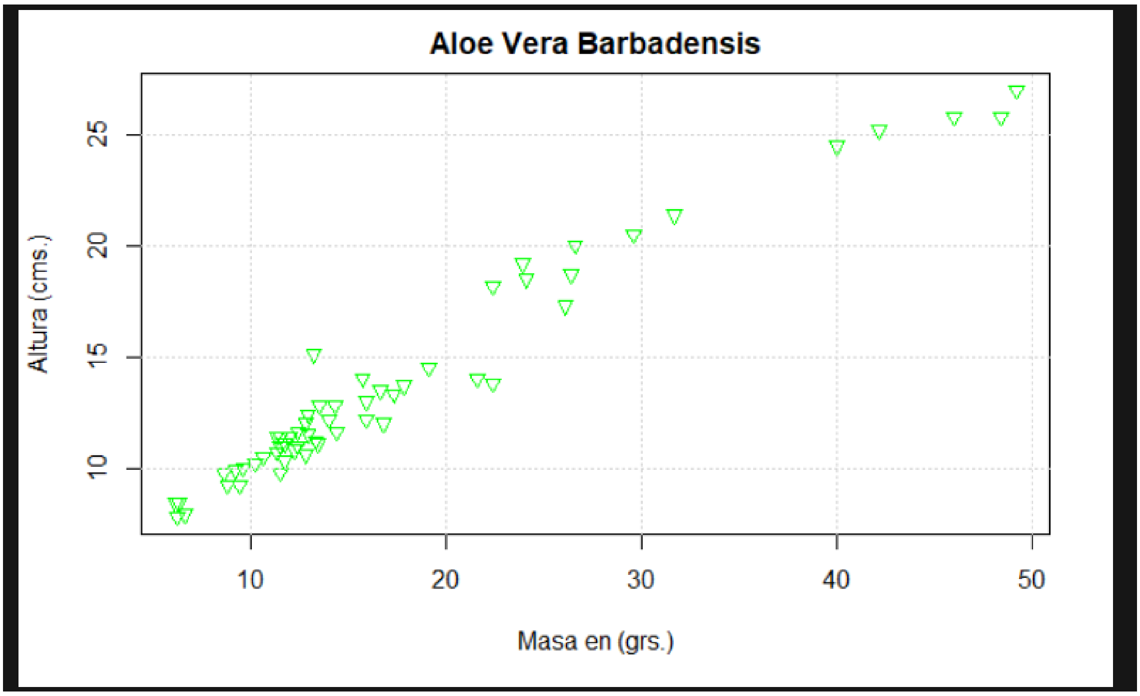

# Con el comando subset seleccionamos la variedad barbadensis

Barbadensis<-subset(Aloe,subset = (Aloe$Variedad=="barbadensis"))

#Ploteamos los datos de la altura en funcion de la masa de la variedad barbadensis

plot(Altura~Masa, data=Barbadensis, pch=25, col="green", xlab = "Masa en (grs.)", ylab="Altura (cms.)",main="Aloe Vera Barbadensis")

grid()
```




```{r}
#c)

#•Seleccionamos la barbadenis del fichero

Barbadensis<-subset(Aloe,subset = (Aloe$Variedad=="barbadensis"))

#Ploteamos la altura en funcion de la masa del aloe barbadensis

plot(Altura~Masa, data=Barbadensis, pch=19, col="blue", xlab = "Masa en (grs.)", ylab="Altura (cms.)", main="Aloe Vera Barbadensis"); grid()

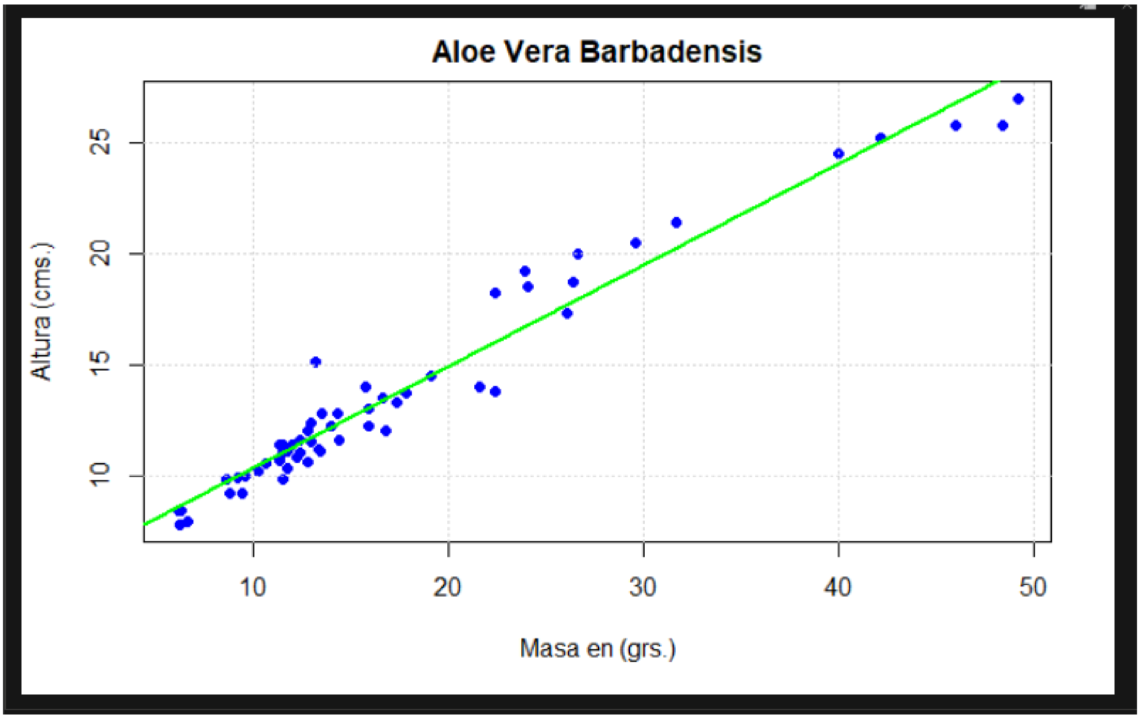
# modelo de regresion lineal simple de la barbadensis
```

modelo1<-lm(Altura~Masa, data=Barbadensis)

representacion de la linea de regresion lineal

abline(modelo1, col="green", lwd=2)

