UNIVERSIDAD DE LAS PALMAS
DE GRAN CANARIA

MÉTODOS ESTADÍSTICOS
LECTURAS 8,9,10
LABS 6,7
TERCERA ENTREGA

AITOR VENTURA DELGADO

GRADO EN INGENIERÍA INFORMÁTICA

04 DE ENERO DE 2020

ESCUELA DE INGENIERÍA INFORMÁTICA

ÍNDICE

LECTURA 8: CUESTIÓN 1

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

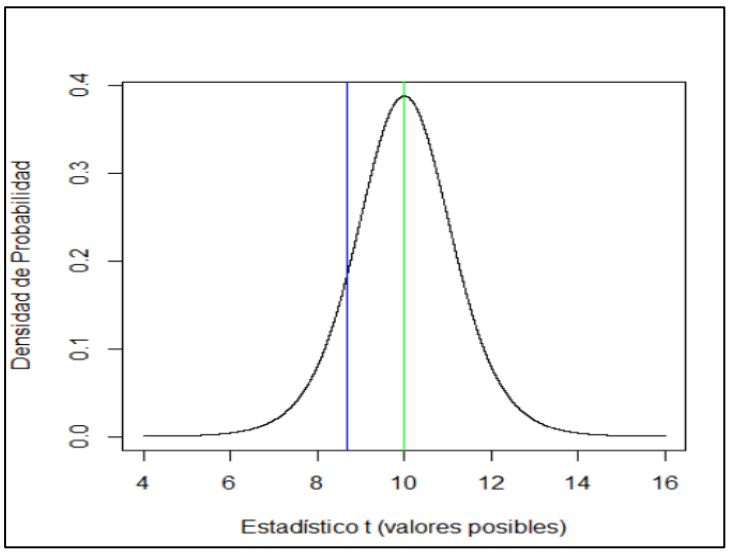
3.58, 10.03, 4.77, 9.71, 10.4, 14.66, 8.45, 5.4, 9.75, 10.1

Utilizando Alpha = 0.05:

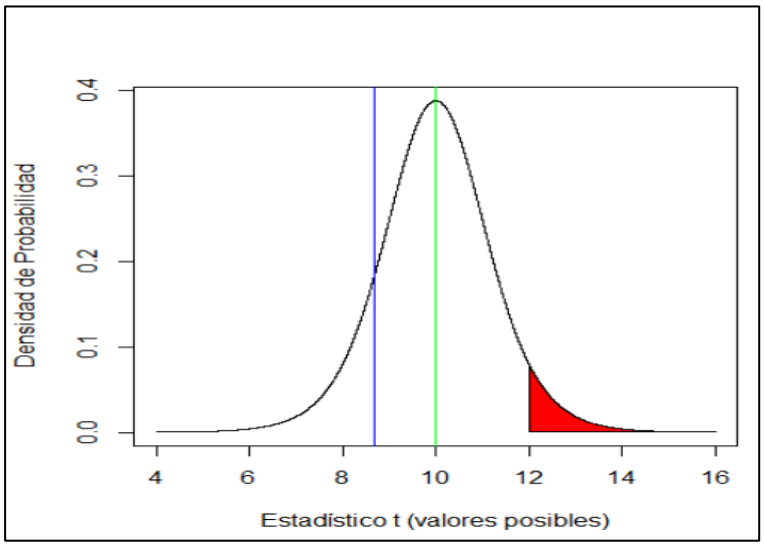
a) ¿Hay evidencias para pensar que la media de la población sea mayor o igual que 10? Razonar la respuesta.

```
> a=0.05
> media_0=10
> datos <- c(3.58, 10.03, 4.77, 9.71, 10.4, 14.66, 8.45, 5.4, 9.75, 10.1)
> U<-mean(datos)
> S<-sd(datos)
> n=10
> t<-(U-media_0)/(S/sqrt(n))
> Zona_critica2<-qt(1-a,n-1)*(S/sqrt(n-1))+media_0
> Zona_critica2
[1] 12.00469
```

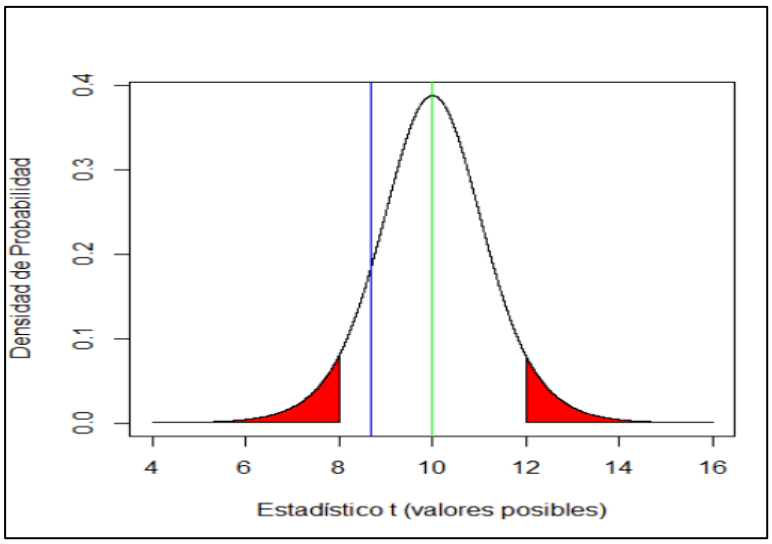
```
> echo=FALSE
> PT<-seq(4,16,0.001)
> TPT<-(PT-media_0)/(S/sqrt(n-1))
> DPO<-dt(TPT, n-1)
> plot(PT,DPO, type = "l", col="black", ylab =
+        "Densidad de Probabilidad", xlab =
+        "Estadístico t (valores posibles)")
> abline(v=media_0, col="green")
> abline(v=U, col="blue")
```



```
> Fliminf<-Zona_critica2
> Flimsup<-16
> xv<-PT[PT>=Fliminf & PT<=Flimsup]
> yv<-DPO[PT>=Fliminf & PT<=Flimsup]
> xv<-c(xv,Flimsup,Fliminf)
> yv<-c(yv,DPO[1],DPO[1])
> polygon(xv,yv,col = "red")
```



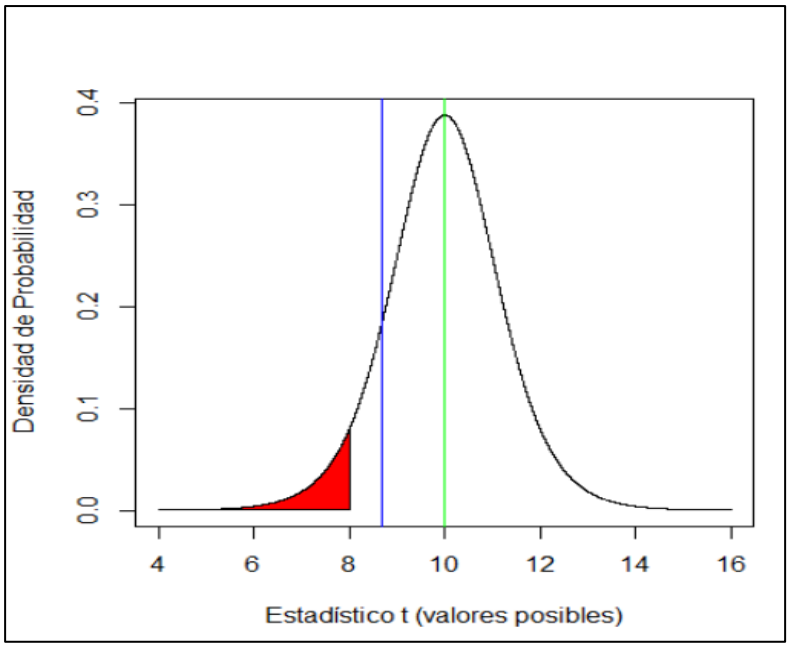

```
> Zona_critica1<-qt(a,n-1)*(S/sqrt(n-1))+media_0
> Fliminf2<-4
> Flimsup2<-Zona_critica1
> xv2<-PT[PT>=Fliminf2 & PT<=Flimsup2]
> yv2<-DPO[PT>=Fliminf2 & PT<=Flimsup2]
> xv2<-c(xv2,Flimsup2,Fliminf2)
> yv2<-c(yv2,DPO[1],DPO[1])
> polygon(xv2,yv2,col = "red")
```



Como observamos en las gráficas nuestra media muestral se encuentra en la región de aceptación evitando las zonas de rechazo por la derecha y por la izquierda lo que nos indica que H0 es válida y la media de la población es igual a 10

b) ¿Podría afirmarse con los datos que la media de la población es inferior a 10?

```
> echo=FALSE
> PT<-seq(4,16,0.001)
> TPT<-(PT-media_0)/(S/sqrt(n-1))
> DPO<-dt(TPT, n-1)
> plot(PT,DPO, type = "l", col="black", ylab =
+        "Densidad de Probabilidad", xlab =
+        "Estadístico t (valores posibles)")
> abline(v=media_0, col="green")
> abline(v=U, col="blue")
>
> Fliminf<-4
> Flimsup<-Zona_critica1
> xv<-PT[PT>=Fliminf & PT<=Flimsup]
> yv<-DPO[PT>=Fliminf & PT<=Flimsup]
> xv<-c(xv,Flimsup,Fliminf)
> yv<-c(yv,DPO[1],DPO[1])
> polygon(xv,yv,col = "red")
```



La media muestral no se encuentra dentro de la zona de rechazo por lo que la hipótesis de que la media es menor a 10 es errónea

c) Calcular los errores tipo I, tipo II, y la potencia de la prueba en su caso.