```




```{r}

#d)

attach(Barbadensis)

# Hacemos un data frame de la masa del barbadensis

gridx <- data.frame(Masa)

gridx

attach(gridx)

#Hacemos una prediccion del modelo segun su masa

CIline <- predict(modelo1,new=gridx,interval="conf",level=0.95)

CIpred <- predict(modelo1,new=gridx,interval="pred",level=0.95)

#Volvemos a plotear la barbadensis

plot(Altura~Masa,pch=15,cex=.5, col="green", xlab = "Altura (cms)", ylab = "Masa (grs)", main="Intervalos de confianza")

~~~
#Ponemos encima la linea del modelo de regresion lineal

abline(modelo1, col="black")

#Y las lineas del intervalo de confianza  y para la prediccion

matlines(gridx,cbind(CIline,CIpred[,-1]), lty=c(1,2,2,3,3),col="red")

legend("topleft",lty=2:3,c("pred","conf"))

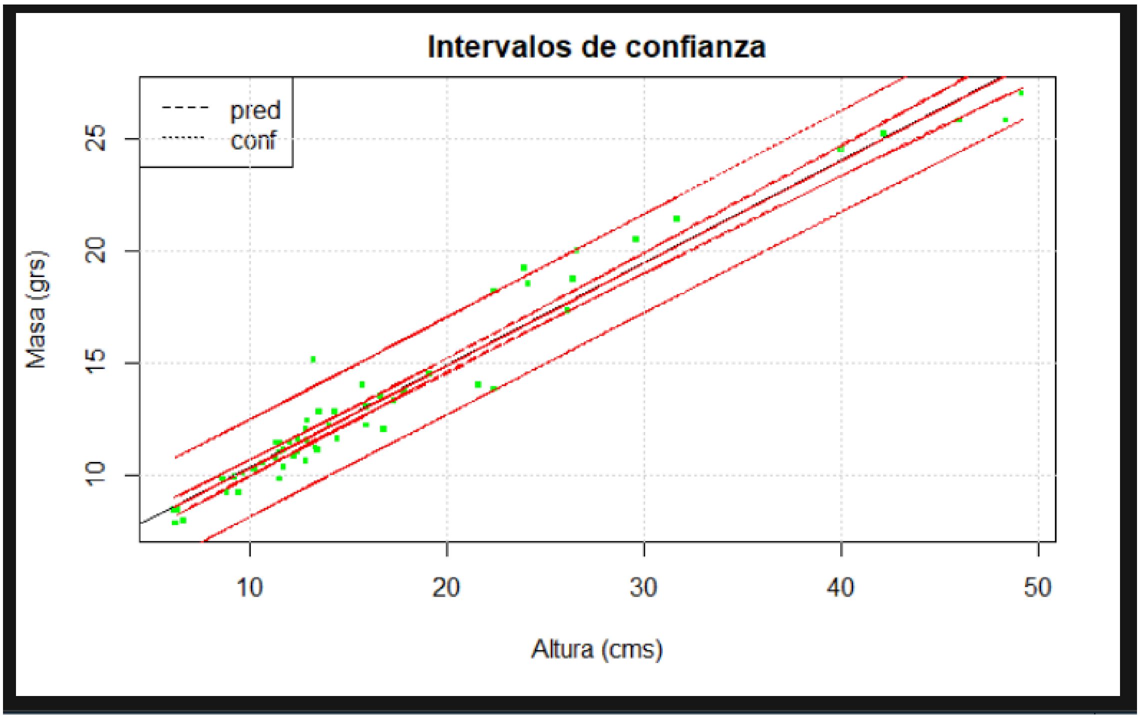
grid()

summary(CIline)

summary(CIpred)

```
~~~

```
      fit              lwr              upr        
 Min.   : 8.533   Min.   : 8.107   Min.   : 8.958  
 1st Qu.:10.998   1st Qu.:10.666   1st Qu.:11.329  
 Median :11.842   Median :11.533   Median :12.151  
 Mean   :13.771   Mean   :13.392   Mean   :14.151  
 3rd Qu.:15.699   3rd Qu.:15.390   3rd Qu.:16.008  
 Max.   :28.206   Max.   :27.294   Max.   :29.118  
      fit              lwr              upr       
 Min.   : 8.533   Min.   : 6.346   Min.   :10.72  
 1st Qu.:10.998   1st Qu.: 8.827   1st Qu.:13.17  
 Median :11.842   Median : 9.674   Median :14.01  
 Mean   :13.771   Mean   :11.588   Mean   :15.95  
 3rd Qu.:15.699   3rd Qu.:13.531   3rd Qu.:17.87  
 Max.   :28.206   Max.   :25.874   Max.   :30.54  
```



Masa <dbl>
40.0
49.2
46.0
26.4
42.2
48.4
23.9
31.7
16.8
21.6

1-10 of 56 rows Previous 1 2 3 4 5 6 Next

Masa <dbl>
24.1
13.5
22.4
26.1
12.9
26.6
29.6
22.4
17.3
16.6

11-20 of 56 rows Previous 1 2 3 4 5 6 Next

Masa
<dbl>
12.8
19.1
12.4
8.8
13.2
15.9
13.3
6.3
12.9
6.2

21-30 of 56 rows Previous 1 2 3 4 5 6 Next

Masa
<dbl>
8.6
14.4
11.5
11.5
12.8
11.7
15.7
12.0
13.4
11.3

31-40 of 56 rows Previous 1 2 3 4 5 6 Next

Masa
<dbl>
6.6
17.8
9.6
14.3
14.0
11.3
10.2
12.2
15.9
11.7

41-50 of 56 rows Previous 1 2 3 4 5 6 Next

Masa
<dbl>
12.4
11.5
10.6
9.4
9.2
6.1

51-56 of 56 rows Previous 1 2 3 4 5 6 Next

```{r}

#e)

#Volvemos ha ahcer lo mismo seleccionamos la barbadensis y plotemos la masa en funcion de la altura de la barbadensis

Barbadensis<-subset(Aloe,subset = (Aloe$Variedad=="barbadensis"))

plot(Altura~Masa, data=Barbadensis, pch=19, col="blue", xlab = "Masa en (grs.)", ylab="Altura (cms.)",main="Aloe Vera Barbadensis"); grid()

# Volvemos ha hacer el modelo de regresion lineal simple que hizimos anteriormente para poder visualizarlo luego para una masa de x0=15.5

modelo1<-lm(Altura~Masa, data=Barbadensis)

abline(modelo1, col="green", lwd=2)

#Prediccion para la masa de x0=15.5 gramos

x0<-15.5

prediccion_masa<-predict(modelo1,list(Masa=x0))

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

lines(c(x0,x0),c(0,prediccion_masa), col="red",lty=3, lwd=3)


#Intervalo confianza predicción

inter_prediccion2<-predict(modelo1, level = 0.95, newdata = data.frame(Masa=x0), interval = "prediction" )

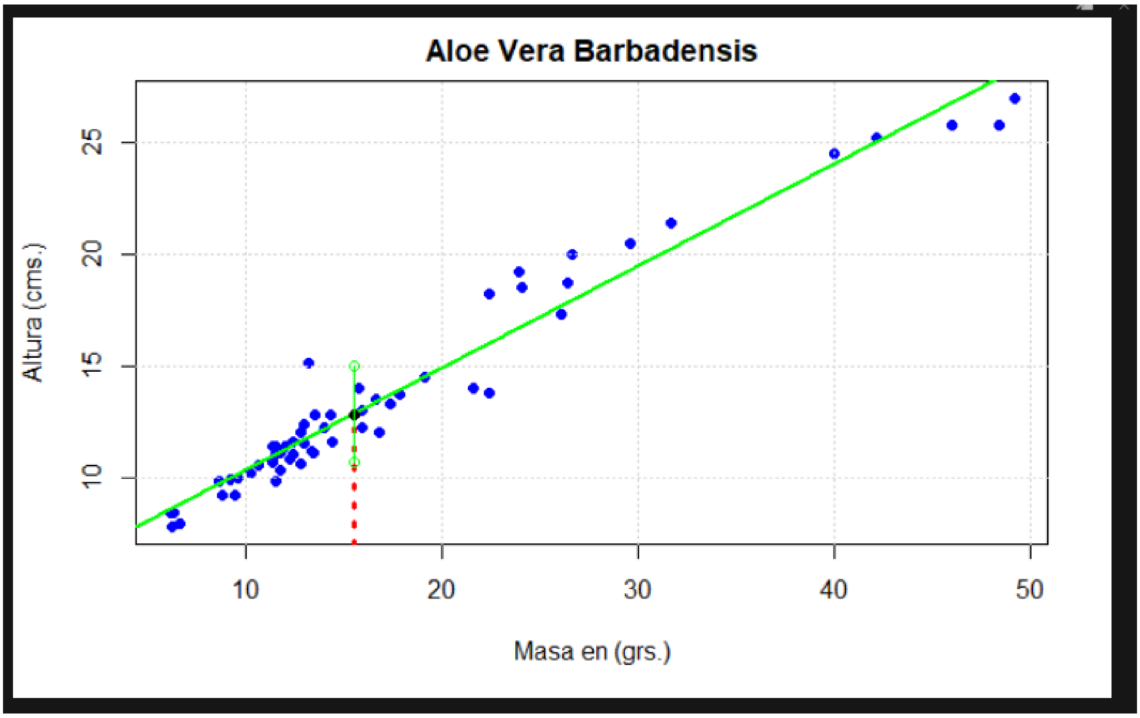
#Lineas de la prediccion del intervalo en funcion  que la masa sea x0=15.5 gramos

lines(c(x0,x0), c(inter_prediccion2[2], inter_prediccion2[3]), col="green")

#puntos de la prediccion del intervalo en funcion  que la masa sea x0=15.5 gramos

points(c(x0,x0), c(inter_prediccion2[2], inter_prediccion2[3]), col="green")