La probabilidad de cometer errores de tipo 1 en el primer y segundo apartado es de 0.05.

LECTURA 8: 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.

Para tener un intervalo de confianza del 95% alpha=0.05, en este caso

n1 = 16:

- A media muestral 1= 1.516,60
- La desviación estándar 1= 61.500

n2 = 10:

- A media muestral 2= 1.298,35
- La desviación estándar 1= 90.201

Las hipótesis de las que disponemos son:

H0: u1-u2 = 0

H1: u1-u2 != 0

Ello implica que para que H0 sea aceptada en un intervalo de confianza debemos calcular el cuantil de 0.95 y 0.05 respecto a26-2.

```
> qt(0.975,16+10-2)
[1] 2.063899
> qt(0.025,16+10-2)
[1] -2.063899
```

Dístico con distribución t debería ser menor que 2.064 y mayor que -2.064.

b) ¿Cuáles serían las diferencias de los límites si se establece al 90%?

```
> qt(0.95,26-2)
[1] 1.710882
> qt(0.05,26-2)
[1] -1.710882
>
> qt(0.95,26-2)-qt(0.975,16+10-2)
[1] -0.3530165
> qt(0.05,26-2)-qt(0.025,16+10-2)
[1] 0.3530165
```

La diferencia en valor absoluto es de 0.353

c) A partir del resultado de a) razonar sobre la existencia de discriminación salarial entre hombres y mujeres en el ayuntamiento de referencia.

```
> sp = (61500*15+90201*8)/24
> t = (1515.6-1298.35)/(sp*sqrt(1/16+1/10))
> show(t)
[1] 0.007867088
```

Debemos calcular el valor del estadístico t, con la fórmula

$$t = \frac{X - \mu}{s/\sqrt{n}}$$

En donde el numerador representa la diferencia a probar y el denominador la desviación estándar de la diferencia llamado también Error Estándar. En esta fórmula t representa el valor estadístico que estamos buscando, X la barra es el promedio de la variable analizada de la muestra, y miu es el promedio poblacional de la variable a estudiar. En el denominador se encuentra s como representante de la desviación estándar de la muestra y n el tamaño de esta.

Al ser un valor comprendido entre las zonas de rechazo H0 se considera verdadera y no existiría una diferencia salarial.

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 diferente?

Efectuaríamos los mismos cálculos que en apartados anteriores, pero para valores de n1 y n2 diferentes, a su vez, se debe ajustar las zonas de rechazo.

Al encontrarnos de nuevo dentro del intervalo de confianza, rechazamos la hipótesis H1.

```
> z1 = qt(0.975,45+30-2)
> z2 = qt(0.025,45+30-2)
> show(z1)
[1] 1.992997
> show(z2)
[1] -1.992997
>
> sp = (61500*44+90201*28)/73
> t = (1515.6-1298.35)/(sp*sqrt(1/45+1/30))
> show(t)
[1] 0.01286122
```
A

LECTURA 8: CUESTIÓN 3

Se desea conocer la media y 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.

```
> rentas <-c(1500.21,880.66,605.22,1210.12,2010.1,701.3,2060.01,
810.10,1012.34,1500.08,2500,917.45,890.5,515.01,820.39,1800.3,62
5.12,1002.2,2015.22,720.25,1102.45,3200,1601.79,1219.7,1005.4,21
50.1,623.56)
> media_rentas<-mean(rentas)
> desviacionsd_retenas<-sd(rentas)
> z1<-qt(0.975,26)
> z2<-qt(0.025,26)
> val_z1=(z1*desviacionsd_retenas)+media_rentas
> val_z2=(z2*desviacionsd_retenas)+media_rentas
> show(val_z2)
[1] -86.45786
> show(val_z1)
[1] 2679.019
```

Los valores para un intervalo de confianza del 5% van de desde una renta de 2679.01 hasta los -86.45 euros.

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.

```
> nuevas_rentas<-c(1500.21,880.66,605.22,1210.12,2010.1,701.3,20
60.01,810.10,1012.34,1500.08,2500,917.45,890.5,515.01,820.39,180
0.3,625.12,1002.2,2015.22,720.25,1102.45,1601.79,1219.7,1005.4,2
150.1,623.56)
> media_nuevas_rentas<-mean(nuevas_rentas)
> desviacionsd_nuevas_retenas<-sd(nuevas_rentas)
> show(desviacionsd_nuevas_retenas)
[1] 565.7507
```

Nos quedamos con los valores que se encuentran dentro del intervalo de confianza y calculamos la desviación típica, que resulta en un valor de 565.751, no supera los 1000 euros de desviación, propuestos en el enunciado.

LECTURA 8: CUESTIÓN 4

El propietario de un vehículo híbrido de la marca Toyota piensa que el consumo medio de gasolina, un 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:

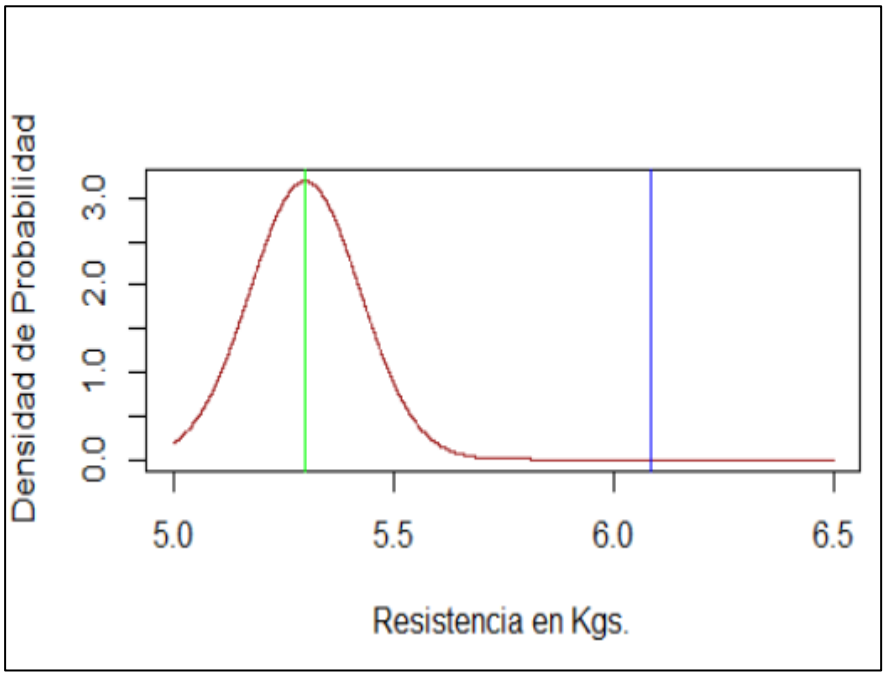
6.2, 6.6, 5.8, 5.4, 5.3, 6.15, 6.68, 7.0, 5.8, 5.6, 5.85, 6.2, 6.4, 6.75, 5.3, 6.3

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 5.35L/100km.

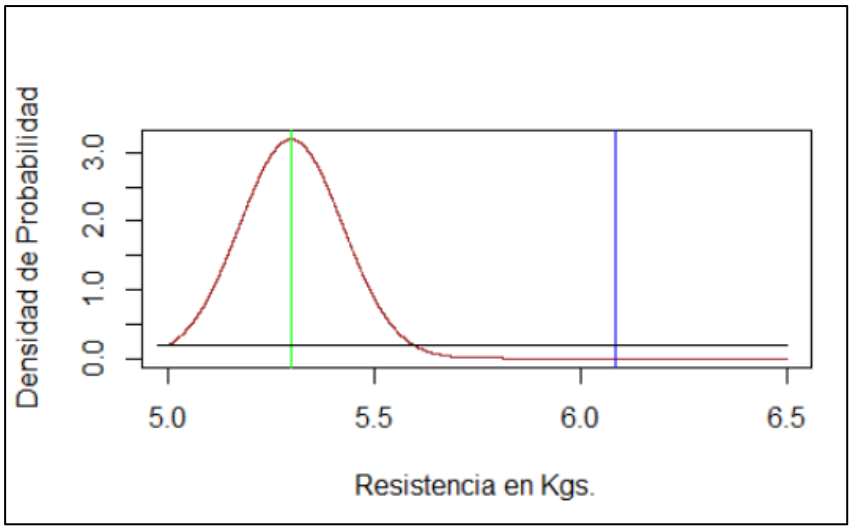
```
> consumo<-c(6.2,6.6,5.8,5.4,5.3,6.15,6.68,7,5.8,5.6,5.85,6.2,
6.4,6.75,5.3,6.3)
> media_consumo=mean(consumo)
> desviacion_consumo=sd(consumo)
> z1<-qnorm(0.995,0,1)
> show(z1)
[1] 2.575829
> z2<-qnorm(0.005,0,1)
> show(z2)
[1] -2.575829
> z_muestra=(media_consumo-5.3)/(desviacion_consumo/sqrt(16))
> show(z_muestra)
[1] 5.899296
```

La zona de rechazo se encuentra entre -2.576 y 2.576, al calcular el estadístico Z se obtiene un valor de 5.899 que nos indica que la hipótesis H0: media= 5.3, no sería cierta.