```




```{r}

# f)


#El coeficiente de determinaciónR2es una medida de la proporción de la variabilidad explicada por el modelo ajustado.

#r2= caso de ajuste casi perfecto la SCEse aproximaría a 0 .


SCE<-sum(residuals(modelo1)^2)

SCE

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

1-(SCE/STCC)

summary(modelo1)


k=2 # (número de regresoresindependientes)

n=length(Altura)

adjusR2<-1-(SCE*(n-1)/(STCC*(n-k))); adjusR2


```

```
[1] 61.83836
[1] 0.9538772

Call:
lm(formula = Altura ~ Masa, data = Barbadensis)

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 ***
Masa         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

[1] 0.9530231
```

```{r}
#g)


# H0: β1= 0

# H1: β1≠ 0

Barbadensis

modelo1

qf(0.95,1,110-2)

anova_XY<-aov(modelo1)

anova_XY

summary(anova_XY)

# #la variación en los resultados Y debida a las fluctuaciones de probabilidad o aleatorias son dependientes de los valores de x
```

	Masa <dbl>	Altura <dbl>	Num_Hojas <int>	Masa_Seca <dbl>	Variedad <fctr>
55	40.0	24.5	7	17.4	barbadensis
56	49.2	27.0	7	16.2	barbadensis
57	46.0	25.8	5	13.9	barbadensis
58	26.4	18.7	3	8.3	barbadensis
59	42.2	25.2	5	15.5	barbadensis
60	48.4	25.8	4	16.2	barbadensis
61	23.9	19.2	4	8.0	barbadensis
62	31.7	21.4	5	10.9	barbadensis
63	16.8	12.0	4	5.3	barbadensis
64	21.6	14.0	5	7.2	barbadensis

1-10 of 56 rows   Previous 1 2 3 4 5 6 Next

```
Call:
lm(formula = Altura ~ Masa, data = Barbadensis)

Coefficients:
(Intercept)         Masa
     5.7485       0.4564

[1] 3.929012
Call:
   aov(formula = modelo1)

Terms:
                     Masa Residuals
Sum of Squares  1278.8959   61.8384
Deg. of Freedom         1        54

Residual standard error: 1.070119
Estimated effects may be unbalanced
            Df Sum Sq Mean Sq F value Pr(>F)
Masa         1 1278.9  1278.9    1117 <2e-16 ***
Residuals   54   61.8     1.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```{r}

# h)

#ho= hay HOMOCEDASTICIDAD

#h1= hay HETEROCEDASTICIDAD

library(lmtest)

prueba_white<- bptest(modelo1, ~I(Altura^2)+I(Masa^2)+Altura*Masa,data =Barbadensis)

print(prueba_white)

# Como   el p value es 8.127e-11 < 0.05 Se rechaza la Ho, por lo tanto hay evidencia de que la varianza de que  en los residuos hay HETEROCEDASTICIDAD

```

```
	studentized Breusch-Pagan test

data:  modelo1
BP = 56, df = 5, p-value = 8.127e-11
```

Ejercicio 2. Un comerciante minorista de la zona de Triana quiere analizar la influencia de los costos de publicidad en sus ventas. Durante 3 meses evalúa los costos semanales correspondientes, que se detallan en la siguiente tabla:

a) Visualizar los datos para disponer de una visión clara de su evolución y estimar posibles relaciones entre las variables implicadas.

b) Calcular la ecuación de la recta de regresión para pronosticar las ventas semanales a partir de los gastos de publicidad.

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

d) Evaluar una predicción para unos costes de publicidad de 35€ y encontrar un intervalo de confianza para la misma.

e) Visualizar los intervalos de confianza para la respuesta media y las predicciones del modelo establecido en b)

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

```{r}

#a)

costes_publicidad=c(40,20,25,30,30,50,40,20,50,40,25,50)

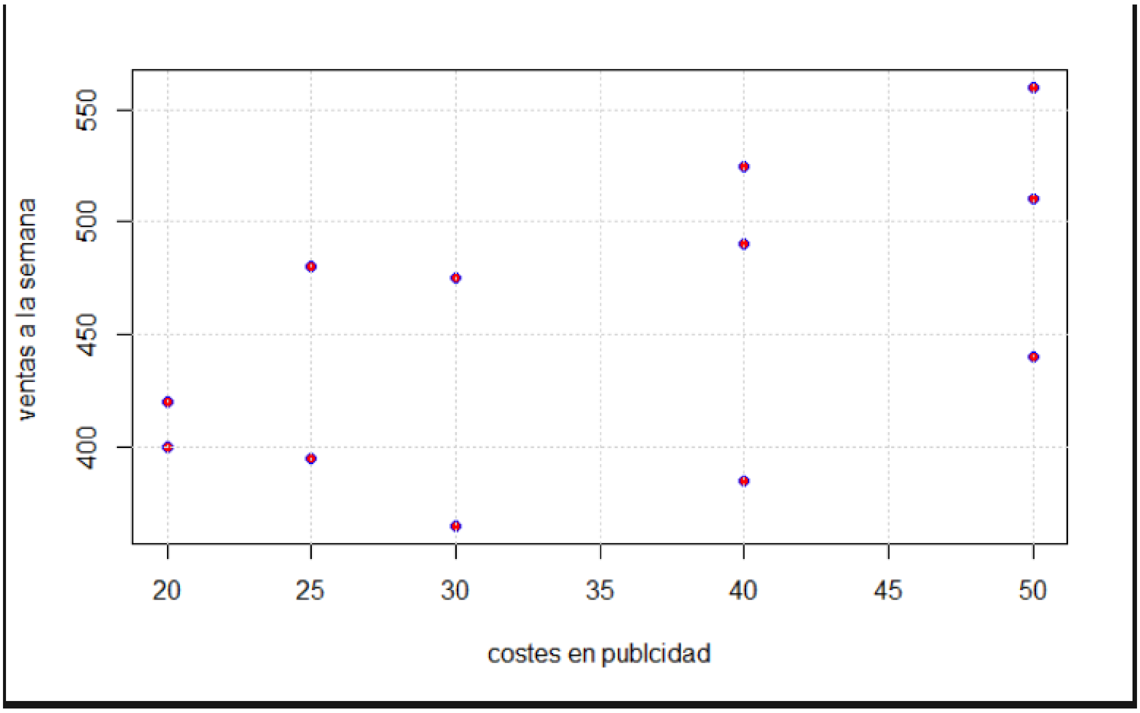
ventas_semana=c(385,400,395,365,475,440,490,420,560,525,480,510)


#Se puede ver claramente como mientras mas costes de publicidad gastes mas ventas a la semanas puedes llegar a tener

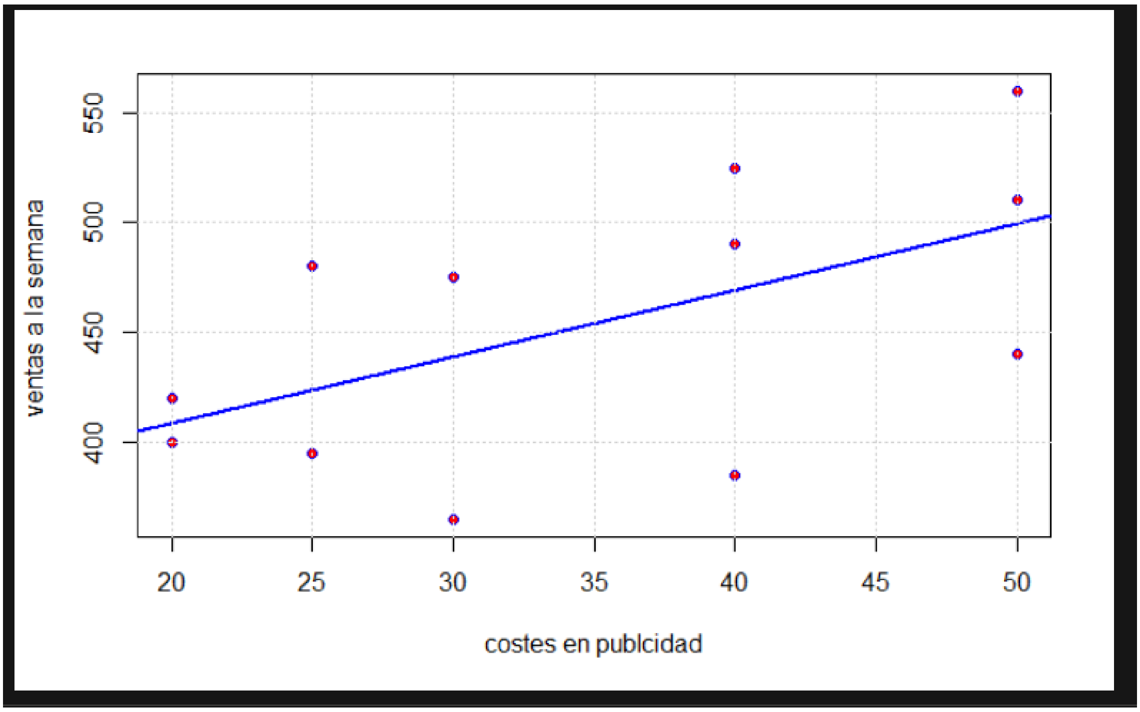
plot(costes_publicidad,ventas_semana, pch=21,col="blue",bg="red", xlab = "costes en publcidad", ylab = "ventas a la semana ")

grid()