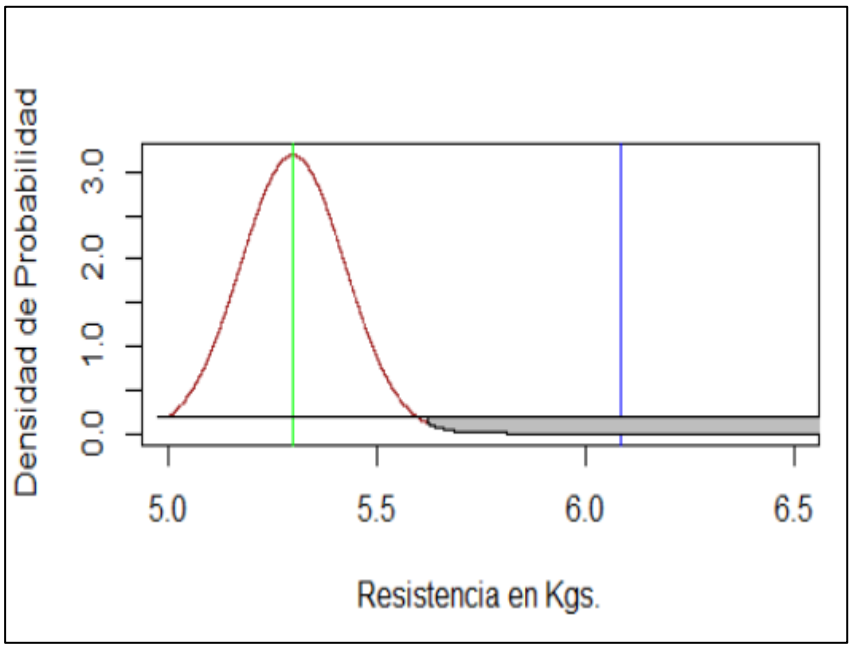
```
> PT<-seq(5,6.5,0.0001)
> n1<-16
> sigma<-0.5
> sigma0<-sigma/sqrt(n1)
> mu0<-5.3
> media_muestra<-media_consumo
> DP0<-dnorm(PT, mu0,sigma0)
> plot(PT,DP0, type = "l", col="brown",
+      ylab = "Densidad de Probabilidad", xlab
+      = "Resistencia en Kgs.")
> abline(v=mu0, col="green")
> abline(v=media_muestra, col="blue")
> alfa<-0.01
```




```
> Zona_critica1<-qnorm((1-alfa/2),mu0,sigma0)
> Zona_critica2<-qnorm(alfa/2,mu0,sigma0)
> Fliminf<-6.5
> Flimsup<-Zona_critica2
> xv<-PT[PT>=Fliminf & PT<=Flimsup]
> yv<-DP0[PT>=Fliminf & PT<=Flimsup]
> xv<-c(xv,Flimsup,Fliminf)
> yv<-c(yv,DP0[1],DP0[1])
> polygon(xv,yv,col = "gray")
```




```
> Fliminf<-Zona_critica1
> Flimsup<-8.4
> xv<-PT[PT>=Fliminf & PT<=Flimsup]
> yv<-DPO[PT>=Fliminf & PT<=Flimsup]
> xv<-c(xv,Flimsup,Fliminf)
> yv<-c(yv,DPO[1],DPO[1])
> polygon(xv,yv,col = "gray")
```



LECTURA 9: 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

```
> library(knitr)
> library(vcd)
> library (PerformanceAnalytics)
```

```
> tabla <- matrix (c(35,47,31,55), 2,2, byrow = TRUE)
> colnames(tabla) <- c("hombres", "mujeres")
> rownames(tabla) <- c ("Interesados", "No interesados")
> datos <- as.table(tabla)
> kable (tabla)


|               | hombres| mujeres|
|:--------------|-------:|-------:|
|Interesados    |      35|      47|
|No interesados |      31|      55|
> summary(tabla)
    hombres      mujeres
 Min.   :31   Min.   :47
 1st Qu.:32   1st Qu.:49
 Median :33   Median :51
 Mean   :33   Mean   :51
 3rd Qu.:34   3rd Qu.:53
 Max.   :35   Max.   :55
```

Se puede decir, a la luz de esos datos, ¿que existe una relación significativa entre el género y esa clasificación?

Se puede observar en la tabla que son más los hombres interesados en participar en la política a pesar de que son menos los hombres encuestados (un total de 82 hombres frente a las 86 mujeres).

*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.*

Pruebas que ya incorpora R:

```
> res1 <- mcnemar.test(tabla, correct = TRUE);
> res1

        McNemar's Chi-squared test with continuity correction

data:  tabla
McNemar's chi-squared = 2.8846, df = 1, p-value = 0.08943

> res2 <- chisq.test(tabla,correct = TRUE);
> res2

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

data:  tabla
X-squared = 0.52181, df = 1, p-value = 0.4701
```

LECTURA 9: CUESTIÓN 2

Las clasificaciones de un grupo de 30 estudiantes de Ingeniería Informática han obtenido en las asignaturas de álgebra y programación en el curso 2017-2018 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

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 del 0.05.

```
> estudiante <- c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30)
> 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, 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)
>
> tabla <- matrix(c(estudiante,algebra,programacion),30,3, byrow = FALSE)
```

```
> colnames(tabla) <- c("Estudiante", "Algebra" , "Programacion")
> datos <- as.table(addmargins(tabla))
> kable(datos)
```

```
> ni <- datos[30,]; ni
  Estudiante      Algebra Programacion          Sum
          30            5            1           36
>
> nj <- datos[,4]; nj

 13.7  15.8   7.9  12.7  20.4  15.0  20.5  22.6  22.7  23.0  23.7

 23.8  28.9  25.7  27.8  35.0  31.7  27.6  30.3  35.6  35.6  37.8
                                                    Sum
 32.0  32.0  39.5  42.0  34.2  39.5  41.8  36.0 834.8
>
> N<- as.numeric(datos[30,4]); N
[1] 36
>
> pxy <- tabla^2
> suma <- 0
> for(i in 1:30) {
+   for(j in 1:4) {
+     suma <- suma+as.numeric(pxy[j,i]/(ni[i]*nj[j]))
+   }
+ }
```

```
> suma
[1] 11.44873
>
> chi2 <- N * (suma-1); chi2
[1] 376.1544
>
> g1 <- (nrow(tabla)-1)*(ncol(tabla)-1); g1
[1] 58
>
> qchisq(0.95,g1)
[1] 76.7778
>
> resultado1 <- chisq.test(tabla, correct = T); resultado1
```

```
        Pearson's Chi-squared test

data:  tabla
X-squared = 113.69, df = 58, p-value = 1.743e-05


>
> mean (algebra)
[1] 6.296552
> mean (programacion)
[1] 6.206667
```

Con este análisis podemos concluir que:

1. El alumno que presenta la mayor puntuación en la asignatura de álgebra es el número 17 con un 9.9.
2. El número con la nota más alta en la asignatura de programación es el número 13 con un 9.2.
3. Las medias en ambas asignaturas son muy similares, casi iguales, con una diferencia de 0.09.
4. Como la distribución de chi cuadrada tiene 58 como grado de libertad se puede concluir que se asemeja a una distribución normal.

LECTURA 9: CUESTIÓN 4

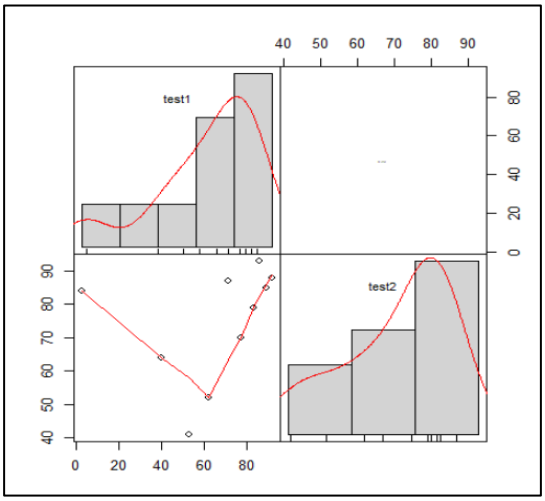
La puntuación de 10 estudiantes en dos pruebas psicológicas se detalla en la tabla siguiente. Calcular el coeficiente de correlación de Pearson y el coeficiente de Spearman para los rangos. ¿Qué conclusiones pueden extraerse de los resultados de ambos exponentes?