```




```{r}
#b)
#Gracias al modelo de regresion lineal podemos predecir segun su coste en publicadad las ventas que puede llegar a tener
plot(costes_publicidad,ventas_semana, pch=21,col="blue",bg="red", xlab = "costes en publcidad", ylab = "ventas a la semana ")
grid()
modelo<-lm(ventas_semana~costes_publicidad)
abline(modelo, col="blue", lwd=2)
grid()
```




```{r}

#c)

#Hacemos un data frame de las ventas

gridx <- data.frame(ventas_semana)

attach(gridx)

#hacemos predicciones para los intervalos de confianzas segun las ventas semanales

CIline <- predict(modelo,new=gridx,interval="conf",level=0.95)

CIpred <- predict(modelo,new=gridx,interval="pred",level=0.95)

#Volvemos a plotear

plot(ventas_semana~costes_publicidad,pch=15,cex=.5, col="red", xlab = "costes en publcidad", ylab = "ventas a la semana ", main="Intervalos de confianza")

abline(modelo, col="blue")

#aplicamos las predicciones

matlines(gridx,cbind(CIline,CIpred[,-1]), lty=c(1,2,2,3,3),col="blue")

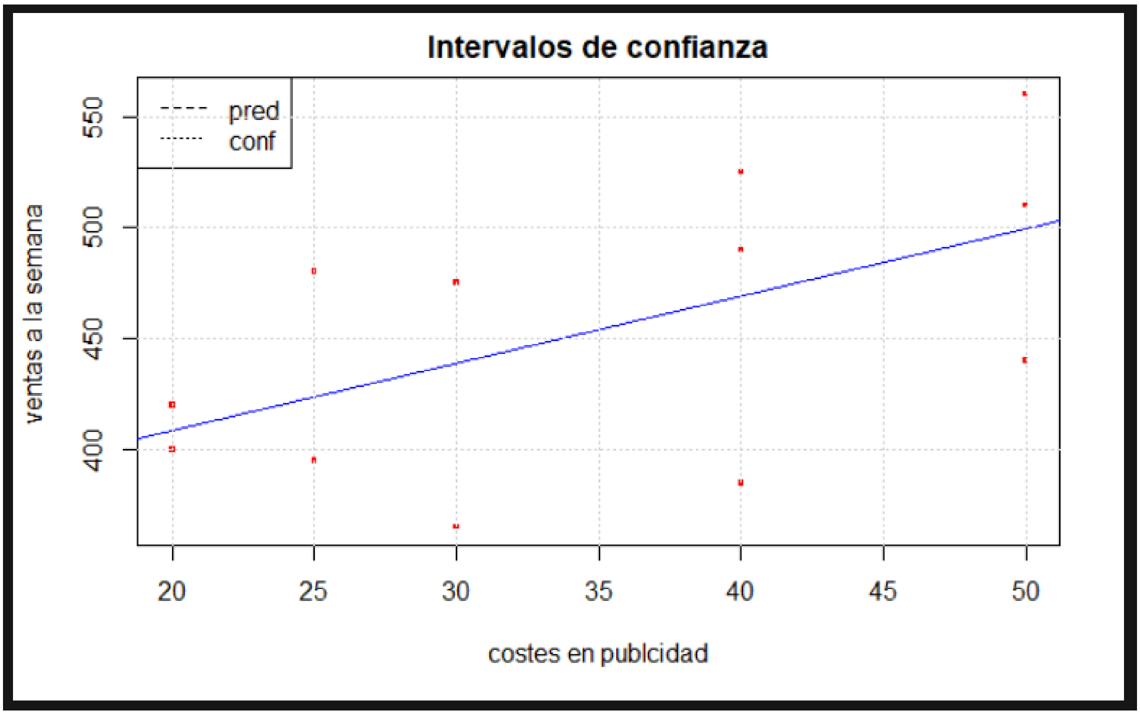
legend("topleft",lty=2:3,c("pred","conf"))

grid()

summary(CIline)

summary(CIpred)
```

```




~~~
      fit              lwr              upr       
 Min.   :408.2   Min.   :349.7   Min.   :466.7  
 1st Qu.:423.4   1st Qu.:376.7   1st Qu.:470.2  
 Median :453.8   Median :415.8   Median :491.7  
 Mean   :453.8   Mean   :405.8   Mean   :501.7  
 3rd Qu.:476.5   3rd Qu.:433.4   3rd Qu.:519.6  
 Max.   :499.3   Max.   :440.8   Max.   :557.8  
      fit              lwr              upr       
 Min.   :408.2   Min.   :274.9   Min.   :541.5  
 1st Qu.:423.4   1st Qu.:294.8   1st Qu.:552.0  
 Median :453.8   Median :328.1   Median :579.4  
 Mean   :453.8   Mean   :324.4   Mean   :583.1  
 3rd Qu.:476.5   3rd Qu.:348.9   3rd Qu.:604.1  
 Max.   :499.3   Max.   :366.0   Max.   :632.6  
~~~

```{r}

#d)

plot(costes_publicidad,ventas_semana, pch=21,col="blue",bg="red", xlab = "costes en publcidad", ylab = "ventas a la semana ")

grid()

modelo<-lm(ventas_semana~costes_publicidad) abline(modelo, col="blue", lwd=2)

grid()

#Vamos a realizar una prediccion para el coste de publicidad =35 $

costespublicdad<-35

prediccion_masa<-predict(modelo,list(costes_publicidad=costespublicdad))

#Representamos la linea y los puntos del coste de 35 $

points(costespublicdad,prediccion_masa, pch=16, col="black")

lines(c(costespublicdad,costespublicdad),c(0,prediccion_masa), col="red",lty=3, lwd=3)