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

```
> estudiante<- c(A,B,C,D,E,F,G,H,I,J,L)
> 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)
> datos <- data.frame(test1, test2)
> chart.Correlation(datos)
> cor(test1,test2)
[1] 0.2907495
> cor.test (test1, test2)

        Pearson's product-moment correlation

data:  test1 and test2
t = 0.85949, df = 8, p-value = 0.4151
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.4148141  0.7779598
sample estimates:
      cor
0.2907495
```

```
> cor(test1, test2, method = "spearman")
[1] 0.6363636
> cor.test (test1, test2, method = "spearman")

        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
```

```
> shapiro.test(test1)

        Shapiro-Wilk normality test

data:  test1
W = 0.8637, p-value = 0.08437

> shapiro.test(test2)

        Shapiro-Wilk normality test

data:  test2
W = 0.88969, p-value = 0.1682
```

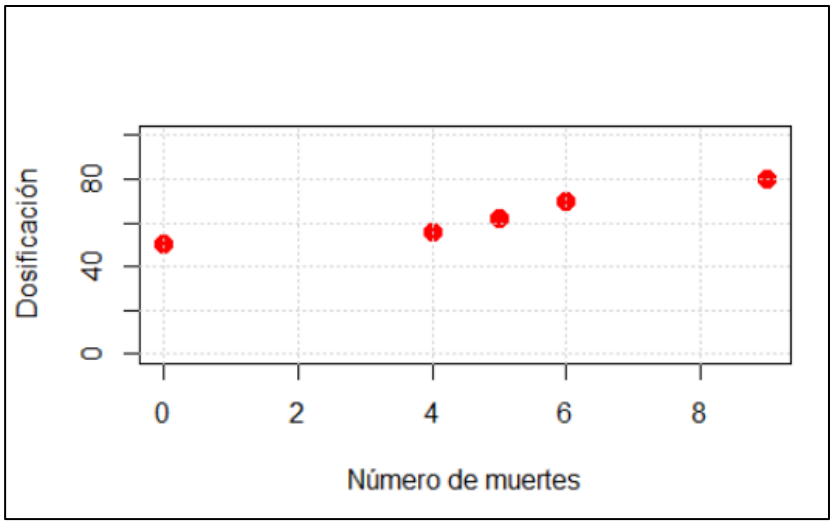


LECTURA 10: CUESTIÓN 1

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

```
> Y<-c(50,56,62,70,80)
> X<- c(0,4,5,6,9)
> plot(X, Y, ylim = c(0,100), pch = 19, col = "red",
+      xlab = "Número de muertes", ylab = "Dosificación", cex = 1.5)
> grid()
```



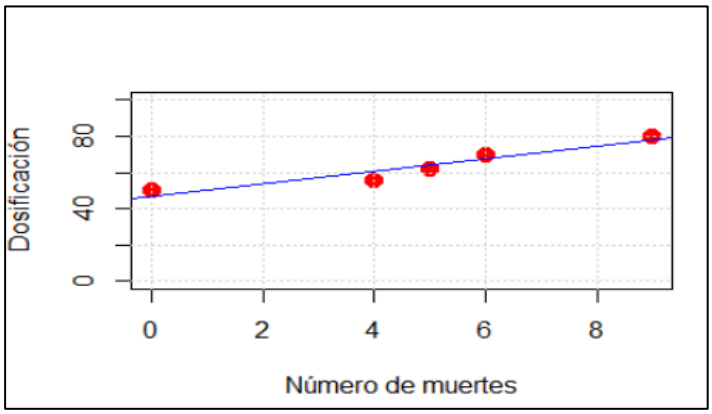
a) Realizar un análisis de regresión simple entre ambas variables.

Con la regresión simple tratamos de explicar la relación que existe entre la variable respuesta Y, y una única variable explicativa X.

Lo que debe ocurrir con el modelo de regresión lineal simple

- Debe existir una tendencia lineal. En el diagrama de dispersión no se debe dar un patrón no lineal.
- La varianza de los residuos ha de ser constante para todos los valores de la variable independiente.
- Las medidas de los residuos deben ser cero para cualquier valor de la variable independiente.

```
> 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
(Intercept)  47.0467     3.3715   13.95
X             3.4486     0.5998    5.75
            Pr(>|t|)
(Intercept) 0.000797 ***
X           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

> b0b1<-coefficients(modelo)
> confint(modelo)
                2.5 %   97.5 %
(Intercept) 36.317177 57.77628
X            1.539896  5.35730
> abline(modelo, col = "blue")
```



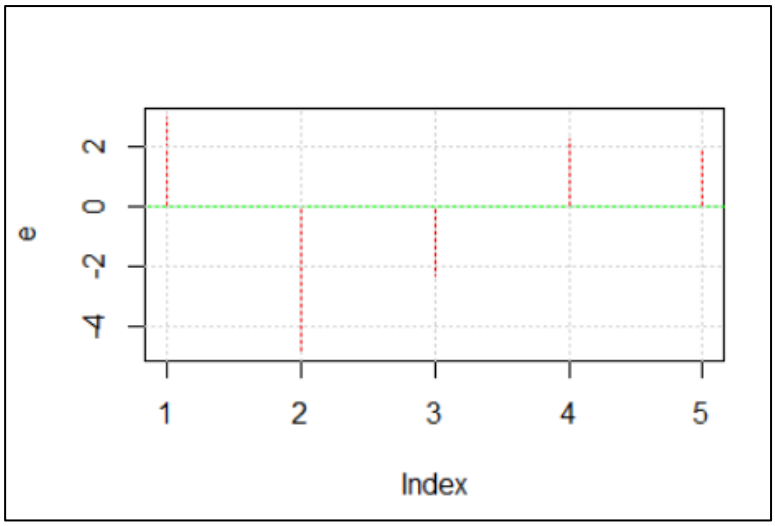
Efectivamente, podemos comprobar con la representación gráfica y los resultados obtenidos numéricamente, que entre ambas variables se da la regresión lineal simple.

Con un error residual de 3.924 con 3 grados de libertad, y un valor p-value = 0.01045

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.

Representación gráfica de la suma de cuadrados del error.

```
> e<-residuals(modelo)
> plot(e, type = "h", pch = 2, col = "red")
> abline(h=0, col = "green")
> grid()
```

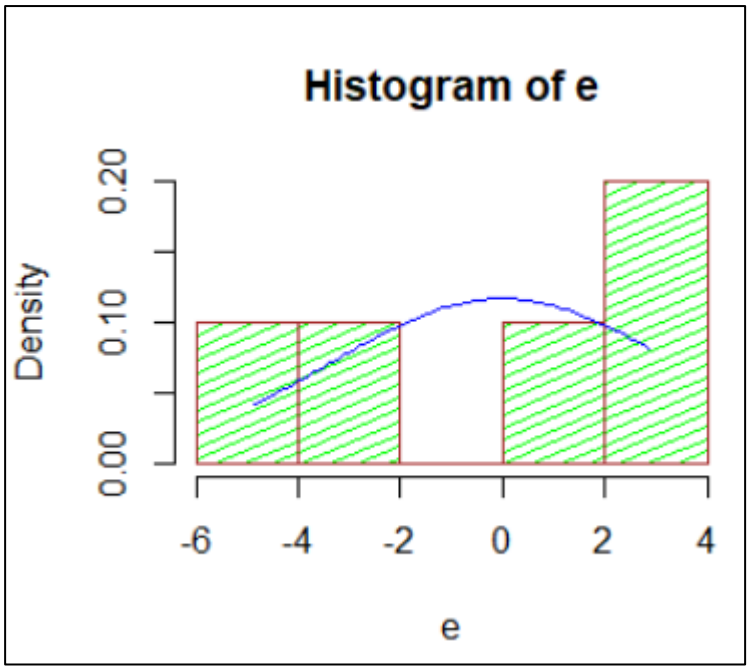


Realizamos diferentes pruebas para la falta de ajuste.

```
> shapiro.test(e)

        Shapiro-Wilk normality test

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

> ks.test(e,"pnorm")

        One-sample Kolmogorov-Smirnov test

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

> jarque.bera.test(e)

        Jarque Bera Test

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

Análisis gráfico de las relaciones y errores