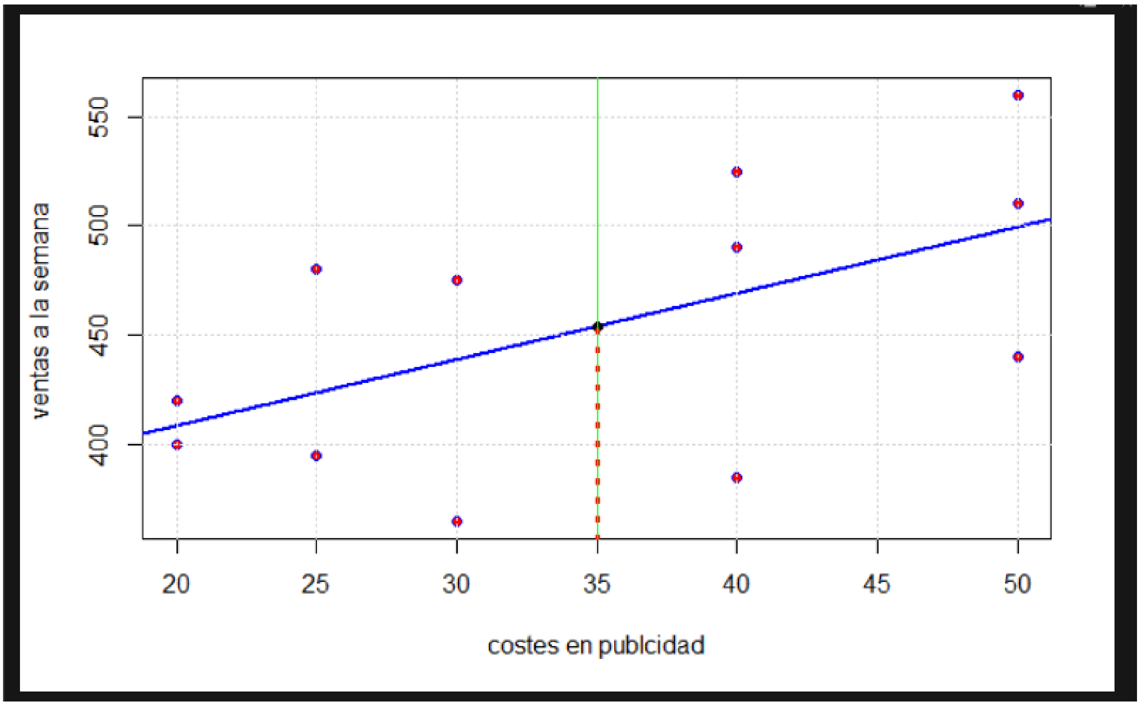
#Intervalo confianza predicción

inter_prediccion2<-predict(modelo, level = 0.95, newdata = data.frame(costes_publicidad=costespublicdad), interval = "prediction" )

lines(c(costespublicdad,costespublicdad), c(inter_prediccion2[2], inter_prediccion2[3]), col="green")

points(c(costespublicdad,costespublicdad), c(inter_prediccion2[2], inter_prediccion2[3]), col="green")

```




```{r}

#e)

# Hacemos un data frame de la masa del barbadensis

```
gridx <- data.frame(ventas_semana)

gridx

attach(gridx)

#Hacemos una prediccion del modelo segun su masa

CIline <- predict(modelo,new=gridx,interval="conf",level=0.95)

CIpred <- predict(modelo,new=gridx,interval="pred",level=0.95)

#Volvemos a plotear la barbadensis

plot(ventas_semana~costes_publicidad,pch=15,cex=.5, col="green", xlab = "costes en publicidad", ylab = "ventas a la semana ", main="Intervalos de confianza")

#Ponemos encima la linea del modelo de regresion lineal

abline(modelo, col="black")

#Y las lineas del intervalo de confianza y para la prediccion

matlines(gridx,cbind(CIline,CIpred[,-1]), lty=c(1,2,2,3,3),col="red")

legend("topleft",lty=2:3,c("pred","conf"))

grid()

summary(CIline)

summary(CIpred)
```

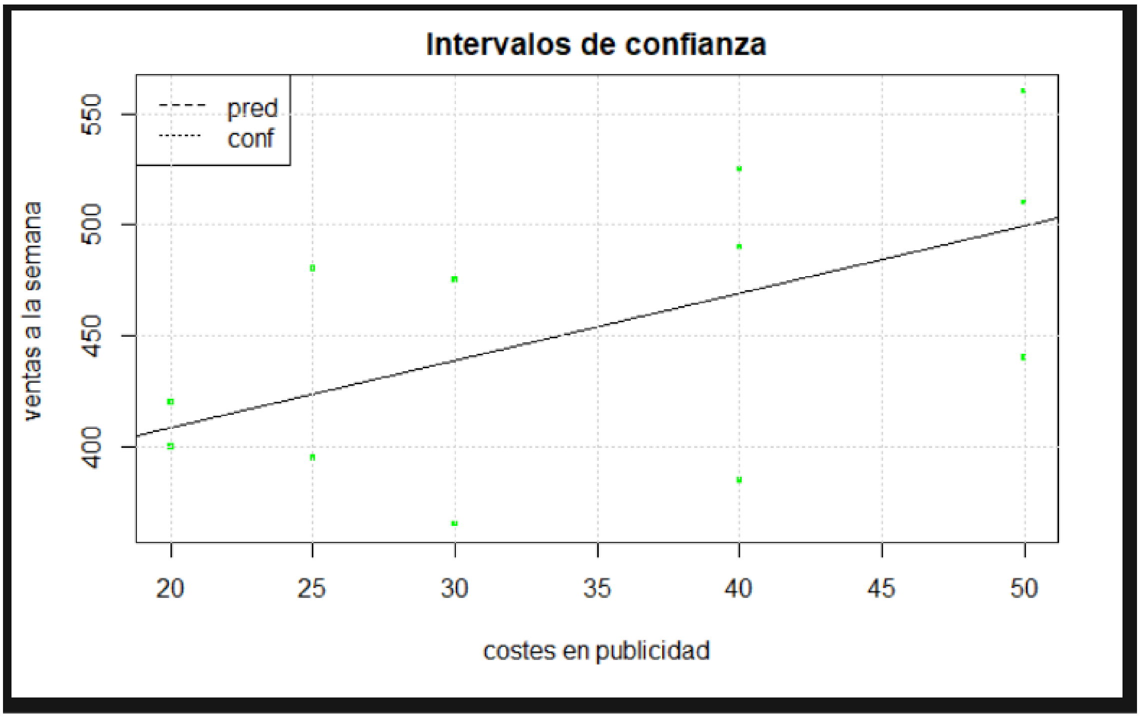
| ventas_semana <dbl> |
|---|
| 385 |
| 400 |
| 395 |
| 365 |
| 475 |
| 440 |
| 490 |
| 420 |
| 560 |
| 525 |

1-10 of 12 rows Previous 1 2 Next

| ventas_semana <dbl> |
|---|
| 480 |
| 510 |

11-12 of 12 rows Previous 1 2 Next

```
      fit              lwr              upr
 Min.   :408.2   Min.   :349.7   Min.   :466.7
 1st Qu.:423.4   1st Qu.:376.7   1st Qu.:470.2
 Median :453.8   Median :415.8   Median :491.7
 Mean   :453.8   Mean   :405.8   Mean   :501.7
 3rd Qu.:476.5   3rd Qu.:433.4   3rd Qu.:519.6
 Max.   :499.3   Max.   :440.8   Max.   :557.8
      fit              lwr              upr
 Min.   :408.2   Min.   :274.9   Min.   :541.5
 1st Qu.:423.4   1st Qu.:294.8   1st Qu.:552.0
 Median :453.8   Median :328.1   Median :579.4
 Mean   :453.8   Mean   :324.4   Mean   :583.1
 3rd Qu.:476.5   3rd Qu.:348.9   3rd Qu.:604.1
 Max.   :499.3   Max.   :366.0   Max.   :632.6
```




```{r}

#f)
```

```
X=c(40,20,25,30,30,50,40,20,50,40,25,50)

Y=c(385,400,395,365,475,440,490,420,560,525,480,510)

qf(0.95,1,12-2)

X_M<-mean(X);X_M

Y_M<-mean(Y);Y_M

X2<-X*X

X_VAR<-sum(X2)/length(X)-X_M^2;X_VAR

Y2<-Y*Y

Y_VAR<-sum(Y2)/length(Y)-Y_M^2;Y_VAR

XY_COVAR<-sum(X*Y)/length(Y)-(X_M*Y_M);XY_COVAR

B1<-XY_COVAR/X_VAR

B1

B0<-Y_M-B1*X_M

B0

boxplot(Y ~ X, col=c("green","blue","yellow","purple","gray"),xlab= "X", ylab= "Y" )

grid()

n<-length(X)

S_2<-(sum((Y-(B0+B1*X))^2))/(n-2)

S_2

SCE<-sum((Y-(B0+B1*X))^2);SCE

STCC<-sum((Y-Y_M)^2);STCC

SCR<-STCC-SCE;SCR

F_SCR<-SCR/S_2;F_SCR

1-pf(F_SCR,1,n-2)


# H0: β1= 0

# H1: β1≠ 0

anova_XY<-aov(modelo)

anova_XY

summary(anova_XY)
```

```
#la variación en los resultadosYdebida a las fluctuaciones de probabilidad o aleatorias son independientes de los valores de x


X_Factor<-as.factor(X)

datos<-data.frame(X_Factor,Y)

Y_M_F<-rep(0,n)

for( i in 1:n) {Y_M_F[i]<-mean(Y[X_Factor==X[i]])}

SCE_Puro<-sum((Y-Y_M_F)^2)

SCE_Puro

SC_Falta_Ajuste<-SCE-SCE_Puro

SC_Falta_Ajuste

k<-nlevels(X_Factor)

S_2_Puro<-SCE_Puro/(n-k)

S_2_Puro

F_SC_Falta_Ajuste<-SC_Falta_Ajuste/(S_2_Puro*(k-2))

F_SC_Falta_Ajuste


1-pf(F_SCR,1,n-2)

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

qf(0.95,1,k-2)



#Cómo el valor del estadístico F de falta de ajuste ("F_SC_Falta_Ajuste") es 0.09744223, inferior al F límite para α=0.05, 10.12796, no puede rechazarse la hipótesis de que el modelo tentativo describe adecuadamente los datos. Además el valor P para este valor de F(0.09744223) es de 0.7753595, muy superior a 0.05 lo que ratifica la anterior suposición.

```

```
[1] 4.964603
[1] 35
[1] 453.75
[1] 120.8333
[1] 3521.354
[1] 366.6667
[1] 3.034483
[1] 347.5431
[1] 2890.453
[1] 28904.53
[1] 42256.25
[1] 13351.72
[1] 4.61925
[1] 0.05713728
```

```
Call:
   aov(formula = modelo)

Terms:
                costes_publicidad Residuals
Sum of Squares           13351.72  28904.53
Deg. of Freedom                 1        10

Residual standard error: 53.76293
Estimated effects may be unbalanced
                  Df Sum Sq Mean Sq F value Pr(>F)
costes_publicidad  1  13352   13352   4.619 0.0571 .
Residuals         10  28905    2890
---
```