```
> hist(e, freq = FALSE, 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")
```



c) Encontrar los intervalos de confianza para los coeficientes de regresión.

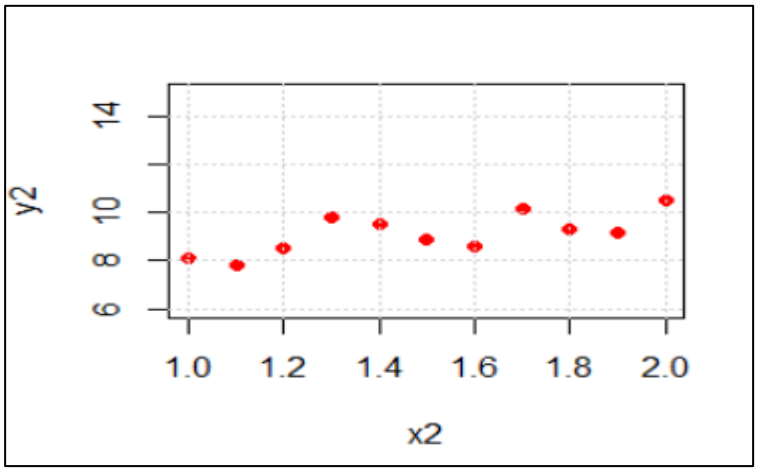
```
> confint(modelo)
                2.5 %   97.5 %
(Intercept) 36.317177 57.77628
X            1.539896  5.35730
> coefficients(modelo)
(Intercept)           X
  47.046729    3.448598
```

LECTURA 10: 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 25ºC y las cantidades de azúcar en miligramos. Los datos se codificaron y registraron como se indica en la siguiente tabla:

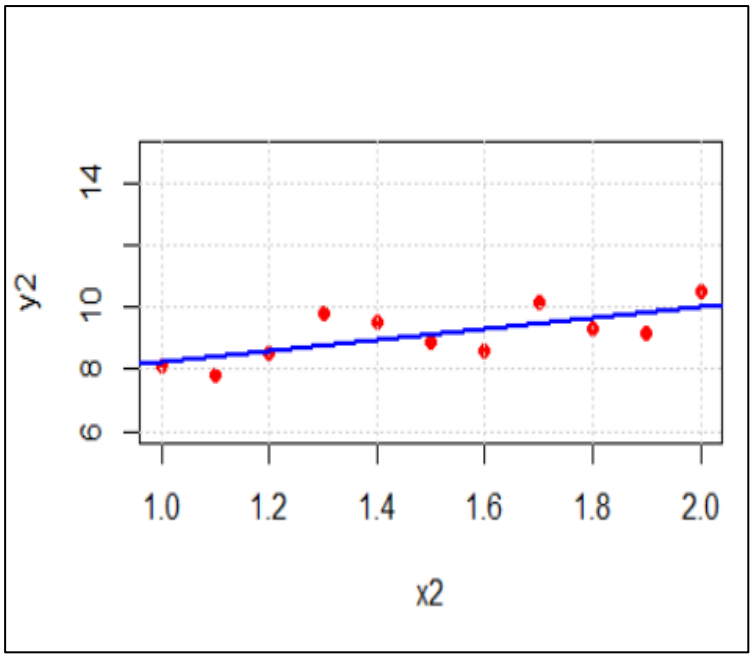
Num. Ensayo	Temperatura codificada *x* (base 25 ºC)	Azúcar convertida *y* (mg)
1	1.0	8.1
2	1.1	7.8
3	1.2	8.5
4	1.3	9.8
5	1.4	9.5
6	1.5	8.9
7	1.6	8.6
8	1.7	10.2
9	1.8	9.3
10	1.9	9.2
11	2.0	10.5

```
> y2<-c(8.1, 7.8, 8.5, 9.8, 9.5, 8.9, 8.6, 10.2, 9.3, 9.2, 10.5)
> x2<-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(x2)
> plot(x2,y2, ylim = c(6,15), pch = 19, col = "red")
> grid()
```



a) Realizar un análisis de regresión lineal simple de y con x.

```
> modelo2<-lm(y2~x2)
> b0b1_2<-coefficients(modelo2)
> abline(modelo2, col = "blue", lwd =2)
```

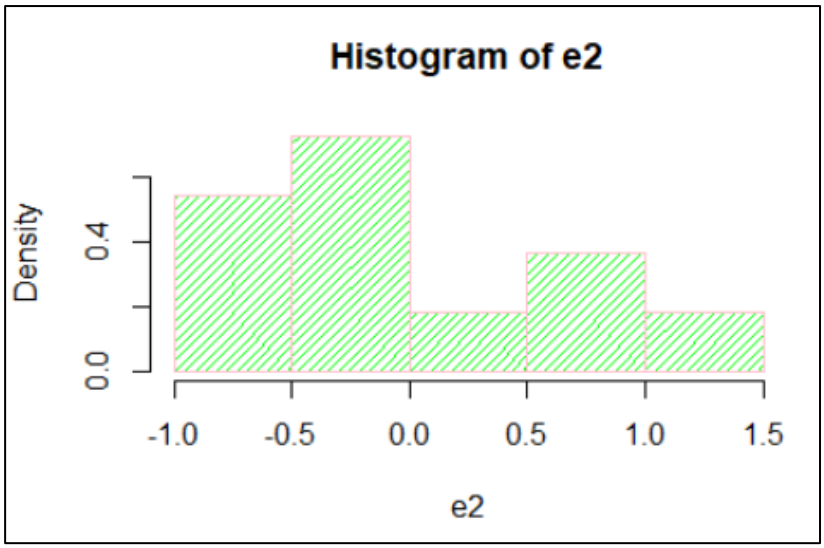


b) Calcular la suma de cuadrados del error y realizar una prueba para la falta de ajuste. Evaluar gráficamente las relaciones y los errores residuales correspondientes.

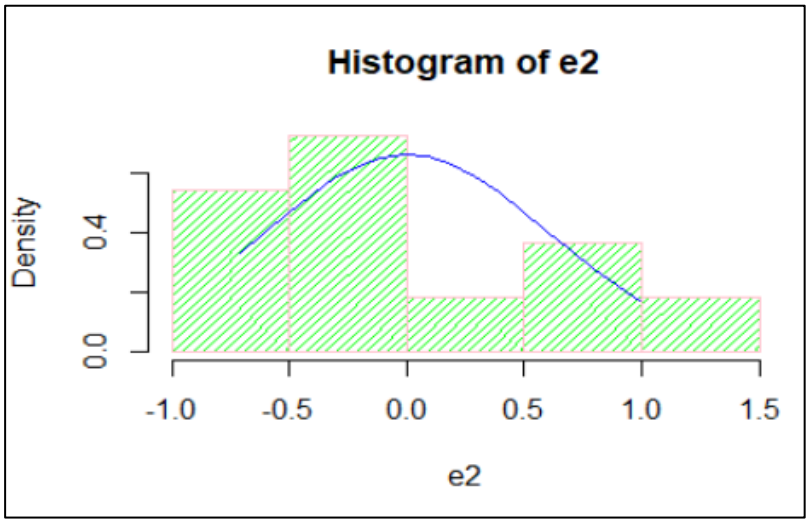
```
> e2<-residuals(modelo2)
> plot(e2, type= "h", lwd=2, col ="red")
> abline(h=0, col = "yellow")
> grid()
> shapiro.test(e2)

        Shapiro-Wilk normality test

data:  e2
W = 0.91869, p-value = 0.3079

> ks.test(e2, "pnorm")

        One-sample Kolmogorov-Smirnov test

data:  e2
D = 0.23942, p-value = 0.4814
alternative hypothesis: two-sided
```

```
> jarque.bera.test(e2)

        Jarque Bera Test

data:  e2
X-squared = 0.94717, df = 2, p-value = 0.6228
```

```
> hist(e2, freq = FALSE, col = "green", density = 25, border = "pink")
```




```
> valores<-seq(min(e2), max(e2), 0.1)
> points(valores, dnorm(valores, mean = mean(e2), sd = sd(e2)), type = "l", col ="blue")
```

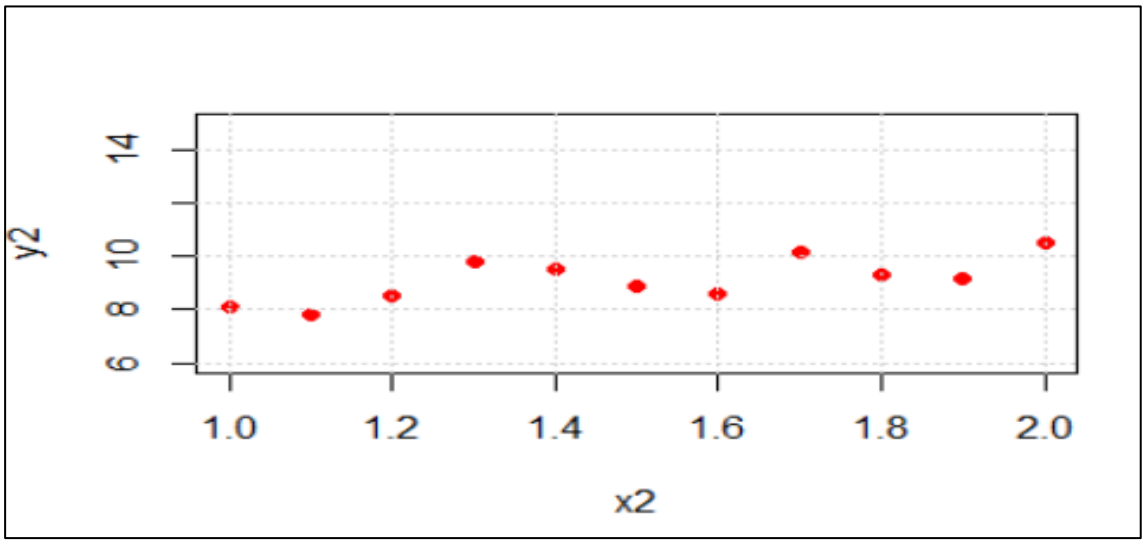


c) Encontrar los intervalos de confianza para los coeficientes de regresión.

```
> confint(modelo2)
                2.5 %   97.5 %
(Intercept) 4.3219598 8.505313
x2          0.4446316 3.173550
> coefficients(modelo2)
(Intercept)          x2
   6.413636    1.809091
```

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.

```
> prediccion<-predict(modelo2,newdata = data.frame(x2=1.75), interval = "pred")
> prediccion
       fit      lwr      upr
1 9.579545 8.046425 11.11267
> plot(x2,y2, ylim = c(6,15), pch = 19, col = "red")
> grid()
```

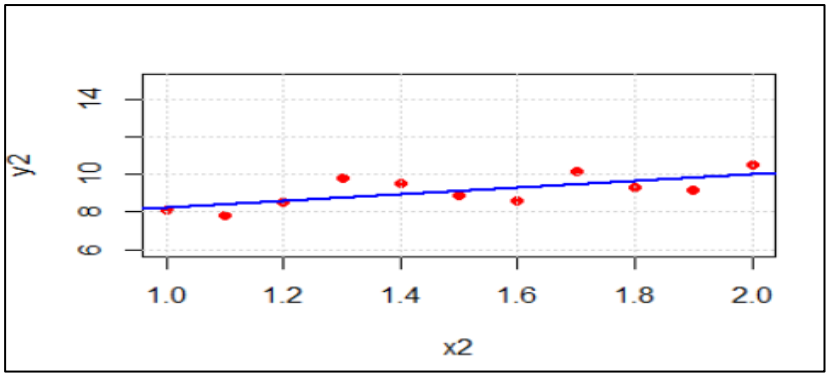


e) Definir el concepto de respuesta media y encontrar los intervalos de confianza para la misma en el apartado anterior.

```
> prediccion_med<-predict(modelo2, newdata = data.frame(x2=1.75, interval = "confidence"))
> prediccion_med
       1
9.579545
```

f) Visualizar los resultados de los apartados a), c), d) y e).

```
> abline (modelo2, col = "blue", lwd = 2)
> lines(c(1.75,1.75), c(prediccion_med[2], prediccion_med[3]), col="brown")
> points(c(1.75,1.75), c(prediccion_med[2], prediccion_med[3]), col="brown")
```



LAB 6: 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 significancia del 5%.

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

```
> library(MASS)
> library(vcd)
> num_dias <- c(41,81,87,54,30,12,3)
> num_sol <- c(0,1,2,3,4,5,6)
> var_poisson <- data.frame(num_sol,num_dias)
> var_p <- c(rep(num_sol,num_dias))
```

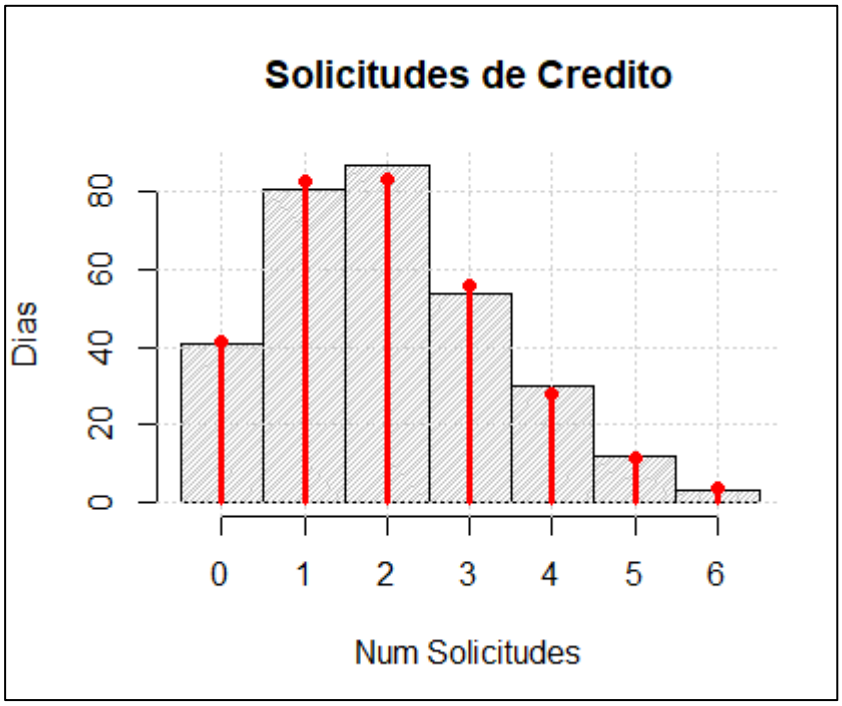
```
> ajuste.poisson <- goodfit(var_p,type="poisson",method
="MinChisq")
> summary(ajuste.poisson)

         Goodness-of-fit test for poisson distribution

              X^2 df P(> X^2)
Pearson 1.385323  5 0.925912
> |
```

Obtenemos, por tanto, que el valor del estadístico de contrates es 1.39, los grados de libertad son 5 y el p-valor es 0.93. Ese p-valor nos permite aceptar la hipótesis nula.

```
> ajuste.poisson$par
$lambda
[1] 2.007465

> ajuste.poisson

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

 count observed    fitted pearson residual
     0       41 41.373253      -0.05802883
     1       81 83.055365      -0.22553039
     2       87 83.365376       0.39807660
     3       54 55.784363      -0.23890592
     4       30 27.996291       0.37869044
     5       12 11.240316       0.22659159
     6        3  3.760757      -0.95958340
```

El ajuste por el método de la mínima χ 2 es una Poisson (2.007).

```
> hist(var_p,breaks=-0.5:6.5,xlab="Num Solicitudes",ylab="Dias",main="Solicitudes de Credito",
+      col="gray",border="black",density=50)
> grid()
> points(0:6,ajuste.poisson$fitted,type="h",lwd=3,col="red")
> points(0:6,ajuste.poisson$fitted,pch=19,col="red")
```



Efectivamente como se puede comprobar con la representación gráfica la distribución que presentan las solicitudes de crédito siguen una distribución de Poisson.

LAB 6: 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 significancia del 5%.

```
> library(knitr)
> Tabla <- matrix(c(34,73,63,16,12,12),2,3,byrow=T)
> colnames(Tabla) <- c("Fertilizante_A","Fertilizante_B","Fertilizante_C")
> rownames(Tabla) <- c("Florecido","No_Florecido")
> Tabla <- as.table(Tabla)
> kable(Tabla)


|             | Fertilizante_A| Fertilizante_B| Fertilizante_C|
|:------------|--------------:|--------------:|--------------:|
|Florecido    |             34|             73|             63|
|No_Florecido |             16|             12|             12|
```

A continuación, se formará la tabla ampliada que incluye una columna con la suma de los valores conseguidos con cada fertilizante de forma individualmente.

```
> Tabla_ampliada <- addmargins(Tabla)
> kable(Tabla_ampliada)


|             | Fertilizante_A| Fertilizante_B| Fertilizante_C| Sum|
|:------------|--------------:|--------------:|--------------:|---:|
|Florecido    |             34|             73|             63| 170|
|No_Florecido |             16|             12|             12|  40|
|Sum          |             50|             85|             75| 210|
>
```

```
> ni <- Tabla_ampliada[3,]
> nj <- Tabla_ampliada[,4]
> N <- as.numeric(Tabla_ampliada[3,4])
> ni <- Tabla_ampliada[3,]
> nj <- Tabla_ampliada[,4]
> N <- as.numeric(Tabla_ampliada[3,4])
>
> pXY <- Tabla^2
> suma <- 0
> for (i in 1:3){
+   for (j in 1:2){
+     suma <- suma+as.numeric(pXY[1,1]/(ni[i]*nj[j]))
+   }
+ }
> CHI2 <- N*(suma-1)
> CHI2
[1] 128.1
```

Grados de libertad y región crítica

```
> gl <- (nrow(Tabla)-1)*(ncol(Tabla)-1)
> qchisq(0.95,gl)
[1] 5.991465
>
```

Como el valor 7.232 es mayor que el valor límite 5.991, el estadístico está dentro de la RC y se rechaza, por tanto, la hipótesis de independencia.

```
> resultado <- chisq.test(Tabla,correct=T)
> resultado

        Pearson's Chi-squared test

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

El estadístico toma el valor 7.232 y el p-value el valor límite 0.0269, que es inferior a 0.05, luego se rechaza la hipótesis de independencia.

LAB 6: 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(knitr)
> Tabla <- matrix(c(75,73,15,32),2,2,byrow=T)
> colnames(Tabla) <- c("Estudiantes II","Otros titulos")
> rownames(Tabla) <- c("Mas de 2 horas","Menos de 2 horas")
> Tabla <- as.table(Tabla)
> kable(Tabla)


|                 | Estudiantes II| Otros titulos|
|:----------------|--------------:|-------------:|
|Mas de 2 horas   |             75|            73|
|Menos de 2 horas |             15|            32|
```

A continuación, se formará la tabla ampliada que incluye una columna con la suma de los valores conseguidos con los alumnos de Ingeniería Informática y los de otros títulos.

```
> Tabla_ampliada <- addmargins(Tabla)
> kable(Tabla_ampliada)


|                 | Estudiantes II| Otros titulos| Sum|
|:----------------|--------------:|-------------:|---:|
|Mas de 2 horas   |             75|            73| 148|
|Menos de 2 horas |             15|            32|  47|
|Sum              |             90|           105| 195|
>
```

```
> ni <- Tabla_ampliada[3,]
> nj <- Tabla_ampliada[,3]
> N <- as.numeric(Tabla_ampliada[3,3])
>
> Tabla_Esperada <- Tabla
> suma <- 0
> for (i in 1:2){
+   for(j in 1:2){
+     Tabla_Esperada[i,j] <- (ni[j]*nj[i])/N
+     suma <- suma+((abs(Tabla[i,j]-Tabla_Esperada[i,j])-0.5)^2)/Tabla_Esperada[i,j]
+   }
+ }
> kable(Tabla_Esperada)


|                 | Estudiantes II| Otros titulos|
|:----------------|--------------:|-------------:|
|Mas de 2 horas   |       68.30769|      79.69231|
|Menos de 2 horas |       21.69231|      25.30769|
> CHI2 <- suma
> CHI2
[1] 4.325313
```

Grados de libertad y región crítica

```
> gl <- (nrow(Tabla)-1)*(ncol(Tabla)-1)
> qchisq(0.95,gl)
[1] 3.841459
```

Como el valor 4.325 es mayor que el valor limite 3.841, el estadístico está dentro de la RC y se rechaza la hipótesis de independencia.

```
> resultado <- chisq.test(Tabla,correct=T)
> resultado

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

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

El estadístico toma el valor 4.325 y el p-value valor límite 0.03755 que es inferior a 0.05, luego se rechaza la hipótesis de independencia.

LAB 6: 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.

```
> library(knitr)
> Tabla <- matrix(c(103,12,18,35),2,2,byrow=T)
> colnames(Tabla) <- c("Brazo_Fracturado_Jefe","Brazo_Normal_Jefe")
> rownames(Tabla) <- c("Brazo_Fracturado_Interno","Brazo_Normal_Interno")
> Tabla <- as.table(Tabla)
> kable(Tabla)


|                        | Brazo_Fracturado_Jefe| Brazo_Normal_Jefe|
|:-----------------------|---------------------:|-----------------:|
|Brazo_Fracturado_Interno |                   103|                12|
|Brazo_Normal_Interno     |                    18|                35|
>
```

a) Explicar la aplicación del test de McNemar para tablas de contingencia que tengan que ver con los resultados de dos pruebas sobre los mismos individuos.

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

        McNemar's Chi-squared test with continuity correction

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

En este caso hemos elegido la corrección por continuidad o prueba exacta de Fisher. El p-value toma un valor de 0.3613.

La prueba exacta de Fisher se utiliza para contrastar la relación entre las dos variables cualitativas que constituyen la tabla de contingencia, pero no admite la posibilidad de contrastar si una proporción es mayor que en otra.

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?

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

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

data:  Tabla
X-squared = 52.941, df = 1, p-value = 3.437e-13
```

Son independientes ambos casos.

LAB 6: EJERCICIO 5

Se llevaron a cabo las pruebas con tres tratamientos para una enfermedad infecciosa leve sobre tres grupos de pacientes. Además, se incluyó un grupo adicional, al cual se le suministró una medicación placebo. 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 Krustal-Walls.

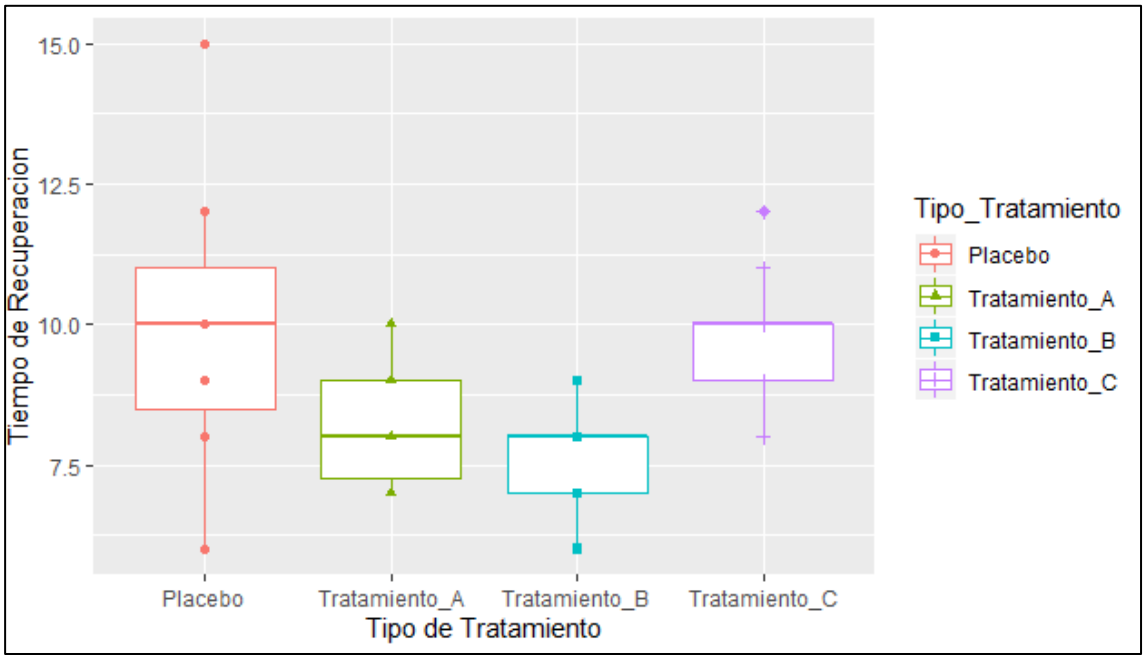
```
> library(knitr)
> library(ggplot2)
> P <- c(15,12,10,8,9,6,10)
> A <- c(7,8,9,8,7,10,9,8,7,10)
> B <- c(8,9,8,6,7,8,9,8,7,6)
> C <- c(10,12,10,8,9,11,10,9,8)
```

Creamos la concatenación de los resultados obtenidos con los diferentes tratamientos de la enfermedad.

```
> tratamientos_f <- factor(rep(1:4,c(length(P),length(A),length(B),length(C))),
+                          labels=c("Placebo","Tratamiento_A","Tratamiento_B","Tratamiento_C"))
> tratamientos_v <- c(P,A,B,C)
> datos_tratamientos <- as.data.frame(tratamientos_v)
> datos_tratamientos[,2] <- tratamientos_f
> names(datos_tratamientos) <- c("Tiempo_Recuperacion","Tipo_Tratamiento")
> tabla <- table(datos_tratamientos)
> tabla
                  Tipo_Tratamiento
Tiempo_Recuperacion 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
```

Comandos necesarios para la representación como cuartiles de los diferentes tratamientos y el placebo

```
> g <- ggplot(data = datos_tratamientos,aes(x=Tipo_Tratamiento,y=Tiempo_Recuperacion,
+                                           color=Tipo_Tratamiento))
> g+geom_boxplot()+xlab("Tipo de Tratamiento")+ylab("Tiempo de Recuperacion")+geom_point(aes(shape=
+                                                                                           Tipo_Tratamiento))
```




```
> kruskal.test(Tiempo_Recuperacion,Tipo_Tratamiento,datos_tratamientos)

        Kruskal-Wallis rank sum test

data:  Tiempo_Recuperacion and Tipo_Tratamiento
Kruskal-Wallis chi-squared = 10.697, df = 3, p-value = 0.01348

>
> wilcox.test(Tiempo_Recuperacion[Tipo_Tratamiento=="Placebo"],
+             Tiempo_Recuperacion[Tipo_Tratamiento=="Tratamiento_A"],
+             alternative="greater")

        Wilcoxon rank sum test with continuity correction

data:  Tiempo_Recuperacion[Tipo_Tratamiento == "Placebo"] and Tiempo_Recuperacion[Tipo_Tratamiento == "Tratamiento_
A"]
W = 49.5, p-value = 0.08222
alternative hypothesis: true location shift is greater than 0
```

LAB 7: EJERCICIO 1

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

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

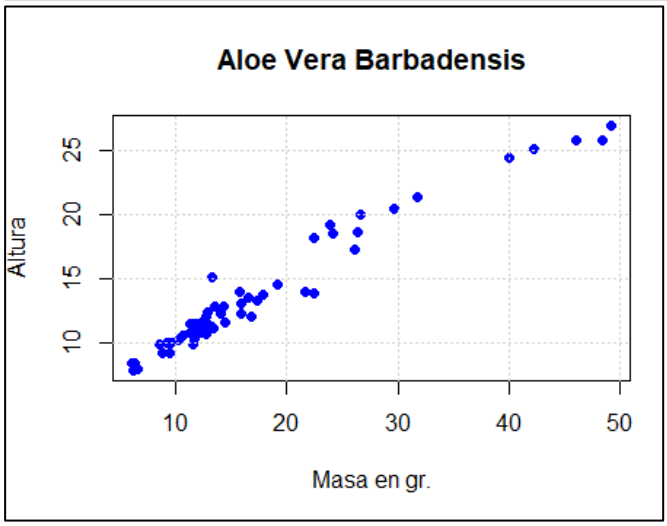
a) *Estudiar las variedades que dan más rendimiento desde el punto de vista de su masa y masa seca.*

```
> aggregate(Aloe$Masa~Aloe$Variedad,Aloe,mean)
  Aloe$Variedad Aloe$Masa
1    brevifolia  4.412500
2   arborescens 18.081481
3   barbadensis 17.576786
4     saponaria  7.521429
> aggregate(Aloe$Masa_Seca~Aloe$Variedad,Aloe,mean)
  Aloe$Variedad Aloe$Masa_Seca
1    brevifolia       1.010145
2   arborescens       6.350000
3   barbadensis       6.110714
4     saponaria       1.650000
```

La que dan más rendimiento son las barbadensis y las arborescens.

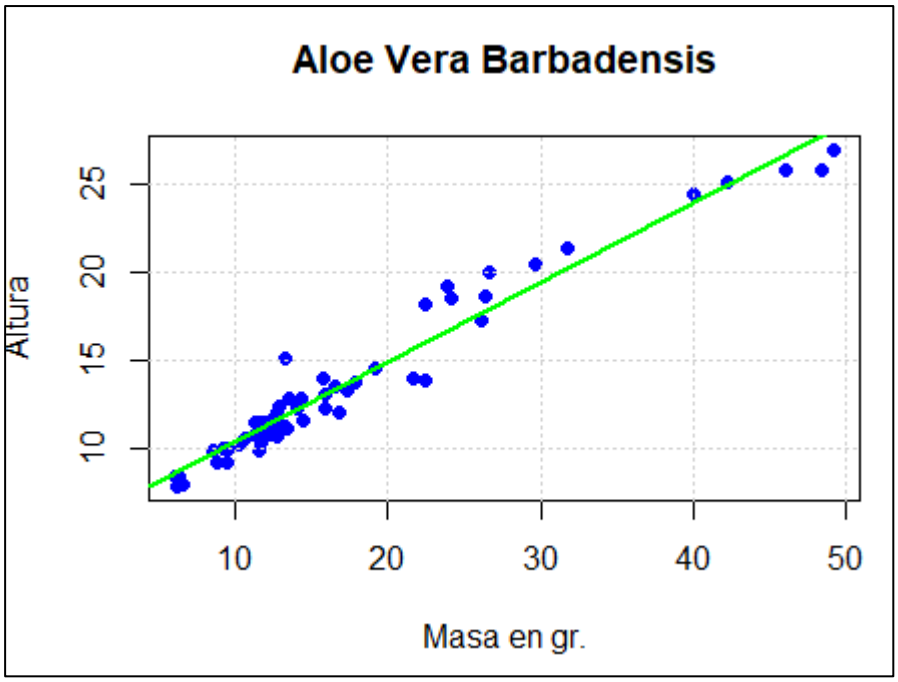
b) Analizar las dependencias entre la masa y la altura de la variedad "bardadensis".

```
> Barbadensis <- subset(Aloe,subset=(Aloe$Variedad=="barbadensis"))
> plot(Altura~Masa, data=Barbadensis,pch=19,col="blue",xlab="Masa en gr.",ylab="Altura",
+      main="Aloe Vera Barbadensis")
> grid()
```



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

```
modelo1 <- lm(Altura~Masa,data=Barbadensis)
abline(modelo1,col="green",lwd=2)
```



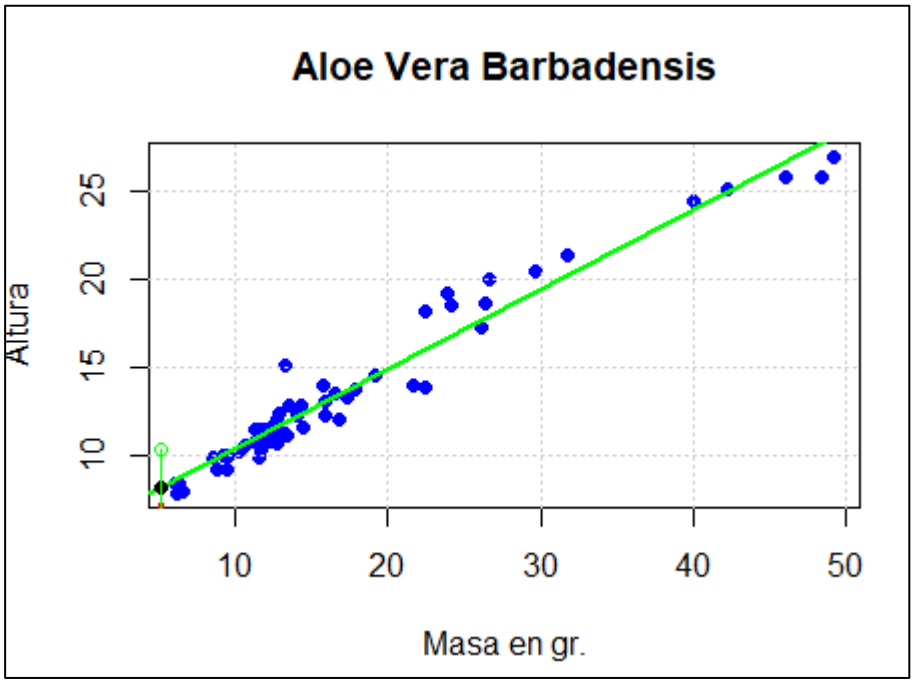
d) Analizar el modelo estimado con la función summary y obtener un posible intervalo de confianza para las conclusiones de los distintos parámetros.

```
> summary(modelo1)

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

> confint(modelo1)
                2.5 %   97.5 %
(Intercept) 5.1883011 6.308774
Masa        0.4290643 0.483832
```

En intervalo de confianza entre 9,42 y 9,48.

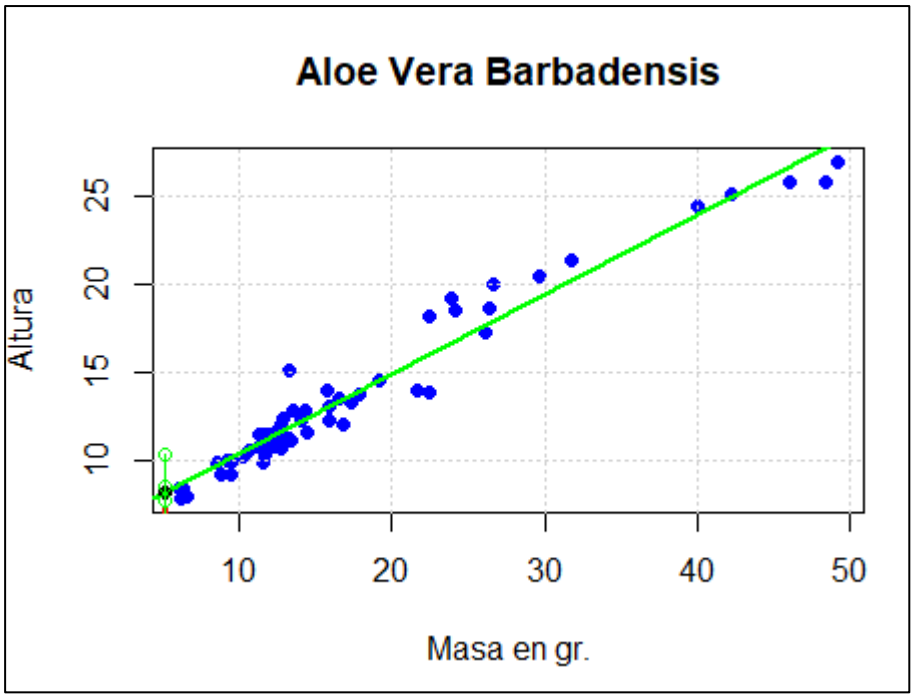
e) Evaluar una predicción para una masa de x0=5.1 gramos y encontrar un intervalo de confianza para la misma.

```
> x0 <- 5.1
> 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)
>
> inter_prediccion2 <- predict(modelo1,level=0.95,newdata=data.frame(Masa=x0),
+                              interval="pred")
> inter_prediccion2
       fit      lwr      upr
1 8.076423 5.885093 10.26775
> lines(c(x0,x0),c(inter_prediccion2[2],inter_prediccion2[3]),col="green")
> points(c(x0,x0),c(inter_prediccion2[2],inter_prediccion2[3]),col="green")
```



El intervalo es [5.88, 10.27]

```
> inter_prediccion3 <- predict(modelo1,level=0.95,newdata=data.frame(Masa=x0),
+                              interval="confidence")
> inter_prediccion3
       fit      lwr      upr
1 8.076423 7.630408 8.522438
> lines(c(x0,x0),c(inter_prediccion3[2],inter_prediccion3[3]),col="green")
> points(c(x0,x0),c(inter_prediccion3[2],inter_prediccion3[3]),col="green")
```



El intervalo de respuesta media es [7.63 ,8.52]

f) Encontrar el coeficiente de determinación R2.

```
> summary(modelo1)

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
```

El coeficiente de determinación R2 es igual a 0.9539

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

El estadístico toma un valor de 1117 y un p-value 2.2e-16 lo que indica que hay una relación muy fuerte, y toma la forma de una recta.

```
> x_factor <- as.factor(Barbadensis$Masa)
> n <- length(x_factor)
> datos2 <- data.frame(x_factor,Barbadensis$Altura)
> Y_M_F <- rep(0,n)
> for(i in 1:n){
+   Y_M_F[i] <- mean(Barbadensis$Altura[x_factor==Barbadensis$Masa[i]])
+ }
> SCE_Puro <- sum((Barbadensis$Altura-Y_M_F)^2)
> SCE_Puro
[1] 13.51667
> SC_Falta_Ajuste <- 61.8-SCE_Puro
> k <- nlevels(x_factor)
> S_2_Puro <- SCE_Puro/(n-k)
> S_2_Puro
[1] 1.501852
> F_SC_Falta_Ajuste <- SC_Falta_Ajuste/(S_2_Puro*(k-2))
> F_SC_Falta_Ajuste
[1] 0.7144266
>
> 1-pf(F_SC_Falta_Ajuste,1,k-2)
[1] 0.40245
```

En un porcentaje del 40%, se encuentra bien ajustado, el modelo sigue bien el proceso.